






Universitat Autònoma de Barcelona

ADVERTIMENT. L'accés als continguts d'aquesta tesi queda condicionat a l'acceptació de les condicions d'ús establertes per la següent llicència Creative Commons:  http://cat.creativecommons.org/?page_id=184

ADVERTENCIA. El acceso a los contenidos de esta tesis queda condicionado a la aceptación de las condiciones de uso establecidas por la siguiente licencia Creative Commons:  <http://es.creativecommons.org/blog/licencias/>

WARNING. The access to the contents of this doctoral thesis it is limited to the acceptance of the use conditions set by the following Creative Commons license:  <https://creativecommons.org/licenses/?lang=en>



Universitat Autònoma de Barcelona
Escola d'Enginyeria

Departament d'Arquitectura de Computadors i Sistemes Operatius (DACSO)

Doctorat en Bioinformàtica

Large scale quantitative assessment of biodiversity using next-generation sequencing

Lidia Garrido Sanz

Tesi doctoral

Octubre 2021

Directors

Dr. Josep Piñol Pascual

Dr. Miquel Àngel Senar Rosell

Tutor

Dr. Mario Cáceres Aguilar

This thesis has been funded by the following grants: Spanish Government grant TIN2017-84553-C2-1-R and Agència de Gestió d'Ajuts Universitaris i de Recerca (AGAUR) grant 2017-SGR-1001

Large scale quantitative assessment of biodiversity using next-generation sequencing

Thesis dissertation presented by Lidia Garrido Sanz for the degree of Doctor of Bioinformatics at the Computing Architecture and Operating Systems (CAOS) Department at the Autonomous University of Barcelona.

Directors:

- Dr. Josep Piñol Pascual
- Dr. Miquel Àngel Senar Rosell

Tutor:

- Dr. Mario Cáceres Aguilar

Acknowledgements

Acknowledgements

I would like to express my sincere gratitude to my PhD supervisors Josep Piñol and Miquel Àngel Senar, for their mentorship and guidance throughout this thesis. I would also like to thank them for giving me the opportunity to do this research. I am also grateful to my PhD tutor, Mario Cáceres, for his valuable advice wherever necessary.

I am extremely grateful to my current and past laboratory colleagues, Jordi Alcaraz, Carles Carrillo, Pau Cebrian, Dani Hernández, Pilar Gómez, Betzabeth León, Verónica Vidal, Felipe Tirado and Àlvaro Wong for their support and coffee chats that gave me strength to complete this thesis work. I cannot express enough thanks to Eduardo César, Gemma Roquet, Dani Ruíz, Ania Sikora and Sergio Villar for the invaluable help when the times got rough. And a huge thank you to every and single one in CAOS department, it was a big pleasure to share this experience with you all.

I am also grateful to all entomologists that provided the specimens used to create the insects DNA libraries, Francisco Beitia, Jacobus J. Boomsma, Luis Calcaterra, Xavier Espadaler, Carlos Hernández-Castellano, Joan Josep Ibañez, Rasmus S. Larsen, Francesc Mestres, Misato O. Miyakawa, Nicolás Pérez, Carlos Pradera, Alfredo Ruíz, and Aleix Valls. I also want to thank Anna Barceló and Roger Lahoz of the Genomics facilities of the UAB for the preparation and sequencing of the DNA libraries.

I also would like to thank the staff of the Australian Antarctic Division and in particular Dr. Bruce E. Deagle to allow me the opportunity to work with them and learn about the metabarcoding method. Similar, I also thank Alfried P. Vogler for the opportunity of collaborating with his team, albeit the stay was cancelled due to the COVID-19 pandemic.

Last but not the least, a huge thank you to my parents, sister, and niece, and also my friends, Núria, Deby, Gabriel, Mónica and Anna for your great support.

Abstract

Abstract in English

Molecular-based methods for the study of biological communities are widely applied today. For eukaryotes, the dominant technique is DNA metabarcoding. It relies on the PCR-amplification of one or a few genomic regions, so-called barcodes. However, the PCR step introduces biases that make difficult to recover the original relative abundance of species in complex mixtures. All PCR-biases can be avoided by shotgun sequencing all the DNA of a sample and comparing the reads to whole genomes (metagenomics) or mitochondrial genomes (mito-metagenomics). Metagenomic methods are currently unfeasible for *real* samples due to the low number of eukaryotes with sequenced genomes, but this situation will probably improve in the future. To explore the capabilities of metagenomic methods with reference databases containing the genomes of most species, we simulated such a future using *artificial* samples of insect species with known genomes.

First, we explored metagenomics and found that the method was perfectly able to recover the species identity and the relative species abundance (RSA). However, an analytical detection limit was needed to reduce the long list of low-abundant false positive species.

Next, we evaluated the mito-metagenomics method; this method is already being applied today, but the estimations are far from perfect despite the avoidance of the PCR step. Results showed that mito-metagenomics recovered all the species in the mixtures with just a few false positives species and robustly estimated the *within*-species RSA (is species i more abundant in sample s than in sample r ?). However, the *across*-species RSA (is species i more abundant than species j in sample s ?) was only correctly estimated when a species-specific correction factor accounting for the mitochondrial copy number was applied.

Finally, we explored the problem of detecting false positive species on the results attributable to the metagenomic classifiers. To this end, we challenged two popular metagenomic classifiers (*i.e.*, BLASTn followed by MEGAN6, and Kraken2) to identify

species in single-species samples using mito-metagenomics. The results showed that when the classifiers were used with default parameters, they reported many false positive species. However, most wrongly assigned species were eliminated by the intersection of the results from both classifiers plus an analytical detection limit.

In summary, this thesis provides an overview of the advantages and limitations of PCR-free metagenomic methods to explore the biodiversity of eukaryotes in complex samples once the genomic repositories contain the genomes of most species.

Resum (Abstract in Catalan)

Actualment, els mètodes moleculars són aplicats àmpliament en l'estudi de comunitats biològiques. Pels organismes eucariotes, la tècnica dominant és el *DNA metabarcoding*. Aquesta tècnica es basa en l'amplificació amb PCR d'una o varies regions del genoma, anomenades *barcodes*. Malauradament, la PCR introdueix biaixos que dificulten l'estimació de l'abundància relativa original de les espècies en mostres complexes. Els biaixos lligats a la PCR es poden evitar mitjançant la seqüenciació aleatòria de tot l'ADN de la mostra (*shotgun sequencing*) i comparant les seqüències obtingudes amb genomes sencers (metagenòmica) o genomes mitocondrials (mito-metagenòmica). El mètode metagenòmic no és factible actualment degut al baix nombre d'espècies eucariotes amb el genoma complet seqüenciat, tot i que aquesta situació sembla que millorarà en el futur. Amb l'objectiu d'explorar les capacitats dels mètodes metagenòmics quan les bases de dades de referència continguin els genomes de la majoria de les espècies, vam simular aquest futur amb mostres *artificials* d'insectes dels quals ja es coneix el seu genoma.

En primer lloc, vam explorar el mètode *metagenòmic* i vam observar que el mètode va ser capaç de recuperar la identitat i l'abundància relativa de les espècies (RSA). De totes maneres, va ser necessari aplicar un límit de detecció per a reduir la llarga llista d'espècies fals-positives i amb baixa concentració.

A continuació vam avaluar el mètode mito-metagenòmic; aquest mètode ja s'aplica avui en dia, però les estimacions són lluny de ser perfectes malgrat l'absència de la PCR. Els resultats van mostrar que el mètode mito-metagenòmic va recuperar totes les espècies en les mostres de barreges d'ADN amb l'addició d'alguns falsos positius i va estimar amb robustesa la RSA dintre de la mateixa espècie (*within-species* RSA; l'espècie *i* és més abundant a la mostra *s* que a la mostra *r*?). En canvi, l'abundància entre diferents espècies (*across-species* RSA; l'espècie *i* és més abundant que l'espècie *j* a la mostra *s*?) només es va recuperar després d'aplicar un factor de correcció específic per a cada espècie que inclou el número de còpies d'ADN mitocondrials.

Finalment, vam explorar el problema de la detecció d'espècies fals-positives als resultats atribuïbles als classificadors metagenòmics. Amb aquest objectiu, vam utilitzar dos classificadors metagenòmics populars (*i.e.*, BLASTn seguit de MEGAN6, i Kraken2) per identificar espècies en mostres que contenen una única espècie mitjançant el mètode mito-metagenòmic. Els resultats van mostrar que quan els classificadors metagenòmics s'utilitzen amb els paràmetres per defecte, aquest retornen moltes espècies fals-positives. No obstant això, la majoria de les espècies assignades erròniament van ser eliminades amb la intersecció dels resultats d'ambdós classificadors i l'addició d'un límit de detecció analític.

En resum, aquesta tesi proporciona una visió general dels avantatges i limitacions dels mètodes metagenòmics independents de la PCR per explorar la biodiversitat dels organismes eucariotes en mostres complexes un cop les bases de dades genètiques continguin els genomes de la majoria de les espècies.

Resumen (Abstract in Spanish)

Actualmente, los métodos moleculares se aplican ampliamente para el estudio de comunidades biológicas. En organismos eucariotas, la técnica predominante es el *DNA metabarcoding*. Esta técnica se basa en la amplificación con PCR de una o varias regiones del genoma llamadas *barcodes*. Sin embargo, la PCR introduce sesgos que dificultan la recuperación de la abundancia relativa original de las especies en muestras complejas. Los sesgos atribuibles a la PCR se pueden evitar mediante la secuenciación aleatoria de todo el ADN de la muestra (*shotgun sequencing*) y comparando las secuencias contra genomas completos (metagenómica) o genomas mitocondriales (mito-metagenómica). El método metagenómico no es viable actualmente debido al número reducido de especies eucariotas con el genoma completo secuenciado, aunque esta situación parece que mejorará en el futuro. Con el objetivo de explorar las capacidades de los métodos metagenómicos cuando las bases de datos de referencia almacenen el genoma de la mayoría de las especies, nosotros simulamos dicho futuro trabajando con muestras *artificiales* de insectos cuyo genoma ya se conoce.

En primer lugar, exploramos el método metagenómico y observamos que fue capaz de recuperar la identidad y la abundancia relativa de las especies (RSA). De todas formas, fue necesario un límite de detección analítico para reducir la larga lista de especies falso-positivas presentes en concentraciones bajas.

A continuación, evaluamos el método mito-metagenómica; este método se aplica actualmente, pero las estimaciones están lejos de ser perfectas, aunque no se utilice la PCR. Los resultados mostraron que el método mito-metagenómico pudo recuperar todas las especies en las muestras de mezclas de ADN, aunque con algunos falsos positivos y también estimó con robustez la RSA dentro de una misma especie (*within-species* RSA; ¿la especie i es más abundante en la muestra s que en la muestra r ?). Pero la abundancia entre diferentes especies (*across-species* RSA; ¿la especie i es más abundante que la especie j en la muestra s ?) sólo se recuperó tras aplicar un factor de

corrección específico para cada especie que incluye el número de copias de ADN mitocondrial.

Para terminal, exploramos el problema de la detección de especies falso-positivas en los resultados atribuibles a los clasificadores metagenómicos. Con este objetivo, utilizamos dos clasificadores metagenómicos populares (*i.e.*, BLASTn seguido de MEGAN6, y Kraken2) para identificar especies en muestras que contienen una única especie mediante el método mito-metagenómico. Los resultados mostraron que cuando los clasificadores metagenómicos se utilizan con los valores por defecto de los parámetros, se recuperan muchas especies falso-positivas. No obstante, la mayoría de las especies asignadas erróneamente fueron eliminadas mediante la intersección de los resultados de ambos clasificadores y un límite de detección analítico.

En resumen, esta tesis proporciona una visión general de las ventajas y limitaciones de los métodos metagenómicos libres de PCR para explorar la biodiversidad de organismos eucariotas en muestras complejas cuando las bases de datos genéticas almacenen el genoma de la mayoría de las especies.

Table of contents, lists of figures, tables, abbreviations, and symbols

Table of Contents

Acknowledgements	i
Abstract	v
Abstract in English.....	vii
Resum (Abstract in Catalan)	ix
Resumen (Abstract in Spanish)	xi
Table of contents, lists of figures, tables, abbreviations, and symbols	xiii
Table of Contents.....	xv
List of Figures	xix
List of Tables	xxi
List of Abbreviations	xxiii
List of Symbols	xxv
1. Introduction.....	1
1.1. Introduction to biodiversity assessment	3
1.1.1. Classical morphology-based identification.....	4
1.1.2. Molecular-based identification	5
1.2. From single species to multiple species assessment with DNA barcodes.....	7
1.2.1. DNA barcoding.....	7
1.2.2. DNA metabarcoding	10
1.3. Metagenomics	11
1.3.1. A brief history of the metagenomic method.....	11
1.3.2. Metagenomics in eukaryotes	14
1.3.3. Mitochondrial metagenomics	15
1.4. Bioinformatic considerations.....	18
1.4.1. Reference databases	18
1.4.2. Bioinformatic tools and pipelines.....	20
1.5. Insects as model organisms	21
1.6. Thesis hypothesis and goals.....	22
1.6.1. Defining the problem and thesis hypothesis.....	22
1.6.2. Research goals	24
2. Methodology.....	27

2.1.	General pipeline.....	29
2.2.	Reference genomes: Whole genomes and mitochondrial genomes	30
2.3.	Preparation of samples: Selection of the species, laboratory treatment and quality control.....	30
2.4.	Classification of reads to species: Matching and assignment steps.....	33
2.4.1.	BLASTn plus MEGAN6.....	33
2.4.2.	Kraken2.....	34
2.4.3.	BWA plus γ - δ algorithm	34
2.5.	Contaminant species.....	37
2.6.	Statistics analysis	38
3.	Estimation of the relative abundance of species in artificial mixtures of insects using low-coverage shotgun metagenomics.....	39
3.1.	Abstract.....	41
3.2.	Introduction	42
3.3.	Material and Methods	43
3.3.1.	Reference genomes	43
3.3.2.	Selection of species and preparation of the DNA libraries	43
3.3.3.	Classification of reads to species.....	45
3.3.4.	Detection limit	45
3.3.5.	Selection of best values of γ , δ and ϵ	46
3.3.6.	Quantification of the relative proportion of the species	47
3.3.7.	Rarefaction of the input samples	47
3.3.8.	Hardware	48
3.4.	Results.....	48
3.4.1.	Single-species libraries	48
3.4.2.	Mixed-species libraries.....	53
3.4.3.	Rarefaction of the reads	56
3.5.	Discussion	56
3.5.1.	Species identification: Spurious species and the need for an analytical limit of detection.....	59
3.5.2.	Quantification of the relative abundance of the species	62
3.5.3.	Data treatment and the assignation of reads to species	63
3.5.4.	Present and future of metagenomics.....	63

3.6.	Concluding remarks	66
4.	Relative species abundance estimation in artificial mixtures of insects using mito-metagenomics and a correction factor for the mitochondrial DNA copy number	67
4.1.	Abstract.....	69
4.2.	Introduction	70
4.3.	Material and Methods	71
4.3.1.	Reference genomes	71
4.3.2.	Selection of species and preparation of the DNA libraries	71
4.3.3.	Input data filtering.....	73
4.3.4.	Classification of reads to species.....	74
4.3.5.	Quantification of the RSA in mixed-species libraries and the need for a species-specific correction factor	75
4.3.6.	Rarefaction of the input samples	77
4.3.7.	Hardware	77
4.4.	Results.....	77
4.4.1.	Species identification.....	77
4.4.2.	Estimation of the RSA in mixed-species libraries	82
4.4.3.	Computer use	86
4.5.	Discussion	86
4.5.1.	Species identification.....	86
4.5.2.	Quantification of the RSA and the need for a species-specific correction factor	87
4.5.3.	Mito-metagenomics in <i>real</i> samples	89
4.6.	Concluding remarks	93
5.	Drastic reduction of false positive species in samples of insects by the combined use of two metagenomic classifiers	95
5.1.	Abstract.....	97
5.2.	Introduction	98
5.3.	Material and Methods	99
5.3.1.	Reference mitogenomes	99
5.3.2.	Selection of species and preparation of the DNA libraries	99
5.3.3.	Classification of reads to species.....	100
5.3.4.	Metrics.....	102

5.3.5.	Hardware	103
5.4.	Results.....	103
5.4.1.	Individual classifiers.....	103
5.4.2.	Combined classifiers	104
5.4.3.	Use of an analytical detection limit.....	104
5.5.	Discussion	107
5.5.1.	Individual metagenomic classifiers with default parameters	107
5.5.2.	Combination of the two metagenomic classifiers.....	108
5.5.3.	The use of an analytic detection limit	109
5.6.	Concluding remarks	109
6.	General discussion	111
6.1.	Identification of species.....	114
6.2.	Quantification of the relative species abundance.....	117
6.3.	A methodological perspective	119
6.3.1.	The workflow	119
6.3.2.	Metagenomic classifiers	121
6.4.	Future of molecular-based methods	124
7.	Conclusions and future works.....	127
	References.....	133
	Annexes	161
	Data Accessibility	163
	Supplementary material	163
	Supplementary Figures	163
	Supplementary Tables.....	163
	Supplementary Methods.....	165

List of Figures

Figure 1.1. NCBI database size (in Tbytes) for Eukaryotic taxa from 1992 to today. Data from NCBI repository (URL: https://www.ncbi.nlm.nih.gov/genome/browse/) consulted on 3rd July 2021.....	18
Figure 2.1. Main steps of the metagenomic method. Dashed arrows indicate an optional path.	29
Figure 2.2. Flow diagram of the computational pipeline used in the $B\gamma\delta$ method. At the top, input data and, below it, the steps and tools needed for the identification procedure.	35
Figure 2.3. Flow diagram of the γ - δ algorithm. Only the two highest mapping ratios to two reference genomes of a single read are required. In the figure, it is assumed that the highest mapped ratio A_i belongs to the reference genome i	36
Figure 3.1. Summary boxplots of the 22 single-species libraries used to search the best combination of parameters γ and δ ; in all cases, a detection limit of $\varepsilon = 0.001$ was used and contaminant species were discarded. (A) Number of identified species in the library. (B) Proportion of the assigned reads allocated to the right species. (C) RPIR. A different letter at the top of the figures indicates significant differences amongst γ - δ combinations.	51
Figure 3.2. Scatter plots between the expected (<i>i.e.</i> , as the mixtures were prepared in the lab; Table 2.2) and the estimated species relative abundance (Table 3.4) following the described bioinformatic pipeline. Each plot corresponds to one mixed-species library (A to F corresponds to libraries <i>no.</i> 1 to <i>no.</i> 6). Each point in the plot indicates one species in the mixture. In each plot, the correlation coefficient (r) and its p -value are also indicated.	55
Figure 3.3. Effect of the rarefaction of reads on the number of species detected (above $\varepsilon = 0.001$ and without contaminants) in the six mixed-species libraries (A to F correspond to libraries <i>no.</i> 1 to <i>no.</i> 6). The x axis indicates the proportion of reads used (when 1, all reads were used, so there is only one value); in the rest of the values, 100 random repetitions were conducted using the indicated proportion of reads. The horizontal dashed line of each plot indicates the actual number of species in the mixture.	57
Figure 3.4. Effect of the rarefaction of reads on the correlation coefficient r between the expected and the recovered relative abundance of the species in the six mixed-species libraries (A to F correspond to libraries <i>no.</i> 1 to <i>no.</i> 6). The x axis indicates the proportion of reads used (when 1, all reads were used, so there is only one value); in the rest of the values, 100 random repetitions were conducted using the indicated proportion of reads. The horizontal dashed line of each plot indicates the critical value of r , above which measured r is statistically significant at $p < 0.05$	58
Figure 4.1. MMG pipeline applied in chapter 4. In brackets, the tools used in each step.	80

Figure 4.2. Scatter plot of the estimated *versus* the actual RSA for each species of the mixed-species libraries (*i.e.*, *within*-species RSA). Each plot shows the Pearson correlation coefficient (r) and the corresponding p -value. The coordinate at the origin of all regression lines was not different to 0. 82

Figure 4.3. Scatter plot of the estimated *versus* the actual RSA in each mixed-species library (*i.e.*, *across*-species RSA). At the top, it is indicated the way we conducted the actual RSA. A: original expected data; B: corrected expected data after applying the N_{Mi} correction factor; C: corrected expected data after applying the \bar{N}_{Mi} correction factor. Rows from top to bottom correspond to mixed-species libraries from *no.* 1 to *no.* 6. Each plot shows the Pearson correlation coefficient (r) and the corresponding p -value. 83

Figure 4.4. Number of identified species using different proportions of reads in the mixed-species libraries. Each simulation was performed 100 times with different subsets, except when the entire library was used. Letters from A to F indicate mixed-species libraries from *no.* 1 to *no.* 6. Grey dashed lines indicate the expected number of recovered species in each library. 84

Figure 6.1. Number of available whole-genomes and mitogenomes of eukaryote species over the last thirty years. Data extracted from NCBI repository (URL: <https://www.ncbi.nlm.nih.gov/genome/browse/>) consulted on 7th June 2021. 114

List of Tables

Table 2.1. Summary table of the species in single-species libraries; the first run (run <i>no.</i> 1) was performed in September 2016 and the second (run <i>no.</i> 2) in July 2018.	31
Table 2.2. Summary table of the species and their relative concentrations in mixed-species libraries; run (run <i>no.</i> 3) was performed in September 2016.	32
Table 3.1. Summary information of the single-species libraries. In chapter 3 sequenced libraries were treated as single-end reads samples.	44
Table 3.2. Proportion of reads assigned to species, in parentheses, for each one of the 10 single-species libraries included in the first sequencing run. The species in each library are shown in the header column. Species assignments are divided in each column into blocks: A, species with abundance higher than $\epsilon = 0.01$; B, with abundance between $\epsilon = 0.01$ and $\epsilon = 0.001$; C, with abundance between $\epsilon = 0.001$ and $\epsilon = 0.0001$; D, with abundance below $\epsilon = 0.0001$; E, potential contaminants. Codes of the species as in Table S3.1.	52
Table 3.3. Summary information of the number of reads in mixed-species libraries (Table 2.2). In this study sequenced libraries were treated as single-end reads samples.	54
Table 3.4. Relative proportion of assigned reads to species, in parentheses, for each one of the six mixed-species libraries of Table 2.2 after applying γ - δ algorithm with parameters $\gamma = 0.99$ and $\delta = 0.98$. Codes of the species as in Table S3.1. In bold, the species whose DNA was actually put in the mixture.	54
Table 4.1. Summary information of single-species libraries used with paired-end reads.	72
Table 4.2. Summary of the results per library (mean \pm SD) on the training dataset of single-species libraries for the four combinations of input data assessed in this study (raw reads and candidate mito-reads mapped to Mito1794 and FilteredMito1794 databases) and using $\gamma = 0.99$ and $\delta = 0.96$. (A) Number of recovered species per library; (B) Relative proportion of informative reads (RPIR) per library; and (C) processing time per library (format h:mm:ss) (the time necessary to find the candidate mito-reads is included in the processing time). Reads from contaminant species have not been considered.	79
Table 4.3. List of species detected on the single-species libraries when the mitogenome of the focal species is in the reference database (column A) and when it is not (column B). For each detected species we indicate its name and the number of assigned reads (in brackets). The number of congeneric species of the focal species included in the database is provided in column C. Libraries are divided into 4 groups: Group 1, species without congeneric species in the database and without FP species; Group 2, species with congeneric species in the database and without FP species; Group 3, species with congeneric species in the database but with FP of the same genus; and Group 4, species with congeneric species in the database but with FP of a different genus.	81

Table 4.4. Summary table of the species in single-species libraries data used for the obtention of the correction factors N_{Mi} and \bar{N}_{Mi} . x_i is the number of reads mapping into the mitogenome of species i divided by the total number of reads from that single-species library; N_{Mi} is the estimated number of mitochondrial DNA copies for species i ; \bar{N}_{Mi} is as N_{Mi} but calculated using the mean length of whole genomes and mitogenomes of all species considered here. 85

Table 5.1. Rules of classification of a read r using the union and the intersection approaches of the metagenomic classifiers p and q . The read r can be assigned to a species (*e.g.*, species s or species n) or can remain not assigned (NA)..... 101

Table 5.2. Benchmark metrics scores for each classifier without detection limit (A), with an analytical detection limit of 0.0001 (B) and with an analytical detection limit of 0.001 (C). For richness, the RPIR, precision and recall we provide the mean and standard deviation (SD) of all 21 samples (format mean \pm SD), and for processing time the sum of the total consumed time when running all the samples sequentially (format hh:mm:ss). The time for creating the databases and running in-house python scripts are omitted. 105

Table 5.3. False positive species detected on each library by the intersection approach. For each library, we indicated the run and library codes, the name of focal species (its order in brackets), the number of congeneric species in the reference database, and a list of the FP species divided in congeneric and non-congeneric to the focal species. The last three columns contain the number of FP species detected with the analytical detection limits (ϵ) of 0, 0.0001 and 0.001. For each species, we indicated, in brackets, the RPIR and its order when it is different from the focal species. Order abbreviations are Col: Coleoptera, Dip: Diptera, Hem: Hemiptera, Hym: Hymenoptera, Lep: Lepidoptera. (Table shown in the next page)..... 105

Table 6.1. Comparison of the MG and MMG approaches. * Indicates that correlation was corrected with \bar{N}_{Mi} value. As the number of sequenced mitogenomes is higher than the number of sequenced whole-genomes, two sets of results are provided for MMG, one with the results presented in chapter 4 and another using just the same number and identity of mitogenomes as the whole genomes available in chapter 3 ($n = 110$). 115

Table 6.2. Comparison of classifiers used in this thesis for MMG analysis of single-species libraries (Table 2.1). 123

List of Abbreviations

B$\gamma$$\delta$	BWA + γ - δ algorithm
BIN	Barcode Index Number
BLAST	Basic local alignment search tool
BM	BLASTn + MEGAN6
BOLD	Barcode of life data
BWA	Burrows-Wheeler aligner
COI	Cytochrome c oxidase subunit I
CV	Coefficient of variation
DNA	Deoxyribonucleic acid
eDNA	Environmental DNA
FN	False negative
FP	False positive
GBS	Genotype by sequencing
gDNA	genomic DNA
HTS	High-throughput sequencing
ITS	Internal transcribed spacer
K2	Kraken2
LCA	Lowest-common ancestor
matK	Maturase K
MB	Metabarcoding
MG	Metagenomics
MMG	Mitochondrial metagenomics
msGBS	Multiple-species genotype by sequencing
NA	Not assigned
NCBI	National Center for Biotechnology Information
NGS	Next-generation sequencing
NUMT	Nuclear mitochondrial pseudogene
OTU	Operational taxonomic unit
PCR	Polymerase chain reaction
QC	Quality control

qPCR	Quantitative PCR
rbcl	Rubisco large subunit
RefSeq	Reference sequence
RevMet	Reverse metagenomics
RNA	Ribonucleic acid
RPIR	Relative proportion of informative reads
rRNA	Ribosomal RNA
RSA	Relative species abundance
SD	Standard deviation
taxID	Taxonomy identifier
TN	True negative
TP	True positive

List of Symbols

γ	Upper threshold of the γ - δ algorithm
δ	Lower threshold of the γ - δ algorithm
A	Mapping ratio
G_i	Length of the haploid nuclear genome of species i
\bar{G}	Mean value of G
M_i	Length of the mitochondrial genome of species i
\bar{M}	Mean value of M
n_m	Number of matching nucleotides between the read and the target sequence
N_M	Mitochondrial copy number
\bar{N}_M	Mean value of N_M
n_t	Total number of nucleotides involved in the alignment of the read and the target sequence
R_{Gi}	Number of reads from the single-species library of species i
R_{Mi}	Number of reads mapping into the mitochondrial genome of species i
x_i	Proportion of reads belonging to the mitochondrial genome of species i

1. Introduction

1. Introduction

1.1. Introduction to biodiversity assessment

Biodiversity is described as a snapshot of the diversity of organisms, their relationships, and functions at a certain location at a particular time (Porter & Hajibabaei 2017; Walker 1992). Biodiversity encompasses multiple levels of organization: ecosystems, populations, species, and/or genetics (Noss 1990). The pivotal question on biodiversity studies regards the identification of the species present in the sample under study. In the identification procedure, specimens are named, generally at species rank (Gaston 2000). This name links the specimen to all the scientific knowledge related to that species; the correct identification of species is therefore crucial.

With an estimated richness of 8.7 million [\pm 1.3 million standard error (SD)] eukaryote species, our knowledge of Earth's biodiversity is incomplete, as most of the species remain undescribed (Costello *et al.*, 2013; Mora *et al.*, 2011; Stork 2018). Although recent research efforts and easy data sharing through online publishing have increased the rate of discovery of new species (\sim 18,000 species per year since 2006; Costello *et al.*, 2013), it will take centuries to characterise all unknown species at the current rate (Bouchet 2006; Costello *et al.*, 2013). Given the decline of biodiversity caused by anthropogenic disturbances (*i.e.*, *biodiversity crisis*; Wilson 1985; Dirzo *et al.*, 2014), time is pressing for prompt and accurate identification techniques to describe the highest number of species before they become extinct (Ebach *et al.*, 2011). In addition, the decline of species number, abundance and distribution may have severe impacts on the ecosystems, in terms of communities' structures and functions (Dirzo *et al.*, 2014; Goulson 2019). Therefore, the rapid identification is also a crucial task for biodiversity monitoring to effectively apply conservation strategies. In this context, the present thesis focuses on the identification of species that are already described.

1.1.1. Classical morphology-based identification

The primary approach used by taxonomists to define and identify species was the morphological method. This method is based on the observation of anatomical features from the organisms. Currently, this method is applied with identification keys based on morphological characters (*e.g.*, presence, shape, or colour of structures) of the organisms under study. Although inventorying species with morphological methods have been useful for many years and different ecological settings, the procedure is constrained by many shortcomings:

- High expertise is required (Hajibabaei *et al.*, 2011; Hebert *et al.*, 2003a).
- Lengthy identification procedures. Importantly, the species discovery rate is lower than the extinction rate (Blaxter 2003; Lebuhn *et al.*, 2013; May 1988).
- Morphological characteristics may be tricky and lead to misidentification; for instance, phenotypic features may not provide enough information to set apart different species (*e.g.*, cryptic species; Korshunova *et al.*, 2019), and phenotypic plasticity may overestimate species diversity (*e.g.*, polyphenism; Yang & Pospisilik 2019).
- Incomplete morphological identification keys, especially for early life stages, consequently, morphological differentiation can only be applied at certain stages (Harvey *et al.*, 2003; Sweeney *et al.*, 2011).
- Organisms can be damaged and, if critical features for species discrimination are missing from the specimen, taxonomic resolution may be limited to higher taxonomic levels (*e.g.*, genus or family) (Reilly 2003).
- Organisms need to be observed or captured and this is not always an easy task; organisms hide, have circadian rhythms, small organisms are likely overlooked (Haase *et al.*, 2010).
- Experimental design, like sampling effort, may bias the detected taxa (Martinez *et al.*, 1999).

In addition to the limitations of the method *per se*, and perhaps more importantly, the specimens can be killed or damaged during the study and so the environment can be

destroyed by intrusive sampling procedures (Baldwin *et al.*, 1996). Despite the listed limitations, morphological identification is still applied for species inventories, as this approach is cheap, in terms of equipment (Buss *et al.*, 2015), and it saves time and money when only a few organisms need to be identified (Erlank *et al.*, 2018).

1.1.2. Molecular-based identification

Alternatively, molecular methods have been proposed for species identification instead of morphology-based methods. Molecular methods mainly consist of the study of a target sequence or the whole genomic content (Mande *et al.*, 2012). In the former method, target sequences are known as “DNA barcodes”. The DNA barcodes are amplified using the Polymerase Chain Reaction (PCR) with primers designed to match specific taxonomic groups (Valentini *et al.*, 2009). The barcodes can be used to identify one species in “DNA barcoding” or multiple specimens in “DNA metabarcoding”. DNA barcoding uses primers for single species PCR amplification to identify one species at a time, while DNA metabarcoding uses universal primers for group-specific amplification to identify multiple species simultaneously (Taberlet *et al.*, 2018).

The second approach targeting all genomic content is so-called “metagenomics”. In the metagenomic method, the whole genomic content within a sample is directly sequenced (*i.e.*, without enrichment of target sequences or organelle DNA before sequencing). Metagenomics uses the total genomic material, not only the barcode regions, to provide an overview of the taxonomic diversity and the functional profile of the sample under study (Porter & Hajibabaei 2017). When this method is applied to organelle genomes, it is named “mitochondrial metagenomics”, or “mito-metagenomics”, for mitochondrial genomes and “chloroplast metagenomics”, for chloroplasts genomes (Crampton-Platt *et al.*, 2016; Piñol 2021). All the above listed molecular-based techniques typically match the query sequences against a database of sequences with a known provenance for classification; thus, these methods are

dependent on the completeness of the reference databases, in terms of number and representation of species (Singer *et al.*, 2020; Taberlet *et al.*, 2012b).

Molecular-based methods are particularly useful to characterise the taxonomic composition and richness of the DNA contained within an environmental sample (Mardis 2008; Taberlet *et al.*, 2012a; Thomsen & Willerslev 2015; Valentini *et al.*, 2016). These samples may be obtained by collecting water, soil, sediments, air, faeces or gut contents (Agustí *et al.*, 2003; Alberdi *et al.*, 2018; Taberlet *et al.*, 2018; Valentini *et al.*, 2016; Venter *et al.*, 2004). The DNA extracted from such samples is referred to as environmental DNA (eDNA). The eDNA samples contain DNA emitted from living organisms into the environment with secretions, dead or quiescent cells (like seeds), and fragments or whole specimens (Porter & Hajibabaei 2017). The eDNA comprises a “soup” intracellular and extracellular DNA from different organisms (Taberlet *et al.*, 2018). The complexity of the eDNA generates particular challenges that are well addressed in Goldberg *et al.* (2015, 2016). Additionally, several studies report significant correlations between DNA quantities and specimens’ abundances (*e.g.*, Goldberg *et al.*, 2013; Kraaijeveld *et al.*, 2015; Saitoh *et al.*, 2016), albeit many factors influence the DNA amounts in samples (like the shedding rates among species, degradation rates under different conditions, collection, conservation and laboratory processing; Goldberg *et al.*, 2015, 2016; McInnes *et al.*, 2016).

While molecular-based methods have gained importance, there are still in their infancy, and the standardization of procedures and terminologies are underway. To this end, DNAqua-Net was formed in 2016 and held its first international conference just a short time ago, on 9-11 March 2021 (Bohmann *et al.*, 2021). The DNAqua-Net is devoted specifically to the standardization and implementation of DNA-based monitoring methods for European waters, but their recommendations can easily be applied to other communities (Bohmann *et al.*, 2021). To avoid confusion, here we use the names (meta)barcoding, mitochondrial/chloroplast metagenomics and metagenomics, for methods using respectively barcodes, organelle genomes or whole genomes as references, regardless of the sequencing technology applied.

1.2. From single species to multiple species assessment with DNA barcodes

1.2.1. DNA barcoding

The term “DNA barcode” was coined by Arnot *et al.* (1993); in that study a hypervariable tandemly repeated regions was used to produce a unique digital code, the “barcode”, for *Plasmodium falciparum*. However, the idea of using molecular methods for species discrimination is older than that. In 1977, Woese and Fox (1977) proposed the 16S ribosomal RNA (rRNA) as a phylogenetic marker for microbial diversity assessment. Since then, there were efforts to develop a large-scale standardized method for species identification (*e.g.*, Agustí *et al.*, 2003; Bartlett & Davidson 1991; Fox *et al.*, 1977; Giovannoni *et al.*, 1990; Ward *et al.*, 1990). There was no agreement until Hebert *et al.* (2003a) proposed the DNA barcode as a small DNA segment of the genome that contains enough taxonomic information to be used as a tag for species identification. The theory behind the Herbert’s proposal is that every species is characterized by a unique DNA barcode that is shared across the individuals within that species but not shared across different species. However, specimens of the same species exhibit some degree of genetic intraspecific variations. Therefore, an ideal DNA barcode is a short genetic fragment that is highly similar within specimens of a species (*i.e.*, low intraspecific genetic variation) and different across species (*i.e.*, high interspecific genetic variation). The DNA barcodes are not informative *per se*; query barcodes may be compared to reference databases populated with barcodes from identified specimens representing the genetic diversity of the taxonomic groups. Established in 2008, the consortium called the International Barcode of Life (iBOL) aims to compile barcodes from all species on Earth. The barcode sequences are placed in the Barcode of Life Data (BOLD) System (Ratnasingham & Hebert 2007, 2013), an online repository that stores barcodes sequences along with specimen information, such as taxonomy, specimens’ collection, preservation, and laboratory treatment. Barcodes from different individuals that belong to the same species are clustered in

BINs (Barcode Index Numbers). Currently, BOLD System contains more than 9,6 million barcodes, comprising 719 thousand BINs from 231 thousand animal species, 70 thousand plant species and 24 thousand fungi and other species (data extracted from BOLD System on 5th July 2021). The number of species represented is expected to continue to grow, thanks to initiatives like the BIOSCAN project (Hobern & Hebert 2019; Pennisi 2019).

The first proposed DNA barcode was the 658bp fragment of the mitochondrial genome, the cytochrome *c* oxidase subunit I (COI) gene, so-called Folmer region (Folmer *et al.*, 1994), for animals (Hebert *et al.*, 2003a, 2003b). The rationale behind this gene selection started with the choice of the mitochondrial genome (also referred as mitogenome). The advantages of the mitogenome, over the nuclear genome, are its lack of introns and its maternal inheritance (Hebert *et al.*, 2003a). The COI gene was subsequently chosen because it is a coding gene, thus indels are rare, and previous work provided a set of robust primers (Hebert *et al.*, 2003a). For species delimitation with the Folmer region, most congeneric species presented a genetic distance of 2%, while the genetic distance between specimens of the same species rarely exceeds such threshold (Hebert *et al.*, 2003b). Such distance is known as the “barcoding gap”.

The DNA barcoding technique for identification works as follows: a single specimen is collected, the DNA of the specimen is extracted, the DNA barcode is PCR-amplified and the amplicons are clustered into Operational Taxonomic Units (OTUs) or directly compared to a reference database (Ratnasingham & Hebert 2013). This straightforward technique rapidly caught on and extended to other taxonomic groups. However, COI does not always resolve species-level identification; consequently, alternative genetic markers have been proposed for particular groups, such as rubisco large subunit (*rbcl*) and maturase K (*matK*) for plants (CBoL Plant Working Group 2009), internal transcribed spacer (ITS) for fungi (Schoch *et al.*, 2012), 16S rRNA for prokaryotic species (Lebonah *et al.*, 2014). The uses of the DNA barcode also expand from species identification to diverse settings, like molecular phylogenetic, biodiversity conservation, environmental monitoring, trophic interactions, food safety, industry

quality assurance and human health (Fišer Pečnikar & Buzan 2014; Kress *et al.*, 2015; Sgamma *et al.*, 2017).

The DNA barcoding for species identification presents several advantages:

- Low, or none, taxonomic expertise is needed (Hebert *et al.*, 2003a).
- Time- and cost-effective method (De Barba *et al.*, 2010). The costs for obtaining barcodes are low (about \$1 per specimen, including collecting specimens, DNA extraction and sequencing) (Pennisi 2019).
- Easily reproducible procedure (Shokralla *et al.*, 2015).
- Specimens are generally identified at species rank (Sweeney *et al.*, 2011).
- Specimens can be identified at any life stage, when they are partially destroyed or when the phenotypic features do not provide enough information (*e.g.*, cryptic species; Hebert *et al.*, 2004a; Kress *et al.*, 2015).
- Inaccessible or elusive specimens can be identified with DNA barcodes recovered from environmental samples (Ficetola *et al.*, 2008; Goldberg *et al.*, 2015).
- Environmental samples are generally obtained in a non-invasive manner (*e.g.*, Deagle *et al.*, 2007; Ficetola *et al.*, 2008).
- A low amount of DNA is needed to detect species (deWaard *et al.*, 2008).

Yet, this technique is not free from shortcomings; the limits that hinder the DNA barcoding method include the following issues:

- Taxonomic resolution at the species level is not guaranteed (Hollingsworth *et al.*, 2011; Little & Stevenson 2007). For instance, recently diverged species are likely to present a low level of differentiation (Hebert *et al.*, 2004b; Meier *et al.*, 2006). Additionally, the barcodes in the repositories are not always identified at species level; for instance, on 20th August 2021, the search of “Insects” term in BOLD System returned 113,439 records, but only ~59% of them had species names.

