Impact de la dérive génétique sur l'évolution de la composition en base des gènes et de la complexité du transcriptome chez les métazoaires

Florian Fabien Bénitière

Université Claude Bernard Lyon 1

THÈSE de DOCTORAT DE
L'UNIVERSITÉ CLAUDE BERNARD LYON 1

Ecole Doctorale 341
Écosystèmes Évolution Modélisation Microbiologie

Discipline : Génomique Évolutive

Soutenue publiquement le 22 mai 2024, par :
Florian Bénitière

Impact de la dérive génétique sur l'évolution de la composition en base des gènes et de la complexité du transcriptome chez les métazoaires

---

Devant le jury composé de :

Stephanie BEDHOMME Chargée de Recherche, CNRS/CEFE, UMR 5175	**Rapportrice**
Hugues ROEST CROLLIUS Directeur de recherche, CNRS/IBENS, UMR 8197	**Rapporteur**
Anna-Sophie FISTON-LAVIER Maîtresse de conférence, Université de Montpellier/ISEM, UMR 5554	**Examinatrice**
Sabine PÉRÈS Professeure, Université Claude Bernard Lyon 1/LBBE, UMR 5558	**Examinatrice**
Laurent DURET Directeur de recherche, CNRS/LBBE, UMR 5558	**Directeur de thèse**
Tristan LEFÉBURE Maître de conférence, Université Claude Bernard Lyon 1/LEHNA, UMR 5023	**Co-directeur de thèse**

We are no better or more evolved than any other living thing. We are not above nature. We are simply part of it. [...] Some of us have evolved enough to realise we have not evolved enough.

Ricky Gervais (2014)

Remerciements

Je tiens à remercier toutes les personnes qui ont fait de ces années un temps d' épanouissement personnel et scientifique.

Aux amis, toujours là, m'inspirant par leurs qualités, avec qui les heures en dehors du labo sont si ressourçantes et nécessaires pour me libérer l'esprit.

À Bobo/Gégé pour ces 28 ans de délires, les missions en Irlande, Pologne, stratedgic object, Italie, les tennis, la spéléo, mais pas trop... Ton goût pour l'aventure me pousse à oser sortir de ma zone de confort. Je garde l'espoir d'une promotion.

Le Doud, depuis le collège t'as toujours été là, merci pour ces journées tt/AOE, pour m'avoir introduit à la team CPE. Ton calme, ton organisation, ton professionnalisme et ton game play à fifa m'ont longtemps inspirés. Merci aussi de me backuper avec ton aide en informatique, en ML, et pour les travaux... On le fera ce micro-ondes.

Lucas, poto depuis mes débuts au LBBE il y a 6 ans, tu es pour beaucoup responsable de mon rapport à la vie, à la relativisation, et au lâcher-prise. Merci pour tous ces moments, pétanque, Sucre, fifa, techno, festoch *etc.* À toi et à Kazou pour m'avoir fourni des nuits d'hôtel accompagnées de grandes discussions savantes pendant le Covid !

Vic, pour tous ces superbes moments qu'on a partagés à Nice. Franchement, lors de mes débuts au laboratoire, passer de tels weeks-ends dans le sud était vraiment top. Tes projets pro à l'étranger ont indéniablement nourri mon envie de partir travailler au Canada.

La team Paris évidemment, Seb, Vince et Flo ; les bivouacs, les ampoules, les lynxs, BK production *etc.* C'est toujours un plaisir de passer des moments avec vous, rires et activités garantis.

Au come-back pour les belles vacances passées chaque été et Nouvel An dans des lieux toujours plus stylés.

Flo la rencontre techno, merci à toi, Val et Florent, à ces dimanches après-midi sur la terrasse du Sucre avant d'attaquer la semaine, tes divines pâtisseries, ta bonne humeur et tes pitreries.