- The “barcoding gap” is not constant, so the same threshold may accurately differentiate species within a certain group but fail in other groups (Bell *et al.*, 2019; Meier *et al.*, 2006; Shearer & Coffroth 2008).
- Populating the reference databases is time-consuming because the barcode must be obtained from single specimens (Stein *et al.*, 2014).
- Reference databases are far from complete (Kwong *et al.*, 2012) and some groups are poorly represented (Chambers & Hebert 2016; Ermakov *et al.*, 2015), which limits the identification success.
- Erroneous DNA barcodes in reference databases are not rare and may generate incorrect identifications (Shen *et al.*, 2013).
- DNA is often degraded (*i.e.*, age, storage, or chemical treatments) which limits the efficiency of PCR to amplify the target barcode and generates false negatives (Valentini *et al.*, 2009).

Despite the strengths of the DNA barcoding technique, the usefulness of the method for biodiversity assessment is limited because it only provides one species per analysis. Therefore, its application is not feasible in specious samples with thousands of specimens, as most environmental samples are, both in terms of time and cost (Shokralla *et al.*, 2015).

1.2.2. DNA metabarcoding

By 2006, the advent of high-throughput sequencing (HTS) technologies offered new avenues for molecular-based biodiversity assessment with the massive sequencing of millions of genetic sequences in parallel from a single DNA sample (Ansorge 2009; Mardis 2008). DNA metabarcoding (MB) couples HTS with DNA barcoding, modified by targeting shorter gene fragments and using universal primers on the PCR step, to identify and quantify species abundance in complex samples (Taberlet *et al.*, 2012b). Thus, MB has the potential to overcome the limitation of barcoding individual specimens per experiment and provide barcodes from a broad spectrum of taxa.

This technique is well-established and used in many ecological settings and with different groups of organisms (Clare *et al.*, 2016; Deagle *et al.*, 2018; Evans *et al.*, 2016; Taberlet *et al.*, 2018). However, the above-mentioned advantages and drawbacks of the DNA barcoding technique are also valid for MB. Still, there are several additional drawbacks:

- The primer binding sites are not always highly conserved among all taxa of the target group and, therefore, universal primers design is not always possible (Deagle *et al.*, 2014).
- The quantitative ability of the method is hampered by PCR dependency (Fonseca 2018; Lamb *et al.*, 2018; Piñol *et al.*, 2019). Amplification is possible even without a perfect match between the primer and the binding site, but in such cases the efficiency of PCR is difficult to predict (Deagle *et al.*, 2014).
- PCR also increases the risk of false positive detections due to amplification of artefactual sequences (*e.g.*, chimeric sequences; Galan *et al.*, 2012).
- NUMTs (nuclear mitochondrial pseudogenes) can be amplified alongside the focal barcode and, consequently, they bias the final results (but see Ermakov *et al.*, 2015).
- OTUs clustering may result in either over or underestimation of species richness (Clare *et al.*, 2016).
- Samples are easily contaminated with exogenous DNA (*e.g.*, laboratory contamination), that can be co-amplified alongside the target barcode (Goldberg *et al.*, 2016; Liu *et al.*, 2013).

1.3. Metagenomics

1.3.1. A brief history of the metagenomic method

The starting point of this history is shared with the DNA barcoding method but evolved differently and goes in hand with microbial community analysis. Since Woese and Fox

(1977) proposed to use the 16S rRNA gene as a phylogenetic marker, it took more than ten years to produce the first microbial community analysis using the 16S rRNA gene (Giovannoni *et al.*, 1990). In Giovannoni's study, the 16S rRNA gene was PCR amplified, cloned by inserting the DNA fragments to vectors and sequenced with the Sanger method (Sanger *et al.*, 1977). Despite these improvements on the culture-independent study of microbial communities, the limitations of amplicon sequencing pushed the field forward with the attempt of the whole genome study (Stein *et al.*, 1996).

In 1998, the term "metagenomics" (MG) was introduced by Handelsman *et al.* (1998) to refer to the direct study of the collective genomes of microorganisms isolated from an environmental sample (*e.g.*, soil, water, sediments, or air). Albeit the first attempt of metagenomic sequencing is attributable to Stein *et al.* (1996). Their premise was that taxonomic information could be obtained using longer genomic fragments (~40 Kbases) than the 16S rRNA. Stein *et al.* (1996) used a "genome walking" upstream and downstream of the 16S rRNA gene to study oceanic planktonic archaea. In that study, they digested the DNA in smaller fragments with restriction enzymes and cloned those fragments using an *Escherichia coli* vector. Next, the fragments that contained the 16S rRNA gene were PCR amplified and sequenced. They finally looked for homologies using BLAST against the NCBI *nr* database. This genome walking constrained the study to sequences adjacent to the 16S rRNA region but led the march of analysing the genome of several species at the same time.

The "metagenomic" concept was quickly adapted to random shotgun sequencing the whole genomic content within a sample (Tyson *et al.*, 2004; Venter *et al.*, 2004). Such sequencing reported a myriad of DNA sequences from diverse genomic regions and organisms. To sort things out, Tyson *et al.* (2004) proposed reconstructing the genomes by assembling the shotgun sequences. However, the number of retrieved DNA sequences by Sanger sequencing was low (~100 sequences in a common sequencing run and up to $\sim 10^5$ when deep sequencing; Not *et al.*, 2009; Tyson *et al.*, 2004). This bottleneck was surpassed by the emergence of next-generation sequencing (NGS) technologies able to sequence millions of reads in parallel (Ansorge 2009; Mardis 2008).

The above-listed advances lead to the current definition of the MG approach; today this method consists of extracting the total genomic content from a sample and randomly sequence DNA from all parts of the genome using NGS technologies. The obtained reads can be assembled into longer contigs or directly compared with sequences stored in genomic repositories to infer the species identity and functional diversity (Breitwieser *et al.*, 2019).

The low cost and high throughput of NGS stimulated the application of PCR-free approaches for biodiversity analysis using the MG method on its own or coupled with PCR MB. In either case, and similarly to MB, the sensitivity of the method relies on the completeness of the reference database (Singer *et al.*, 2020); but, unlike MB, the identification is not restricted to particular genetic markers, which confers several advantages over amplicon-based methods:

- Because the MG method is PCR independent, the sequenced reads should be, statistically, a faithful representation of the original DNA composition of the sample.
- The MG methods do not require prior knowledge of the taxonomic group under study, while DNA (meta)barcoding does (*e.g.*, for primer design).
- The shotgun sequencing recovers different loci that increase the taxonomic information recovered from the dataset, enabling higher resolution and better discriminatory power between genetically close species (Srivathsan *et al.*, 2016).
- Sequenced reads belong to organisms from all kingdoms, enabling extensive characterization of biodiversity (Ranjan *et al.*, 2016).
- The approach also gives insights into the functional role of the microbial community (Escobar-Zepeda *et al.*, 2015).
- Shotgun sequencing can recover DNA sequences even on highly degraded samples (Parducci *et al.*, 2019; but see Chua *et al.*, 2021).

However, the technique has also some limitations:

- Non target DNA (*e.g.*, contamination and host) is common and sometimes may outnumber the focal reads (McArdle & Kaforou 2020; Pereira-Marques *et al.*, 2019).
- Low abundant taxa are likely to be overlooked (Kuczynski *et al.*, 2012; Pereira-Marques *et al.*, 2019).
- Sequences with high and low GC content may be under-represented (Benjamini & Speed 2012; Ross *et al.*, 2013).
- Read length affect classification accuracy. Long-reads are more error-prone but they provide more accurate assignments than short-reads (Pearman *et al.*, 2020).
- Species detection success wanes with sample age (Chua *et al.*, 2021).
- Today, shotgun sequencing is more expensive and requires a more extensive database than amplicon metabarcoding (Fonseca 2018; Ranjan *et al.*, 2016).

1.3.2. Metagenomics in eukaryotes

In prokaryotes, shotgun metagenomics provides more accurate taxonomic identification than the classical 16S amplicon metabarcoding (Chen & Pachter 2005). However, in eukaryotes, shotgun metagenomics is hindered by the scarcity of eukaryote species with sequenced genomes (8,417 eukaryote *versus* 203,148 prokaryote genomes; NCBI database, accessed on 29th April 2019). There are good reasons for the lack of eukaryote genomes on reference repositories, being the most important the longer size and complexity of eukaryote genomes (400.2 Mb \pm 1,106.2 Mb in eukaryotes and 3.9 Mb \pm 3.7 Mb on prokaryotes) that hamper the reconstruction of individual genomes. However, the number of sequenced genomes is quickly increasing, as there are several ongoing projects devoted to obtain complete genomes of several groups of organisms: G10K for vertebrates (Genome 10K Community of Scientists 2009), GIGA for marine invertebrates (GIGA Community of Scientists 2014), GAGA for ants (Boomsma *et al.*, 2017), i5K for arthropods (i5K Consortium 2013; Levine 2011; Robinson *et al.*, 2011), 10KP for plants (Cheng *et al.*,

2018) and 1KFG for fungi (Grigoriev *et al.*, 2014), amongst others. There is even a proposal to sequence the genomes of all eukaryotic species in ten years for *ca.* 3 billion dollars (Lewin *et al.*, 2018); this estimate could be optimistic, but it probably means that the objective is within reach in a few decades, not more.

In eukaryotes, shotgun metagenomics has been mainly applied using chloroplasts and mitochondrial genomes (Srivathsan *et al.*, 2015; Tang *et al.*, 2014), but also using nuclear genomic regions with multiple-copy number (Linard *et al.*, 2015). The studies, using mitochondrial genomes, showed that quantitative information could be obtained from heterogeneous samples (Bista *et al.*, 2018; Tang *et al.*, 2015; Zhou *et al.*, 2013), but it is fair to assume that the use of complete genomes would provide better quantitative results.

1.3.3. Mitochondrial metagenomics

When the MG approach is restricted to the analysis of the mitochondrial sequences contained within a shotgun sample, the approach is referred to as “mitochondrial metagenomics” or “mito-metagenomics” (MMG) (Crampton-Platt *et al.*, 2016). The MMG is more reliable today than the MG method for the biodiversity assessment for eukaryote species because the number of mitochondrial genomes (mitogenomes) is much higher than the number of complete genomes. On 13th June 2021, the eukaryote domain had ~2,700 whole-genomes and ~12,400 mitochondrial genomes on the NCBI RefSeq (Reference Sequence) repository. Besides, thanks to the natural enrichment of the mitochondrial DNA in cells, multiple genetic markers from the mitochondria, or even whole mitogenomes, can be recovered without PCR enrichment (Tang *et al.*, 2015; Zhou *et al.*, 2013).

Recent studies have exploited the MMG method as a powerful tool for phylogenetic and taxonomic studies of eukaryote species (Crampton-Platt *et al.*, 2016). Such studies validate the potential of the MMG method as a bridge between MB and MG methods.

Indeed, MMG addresses some of the problems associated with MB but other challenges arise. On the positive side, the MMG have more discriminatory power than individual barcode loci (Gómez-Rodríguez *et al.*, 2015) and also to avoid PCR associated biases that hampers quantitative interpretation of MB results (Tang *et al.*, 2015). On the negative side, the mitochondrial proportion only represents about the 0.5-4% from the total shotgun reads, so, most of the shotgun read are not used in the analysis (Bista *et al.*, 2018; Gómez-Rodríguez *et al.*, 2015).

The relative species abundance concept in mitochondrial metagenomics

It is generally assumed that MMG quantifies satisfactorily the relative species abundance (RSA) of complex mixtures (Gómez-Rodríguez *et al.*, 2015; Zhou *et al.*, 2013). However, as far as we know, there are only five studies that tested the MMG method (plus one using chloroplast metagenomics in plants) using shotgun samples of known composition (Bista *et al.*, 2018; Gómez-Rodríguez *et al.*, 2015; Gueuning *et al.*, 2019; Ji *et al.*, 2020; Lang *et al.*, 2019; Tang *et al.*, 2015). In general, the relationship between the expected and estimated RSA is statistically significant, but with high variability in the goodness of fit.

How the RSA of complex mixtures is presented in the literature needs some clarification. First, the RSA can be expressed as a proportion of the species biomass (*e.g.*, Gueuning *et al.*, 2019) or individual counts (*e.g.*, Lang *et al.*, 2019); alternatively, the RSA can refer to the proportion of the DNA amount of each species in the mixture. Whilst the former approach is more meaningful for most ecological studies, in this thesis we adopt the latter approach because it allows the independent evaluation of different sources of bias on the RSA estimation. Second, some studies provide the relative abundance of one species in different samples (*e.g.*, Bista *et al.*, 2018), whereas others report the abundance of several species in a single sample (*e.g.*, Saitoh *et al.*, 2016). Ji *et al.* (2020) named *within*-species estimation the former (is species *i* more abundant in sample *s* than in sample *r*?) and *across*-species the latter (is species *i* more abundant than species *j* in sample *s*?). This distinction is important because there

are species-specific characteristics that influence the *across*-species estimation but not the *within*-species estimation.

For MMG studies, the most important of these characteristics is the variable number of mitogenomes per nuclear genome (mitochondrial DNA copy number). Thus, a species *i* with twice the number of mitogenomes per nuclear genome than another species *j* will produce twice many mitochondrial reads as well; without a proper correcting factor, species *i* would, apparently, be twice more abundant in the mixture than species *j*. This fact is known (Bista *et al.*, 2018; Piñol *et al.*, 2015; Tang *et al.*, 2014), but there are not reliable solutions to the problem because little is known about the causes of the variation of the mitochondrial copy number *across*-species (but see Liu *et al.*, 2018, that reported a higher mitochondrial copy number in organs with a high metabolic rate and in species living at low altitude than in their counterparts at high altitude in the Tibetan Plateau).

The size of the nuclear genome of the species also affects the *across*-species RSA estimation in MMG studies. Being all other things equal, a species *r* with a nuclear genome half as big as that of another species *s* will produce twice many mitochondrial reads because the mitochondrial DNA is diluted in a smaller amount of nuclear DNA. Therefore, without a proper correcting factor, species *r* would, apparently, be twice more abundant in the mixture than species *s*. The effect of genome size on RSA estimation is also known (Crampton-Platt *et al.*, 2016; Krehenwinkel *et al.*, 2017; Tang *et al.*, 2014), but it is difficult to consider it because measuring the genome size is not an easy task (there is a database of genomes sizes with 1344 insect species on it; Gregory (2020), accessed on 25th March 2020). Both the variation *across*-species of mitochondrial copy number and genome size affect MMG, but also any amplicon MB method that targets genomic regions with a variable copy number, like COI in animals (Hebert *et al.*, 2003b), ITS in fungi (Schoch *et al.*, 2012), or *rbcl* + *matK* in plants (CBoL Plant Working Group 2009).

1.4. Bioinformatic considerations

The increasing throughput of sequencing technologies was an important stepping stone to the molecular-based methods, by increasing massively the amount of genetic data recovered from a single experiment while reducing sequencing costs. This large amount of data opens two major challenges: data storage and processing.

1.4.1. Reference databases

Reference databases are essential to infer the phylogenetic relationship of genomic sequences. Thanks to the improvement of the sequencing technologies mentioned earlier, the number of genomic data on online repositories is increasing quickly (Figure 1.1). On the positive side, as the number of complete genomes grows, the ability of molecular-based methods for taxonomic profiling should also improve. On the negative side, the computational workload increases with the increasing size of the reference database.

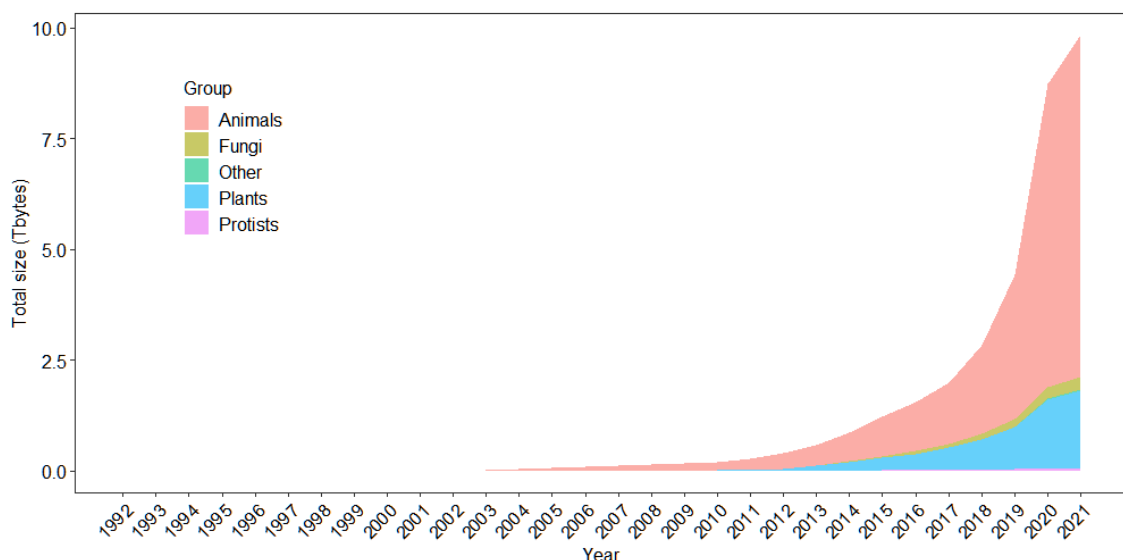


Figure 1.1. NCBI database size (in Tbytes) for Eukaryotic taxa from 1992 to today. Data from NCBI repository (URL: <https://www.ncbi.nlm.nih.gov/genome/browse/>) consulted on 3rd July 2021

Today there are several repositories for genomic data storage, some of them devoted to a particular data type, like the above-mentioned BOLD System that contains barcodes from COI, ITS, *rbcl* + *matK* genes (Ratnasingham & Hebert 2007), or SILVA repository that stores ribosomal genes (Quast *et al.*, 2013); or diverse data types, like NCBI *nt* containing from whole genomes to scaffolds and genes. In metagenomics studies, the NCBI database is the common choice for phylogenetic inference as it is the most complete database with high-quality standards that ensure curated reference sequences.

Other challenges that arise related to data storage include storage space, data access and standardized format of data. Some online repositories like NCBI and EMBL-EBI allow the blasting of samples using online servers; this is practical for a general search of a reduced number of sequences. However, in general, the interest is to compare relatively huge query samples to restricted taxonomic groups. In such a case, a local database is built with selected references from online repositories. Thus, repositories must have efficient ways to download their data. The data might be also stored in a standardized format, this includes the representation of the data (*e.g.*, FASTA format), and linked to its description including, for example, specimens' details, data collection, storage, and laboratory treatments that may help to the interpretation of the results. Once the reference data is selected, there should be enough disk space available for the raw sequences and for the indexes that most metagenomic classifiers use when searching for homologies. To give some numbers, on 12th July 2021, the NCBI *nt* database occupied 488 Gbytes (in the uncompressed format, and 132 Gbytes the pre-compiled database for BLAST). In a modest cluster, or even desktop computer, there is enough disk space to store such an amount of data, but the storage space needed increases when several classifiers are used, as mappers require custom indexes, by aggregating references from different repositories and keeping the different versions of the same reference database.

1.4.2. Bioinformatic tools and pipelines

The bioinformatic tools and pipelines used to assign HTS reads to species come with many names, but we refer to them here generically as metagenomic classifiers. There is a myriad of such tools that use a wide variety of strategies, like read alignment, k -mer mapping, marker genes alignment or sequence composition (Breitwieser *et al.*, 2019; Mande *et al.*, 2012). Considering only classifiers that assign individual query sequences to reference sequences by similarity (*i.e.*, taxonomic binning), there are two general strategies in a compromise between accurate results and reasonable execution times. Tools specifically designed to provide highly precise classification are built on aligning reads against reference sequences to return the most similar matches. Despite significant improvements in aligners performance, this approach is computationally intensive. Popular tools of this group are BLAST (Altschul *et al.*, 1990), Bowtie2 (Langmead & Salzberg 2012) and BWA (Li & Durbin 2009). Alternatively, classifiers can reduce the complexity of the alignment at the expense of sensitivity. A very efficient strategy is based on k -mers (read sub-strings of length k); rather than mapping whole read, the k -mers of a query read are directly associated with taxa that contain the same k -mers. Examples of classifiers of this group are Kraken (Wood & Salzberg 2014), CLARK (Ounit *et al.*, 2015) and Kallisto (Bray *et al.*, 2016). In both cases (whole-read alignment and k -mer-based) several taxa can be associated with a read, so an algorithm is needed to assign a taxon to each read; the most common approach is the so called lowest-common ancestor (LCA) algorithm, implemented, among many others, in MEGAN (Huson *et al.*, 2007).

With the overwhelming variety of software and bioinformatic pipelines for sequence identification, a major issue arises from the lack of standard procedures, that is different pipelines may lead to different conclusions (Harbert 2018; Lindgreen *et al.*, 2016); besides, results from different studies can neither be directly compared. This myriad of pipelines also reflects the uncertainty upon the most appropriate bioinformatic pipeline for assessing a particular sample type. Therefore, it is of foremost importance that metagenomic classifiers are continually benchmarked to evaluate their performance and compare to new tools. Indeed, there are plenty of

published papers devoted to such comparison (*e.g.*, Almeida *et al.*, 2018; Hleap *et al.*, 2021; Lindgreen *et al.*, 2016; McIntyre *et al.*, 2017; Meyer *et al.*, 2019; Peabody *et al.*, 2015; Siegwald *et al.*, 2017; Ye *et al.*, 2019).

1.5. Insects as model organisms

Insects are the largest and more diverse group in the world. To date, there are nearly one million insect species described and many millions are still to be discovered (Gullan & Cranston 2014). Estimations suggest that insects constitute about half of all species on Earth (May 1988; Sabrosky 1952). Their distribution covers all continents on Earth (Gressitt & Leech 1961). Insects display an enormous variety of forms and structures adapted to incredibly variable environmental types, from water to terrestrial ecosystems, including above and underground (Gullan & Cranston 2014; Kingsolver *et al.*, 2011). Moreover, insects may live on the same or different ecosystem during their entire life or a particular life stage (Gullan & Cranston 2014).

Insects play fundamental roles in ecosystems, such as plant propagation, via pollination of flowering plants and seed dispersal (Bronstein *et al.*, 2006); nutrient recycling, including disposal of detritus and feeding on dead organisms (Adamski *et al.*, 2019); and establishing crucial links in food webs as predators and/or preys (Goulson 2019), and also as a vector on diseases transmission (Carn 1996). From an anthropocentric point of view, many popular insect species are known for their harmful effect as pests on agriculture (*e.g.*, the sweet potato whitefly, *Bemisia tabaci*, which is a crop pest of major food staples, like tomato, cucumber, and zucchini; Gullan & Cranston 2014), or on human health (*e.g.*, the house fly, *Musca domestica*, which is a vector of pathogens that cause serious infections to humans; Khamesipour *et al.*, 2018). Whereas other insect species are beneficial for humans by providing food, directly or indirectly; for instance, the honeybee, *Apis mellifera*, is valuable for honey production, and as a flower pollinator (Paudel *et al.*, 2015). Additionally, insects are used for biological control of pests (Kulkarni *et al.*, 2015), for extraction of chemicals

(like chitin used as an anticoagulant), and primary materials (*e.g.*, silks from the cocoons of *Bombyx mori*) (Gullan & Cranston 2014). Additionally, the diversity and abundance of insect species vary according to biotic and abiotic factors (Adamski *et al.*, 2019; Kingsolver *et al.*, 2011). Such characteristic makes insect species perfect for assessing the conservation status of habitats and monitoring ecosystems.

Several molecular-based methods for the biodiversity assessment have been tested on insects and argue that the method can be extrapolated to other animals (*e.g.*, Crampton-Platt *et al.*, 2016; Hebert *et al.*, 2003a). This great interest in insects is justified by the above-listed reasons, plus the fact that the identification and description of insect species with morphological keys is particularly challenging in some groups. In this thesis, we also use insect species as model organisms and argue that the strategies and conclusions yield from the presented research can be inferred to other Metazoans.

1.6. Thesis hypothesis and goals

1.6.1. Defining the problem and thesis hypothesis

In the last two decades, the molecular-based method has encompassed, or even replaced, the classical morphological-based identification of species because it allows the detection and identification of species even when morphology study is impossible (Barrett & Hebert 2005). Besides, advents of throughput in sequencing technologies together with the reduction of sequencing costs and time have accelerated the production of genetic data (Figure 1.1) that allows for a more in-depth study of a DNA sample.

In eukaryotes species, DNA MB is the common approach for characterizing the species composition. In DNA MB, a single or a few marker genes are PCR amplified and used as tags for species identification (Hebert *et al.*, 2003a; Taberlet *et al.*, 2012b). The

amplicon number is subsequently used to infer the RSA (Deagle *et al.*, 2018). Despite its diverse application in different organisms and ecological settings, several studies reported that this technique can recover the species list but not the RSA (Elbrecht & Leese 2015; Piñol *et al.*, 2015). The most important reason of this problem is the dependency of the method on PCR. The PCR is applied with universal primers to enrich the sample from the target genetic markers. However, the universal primers may not perfectly match the primers binding sides of all the target species (Deagle *et al.*, 2014), therefore the unpredictable efficiency of amplification may blur the original species composition.

There is some consensus in the literature that, if the PCR step could be avoided, then the MB process would be much more quantitative (Bista *et al.*, 2018; Taberlet *et al.*, 2012b; Zhou *et al.*, 2013). One PCR-free approach is shotgun MG, where the extracted DNA is sequenced directly, so all PCR-generated biases are avoided (Elbrecht & Leese 2015; Yu *et al.*, 2012; Zhou *et al.*, 2013). The MG method is widely applied by the time of writing this thesis but mainly used for taxonomic and functional genes profiling of microbial communities. The restriction of the MG method to microbial communities is justified by our limited knowledge of eukaryote species' whole-genomes, and arguably unequalled among groups, because of the length and complexity of the eukaryote species hampers the reconstruction of whole genomes (Escobar-Zepeda *et al.*, 2015).

In this thesis is hypothesized that in the future, when the entire genome of most species would be sequenced and available in public databases, the shotgun MG could be used to characterize, quantitatively, the biodiversity of Metazoan species. To test this hypothesis, we simulate highly complete online repositories by analysing *artificial* samples of insect species whose whole genomes are assembled to an advanced degree and available on online repositories.

1.6.2. Research goals

The present thesis has the core objective of developing a PCR-free molecular-based method that provides robust identification and quantification of Metazoan species, when the genomes of all species are known and available on public repositories. This broad objective comes with several specific goals that we list below:

- I. Generate a relevant method for the robust identification of species, starting with insects as a particular case.
- II. Explore the ability of the metagenomic and mito-metagenomic methods to identify and recover the original composition of multiple species samples.
- III. Evaluate the performance of metagenomic classifiers to classify reads to species.
- IV. Test strategies to optimize the performance of the bioinformatic pipelines in terms of accuracy and execution times.
- V. Identify the difficulties that future molecular-based studies would face when assessing *real* samples of unknown composition and provide solutions to such difficulties, if possible.

The research done in this thesis is of the exploratory kind, therefore, the specific goals may not be chapter-specific, they are rather selected and blended whenever we deemed appropriate.

The rest of this thesis is structured as follows. Chapter 2 holds the common methodology for the research studies done in this thesis. Chapters 3, 4 and 5 contain one particular study each: the MG and MMG methods are tested in chapters 3 and 4, respectively; chapter 5 presents the results of benchmarking two popular metagenomic classifiers with the MMG approach. Chapter 6 provides a discussion that intertwines the individual experiment from chapters 3, 4 and 5. Finally, chapter 7 summarises the main conclusions and future works. Importantly, the studies presented in chapters 3 and 4 have already been published (Garrido-Sanz *et al.*, 2020, 2021) and

the article corresponding to chapter 5 is currently under review at *BMC Bioinformatics*.
So, parts of the manuscripts are common with chapters 1-5.

2. Methodology

2. Methodology

2.1. General pipeline

In the chapters that follow we applied the metagenomic (MG) method in chapter 3 and the mito-metagenomic (MMG) method in chapters 4 and 5. Albeit specific pipelines were applied in each chapter, using different software, algorithms and steps, the main steps are shared between the three studies (Figure 2.1). Broadly, reference genomes were downloaded from online repositories and query samples were processed through a quality control (QC) filtering. High quality query reads were compared against the reference database. The matching reads were assigned to a species. Finally, read counts were used to recover the species list together with their estimated concentration. Additionally, results can also be filtered in order to get more confident results.

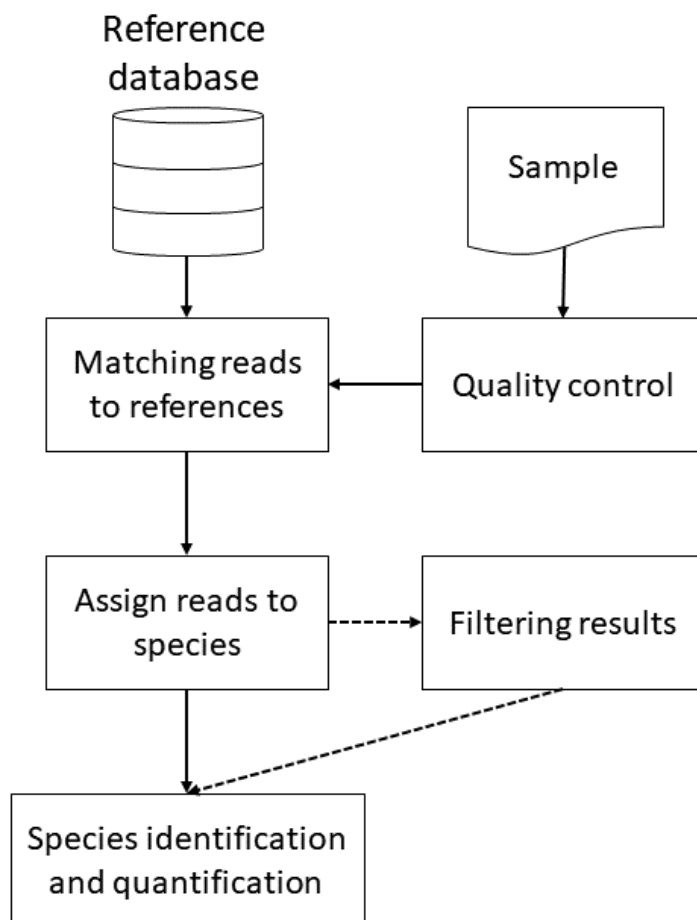


Figure 2.1. Main steps of the metagenomic method. Dashed arrows indicate an optional path.

2.2. Reference genomes: Whole genomes and mitochondrial genomes

In this thesis, we used whole genomes and mitogenomes of insect species as reference sequences. References of both kinds were downloaded from NCBI RefSeq repository. We also considered the GenBank repository to download the mitogenomes of those species having their whole genome available at RefSeq but without a differentiated mitogenome. We used high-quality whole genomes and mitogenomes of as many species as possible to test the ability of the methods to find the selected species among the many others in the reference database. Detailed information about specific reference databases used in every chapter is provided in the Material and Methods section of the corresponding chapter.

2.3. Preparation of samples: Selection of the species, laboratory treatment and quality control

From the list of insect species with available whole genomes, we selected 18 species for shotgun sequencing, based on the availability of fresh specimens (Table 2.1). In general, the specimens were captured alive, but for two dipterans, *Ceratitis capitata* and *Bactrocera oleae*, that came together from fly traps and for the bed bug *Cimex lectularius* that was captured and stored by a pest-control company. The specimens were preserved in 70% ethanol at 4°C for no longer than a few weeks and high-quality DNA was extracted from ca. 20 mg of material of each species. In some cases, multiple extractions were done to obtain the minimum amount of DNA required for library preparation. We used the DNeasy Blood & Tissue Kit (Qiagen) to extract the DNA.

With the DNA extracts, we prepared two kinds of libraries: 22 libraries with DNA of a single species (Table 2.1) and 6 libraries with a mixture of DNA of several species (Table 2.2). All libraries were prepared using the TruSeq DNA PCR-Free LT Kit of Illumina following the manufacturer's instructions (Ref. 15037063). Libraries were subsequently

sequenced using an Illumina MiSeq with the 2x150 chemistry in three different runs, two runs for single-species (Table 2.1) and one for mixed-species libraries (Table 2.2). Four species (*Drosophila melanogaster*, *D. mojavensis*, *D. virilis* and *Linepithema humile*) were sequenced twice in single-species libraries (using different DNA extractions in all cases and different populations for *D. melanogaster* and *L. humile*) to evaluate the repeatability of the method. The same extracts used for the first run of single-species libraries were also used to create six artificial mixed-species libraries of 8-9 species at known relative DNA concentrations to test the ability of the method to estimate the relative species abundance (RSA) (Table 2.2).

Table 2.1.1. Summary table of the species in single-species libraries; the first run (run no. 1) was performed in September 2016 and the second (run no. 2) in July 2018.

Run-Lib.	Species	Order	Family	Cultured/Wild	Origin (Country)
1-1	<i>Papilio machaon</i>	Lepidoptera	Papilionidae	Wild	Spain
1-2	<i>Drosophila virilis</i>	Diptera	Drosophilidae	Cultured	Spain
1-3	<i>Drosophila melanogaster</i>	Diptera	Drosophilidae	Cultured	Spain
1-4	<i>Drosophila mojavensis</i>	Diptera	Drosophilidae	Cultured	Spain
1-5	<i>Bactrocera oleae</i>	Diptera	Tephritidae	Wild	Spain
1-6	<i>Linepithema humile</i>	Hymenoptera	Formicidae	Wild	Spain
1-7	<i>Acromyrmex echinator</i>	Hymenoptera	Formicidae	Cultured	Denmark
1-8	<i>Bombus terrestris</i>	Hymenoptera	Apidae	Wild	Spain
1-9	<i>Apis mellifera</i>	Hymenoptera	Apidae	Wild	Spain
1-10	<i>Acyrtosiphon pisum</i>	Hemiptera	Aphididae	Cultured	USA
2-1	<i>Atta colombica</i>	Hymenoptera	Formicidae	Cultured	Denmark
2-2	<i>Bemisia tabaci</i>	Hemiptera	Aleyrodidae	Cultured	Spain
2-3	<i>Cimex lectularius</i>	Hemiptera	Cimicidae	Wild	Spain
2-4	<i>Drosophila melanogaster</i>	Diptera	Drosophilidae	Cultured	Spain
2-5	<i>Drosophila mojavensis</i>	Diptera	Drosophilidae	Cultured	Spain
2-6	<i>Drosophila virilis</i>	Diptera	Drosophilidae	Cultured	Spain
2-7	<i>Drosophila suzukii</i>	Diptera	Drosophilidae	Cultured	Spain
2-8	<i>Linepithema humile</i>	Hymenoptera	Formicidae	Wild	Spain
2-9	<i>Plutella xylostella</i>	Lepidoptera	Plutellidae	Wild	Spain
2-10	<i>Solenopsis invicta</i>	Hymenoptera	Formicidae	Wild	Argentina
2-11	<i>Vollenhovia emeryi</i>	Hymenoptera	Formicidae	Wild	Japan
2-12	<i>Wasmannia auropunctata</i>	Hymenoptera	Formicidae	Wild	Spain

Table 2.2. Summary table of the species and their relative concentrations in mixed-species libraries; run (run no. 3) was performed in September 2016.

Library	<i>Acromyrmex echinator</i>	<i>Acyrtosiphon pisum</i>	<i>Apis mellifera</i>	<i>Bactrocera oleae</i>	<i>Bombus terrestris</i>	<i>Drosophila melanogaster</i>	<i>Drosophila mojavensis</i>	<i>Linepithema humile</i>	<i>Papilio machaon</i>
1	0.5010	0.0078	0.0626	0.2505	0.1252	0.0157	0.0313	0.0039	0.0020
2	0.2505	0.0020	0.1252	0.5010	0.0626	0.0313	0.0157	0.0078	0.0039
3	0.3039	0.0389	0.1088	0.2158	0.1532	0.0548	0.0772	0.0276	0.0196
4	0.2158	0.0196	0.1532	0.3039	0.1088	0.0772	0.0548	0.0389	0.0276
5	0.2127	0.0747	0.1261	0.1787	0.1501	0.0890	0.1059		0.0628
6	0.1787	0.0628	0.1501	0.2127	0.1261	0.1059	0.0890		0.0747

The target concentration of each species in the mixtures was calculated using a geometric law of parameter k (Magurran 2004): the abundance of the most abundant species is k ; the abundance of the second most abundant one is $k \cdot (1-k)$ and so on. The higher the k value, the greater the difference in concentration between species. In the mixtures, we used the following values of k : 0.50 (libraries no. 1 and no. 2), 0.30 (no. 3 and no. 4) and 0.20 (no. 5 and no. 6). In each library, the order of the species in terms of abundance varied, but several species were only used at low or at high DNA concentrations because of a limitation on the amount of DNA available for each species. Libraries 1-4 contained DNA of nine species and libraries no. 5 and no. 6 of eight species (Table 2.2). It is important to notice that here the RSA is the relative DNA concentration of the species in the mixture, not their relative biomass. Consequently, all sources of variation between the fresh biological material and the extracted DNA (e.g., DNA-to-biomass ratio) are ignored (Matesanz *et al.*, 2019; Tang *et al.*, 2015).

For the sequences generated in both sample types, we assessed the quality of raw reads with FastQC v0.11.7 (Andrews 2015). Trimmomatic v0.36 (Bolger *et al.*, 2014) was subsequently used to trim the reads to the specified length of 150 bp and to discard those shorter than 140 bp. This QC step was applied for sample libraries of single-end reads in chapters 3 and 5 libraries, whereas in chapter 4 libraries consisted of paired-end reads; in this last case, only pairs of reads were kept for downstream analyses.

2.4. Classification of reads to species: Matching and assignment steps

To classify reads to species we implemented three different methods: (1) BLASTn (Altschul *et al.*, 1990) plus MEGAN6 (Huson *et al.*, 2016) (here after referred as BM); (2) Kraken2 (Wood *et al.*, 2019) (here after referred as K2); and (3) BWA (Li & Durbin 2009) followed by the γ - δ algorithm (here after referred as B γ δ). All three classifiers implement a three-steps strategy: first, the building of a custom reference database; second, reads from samples are compared to the reference database; third, reads are associated to taxa. The BM and K2 classifiers are applied in chapter 5, while B γ δ is used in chapters 3 and 4.

2.4.1. BLASTn plus MEGAN6

Classification of the sequencing reads with BLASTn v2.10.0 (Altschul *et al.*, 1990) requires two main steps: database building and alignment of reads. For creating the reference database, all reference sequences were stored in a single file, and this file was subsequently used to build the database with the default parameters. Next, the reads were aligned to the database and output format was set to number 6.

For every read, BLASTn returns multiple hits, that are the best alignments of the query read to reference sequences. To find the best assignment of every read, the LCA algorithm is implemented with MEGAN6 v6.18.11 (Huson *et al.*, 2016) with the naive algorithm. As we are only interested in species level assignment, we discarded any classification at higher taxonomic ranks.

2.4.2. Kraken2

For Kraken2 v2.0.8-beta (Wood *et al.*, 2019), the database creation requires two main steps: adding the reference sequences to the database and the subsequent building of a compact hash table. Importantly, the header line of the reference sequences must fulfil Kraken2's format "kraken:taxid|XXX", where XXX have to be replaced by the NCBI taxonomy identifier (taxID). Kraken2 starts the creation of its database by extracting all k -mers (*i.e.*, substrings from sequences of length k , by default $k=35$) from the reference sequence and assigns each of them to a single taxon with the LCA algorithm. Thereby, if the k -mer is uniquely found in a single species, it is assigned to that species; whereas, if the k -mer is found in more than one reference species, it is assigned to the LCA of those species. The k -mer/LCA pairs are used to create a memory efficient hash table.

Once the database is created, query reads are classified by splitting each read in k -mers, which are looked up into the hash table. Thus, for instance, a read of 150bp length will have 116 k -mers of default length 35bp; each of those k -mers is assigned independently, therefore, the final set of LCAs may compress various taxa with different taxonomic ranks. Kraken2 subsequently uses a scoring scheme to assign the query read to a single taxon. Again, as we are only interested in species rank, assignments at higher taxonomic ranks are removed.

2.4.3. BWA plus γ - δ algorithm

We used a third classification method based on the combination of the BWA mapper (Li & Durbin 2009) and a newly developed algorithm that we called γ - δ (Figure 2.2). BWA (v. 0.7.15-r1140) was used to map each read to all reference genomes individually. For each reference, the BWA index was constructed using the index command with default settings. The mapping was conducted with the *mem* algorithm (Li 2013) with the default options. As the mapping of a read was performed independently for each reference, we acquired as many alignment files as references

used. We used SAMtools (Li *et al.*, 2009) to remove reads that did not map to any reference.

In general, one-read maps into several reference genomes (*e.g.*, homologous sequences in several species), so an algorithm is needed to decide between alternative assignments of a read (Figure 2.2). In metabarcoding (MB) and MG studies, reads are commonly assigned to taxa using the LCA (as previously seen in MEGAN6 and Kraken 2). The LCA algorithm intends to extract as much taxonomic information as possible from a set of reads, so, if one-read maps well enough in two or more different reference genomes, the LCA assigns the read to their common ancestor in the phylogenetic tree. In this thesis, the interest is different, as we intend to only use genomic regions that are useful for species-level identification; thus, if one read maps well enough in two different reference genomes, we deem it as non-informative and ignore it, instead of assigning it to its common genus or family. Having this objective in mind, we devised the simple γ - δ algorithm to accomplish it while parsing multiple SAM files returned by BWA (Figure 2.3).

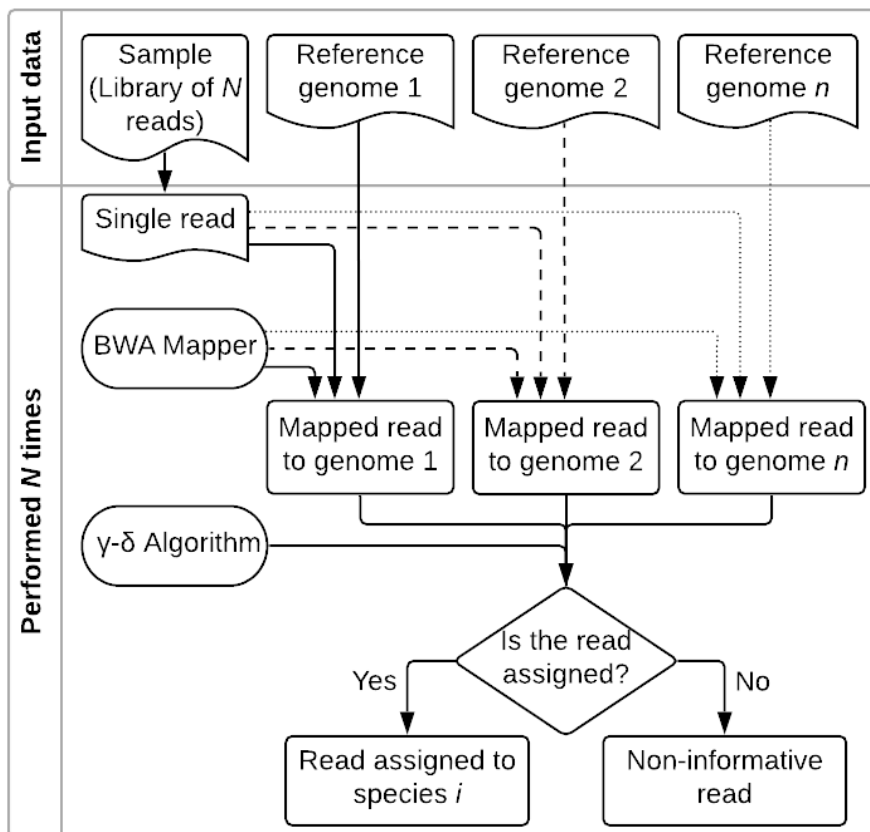


Figure 2.2. Flow diagram of the computational pipeline used in the $B\gamma\delta$ method. At the top, input data and, below it, the steps and tools needed for the identification procedure.

The γ - δ algorithm

Basically, what the γ - δ algorithm does is to assign a read to species i when it maps well to species i and bad to the rest of species; on the contrary, when a read maps well in two or more species, we declare it non-informative. The algorithm uses two thresholds (γ and δ) to determine the read mapping quality. In all cases, γ and δ satisfy that $1 > \gamma$, $\delta > 0$ and $\gamma > \delta$. The quality of the mapping is measured as the mapping ratio A and it is calculated as the sum of read's matching nucleotides to the target sequence (n_m), divided by the total number of nucleotides in the alignment (n_t) (Eq. 2.1).

$$A = \frac{n_m}{n_t}$$

Equation 2.1

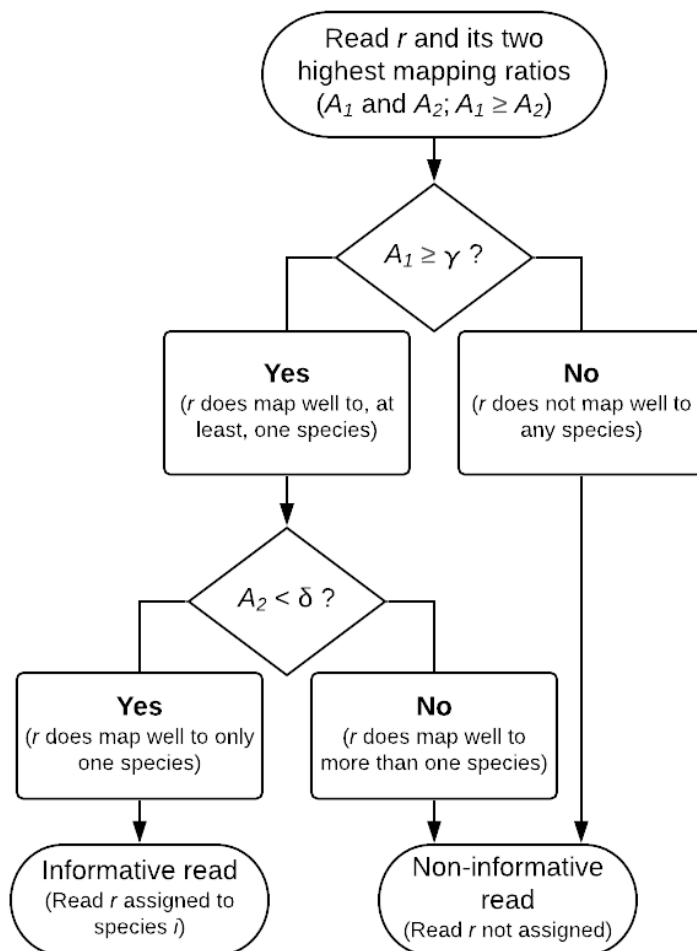


Figure 2.3. Flow diagram of the γ - δ algorithm. Only the two highest mapping ratios to two reference genomes of a single read are required. In the figure, it is assumed that the highest mapped ratio A_1 belongs to the reference genome i .

Even though a read r can be assigned to many references, the γ - δ algorithm only needs the two highest mapping ratios. Let A_1 and A_2 be the highest and second highest mapping ratios of r . We assume that A_1 corresponds to the mapping ratio of r to the reference genome of species i . Then the assignment algorithm works in the following way (Figure 2.3):

- If $A_1 < \gamma$, then r is non-informative (because it does not map well enough to any species).
- If $A_1 \geq \gamma$ and $A_2 \geq \delta$, then r is non-informative (because it maps too well to two different species).
- If $A_1 \geq \gamma$ and $A_2 < \delta$, then r is informative, and it is assigned to species i (because it maps well enough in one species and not in any other one).

2.5. Contaminant species

In *real* samples, contaminants are hard to detect, but in our libraries, they are not. If a read can be genuinely attributed to cross-contamination, either from the laboratory, or from the field sampling, then it is probably a genuine contamination problem and could be removed from the analysis. In our case, these contaminant species are species sequenced on the same sequencing run (Table 2.1 and Table 2.2) and species handled in the lab but finally not sequenced, these species included *Ceratitis capitata* (Diptera: Tephritidae), *Bombyx mori* (Lepidoptera: Bombycidae) and *Tribolium castaneum* (Coleoptera: Tenebrionidae) in the first single-species run and *B. mori* again in the second run; in the mixed-species samples, these species are the same species than in the first single-species run, plus *Drosophila virilis*, and also *Linepithema humile* in libraries *no.* 5 and *no.* 6.

2.6. Statistics analysis

All statistical analyses were performed using R v3.4.2 (R Core Team 2016) in RStudio v1.0.143 (RStudio Team 2015). Permutational analysis of variance (PERMANOVA; Anderson 2001) and subsequent pairwise comparison with Bonferroni correction were conducted using the package “vegan” (Oksanen *et al.*, 2018). Z-test for the comparison of correlations was conducted using the package “psych” (Revelle 2021). Plots were created using the packages “ggplot2” (Wickham 2016) and “ggpubr” (Kassambara 2018).

3. Estimation of the relative abundance of species in artificial mixtures of insects using low-coverage shotgun metagenomics

Chapter published as: Garrido-Sanz L, Senar MÀ, Piñol J (2020) Estimation of the relative abundance of species in artificial mixtures of insects using low-coverage shotgun metagenomics. Metabarcoding and Metagenomics, 4: e48281. <https://doi.org/10.3897/mbmg.4.48281>

3. Estimation of the relative abundance of species in artificial mixtures of insects using low-coverage shotgun metagenomics

3.1. Abstract

Amplicon metabarcoding is an established technique to analyse the taxonomic composition of communities of organisms using high-throughput DNA sequencing, but there are doubts about its ability to quantify the relative proportions of the species, as opposed to the species list. Here, we bypass the enrichment step and avoid the PCR-bias, by directly sequencing the extracted DNA using shotgun metagenomics. This approach is common practice in prokaryotes, but not in eukaryotes, because of the low number of sequenced genomes of eukaryotic species. We tested the metagenomics approach using insect species whose genome is already sequenced and assembled to an advanced degree. We shotgun-sequenced, at low-coverage DNA, 18 species of insects in 22 single-species and six mixed-species libraries and mapped the reads against 110 reference genomes of insects. We used the single-species libraries to calibrate the process of assignation of reads to species and the libraries created from species mixtures to evaluate the ability of the method to quantify the relative species abundance. Our results showed that the shotgun metagenomic method is easily able to set apart closely-related insect species, like four species of *Drosophila* included in the artificial libraries. However, to avoid the counting of rare misclassified reads in samples, it was necessary to use a rather stringent detection limit of 0.001, so species with a lower relative abundance are ignored. We also identified that approximately half the raw reads were informative for taxonomic purposes. Finally, using the mixed-species libraries, we showed that it was feasible to quantify with confidence the relative abundance of individual species in the mixtures.

Keywords Eukaryotes, Metazoa, genome skimming, PCR-free, mock sample, sequenced genomes

3.2. Introduction

DNA metabarcoding (MB) is currently the method of choice for DNA-based surveys of biodiversity, but its dependency on PCR to amplify the “barcodes” hampers the ability of the method to estimate the relative proportion of the species (Piñol *et al.*, 2019). Alternatively, metagenomics (MG) is a technique used to explore community’s diversity by the direct analysis of the entire genetic content within a sample (Mardis 2008). In this approach, DNA sequences across the genome provide information about the taxonomic profile and functions of the organisms in the community under study (Venter *et al.*, 2004). As this method does not sequence specific genetic markers, the PCR associated biases are avoided. Albeit the huge potential of MG approach on biodiversity assessment, this technique is currently constrained to prokaryote studies because the genome of most eukaryotic species is still to be sequenced.

In this chapter, we imagine a world in which the complete genomes of all the species are known to implement a shotgun MG method on Metazoan species. We simulate this future world by preparing a reference database of insect species whose genome is already assembled to an advanced degree and available at the NCBI RefSeq repository. We shotgun-sequenced DNA from some of these species in low-coverage single-species libraries, prepared without any PCR step, and calibrated our new approach, BWA followed by the γ - δ algorithm (hereafter, $B\gamma\delta$) to go from raw reads to species assignation. Subsequently, we apply the bioinformatic pipeline, on controlled mixtures of insects to see if the method produces a quantitative estimate of the insect species present.

This exercise is a preliminary test of the difficulties likely to be faced in the future when an important number of complete genomes becomes available. In particular, we

address here the following questions: (1) Is the MG method useful to set apart closely related insect species; (2) What is the proportion of reads that is truly informative for species identification; (3) How many reads are necessary to achieve a reasonable level of confidence to provide quantitative estimates of the relative species abundance (RSA)?

3.3. Material and Methods

3.3.1. Reference genomes

We considered all insect species whose genome was sequenced, assembled and available at the NCBI RefSeq Database on 2nd August 2018. In total, 115 representative genomes of insect species were downloaded; of those genomes, five were removed for different reasons (Table S3.1). The remaining 110 species belonged to 7 orders and 43 families; 28 of them were of the genus *Drosophila*.

3.3.2. Selection of species and preparation of the DNA libraries

From this group of 110 species with whole genomes, we selected 18 species (Table 3.1). With the 18 insect species, two different kinds of samples were generated: The single-species samples, containing one species each, and the mixed species samples, containing 8-9 species each at relative known concentration (Table 2.2). The single-species libraries were used to calibrate the bioinformatic pipeline that assigns reads to species; the mixed-species libraries were used to test the ability of the calibrated method to estimate the relative abundance of individual species in mixtures. Detailed information regarding specimens' collection, laboratory treatment and quality control are provided at subchapter 2.3. Preparation of samples: Selection of the species, laboratory treatment and quality control from chapter 2. Methodology.

Table 3.1. Summary information of the single-species libraries. In chapter 3 sequenced libraries were treated as single-end reads samples.

Run-Library	Code	Species	Number of raw reads (single-end)	Number of reads after QC step	Number of reads after mapping step	Genome coverage
1-1	PM	<i>Papilio machaon</i>	217,260	216,425	208,832	0.117
1-2	DV	<i>Drosophila virilis</i>	2,357,451	2,327,325	2,088,898	1.717
1-3	DMe	<i>Drosophila melanogaster</i>	1,141,884	1,137,997	1,094,403	1.195
1-4	DMo	<i>Drosophila melanogaster</i>	834,212	823,555	787,688	0.646
1-5	BO	<i>Bactrocera oleae</i>	290,498	288,973	279,835	0.093
1-6	LH	<i>Linepithema humile</i>	711,171	701,630	683,311	0.486
1-7	AE	<i>Acromyrmex echinator</i>	116,597	112,337	110,086	0.059
1-8	BT	<i>Bombus terrestris</i>	997,469	994,493	972,727	0.603
1-9	AM	<i>Apis mellifera</i>	631,194	618,649	607,965	0.378
1-10	AP	<i>Acyrtosiphon pisum</i>	342,344	299,234	282,940	0.092
2-1	ACo	<i>Atta colombica</i>	1,636,355	1,634,216	1,607,703	0.845
2-2	BTa	<i>Bemisia tabaci</i>	1,256,606	1,251,241	1,170,057	0.307
2-3	CL	<i>Cimex lectularius</i>	1,753,361	1,737,768	1,703,285	0.515
2-4	DMe	<i>Drosophila melanogaster</i>	1,454,804	1,452,075	1,428,616	1.523
2-5	DMo	<i>Drosophila melanogaster</i>	898,735	893,786	820,019	0.697
2-6	DV	<i>Drosophila virilis</i>	668,442	666,621	619,733	0.488
2-7	DSu	<i>Drosophila suzukii</i>	1,255,192	125,1280	1,178,528	0.811
2-8	LH	<i>Linepithema humile</i>	1,082,202	1,078,804	1,047,744	0.742
2-9	PXY	<i>Plutella xylostella</i>	2,125,062	2,122,603	1,913,733	0.813
2-10	SI	<i>Solenopsis invicta</i>	1,830,687	1,825,077	1,772,743	0.695
2-11	VE	<i>Vollenhovia emeryi</i>	1,743,917	1,740,846	1,679,101	0.912
2-12	WA	<i>Wasmannia auropunctata</i>	1,667,606	1,647,411	1,613,371	0.772

Here, we did not use the paired-end reads provided by the Illumina sequencer, but only the first set of single-end reads (the R1 FASTQ files), because, in many eDNA applications, the fragments were rather short, so the advantage of having paired reads in longer fragments was reduced in actual samples. Considering the sequencing depth and the genome size of the studied species, the mean coverage obtained was below 1 (Table 3.1). Therefore, our approach can be qualified as low-coverage shotgun MG.

3.3.3. Classification of reads to species

To classify reads to species, we applied the $B\gamma\delta$ pipeline described in the 2. Methodology chapter (2.4.3. BWA plus γ - δ algorithm) and at supplementary material (Methodology S3.1). Briefly, BWA was used for matching query reads to reference sequences. Unmapped reads were eliminated from the SAM files with SAMtools. The reported alignments were parsed with the γ - δ algorithm to return the best assignment for every read.

As this is the first time that the γ - δ algorithm is used, the best values of γ and δ are unknown. For this reason, we used the single-species libraries to find the best combination of γ and δ . The reads were divided into a training set (75% of the values chosen randomly) to find the best γ - δ and a test set (the remaining 25% of reads) to independently calculate the goodness of fit of the model. The tested values of γ and δ were all the combinations of $\gamma = \{0.99, 0.98, 0.97\}$ and $\delta = \{0.98, 0.97, 0.96\}$ where $\gamma > \delta$.

3.3.4. Detection limit

The γ - δ algorithm produced a list of species assigned to each read of a library. In single-species libraries, ideally, all reads should belong to the same species (from now on, the focal species). However, detection of additional species could occur for several

reasons, such as contamination from the lab, sequencing errors and even tag jumping between multiplexed libraries (Schnell *et al.*, 2015).

As stated previously in subchapter 2.5. **Contaminant species**, contamination is easy to detect in our samples, because we know the species handled in the lab, so we removed them from the analysis. The other kinds of wrongly assigned species likely produce a very low number of reads. The simplest way to deal with them is to set a detection limit (ϵ), so the species with a proportion of reads lower than ϵ are ignored. Here, we present results using the detection limits of 10^{-2} , 10^{-3} and 10^{-4} .

3.3.5. Selection of best values of γ , δ and ϵ

With the single-species libraries, we used three different criteria to decide which values of γ , δ and ϵ provided best results. The most important one was that the number of species reported had to be one for single-species libraries. In addition, we wanted to maximise the proportion of reads assigned to the focal species (*i.e.*, the precision) and the relative proportion of informative reads (RPIR) assigned to any species.

In practical terms, we first fixed the ϵ parameter. Next, we compared the different γ - δ combinations using the PERMANOVA test (Anderson 2001), followed by a *post hoc* multiple comparison with the Bonferroni test.

The final output of the above analysis is a combination of values of γ , δ and ϵ that were best for the single-species libraries analysed in this study. The goodness of fit of this set of parameters was evaluated using the test set, *i.e.*, the remaining 25% of reads were not used for the calibration.

As will be shown in the results, using the best values of γ , δ and ϵ , we still found in the single species libraries some reads that were wrongly assigned to non-focal species. To

explore the identity of all these misidentified reads, we blasted them (or a subset of 100 reads when the total number of misclassified reads was higher) with megablast (Morgulis *et al.*, 2008) against the NCBI nucleotide collection (*nt*) database (Wheeler *et al.*, 2007).

3.3.6. Quantification of the relative proportion of the species

Mixed-species mock samples were processed following the same computational pipeline as outlined above (Figure 2.2), using the best combination of γ , δ and ϵ values determined in the previous step. The estimated proportion of reads, assigned to each reference genome, was calculated without considering the rejected reads (not mapped or not assigned reads). This estimated proportion was compared with the actual one (Table 2.2), using the Pearson correlation coefficient.

3.3.7. Rarefaction of the input samples

As can be seen from the results, we obtained a good quantitative estimation of species abundance in all mixed-species libraries. However, from a practical point of view, it would be interesting to investigate if sequencing depth can be reduced and a robust quantitative estimation of the RSA maintained. Thus, more libraries could be multiplexed in a single run and so reduce the overall cost. To evaluate this possibility, we ran the same computational pipeline as before (using the chosen parameters γ , δ and ϵ) but randomly reducing the number of reads to a proportion of 0.1, 0.01 and 0.001 of the original ones. Each simulation was repeated 100 times, using a different random set of reads. Afterwards, we estimated the number and relative abundance of the recovered species in each rarefied sample and calculated the Pearson correlation between the actual and the estimated RSA.

3.3.8. Hardware

We run the entire pipeline on a server with two Intel Xeon E5-2620 v3 processors with six cores each, which allowed a maximum of 24 threads, thanks to their hyper-threading technology.

3.4. Results

3.4.1. Single-species libraries

The 22 libraries prepared from DNA of single-species of insects (Table 3.1) generated $1,136,957 \pm 633,142$ (mean \pm SD) reads, with a coverage of 0.65 ± 0.43 . A proportion of 0.013 ± 0.026 reads were eliminated in the trimming step and 0.042 ± 0.026 in the mapping step, so a proportion of 0.95 ± 0.04 of the raw reads remained for further analysis.

The most important characteristic of these libraries is that the number of species recovered, in theory, must be one. With these libraries, we parameterised two aspects of MG species assignment: first, what is the appropriate detection limit (ϵ) for removal of spurious species (*i.e.*, cut-off for minimum proportion of reads for species to be retained) and second, which are the best values of γ and δ .

For the detection limit ϵ , we describe in detail the process followed for the analysis of the first run of single-species libraries using the values of $\gamma = 0.99$ and $\delta = 0.98$. The rest of the single-species libraries and all the other γ - δ combinations produced relatively similar results and are provided as supplementary material (Tables S3.2 and S3.3).

After the application of the γ - δ algorithm, there were 19.6 ± 8.0 reference genomes (species) per library (Table 3.2). The most abundant one was the focal species, always

above 0.98, except for *B. oleae* (0.93). Obviously, this high number of recovered species is unacceptable for single-species libraries.

Some of these additional species were handled in the same lab, but finally were not sequenced because of their poor quality or for other reasons. Thus, they can legitimately appear in the species list because of lab contamination or tag-jumping. If we eliminate these species (Table 3.2E), the number of species per library is still high at 9.5 ± 11.4 (Table 3.2A-D).

The next step is the removal of the species below a certain detection limit. If the species having a relative proportion below $\epsilon = 0.0001$ were discarded (Table 3.2D), the remaining number of recovered species would be reduced to 3.1 ± 2.6 (Table 3.2A-C). An increase in the detection limit to $\epsilon = 0.001$ reduced the number of recovered species to one, but for *Drosophila virilis* (*Lucilia cuprina*, same order) and *Apis mellifera* libraries (*Apis florea*, same genus) (Table 3.2A-B). A further increase in the detection limit to $\epsilon = 0.01$ eliminated all non-focal species. In summary, the use of a detection limit of $\epsilon = 0.001$ almost eliminates all undesired species from the list (Table 3.2A). A very similar result was observed with the single-species libraries of the second run: again, *L. cuprina* appeared in the library of *D. virilis* and *Atta cephalotes* in the library of *A. colombica* (Table S3.3). Therefore, considering these results, we will use a detection limit of $\epsilon = 0.001$ in all further analyses.

The exploration of the misidentified reads in Table 3.2 against the NCBI *nt* database produced different kinds of results depending on the species considered (Table S3.4).

(1) The reads assigned to the dipteran *Lucilia cuprina* in the libraries of the three species of *Drosophila* were assigned to bacteria, mostly *Providencia* sp. and *Morganella* sp. (2) Ninety-nine percent of the reads assigned to *Apis florea* in the libraries of *Bombus terrestris* and *A. mellifera*, were rRNA and other kinds of RNA. (3) Many reads of *Drosophila melanogaster* and *D. mojavensis*, assigned to a wrong species of *Drosophila*, mapped into bacteria of the genus *Lactobacillus* and *Acetobacter* and a few were RNAs or transposons. (4) All seven reads of *Acromyrmex echinator*, wrongly assigned to *Vollenhovia emeryi*, mapped to the bacteria *Wolbachia*

sp. (5) Approximately a quarter of the wrongly assigned reads of *Bombus terrestris* to *B. impatiens* were RNAs of *Bombus* or *Apis* and (6) About half of the reads of *Apis mellifera*, assigned to *A. cerana* and *A. dorsata*, were RNAs (Table 3.2 and Table S3.4).

The final step is to decide which of the tested γ - δ combinations provided better results. First, the number of identified species was closer to 1 for $\gamma = 0.99$ than for $\gamma = 0.98$ or $\gamma = 0.97$ (Figure 3.1A). Even though the combination of $\gamma = 0.98$ and $\delta = 0.96$ was not significantly different from those with $\gamma = 0.99$, that combination had a higher data dispersion of detected species (maximum value is 1 versus 3). Next, we observed neither differences amongst three γ - δ combinations with $\gamma = 0.99$ in the proportion of correctly assigned reads (Figure 3.1B; $p > 0.99$), nor in the proportion of the informative reads (Figure 3.1C; $p > 0.91$). Albeit non-significantly, the combination of parameters $\gamma = 0.99$ and $\delta = 0.98$ was slightly better than for $\delta = 0.97$ or $\delta = 0.96$ and therefore, we will use them in all the following analyses.

The replicated single-species libraries (four species that were analysed in two separate runs) produced remarkably similar results. For example, both libraries of *D. virilis* had *L. cuprina* at a relative concentration higher than $\epsilon = 0.001$; in the other three libraries, only the focal species was recovered above a value of $\epsilon = 0.001$ (see Table S3.5, for a direct comparison of the duplicated single-species libraries).

We tested the quality of the adjusted parameter set $\epsilon = 0.001$, $\gamma = 0.99$ and $\delta = 0.98$ obtained with the training set, using the remaining 25% of reads (*i.e.*, the test set). The results were not statistically different between the test set and the training set for any of the analysed variables (Figure S3.1). Using the test set and the above parameter values, the number of identified species per library was 1.09 ± 0.29 , the proportion of correctly assigned reads was 0.99 ± 0.01 and the RPIR per sample was 0.47 ± 0.15 . It is worth noting that the proposed algorithm with the above parameter set was perfectly able to set apart closely-related species, like the species of *Drosophila* (three in the first run and four in the second one).

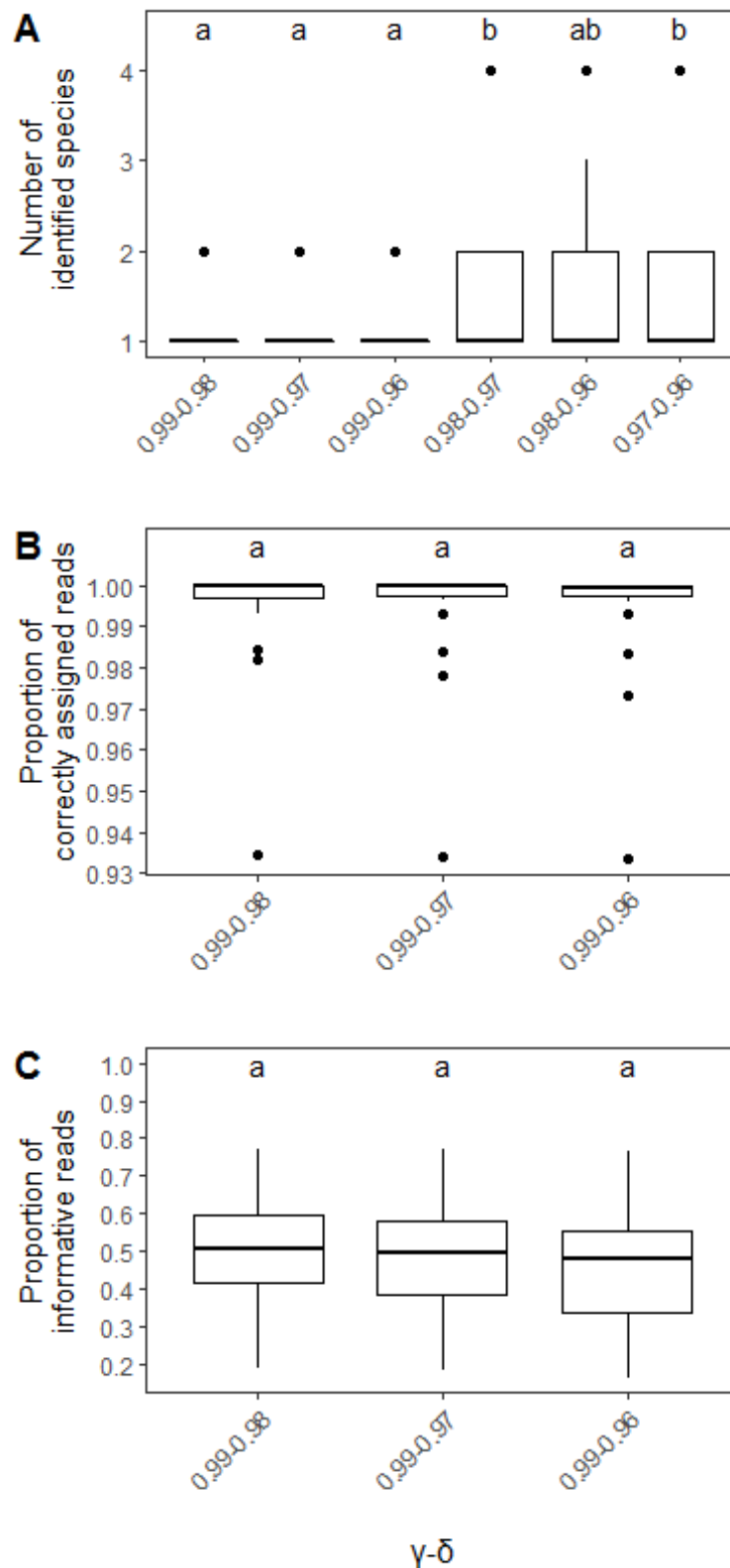


Figure 3.1. Summary boxplots of the 22 single-species libraries used to search the best combination of parameters γ and δ ; in all cases, a detection limit of $\epsilon = 0.001$ was used and contaminant species were discarded. (A) Number of identified species in the library. (B) Proportion of the assigned reads allocated to the right species. (C) RPIR. A different letter at the top of the figures indicates significant differences amongst γ - δ combinations.

Table 3.2. Proportion of reads assigned to species, in parentheses, for each one of the 10 single-species libraries included in the first sequencing run. The species in each library are shown in the header column. Species assignments are divided in each column into blocks: A, species with abundance higher than $\epsilon = 0.01$; B, with abundance between $\epsilon = 0.01$ and $\epsilon = 0.001$; C, with abundance between $\epsilon = 0.001$ and $\epsilon = 0.0001$; D, with abundance below $\epsilon = 0.0001$; E, potential contaminants. Codes of the species as in Table S3.1.

Criteria	Lib. 1 - PM	Lib. 2 - DV	Lib. 3 - DMe	Lib. 4 - DMo	Lib. 5 - BO	Lib. 6 - LH	Lib. 7 - AE	Lib. 8 - BT	Lib. 9 - AM	Lib. 10 - AP
A: above $\epsilon = 0.01$	PM (0.98249)	DV (0.9958)	DMe (0.99812)	DMo (0.99627)	BO (0.93349)	LH (0.99877)	AE (0.99862)	BT (0.99777)	AM (0.99627)	AP (0.99557)
B: from $\epsilon = 0.01$ to 0.001	LC (0.00259)									
C: from $\epsilon = 0.001$ to 0.0001	Dsi, DSe, DS, DSu, LC									
D: below $\epsilon = 0.0001$	NVi, PP	CQ, NVi, MP, OT, CS, Cl, BL, PXY, Bi, Dar, DF, DSe, DSu, DT, MDe, DO, Nlu	DNo, DY, CL, DW, CQ, DEr, NVi, MP, DF, DO, NLu, Deu	Dar, Del, DB, LC, DO, DEu, DF, DBi	NVi, MDe, LC	BT, CF, SI, CC, TZ	WA, DNo, Bi, TS	SI, LC, CCal	EM, DBi	DN, MS, F
E: potential contaminants:	DV (0.01359)	BM (0.00078)	BM (0.00046)	BM (0.00036)	CCap (0.0655)	AE (0.0006)	BM (0.00055)	BM (0.00078)	BM (0.00062)	DMe (0.00288)
	BM (0.0021)	TCa (0.00025)	DV (0.00031)	DV (0.00008)	BM (0.00043)	BM (0.00035)	LH (0.00019)	DV (0.00025)	DMe (0.00014)	BM (0.00071)
Bold: species handled in the lab and sequenced	DMe (0.00057)	DMe (0.00012)	LH (0.00004)	DMe (0.00004)	BT (0.00015)	DV (0.00005)	DV (0.00018)	DMe (0.00024)	DV (0.00006)	LH (0.00012)
	BT (0.00047)	PM (0.00011)	BO (0.00004)	LH (0.00001)	DV (0.00012)	DMe (0.00005)	DMe (0.00005)	DMo (0.00008)	BT (0.00004)	DV (0.00011)
<i>Italic:</i> species handled in the lab but not sequenced	DMo (0.0002)	LH (0.00008)	BT (0.00003)	BT (0.00001)	LH (0.00012)	AP (0.00003)	DMo (0.00005)	LH (0.00004)	BO (0.00004)	BT (0.00009)
	AE (0.00013)	DMo (0.00003)	DMo (0.00003)	PM (0.00001)	DMe (0.00006)	AM (0.00003)	BT (0.00003)	BO (0.00004)	LH (0.00003)	DMo (0.00009)
	LH (0.0001)	BT (0.00002)	PM (0.00002)	AM (0.00001)	PM (0.00002)	DMo (0.00002)	TCa (0.00003)	AM (0.00002)	DMo (0.00002)	AM (0.00003)
	AM (0.00006)	AE (0.00001)	AE (0.00001)	BO (<0.00001)	DMo (<0.00001)	BO (<0.00001)	BO (0.00001)	AP (0.00002)	PM (0.00002)	BO (0.00003)
	AP (0.00006)	AM (0.00001)	AM (0.00001)	AE (<0.00001)		PM (<0.00001)		PM (0.00001)	AP (0.00001)	PM (0.00003)
	CCap (0.00006)	BO (0.00001)	AP (0.00001)	AP (<0.00001)		AE (0.00001)	CCap (<0.00001)	AE (0.00001)	CCap (<0.00001)	AE (0.00001)
	BO (0.00003)	AP (<0.00001)	TCa (<0.00001)	CCap (<0.00001)		CCap (<0.00001)		CCap (<0.00001)		
	CCap (<0.00001)									
Total number of species	14	31	30	32	12	15	14	17	16	15

3.4.2. Mixed-species libraries

The six libraries prepared from DNA of multiple species of insects generated 1,688,044 \pm 212,119 reads (Table 3.3). A proportion of 0.003 \pm 0.001 reads were eliminated in the trimming step and of 0.035 \pm 0.012 in the mapping step, so there remained a proportion of 0.962 \pm 0.012 of the raw reads for further analysis.

As in the single-species libraries, in the mixed-species libraries, there were also contaminants handled in the laboratory, but not sequenced. To the already mentioned *C. capitata*, *B. mori* and *T. castaneum*, we must add *D. virilis* (that was not sequenced in any mixed-species library) and *L. humile* (not sequenced in libraries *no.* 5 and *no.* 6). As we did before, we eliminated all these species as genuine contaminants (Table 3.4E and Table S3.6). Even after removing these contaminants, the number of species in the mixture was still very high (47-55), so it was mandatory to apply the proposed detection limit of $\epsilon = 0.001$ values. By doing this, we recovered all the expected species in the mixtures, nine in libraries *no.* 3 and *no.* 4 and eight in libraries *no.* 5 and *no.* 6; Table 3.4A-B), except in libraries *no.* 1 and *no.* 2 where *P. machaon*, the species with the actual lowest abundance in the mixture, was present in a proportion slightly below $\epsilon = 0.001$ (Table 3.4C). The correlation coefficient between actual and estimated RSA was statistically significant in all mixtures (Figure 3.2), so the method was able to quantify the relative proportions of the species. The fitting was better for high values of k (more difference in the relative abundance of species; libraries *no.* 1 and *no.* 2, $k = 0.50$) than for low values of k (less difference in the relative abundance; libraries *no.* 5 and *no.* 6, $k = 0.20$) (Figure 3.2).

The total processing time varied between 54 min (library *no.* 6, 1.3 raw million reads) and 1 h 16 min (library *no.* 1, 1.9 raw million reads), most of it (89%) consumed by the mapping of the reads into the reference genomes and very little (3-4%) by the γ - δ algorithm (see Table S3.7 for the processing time of each step of the pipeline for six mixed-species libraries).

Table 3.3. Summary information of the number of reads in mixed-species libraries (Table 2.2). In this study sequenced libraries were treated as single-end reads samples.

Library	Number of raw reads	Number of reads after quality control step	Number of reads after mapping step
1	1,897,302	1,894,232	1,842,838
2	1,674,754	1,671,172	1,597,601
3	1,887,006	1,881,011	1,829,291
4	1,557,348	1,552,940	1,511,536
5	1,767,384	1,761,238	1,709,991
6	1,344,467	1,339,919	1,264,242

Table 3.4. Relative proportion of assigned reads to species, in parentheses, for each one of the six mixed-species libraries of Table 2.2 after applying γ - δ algorithm with parameters $\gamma = 0.99$ and $\delta = 0.98$. Codes of the species as in Table S3.1. In bold, the species whose DNA was actually put in the mixture.

Criteria	Lib. 1	Lib. 2	Lib. 3	Lib. 4	Lib. 5	Lib. 6
A: above $\epsilon = 0.01$	AE (0.64971)	AE (0.37924)	AE (0.44486)	AE (0.30977)	AE (0.32342)	AE (0.27894)
	BO (0.17167)	BO (0.37805)	BO (0.16181)	BO (0.2496)	BO (0.1536)	BO (0.18874)
	BT (0.06582)	AM (0.12349)	AM (0.10563)	AM (0.15797)	AM (0.13569)	AM (0.16409)
	AM (0.05033)	BT (0.03834)	BT (0.08595)	BT (0.06737)	DMo (0.12864)	DMo (0.10473)
	DMo (0.02998)	DMe (0.02339)	DMo (0.0789)	DMo (0.05985)	BT (0.09288)	DMe (0.08641)
	DMe (0.01052)	DMo (0.01707)	DMe (0.03823)	DMe (0.05742)	DMe (0.0694)	BT (0.082)
B: from $\epsilon = 0.01$ to 0.001	AP (0.00577)	LH (0.00997)	PM (0.00434)	PM (0.00697)		
	LH (0.00245)	AP (0.0017)				
C: from $\epsilon = 0.001$ to 0.0001	PM (0.00082)	PM (0.0009)	AF (0.00024)	AF (0.00034)	AF (0.00029)	AF (0.00036)
	VE (0.00013)	AF (0.00026)	VE (0.00011)			
D: below $\epsilon = 0.0001$	WA, BI, Dar, TCo, ACer, DB, TS, Del, LC, DO, AD, DBi, TZ, DEu, MDe, CCal, DSi, ACep, CC, DN, DF, DK, DSe, EM, MP, ACo, BD, BL, CL, DR, DSu, NVi, PH, ZC	WA, ACer, BI, LC, AD, Dar, DB, TS, TCo, MDe, Del, EM, NVi, DO, DSi, DS, DEu, ACo, BL, TZ, ACep, ZC, CQ, DBi, DSe, MP, BD, API, CF, DT, DW, DY, HL, SI, SC	Dar, WA, Del, DB, LC, BI, DO, ACer, DBi, MP, TCo, DSi, AD, TS, DEu, DF, MDe, DS, DSe, DT, ACo, TZ, BD, NVi, ACep, SI, DR, BL, DW, CCal, DN, DSu, DNa, DNo, BA, LD, NL, SL	VE, WA, ACer, Dar, Del, DB, BI, DO, LC, AD, DSi, TS, MP, TCo, DEu, NVi, DF, DSe, MDe, DSu, DY, DBi, DS, ACo, ACep, DNo, CL, PXy, SF, DT, TZ, SI, BL, DW, DN, DNa, CQ, CF, DK, PP	Dar, Del, VE, DB, DO, WA, LC, BI, ACer, MP, DSi, DF, DBi, DEu, AD, TS, TCo, NVi, DS, TZ, DSe, DW, NL, EM, MDe, DSu, SF, DNa, CQ, PP, ACep, DNo, CL, DT, DN, DK, CCal, BA, LD, API, CC, DC, DER	Dar, VE, Del, ACer, DB, DO, LC, WA, BI, DEu, MP, DSi, DF, DS, DBi, AD, MDe, TCo, DSu, TS, DSe, EM, DNo, NVi, NL, DNa, ACep, DT, ACo, DH, TZ, CQ, PP, CL, DN, ZC, HL, AGa, Dan, MS, PR
	E: potential contaminants	CCap (0.01185) BM (0.0005)	CCap (0.02603) BM (0.00112) TCa (< 0.00001)	CCap (0.01131) BM (0.00051) TCa (< 0.00001) DV (< 0.00001)	CCap (0.01743) BM (0.00044)	CCap (0.01064) BM (0.00055) LH (< 0.00001)
Total number of species	47	49	53	52	55	54

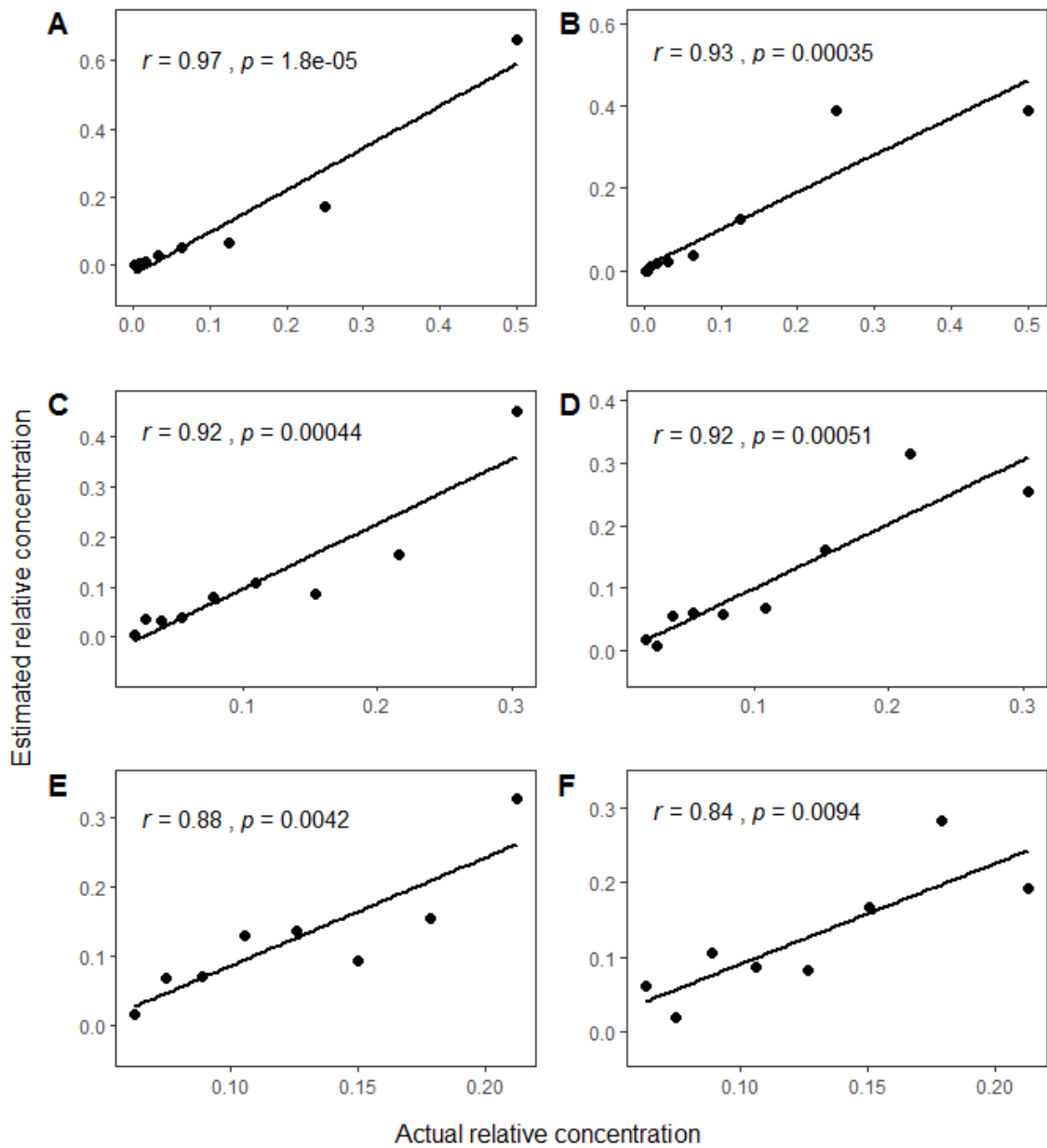


Figure 3.2. Scatter plots between the expected (*i.e.*, as the mixtures were prepared in the lab; Table 2.2) and the estimated species relative abundance (Table 3.4) following the described bioinformatic pipeline. Each plot corresponds to one mixed-species library (A to F corresponds to libraries *no.* 1 to *no.* 6). Each point in the plot indicates one species in the mixture. In each plot, the correlation coefficient (r) and its p -value are also indicated.