Aux collègues devenus amis, Thibault, pour m'avoir intégré dans le domaine de la bioinfo avec tant de bienveillance, me délivrant tant de tips. À cette fameuse baignade dans la grotte lors de la 1ère inté avec Judith, Alex et Théo. À nos pique-niques/pétanque sur les quais au coucher de soleil.

Merci à la team des docs/postdoc du 2ème à mon arrivée au labo, Alexia, Alex, Dji, Théo, Diego, Marina. Vraiment de bons moments passés. Rien que cette première conférence Lille into Bruxelles était exceptionnelle (surface plane). Je me suis rapidement aperçu que les années à venir allaient être formidables avec vous dans cet environnement de travail.

Gasp, l'élève à dépasssé le maître sur RL, tu peux remercier le Covid. C'est regrettable que tu en sois venu à le renier et gâcher tant de talent. Merci pour ton aide à l'animation au labo, et pour ces games avec Abi et Thibaud, que je remercie également.

Aux potes plus récents, à la team Mroc, aux soirées jeux, aux crémaillères désordonnées *etc.*: Mélo, Matthieu, Soso, Lulu, Jess, Emma, Adrian, PA, Luchad le pgm EVA, Amandine, Rémi, Léa, Max, Valoche.

À ma famille, Mams, Pat, merci de m'avoir donné l'opportunité de faire tant d'études, dans un cadre si privilégié, de m'avoir toujours poussé vers l'avant et donné les moyens d'y arriver. Ton carriérisme, Mams, m'apprend à persévérer pour suivre mes passions. Pat, ton sens extrême de l'organisation coule dans mes veines et me permet de rester ordonné face aux grandes quantités de données que j'ai à gérer.

Quentine/bb mousse pour m'avoir initié aux basses, et Pau, pour ces étés au soleil et nos moments en famille.

Papi, Mamie, c'est bien dans votre campagne que j'ai commencé à m'intéresser à la biologie, en passant tant de temps dans la nature, à courir après les vaches et à pêcher la truite... Merci.

Didou, merci pour ton soutien et les jours passés à réviser dans l'Antre des verts. Ce fut un plaisir d'avoir été salarié d'ELCO.

À la team LEHNA, malgré le peu de temps passé dans vos locaux et des sujets assez différents du reste de l'équipe, je garde de très bons souvenirs. Vous avez été très accueillants, et j'ai pu avoir des échanges enrichissants et bienveillants. Merci pour la journée d'équipe, la sortie canoë *etc.* Spécial merci à Clémentine, Maïlys, Axelle, Gautier, Nans, Paul, Héloïse, et Emma.

Au LBBE, je ne compte plus les personnes qui m'ont aidées, soutenues, avec qui j'ai eu des discussions, qui m'ont permis d'évoluer, et de faire évoluer ma façon de faire de la recherche. Les pôles du LBBE, administratif évidemment: Nathalie merci de gérer tout cela à la perfection malgré la multitude de contrats que j'ai eu dans ce labo. Merci pour ce travail qui nous déleste de ces tracas administratifs. Bien sûr, merci au pôle informatique, Stéphane et Bruno, je vous ai suffisamment dérangés avec le cluster, iRods, le pbil, les VMs, Shiny *etc.* Si j'ai pu réaliser autant d'analyses c'est parceque l'infrastructure à laquelle nous avons accès est vraiment entre de bonnes mains. Au fil des années j'ai pu apprécier les efforts d'amélioration qui ont été proposés, et c'est un bonheur de savoir que l'on peut compter sur vous et venir vous déranger n'importe quand pour la moindre question, et d'en sortir avec plus de solutions qu'on avait de questions. Merci évidemment aussi à Philippe de prendre autant de temps pour m'expliquer les choses avec bienveillance et pédagogie.

Merci à la team HH/LBBistrot, à Théo pour nous avoir passé la main, et donc à Gasp, Nat et Pascal de m'avoir rejoint. Si c'est tant une réussite c'est bien grâce à vous, et c'était un plaisir de gérer ça avec vous.

Aux Pinsons, merci de proposer autant d'activités renforçant la cohésion au labo.

Merci à nos DU Fabrice et Manu pour faire de ce labo un espace où il fait bons vivre et laisser/encourager les prises d'initiatives.

La team BPGE, je garde de superbes souvenirs, à mon arrivée en 2018, de cette retraite au lac de Cublize, qui m'a directement plongé dans une équipe top, pleine de bienveillance et de projets. Avec tant d'experts dans différentes disciplines de la biologie évolutive, ce fut un cadre de travail dynamisé par de nombreuses réunions, scientifiquement enrichissantes, avec une recherche constante de rigueur.

Merci à ceux à qui j'ai pu demander de l'aide et échanger au cours de ces années : Alba Marino, Olivier Arnaiz, Olivier Tenaillon, Maud Tenaillon, Ignacio Bravo, Julien Clavel, Laurent Gueguen, Guillame Achaz *etc.*

Merci à ceux qui m'ont ouvert les portes de la recherche et de la science : Elie Maza, Roland Barriot, Laurent Bezin, Simon Hardy, et Patrick Desrosiers.

Merci aux membres de mon comité de pilotage, Benoit, Emiliano et Vincent, pour vos retours constructifs et les discussions sur les perspectives de carrière encourageantes. Merci Rita, tu as été une super tutrice, disponible, à l'écoute, motivante, bienveillante et de bon conseil.

Aux rapporteurs de ce manuscrit, Stéphanie et Hugues, merci d'avoir accepté de relire ce manuscrit et pris le temps de le corriger. Anna-Sophie et Sabine, merci d'avoir accepté d'être membre de mon jury.

Tristan, merci de m'avoir donné l'oppportunité de réaliser cette thèse, merci pour nos échanges aussi bien scientifiques qu'humains, et surtout merci pour les relectures de ce manuscrit.

Anouk, que te dire ? Depuis le début de cette aventure, tu es très, voire trop bienveillante avec moi. Merci d'avoir pris le temps de me former techniquement, de m'avoir transmis l'importance d'être organisé. Merci pour nos collaborations, pour toutes les discussions, les quelques matchs de tennis, les paniers légumes *etc.* Merci d'avoir été à l'écoute, une confidente à qui je pouvais parler si ça n'allait pas. Tu es grandement responsable de mon bien-être au labo, et de mon épanouissement scientifique. Finalement, je te considère comme une encadrante de cette thèse tant tu y as contribué.

Laurent, alias chef, évidemment que je te dois vraiment beaucoup voir tout. Merci de m'avoir accordé autant de confiance et de m'avoir gardé à tes côtés pendant toutes ces années. J'ai eu la chance d'avoir un superviseur toujours disponible, malgré les responsabilités que tu as, accessible d'une simple rotation de chaise pour discuter que ce soit science, carrière ou perso. Notamment, j'ai souvenir pendant le Covid où je n'avais qu'à envoyer un '*dans 5min sur discord ?*' par mail, et nous voilà en visio pour discuter science ou société, environ une fois par jour voire plus. Tu as fait de cette période une phase épanouissante scientifiquement. Tu es extrêmement patient et pédagogue, m'expliquant les choses sans jamais paraître agacé et en prenant le temps. Tu m'as toujours fasciné par la richesse de ton savoir, et surtout par ton humilité, ce qui m'a nécessairement fait grandir scientifiquement et humainement. J'étais constamment en présence d'une montagne de savoir pour laquelle je me plaisais à être l'expert technique et à contribuer à ses interrogations. Nous avons certes pu avoir des désaccords sur la facon de valoriser nos résultats dans un domaine toujours plus compétitif, mais tu m'as fourni un cadre de travail incroyable, que j'espère retrouver ailleurs. Ce fut un vrai plaisir de travailler pour et avec toi pendant ces 6 années. Enfin, merci pour le nombre de relectures, que ce soit pour ce manuscrit ou bien nos articles scientifiques. Merci pour tous tes retours. Un jour, peut-être mais j'en doute, j'atteindrai ce niveau d'excellence.