3.4.3. Rarefaction of the reads

When only a proportion of 0.1 or even 0.01 of the initial reads was used, the number of recovered species was the same in libraries 3-6 as when all reads were used (Figure 3.3C-F). In libraries *no.* 1 and *no.* 2, there was some discrepancy, but it was caused by the estimated relative abundance of *P. machaon* being sometimes slightly below and sometimes slightly above 0.001 and so our detection limit of $\epsilon = 0.001$ discarded or accepted the species accordingly (Figure 3.3A-B). A further reduction in the proportion of used reads to 0.001 made the number of identified species less predictable (Figure 3.3). However, the correlation coefficient r between the observed and the expected RSA was always significant at all rarefaction levels (Figure 3.4).

3.5. Discussion

MG is a technology devised to obtain both taxonomic and functional gene information for entire communities of organisms (Thomas *et al.*, 2012; Zepeda Mendoza *et al.*, 2015) and its use is more common in prokaryotes than in eukaryotes. Here, we focused on the taxonomic aspect of MG and applied it to Metazoa. We evaluated the technique using artificial mixtures of DNA consisting from one to nine insect species whose complete genome has been sequenced to an advanced degree. The single-species libraries proved to be very useful in showing the limitations of the technique: in these libraries, the number of expected species is one, but we found between 12 and 32 species per library, so it was mandatory to establish a detection limit for a species to be included in the species list. The mixed-species libraries showed that the technique is perfectly able to quantitatively determine the relative abundance of individual species in mixtures. Given the scarcity of assembled genomes of Metazoa, the proposed methodology is a proof of concept of the MG approach rather than a method to be applied immediately to actual environmental samples.

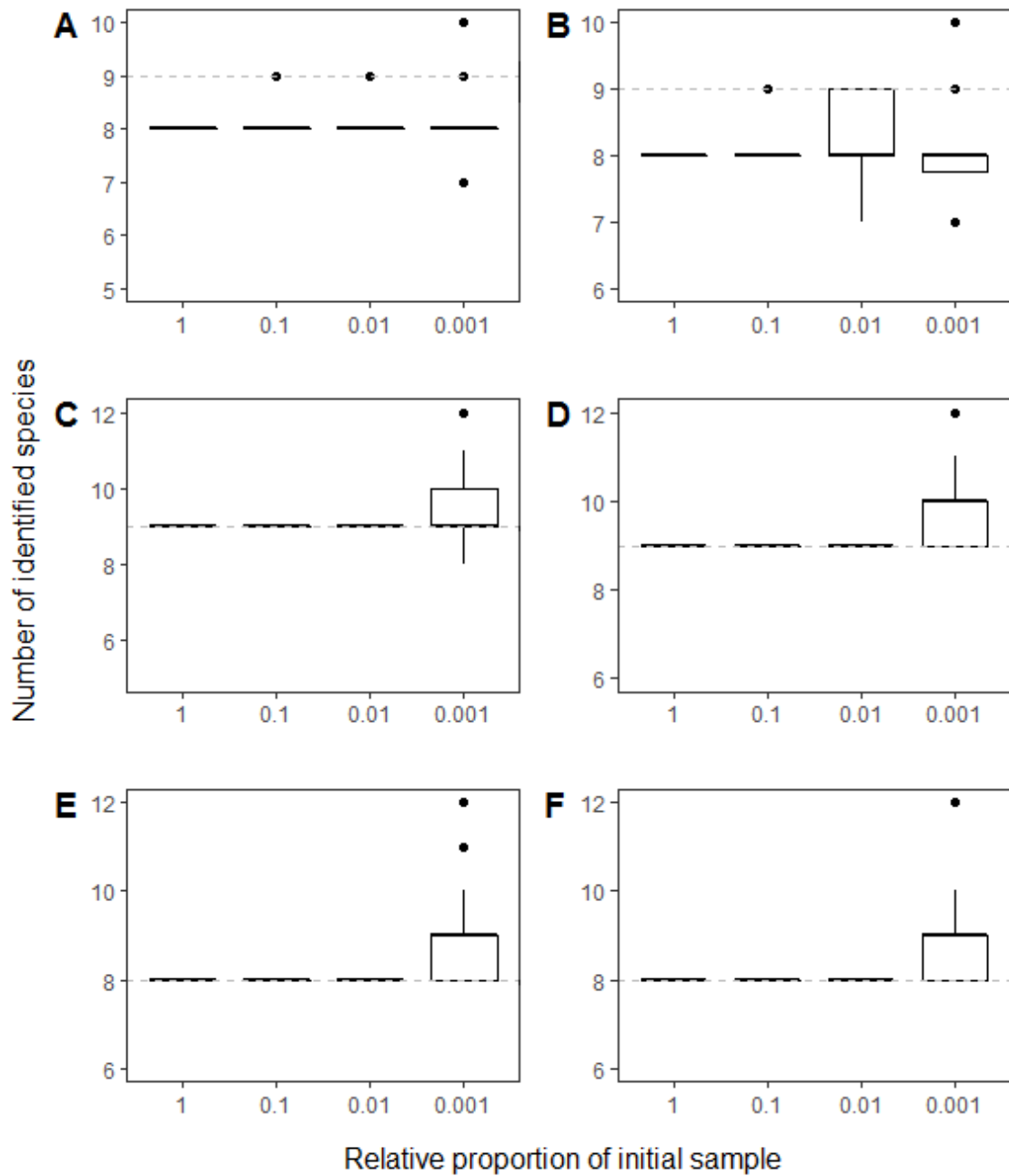


Figure 3.3. Effect of the rarefaction of reads on the number of species detected (above $\epsilon = 0.001$ and without contaminants) in the six mixed-species libraries (A to F correspond to libraries *no.* 1 to *no.* 6). The x axis indicates the proportion of reads used (when 1, all reads were used, so there is only one value); in the rest of the values, 100 random repetitions were conducted using the indicated proportion of reads. The horizontal dashed line of each plot indicates the actual number of species in the mixture.

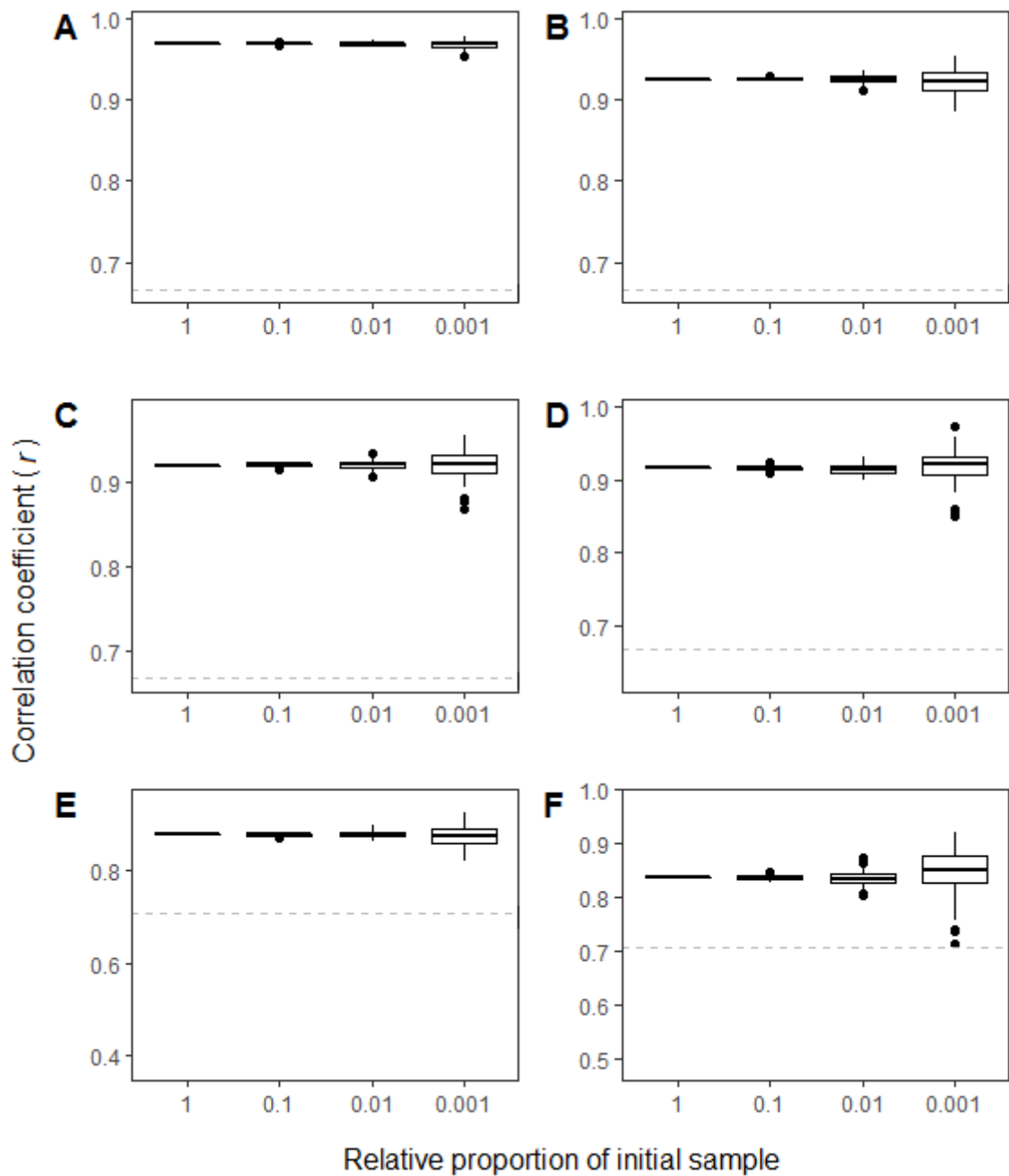


Figure 3.4. Effect of the rarefaction of reads on the correlation coefficient r between the expected and the recovered relative abundance of the species in the six mixed-species libraries (A to F correspond to libraries *no.* 1 to *no.* 6). The x axis indicates the proportion of reads used (when 1, all reads were used, so there is only one value); in the rest of the values, 100 random repetitions were conducted using the indicated proportion of reads. The horizontal dashed line of each plot indicates the critical value of r , above which measured r is statistically significant at $p < 0.05$.

3.5.1. Species identification: Spurious species and the need for an analytical limit of detection

Our data, collected from single insect specimens, produced assignments to *ca.* 20 species; similarly, in each mixed-species library (8-9 species), *ca.* 50 species were recovered, so most of the listed species were spurious (Table 3.2 and Table 3.4). These extra species can be divided into two groups; contaminants (species handled simultaneously in the same lab) and species for which there is no known reason for their presence.

There are two possible causes for contaminant DNA. The first one is physical contamination in the preparation of the libraries in the lab; the second one is the index-hopping effect during the sequencing reaction (Schnell *et al.*, 2015). We had examples of both kinds.

Three species were handled simultaneously in the lab, but not sequenced. All these species appear in most libraries, generally in a proportion lower than 0.001 (Table 3.2, Table 3.4 and Table S3.3). Most of the reported contaminants cannot be attributed to specific issues in the lab workflow. However, the presence of *Ceratitis capitata* in libraries of *Bractocera oleae* (it accounts for 6.6% in the single-species library and above 1% in the mixed-species ones) may have occurred during sample collection. These two dipterans were trapped together in agricultural fields and also transported together to the lab. There, a trained entomologist separated the individuals of the two species; it is very unlikely that this person could have made an identification mistake, but it is possible that fragments of *C. capitata* (legs, antennae, wings) ended up in the *B. oleae* tube. In addition, the two species were, for a certain period, suspended in the same ethanol solution. In the analysis of our artificial mock samples (both single and mixed-species), we eliminated all the contaminant species because we knew that they were contaminants. However, in actual environmental samples, it can be challenging to set apart contaminants from species belonging to the community.

Another possibility for inter-sample contamination is the worrisome tag-jumping effect (Schnell *et al.*, 2015), in which a read from one library is mistakenly taken as belonging to another one because the tag, used to identify each multiplexed library, is sequenced erroneously. Of course, it is not possible to distinguish this process from the genuine contamination discussed above. Again, we could safely ignore these species in the single-species libraries, but we cannot do anything about them in the mixed-species libraries nor in the *real* samples.

In addition to the contaminants, many other species appeared in the lists of both the single and mixed-species libraries (Table 3.2 and Table 3.4) that were never handled in our lab nor could be found in the area. All these species appeared at small relative abundances, almost always below the threshold of $\epsilon = 0.001$. The cause of these misclassifications is probably a sequencing error in our samples, but there must also be errors and missing sequences in the reference genomes themselves (Donovan *et al.*, 2018; Lu & Salzberg 2018). For example, some of the wrongly assigned reads were of mutualistic or parasitic bacteria of insects, like *Providencia* sp., *Morganella* sp., *Lactobacillus* sp., *Acetobacter* sp. and *Wolbachia* sp. (Chandler *et al.*, 2011; Simhadri *et al.*, 2017; Singh *et al.*, 2015). Thus, it is reasonable to assume that they were in our samples alongside the insects, but also in the specimens used to generate the reference genomes. Several other wrongly assigned reads were of conserved RNA sequences that are difficult to set apart from phylogenetically similar species. In addition, there is always some intraspecific genetic variability in all species and the specimens that we sequenced likely come from a different population from the one used to obtain the reference genome.

The only way to eliminate these species from the species list of each library is to set a threshold for the relative abundance of the species, *i.e.*, an analytical detection limit. A detection limit of $\epsilon = 0.001$ eliminated all the unwanted species in all but 3 of the 28 artificial libraries (Table 3.2, Table 3.4 and Table S3.3). There is a reasonable explanation for two of these three misplaced species, as they were congeneric species in the honey bee *Apis mellifera* (*A. florea*) and in *Atta colombica* (*A. cephalonica*) libraries. The presence of the Dipteran *Lucilia cuprina* in the two libraries of *Drosophila*

virilis seems to be mediated by two bacteria (*Providencia* sp. and *Morganella* sp.) associated with the microbiome of dipterans (Chandler *et al.*, 2011; Singh *et al.*, 2015) that appear in the published genome of *L. cuprina*. Merchant *et al.* (2014) show that this problem is widespread, as they found bacterial contamination in five out of nine eukaryotic assembled and published genomes. New bioinformatic tools for the decontamination of eukaryotic genome assemblies from bacterial contaminants (Fierst & Murdock 2017) are likely to alleviate this problem.

In the mixed-species libraries, the detection limit of $\epsilon = 0.001$ removed all spurious species, with no exceptions. However, in libraries *no. 1* and *no. 2*, DNA of *Papilio machaon* was used to prepare the mixtures at a low concentration (Table 2.2) but was excluded from the species list (Table 3.4 and Table S3.6). Therefore, on one hand, the use of a detection limit has the desired effect of eliminating false positives (FP) but, on the other hand, can generate false negatives (FN). In our mixed-species libraries, the balance was favourable, as there were no FP and only two FN.

We do not think that the presence of spurious species in our artificial libraries is specific to the way that we handled the DNA in the lab or to our species assignment algorithm. The problem is probably more general, but it is only exposed when artificial samples are analysed, especially in those consisting of only one species. Other researchers have found similar results using prokaryotes (Pereira *et al.*, 2018). Consequently, we recommend the use of a stringent detection limit (*e.g.*, $\epsilon = 0.001$) to avoid a long list of spurious species. Of course, this will have the negative effect of excluding some species that actually are present at low abundance, but this trade-off between FP and FN is inevitable (Alberdi *et al.*, 2018). To be fair, most studies already do this but in a rather unsystematic way. For instance, MEGAN (Huson *et al.*, 2007) and many other studies always ignore singletons. Other studies increase the minimum number of reads to keep a taxon in the list (five in Piñol *et al.*, 2014; ten in Gibson *et al.*, 2015 and in Lee *et al.*, 2018). As we do here, Pompanon *et al.* (2012) and Alberdi *et al.* (2018) suggest that a relative threshold can be more appropriate than absolute read count thresholds.

3.5.2. Quantification of the relative abundance of the species

The main objective of using MG for the quantification of the species abundance and hence of this study, was to overcome the PCR-biases of amplicon MB. Here, we showed that the MG approach completely fulfilled this objective, whereas in amplicon MB, the quantification of the abundances of the species is sometimes good (Kraaijeveld *et al.*, 2015; Saitoh *et al.*, 2016), but in others, it is very poor (Leray & Knowlton 2017; Piñol *et al.*, 2015).

Our mixed-species libraries comprised *ca.* 1.7 million reads each, but the rarefaction experiment showed that, even with 100 times less reads (*ca.* 17,000 per library), the quantification would still be good (Figure 3.4). Thus, many more samples could be multiplexed in one single Illumina MiSeq run and, consequently, reduce the mean cost per library. Of course, if the mixtures were richer in species, more reads per sample would be needed. Greenwald *et al.* (2017) applied shotgun MG in prokaryotes and was also able to estimate RSA with high fidelity ($r^2 > 0.92$).

However, it is important to remember that not all biases are corrected by shotgun MG. Here, we began the process using extracted DNA, so all the biases in the generation of eDNA sequences (*i.e.*, digestion rates in dietary studies or DNA degradation in the soil or in the water, or in the DNA extraction) are not accounted for. In particular, the same amount of biomass does not always render the same amount of DNA (Pornon *et al.*, 2016); thus, as the usual goal is the estimation of species biomass, a biomass-to-DNA factor should be estimated for each species or, alternatively, the artificial mixtures should be prepared from a known biomass of each species rather than from a known DNA amount, as some authors already do (Tang *et al.*, 2015; Zhou *et al.*, 2013).

3.5.3. Data treatment and the assignation of reads to species

In MG, there are, basically, two methods to assign reads to species; the assembly-based and the read-based approaches (Thomas *et al.*, 2012). In the former, the reads are assembled using a *de novo* assembler into contigs and these are mapped into reference genomes; the quantification of the species is achieved by counting the number of reads assembled in contigs that map into a given species. This approach is commonly used in prokaryote and in mitochondrial metagenomics (MMG), but it was not useful in this application because of the low coverage of our sequencing: with so few overlapping reads, many very small contigs would be obtained.

Consequently, we used here the read-based approach that assigns a species to every read by mapping it into a reference genome. As a mapper, we used BWA, but other possibilities would probably be good choices too (*e.g.*, Bowtie2: Boratyn *et al.*, 2018; MagicBlast: Langmead & Salzberg 2012; GEM: Marco-Sola *et al.*, 2012). In any case, all mappers normally produce hits of one read into several reference genomes, so an algorithm is needed to assign a species to a read. By far the most common algorithm used in metabarcoding (MB) and MG studies is the LCA algorithm (*e.g.*, MEGAN: Huson *et al.*, 2007; KRAKEN: Wood & Salzberg 2014); albeit there are other alternatives (Hanson *et al.*, 2016; Sarmashghi *et al.*, 2019). However, we used here our own γ - δ algorithm that sets species apart rather than extracting as much taxonomic information as possible from a set of reads, as the LCA algorithm does. The γ - δ algorithm declared, as informative, approximately half of the reads. This algorithm is extremely straightforward and easy to implement.

3.5.4. Present and future of metagenomics

The MG approach presented here for eukaryotic species will not be a realistic option until the number of sequenced genomes is a substantial fraction of the total biodiversity. Today, the MG method for taxonomic purposes is used mostly with

genomes of organelles instead of whole genomes, because the number of sequenced organelle genomes is much higher than the number of whole genomes (*e.g.*, today there are roughly, in the NCBI RefSeq database, 14 times more mitogenomes than whole genomes of insects). In addition, the number of sequenced organelle genomes is increasing quickly with new easier and faster methods, based on next generation sequencing and *de novo* assembly (Cameron 2014).

MMG has proved to be better than amplicon MB for quantification purposes (Bista *et al.*, 2018; Gómez-Rodríguez *et al.*, 2015; Gueuning *et al.*, 2019), but estimation of relative abundance amongst species (*i.e.*, in a given sample, species *i* is more abundant than species *j*) is not always high (Krehenwinkel *et al.*, 2017; Tang *et al.*, 2015). The quantification power of MMG is likely bounded for two reasons. First, when there is no mitochondrion enrichment (as in Zhou *et al.*, 2013), only a small proportion of the shotgun reads map into the mitogenome (~0.5 % in insects; Tang *et al.*, 2014), so a high sequencing depth is necessary to obtain good quantitative results (Gueuning *et al.*, 2019). Second and most important, the number of mitogenomes per nuclear genome (mitochondrial copy number) is variable amongst species and even between tissues. Consequently, in a given amount of DNA (and using it as a *proxy* of biomass), the mitochondrial copy number will vary across species, so the estimation of the RSA will be affected. This problem is known (Crampton-Platt *et al.*, 2016; Krehenwinkel *et al.*, 2017) and applies not only to MMG, but also to amplicon MB targeting mitochondrial markers. The solution is to use an independent estimation of the mitochondrial copy number for each species, but, at this point, we are not aware of any reliable data of this variable across arthropod species.

It is also fair to question about the computational problems that would pose a future with huge reference databases when the genomes of most species are sequenced. Perhaps our implementation of the method could be so computationally costly that it would be inapplicable in practice. In our opinion, the method is perfectly manageable today in a modest computer server and will remain so in the foreseeable future. In the reported experiments, the maximum processing time of the entire pipeline per mixed-species library was of 1.3 hours, most of it being devoted to the mapping of reads into

the reference genomes. This mapping of reads into genomes is a problem that fits into the category known as “embarrassingly parallel” applications (McCool *et al.*, 2012), in which a read can be processed simultaneously with different references and, therefore, the complexity of the algorithm increases linearly with the number of reads n and the number of genomes g . Thus, using library *no.* 1 as an example, multiplying g by 100 (~ 11,000 genomes in our case) and decreasing n by 100 (~ 19,000 reads) should keep the execution time roughly at the same 1.3 hours (we showed here that it was possible to reduce the number of reads without loss of identification and quantification power; [Figure 3.3](#) and [Figure 3.4](#)).

In addition, the pipeline could eventually be modified in several ways to further reduce the execution time. (1) Selection of reference genomes in the database: when processing a sample, there is no need to compare the reads with all the animal genomes (or plant or fungi) in the world: if the interest is in insects, then only the genomes of insects known to occur in a certain geographical region should become the reference database. Thus, even in a future with the genomes of *all* species already sequenced, the number of genomes of interest will never be of millions, but of 10^3 to 10^5 genomes at most. (2) Filtering of the reference genomes database: we showed here that only approximately half of the reads were informative. The non-informative reads probably belong to certain regions of the genome that, when identified, could be filtered out from the reference database with appropriate programmes. (3) Elimination of non-informative reads in running time: in the γ - δ algorithm, a read that maps better than δ in two different genomes is declared non-informative. Once that read is detected, the mapping of it against the remaining reference genomes is not necessary anymore and finally (4) It is reasonable to assume that the power of the computers will continue to increase in the future as it has done in the past (Williams 2017). It is even possible that, in the next decades, unimagined computational capabilities become available with the advent of quantic processors (Arute *et al.*, 2019).

3.6. Concluding remarks

According to our results, the low-coverage shotgun MG method is perfectly capable to set apart closely related insect species, like the four species of the genus *Drosophila* that we included in the artificial libraries. We also saw that, despite the risk that some reads were not in the reference databases that we used (reads of commensal or parasites species; parts of the genome not yet sequenced) or that some reads were very similar in more than one reference genome, we achieved a reasonable proportion (*ca.* 0.50) of truly informative reads. By using mixtures, we showed that it is possible with this technique to quantify with confidence the relative abundance of individual species in the mixtures and that, with much less sequencing depth than the one used here, it was possible to obtain comparable results (*ca.* 17,000 reads in mixtures of *ca.* 10 species). Finally, a word of caution. The “dream” of getting an eDNA sample, sequencing it, mapping it against a growing DNA database and obtaining the species names and relative abundance of all species in the mixture that we tried to simulate in this study, is not without hurdles. The main one is obviously the low number and quality of eukaryote genomes sequenced so far, but also the impossibility of identifying, with confidence, species below a certain detection limit and the need to improve the algorithms in a future with huge genome databases and increased sequencing depth.

4. Relative species abundance estimation in artificial mixtures of insects using mito-metagenomics and a correction factor for the mitochondrial DNA copy number

Chapter published as: Garrido-Sanz L, Senar MÀ, Piñol J. (2021) Relative species abundance estimation in artificial mixtures of insects using mito-metagenomics and a correction factor for the mitochondrial DNA copy number. Molecular Ecology Resources, 00: 1-15. <https://doi.org/10.1111/1755-0998.13464>

4. Relative species abundance estimation in artificial mixtures of insects using mito-metagenomics and a correction factor for the mitochondrial DNA copy number

4.1. Abstract

Mito-metagenomics (MMG) is becoming an alternative to amplicon metabarcoding for the assessment of biodiversity in complex biological samples using high-throughput sequencing (HTS). Whereas MMG overcomes the biases introduced by the PCR step in the generation of amplicons, it is not yet a technique free of shortcomings. First, as the reads are obtained from shotgun sequencing, a very low proportion of reads map into the mitogenomes, so a high sequencing effort is needed. Second, as the number of mitogenomes per cell can vary among species, the relative species abundance (RSA) in a mixture could be wrongly estimated. Here, we challenge the MMG method to estimate the RSA using artificial libraries of 17 insect species whose complete genomes are available on public repositories. With fresh specimens of these species, we created single-species libraries to calibrate the bioinformatic pipeline and mixed-species libraries to estimate the RSA. Our results showed that the MMG approach confidently recovers the species list of the mixtures, even when they contain congeneric species. The method was also able to estimate the abundance of a species across different samples (*within*-species estimation) but failed to estimate the RSA within a single sample (*across*-species estimation) unless a correction factor accounting for the variable number of mitogenomes per cell was used. To estimate this correction factor, we used the proportion of reads mapping into mitogenomes in the single-species libraries and the lengths of the whole genomes and mitogenomes.

Keywords Metazoa, mitochondrial genomes, mitogenome skimming, next-generation sequencing, PCR-free, mock sample

4.2. Introduction

Mitochondrial metagenomics or mito-metagenomics (MMG) is becoming an alternative to the classical amplicon metabarcoding (MB) for the large-scale assessment of biodiversity of Metazoa (Crampton-Platt *et al.*, 2015; Tang *et al.*, 2014; Zhou *et al.*, 2013). MMG consists in the shotgun sequencing of a DNA sample followed by the mapping of the reads to mitochondrial genomes (also referred to as mitogenomes) obtained from online repositories or *ad hoc* assemblages (Crampton-Platt *et al.*, 2016). On the positive side, MMG avoids the amplification biases caused by the PCR step (Elbrecht & Leese 2015; Piñol *et al.*, 2015; Taberlet *et al.*, 2012b); thanks to the natural enrichment in cell and short length of the mitogenomes, a high number of mitochondrial reads are present in the sample, such assembly of mitogenomes is sometimes possible (Crampton-Platt *et al.*, 2016). On the negative side, the tiny size of the mitogenome compared to the nuclear genome produces a high number of non-informative reads (Tang *et al.*, 2014), so a great sequencing depth is needed; besides, the number of mitochondrial genomes per nuclear genomes (mito-to-nuclear ratio) varies largely between species (Bista *et al.*, 2018) (see subsubchapter [1.3.3. Mitochondrial metagenomics](#) for a more thoughtful explanation about this problematic).

In this chapter we explore the quantitative capabilities of MMG for the estimation of relative species abundance (RSA) of heterogeneous mixtures of insects. For this purpose, we prepared single-species and artificial mixed-species libraries with several species of insects whose entire genome has already been sequenced. The single-species libraries allowed the calculation of a reliable mitochondrial DNA copy number (N_M) for each species that was further used as a correction factor for the *across*-species estimation of RSA of the mixed-species libraries.

This study aims to clarify some aspects of the MMG and to test the difficulties likely to be faced in *real* samples when the original composition of the sample is unknown. In particular, we addressed the following questions: (1) is the MMG method able to identify species in complex mixtures, even when they are of the same genus? As an approach to *real* samples, we investigated the robustness of the method in the absence of the mitogenome of the focal species. (2) Can MMG estimate the RSA of complex samples? Is it necessary the use of the N_M correction factor for the *across*-species estimation of RSA? (3) Finally, can the number of sequenced reads be reduced and still recover all species in a complex sample of insects?

4.3. Material and Methods

4.3.1. Reference genomes

We downloaded all mitogenomes of insects available at the NCBI RefSeq database on 1st August 2019, plus the complete genomes of the 17 species selected for the study from the same database (the 17 selected species are indicated below). Ten species had the complete genomes but not the mitogenomes on RefSeq, so we downloaded their mitogenomes from GenBank (accessed on 2nd August 2019). Species with several mitogenomes were deduplicated, so we obtained the mitogenomes of 1794 species of insects (hereafter, Mito1794), comprising 1174 genera, 331 families, and 27 orders (Table S4.1).

4.3.2. Selection of species and preparation of the DNA libraries

We selected 17 species of insects whose complete genome and mitogenomes are already sequenced and available on the RefSeq repository (Table 4.1). With the 17 insect species, we prepared two kinds of libraries: 21 single-species libraries and 6

mixed-species libraries. The same DNA extracts used for the first run of single-species libraries were also used to create six artificial mixed-species libraries of 7-8 species at known relative DNA concentrations (Table S4.2). As we did before in chapter 3, we used the single-species samples to calibrate the bioinformatic pipeline and the mixed-species samples to test the ability of the method to estimate the RSA. Detailed information regarding specimens' collection, laboratory treatment and quality control are provided in subchapter 2.3. **Preparation of samples: Selection of the species, laboratory treatment and quality control.**

Albeit the same samples were used in the previous chapter 3, we excluded the Panamanian leafcutter ant (*Acromyrmex echinator*) from the current study because its mitogenome was absent from RefSeq and GenBank repositories.

Table 4.1. Summary information of single-species libraries used with paired-end reads.

Run-Library	Species	Number of raw reads	Number of reads after quality control step	Number of candidate mito-reads
1-1	<i>Papilio machaon</i>	434,520	432,712	14,774
1-2	<i>Drosophila virilis</i>	4,714,902	4,652,442	152,568
1-3	<i>Drosophila melanogaster</i>	2,283,768	2,275,386	50,944
1-4	<i>Drosophila mojavensis</i>	1,668,424	1,646,524	109,458
1-5	<i>Bactrocera oleae</i>	580,996	577,720	23,204
1-6	<i>Linepithema humile</i>	1,422,342	1,402,800	73,580
1-8	<i>Bombus terrestris</i>	1,994,938	1,987,920	43,872
1-9	<i>Apis mellifera</i>	1,262,388	1,236,630	223,202
1-10	<i>Acyrtosiphon pisum</i>	684,688	596,972	104,344
2-1	<i>Atta colombica</i>	3,272,710	3,267,688	333,064
2-2	<i>Bemisia tabaci</i>	2,513,212	2,501,586	23,096
2-3	<i>Cimex lectularius</i>	3,506,722	3,474,010	122,826
2-4	<i>Drosophila melanogaster</i>	2,909,608	2,903,546	74,984
2-5	<i>Drosophila mojavensis</i>	1,797,470	1,786,926	82,794
2-6	<i>Drosophila virilis</i>	1,336,884	1,332,888	49,146
2-7	<i>Drosophila suzukii</i>	2,510,384	2,501,862	79,498
2-8	<i>Linepithema humile</i>	2,164,404	2,157,154	95,936
2-9	<i>Plutella xylostella</i>	4,250,124	4,244,328	85,882
2-10	<i>Solenopsis invicta</i>	3,661,374	3,648,438	146,812
2-11	<i>Vollenhovia emeryi</i>	3,487,834	3,480,802	66,664
2-12	<i>Wasmannia auropunctata</i>	3,335,212	3,293,182	179,000

4.3.3. Input data filtering

In MMG studies, a small proportion of shotgun reads map into the mitogenome (*e.g.*, Tang *et al.*, 2014), hence most reads are not useful and slow down the mapping process. Thus, it seems reasonable to eliminate the reads that are not mitochondrial before the mapping step (Crampton-Platt *et al.*, 2015, 2016; Zhou *et al.*, 2013). For this purpose, we created a reference database with one mitogenome per family (hereafter, Mito331) where the representative species per family was chosen randomly. Then, we mapped the raw reads against the Mito331 reference dataset using a permissive criterion and kept the putative mitochondrial reads (hereafter, candidate mito-reads). The mapping was done using BWA v0.7.15-r1140 (Li 2013) with *mem* algorithm and an alignment score of zero. SAMtools v1.10 (Li *et al.*, 2009) was subsequently used to filter the paired-end reads with no mapping reads and recovered the mito-reads in FASTQ format (*bam2fq*).

As low-complexity regions are prone to misclassify the reads (Lu & Salzberg 2018; Pearman *et al.*, 2020), we prepared a new set of filtered mitogenome references by removing low-complexity regions from the Mito1794 reference (hereafter, FilteredMito1794). Low-complexity regions were identified using dustmasker (-level 45) (Morgulis *et al.*, 2006) and replaced with Ns using an in-house python script.

To avoid confusions, we recapitulate below the name and meaning of the three different databases of mitochondrial genomes that we used for the mapping of reads:

- Mito1794: the original mitogenomes of 1794 species.
- FilteredMito1794: as Mito1794, but with the low complexity regions removed.
- Mito331: a subset of Mito1794 with only one mitogenome per family; this reference was only used to obtain the candidate mito-reads from the total of reads of each sample.

As we did not know to which extent the filtering of raw reads and mitogenomes was useful, we conducted four different kinds of mapping of reads to reference mitogenomes in the single-species libraries:

- Raw reads against Mito1794 reference database
- Raw reads against FilteredMito1794 reference database
- Candidate mito-reads against Mito1794 reference database
- Candidate mito-reads against FilteredMito1794 reference database

4.3.4. Classification of reads to species

In all input data combinations, data was processed using the $B\gamma\delta$ classifier defined in the preceding subchapter 2.4.3. **BWA plus γ - δ algorithm** and at supplementary material (Methodology S3.1). Briefly, mapping was conducted with BWA, and no mapping reads were filtered out with SAMtools. As reads may map to several references, we used the γ - δ algorithm to seek robust identifications at species rank.

The γ - δ algorithm has never been applied before to MMG data, hence the appropriate values of γ and δ are unknown, so the single-species libraries were used to find the best combination of the parameters γ and δ . The tested values were all the combinations of $\gamma = \{0.99, 0.98, 0.97\}$ and $\delta = \{0.98, 0.97, 0.96\}$ provided that $\gamma > \delta$. To find the best values of γ and δ we relied on the criterion that the number of recovered species had to be one in the single-species libraries.

Here we used the same single-species libraries that we created in chapter 3, *i.e.*, a training set with 75% of the reads randomly selected, and a test set with the remaining 25% of the reads. The training set was used for the calibration of the procedure; the test set was used to assess the goodness of fit of the model and to calculate the summary statistics.

A situation that can arise in *real* samples, as opposed to the *artificial* samples used here, is that the mitogenomes of some species in the sample are not in the reference database. We explored this situation by running again the complete pipeline with all the single-species libraries using the best set of input data and parameters and without

the mitogenome of the species actually in each library. Ideally, no read should be assigned to any species because the mitogenome of the only species in the library is not in the database. However, the reads might eventually be wrongly assigned to other species in the database and, thus, generate false positives (FP). The outcome of this experiment should reveal the robustness of the γ - δ algorithm in the assignment of reads to species.

4.3.5. Quantification of the RSA in mixed-species libraries and the need for a species-specific correction factor

In the literature, the relative abundance of one species is sometimes compared among different samples and on other occasions, the relative abundance of several species is compared within one sample (*within*-species and *across*-species RSA, respectively, following Ji *et al.*, 2020). Here, we present the comparison of actual *versus* estimated RSA in the mixed-species libraries using both approaches. As we observed in the subsubchapter 1.3.3. **Mitochondrial metagenomics**, from a conceptual point of view, the quantitative estimation of *within*-species RSA in MMG is easier than the *across*-species RSA, because in the latter the mitochondrial DNA copy number can vary widely between species.

With the single-species libraries we estimated the mitochondrial DNA copy number (N_M) of each species in the following way:

1. Let x_i be the ratio of the genomic mitochondrial information divided by the total (haploid) genomic information for species i . The mitochondrial information is the mitogenome length (M_i) times N_{Mi} ; the total genomic information is the sum of the nuclear genome length (G_i) and the mitochondrial information.

$$x_i = \frac{M_i \cdot N_{Mi}}{G_i + (M_i \cdot N_{Mi})} \quad \text{Equation 4.1}$$

2. The re-arrangement of Eq. 4.1 allows the estimation of N_{Mi} .

$$N_{Mi} = \frac{x_i \cdot G_i}{M_i \cdot (1 - x_i)} \quad \text{Equation 4.2}$$

3. G_i and M_i are known for species with sequenced genomes, but x_i is not. In our experimental setting, x_i can be estimated in the single-species libraries as the ratio between the number of reads that map into the mitogenome (R_{Mi}) divided by the total number of reads of species i (R_{Gi}).

$$x_i = \frac{R_{Mi}}{R_{Gi}} \quad \text{Equation 4.3}$$

We obtained R_{Mi} by mapping the reads of species i to its mitogenome when this mitogenome was the only one used as the reference in the mapping. Regarding R_{Gi} , we assumed that all reads of the single-species library of species i belong to species i .