Et enfin, merci à **mon Beb**. Merci **Mel** d'être toujours là, radieuse, solaire, altruiste, et tellement joviale. Si je peux avoir des hauts et des bas, que ce soit lié au boulot ou autre, tu es, toi, toujours souriante et à l'écoute. Tu es en quelque sorte ma psy personnelle, et une scientifique/collaboratrice à l'écoute de mes projets. Finalement, nous avons acheté et vécu ensemble dès le début de ma thèse, et ce fut l'une de mes meilleures expériences. C'est si ressourçant de vivre avec toi, notamment pendant cette thèse où moralement, c'était une période moins facile que lors de mes précédents contrats. Tu es toujours partante pour faire des activités, toujours enthousiaste, même lorsqu'il s'agit de partir sur un autre continent ensemble. Je partage tellement de bons moments avec toi, et j'espère en créer tant d'autres. Merci, Je t'aime.

Merci à tous ceux que j'aurais pu oublier mais qui, au cours de ces années, ont contribués, directement ou indirectement, à ces travaux de recherche.

Depuis ces lignes, je laisse la place à la science.

Florian Bénitière

Impact de la dérive génétique sur l'évolution de la composition en base des gènes et de la complexité du transcriptome chez les métazoaires

Résumé

Les génomes présentent une diversité remarquable, variant en taille, composition en base, expression, nombre de gènes *etc.* Comprendre l'origine de ces changements captive. Trois forces évolutives sont à l'œuvre: la mutation, *i.e.* la création de nouveaux allèles; la sélection, *i.e.* l'impact des allèles sur la fitness; la dérive génétique aléatoire, *i.e.* l'échantillonnage aléatoire des allèles au fil des générations. Si, au cours de l'évolution, des mutations bénéfiques (*i.e.* qui ont contribué à l'adaptation des espèces à leur environnement) se sont fixées dans les génomes, la sélection naturelle ne peut à elle seule expliquer toute la diversité observée au niveau moléculaire.

La théorie neutraliste de l'évolution nous apprend que les mutations ayant des effets négligeables sur la fitness d'un organisme jouent également un rôle important dans l'évolution des génomes. Plus précisément, la théorie quasi-neutraliste suggère que l'efficacité à purger (ou fixer) des mutations légèrement délétères (ou avantageuses), dépend de la capacité de la sélection à dominer les effets aléatoires de la dérive. À pression sélective égale, la capacité d'un génome à atteindre l'optimal est ainsi limitée par l'intensité de la dérive à laquelle il est soumis. Cette hypothèse, connue sous le nom de "barrière de la dérive", prédit que les espèces à forte dérive, et donc à faible taille efficace (N_e), ont un génome moins bien optimisé que celles qui présentent une N_e plus grande.

Au cours de ma thèse, j'ai étudié l'impact de la dérive génétique sur l'architecture des génomes chez les animaux. Pour ce faire, j'ai collecté de nombreux génomes et transcriptomes, provenant de 1,507 espèces et 15,935 échantillons RNA-seq, afin d'analyser la diversité transcriptomique et la composition des séquences codantes. Ces analyses sont présentées dans une base de données bio-informatique, incluant des estimateurs de N_e.

Ainsi, j'ai étudié l'influence de la dérive sur la diversité transcriptomique, à travers la quantification des variants d'épissage. Nos résultats ont démontré que la dérive limite la capacité de la sélection à optimiser la machinerie d'épissage dans les génomes. Ce qui provoque la production de nombreux transcrits erronés chez les espèces à faible N_e, comme l'Homme.

Je me suis intéressé à un autre trait de la composition en base des génomes, et en particulier l'usage des codons synonymes chez les métazoaires. Nos résultats ont révélé que chez les espèces où la sélection traductionnelle (*i.e.* favorisant l'utilisation de codons optimisant le processus de traduction) est détectée, le coefficient de sélection à l'échelle populationnelle est faible. Cela suggère que la sélection traductionnelle ne peut être efficace que chez les espèces à grande N_e, justifiant sa rareté chez les métazoaires. Néanmoins, certaines espèces ayant une grande N_e ne montrent pas de signaux de sélection traductionnelle, suggérant que l'avantage sélectif à optimiser le processus de traduction varie.

Finalement, cette thèse illustre l'impact de la dérive sur l'architecture des génomes et fournit un cadre conceptuel intéressant, ainsi qu'une collection de données réutilisables, pour examiner ce qui est ou n'est pas soumis à la sélection dans nos génomes.

The impact of random genetic drift on the evolution of genes base composition and expression complexity in metazoans

Abstract

Genomes exhibit remarkable diversity, varying in size, base composition, expression, number of genes *etc.* Understanding the origin of these changes captivates. Three evolutionary forces have shaped genomes: mutation, *i.e.* the creation of new alleles; selection, *i.e.* the impact of alleles on fitness; random genetic drift, *i.e.* the random sampling of alleles through generations. While, over time, beneficial mutations (*i.e.* which have contributed to the adaptation of species to their environment) have become fixed in the genomes, natural selection alone cannot explain all the diversity observed at the molecular level.

Instead, the neutral theory of evolution posits that mutations with negligible effects on individual fitness also play a significant role in shaping genome evolution. Specifically, the nearly neutral theory suggests that the efficiency to purge (or fix) slightly deleterious (or advantageous) mutations, depends on the ability of selection to overcome drift hazard. If selective pressure on specific biological traits remains constant, a genome's ability to attain optimality becomes limited by drift. This hypothesis known as the "drift barrier" predicts that species with strong drift, and thus small effective population size (N_e), have a poorly optimized genome compared to those with larger effective population sizes.

Throughout my thesis I studied the impact of random genetic drift on genomes architecture in animals. To do so, I collected numerous genomes, from 1,507 species, and transcriptomes, from 15,935 RNA-seq samples, to analyze their transcriptomic diversity and coding sequences composition. These analyses are shared in a bio-informatic data resource along with effective population size proxies.

One key aspect of my research involved investigating how increasing drift intensity affects transcriptomic diversity, through the study of splicing variants. Our results demonstrated that drift limits the capacity of selection to optimize the splicing machinery in genomes. It ultimately leads to the production of many spurious transcripts in species with small N_e, such as human.

I investigated another characteristic of genomes base composition, the use of synonymous codons in metazoans. Our research revealed that in species where translational selection (*i.e.* promoting the use of codons optimizing translation process) is detected, the population-scaled selection coefficient is small. This suggests that translational selection can be efficient only in species with large effective population size, elucidating its rarity in metazoans. Nevertheless, intriguingly, certain species with large N_e did not show translational selection signals, which implies that the selective advantage in optimizing the translation process varies across species.

Overall, this thesis underscores the impact of drift on genome architecture and provides an interesting conceptual framework, along with a collection of reusable data, to examine what is or is not under selection in our genomes.

Contents

List of Figures

Acronyms

N_e Effective population size. 19, 20, 23–28, 40, 45, 52, 53, 57, 66, 70–72, 74, 76–80, 82, 87–93, 129–132, 136–139

K Substitution rate *per* generation. 23–25

N Census population size. 24, 70, 137

S Population-scaled selection coefficient. 24, 25, 130, 132, 137

μ Mutation rate *per* base pair *per*. 23, 25, 27, 138

π_s Synonymous nucleotide diversity. 25, 138

s Selection coefficient. 22–27, 44, 45, 52, 70, 131, 132, 138

A Adenine. 4, 10

ADW Animal Diversity Web. 54–56, 66

AS Alternative Splicing. 9, 38, 44, 45, 52, 53, 65, 74, 76–78, 129, 131–133, 136, 138, 139

bp Base pairs. 6, 7, 20–22, 35, 139

BUSCO Benchmarking Universal Single-Copy Orthologs. 36, 50, 53, 57, 58, 62–72, 78, 80, 82–86, 89, 91, 94, 96, 97