In the comparison of actual *versus* estimated RSA using the mixed-species libraries, we multiplied the actual relative abundance of species i (Table S4.2) by N_{Mi} , and then renormalized the values to sum 1.

Finally, we estimated the importance of knowing or ignoring the individual values of G_i and M_i of each species in the mixture in the estimated RSA by comparing the results obtained with the correction factor of Eq. 4.2 (N_{Mi}) with another factor that uses the mean value of G ($\bar{G} = 338$ Mbp) and M ($\bar{M} = 16.3$ kbp) for all the species in the mixture (\bar{N}_{Mi}):

$$\bar{N}_{Mi} = \frac{x_i \cdot \bar{G}}{\bar{M} \cdot (1 - x_i)} \quad \text{Equation 4.4}$$

4.3.6. Rarefaction of the input samples

We only multiplexed six mixed-species libraries in a single Illumina MiSeq run (Table S4.2), with the consequence of a high economic cost per library. However, from a practical point of view, it would be interesting to use fewer reads per library and still have a good quantitative estimation of RSA. To test this possibility, we randomly rarefacted the mixed-species samples at various proportions of the original number of reads {0.5, 0.1, 0.05, 0.01} and run the new datasets through the entire pipeline. We repeated each simulation 100 times using different subsets. We recorded from every simulation the number of recovered species.

4.3.7. Hardware

We run the complete pipeline on a server with two Intel Xeon E5-2620 v3 processors with six cores each and hyper-threading technology, so a maximum of 24 threads were available.

4.4. Results

4.4.1. Species identification

The 21 single-species libraries (Table 4.1) generated $2,371,091 \pm 1,210,091$ (mean \pm SD) paired-end reads. A proportion of 0.012 ± 0.027 reads were eliminated in the trimming step, remaining a proportion of 0.987 ± 0.027 reads available for further analysis.

The results that follow correspond to the application of the γ - δ algorithm for the assignation of reads to species on the training set (*i.e.*, 75% of the sequenced data).

We also eliminated from the following results the reads assigned to species that could legitimately be attributed to physical contamination in the lab or the sequencing. These contaminants were species sequenced in different libraries of the same Illumina run and the fly *Ceratitis capitata* that contaminated the library of *Bactrocera oleae* (see subchapter 2.5. Contaminant species and the discussion for the reason behind this contamination).

Because the MMG method must recover only one species in single-species libraries, we fixed the values of $\gamma = 0.99$ and $\delta = 0.96$ in the γ - δ algorithm as this combination was the only one to provide the expected result (Table S4.3). All the other tested combinations reported FP, like *Bactrocera biguttula* in libraries of *B. oleae*, *Drosophila formosana* in libraries of *D. melanogaster* and *Solenopsis richteri* in libraries of *S. invicta*. Results of all tested γ , δ and input data combinations (raw reads versus candidate mito-reads and Mito1794 versus FilteredMito1794 references) are provided as supplementary material (Tables S4.4-S4.7).

Filtering out the repetitive regions of the mitogenomes (*i.e.*, FilteredMito1794) had a dramatic effect on the number of identified species. With the FilteredMito1794 database, we only detected the focal species in all the libraries, whereas with the Mito1794 database there appeared several FP in many libraries (2.5 ± 1.2 species per library using all raw reads or 1.7 ± 0.9 species using only candidate mito-reads) (Table 4.2A). The masked regions mostly belonged to non-coding regions of the mitogenome, including the control region (Table S4.8). The use of only candidate mito-reads instead of all reads produced a loss of *ca.* 6% of informative reads (Table 4.2B) but reduced ~ 18 times the execution time (Table 4.2C). In summary, the elimination of the repetitive regions from the genomes removed all the FP and the mining of candidate mito-reads reduced 18-fold the execution time of the pipeline with a moderate loss of informative reads. Therefore, in the subsequent steps, we used both the FilteredMito1794 database and only the candidate mito-reads (Figure 4.1).

We evaluated the goodness of fit of the model with the test set (*i.e.*, the remaining 25% of reads not used in the previous calibration) using the best set of input data and

parameters (*i.e.*, the FilteredMito1794 database, the candidate mito-reads and the parameters $\gamma = 0.99$ and $\delta = 0.96$). The number of identified species per library was one in all cases (Table S4.9) and the proportion of informative reads was 0.0046 ± 0.0056 .

The absence of the mitogenome of the focal species in the reference database did not produce many FP in the single-species libraries (Table 4.3). When there were no congeneric species of the focal species in the reference database (six out of 17 species), no read was assigned to any species; when there were congeneric species in the database, in six cases no reads were assigned to any species and in four cases some reads were assigned to another species of the same genus (*Bactrocera*, *Drosophila*, *Plutella*, and *Solenopsis*); only in one species (*Drosophila melanogaster*) appeared some reads belonging to species of a different genus (*Exorista sorbillans*, Diptera:Tachinidae).

The six mixed-species libraries (Table S4.2) generated $3,376,087 \pm 424,238$ paired-end reads. A proportion of 0.003 ± 0.001 reads were eliminated in the trimming step and a proportion of 0.925 ± 0.006 in the mito-reads mining step. Therefore, only a proportion of 0.075 ± 0.006 of the raw reads were candidate mito-reads retained for further analysis.

Table 4.2. Summary of the results per library (mean \pm SD) on the training dataset of single-species libraries for the four combinations of input data assessed in this study (raw reads and candidate mito-reads mapped to Mito1794 and FilteredMito1794 databases) and using $\gamma=0.99$ and $\delta=0.96$. (A) Number of recovered species per library; (B) Relative proportion of informative reads (RPIR) per library; and (C) processing time per library (format h:mm:ss) (the time necessary to find the candidate mito-reads is included in the processing time). Reads from contaminant species have not been considered.

Metric	Input data	Reference database	
		Mito1794	FilteredMito1794
(A) Number of identified species	Raw reads	2.52 ± 1.21	1 ± 0
	Mito-reads	1.71 ± 0.90	1 ± 0
(B) RPIR	Raw reads	0.0049 ± 0.0057	0.0047 ± 0.0056
	Mito-reads	0.0049 ± 0.0057	0.0046 ± 0.0057
(C) Processing time	Raw reads	7:19:43 \pm 3:46:53	7:21:10 \pm 3:53:34
	Mito-reads	0:25:13 \pm 0:17:07	0:24:37 \pm 0:16:37

In the mixed-species libraries, we recovered all species included in the libraries except *Papilio machaon* in library no. 2 (Table S4.10). As in the single-species libraries, in the mixed-species libraries, we found reads of *Ceratitis capitata* and discarded them as lab contamination. Besides, in libraries no. 2 and no. 3 one single read was attributed to *Bactrocera biguttula* (Table S4.10), a species not handled in the laboratory; so, after all, an analytical detection limit of $\epsilon = 0.0001$ would be useful for the elimination of all FP.

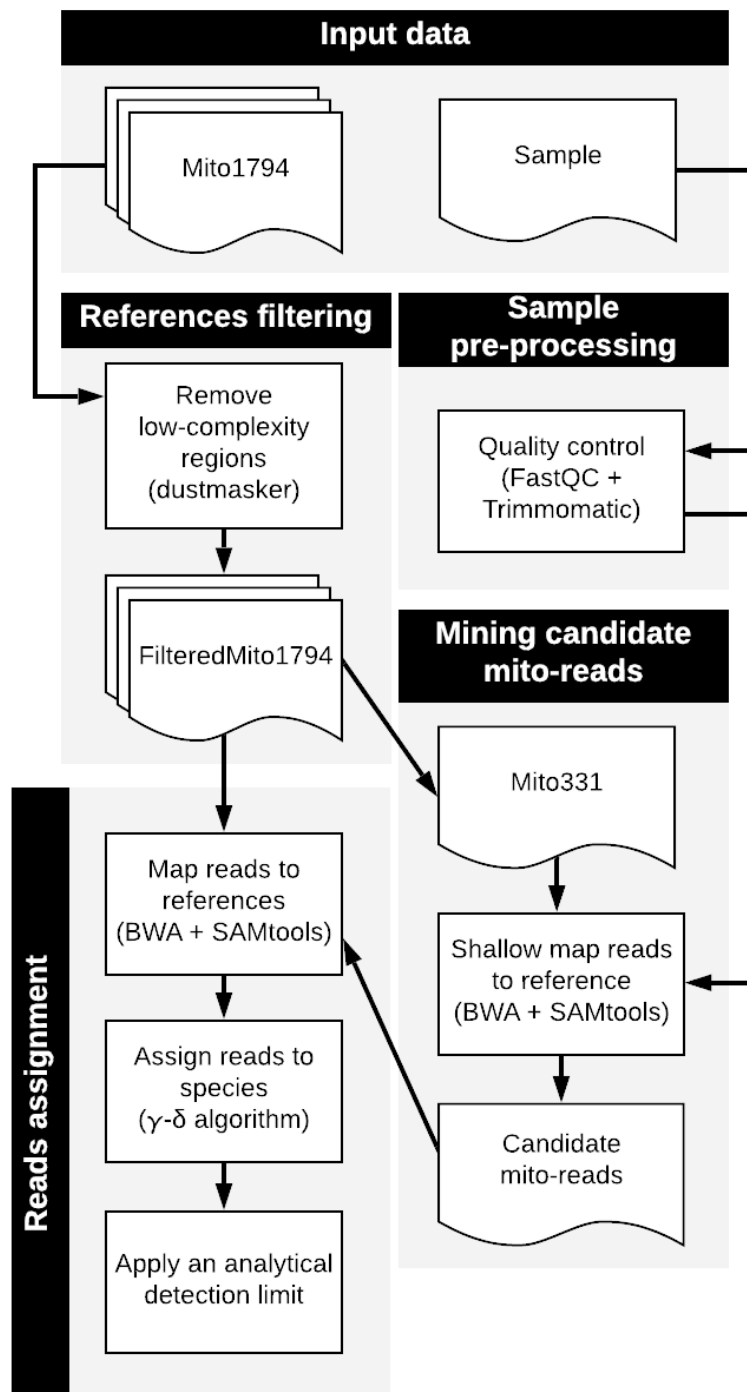


Figure 4.1. MMG pipeline applied in chapter 4. In brackets, the tools used in each step.

Table 4.3. List of species detected on the single-species libraries when the mitogenome of the focal species is in the reference database (column A) and when it is not (column B). For each detected species we indicate its name and the number of assigned reads (in brackets). The number of congeneric species of the focal species included in the database is provided in column C. Libraries are divided into 4 groups: Group 1, species without congeneric species in the database and without FP species; Group 2, species with congeneric species in the database and without FP species; Group 3, species with congeneric species in the database but with FP of the same genus; and Group 4, species with congeneric species in the database but with FP of a different genus.

Group	Run - Lib.	Species used to prepare the library	(A) Focal species mitogenome present in	(B) Focal species mitogenome not present in	(C) Number of congeneric species in the database	
1	1-10	<i>Acyrtosiphon pisum</i>	<i>Acyrtosiphon pisum</i> (154)	None	0	
	2-1	<i>Atta colombica</i>	<i>Atta colombica</i> (19412)	None	0	
	2-3	<i>Cimex lectularius</i>	<i>Cimex lectularius</i> (1082)	None	0	
	1-6	<i>Linepithema humile</i>	<i>Linepithema humile</i> (218)	None	0	
	2-8	<i>Linepithema humile</i>	<i>Linepithema humile</i> (913)	None	0	
	2-11	<i>Vollenhovia emeryi</i>	<i>Vollenhovia emeryi</i> (1399)	None	0	
	2-12	<i>Wasmannia auropunctata</i>	<i>Wasmannia auropunctata</i>	None	0	
	2	1-9	<i>Apis mellifera</i>	<i>Apis mellifera</i> (3597)	None	7
		2-2	<i>Bemisia tabaci</i>	<i>Bemisia tabaci</i> (349)	None	1
		1-8	<i>Bombus terrestris</i>	<i>Bombus terrestris</i> (1877)	None	2
		1-4	<i>Drosophila mojavensis</i>	<i>Drosophila mojavensis</i>	None	18
		2-5	<i>Drosophila mojavensis</i>	<i>Drosophila mojavensis</i>	None	18
2-7		<i>Drosophila suzukii</i>	<i>Drosophila suzukii</i> (4664)	None	18	
3		1-1	<i>Papilio machaon</i>	<i>Papilio machaon</i> (312)	None	13
		1-5	<i>Bactrocera oleae</i>	<i>Bactrocera oleae</i> (425)	<i>Bactrocera biguttula</i> (137)	14
		1-2	<i>Drosophila virilis</i>	<i>Drosophila virilis</i> (8070)	<i>Drosophila littoralis</i> (20)	18
		2-6	<i>Drosophila virilis</i>	<i>Drosophila virilis</i> (2704)	<i>Drosophila littoralis</i> (6)	18
		2-9	<i>Plutella xylostella</i>	<i>Plutella xylostella</i> (2371)	<i>Plutella australiana</i> (127)	1
		2-10	<i>Solenopsis invicta</i>	<i>Solenopsis invicta</i> (131)	<i>Solenopsis richteri</i> (236)	2
	4	1-3	<i>Drosophila melanogaster</i>	<i>Drosophila melanogaster</i> (979)	<i>Drosophila formosana</i> (312) <i>Exorista sorbillans</i> (26) <i>Drosophila mauritiana</i> (3)	18
		2-4	<i>Drosophila melanogaster</i>	<i>Drosophila melanogaster</i> (1384)	<i>Drosophila formosana</i> (478) <i>Exorista sorbillans</i> (53) <i>Drosophila mauritiana</i> (3)	18

4.4.2. Estimation of the RSA in mixed-species libraries

The *within*-species RSA was well estimated for all species ($r \geq 0.97$ and $p < 0.05$ for all species; Figure 4.2), but the *across*-species RSA estimation was very poor ($r \leq 0.67$ and $p > 0.05$ for all samples; Figure 4.3A). Thus, it seems clear the need for a species-specific correction factor that considers a variable ratio of mitochondrial to nuclear DNA (Table 4.4). When we modified the actual RSA with the N_{Mi} correction factor (Eq. 4.2), the correlation between actual and estimated RSA *across*-species became significant in all samples ($r \geq 0.84$ and $p < 0.05$ for all libraries; Figure 4.3B). The use of the \bar{N}_{Mi} correction factor (Eq. 4.4) instead of N_{Mi} provided an even better quantitative estimation of RSA *across*-species ($r \geq 0.91$ and $p < 0.005$ for all libraries; Figure 4.3C).

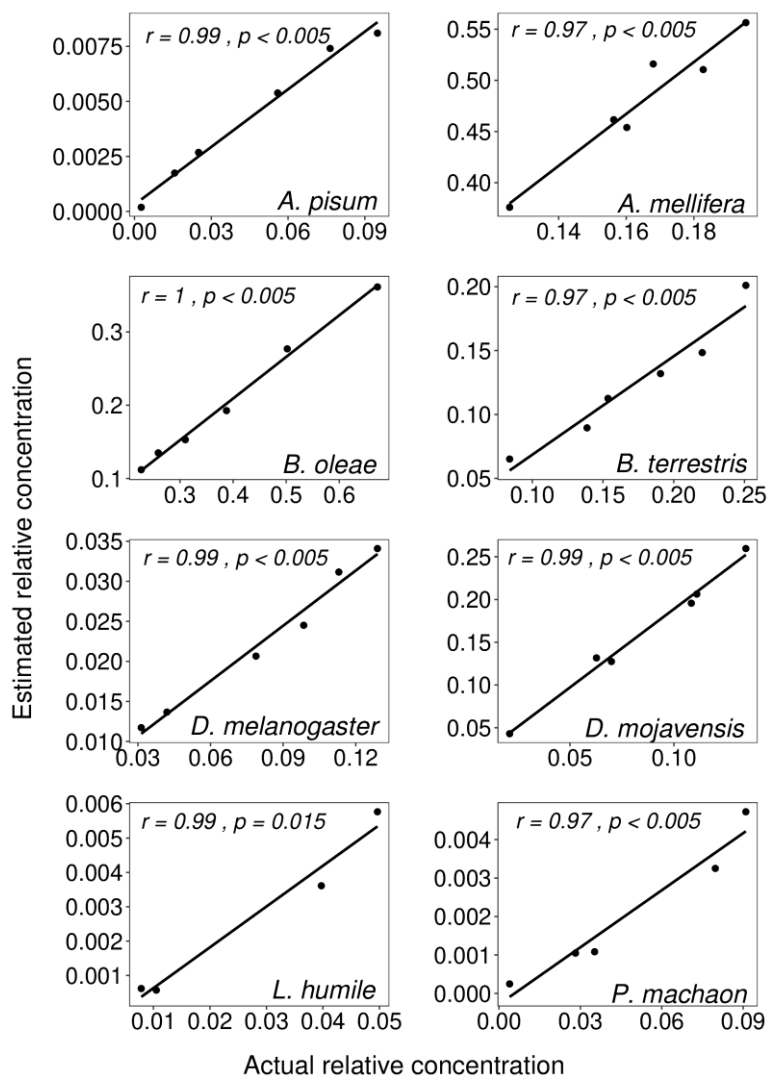


Figure 4.2. Scatter plot of the estimated versus the actual RSA for each species of the mixed-species libraries (*i.e.*, *within*-species RSA). Each plot shows the Pearson correlation coefficient (r) and the corresponding p -value. The coordinate at the origin of all regression lines was not different to 0.

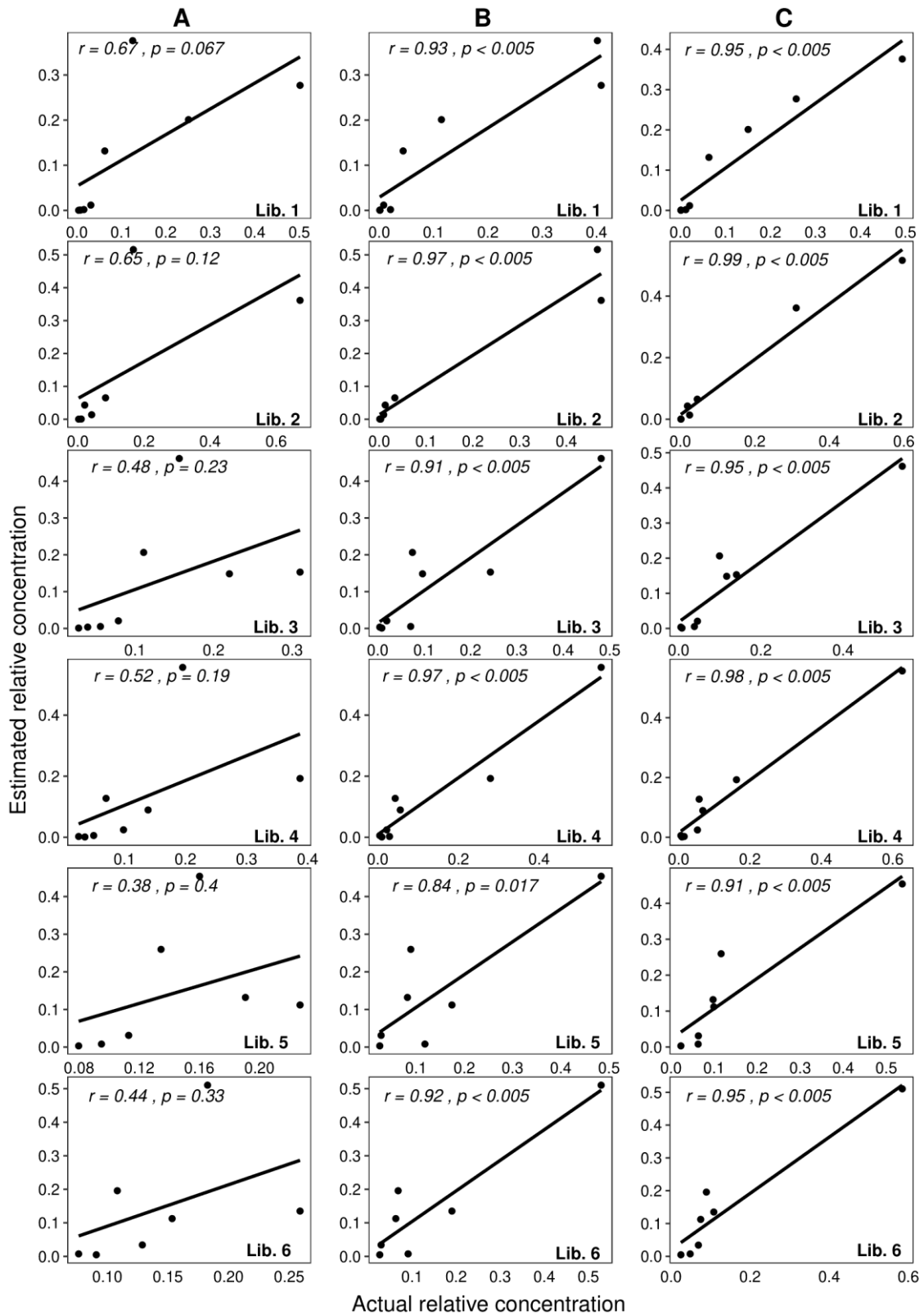


Figure 4.3. Scatter plot of the estimated *versus* the actual RSA in each mixed-species library (*i.e.*, across-species RSA). At the top, it is indicated the way we conducted the actual RSA. A: original expected data; B: corrected expected data after applying the N_{Mi} correction factor; C: corrected expected data after applying the \bar{N}_{Mi} correction factor. Rows from top to bottom correspond to mixed-species libraries from *no.* 1 to *no.* 6. Each plot shows the Pearson correlation coefficient (r) and the corresponding p -value.

The use of rarefacted samples showed that in the libraries with a more variable species abundance (libraries *no. 1* and *no. 2*), the use of just half of the total available reads reduced the number of the identified species (Figure 4.4A) and promoted the presence of low-abundant FP, like *Bactrocera biguttula* above the detection limit $\epsilon = 0.0001$ in library *no. 2*. On the contrary, when the abundance of species was less variable (libraries *no. 5* and *no. 6*), the expected number of species was obtained with half the reads (Figure 4.4F) or even with 10% of reads (Figure 4.4E).

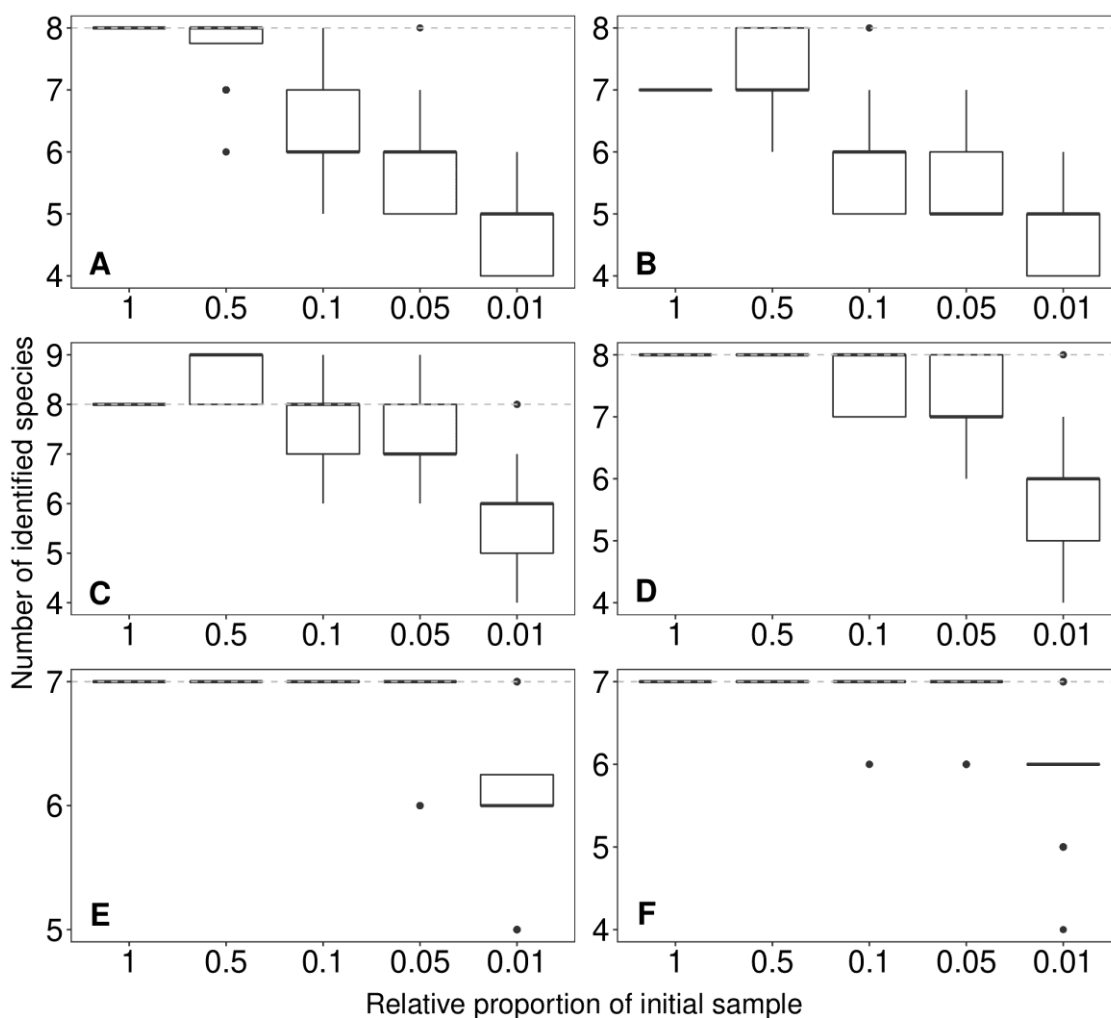


Figure 4.4. Number of identified species using different proportions of reads in the mixed-species libraries. Each simulation was performed 100 times with different subsets, except when the entire library was used. Letters from A to F indicate mixed-species libraries from *no. 1* to *no. 6*. Grey dashed lines indicate the expected number of recovered species in each library.

Table 4.4. Summary table of the species in single-species libraries data used for the obtention of the correction factors M_{Mi} and \bar{N}_{Mi} . x_i is the number of reads mapping into the mitogenome of species i divided by the total number of reads from that single-species library; N_{Mi} is the estimated number of mitochondrial DNA copies for species i ; \bar{N}_{Mi} is as N_{Mi} but calculated using the mean length of whole genomes and mitogenomes of all species considered here.

Species	Complete genome size (Mbp)	Mitochondrial genome size (kbp)	Number of raw reads (paired-end)	Number of reads mapped to the mitogenome	x_i (from Eq. 4.3)	N_{Mi} (from Eq. 4.2)	\bar{N}_{Mi} (from Eq. 4.4)
<i>Papilio machaon</i>	278	15.2	434,520	1,992	0.0046	84.4	95.5
<i>Drosophila virilis</i>	206	14.9	4,714,902	48,967	0.0104	144.6	217.6
<i>Drosophila melanogaster</i>	144	20	2,283,768	20,424	0.0089	66.4	187.1
<i>Drosophila mojavensis</i>	197	14.9	1,668,424	22,839	0.0137	180.5	287.8
<i>Bactrocera oleae</i>	472	15.8	580,996	4,052	0.007	209.5	145.6
<i>Linepithema humile</i>	220	16.1	1,422,342	3,151	0.0022	30.3	46
<i>Bombus terrestris</i>	249	17.4	1,994,938	16,262	0.0082	117.4	170.4
<i>Apis mellifera</i>	250	16.3	1,262,388	64,448	0.0511	823.9	1115.6
<i>Acyrtosiphon pisum</i>	542	17	684,688	7,218	0.0105	340.1	220.9

4.4.3. Computer use

The total consumed time by running the entire pipeline with the mixed-species libraries ranged between 66 and 82 minutes (Table S4.11). Most of the processing time was devoted to the mapping of the reads to the references (95%) and only 3% of the time was used by the γ - δ algorithm (Table S4.11).

4.5. Discussion

MMG proved to be able to set apart and quantify the relative DNA abundance of insect species in artificial mixtures, even when the species were congeneric. The estimation was as good as the one obtained with whole genomes instead of mitogenomes in chapter 3, using the same DNA libraries and bioinformatic methods as here. However, to be able to quantify the RSA in a sample with several species (*across-species* RSA) it was necessary to correct the raw reads by the variable amount of mitochondrial to nuclear DNA (mitochondrial DNA copy number) among species.

4.5.1. Species identification

The MMG approach used in this chapter recovered only the focal species from all the single-species libraries, with no FP when low-complexity regions were filtered out from the mitogenomes (except for genuine contaminants, see below). Without this filtering step, some reads were attributed to non-focal species; these reads were sequences with biased composition, likely from repetitive regions (*e.g.*, microsatellites) that mostly matched non-coding regions on the reference mitogenomes (Table S4.8) (Faber & Stepien 1998; Wolff *et al.*, 2012). Some popular tools do implicitly or explicitly filter out low-complexity regions from the reference genomes: Kraken masks low-

complexity regions when adding references to the database (Wood & Salzberg 2014); and BLAST filters both query sequences and references (Altschul *et al.*, 1990, 1997; Camacho *et al.*, 2009).

We also found contaminant species in all the sequenced libraries (Tables S4.9 and S4.10). The origin of these reads of contaminants can be tag jumping during the sequencing reaction (Schnell *et al.*, 2015) or actual contamination in the lab. The first reason is probably the cause of finding reads in single-species libraries belonging to other species sequenced in the same run but different libraries. The second reason is behind the presence of reads attributed to *Ceratitis capitata* in libraries where *Bactrocera oleae* was also present, because the two dipterans, which are agricultural pests, were captured together in fly traps. A more throughout discussion about this problem is provided in chapter 3. The removal of the genuine contaminant species in artificial libraries as we did here was possible because we knew the identity of the species in the mixture, but it is impossible in *real* samples.

4.5.2. Quantification of the RSA and the need for a species-specific correction factor

With the mixed-species libraries, we estimated the *within*-species RSA with high statistical confidence (Figure 4.2). Similar good results have been reported in previous studies that used mock samples (Bista *et al.*, 2018; Ji *et al.*, 2020). On the contrary, the *across*-species RSA estimation within a sample was not statistically significant in any sample (Figure 4.3A). These results contrast with the study of Gueuning *et al.* (2019) that reported good quantitative estimations of *across*-species RSA in artificial mixtures of wild bees.

Using the mitochondrial DNA copy number correction factor, the RSA *across*-species correlated significantly with the real values in the six artificial samples analysed (Eq. 4.2; Figure 4.3B), even when the mean genome and mitogenome sizes were used

instead of the species-specific value (Eq. 4.4; Figure 4.3C). Other studies reporting RSA estimation *across*-species also used some correction factor before comparing the expected and observed number of reads, but none included the genome size, as we did here. For instance, Gómez-Rodríguez *et al.* (2015) and Tang *et al.* (2015) considered the mitogenome size of the species and Tang *et al.* (2015) and Lang *et al.* (2019) the number of reads from the genome. Tang *et al.* (2015) is the only study that provides both the goodness of fit with and without the use of the correction factor, and the effect is very different from the one reported here, as the result was almost the same in both cases. One possible explanation is that Tang *et al.* (2015) dealt only with wild bees (a group of a few Hymenoptera families), so the interspecific differences might be low compared to our study which included species from four insect orders. Nevertheless, the effect of the mitochondrial DNA copy number on the estimation of RSA deserves more research effort if DNA-based techniques are to provide good quantitative results, both using MMG and amplicon metabarcoding targeting variable copy number regions.

The method used here to estimate the correction factor for the mitochondrial DNA copy number (*i.e.*, preparation and sequencing of a single-species library to a depth of *ca.* one million reads) has a cost that is not negligible. Ideally, there should be a method to independently estimate the mitochondrial DNA copy number of each species that did not involve sequencing. Such methods do exist because there is an interest in medicine to measure the mitochondrial DNA copy number for its relationship with several diseases. In medicine, the mitochondrial DNA copy number is usually estimated using quantitative PCR (qPCR) (Thyagarajan *et al.*, 2012); this method requires two primer pairs, one for a mitochondrial marker and one for a single-copy nuclear marker. These primers are known for humans, but it would be costly to generate them for every species in an environmental sample, especially for those species whose genome has not yet been sequenced (but see Liu *et al.*, 2018). The qPCR itself is cheap, but the preparative work for each species would be long.

It is important to emphasize that the RSA used here is based on the relative proportions of DNA of the species in the mixture, but what is needed in most

ecological applications is the relative proportions of biomass (or individual counts) of the different species. The reason behind our choice was to simplify the problem of obtaining the actual RSA from HTS reads in several steps. There is one bias caused by the variable mitochondrial-copy-number of the different species and there is another, independent, bias caused by the variable DNA content of the biomass of different species. We addressed here the first bias and obviated the second one. From our results, the RSA based on the biomass could be obtained by multiplying our estimates by the biomass-to-DNA ratio of each species, if known. There are very little data in the literature about the proportion of DNA to biomass in different species, but it can be very variable; for instance, Pornon *et al.* (2016) reports a very different DNA yield from the same number of pollen grains of three plant species.

In metabarcoding applications, some authors use empirical correction factors based on mixtures of known relative biomass of several species rather than in mixtures of DNA (*e.g.*, Matesanz *et al.*, 2019; Thomas *et al.*, 2016). In these cases, the correction factor solves for the mitochondrial copy number among species and also for the DNA-to-biomass ratio and the differential amplification efficiency caused by PCR. This method is undoubtedly practical but does not differentiate the relative importance of each source of bias.

4.5.3. Mito-metagenomics in *real* samples

The present study is based on artificial mixtures of a low number of species whose mitogenomes are already assembled. Thus, it is fair to question the value of our proposal in *real* samples with many more species, with a limited amount of DNA, where the prior species composition is unknown, or when the reference mitogenomes are obtained in the same experiment and are only partially assembled.

More complex mixtures

Real samples can contain hundreds of species, and that might affect the ability of the method to detect the less abundant ones. With the sequencing depth achieved here (~3.4 million raw reads per sample) (Table S4.2) we were able to detect three (out of four) species with an expected RSA below 1‰ (Table S4.10). The subsequent rarefaction experiment showed that with fewer reads more species become undetected (Figure 4.4). In consequence, it seems that at least $3.4 \cdot 10^6$ reads are needed to detect most species with an RSA above 1‰. Having hundreds of species in the mixture would not hamper the quantitative ability of the method, as most species would be above a 1‰ abundance. However, ultra-rich samples with thousands of species would require a higher sequencing depth to ensure the detection of most species.

Limited amount of DNA available

The Illumina TruSeq kit used here to prepare the libraries requires 1 µg of DNA and this might be a problem with small specimens or in DNA-poor samples. However, today there are alternative methods that provide good results with just 1 ng of DNA, like the Illumina Nextera DNA Flex kit (Sato *et al.*, 2019), albeit potential biases should be tested in future experiments for these kits.

Absence of mitogenomes in the reference database

Our results showed that the proposed methodology was robust in the absence of the mitogenomes of species in the reference database. Of course, the species without their mitogenome in the reference database will never be found, but their reads will not generate many FP, even for species with close relatives in the reference database (Table 4.3). The presence of species without their reference genome in the mixture is likely to occur frequently in *real* samples. The unassigned reads (or also when the prior

composition of the sample is unknown) can be further explored by mapping them against other databases, like COI barcodes from BOLD System; thus, the identity of more species will be revealed, albeit not their relative abundance.

Incomplete genomes

In several MMG studies, the reference mitogenomes are assembled from the same mixtures in which the RSA is intended to be quantified (Crampton-Platt *et al.*, 2016; Zhou *et al.*, 2013). In these cases, the mitogenomes are assembled *de novo* and, normally, they are incomplete. We do not see any impediment in using mitogenomes assembled in this way if all of them have a similar length and quality. However, we would advise against the simultaneous use of mitogenomes with disparate length or quality for quantification purposes, because that would bias the RSA towards the species with better mitogenomes (Tang *et al.*, 2015). On the contrary, the use of all available partial mitogenomes would be fine for identification purposes.

Estimation of the mitochondrial DNA copy number (N_{Mi})

Perhaps the most difficult problem in *real* samples is the estimation of N_{Mi} . The rationale that we propose for the estimation of N_{Mi} (Eq. 4.1 to 4.4) seems reasonable, but the devil is in the detail: the estimation of the variables needed to calculate N_{Mi} is paved with difficulties for species without a reference genome. First, the estimation of the proportion of reads that belong to the mitogenomes (x_i , Eq. 4.3) is biased, because we assumed that all reads belong to the same species (R_{Gi} , Eq. 4.3); however, there is always DNA that comes from other sources, like food, gut bacteria, parasites, etc. Consequently, the number of reads of the entire genome R_{Gi} is overestimated and, hence, x_i underestimated. Second, the size of the mitogenome and the whole genome is generally unknown for most species; even for the best studied species their whole genome is far from complete (*e.g.*, Paris *et al.*, 2020), so the estimated size (G_i) is an underestimation of the real size.

Nevertheless, despite the above problems, the correction factor N_{Mi} helped to reproduce the expected *across-species* RSA in our libraries. Similar results were obtained using the mean values of the mitogenome and whole genome sizes (\bar{N}_{Mi} , Eq. 4.4). The apparent lack of effect of the species-specific genome and mitogenome sizes might be caused to the low variability of the mitogenome size [coefficient of variation (CV) = 9%] and moderate variability of the whole genome size (CV = 47%) (Table 4.4). On the contrary, the proportion of reads mapping into the mitogenome (x_i , Eq. 4.3) was much more variable among species (CV = 113%).

Given the previous considerations, we suggest the use of the correction factor \bar{N}_{Mi} instead of N_{Mi} , for species without a reference genome and to estimate the three necessary variables (x_i , M_i , G_i) in the following way.

- x_i . The proportion of reads belonging to the mitogenome of species i could be estimated by shotgun sequencing a single-species DNA extract. The value of x_i would be an underestimation of the real value but given the high interspecific variability, the obtained x_i values should still be useful for correction purposes.
- M_i . Ninety per cent of the 1794 mitogenomes used here have a length of 14.9 to 17.0 kpb (*i.e.*, a rank of 2.1 kbp or 13% of the mean M_i) (Table S4.2). Consequently, we recommend the use of the mean value \bar{M} for the group of species of interest (Table S4.12).
- G_i . The length of the whole genome is more variable across species than the length of the mitogenomes: 90% of the 115 whole genomes of insects available at RefSeq (Table S3.1) have a length between 0.14 and 0.98 Gbp. However, if the insects are split by orders the variability of G_i is smaller for most insect orders (Table S4.13). Consequently, we would advise using the mean value \bar{G} for each group of taxa (*e.g.*, insect orders).

4.6. Concluding remarks

The approach presented here to identify insect species and to estimate their relative abundance in complex mixtures using MMG worked well with artificial samples of known composition for a select group of species whose mitogenomes are sequenced to an advanced degree. The key for the accurate estimation of the *across-species* RSA was a correction factor for the mitochondrial copy number of each species. We are aware that the proposed methodology is not immediately applicable to most *real* samples, so its real value should be tested on more of such samples.

5. Drastic reduction of false positive species in samples of insects by the combined use of two metagenomic classifiers

Chapter under review at BMC Bioinformatics

5. Drastic reduction of false positive species in samples of insects by the combined use of two metagenomic classifiers

5.1. Abstract

The use of high-throughput sequencing (HTS) to recover short DNA reads of many species has been widely applied on biodiversity studies, either as amplicon metabarcoding or shotgun metagenomics. These reads are assigned to species or other taxa using bioinformatic tools generically known as metagenomic classifiers. However, for different reasons, the final results often contain many false positive (FP) species. Here we focus on the reduction of FP species attributable to the classifiers. We benchmarked two popular classifiers, BLASTn followed by MEGAN6 (BM) and Kraken2 (K2), to analyse shotgun sequenced artificial single-species samples of insects. To reduce the number of FP, we combined the output of the two classifiers in two different ways: (1) by keeping only the reads that were attributed to the same species by both classifiers (intersection approach); and (2) by keeping the reads from the intersection approach plus all the reads assigned to some species by either classifier but not assigned to any species by the other one (union approach). In addition, we applied an analytical detection limit to further reduce the number of FP. As expected, both metagenomic classifiers used with default parameters generated an unacceptably high number of FP species (tens with BM, hundreds with K2). The FP species were not necessarily phylogenetically close, as some of them belonged to different orders of insects. The union approach failed to significantly reduce the number of FP, but the intersection approach got rid of most of them. Of the 21 single-species samples used, in 11 of them, there were no FP and in four of them, the FP were congeneric of the species used to prepare the sample. The addition of an analytic detection limit of 0.001 (0.1%) further reduced the number to *ca.* 0.5 FP species per sample. The almost universal fact that most metagenomic classifiers report many FP species hampers the

confidence of the DNA-based methods for assessing the biodiversity of biological samples. Our approach to the problem is extremely straightforward and significantly reduced the number of reported FP species.

Keywords BLAST, ensemble of classifiers, Insects, Kraken2, MEGAN, mitochondrial metagenomics, single-species libraries, species identification

5.2. Introduction

Metagenomic classifiers are widely used to analyse genetic data on biodiversity studies. Today there is an overwhelming variety of such tools and algorithms designed to provide accurate taxonomic identifications (Breitwieser *et al.*, 2019; Mande *et al.*, 2012). These classifiers are generally provided with default parameters to make them user-friendly but do not ensure the optimal performance of the classifier on the query dataset. Previous studies prove that when the metagenomic classifiers are used with the default parameters, they normally produce false positives (FP) species (*i.e.*, they detect species that are not present in the sample) (Harbert 2018; Peabody *et al.*, 2015). Albeit parameters optimization is the intuitive procedure for improving the accuracy of the results, in general, it is difficult to tune the parameters without a lengthy calibration process. Consequently, most applications still use default parameters of the tools (*e.g.*, Cribdon *et al.*, 2020; Harbert 2018; Piro *et al.*, 2017; Rodríguez-Martínez *et al.*, 2020). In such circumstances, there are two main approaches to reduce the number of FP. (1) The filtering or post-processing of the classifier's output to refine the assignment (Huson *et al.*, 2007; Paula *et al.*, 2021; Velsko *et al.*, 2018) and (2) the simultaneous use of several metagenomic classifiers that independently assess each sample and produce a combined result (Bazinet *et al.*, 2018; McIntyre *et al.*, 2017; Teeling & Glöckner 2012).

In this chapter we benchmark two popular metagenomic classifiers [BLASTn followed by MEGAN6 (Altschul *et al.*, 1990; Huson *et al.*, 2016); and Kraken2 (Wood *et al.*,

2019)] to identify the species contained in single-species samples. The classifiers compared the reads with a reference database of mitochondrial genomes, in what is so-called mitochondrial metagenomics (MMG) (Crampton-Platt *et al.*, 2016). Subsequently, we post-processed the results with simple techniques to see if the number of FP identifications decrease.

The main questions of the present work are two: (1) Do the metagenomic classifiers generate FP species when they are used with default parameters? (2) Can the number of FP species be reduced with simple post-processing methods?

5.3. Material and Methods

5.3.1. Reference mitogenomes

We downloaded all mitogenomes of insect species available on RefSeq repository plus 11 mitogenomes from GenBank of insect species whose complete genomes were available on RefSeq but that their mitogenomes were not (both repositories were consulted on 3rd May 2020). Species with more than one mitogenome were randomly dereplicated. We obtained a total of 1934 mitogenomes (Table S5.1).

5.3.2. Selection of species and preparation of the DNA libraries

Despite using the data of previous chapters, the present one is fully independent. Instead of assessing the capabilities of whole-genome and mitochondrial genome for the classification of species of insects as we did before, here we compare the performance of popular bioinformatic tools, like BLASTn (Altschul *et al.*, 1990), MEGAN6 (Huson *et al.*, 2016) and Kraken2 (Wood *et al.*, 2019), that we did not use before. Below we provide a short explanation on data gathering, but for more detailed

information the reader is referred to subchapter 2.3. Preparation of samples: Selection of the species, laboratory treatment and quality control.

From the list of 1934 species with mitogenomes, we selected 17 species. With real specimens of these species (as opposed to *in-silico* generated), we created 21 single-species libraries, each one of them containing DNA of one insect species (Table 2.1). So, four species were sequenced twice. Arguably, this kind of sample with only one species is especially suited for the test, because we know for every read the species to which it belongs; on the contrary, in artificial samples of several species, it is known the relative proportion of every species, but not the identity of every read. Albeit libraries were sequenced with a paired-end technology, we only used here the forward read of the pair (R1 files) because many *real* eDNA samples are likely to have very fragmented DNA.

5.3.3. Classification of reads to species

Individual classifiers

We selected two pipelines to assign species to DNA reads, (1) BLASTn (Altschul *et al.*, 1990) followed by MEGAN6 (Huson *et al.*, 2016) (BM) and (2) Kraken2 (Wood *et al.*, 2019) (K2). These tools were chosen because they are widespread among the bioinformatic community and because the underlying algorithms belong to very different approaches. Briefly, BLASTn search for similarities between the query and the reference sequences with local alignments from short exact matches and then extends the alignment to the rest of the query sequence (seed-and-extension algorithm); as multiple matches are reported, MEGAN6 is subsequently used to assign the query reads to taxa using the lowest-common ancestor (LCA) algorithm. On the contrary, Kraken2 seeks for exact matchings between the read's *k*-mers and reference taxa *k*-mers; then, it uses an LCA algorithm to assign a taxon to each read. As we are only interested in the classification at the species level, we ignored in both classifiers the

assignments to superior levels of taxa. We provide the results of the assignation of species by the BM and K2 to the reads of all the single-species libraries. Detailed information about the commands used to run the complete pipelines are provided at the subchapter 2.4. *Classification of reads to species: Matching and assignment steps* and at supplementary material (Methodology S5.1).

Combination of results: Union and intersection of classifiers' results

We combined the results from BM and K2 in a single common output in two distinct ways: union and intersection. In the union approach, a read is assigned to the species identified by any classifier unless both classifiers assign the read to different species, in which case it is discarded (Table 5.1). Thus, the union approach reduces the number of informative reads by eliminating those assigned to different species by the two pipelines. However, it also increases the number of informative reads by keeping those assigned to a species by any classifier, even if the other one did not assign the read. The intersection approach is much more restrictive, as only assign a read to a species when both BM and K2 provide the same result (Table 5.1).

Table 5.1. Rules of classification of a read r using the union and the intersection approaches of the metagenomic classifiers p and q . The read r can be assigned to a species (*e.g.*, species s or species n) or can remain not assigned (NA).

Case	Classification of read r by		Classification of read r when merging results with	
	Classifier p	Classifier q	Union	Intersection
no. 1	NA	NA	NA	NA
no. 2	Species s	NA	Species s	NA
no. 3	NA	Species s	Species s	NA
no. 4	Species s	Species n	NA	NA
no. 5	Species n	Species s	NA	NA
no. 6	Species s	Species s	Species s	Species s

Analytical detection limit

We further refined the above results (both from individual classifiers and the combination of results) by using an analytic detection limit (ϵ). Thus, to include a species in the species list of a sample, its abundance must be above the threshold or detection limit. We report the results without a detection limit and with the detection limits of 0.0001 (0.01%) and 0.001 (0.1%).

5.3.4. Metrics

As each sample belongs to only one species (*i.e.*, the focal species), we assumed that all reads belong to this species. However, this is not exactly true, because the samples also contain exogenous DNA that is also sequenced (*e.g.*, gut content, parasites, food, etc.). Nevertheless, we classified the reads into three categories: true positive (TP, when the read was assigned to the focal species), false positive (FP, when the read was assigned to a different species) and false negative (FN). The consideration of a read as a FN is tricky, because, in addition to the exogenous DNA mentioned above, most of the reads correspond to nuclear DNA and, therefore, will not map into the mitochondrial genomes. Thus, here we declared as a FN a read not assigned to any species by one classifier but assigned to the focal species by the other classifier. As an example, let's consider a read r assigned to the focal species by BM and not assigned (NA) by K2; this read r would be labelled as TP by BM and as FN by K2.

The true negative category (TN, when a read did not belong to any species was not assigned) is omitted, because all DNA sequences may be originated from a specimen (either from the focal species or exogenous DNA); one may argue that not assigned artefactual reads (*e.g.*, chimeric reads or reads loaded with sequencing errors) belong to this group, but we cannot distinguish them from not assigned reads due to database incompleteness. So, we prefer to ignore the TN in the analysis.

We used the following five metrics to evaluate the performance of each classifier and their combinations.

- Richness: Number of species assigned in each library.
- Relative proportion of informative reads (RPIR): proportion of assigned reads (TP + FP) over the total number of reads in the sample (after quality control).
- Precision: ratio of TP reads to the total assigned reads (TP + FP).
- Recall: ratio of TP reads of the assessed classifier and TP reads by any tool (TP + FN).
- Execution time: total consumed time by running the complete pipeline.

5.3.5. Hardware

Pipelines were run on a cluster with 12 identical compute nodes, each with the same architecture: two AMD Opteron(tm) Processor 4180 with 6 cores each, so 12 threads per node were available.

5.4. Results

5.4.1. Individual classifiers

The use of the two metagenomic classifiers with the default parameters detected a high number of species in the single-species libraries, where in theory there should have been only one. The BM method produced 13.2 ± 7.7 species per sample (Table 5.2A) belonging to 11.0 ± 7.7 families and 5.0 ± 2.4 orders per sample (Table S5.2). K2 produced an even higher value of 321.7 ± 122.7 species per sample belonging to 142.1 ± 38.3 families and 21.9 ± 3.0 orders per sample. The precision was higher for BM than for K2 (BM: 0.986 ± 0.015 ; K2: 0.757 ± 0.127) (Table 5.2A). As only the reads mapping into the mitogenome are useful in MMG, both classifiers used a very low RPIR (BM: 0.0069 ± 0.0069 ; K2: 0.0063 ± 0.0056). The recall was also higher with BM than with K2

(BM: 0.864 ± 0.158 ; K2: 0.820 ± 0.097) (Table 5.2A and Table S5.3). Finally, K2 was *ca.* 60 times faster than BM (Table 5.2A).

There were also reads that could be genuinely attributed to contamination, both from the lab and from the field sampling. The reads assigned to contaminant species are neither reported in the above results nor in Table 5.2 and Table 5.3, but they are provided as supplementary material (Table S5.2). More details about the contaminant species are provided in subchapter 2.5. *Contaminant species*.

5.4.2. Combined classifiers

The combination of the outputs of the two classifiers with the union method still produced a richness much higher than expected (316.7 ± 122.4 species per sample) (Table 5.2A). This value is just slightly lower than the one produced by K2 alone, so the union did not help to get rid of FP species.

On the contrary, the combination of the outputs of the two classifiers with the intersection method drastically reduced the number of FP species. The recovered richness decreased to 2.3 ± 1.9 species per sample and the precision was also much higher (0.998 ± 0.005) (Table 5.2A). In fact, there were no FP species in 11 samples (out of 21) (Table 5.3). In four of the remaining samples, the FP were of the same genus, whereas in the last six samples there were species of a different genus or even of a different order (Table 5.3). On the negative side, the elimination of reads reduced the RPIR to 0.0055 ± 0.0054 (Table 5.2A).

5.4.3. Use of an analytical detection limit

The use of an analytical detection limit of 0.0001 (0.01%) reduced the number of detected species, but the reduction was modest (Table 5.2B). The more stringent

detection limit of $\epsilon = 0.001$ (0.1%) removed many more FP species (Table 5.2C). Indeed, the combined use of the $\epsilon = 0.001$ detection limit with the intersection approach reduced the number of recovered species per sample to 1.5 ± 1.1 (16, out of 21, samples were free from FP species; Table 5.3) (results of all libraries and methods are provided in Tables S5.2 and S5.3).

Table 5.2. Benchmark metrics scores for each classifier without detection limit (A), with an analytical detection limit of 0.0001 (B) and with an analytical detection limit of 0.001 (C). For richness, the RPIR, precision and recall we provide the mean and standard deviation (SD) of all 21 samples (format mean \pm SD), and for processing time the sum of the total consumed time when running all the samples sequentially (format hh:mm:ss). The time for creating the databases and running in-house python scripts are omitted.

(A) Metric	BM	K2	Union	Intersection
Richness	13.2 \pm 7.7	321.7 \pm 122.7	316.7 \pm 122.4	2.3 \pm 1.9
RPIR	0.0069 \pm 0.0069	0.0063 \pm 0.0056	0.0072 \pm 0.0065	0.0055 \pm 0.0054
Precision	0.986 \pm 0.015	0.757 \pm 0.127	0.822 \pm 0.098	0.998 \pm 0.005
Recall	0.864 \pm 0.158	0.820 \pm 0.097	0.926 \pm 0.053	0.684 \pm 0.174
Processing time	01:51:30	00:01:55		01:52:42

(B) Metric	BM	K2	Union	Intersection
Richness	13.0 \pm 8.0	232.0 \pm 129.8	215.1 \pm 122.4	2.1 \pm 1.5
RPIR	0.0069 \pm 0.0069	0.0063 \pm 0.0056	0.0071 \pm 0.0065	0.0055 \pm 0.0054
Precision	0.986 \pm 0.015	0.762 \pm 0.131	0.827 \pm 0.103	0.998 \pm 0.005
Recall	0.864 \pm 0.158	0.820 \pm 0.097	0.926 \pm 0.053	0.684 \pm 0.174

(C) Metric	BM	K2	Union	Intersection
Richness	4.2 \pm 3.6	36.5 \pm 38.1	32.1 \pm 38.2	1.5 \pm 1.1
RPIR	0.0069 \pm 0.0069	0.0063 \pm 0.0056	0.0071 \pm 0.0065	0.0055 \pm 0.0054
Precision	0.989 \pm 0.015	0.806 \pm 0.141	0.872 \pm 0.105	0.998 \pm 0.005
Recall	0.864 \pm 0.158	0.820 \pm 0.097	0.926 \pm 0.053	0.684 \pm 0.174

Table 5.3. False positive species detected on each library by the intersection approach. For each library, we indicated the run and library codes, the name of focal species (its order in brackets), the number of congeneric species in the reference database, and a list of the FP species divided in congeneric and non-congeneric to the focal species. The last three columns contain the number of FP species detected with the analytical detection limits (ϵ) of 0, 0.0001 and 0.001. For each species, we indicated, in brackets, the RPIR and its order when it is different from the focal species. Order abbreviations are Col: Coleoptera, Dip: Diptera, Hem: Hemiptera, Hym: Hymenoptera, Lep: Lepidoptera. (Table shown in the next page)

Run - Library	Name of the focal species (Order)	Num. congeneric species within the database	False positive species		Number of false positive species		
			Congeneric to focal species	Non-congeneric to focal species	$\epsilon=0$	$\epsilon=0.0001$	$\epsilon=0.001$
1-1	<i>Papilio machaon</i> (Lep)	14			0	0	0
1-3	<i>Drosophila melanogaster</i> (Dip)	20			0	0	0
1-4	<i>Drosophila mojavensis</i> (Dip)	20			0	0	0
1-6	<i>Linepithema humile</i> (Hym)	0			0	0	0
1-10	<i>Acyrtosiphon pisum</i> (Hem)	0			0	0	0
2-1	<i>Atta colombica</i> (Hym)	0			0	0	0
2-4	<i>Drosophila melanogaster</i> (Dip)	20			0	0	0
2-5	<i>Drosophila mojavensis</i> (Dip)	20			0	0	0
2-7	<i>Drosophila suzukii</i> (Dip)	20			0	0	0
2-8	<i>Linepithema humile</i> (Hym)	0			0	0	0
2-11	<i>Vollenhovia emeryi</i> (Hym)	0			0	0	0
1-2	<i>Drosophila virilis</i> (Dip)	20	<i>D. littoralis</i> (0.0008)		2	1	0
1-5	<i>Bactrocera oleae</i> (Dip)	14	<i>B. biguttula</i> (0.0018)		1	1	1
1-8	<i>Bombus terrestris</i> (Hym)	3	<i>B. hypocrita</i> (0.0035) <i>B. waltoni</i> (0.0023) <i>B. ignitus</i> (0.0015)		3	3	3
1-9	<i>Apis mellifera</i> (Hym)	7	<i>A. nigrocincta</i> (0.0011) <i>A. florea</i> (0.0006) <i>A. laboriosa</i> (0.0002) <i>A. nuluensis</i> (0.0002) <i>A. andreniformis</i> (<0.0001) <i>A. cerana</i> (<0.0001) <i>A. dorsata</i> (<0.0001)		7	4	1
2-3	<i>Cimex lectularius</i> (Hem)	0	<i>Reduvius tenebrosus</i> (Hem) (0.0009) <i>Aquatica wuhana</i> (Col) (0.0001) <i>Prismognathus prossi</i> (Col) (0.0002)		2	2	0
2-9	<i>Plutella xylostella</i> (Lep)	1	<i>Vespa orientalis</i> (Hem) (0.0028) <i>Erigyna pyretorum</i> (Lep) (0.0013) <i>Pristomyrmex punctatus</i> (Hym) (0.001) <i>Allocaisidara bakeri</i> (Hem) (0.0001)		1	1	0
2-12	<i>Wasmannia auropunctata</i> (Hym)	0	<i>Barca bicolor</i> (Lep) (0.0004) <i>Pachycerina decemlineata</i> (Dip) (0.0002) <i>Myrmica scabrinodis</i> (Hym) (0.0013)		4	4	3
2-2	<i>Bemisia tabaci</i> (Hem)	1	<i>B. afer</i> (0.0008)		3	3	0
2-6	<i>Drosophila virilis</i> (Dip)	20	<i>D. littoralis</i> (0.0003)		2	2	0
2-10	<i>Solenopsis invicta</i> (Hym)	2	<i>S. richteri</i> (0.0183) <i>S. geminata</i> (0.0039)		3	3	3

5.5. Discussion

The occurrence of FP species in shotgun sequenced DNA samples seems to be a universal feature that compromises the reliability of the method. Whereas some FP species are produced by contamination during the sampling, in the lab or during the sequencing (Ficetola *et al.*, 2015; Hornung *et al.*, 2019; Kunin *et al.*, 2010), many others are produced by the bioinformatic tools used to assign species to reads (Escobar-Zepeda *et al.*, 2018; Hleap *et al.*, 2021; Walsh *et al.*, 2018). In this study, we have examples of both kinds, but we were able to avoid the contaminant species because we knew which ones were handled simultaneously in the lab. Regarding the misclassifications caused by the bioinformatic tools, we were able to almost eliminate all FP species by combining two popular metagenomic classifiers in a very simple way.

5.5.1. Individual metagenomic classifiers with default parameters

In the literature, different metagenomic classifiers have been compared against each other many times to seek the most suited one depending on the characteristics of the target organisms, laboratory treatment, sequencing technologies, read length, taxonomic rank, database completeness, etc. (Cribdon *et al.*, 2020; Escobar-Zepeda *et al.*, 2018; Lindgreen *et al.*, 2016; Velsko *et al.*, 2018; Walsh *et al.*, 2018). Our approach using the individual classifiers produced results similar to those reported in the literature. Thus, studies running BLAST, with or without MEGAN, had a precision above 90% (Cribdon *et al.*, 2020; McIntyre *et al.*, 2017; Paula *et al.*, 2021; Pearman *et al.*, 2020), as we report here. Similarly, the precision reported with Kraken2 is lower, 75-85% (Marcelino *et al.*, 2020a; McIntyre *et al.*, 2017; Sun *et al.*, 2021); again, these results are in concordance with our findings. Other studies also report a very long list of FP species for Kraken2 (Marcelino *et al.*, 2020a; Ye *et al.*, 2019) as we do here (Table S5.2). The reasons that explain why some methods work better for a particular kind of sample (*e.g.*, BM works better than K2 with our insect samples) depend on many

factors (Harbert 2018; McIntyre *et al.*, 2017; Pearman *et al.*, 2020; Ye *et al.*, 2019) but such analysis is beyond the scope of the present study. The simple truth is that both metagenomic classifiers performed poorly on their own because of an unacceptable number of FP species in samples consisting of DNA of only one species of insect.

5.5.2. Combination of the two metagenomic classifiers

The intersection method used to combine the two metagenomic classifiers significantly reduced the number of FP species. This result suggests that different classifiers misidentify reads in different ways, so the most robust way to present the results is to keep as informative only the reads assigned to the same species by the two classifiers. The important reduction in FP species is accompanied by a modest reduction of FP reads (Table 5.2A), as most FP species were represented by a low number of reads. The alternative union method kept a higher proportion of informative reads but failed in the elimination of FP species.

There are several other tools devised to unify results from several classifiers but, in general, they are more complex or require the use of specific software. These tools either combine profiling [*e.g.*, MetaMeta (Piro *et al.*, 2017) merges six tools] or read-a-read assignments [*e.g.*, WEVOTE (Metwally *et al.*, 2016) combines five tools by default and PhymmBL (Brady & Salzberg 2009) combines Phymm and BLAST]. The strategies used to merge tools can vary widely, but they generally infer taxa with a voting system or rank taxa with probabilistic scores. In general, these tools show that precision is higher when multiple classifiers are combined. Similarly, McIntyre *et al.* (2017) applied various ensemble approaches (*e.g.*, maximum-voting and abundance ranking) that outperformed individual tools. In terms of precision, our results are similar or even better than those reported by other studies (McIntyre *et al.*, 2017; Metwally *et al.*, 2016; Piro *et al.*, 2017).

5.5.3. The use of an analytic detection limit

The reported results showed that the number of FP species could be further reduced by using the simplest of the filtering or post-processing tools, the use of a threshold below which the occurrence of a species in the species list of a sample is ignored. As noted above, FP species generally have a low number of assigned reads, so the use of a simple threshold or detection limit helped to reduce the number of FP species. This approach is by no means new, as many authors use a detection limit to get rid of species, either in absolute terms (species must be above a certain number of reads) or in relative terms, as we do here (Alberdi *et al.*, 2018; Pompanon *et al.*, 2012; Velsko *et al.*, 2018).

In addition, there are other methods to discard unwanted species that have not been considered in this chapter, like the analysis of the distribution of reads across the genome (Breitwieser & Salzberg 2020; Crampton-Platt *et al.*, 2016; Donovan *et al.*, 2018), the calibration or tuning of the parameters of the metagenomic classifier (Bazinet *et al.*, 2018; Hleap *et al.*, 2021), the replication of samples (Ficetola *et al.*, 2015, 2016), the use of negative and positive controls (De Barba *et al.*, 2014; Ficetola *et al.*, 2016; Gardner *et al.*, 2019; Hornung *et al.*, 2019), the removal of low complexity sequences (Lu & Salzberg 2018; Piro *et al.*, 2017), cleaning reference database from contaminants (Lu & Salzberg 2018), limiting the reference database to target species or sequences (Arribas *et al.*, 2016; Paula *et al.*, 2015; Srivathsan *et al.*, 2016) or removing FP species that are unlikely present in the sample (Hornung *et al.*, 2019). All these methods would probably further reduce the number of FP species but at the cost of a more lengthy or more expensive process.

5.6. Concluding remarks

DNA-based identification methods based on HTS holds great potential for the study of biodiversity and interactions in ecological communities, yet this approach is not free

from shortcomings. One important of such shortcomings is the ubiquitous FP species reported by most metagenomic classifiers (Peabody *et al.*, 2015). Unless we find ways to reduce the number of FP in samples of known composition there will always be a shadow of a doubt about the high diversity reported in many field studies (Gonzalez *et al.*, 2016). Here we showed that the simple intersection of the output of two very different metagenomic classifiers drastically reduced the number of FP. When this result was combined with the application of an analytic detection limit of 0.001 (*i.e.*, species below an abundance of 0.1% are not considered), the number of FP species was reduced to a manageable figure of *ca.* 0.5 FP species per sample. All this was accomplished using the default parameters of the two classifiers, making our approach extremely straightforward and at reach to most research labs, even to those without strong bioinformatic expertise.

6. General discussion

6. General discussion

Metagenomics (MG) and metabarcoding (MB) are competing methodologies for the study of biodiversity in natural communities. Today, MB is the dominant technique for the analysis of eukaryote species because it is efficient and affordable (Chua *et al.*, 2021). MG is almost limited to prokaryotes due to their small genome size (thousands of times smaller than the eukaryote genomes) and highly populated reference databases (Escobar-Zepeda *et al.*, 2015; Singer *et al.*, 2020).

In theory, the MG approach should produce better quantitative results than MB, because the direct shotgun sequencing of DNA samples avoids the biases associated with PCR (Bista *et al.*, 2018; Taberlet *et al.*, 2012b; Zhou *et al.*, 2013). Additionally, the DNA from the whole genome provides orders of magnitude more genetic information than a single, or few, genetic markers as those used in MB (Srivathsan *et al.*, 2016), and greater taxonomic resolution (Coissac *et al.*, 2016; Papadopoulou *et al.*, 2015).

Despite the advantages of MG, this method is not very useful for eukaryotes because of the lack of sequenced genomes. However, the number of available whole-genomes is rising quickly (Figure 6.1), thanks to the drop in sequencing costs and by sequencing initiatives like the Earth BioGenome Project (Lewin *et al.*, 2018). The MG method can be also applied to mitochondrial genomes, in what is termed mitochondrial metagenomics (MMG). MMG method shares benefits with MG from shotgun sequencing data, plus the advantage of the mitogenomes' short length, natural enrichment in the cell, and many more available mitogenomes on public repositories (Figure 6.1) (Crampton-Platt *et al.*, 2016).

In this thesis, we evaluate the MG method as a PCR-free alternative for the biodiversity assessment of Metazoan species. To explore the limits of MG we simulate a future in which the whole genomes of all species are known by analysing *artificial* samples of insect species with highly complete genomes available on public repositories. We evaluated the MG (chapter 3) and MMG (chapter 4) methods to quantitatively

estimate species diversity in those artificial samples. In these two studies we applied our new γ - δ algorithm to assign species to reads and in chapter 5 we used two popular metagenomic classifiers for the same purpose using MMG.

In the discussion that follows we further review the MG and MMG methods seen in this thesis. The first (6.1) and second (6.2) subchapters are devoted to compare the ability of MG and MMG to identify species and to estimate their relative abundance (RSA) in complex samples, respectively. In the third subchapter (6.3) we examine the workflow regarding sample types, and metagenomic classifiers. Finally, we humbly attempt to forecast the short and long terms future of MG against competing molecular technologies for the biodiversity assessment of eukaryotes (6.4).

6.1. Identification of species

As we applied MG (chapter 3) and MMG (chapter 4) using both the same pipeline and the same dataset we have the opportunity to conduct a close comparison of the two techniques for the very first time on Metazoan species (Table 6.1).

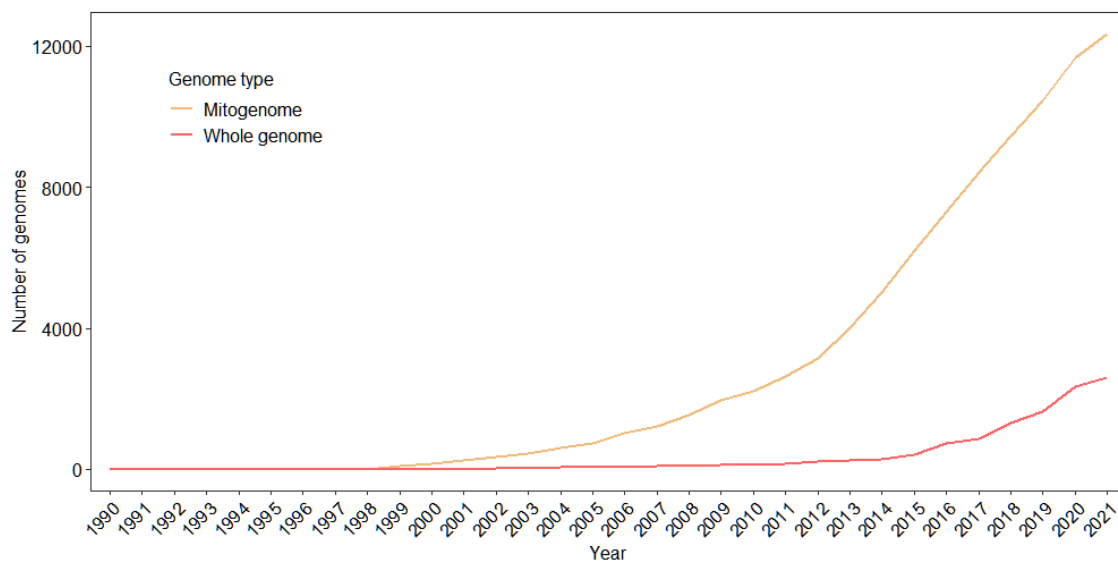


Figure 6.1. Number of available whole-genomes and mitogenomes of eukaryote species over the last thirty years. Data extracted from NCBI repository (URL: <https://www.ncbi.nlm.nih.gov/genome/browse/>) consulted on 7th June 2021.

Table 6.1. Comparison of the MG and MMG approaches. * Indicates that correlation was corrected with \bar{N}_{MI} value. As the number of sequenced mitogenomes is higher than the number of sequenced whole-genomes, two sets of results are provided for MMG, one with the results presented in chapter 4 and another using just the same number and identity of mitogenomes as the whole genomes available in chapter 3 ($n = 110$).

Metric	MG	MMG	MMG
Reads	Single-end raw reads	Paired-end candidate mito-reads	Paired-end candidate mito-reads
Number of reference genomes	110	1794	110
Reference genomes size (mean \pm SD)	318 \pm 188 MBytes	16.0 \pm 0.81 KBytes	16.2 \pm 0.99 KBytes
Filtering out the low-complexity regions from genomes	No	Yes	Yes
γ	0.99	0.99	0.99
δ	0.98	0.96	0.96
Detection limit (ϵ)	0.001	0.0001	Not needed
Richness in single-species libraries (mean \pm SD)	1.09 \pm 0.29	1 \pm 0	1 \pm 0
RPIR in single-species libraries (mean \pm SD)	0.466 \pm 0.150	0.0046 \pm 0.0057	0.0038 \pm 0.0042
Is a correction factor needed for <i>across-species</i> RSA estimation?	No	Yes	Yes
Correlation coefficient of <i>across-species</i> RSA estimation in mixed species libraries (max. - min.)	0.97-0.84	0.99-0.91*	0.99-0.91*
Total execution time (h:mm:ss) of mixed-species libraries (mean \pm SD)	1:22:29 \pm 0:09:45	1:14:24 \pm 0:06:29	0:05:41 \pm 0:00:32
Is it possible to use fewer reads?	Yes (~100 times)	No	No

From a theoretical point of view, the MG approach should produce better results, because whole genomes have several orders of magnitude more genetic information than mitogenomes: the taxonomic resolution should be greater in MG than in MMG, and the number of informative reads too. Somehow surprisingly, MMG was better in the identification of the species than MG, as MMG identified just the focal species in

the single-species libraries, whereas MG identified the focal species alongside some more (Table 6.1). In the mixed-species libraries, false positive (FP) species popped up in both methods; to get rid of FP species, in MG we had to impose a higher detection limit ($\epsilon = 0.001$) than in MMG ($\epsilon = 0.0001$) with the undesired effect of keeping out from the list species that were there (false negatives, FN) (Table 3.4). As expected, the number of informative reads was much higher in MG (~47% of the raw reads) than in MMG (~0.46%) (Table 6.1).

The misidentification of species has been a ubiquitous problem in this thesis, but also in many other studies (Bell *et al.*, 2019; Donovan *et al.*, 2018; Harbert 2018; Lu & Salzberg 2018; Peabody *et al.*, 2015). FP species are generated by contamination and wrong identification. Here we showed examples of both kinds. We could eliminate the contaminant species because we knew them, but for the removal of misidentified species during the bioinformatic analysis, we had to test different strategies throughout the different chapters of this thesis. In particular, we calibrated the bioinformatic tools (chapters 3 and 4), applied an analytical detection limit (chapters 3, 4, and 5), filtered out low-complexity regions from references (chapter 4), removed contaminated references (chapter 3), combined tools' outputs (chapter 5) and also combined the preceding strategies (*e.g.*, calibration of tools plus analytical detection limit in chapters 3 and 4). All of these strategies substantially reduced the number of FP. We applied the above-listed strategies to MMG and MG when deemed appropriate, but we never used them all at once nor in the same experiment.

One possible explanation for MMG being better than MG to inferring species is that we did not filter out low-complexity regions from the genomes in the MG approach. As we saw in chapter 4, most FP disappeared after filtering the low-complexity regions of the mitogenomes (Table 4.2A), the trick could have also worked for complete genomes.

The presence of DNA sequences from contaminating organisms, like bacterial, viral, and human DNA, on the published whole genomes, may also generate erroneous identifications (Marcelino *et al.*, 2020b; Merchant *et al.*, 2014). It is fair to assume that contaminant sequences are more likely to be unnoticed on whole-genomes references

rather than in mitogenomes. In fact, in chapter 3 we removed *Papilio xuthus* genome as contaminated (Table S3.1), but other contaminated references could have remained undetected. Perhaps, more rigorous cleaning of whole genomes from contaminant DNA sequences with specific tools (Lu & Salzberg 2018), would have decreased the number of FP.

Another possible explanation for MG producing more FP than MMG is that most of the available mitogenomes are complete, whereas most whole genomes are not (Breitwieser *et al.*, 2019; Escobar-Zepeda *et al.*, 2015; Paris *et al.*, 2020). The problem with incomplete reference genomes arises when the region of the genome that originated the read r is absent from the reference database. In such circumstances, the read r can match other sequences present in the reference database, like homologous sequences from closely related species or conserved genetic regions from distantly related species, and therefore generate a FP (Marcelino *et al.*, 2020b). This problem is likely to be alleviated with the growing number and completion of genomes available in public databases (Porter & Hajibabaei 2017).

6.2. Quantification of the relative species abundance

Both MG and MMG provided excellent results in the estimation of the RSA in the mixtures, with high correlations on the comparison of expected *versus* estimated RSA (Table 6.1); indeed, the z-test showed that no significant differences exist between the correlations provided by the two methods. Yet, each method has its own strengths and weaknesses that we consider below.

MMG needed a species-specific correction factor (N_{Mi} and \bar{N}_{Mi}) to estimate the *across*-species RSA; consequently, when the study aims to recover the RSA, the MG is the most appealing method, as it requires less *prior* information (*i.e.*, size of the genome and the mitogenome) and fewer steps in the computational pipeline.

Another key advantage of MG is that it produced accurate estimations of the RSA regardless of the degree of completeness of the reference genomes. In the mixed-species samples, there were species with reference whole genomes assembled at different levels; in particular, all species had their genome assembled at scaffold level but for *A. mellifera*, *B. terrestris* and *D. melanogaster* whose genomes were assembled at chromosome level. This result suggests that references assembled at scaffold level probably contain as much genetic information as chromosome assembled references do, albeit in a more fragmented presentation.

Regarding the execution time of the pipeline, MG is much slower than MMG when the same number of genomes are used, but the difference vanishes when all the available mitogenomes were included (Table 6.1); therefore, for a similar execution time, the MMG compares the samples against 16 times more references species. In our MMG implementation, the reduction in the execution time comes mostly from the selection of candidate mito-reads in MMG (~18 times; Table 4.2C) rather than from the smaller size of mitogenomes. It is possible that the mapping step with BWA, which is what takes longer to execute (Table S4.11), could be improved, but the optimization of the execution times was beyond the main objectives of this thesis.

Thanks to the high number of informative reads in MG, it was possible to reduce the number of raw reads 100 times and obtain similar results (Figure 3.3), whereas in MMG it was not possible to reduce, or just slightly, the number of reads (Figure 4.4). A practical corollary of this result is that when both whole genomes and mitogenomes become available in the future, it would be possible to multiplex many more libraries in a sequencing run in MG than in MMG. Importantly, as genomic repositories grow, the presence of homologous sequences between distinct species will increase, consequently, fewer reads will be assigned uniquely to a single species and higher sequencing depth will be required to detect a given species (Bohmann *et al.*, 2020). However, more research is needed to explore the trade-off between reference database size and the required depth of coverage for different species and methods.

6.3. A methodological perspective

6.3.1. The workflow

In this study, we created two kinds of artificial samples of insects, the single-species samples with DNA of only one species (Table 2.1), and the mixed-species samples with DNA of 8-9 species (Table 2.2). The single-species samples were used to calibrate the methods, like defining the optimal parameters (chapters 3 and 4), analysing strategies to eliminate FP from the results (chapters 3 and 5), and estimating the species-specific mitochondrial copy number (chapter 4). This information was subsequently applied to the study of the mixed-species samples to test the ability of the method to estimate the RSA. Ideally, the method should recover all species in the sample alongside their relative abundance, avoiding at the same time the emergence of FP species.

The idea of calibrating and testing the calibrated parameters to optimize the performance of the classification of reads is not new (*e.g.*, Hleap *et al.*, 2021); the novelty of our method relies on the ground-truth samples. Single-species samples are rarely used in molecular-based studies as positive controls; however, only with this kind of single-species samples we do truly know the origin of every read, whilst in mixed-species samples, only the overall composition is known. In addition, the single-species libraries proved to be useful to solve many calibration problems outlined above.

Alternatively, we could have used *in silico* generated datasets. *In silico* datasets are easier and cheaper to generate but may not be a truthful representation of a *real* scenario, because biological biases are difficult to simulate. Some of these biological biases are the proportion of focal species' DNA due to exogenous DNA (like parasites and food) being sequenced alongside the focal DNA; the mitochondrial copy number (seen in chapter 4), and the low (intraspecific) genetic distance between the query sequence and the reference because the same reference is used to create the query read and to populate the database [see Martos' Master Thesis (2020) that replicated

the MG experiment (chapter 3) with simulated datasets and yield much higher RPIR and better estimation of species composition].

The present workflow is based on *artificial* samples of DNA extracted from *real* organisms of a group of selected species. Consequently, the proposed workflow of this thesis may only be applied to mixtures of species whose whole-genome (or mitogenome) is available on online repositories. However, as we said before, not many species of Metazoan have their genomes already sequenced. Thus, from a practical point of view, an important question arises: how can this methodology be transferred to *real* samples where the identity of the species is unknown, and their genomes are not available? The easiest way to transfer our ground-truth samples to *real* samples is to collect and isolate a set of single specimens (or group of specimens of the same species) from the same location as the *real* sample. This subset may be representative of the total biodiversity and should be used to calibrate the pipeline. Then, single- and mixed-species samples can be created from those specimens, and subsequently run the entire workflow. Importantly, at first, the selected species may be unknown, but their identity can be revealed by mapping the single-species samples against a COI database. For practitioners, the validation with mixed-species samples can be avoided, and directly run the calibrated pipeline on the *real* samples. Second, in our experimental design we controlled that the genomes of the focal species were present in the reference database, but this may not be the case on *real* samples; therefore, we recommend downloading the genomes from various repositories (*e.g.*, RefSeq and GenBank) to generate a comprehensive database. Yet, if the reference genome of a given species is absent from the reference database, that species will never be found; so, as stated in subsubchapter 4.5.3. *Mito-metagenomics in real samples*, its presence (not its RSA) can be checked by mapping the sample against a barcode database. Third, if the *across*-species RSA is to be recovered with MMG, the N_{Mi} must be estimated for all detected species (see subsubchapter 4.5.3. *Mito-metagenomics in real samples*).

6.3.2. Metagenomic classifiers

In this thesis, we run three different metagenomic classifiers in the MMG method: $B\gamma\delta$ in chapter 4, BM and K2 in chapter 5. To compare all three methods on the same grounds, we run again the complete $B\gamma\delta$ method using the same single-species libraries, reference genomes, and informatic server that was used in chapter 5 and the $B\gamma\delta$ optimal parameters calibrated in chapter 4.

Individual classifiers

As stated before, the foremost characteristic of the single-species libraries used in this thesis is that the number of recovered species must be one. Focusing on this feature, $B\gamma\delta$ pipeline performed better than any other pipeline as it recovered only the expected species in all samples (Table 6.2). The other pipelines always produced some FP species, even when an analytical detection limit of 0.001 was applied. In terms of precision, $B\gamma\delta$ achieved the topmost value of 1, followed by BM (~ 0.989), and K2 (~ 0.8). Probably, the higher precision of $B\gamma\delta$ can be explained by the calibration of the γ and δ parameters with the same single-species samples that are analysed, whereas BM and K2 were used with default options that might provide suboptimal results. A preliminary analysis on the performance of BM and K2 when they are used after calibration, showed that results improved with optimal parameters, as precision were higher (BM: 0.998 ± 0.006 , K2: 0.999 ± 0.002) and the FP species were highly reduced (BM produced 1.350 ± 0.587 species per sample, and K2 produced 1.200 ± 0.523), yet the $B\gamma\delta$ remained as the most precise method while consumed the higher execution time (results not shown). Another possible explanation is the presence of low complexity regions that are prone to FP (Lu & Salzberg 2018; Pearman *et al.*, 2020). All classifiers masked the reference genomes but default masking in BLAST and Kraken2 may not remove all repetitive regions (dustmasker level 20 *versus* 45 by default and custom for $B\gamma\delta$, respectively). As expected, the RPIR was very low in all cases (Table 6.2), because, as pointed before, in shotgun samples the mitochondrial DNA represents a low proportion from the total genomic content. In terms of execution

time, K2 was the fastest method by far (Table 6.2). Indeed, this is an expected result as aligning reads is computationally more intensive than matching short k -mers.

Improved classifiers performance

To avoid classifier-specific pitfalls, we applied a few straightforward filtering steps that yield very good results. The major drawback of BM and K2 was the presence of many FP; to address this problem we combined their results with the intersection approach (*i.e.*, we kept only those reads assigned to the same species by both methods). The combined result was much better, albeit still remained about 0.5 misclassified species per sample (Table 6.2).