C Cytosine. 4, 11

CDS Coding Sequence. 5, 6

CU Codon Usage. 11, 47

DNA Deoxyribonucleic acid. 4, 7, 13, 16, 28, 30–33

EBI European Bioinformatics Institute. 35

EMBL European Molecular Biology Laboratory. 35

ENA European Nucleotide Archive. 35, 36

EOL Encyclopedia Of Life. 54, 56, 66

FPKM Fragment *Per* Kilobase of exon *per* Million mapped reads. 62, 64, 65

G Guanine. 4, 11

Gb Giga base pairs. 4, 38

gBGC GC-Biased Gene Conversion. 28, 45, 47

GC3 GC-content at third codon position. 11, 46, 59

I Inosine. 10

Indel Insertion or deletion. 20, 21

kb Kilo base pairs. 5, 38, 59, 136

lncRNA Long non-coding RNA. 5

Mb Mega base pairs. 4, 38

ML Machine Learning. 58, 66

mRNA Messenger RNA. 6–8, 10, 11, 35, 133, 136

NCBI National Center for Biotechnology Information. 35, 36

NSP Neutral Substitution Patterns. 46, 48, 129, 130

POC Putative-Optimal Codon. 130, 137

RNA Ribonucleic Acid. 5, 7, 8, 10, 11, 30, 31, 48, 50, 60, 137

SNP Single-Nucleotide Polymorphism. 35

SV Spliced Variant. 80–88, 90–92, 98, 129

T Thymine. 4, 7

tRNA transfer RNA. 5, 10–12, 47, 48, 130, 133

TS Translational Selection. 12, 46–48, 130, 131, 136, 137

TSS Transcription Start Site. 6

U Uracil. 7

UTR Untranslated Region. 6, 10

Glossary

Allele A variant form of a given gene. 21, 22, 45, 47

Biased gene conversion Process by which gene conversion is biased towards a given outcome. It occurs when one haplotype has a higher probability of being the donor. 22, 28, 45, 47

Codon Sequence of three nucleotides coding for a given amino acid. 10, 20, 21, 27, 45, 47, 48, 129, 130, 133

Codon usage Usage of the synonymous codons. 11, 129, 133, 134

CpG Region of DNA where a cytosine is followed by a guanine in the 5'-to-3' direction.. 21, 85, 98

Diploïd Organism displaying a ploidy of 2 (n = 2), *i.e.* two sets of chromosomes (which are paired). 23, 25, 26, 28

dN Non-synonymous substitutions rate. 26, 27, 71, 72

dS Synonymous substitutions rate. 71, 72

Effective population size The number of individuals in a Hardy–Weinberg population who contribute to the next generation. 19, 23–28, 128, 129, 138

GC-content The percentage of G or C nucleotides in a DNA sequence. 11, 137, 138

Genetic drift The random fluctuation in allele frequencies due to random sampling of individuals. 19, 22, 23, 128, 129, 131, 132, 135, 137

Non-synonymous Mutation that modifies the encoded amino acid. 20–22, 26, 27, 138

Phenotype The composite of observable traits. 22, 44, 133

Substitution Point mutation that appeared in only one individual in the population, and subsequently reached fixation in the population. 18, 22, 24–27, 46, 48, 69, 71, 90, 96, 129, 130, 138

Synonymous Mutation that does not modify the encoded amino acid. 11, 12, 18, 27, 133, 138

Translational selection Selection optimizing the speed and accuracy of translation. 12, 46, 47, 128, 129, 133, 136, 137

Preamble

Four billion years of evolution have shaped animal genetic materials, *i.e.* genomes, and human seeks to uncover what is biologically functional in them. Notably, metazoan genomes are complex and vary in size, base composition, expression, number of genes, structure *etc.* I will open the introduction in Part I with Chapter 1 by illustrating the complexity and diversity of metazoan genomes architecture and expression. In order to discern what is functional in these genomes, it is imperative to understand the evolutionary forces that have shaped them over time.