The main shortcoming of $B\gamma\delta$ was the long execution time of the mapping step. To alleviate this issue, we reduced the query input data by mining candidate mito-reads. Thus, we reduced 7.3 times the input sample size, and ~ 5.7 times the total consumed time, however, $B\gamma\delta$ was still the slowest pipeline (Table 6.2).

Unlike most previous comparison works, we limited the classifiers to perform a species-level identification which is particularly important for the assessment of biodiversity (Bertrand *et al.*, 2006) and also difficult for short reads (McIntyre *et al.*, 2017). We observed that no classifier was the best in all circumstances; nonetheless, all classifiers recovered the focal species as the only one present in the sample or the most abundant one (Table 6.2).

Table 6.2. Comparison of classifiers used in this thesis for MMG analysis of single-species libraries (Table 2.1).

Metric	BM	K2	Intersection	Byδ	Byδ
Reads	Single-end raw reads	Single-end raw reads	Single-end raw reads	Single-end raw reads	Single-end candidate mito-reads
Total number of reads	24,710,009	24,710,009	24,710,009	24,710,009	3,400,453
Detection limit (ε)	0.001	0.001	0.001	Not needed	Not needed
Richness (mean ± SD)	4.2 ± 3.6	36.5 ± 38.1	1.5 ± 1.1	1 ± 0	1 ± 0
RPIR (mean ± SD)	0.0069 ± 0.0069	0.0063 ± 0.0056	0.0055 ± 0.0054	0.0051 ± 0.0061	0.0050 ± 0.0060
Precision (mean ± SD)	0.989 ± 0.015	0.806 ± 0.141	0.998 ± 0.005	1 ± 0	1 ± 0
Total execution time	1:51:30	0:01:55	1:52:42	171:46:35	29:27:08

6.4. Future of molecular-based methods

Molecular-based methods for the biodiversity assessment of natural communities are transitioning from amplicon metabarcoding to whole genome MG. Today, DNA MB is still the current choice for most researchers because the procedure has been standardized, it is easy to apply, and it is time- and cost-efficient (Bohmann *et al.*, 2021). Probably, MG will be the technology of choice in the long term when the drawbacks of MG are overcome.

In the meantime, MMG will perhaps be more used. MMG has proved to infer species composition better than MB method (Bista *et al.*, 2018; Gueuning *et al.*, 2019; Tang *et al.*, 2015), probably because PCR biases makes it difficult to estimate the RSA. MMG also provides more genetic information because shotgun sequences may belong to multiple coding genes rather than a single barcode, providing a higher taxonomic resolution. However, while most of the sequence data is used in the MB method, only ~0.5-4% is useful in the MMG (Bista *et al.*, 2018; Gómez-Rodríguez *et al.*, 2015; Tang *et al.*, 2014). To increase the number of mitochondrial DNA sequences, mitochondrial enrichment methods have been applied, like centrifugation (Macher *et al.*, 2018; Zhou *et al.*, 2013) or target capture (Liu *et al.*, 2016; Wilcox *et al.*, 2018). Regardless of the method, such enrichment should not introduce additional biases and, indeed, listed methods presented potential unpredictable efficiencies on different taxa; so further optimization of current laboratory protocols is required for unbiased mitochondrial enrichment. In the absence of a reliable solution for MMG pitfalls, MMG can be used as a robust method to identify species in samples, regardless of their RSA (Crampton-Platt *et al.*, 2016).

Over the duration of this thesis, new strategies have been proposed for metagenomic approaches, involving different HTS technologies (*e.g.*, Illumina, Sanger, and MinION) (Oliveira *et al.*, 2018) and targeting different genomic regions (*e.g.*, few DNA metabarcodes, thousands of loci from RADseq or mitogenomes, millions from whole genomes) (Andrews *et al.*, 2018). In eukaryote species assessment, two new methods

stand out, and both have been applied only to plant samples: Reverse Metagenomics (Peel *et al.*, 2019) and Genotype by Sequencing (Wagemaker *et al.*, 2021).

Peel *et al.* (2019) Reverse Metagenomics (RevMet) is an innovative reference-free metagenomic method that combines genome skimming and nanopore sequencing. The RevMet method is applied in two steps. First, it creates reference sequences by genome skimming single-species samples. Second, the query samples are sequenced with MinION nanopore sequencing; the genome skims are then mapped against the MinION long reads. Peel *et al.* (2019) show that the technique uses ~50-65% of the long reads, and also identifies species at the lowest proportion of 1% and estimates the RSA. Yet, the identification and quantitative ability of the method is limited by the coverage of the reference-skim, FP detection of congeneric species and high sequencing error rates.

Another promising approach is the multiple-species Genotype by Sequencing (msGBS) method designed by Wagemaker *et al.* (2021). msGBS reduces the complexity of the genomic DNA (gDNA) using restriction enzymes to fragment the gDNA so only a subset of the whole genome is sequenced. Importantly, this subset is always the same for a given species. The msGBS method is run in two steps. First, single-species samples are used to generate gDNA clusters to create a reference database. Second, query samples are sequenced with the same laboratory protocol, and the resulting query gDNA clusters are mapped against the database. Wagemaker *et al.* (2021) show that most of the query gDNA clusters are assigned to species. However, this method is limited by not having a reference database and the need for species-specific correction factors that accounted for the varying DNA-to-biomass ratio to yield quantitative estimation of *across-species* RSA.

Regardless of the method applied, four main problems should be addressed if metagenomics is to be used to assessing *real* samples at a large scale.

- Sequencing enough good quality and quantity DNA from the query samples. Without a reliable solution, the high sequencing depth necessary to yield

enough DNA may result in increasing sampling and laboratory workloads, and economic costs.

- Populate genomic repositories with as much diverse genetic information as possible, both inter- and intraspecific.
- Ensure well-curated genomic records, free from contaminant sequences. As exogenous DNA sequences on references may generate wrong identifications and, therefore, misleading conclusions.
- With the growing size of reference repositories, computational challenges arise and press for highly efficient memory systems and software for data processing.

7. Conclusions and future works

7. Conclusions and future works

The research done in this thesis work contributes to the advance of molecular methods for the assessment of eukaryote species in natural communities. With *artificial* samples of insect species whose genomes are already sequenced, we simulated a future with complete online genomic repositories. Within this scenario, we applied the metagenomic (MG) and mito-metagenomic (MMG) methods, and we extracted the following conclusions:

- The MG and MMG methods hold great potential for the assessment of biodiversity of Metazoan species. Indeed, we showed that MG and MMG methods can infer species composition, even when congeneric species are present in the sample.
- The shotgun sequencing of single-species samples recovered genomic information from the focal species but also DNA from diverse origins (like parasites, and gut content); from the “soup” of DNA, about 47% of reads in MG and 0.5% in MMG were truly informative for species identification.
- We proved that MG and MMG can recover with confidence the relative species abundance (RSA) above a detection limit of 0.1% in MG and 0.01% in MMG. For the MG method, robust RSA was still estimated with a reduction of 100-fold of the input samples (*ca.* 17,000 single-end reads per sample); no reduction was possible in the MMG method (*ca.* 3.4 million paired-end reads per sample were needed).
- We demonstrated that the RSA reported by MMG can change according to the way we look at the species composition, that is *within*-species or *across*-species. We recovered the RSA with both approaches, yet for the *across*-species RSA a species-specific correction factor was required to overcome the mito-to-nuclear DNA ratio.

- When the metagenomic classifiers BM and K2 were used in default mode, they generated an unacceptable high number of false positive species. The unpredictable performance of metagenomic classifiers highlights the importance of calibrating the computational pipelines with samples of known composition prior to analysing *real* samples.
- We developed the γ - δ algorithm as a new metagenomic classifier and showed that it was more precise than BM and K2 but classified less of reads and consumed more computer resources.
- The presence of false positive was the rule rather than the exception in all methods and metagenomic classifiers. We applied a variety of strategies to avoid the detection of false positive species and demonstrated that false positive levels can effectively be reduced with straightforward techniques.
- Even though the proposed methodologies are not immediately applicable because of incomplete databases, we give some advice about the applicability of the methods to *real* samples.

Despite the above listed conclusions, our research also opens new problems and questions. Below we list some of these issues and propose possible ways to address them:

- When the genome of most species would be available, the application of the $B\gamma\delta$ classifier (as used here) is unpractical due to long execution times; thus, it is mandatory to reduce the execution time. Because the mapping with BWA is the most time-consuming step, the search step should be optimized. Some options could be the study of BWA performance for optimise the usage of computational resources, seek for alternative and faster aligners like GEM, or the reduction of the input data (both references and query samples).

- To explore whether the proposed pipeline can also recover the species biomass. For this purpose, we suggest creating single- and mixed-species samples with a known amount of biomass and repeat the complete workflow (*i.e.*, calibration and validation of the pipeline). Additionally, the single-species samples will virtually facilitate the design of a correction factor for the DNA-to-biomass ratio, if needed.
- To study the impact of growing online repositories on the false positive rates and also on the depth of coverage required for detecting a species. To this end, we suggest the study of the same query dataset on a simulated scenario of rising number of reference genomes in the repositories which can be created by controlling the number of available reference genomes.
- To reduce the false positive rates in the metagenomics methods, a comprehensive analysis of different filtering approaches (*e.g.*, the addition of positive and negative controls, the analysis of the distribution of reads over the genome or cleaning the reference database) needs to be carried out.
- To compare the performance of B γ δ classifier against other metagenomic classifiers when all classifiers are used with optimal parameters. Indeed, we presented preliminary results of the identification capabilities when BM, K2, and B γ δ are used with calibrated parameters on single-species samples, yet the comparison of their ability to recover the RSA is still missing.

References

References

- Adamski Z, Bufo SA, Chowański S, Falabella P, Lubawy J, Marciniak P, Pacholska-Bogalska J, Salvia R, Scrano L, Słocińska M, Spochacz M, Szymczak M, Urbański A, Walkowiak-Nowicka K, Rosiński G (2019) Beetles as model organisms in physiological, biomedical and environmental studies - A review. *Frontiers in Physiology* 10: 319. <https://doi.org/10.3389/fphys.2019.00319>
- Agustí N, Shayler SP, Harwood JD, Vaughan IP, Sunderland KD, Symondson WOC (2003), Collembola as alternative prey sustaining spiders in arable ecosystems: prey detection within predators using molecular markers. *Molecular Ecology* 12: 3467-3475. <https://doi.org/10.1046/j.1365-294X.2003.02014.x>
- Alberdi A, Aizpurua O, Gilbert MTP, Bohmann K (2018) Scrutinizing key steps for reliable metabarcoding of environmental samples. *Methods in Ecology and Evolution* 9: 134-147. <https://doi.org/10.1111/2041-210X.12849>
- Almeida A, Mitchell AL, Tarkowska A, Finn RD (2018) Benchmarking taxonomic assignments based on 16S rRNA gene profiling of the microbiota from commonly sampled environments. *GigaScience* 7(5): giy054. <https://doi.org/10.1093/gigascience/giy054>
- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *Journal of Molecular Biology* 215(3): 403-410. [https://doi.org/10.1016/s0022-2836\(05\)80360-2](https://doi.org/10.1016/s0022-2836(05)80360-2)
- Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* 25(17): 3389-3402. <https://doi.org/10.1093/nar/25.17.3389>
- Anderson MJ (2001) A new method for non-parametric multivariate analysis of variance. *Austral Ecology* 26: 32-46. <https://doi.org/10.1111/j.1442-9993.2001.01070.pp.x>
- Andrews KR, De Barba M, Russello MA, Waits LP (2018) Advances in using non-invasive, archival, and environmental samples for population genomic studies. In: Hohenlohe PA, Rajora OP (Eds) *Population Genomics: Wildlife*. Population Genomics. Springer, Cham, 63-99. https://doi.org/10.1007/13836_2018_45
- Andrews S (2015) FastQC: a quality control tool for high throughput sequence data. URL: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc>
- Ansorge WJ (2009) Next-generation DNA sequencing techniques. *New Biotechnology* 25(4): 195-203. <https://doi.org/10.1016/j.nbt.2008.12.009>

- Arnot DE, Roper C, Bayoumi RA (1993) Digital codes from hypervariable tandemly repeated DNA sequences in the *Plasmodium falciparum* circumsporozoite gene can genetically barcode isolates. *Molecular and Biochemical Parasitology* 61(1):15-24. [https://doi.org/10.1016/0166-6851\(93\)90154-P](https://doi.org/10.1016/0166-6851(93)90154-P)
- Arribas P, Andújar C, Hopkins K, Shepherd M, Vogler AP (2016) Metabarcoding and mitochondrial metagenomics of endogean arthropods to unveil the mesofauna of the soil. *Methods in Ecology and Evolution* 7(9): 1071-1081 <https://doi.org/10.1111/2041-210X.12557>
- Arute F, Arya K, Babbush R, Bacon D, Bardin JC, Barends R, Biswas R, Boixo S, Brandao FGSL, Buell DA, Burkett B, Chen Y, Chen Z, Chiaro B, Collins R, Courtney W, Dunsworth A, Farhi E, Foxen B, Fowler A, Gidney C, Giustina M, Graff R, Guerin K, Habegger S, Harrigan MP, Hartmann MJ, Ho A, Hoffmann M, Huang T, Humble TS, Isakov SV, Jeffrey E, Jiang Z, Kafri D, Kechedzhi K, Kelly J, Klimov PV, Knysh S, Korotkov A, Kostrița F, Landhuis D, Lindmark M, Lucero E, Lyakh D, Mandrà S, McClean JR, McEwen M, Megrant A, Mi X, Michielsen K, Mohseni M, Mutus J, Naaman O, Neeley M, Neill C, Niu MY, Ostby E, Petukhov A, Platt JC, Quintana C, Rieffel EG, Roushan P, Rubin NC, Sank D, Satzinger KJ, Smelyanskiy V, Sung KJ, Trevithick MD, Vainsencher A, Villalonga B, White T, Yao ZJ, Yeh P, Zalcman A, Neven H, Martinis JM (2019) Quantum supremacy using a programmable superconducting processor. *Nature* 574: 505-510. <https://doi.org/10.1038/s41586-019-1666-5>
- Baldwin CC, Collette BB, Parenti LR, Smith DG, Springer VG (1996) Collecting fishes. In: Lang MA, Baldwin CC (Eds) *Methods and Techniques of Underwater Research, Proceedings of the 16th Annual Scientific Diving Symposium, American Academy of Underwater Sciences*, 11-33.
- Barrett RDH, Hebert PDN (2005) Identifying spiders through DNA barcodes. *Canadian Journal of Zoology* 83(3): 481-491. <https://doi.org/10.1139/z05-024>
- Bartlett SE, Davidson WS (1991) Identification of *Thunnus* Tuna Species by the Polymerase Chain Reaction and Direct Sequence Analysis of their Mitochondrial Cytochrome *b* Genes. *Canadian Journal of Fisheries and Aquatic Sciences* 48(2): 309-317. <https://doi.org/10.1139/f91-043>
- Bazinet AL, Ondov BD, Sommer DD, Ratnayake S (2018) BLAST-based validation of metagenomic sequence assignments. *PeerJ* 6: e4892. <https://doi.org/10.7717/peerj.4892>
- Bell KL, Burgess KS, Botsch JC, Dobbs EK, Read TD, Brosi BJ (2019) Quantitative and qualitative assessment of pollen DNA metabarcoding using constructed species mixtures. *Molecular Ecology* 28: 431-455. <https://doi.org/10.1111/mec.14840>

- Benjamini Y, Speed TP (2012) Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Research* 40: e72. <https://doi.org/10.1093/nar/gks001>
- Bertrand Y, Pleijel F, Rouse GW (2006) Taxonomic surrogacy in biodiversity assessment, and the meaning of Linnean ranks. *Systematics and Biodiversity* 4(2): 149-159. <https://doi.org/10.1017/S1477200005001908>
- Bista I, Carvalho GR, Tang M, Walsh K, Zhou X, Hajibabaei M, Shokralla S, Seymour M, Bradley D, Liu S, Christmas M, Creer S (2018) Performance of amplicon and shotgun sequencing for accurate biomass estimation in invertebrate community samples. *Molecular Ecology Resources* 00: 1-18. <https://doi.org/10.1111/1755-0998.12888>
- Blaxter M (2003) Molecular systematics: Counting angels with DNA. *Nature* 421: 122-124. <https://doi.org/10.1038/421122a>
- Bohmann K, Chua P, Holman LE, Lynggaard C (2021) DNAqua-Net conference unites participants from around the world with the quest to standardize and implement DNA-based aquatic biomonitoring. *Environmental DNA* 00: 1-5. <https://doi.org/10.1002/edn3.207>
- Bohmann K, Mirarab S, Bafna V, Gilbert MTP (2020) Beyond DNA barcoding: The unrealized potential of genome skim data in sample identification. *Molecular Ecology* 29: 2521-2534. <https://doi.org/10.1111/mec.15507>
- Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30(15): 2114-2120. <https://doi.org/10.1093/bioinformatics/btu170>
- Boomsma JJ, Brady SáG, Dunn RR, Gadau J, Heinze J, Keller L, Sanders NJ, Schrader L, Schultz TR, Sundström L, Ward PS, Wcislo WT, Zhang G, The GAGA Consortium (2017) The Global Ant Genomics Alliance (GAGA). *Myrmecological News* 25: 61-66. URL: <http://antgenomics.dk/>
- Boratyn GM, Thierry-Mieg J, Thierry-Mieg D, Busby B, Madden TL (2018) Magic-BLAST, an accurate DNA and RNA-seq aligner for long and short reads. *bioRxiv*. <https://doi.org/10.1101/390013>
- Bouchet P (2006) The magnitude of marine biodiversity. In: Duarte CM (Eds) *The Exploration of Marine Biodiversity. Scientific and Technological Challenges*. Fundación BBVA, Bilbao, 31-64.
- Brady A, Salzberg SL (2009) Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. *Nature Methods* 6(9):673-676. <https://doi.org/10.1038/nmeth.1358>

- Bray NL, Pimentel H, Melsted P, Pachter L (2016) Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology* 34: 525-527. <https://doi.org/10.1038/nbt.3519>
- Breitwieser FP, Lu J, Salzberg SL (2019) A review of methods and databases for metagenomic classification and assembly. *Briefings in Bioinformatics* 20(4): 1125-1136. <https://doi.org/10.1093/bib/bbx120>
- Breitwieser FP, Salzberg SL (2020) Pavian: Interactive analysis of metagenomics data for microbiomics and pathogen identification. *Bioinformatics* 36(4): 1303-1304 <https://doi.org/10.1093/bioinformatics/btz715>
- Bronstein JL, Alarcon R, Geber M (2006) The evolution of plant-insect mutualisms. *New Phytologist* 172: 412-428. <https://doi.org/10.1111/j.1469-8137.2006.01864.x>
- Buss DF, Carlisle DM, Chon TS, Culp J, Harding JS, Keizer-Vlek HE, Robinson WA, Strachan S, Thirion C, Hughes RM (2015) Stream biomonitoring using macroinvertebrates around the globe: a comparison of large-scale programs. *Environmental Monitoring and Assessment* 187: 4132. <https://doi.org/10.1007/s10661-014-4132-8>
- Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL (2009). BLAST+: architecture and applications. *BMC Bioinformatics* 10: 421. <https://doi.org/10.1186/1471-2105-10-421>
- Cameron SL (2014) How to sequence and annotate insect mitochondrial genomes for systematic and comparative genomics research. *Systematic Entomology* 39(3): 400-411. <https://doi.org/10.1111/syen.12071>
- Carn VM (1996) The role of dipterous insects in the mechanical transmission of animal viruses. *British Veterinary Journal* 152(4): 377-393. [https://doi.org/10.1016/S0007-1935\(96\)80033-9](https://doi.org/10.1016/S0007-1935(96)80033-9)
- CBoL Plant Working Group (2009). A DNA barcode for land plants. *Proceedings of the National Academy of Sciences of the United States of America* 106(31): 12794-12797. <http://dx.doi.org/10.1073/pnas.0905845106>
- Chambers EA, Hebert PDN (2016) Assessing DNA barcodes for species identification in North American reptiles and amphibians in natural history collections. *PLoS ONE* 11: e0154363. <https://doi.org/10.1371/journal.pone.0154363>
- Chandler JA, Morgan Lang J, Bhatnagar S, Eisen JA, Kopp A (2011) Bacterial Communities of Diverse *Drosophila* Species: Ecological Context of a Host-Microbe Model System. *PLoS Genetics* 7(9): e1002272. <https://doi.org/10.1371/journal.pgen.1002272>

- Chen K, Pachter L (2005) Bioinformatics for whole-genome shotgun sequencing of microbial communities. *PLoS Computational Biology* 1(2): e24. <https://doi.org/10.1371/journal.pcbi.0010024>
- Cheng S, Melkonian M, Smith SA, Brockington S, Archibald JM, Delaux P-M, Li F-W, Melkonian B, Mavrodiev EV, Sun W, Fu Y, Yang H, Soltis DE, Graham SW, Soltis PS, Liu X, Xu X, Wong GK-S (2018) 10KP: A phylodiverse genome sequencing plan. *GigaScience* 7(3): giy013. <https://doi.org/10.1093/gigascience/giy013>
- Chua PY, Crampton-Platt A, Lammers Y, Alsos IG, Boessenkool S, Bohmann K (2021) Metagenomics: A viable tool for reconstruction herbivore diet. *Molecular Ecology Resources* 00: 1-15. <https://doi.org/10.1111/1755-0998.13425>
- Clare EL, Chain FJJ, Littlefair JE, Cristescu ME (2016) The effects of parameter choice on defining molecular operational taxonomic units and resulting ecological analyses of metabarcoding data. *Genome* 59(11): 981-990. <https://doi.org/10.1139/gen-2015-0184>
- Coissac E, Hollingsworth PM, Lavergne S, Taberlet P (2016) From barcodes to genomes: extending the concept of DNA barcoding. *Molecular Ecology* 25: 1423-1428. <https://doi.org/10.1111/mec.13549>
- Costello MJ, May RM, Stork NE. (2013) Can we name Earth's species before they go extinct?. *Science* 339: 413-416. <https://doi.org/10.1126/science.1230318>
- Crampton-Platt A, Timmermans MJTN, Gimmel ML, Kutty SN, Cockerill TD, Khen CV, Vogler AP (2015). Soup to tree: The phylogeny of beetles inferred by mitochondrial metagenomics of a borean rainforest sample. *Molecular Biology and Evolution* 32(9): 2302-2316. <https://doi.org/10.1093/molbev/msv111>
- Crampton-Platt A, Yu DW, Zhou X, Vogler AP (2016) Mitochondrial metagenomics: letting the genes out of the bottle. *Gigascience* 5(1): 15. <https://doi.org/10.1186/s13742-016-0120-y>
- Cribdon B, Ware R, Smith O, Gaffney V, Allaby RG (2020) PIA: More accurate taxonomic assignment of metagenomic data demonstrated on sedaDNA from the North Sea. *Frontiers in Ecology and Evolution* 8: 84. <https://doi.org/10.3389/fevo.2020.00084>
- De Barba M, Miquel C, Boyer F, Mercier C, Rioux D, Coissac E, Taberlet P (2014) DNA metabarcoding multiplexing and validation of data accuracy for diet assessment: application to omnivorous diet. *Molecular Ecology Resources* 14(2): 306-323. <https://doi.org/10.1111/1755-0998.12188>
- De Barba M, Waits LP, Genovesi P, Randi E, Chirichella R, Cetto E (2010) Comparing opportunistic and systematic sampling methods for non-invasive genetic

monitoring of a small translocated brown bear population. *Journal of Applied Ecology* 47: 172-181. <https://doi.org/10.1111/j.1365-2664.2009.01752.x>

Deagle BE, Gales NJ, Evans K, Jarman SN, Robinson S, Trebilco R, Hindell MA (2007) Study seabird diet through genetic analysis of faeces: a case study on macaroni penguins (*Eudyptes chrysolophus*). *PLoS ONE* 2(9): e831. <https://doi.org/10.1371/journal.pone.0000831>

Deagle BE, Jarman SN, Coissac E, Pompanon F, Taberlet P (2014) DNA metabarcoding and the cytochrome *c* oxidase subunit I marker: not a perfect match. *Biology Letters* 10: 20140562. <http://dx.doi.org/10.1098/rsbl.2014.0562>

Deagle BE, Thomas AC, McInnes JC, Clarke LJ, Vesterinen EJ, Clare EL, Kartzinel TR, Eveson JP (2018) Counting with DNA in metabarcoding studies: How should we convert sequence reads to dietary data?. *Molecular Ecology* 28: 391-406. <https://doi.org/10.1111/mec.14734>

deWaard JR, Ivanova NV, Hajibabaei M, Hebert PDN (2008) Assembling DNA barcodes. In: Martin CC, Martin CC (Eds) Environmental Genomics. Methods in Molecular Biology, vol 410. Humana Press. https://doi.org/10.1007/978-1-59745-548-0_15

Dirzo B, Young HS, Galetti M, Ceballos G, Isaac NJB, Collen B (2014) Defaunation in the Anthropocene. *Science* 345: 401-406. <https://doi.org/10.1126/science.1251817>

Donovan PD, Gonzalez G, Higgins DG, Butler G, Ito K (2018) Identification of fungi in shotgun metagenomics datasets. *PLoS ONE* 13(2): e0192898. <https://doi.org/10.1371/journal.pone.0192898>

Ebach MC, Valdecasas AG, Wheeler QD (2011) Impediments to taxonomy and users of taxonomy: accessibility and impact evaluation. *Cladistics* 27: 550-557. <https://doi.org/10.1111/j.1096-0031.2011.00348.x>

Elbrecht V, Leese F (2015) Can DNA-based ecosystem assessments quantify species abundance? Testing primer bias and biomass-sequence relationships with an innovative metabarcoding protocol. *PLoS ONE* 10(7): e0130324. <https://doi.org/10.1371/journal.pone.0130324>

Erlank E, Koekemoer LL, Coetzee M (2018) The importance of morphological identification of African anopheline mosquitoes (Diptera: Culicidae) for malaria control programmes. *Malaria Journal* 17: 43. <https://doi.org/10.1186/s12936-018-2189-5>

Ermakov OA, Simonov E, Surin VL, Titov SV, Brandler OV, Ivanova NV, Borisenko AV (2015) Implications of Hybridization, NUMTs, and Overlooked Diversity for DNA Barcoding of Eurasian Ground Squirrels. *PLoS ONE* 10(1): e0117201. <https://doi.org/10.1371/journal.pone.0117201>

- Escobar-Zepeda A, Godoy-Lozano EE, Raggi L, Segovia L, Merino E, Gutiérrez-Rios RM, Juarez K, Licea-Navarro AF, Pardo-Lopez L, Sanchez-Flores A (2018) Analysis of sequencing strategies and tools for taxonomic annotation: Defining standards for progressive metagenomics. *Scientific Reports* 8: 12034
<https://doi.org/10.1038/s41598-018-30515-5>
- Escobar-Zepeda A, Vera-Ponce de León A, Sanchez-Flores A (2015) The road to metagenomics: from microbiology to DNA sequencing technologies and bioinformatics. *Frontiers in Genetics* 6: 348.
<https://doi.org/10.3389/fgene.2015.00348>
- Evans NT, Olds BP, Renshaw MA, Turner CR, Li Y, Jerde CL, Mahon AR, Pfrender ME, Lamberti GA, Lodge DM (2016) Quantification of mesocosm fish and amphibian species diversity via environmental DNA metabarcoding. *Molecular Ecology Resources* 16(1): 29- 41. <https://doi.org/10.1111/1755-0998.12433>
- Faber JE, Stepien CA (1998) Tandemly repeated sequences in the mitochondrial DNA control region and phylogeography of the pike-perches *Stizostedion*. *Molecular Phylogenetics and Evolution* 10(3): 310-322.
<https://doi.org/10.1006/mpev.1998.0530>
- Ficetola GF, Miaud C, Pompanon F, Taberlet P (2008) Species detection using environmental DNA from water samples. *Biology Letters* 4: 423-425.
<https://doi.org/10.1098/rsbl.2008.0118>
- Ficetola GF, Pansu J, Bonin A, Coissac E, Giguët-Covex C, De Barba M, Gielly L, Lopes CM, Boyer F, Pompanon F, Rayé G, Taberlet P (2015) Replication levels, false presences and the estimation of the presence/absence from eDNA metabarcoding data. *Molecular Ecology Resources* 15: 543-556.
<https://doi.org/10.1111/1755-0998.12338>
- Ficetola CF, Taberlet P, Coissac E (2016) How to limit false positives in environmental DNA and metabarcoding?. *Molecular Ecology Resources* 16: 604-607.
<https://doi.org/10.1111/1755-0998.12508>
- Fierst J, Murdock DA (2017) Decontaminating eukaryotic genome assemblies with machine learning. *BMC Bioinformatics* 18: 533. <https://doi.org/10.1186/s12859-017-1941-0>
- Fišer Pečnikar Ž, Buzan EV (2014) 20 years since the introduction of DNA barcoding: from theory to application. *Journal of Applied Genetics* 55: 43-52.
<https://doi.org/10.1007/s13353-013-0180-y>
- Folmer O, Black M, Hoeh W, Lutz R, Vrijenhoek R (1994). DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. *Molecular Marine Biology and Biotechnology* 3(5): 294-299.

- Fonseca VG (2018) Pitfalls in relative abundance estimation using eDNA metabarcoding. *Molecular Ecology Resources* 18: 923-926. <https://doi.org/10.1111/1755-0998.12902>
- Fox GE, Magrum LJ, Balch WE, Wolf RS, Woese CR (1977) Classification of methanogenic bacteria by 16S ribosomal RNA characterization. *Proceedings of the National Academy of Sciences of the United States of America* 74: 4537-4541. <https://doi.org/10.1073/pnas.74.10.4537>
- Galan M, Pagès M, Cosson J-F (2012) Next-generation sequencing for rodent barcoding: Species identification from fresh, degraded and environmental samples. *PLoS ONE* 7(11): e48374. <https://doi.org/10.1371/journal.pone.0048374>
- Gardner PP, Watson RJ, Morgan XC, Draper JL, Finn RD, Morales SE, Stott MB (2019) Identifying accurate metagenome and amplicon software via a meta-analysis of sequence to taxonomy benchmarking studies. *PeerJ* 7: e6160 <https://doi.org/10.7717/peerj.6160>
- Garrido-Sanz L, Senar MÀ, Piñol J (2020) Estimation of the relative abundance of species in artificial mixtures of insects using low-coverage shotgun metagenomics. *Metabarcoding and Metagenomics* 4: e48281. <https://doi.org/10.3897/mbmg.4.48281>
- Garrido-Sanz L, Senar MÀ, Piñol J (2021) Relative species abundance estimation in artificial mixtures of insects using mito-metagenomics and a correction factor for the mitochondrial DNA copy number. *Molecular Ecology Resources* 00: 1-15. <https://doi.org/10.1111/1755-0998.13464>
- Gaston KJ (2000) Global patterns in biodiversity. *Nature* 405: 220-227. <https://doi.org/10.1038/35012228>
- Genome 10K Community of Scientists (2009) Genome 10K: a proposal to obtain whole-genome sequence for 10 000 vertebrate species. *Journal of Heredity* 100(6): 659-674. <https://doi.org/10.1093/jhered/esp086>
- Gibson JF, Shokralla S, Curry C, Baird DJ, Monk WA, King I, Hajibabaei M (2015) Large-scale biomonitoring of remote and threatened ecosystems via high-throughput sequencing. *PLoS ONE* 10(10): e0138432. <https://doi.org/10.1371/journal.pone.0138432>
- GIGA Community of Scientists (2014) The global invertebrate genomics alliance (GIGA): Developing community resources to study diverse invertebrate genomes. *Journal of Heredity* 105(1): 1-18. <https://doi.org/10.1093/jhered/est084>
- Giovannoni SJ, Britschgi TB, Moyer CL, Field KG (1990) Genetic diversity in Sargasso Sea bacterioplankton. *Nature* 345: 60-63. <https://doi.org/10.1038/345060a0>

- Goldberg CS, Sepulveda A, Ray A, Baumgardt J, Waits LP (2013) Environmental DNA as a new method for early detection of New Zealand mudsnails (*Potamopyrgus antipodarum*). *Freshwater Science* 32(3): 792-800. <https://doi.org/10.1899/13-046.1>
- Goldberg CS, Strickler KM, Pilliod DS (2015) Moving environmental DNA methods from concept to practice for monitoring aquatic macroorganisms. *Biological Conservation* 183: 1-3. <https://doi.org/10.1016/j.biocon.2014.11.040>
- Goldberg CS, Turner CR, Deiner K, Klymus KE, Thomsen PF, Murphy MA, Spear SF, McKee A, Oyler-McCance SJ, Cornman RS, Laramie MB, Mahon AR, Lance RF, Pilliod DS, Strickler KM, Waits LP, Fremier AK, Takahara T, Herder JE, Taberlet P (2016) Critical considerations for the application of environmental DNA methods to detect aquatic species. *Methods in Ecology and Evolution* 7(11): 1299-1307. <https://doi.org/10.1111/2041-210X.12595>
- Gómez-Rodríguez C, Crampton-Platt A, Timmermans MJTN, Baselga A, Vogler AP (2015) Validating the power of mitochondrial metagenomics for community ecology and phylogenetics of complex assemblages. *Methods in Ecology and Evolution* 6(8): 883-894. <https://doi.org/10.1111/2041-210X.12376>
- Gonzalez A, Vázquez-Baeza Y, Pettengill JB, Ottesen A, McDonald D, Knight R (2016) Avoiding pandemic fears in the subway and conquering the platypus. *mSystems* 1(3): e00050-16. <https://doi.org/10.1128/mSystems.00050-16>
- Goulson D (2019) The insect apocalypse, and why it matters. *Current Biology* 29(19): R942-R995. <https://doi.org/10.1016/j.cub.2019.06.069>
- Greenwald WW, Klitgord N, Seguritan V, Yooseph S, Venter JC, Garner C, Nelson KE, Li W (2017) Utilization of defined microbial communities enables effective evaluation of meta-genomic assemblies. *BMC Genomics* 18: 296. <https://doi.org/10.1186/s12864-017-3679-5>
- Gregory TR (2020) Animal genome size database. URL: <http://www.genomesize.com>
- Gressitt J, Leech R (1961) Insect habitats in Antarctica. *Polar Record* 10(68): 501-504. <https://doi.org/10.1017/S0032247400051871>
- Grigoriev IV, Nikitin R, Haridas S, Kuo A, Ohm R, Otilar R, Riley R, Salamov A, Zhao X, Korzeniewski F, Smirnova T, Nordberg H, Dubchak I, Shabalov I (2014) MycoCosm portal: gearing up for 1000 fungal genomes. *Nucleic Acids Research* 42(D1): D699-D704. <https://doi.org/10.1093/nar/gkt1183>
- Gueuning M, Ganser D, Blaser S, Albrecht M, Knop E, Praz C, Frey JE (2019) Evaluating next-generation sequencing (NGS) methods for routine monitoring of wild bees:

Metabarcoding, mitogenomics or NGS barcoding. *Molecular Ecology Resources* 19(4): 847-862. <https://doi.org/10.1111/1755-0998.13013>

Gullan PJ, Cranston PS (2014) *The Insects: An outline of entomology*. 5th Edition. Wiley-Blackwell, West Sussex, UK. 624pp.

Haase P, Pauls SU, Schindehütte K, Sundermann A (2010) First audit of macroinvertebrate samples from an EU Water Framework Directive monitoring program: human error greatly lowers precision of assessment results. *Journal of the North American Benthological Society* 29(4): 1279-1291. <https://doi.org/10.1899/09-183.1>

Hajibabaei M, Shokralla S, Zhou S, Singer GAC, Baird DJ (2011) Environmental barcoding: A next generation sequencing approach for biomonitoring applications using river benthos. *PLoS ONE* 6: e17497. <https://doi.org/10.1371/journal.pone.0017497>

Handelsman J, Rondon MR, Brady SF, Clardy J, Goodman RM (1998) Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chemistry & Biology* 5(10): R245-R249. [https://doi.org/10.1016/s1074-5521\(98\)90108-9](https://doi.org/10.1016/s1074-5521(98)90108-9)

Hanson NW, Konwar KM, Hallam SJ (2016) LCA*: an entropy-based measure for taxonomic assignment within assembled metagenomes. *Bioinformatics* 32(23): 3535-3542. <https://doi.org/10.1093/bioinformatics/btw400>

Harbert RS (2018) Algorithms and strategies in short-read shotgun metagenomic reconstruction of plant communities. *Applications in Plant Sciences* 6(3): e1034 <https://doi.org/10.1002/aps3.1034>

Harvey ML, Dadour IR, Gaudieri S (2003) Mitochondrial DNA cytochrome oxidase I gene: potential for distinction between immature stages of some forensically important fly species (Diptera) in western Australia. *Forensic Science International* 131: 134-139. [https://doi.org/10.1016/s0379-0738\(02\)00431-0](https://doi.org/10.1016/s0379-0738(02)00431-0)

Hebert PDN, Cywinska A, Ball SL, deWaard JR (2003a) Biological identifications through DNA barcodes. *Proceedings of the Royal Society B - Biological Sciences* 270(1512): 313-321. <https://doi.org/10.1098/rspb.2002.2218>

Hebert PDN, Penton EH, Burns JM, Janzen DH, Hallwachs W (2004a) Ten species in one: DNA barcoding reveals cryptic diversity in the neotropical skipper butterfly *Astraptes fulgerator*. *Proceedings of the National Academy of Sciences of the United States of America* 101(41): 14812-4817. <https://doi.org/10.1073/pnas.0406166101>

Hebert PDN, Ratnasingham S, de Waard JR (2003b) Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. *Proceedings of*

the Royal Society B - Biological Sciences 270: S96-S99.
<https://doi.org/10.1098/rsbl.2003.0025>

Hebert PDN, Stoeckle MY, Zemplak TS, Francis CM (2004b) Identification of birds through DNA barcodes. *PLoS Biology* 2(10): e312.
<https://doi.org/10.1371/journal.pbio.0020312>

Hleap JS, Littlefair JE, Steinke D, Hebert PDN, Cristescu ME (2021) Assessment of current taxonomic assignment strategies for metabarcoding eukaryotes. *Molecular Ecology Resources*. <https://doi.org/10.1111/1755-0998.13407>

Hobern D, Hebert P (2019) BIOSCAN - Revealing Eukaryote Diversity, Dynamics, and Interactions. *Biodiversity Information Science and Standards* 3: e37333.
<https://doi.org/10.3897/biss.3.37333>

Hollingsworth PM, Graham SW, Little DP (2011) Choosing and using a plant DNA barcode. *PLoS ONE* 6(5): e19254.
<https://doi.org/10.1371/journal.pone.0019254>

Hornung BVH, Zwittink RD, Kuijper EJ (2019) Issues and current standards of controls in microbiome research. *FEMS Microbiology Ecology* 95(5): fiz045.
<https://doi.org/10.1093/femsec/fiz045>

Huson DH, Auch AF, Qi J, Schuster SC (2007) MEGAN analysis of metagenomic data. *Genome Research* 17: 377-386.
<http://www.genome.org/cgi/doi/10.1101/gr.5969107>

Huson DH, Beier S, Flade I, Górska A, El-Hadidi M, Mitra S, Ruscheweyh H-J, Tappu R (2016) MEGAN Community Edition - Interactive exploration and analysis of large-scale microbiome sequencing data. *PLoS Computational Biology* 12(6): e1004957.
<https://doi.org/10.1371/journal.pcbi.1004957>

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

Ji Y, Huotari T, Roslin T, Schmidt NM, Wang J, Yu DW, Ovaskainen O (2020) SPIKEPIPE: A metagenomic pipeline for the accurate quantification of eukaryotic species occurrences and intraspecific abundance change using DNA barcodes or mitogenomes. *Molecular Ecology Resources* 20: 256-267.
<https://doi.org/10.1111/1755-0998.13057>

Kassambara A (2018) ggpubr: 'ggplot2' Based Publication Ready Plots. R package version 0.2. URL: <https://CRAN.R-project.org/package=ggpubr>

- Khamesipour F, Lankarani KB, Honarvar B, Kwenti TE (2018) A systematic review of human pathogens carried by the housefly (*Musca domestica* L.). *BMC Public Health* 18(1): 1049. <https://doi.org/10.1186/s12889-018-5934-3>
- Kingsolver JG, Woods HA, Buckley LB, Potter KA, MacLean HJ, Higgins JK (2011) Complex life cycles and the responses of Insects to climate change. *Integrative and Comparative Biology* 51(5): 719-732. <https://doi.org/10.1093/icb/icr015>
- Korshunova T, Picton B, Furfaro G, Mariottini P, Pontes M, Prkić J, Fletcher K, Malmberg K, Lundin K, Martynov A (2019) Multilevel fine-scale diversity challenges the 'cryptic species' concept. *Scientific Reports* 9: 6732. <https://doi.org/10.1038/s41598-019-42297-5>
- Kraaijeveld K, de Weger LA, Ventayol García M, Buermans H, Frank J, Hiemstra PS, den Dunnen JT (2015) Efficient and sensitive identification and quantification of airborne pollen using next-generation DNA sequencing. *Molecular Ecology Resources* 15(1): 8-16. <https://doi.org/10.1111/1755-0998.12288>
- Krehenwinkel H, Wolf M, Lim JY, Rominger AJ, Simison WB, Gillespie RG (2017) Estimating and mitigating amplification bias in qualitative and quantitative arthropod metabarcoding. *Scientific Reports* 7: 17668. <https://doi.org/10.1038/s41598-017-17333-x>
- Kress WJ, García-Robledo C, Uriarte M, Erickson DL (2015) DNA barcodes for ecology, evolution, and conservation. *Trends in Ecology & Evolution* 30(1): 25-35. <http://dx.doi.org/10.1016/j.tree.2014.10.008>
- Kuczynski J, Lauber CL, Walters WA, Wegener Parfrey L, Clemente JC, Gevers D, Knight R (2012) Experimental and analytical tools for studying the human microbiome. *Nature Reviews Genetics* 13: 47-58. <https://doi.org/10.1038/nrg3129>
- Kulkarni SS, Dosedall LM, Willenborg CJ (2015) The role of ground beetles (Coleoptera: Carabidae) in weed seed consumption: a review. *Weed Science* 63: 355-376. <https://doi.org/10.1614/WS-D-14-00067.1>
- Kunin V, Engelbrektsen A, Ochman H, Hugenholtz P (2010) Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. *Environmental Microbiology* 12(1): 118-123. <https://doi.org/10.1111/j.1462-2920.2009.02051.x>
- Kwong S, Srivathsan A, Meier R (2012) An update on DNA barcoding: low species coverage and numerous unidentified sequences. *Cladistics* 28: 639-644. <https://doi.org/10.1111/j.1096-0031.2012.00408.x>
- Lamb PD, Hunter E, Pinnegar JK, Creer S, Davies RG, Taylor MI (2018) How quantitative is metabarcoding: A meta-analytical approach. *Molecular Ecology* 28(2): 420-430. <https://doi.org/10.1111/mec.14920>

- Lang D, Tang M, Hu J, Zhou X (2019). Genome-skimming provides accurate quantification for pollen mixtures. *Molecular Ecology Resources* 19(6): 1433-1446. <https://doi.org/10.1111/1755-0998.13061>
- Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nature Methods* 9: 357-359. <https://doi.org/10.1038/nmeth.1923>
- Lebonah DE, Dileep A, Chandrasekhar K, Sreevani S, Sreedevi B, Kumari JP (2014) DNA barcoding on bacteria: A review. *Advances in Biology* 2014: 541787. <https://doi.org/10.1155/2014/541787>
- Lebuhn G, Droege S, Connor EF, Gemmill-Herren B, Potts SG, Minckley RL, Griswold T, Jean R, Kula E, Roubik DW, Cane J, Wright KW, Frankie G, Parker F (2013) Detecting insect pollinator declines on regional and global scales. *Conservation Biology* 27: 113-120. <https://doi.org/10.1111/j.1523-1739.2012.01962.x>
- Lee TR, Alemseged Y, Mitchell A (2018) Dropping Hints: Estimating the diets of livestock in rangelands using DNA metabarcoding of faeces. *Metabarcoding and Metagenomics* 2: 1-17. <https://doi.org/10.3897/mbmg.2.22467>
- Leray M, Knowlton N (2017) Random sampling causes the low reproducibility of rare eukaryotic OTUs in Illumina COI metabarcoding. *PeerJ* 5: e3006. <https://doi.org/10.7717/peerj.3006>
- Levine R (2011) i5k: the 5,000 insect genome project. *American Entomologist* 57(2): 110-113. <https://doi.org/10.1093/ae/57.2.110>
- Lewin HA, Robinson GE, Kress WJ, Baker WJ, Coddington J, Crandall KA, Durbin R, Edwards SV, Forest F, Gilbert MTP, Goldstein MM, Grigoriev IV, Hackett KJ, Haussler D, Jarvis ED, Johnson WE, Patrinos A, Richards S, Castilla-Rubio JC, van Sluys M-A, Soltis PS, Xu X, Yang H, Zhang G (2018) Earth BioGenome Project: Sequencing life for the future of life. *Proceedings of the National Academy of Sciences of the United States of America* 115(17): 4325-4333. <https://doi.org/10.1073/pnas.1720115115>
- Li H (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv* 1303.3997v1 [q-bio.GN].
- Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25(14): 1754-1760. <https://doi.org/10.1093/bioinformatics/btp324>
- Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup (2009) The sequence alignment/map format and SAMtools. *Bioinformatics* 25(16): 2078-2079. <https://doi.org/10.1093/bioinformatics/btp352>

- Linard B, Crampton-Platt A, Cillett CPDT, Timmermans MJTN, Vogler AP (2015) Metagenome skimming of insect specimen pools: potential for comparative genomics. *Genome Biology and Evolution* 7(6): 1474-1489. <https://doi.org/10.1093/gbe/evv086>
- Lindgreen S, Adair K, Gardner P (2016) An evaluation of the accuracy and speed of metagenome analysis tools. *Scientific Reports* 6: 19233 <https://doi.org/10.1038/srep19233>
- Little DP, Stevenson DWm (2007) A comparison of algorithms for the identification of specimens using DNA barcodes: Examples from gymnosperms. *Cladistics* 23: 1-21. <https://doi.org/10.1111/j.1096-0031.2006.00126.x>
- Liu R, Jin L, Long K, Tang Q, Ma J, Wang X, Zhu L, Jiang A, Tang G, Jiang Y, Li X, Li M (2018) Analysis of mitochondrial DNA sequence and copy number variation across five high-altitude species and their low-altitude relatives. *Mitochondrial DNA Part B* 3(2): 847-851. <https://doi.org/10.1080/23802359.2018.1501285>
- Liu S, Li Y, Lu J, Su X, Tang M, Zhang R, Zhou L, Zhou C, Yang Q, Ji Y (2013) SOAPBarcode: revealing arthropod biodiversity through assembly of Illumina shotgun sequences of PCR amplicons. *Methods in Ecology and Evolution* 4: 1142-1150. <https://doi.org/10.1111/2041-210X.12120>
- Liu S, Wang X, Xie L, Tan M, Li Z, Su X, Zhang H, Misof B, Kjer KM, Tang M, Niehuis O, Jiang H, Zhou X (2016) Mitochondrial capture enriches mito-DNA 100 fold, enabling PCR-free mitogenomics biodiversity analysis. *Molecular Ecology Resources* 16: 470-479. <https://doi.org/10.1111/1755-0998.12472>
- Lu J, Salzberg SL (2018) Removing contaminants from metagenomic databases. *PLoS Computational Biology* 14(6): e1006277. <https://doi.org/10.1371/journal.pcbi.1006277>
- Macher J-N, Zizka VMA, Weigand AM, Leese F (2018) A simple centrifugation protocol for metagenomic studies increases mitochondrial DNA yield by two orders of magnitude. *Methods in Ecology and Evolution* 9: 1070-1074. <https://doi.org/10.1111/2041-210X.12937>
- Magurran AE (2004) Measuring biological diversity. *The Journal of the Torrey Botanical Society* 131(3): 277-278. <https://doi.org/10.2307/4126959>
- Mande SS, Mohammed MH, Ghosh TS (2012) Classification of metagenomic sequences: methods and challenges. *Briefings in Bioinformatics* 13(6): 669-681 <https://doi.org/10.1093/bib/bbs054>
- Marcelino VR, Clausen PTLC, Buchmann JP, Wille M, Iredell JR, Meyer W, Lund O, Sorrell TC, Holmes EC (2020a) CCMetagen: comprehensive and accurate

identification of eukaryotes and prokaryotes in metagenomic data. *Genome Biology* 21: 103. <https://doi.org/10.1186/s13059-020-02014-2>

- Marcelino VR, Holmes EC, Sorrell TC (2020b) The use of taxon-specific reference databases compromises metagenomic classification. *BMC Genomics* 21: 184. <https://doi.org/10.1186/s12864-020-6592-2>
- Marco-Sola S, Sammeth M, Guigó R, Ribeca P (2012) The GEM mapper: fast, accurate and versatile alignment by filtration. *Nature Methods* 9: 1185-1188. <https://doi.org/10.1038/nmeth.2221>
- Mardis ER (2008) Next-generation DNA sequencing methods. *The Annual Review of Genomics and Human Genetics* 9: 387-402. <https://doi.org/10.1146/annurev.genom.9.081307.164359>
- Martinez ND, Hawkins BA, Dawah HA, Feifarek BP (1999) Effects of sampling effort on characterization of food-web structure. *Ecology* 80(3): 1044-1055. [https://doi.org/10.1890/0012-9658\(1999\)080\[1044:EOSEOC\]2.0.CO;2](https://doi.org/10.1890/0012-9658(1999)080[1044:EOSEOC]2.0.CO;2)
- Martos MC (2020) *In silico* assessment of metagenomic-based method for quantitative identification of species. Master Thesis, Universitat Autònoma de Barcelona, Cerdanyola del Vallès, Spain.
- Matesanz S, Pescador DS, Pías B, Sánchez AM, Chacón-Labela J, Illuminati A, de la Cruz M, López-Angulo J, Marí-Mena N, Vizcaíno A, Escudero A (2019) Estimating belowground plant abundance with DNA metabarcoding. *Molecular Ecology Resources* 19(5): 1265-1277. <https://doi.org/10.1111/1755-0998.13049>
- May RE (1988) How many species are there on Earth?. *Science* 241(4872): 1441-1449. <https://doi.org/10.1126/science.241.4872.1441>
- McArdle AJ, Kaforou M (2020) Sensitivity of shotgun metagenomics to host DNA: abundance estimates depend on bioinformatic tools and contamination is the main issue. *Access Microbiology* 2(4): acmi000104. <https://doi.org/10.1099/acmi.0.000104>
- McCool M, Reinders J, Robison A (2012) Structured Parallel Programming: Patterns for Efficient Computation. Morgan Kaufmann Publishers Inc., USA, 432 pp.
- McInnes JC, Alderman R, Deagle BE, Lea M-A, Raymond B, Jarman SN (2016) Optimised scat collection protocols for dietary DNA metabarcoding in vertebrates. *Methods in Ecology and Evolution* 8(2): 192-202. <https://doi.org/10.1111/2041-210X.12677>
- McIntyre ABR, Ounit R, Afshinnekoo E, Prill RJ, Hénaff E, Alexander N, Minot SS, Danko D, Foox J, Ahsanuddin S, Tighe S, Hasan NA, Subramanian P, Moffat K, Levy S, Lonardi S, Greenfield N, Colwell RR, Rosen GL, Mason, Christopher E (2017)

Comprehensive benchmarking and ensemble approaches for metagenomic classifiers. *Genome Biology* 18: 182. <https://doi.org/10.1186/s13059-017-1299-7>

Meier R, Shiyang K, Vaidya G, Ng PKL (2006) DNA barcoding and taxonomy in diptera: A tale of high intraspecific variability and low identification success. *Systematic Biology* 55(5): 715-728. <https://doi.org/10.1080/10635150600969864>

Merchant S, Wood DE, Salzberg SL (2014) Unexpected cross-species contamination in genome sequencing projects. *PeerJ* 2: e675. <https://doi.org/10.7717/peerj.675>

Metwally AA, Dai Y, Finn PW, Perkins DL (2016) WEVOTE: Weighted voting taxonomic identification method of microbial sequences. *PLoS ONE* 11(9): e0163527 <https://doi.org/10.1371/journal.pone.0163527>

Meyer F, Bremges A, Belmann P, Janssen S, McHardy AC, Koslicki D (2019) Assessing taxonomic metagenome profilers with OPAL. *Genome Biology* 20: 51. <https://doi.org/10.1186/s13059-019-1646-y>

Mora C, Tittensor DP, Adl S, Simpson AGB, Worm B (2011) How many species are there on Earth and in the ocean?. *PLoS Biology* 9: e1001127. <https://doi.org/10.1371/journal.pbio.1001127>

Morgulis A, Coulouris G, Raytselis Y, Madden TL, Agarwala R, Schäffer AA (2008) Database indexing for production MegaBLAST searches. *Bioinformatics* 24(16): 1757-1764. <https://doi.org/10.1093/bioinformatics/btn322>

Morgulis A, Gertz EM, Schäffer AA, Agarwala R (2006) A fast and symmetric DUST implementation to mask low-complexity DNA sequences. *Journal of Computational Biology* 13(5): 1028-1040. <https://doi.org/10.1089/cmb.2006.13.1028>

Noss RF (1990) Indicators for monitoring biodiversity: A hierarchical approach. *Conservation Biology* 4: 355-364. <https://doi.org/10.1111/j.1523-1739.1990.tb00309.x>

Not F, del Campo J, Balagué V, de Vargas C, Massana R (2009) New Insights into the Diversity of Marine Picoeukaryotes. *PLoS ONE* 4(9): e7143. <https://doi.org/10.1371/journal.pone.0007143>

Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, McGlinn D, Minchin PR, O'Hara RB, Simpson GL, Solymos P, Stevens MHH, Szoecs E, Wagner H (2018) Vegan: Community Ecology Package. R package version 2.5-4. URL: <https://CRAN.R-project.org/package=vegan>

Oliveira MC, Repetti SI, Iha C, Jackson CJ, Díaz-Tapia P, Lubiana KMF, Cassano V, Costa JF, Cremen MaCM, Marcelino VR, Verbruggen H (2018) High-throughput

sequencing for algal systematics. *European Journal of Phycology* 53(3): 256-272. <https://doi.org/10.1080/09670262.2018.1441446>

- Ounit R, Wanamaker S, Close TJ, Lonardi S (2015) CLARK: Fast and accurate classification of metagenomic and genomic sequences using discriminative *k*-mers. *BMC Genomics* 16: 236. <https://doi.org/10.1186/s12864-015-1419-2>
- Papadopoulou A, Taberlet P, Zinger L (2015) Metagenome skimming for phylogenetic community ecology: A new era in biodiversity research. *Molecular Ecology* 24(14): 3515-3517. <https://doi.org/10.1111/mec.13263>
- Parducci L, Alsos IG, Unneberg P, Pedersen MW, Han LU, Lammers Y, Salonen JS, Väiliranta MM, Slotte T, Wohlfarth B (2019) Shotgun environmental DNA, pollen, and macrofossil analysis of Lateglacial lake sediments from Southern Sweden. *Frontiers in Ecology and Evolution* 7: 189. <https://doi.org/10.3389/fevo.2019.00189>
- Paris M, Boyer R, Jaenichen R, Wolf J, Karageorgi M, Green J, Cagnon M, Parinello H, Estoup A, Gautier M, Gompel N, Prud'homme B (2020) Near-chromosome level genome assembly of the fruit pest *Drosophila suzukii* using long-read sequencing. *Scientific Reports* 10: 11227. <https://doi.org/10.1038/s41598-020-67373-z>
- Paudel YP, Mackereth R, Hanley R, Qin W (2015) Honey bees (*Apis mellifera* L.) and pollination issues: Current status, impacts, and potential drivers of decline. *Journal of Agricultural Science* 7(6): 93-109. <https://doi.org/10.5539/jas.v7n6p93>
- Paula DP, Linard B, Andow DA, Sujii ER, Pires CS, Vogler AP (2015) Detection and decay rates of prey and prey symbionts in the gut of a predator through metagenomics. *Molecular Ecology Resources* 15: 880-892. <https://doi.org/10.1111/1755-0998.1236>
- Paula DP, Timbó RV, Togawa RC, Vogler AP, Andow DA (2021) Quantitative prey species detection in predator guts across multiple trophic levels by DNA shotgun sequencing. *bioRxiv*. <https://doi.org/10.1101/2021.04.01.438119>
- Peabody MA, Van Rossum T, Lo R, Brinkman FSL (2015) Evaluation of shotgun metagenomics sequence classification methods using *in silico* and *in vitro* simulated communities. *BMC Bioinformatics* 16: 362 <https://doi.org/10.1186/s12859-015-0788-5>
- Pearman WS, Freed NE, Silander OK (2020) Testing the advantages and disadvantages of short- and long- read eukaryotic metagenomics using simulated reads. *BMC Bioinformatics* 21: 220. <https://doi.org/10.1186/s12859-020-3528-4>
- Peel N, Dicks LV, Clark MD, Heavens D, Percival-Alwyn L, Cooper C, Davies RG, Leggett RM, Yu DW (2019). Semi-quantitative characterisation of mixed pollen samples using MinION sequencing and Reverse Metagenomics (RevMet). *Methods in*

Ecology and Evolution 10(10): 1690-1701. <https://doi.org/10.1111/2041-210X.13265>

Pennisi E (2019) \$180 million DNA 'barcode' project aims to discover 2 million new species. *Science* 364: 920-921. <https://doi.org/10.1126/science.aay2877>

Pereira MB, Wallroth M, Jonsson V, Kristiansson E (2018) Comparison of normalization methods for the analysis of metagenomic gene abundance data. *BMC Genomics* 19: 274. <https://doi.org/10.1186/s12864-018-4637-6>

Pereira-Marques J, Hout A, Ferreira RM, Weber M, Pinto-Ribeiro I, van Doorn L-J, Knetsch CW, Figueiredo C (2019) Impact of host DNA and sequencing depth on the taxonomic resolution of whole metagenome sequencing for microbiome analysis. *Frontiers in Microbiology* 10: 1277. <https://doi.org/10.3389/fmicb.2019.01277>

Piñol J (2021) Genotype by sequencing: An alternative new method to amplicon metabarcoding and shotgun metagenomics for the assessment of eukaryote biodiversity. *Molecular Ecology Resources* 21(4): 1001-1004. <https://doi.org/10.1111/1755-0998.13320>

Piñol J, Mir G, Gomez-Polo P, Agustí N (2015) Universal and blocking primer mismatches limit the use of high-throughput DNA sequencing for the quantitative metabarcoding of arthropods. *Molecular Ecology Resources* 15(4): 819-830. <https://doi.org/10.1111/1755-0998.12355>

Piñol J, San Andrés V, Clare EL, Mir G, Symondson WOC (2014) A pragmatic approach to the analysis of diets of generalist predators: the use of next-generation sequencing with no blocking probes. *Molecular Ecology Resources* 14(1): 18-26. <https://doi.org/10.1111/1755-0998.12156>

Piñol J, Senar MA, Symondson WOC (2019) The choice of universal primers and the characteristics of the species mixture determines when DNA metabarcoding can be quantitative. *Molecular Ecology* 28(2): 407-419. <https://doi.org/10.1111/mec.14776>

Piro V, Matschkowski M, Renard B (2017) MetaMeta: integrating metagenome analysis tools to improve taxonomic profiling. *Microbiome* 5: 101 <https://doi.org/10.1186/s40168-017-0318-y>

Pompanon F, Deagle BE, Symondson WOC, Brown DS, Jarman SN, Taberlet P (2012) Who is eating what: Diet assessment using next generation sequencing. *Molecular Ecology* 21(8): 1931-1950. <https://doi.org/10.1111/j.1365-294X.2011.05403.x>

Pornon A, Escaravage N, Burrus M, Holota H, Khimoun A, Mariette J, Pellizzari C, Iribar A, Etienne R, Taberlet P, Vidal M, Winterton P, Zinger L, Andalo C (2016) Using

metabarcoding to reveal and quantify plant-pollinator interactions. *Scientific Reports* 6: 27282. <https://doi.org/10.1038/srep27282>

- Porter TM, Hajibabaei M (2017) Scaling up: A guide to high-throughput genomic approaches for biodiversity analysis. *Molecular Ecology* 27: 313-338. <https://doi.org/10.1111/mec.14478>
- Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research* 41: D590-D596. <https://dx.doi.org/10.1093%2Fnar%2Fgks1219>
- R Core Team (2016) R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria. URL: <https://www.R-project.org/>
- Ranjan R, Rani A, Metwally A, McGee HS, Perkins DL (2016) Analysis of the microbiome: advantages of whole genome shotgun versus 16S amplicon sequencing. *Biochemical and Biophysical Research Communications* 469: 967-977. <https://dx.doi.org/10.1016%2Fj.bbrc.2015.12.083>
- Ratnasingham S, Hebert PD (2007) BOLD: The Barcode of Life Data System (<http://www.barcodinglife.org>). *Molecular Ecology notes* 7(3): 355-364. <https://doi.org/10.1111/j.1471-8286.2007.01678.x>
- Ratnasingham S, Hebert PDN (2013) A DNA-Based Registry for All Animal Species: The Barcode Index Number (BIN) System. *PLoS ONE* 8(7): e66213. <https://doi.org/10.1371/journal.pone.0066213>
- Reilly E (2003) The contribution of insect remains to an understanding of the environment of Viking-age and medieval Dublin. In: Duffy S (Eds) *Medieval Dublin IV*. Four Courts Press, Dublin, 40-62.
- Revelle W (2021). psych: Procedures for Psychological, Psychometric, and Personality Research. R package version 2.1.6. URL: <https://CRAN.R-project.org/package=psych>
- Robinson GE, Hackett KJ, Purcell-Miramontes M, Brown SJ, Evans JD, Goldsmith MR, Lawson D, Okamuro J, Robertson HM, Schneider DJ (2011) Creating a buzz about insect genomes. *Science* 331(6023): 1386. <https://doi.org/10.1126/science.331.6023.1386>
- Rodríguez-Martínez R, Leonard G, Milner DS, Sudek S, Conway M, Moore K, Hudson T, Mahé F, Keeling PJ, Santoro AE, Worden AZ, Richards TA (2020) Controlled sampling of ribosomally active protistan diversity in sediment-surface layers identifies putative players in the marine carbon sink. *The ISME Journal* 14(4): 984-998. <https://doi.org/10.1038/s41396-019-0581-y>

- Ross MG, Russ C, Costello M, Hollinger A, Lennon NJ, Hegarty R, Nusbaum C, Jaffe DB. (2013) Characterizing and measuring bias in sequence data. *Genome Biology* 14: R51. <https://doi.org/10.1186/gb-2013-14-5-r51>
- RStudio Team (2015) RStudio: Integrated Development for R. RStudio, Inc, Boston, MA. URL: <http://www.rstudio.com/>
- Sabrosky CW (1952) How many insects are there?. *Systematic Biology* 2(1):31-36. <https://doi.org/10.2307/2411567>
- Saitoh S, Aoyama H, Fujii S, Sunagawa H, Nagahama H, Akutsu M, Shinzato N, Kaneko N, Nakamori T (2016) A quantitative protocol for DNA metabarcoding of springtails (Collembola). *Genome* 59(9): 705-723. <https://doi.org/10.1139/gen-2015-0228>
- Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America* 74(12): 5463-5467. <https://doi.org/10.1073/pnas.74.12.5463>
- Sarmashghi S, Bohmann K, Gilbert MTP, Bafna V, Mirarab S (2019). Skmer: Assembly-free and alignment-free sample identification using genome skims. *Genome Biology* 20: 34. <https://doi.org/10.1186/s13059-019-1632-4>
- Sato MP, Ogura Y, Nakamura K, Nishida R, Gotoh Y, Hayashi M, Hisatsune J, Sugai M, Takehiko I, Hayashi T. (2019) Comparison of the sequencing bias of currently available library preparation kits for Illumina sequencing of bacterial genomes and metagenomes. *DNA Research* 26(5): 391-398. <https://doi.org/10.1093/dnares/dsz017>
- Schnell IB, Bohmann K, Gilbert MTP (2015) Tag jumps illuminated - reducing sequence-to-sample misidentifications in metabarcoding studies. *Molecular Ecology Resources* 15(6): 1289-1303. <https://doi.org/10.1111/1755-0998.12402>
- Schoch CL, Seifert KA, Huhndorf S, Robert V, Spouge JL, Levesque CA, Chen W, Fungal Barcoding Consortium (2012). Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proceedings of the National Academy of Sciences of the United States of America* 109(16): 6241-6246. <https://doi.org/10.1073/pnas.1117018109>
- Sgamma T, Lockie-Williams C, Kreuzer M, Williams S, Scheyhing U, Koch E, Slater A, Howard C (2017) DNA barcoding for industrial quality assurance. *Planta Medica* 83(14/15): 1117-1129. <https://doi.org/10.1055/s-0043-113448>
- Shearer TL, Coffroth MA (2008) DNA BARCODING: Barcoding corals: limited by interspecific divergence, not intraspecific variation. *Molecular Ecology Resources* 8(2): 247-255. <https://doi.org/10.1111/j.1471-8286.2007.01996.x>

- Shen YY, Chen X, Murphy RW (2013) Assessing DNA barcoding as a tool for species identification and data quality control. *PLoS ONE* 8(2): e57125. <https://doi.org/10.1371/journal.pone.0057125>
- Shokralla S, Porter T, Gibson J, Dobosz R, Janzen DH, Hallwachs W, Brian Golding G, Hajibabaei M (2015) Massively parallel multiplex DNA sequencing for specimen identification using an Illumina MiSeq platform. *Scientific Reports* 5: 9687. <https://doi.org/10.1038/srep09687>
- Siegwald L, Touzet H, Lemoine Y, Hot D, Audebert C, Caboche S (2017) Assessment of common and emerging bioinformatics pipelines for targeted metagenomics. *PLoS ONE* 12: e0169563. <https://doi.org/10.1371/journal.pone.0169563>
- Simhadri RK, Fast EM, Guo R, Schultz MJ, Vaisman N, Ortiz L, Bybee J, Slatko BE, Frydman HM (2017) The gut commensal microbiome of *Drosophila melanogaster* is modified by the endosymbiont *Wolbachia*. *mSphere* 2(5): e00287-17. <https://doi.org/10.1128/mSphere.00287-17>
- Singer GAC, Shekarriz S, McCarthy A, Fahner N, Hajibabaei M (2020) The utility of a metagenomics approach for marine biomonitoring. *bioRxiv*. <https://doi.org/10.1101/2020.03.16.993667>
- Singh B, Crippen TL, Zheng L, Fields AT, Yu Z, Ma Q, Wood TK, Dowd SE, Flores M, Tomberlin JK, Tarone AM (2015) A metagenomic assessment of the bacteria associated with *Lucilia sericata* and *Lucilia cuprina* (Diptera: Calliphoridae). *Applied Microbiology and Biotechnology* 99(2): 869-883. <https://doi.org/10.1007/s00253-014-6115-7>
- Srivathsan A, Ang A, Vogler AP, Meier R (2016) Fecal metagenomics for the simultaneous assessment of diet, parasites, and population genetics of an understudied primate. *Frontiers in Zoology* 13: 17. <https://doi.org/10.1186/s12983-016-0150-4>
- Srivathsan A, Sha JCM, Vogler AP, Meier R (2015) Comparing the effectiveness of metagenomics and metabarcoding for diet analysis of a leaf-feeding monkey (*Pygathrix nemaeus*). *Molecular Ecology Resources* 15(2): 250-261. <https://doi.org/10.1111/1755-0998.12302>
- Stein ED, Martinez MC, Stiles S, Miller PE, Zakharov EV (2014) Is DNA barcoding actually cheaper and faster than traditional morphological methods: Results from a survey of freshwater bioassessment efforts in the United States?. *PLoS ONE* 9(4): e95525. <https://doi.org/10.1371/journal.pone.0095525>
- Stein JL, Marsh TL, Wu KY, Shizuya H, DeLong EF (1996) Characterization of uncultivated prokaryotes: isolation and analysis of a 40-kilobase-pair genome

fragment from a planktonic marine archaeon. *Journal of Bacteriology* 178(3): 591-599. <https://doi.org/10.1128/jb.178.3.591-599.1996>

Stork NE (2018) How many species of insects and other terrestrial arthropods are there on earth?. *Annual Review of Entomology* 63: 31-45. <https://doi.org/10.1146/annurev-ento-020117-043348>

Sun Z, Huang S, Zhang M, Zhu Q, Haiminen N, Carrieri AP, Vázquez-Baeza Y, Parida L, Kim H-C, Knight R, Liu Y-Y (2021) Challenges in benchmarking metagenomic profilers. *Nature Methods* 18: 618-626. <https://doi.org/10.1038/s41592-021-01141-3>

Sweeney BW, Battle JM, Jackson JK, Dapkey T (2011) Can DNA barcodes of stream macroinvertebrates improve descriptions of community structure and water quality?. *Journal of the North American Benthological Society* 30(1): 195-216. <https://doi.org/10.1899/10-016.1>

Taberlet P, Bonin A, Zinger L, Coissac E (2018) Environmental DNA: For biodiversity research and monitoring. Oxford University Press, UK, 268pp. <https://doi.org/10.1093/oso/9780198767220.001.0001>

Taberlet P, Coissac E, Hajibabaei M, Riesenberger LH (2012a) Environmental DNA. *Molecular Ecology* 21: 1789-1793. <https://doi.org/10.1111/j.1365-294X.2012.05542.x>

Taberlet P, Coissac E, Pompanon F, Brochmann C, Willerslev E (2012b) Towards next-generation biodiversity assessment using DNA metabarcoding. *Molecular Ecology* 21(8): 2045-2050. <https://doi.org/10.1111/j.1365-294X.2012.05470.x>

Tang M, Hardman CJ, Ji Y, Meng G, Liu S, Tan M, Yang S, Moss ED, Wang J, Yang C, Bruce C, Nevard T, Potts SG, Zhou X, Yu DW (2015) High-throughput monitoring of wild bee diversity and abundance via mitogenomics. *Methods in Ecology and Evolution* 6: 1034-1043. <https://doi.org/10.1111/2041-210X.12416>

Tang M, Tan M, Meng G, Yang S, Su X, Liu S, Song W, Li Y, Wu Q, Zhang A, Zhou X (2014) Multiplex sequencing of pooled mitochondrial genomes-a crucial step toward biodiversity analysis using mito-metagenomics. *Nucleic Acids Research* 42(22): e166. <https://doi.org/10.1093/nar/gku917>

Teeling H, Glöckner FO (2012) Current opportunities and challenges in microbial metagenome analysis-a bioinformatic perspective. *Briefings in Bioinformatics* 13(6): 728-742 <https://doi.org/10.1093/bib/bbs039>

Thomas AC, Deagle BE, Everson JP, Harsch CH, Trites AW (2016) Quantitative DNA metabarcoding: improved estimates of species proportional biomass using correction factors derived from control material. *Molecular Ecology Resources* 16(3): 714-726. <https://doi.org/10.1111/1755-0998.12490>

- Thomas T, Gilbert J, Meyer F (2012) Metagenomics - a guide from sampling to analysis. *Microbial Informatics and Experimentation* 2: 3. <https://doi.org/10.1186/2042-5783-2-3>
- Thomsen PF, Willerslev E (2015) Environmental DNA - An emerging tool in conservation for monitoring past and present biodiversity. *Biological Conservation* 183: 4-18. <https://doi.org/10.1016/j.biocon.2014.11.019>
- Thyagarajan B, Wang R, Barcelo H, Koh W-P, Yuan J-M (2012) Mitochondrial copy number is associated with colorectal cancer risk. *Cancer Epidemiology, Biomarkers & Prevention* 21(9): 1574-1581. <https://doi.org/10.1158/1055-9965.EPI-12-0138-T>
- Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, Solovyev VV, Rubin EM, Rokhsar DS, Banfield JF (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* 428(6978): 37-43. <https://www.nature.com/articles/nature02340>
- Valentini A, Pompanon F, Taberlet P (2009) DNA barcoding for ecologists. *Trends in Ecology and Evolution* 24(2): 110-117. <https://doi.org/10.1016/j.tree.2008.09.011>
- Valentini A, Taberlet P, Miaud C, Civade R, Herder J, Thomsen PF, Bellemain E, Besnard A, Coissac E, Boyer F, Gaboriaud C, Jean P, Poulet N, Roset N, Copp GH, Geniez P, Pont D, Argillier C, Baudoin J-M, Peroux T, Crivelli AJ, Olivier A, Acqueberge M, Le Brun M, Møller PR, Willerslev E, Gaboriaud C (2016). Next-generation monitoring of aquatic biodiversity using environmental DNA metabarcoding. *Molecular Ecology* 25(4): 929-942. <https://doi.org/10.1111/mec.13428>
- Velsko IM, Frantz LAF, Herbig A, Larson G, Warinner C (2018) Selection of appropriate metagenome taxonomic classifiers for ancient microbiome research. *mSystems* 3(4): e00080-18. <https://doi.org/10.1128/mSystems.00080-18>
- Venter JC, Remington K, Heidelberg JF, Aaron L, Halpern, Doug Rusch, Eisen JA, Wu D, Paulsen I, Nelson KE, Nelson W, Fouts DE, Levy S, Knap AH, Lomas MW, Nealson K, White O, Peterson J, Hoffman J, Parsons R, Baden-Tillson H, Pfannkoch C, Rogers Y-H, Smith HO (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science* 304(5667): 66-74. <https://doi.org/10.1126/science.1093857>
- Wagemaker CAM, Mommer L, Visser JW, Weigelt A, van Gorp TP, Postuma M, Smit-Tiekstra AE, de Kroon H (2021) msGBS: A new high-throughput approach to quantify the relative species abundance in root samples of multispecies plant communities. *Molecular Ecology Resources* 21: 1021-1036. <https://doi.org/10.1111/1755-0998.13278>

- Walker BH (1992) Biodiversity and ecological redundancy. *Conservation Biology* 6(1): 18-23.
- Walsh AM, Crispie F, O'Sullivan O, Finnegan L, Claesson MJ, Cotter PD (2018) Species classifier choice is a key consideration when analysing low-complexity food microbiome data. *Microbiome* 6: 50. <https://doi.org/10.1186/s40168-018-0437-0>
- Ward D, Weller R, Bateson M (1990) 16S rRNA sequences reveal numerous uncultured microorganisms in a natural community. *Nature* 345: 63-65. <https://doi.org/10.1038/345063a0>
- Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Edgar R, Federhen S, Geer LY, Kapustin Y, Khovayko O, Landsman D, Lipman DJ, Madden TL, Maglott DR, Ostell J, Miller V, Pruitt KD, Schuler GD, Sequeira E, Sherry ST, Sirotkin K, Souvorov A, Starchenko G, Tatusov RL, Tatusova TA, Wagner L, Yaschenko E (2007) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research* 35(1): D5-D12. <https://doi.org/10.1093/nar/gkl1031>
- Wickham H (2016) ggplot2: Elegant graphics for data analysis. Springer-Verlag New York (New York): 1-166. URL: <http://ggplot2.org>.
- Wilcox TM, Zarn KE, Piggott MP, Young MK, McKelvey KS, Schwartz MK (2018) Capture enrichment of aquatic environmental DNA: A first proof of concept. *Molecular Ecology Resources* 18: 1392-1401. <https://doi.org/10.1111/1755-0998.12928>
- Williams RS (2017) What's next? [The end of Moore's law]. *Computing in Science & Engineering* 19(2): 7-13. <https://doi.org/10.1109/MCSE.2017.31>
- Wilson EO (1985) The global biodiversity crisis: a challenge to science. *Issues in Science & Technology* 2: 20-29.
- Woese CR, Fox GE (1977) Phylogenetic structure of the prokaryotic domain: The primary kingdoms. *Proceedings of the National Academy of Sciences of the United States of America* 74(11): 5088-5090. <https://doi.org/10.1073/pnas.74.11.5088>
- Wolff JN, Shearman DCA, Brooks RC, Ballard JWO (2012) Selective enrichment and sequencing of whole mitochondrial genomes in the presence of nuclear encoded mitochondrial pseudogenes (Numts). *PLoS ONE* 7(5): e37142. <https://doi.org/10.1371/journal.pone.0037142>
- Wood DE, Lu J, Langmead B (2019) Improved metagenomic analysis with Kraken 2. *Genome Biology* 20: 257. <https://doi.org/10.1186/s13059-019-1891-0>

- Wood DE, Salzberg SL (2014) Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology* 15: R46. <https://doi.org/10.1186/gb-2014-15-3-r46>
- Yang C-H, Pospisilik JA (2019) Polyphenism - A window into gene-environment interactions and phenotypic plasticity. *Frontiers in Genetics* 10: 132. <https://doi.org/10.3389/fgene.2019.00132>
- Ye SH, Siddle KJ, Park DJ, Sabeti PC (2019) Benchmarking metagenomics tools for taxonomic classification. *Cell* 178: 779-794 <https://doi.org/10.1016/j.cell.2019.07.010>
- Yu DW, Ji Y, Emerson BC, Wang X, Ye C, Yang C, Ding Z (2012) Biodiversity soup: Metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring. *Methods in Ecology and Evolution* 3: 613-623. <https://doi.org/10.1111/j.2041-210X.2012.00198.x>
- Zepeda Mendoza ML, Sicheritz-Pontén T, Gilbert MTP (2015) Environmental genes and genomes: understanding the differences and challenges in the approaches and software for their analyses. *Briefings in Bioinformatics* 16(5): 745-758. <https://doi.org/10.1093/bib/bbv001>
- Zhou X, Li Y, Liu S, Yang Q, Su X, Zhou L, Tang M, Fu R, Li J, Huang Q (2013) Ultra-deep sequencing enables high-fidelity recovery of biodiversity for bulk arthropod samples without PCR amplification. *Gigascience* 2(1): 4. <https://doi.org/10.1186/2047-217X-2-4>

Annexes

Annexes

Data Accessibility

- The datasets used in the analyses reported in this thesis are deposited in the Dryad repository <https://doi.org/10.5061/dryad.t1g1jwsz7>.
- The γ - δ algorithm script is available at GitHub: https://github.com/LidiaGS/g-d_algorithm.
- The software used in chapter 5 is available at GitHub: https://github.com/LidiaGS/ensemble_BM_K2.

Supplementary material

Supplementary material may be found in pen-drive attached.

Supplementary Figures

Figure S3.1. Summary boxplots of the 22 single-species libraries used to test the quality of the adjusted parameters.

Supplementary Tables

Table S3.1. List of reference genomes used in the MG study.

Table S3.2. Proportion of reads assigned to species, on single-species libraries from run *no.* 1.

Table S3.3. Proportion of reads assigned to species, on single-species libraries from run *no. 2*.

Table S3.4. Assignment of misidentified reads from run *no. 1* using megablast and NCBI nucleotide collection (*nt*) database.

Table S3.5. Comparison of the proportion of reads assigned for species sequenced twice.

Table S3.6. Relative proportion of assigned reads to species for each one of the six mixed-species libraries.

Table S3.7. Summary table of the execution times of mixed-species library.

Table S4.1. List of reference mitogenomes used in chapter 4.

Table S4.2. Relative Species Abundance (RSA) and summary information of the number of processed reads in the mixed-species libraries.

Table S4.3. Comparison of the number of identified species per library using different combinations of γ and δ .

Table S4.4. Estimation of the species included on the training dataset of single-species libraries by mapping raw reads to Mito1794 database and using different combinations of γ and δ .

Table S4.5. Estimation of the species included on the training dataset of single-species libraries by mapping raw reads to FilteredMito1794 database and using different combinations of γ and δ .

Table S4.6. Estimation of the species included on the training dataset of single-species libraries by mapping candidate mito-reads to Mito1794 database and using different combinations of γ and δ .

Table S4.7. Estimation of the species included on the training dataset of single-species libraries by mapping candidate mito-reads to FilteredMito1794 database and using different combinations of γ and δ .

Table S4.8. Summary table of the exploration of masked regions in 23 mitogenomes (17 focal species plus 6 wrong species) of species detected when mapping single-species libraries against Mito1794 database.

Table S4.9. Estimation of the species included on the test dataset of single-species libraries by mapping candidate mito-reads to FilteredMito1794 database and using $\gamma = 0.99$ and $\delta = 0.96$.

Table S4.10. Estimation of the *across*-species RSA on mixed-species libraries by mapping candidate mito-reads to FilteredMito1794 database and using $\gamma = 0.99$ and $\delta = 0.96$.

Table S4.11. Summary table of the execution times of each step in the MMG pipeline for the six mixed-species libraries.

Table S4.12. Summary table of mitogenomes size for insect orders.

Table S4.13. Summary table of whole-genome size for insect orders on Table S3.1.

Table S5.1. List of reference mitogenomes used in chapter 5.

Table S5.2. List of assigned species for each single-species library and BM, K2, and combinations pipelines.

Table S5.3. Results for metrics (richness, RPIR, precision, recall and execution time) for all pipelines (*i.e.*, BM and K2) and libraries.

Supplementary Methods

Methodology S3.1. Commands used for running the complete $B\gamma\delta$ pipeline.

Methodology S5.1. Commands used for running the complete pipeline with BM and K2 metagenomic classifiers.