'Nothing in Biology Makes Sense Except in the Light of Evolution' T.Dobzhansky (1973)

I present in Chapter 2 my scientific area, evolutionary biology, where we explore genomes and the main evolutionary forces that affect them: mutation, selection and drift. Notably, since Darwin in 1859, who elucidated the role of natural selection in species evolution, successive theories highlighted that natural selection cannot be responsible for all changes. Indeed, neutral or slightly neutral changes affect genomes, and drift hazard may disturb the efficiency of selection to promote slightly advantageous mutations or purge deleterious ones. However, because of the innate desire to discover meaningful characteristics of genomes, we tend to jump to conclusions and prematurely attribute changes solely to natural selection. This, eventually, created vigorous debates within the scientific community between neutralist and selectionist views of evolution.

Fortunately, we live in a period of surrounding data, with an unprecedented amount of sequenced genetic materials available for study, that I present in Chapter 3. This wealth of data, combined with advances in methodology and technology, allowed the biologist to re-investigate evolutionary theories.

In this powerful context, I present in Chapter 4 how I intend to bring new answers to the neutralist/selectionist debate. Indeed, the "drift barrier" hypothesis, which predicts that reduced drift leads to more optimized genomes, offers me a great conceptual framework for discerning adaptive and non-adaptive traits by studying the impact of random genetic drift on genomes architecture.

The Part II outlines the methodology and findings of my thesis in three articles. Initially, I gathered genomic and transcriptomic data, which underwent integrative bioinformatic analyses to facilitate comparative studies. Thus, I introduce GTDrift in Chapter 5, a comprehensive data resource housing transcriptomes analyses alongside effective population size proxies (N_e). This resource serves for exploring the impact of drift on genome architecture. In Chapter 6, I study how transcriptome diversity is affected by drift, through the quantification of alternative splicing in metazoans. Chapter 7 focuses on elucidating the reasons behind the paucity of translational selection in metazoans.

Finally, I discuss in Part III the implications of this thesis, emphasizing the necessity for reproducibility in a world surrounded by data and analyses, to maintain scientific integrity and ensure qualitative research.

Part I

Introduction

1

Metazoan genome structure and expression

Contents

My thesis work is focused on the metazoans, commonly known as animals, consisting of arthropods and other clades such as sponges, mollusks, vertebrates, which include fishes, birds, mammals and reptiles. This group appeared around 543-510 million years ago during the cambrian explosion, based on fossil datation (Tikhonenkov *et al.*, 2020). Initially, they were characterized based on their morphology. Their primary attributes include being eukaryotes, which implies that their cells possess a complex structure with distinct organelles and a compartmentalized nucleus containing the genetic material. In contrast, prokaryotic species have their genetic material dispersed within the cell's cytoplasm. Another key feature of metazoans is their multicellular nature, wherein an individual consists of multiple cells capable of forming various tissues, and acquiring nutrients by consuming organic matter from other organisms (heterotrophy).

Because metazoans present a wide variety of genomes architecture and biological traits, they provide an exciting opportunity to explore the interplay between genomes evolution and the biology of organisms. Moreover, many data are available for this group of species.

This section aims to explore the fundamental aspects of metazoans' genome structure and the mechanisms that determine its expression. The objective is to gain a comprehensive understanding of the genomic components investigated in my study.

1.1 Genomes content

The genome, which can be subdivided into one or more fragments called chromosomes, contains the individual genetic material and is inherited by future generations through reproduction. In metazoans, as in most/all living beings, the genome consists of double-stranded molecules of deoxyribonucleic acid (DNA), this structure was untangled by several scientists in 1953 (Franklin and Gosling, 1953; Watson and Crick, 1953; Wilkins *et al.*, 1953). The genetic information is composed of four nucleotides: pyrimidine bases cytosine (C), thymine (T), and purine bases adenine (A), guanine (G). These nucleotides A, G, C, and T are present in DNA, with T bonding to A and G bonding to C through hydrogen bonds. The directionality of a single strand of DNA, called 5' and 3' ends, is determined by the position of chemical elements (Fig. 1.1). This orientation and the sequence of nucleotides determine the genetic information encoded in the genome.

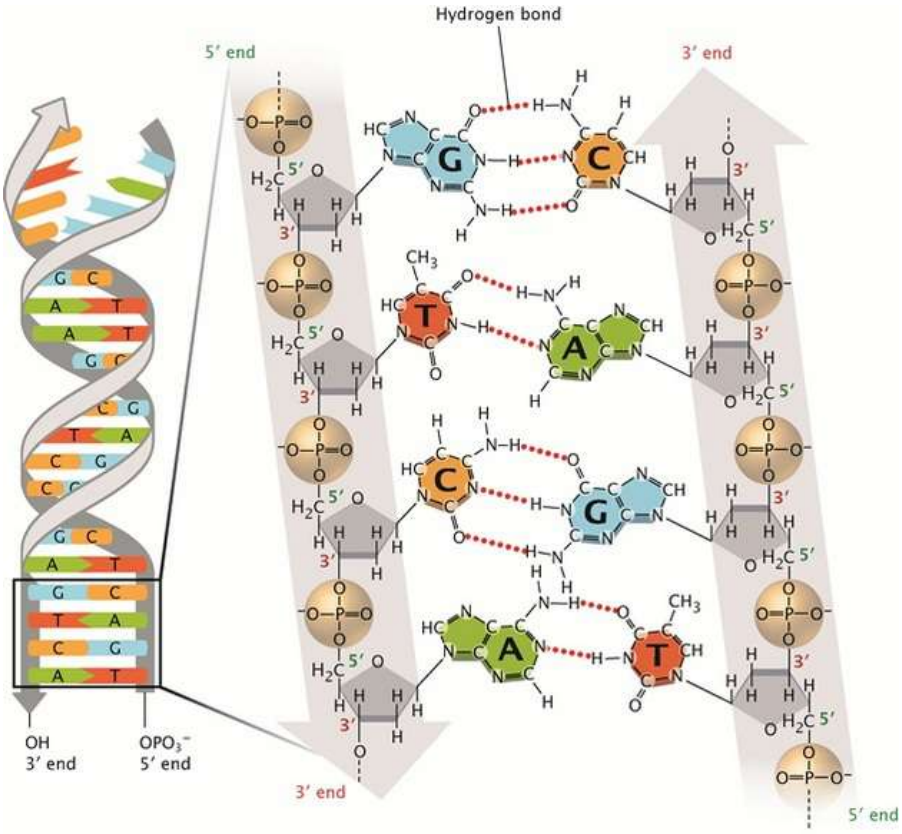


Figure 1.1: DNA molecule. *Double-stranded molecules of deoxyribonucleic acid (DNA), illustrating the structural and atomic bonds connecting each nucleobase.*
Reproduced from Springer Nature for a noncommercial use.

The size of the haploïd genome in metazoans varies significantly, ours is 3.2 Gb (104 cm, Piovesan *et al.* (2019b)). On average, metazoan genomes tend to be around 1 Gb in size (Hotaling *et al.*, 2021). Remarkably, the *Australian lungfish* boasts a genome 14 times larger than that of humans, measuring approximately 43 Gb (Meyer *et al.*, 2021). On the other hand, the marine parasite *Intoshia variabili* possesses the smallest known animal genome, measuring only 15.3 Mb (Slyusarev *et al.*, 2020). Nevertheless, it is

essential to acknowledge that not all genomes have been measured to date, leaving room for further discoveries and variations in genome sizes across the animal kingdom.

The genome comprises distinct regions with diverse functions, including coding regions that oversee protein synthesis and non-coding regions. Remarkably, a significant proportion of our genome, approximately 98.5%, is comprised of non-coding regions. In stark contrast, only a mere 1.5% of the genome encompasses coding regions responsible for protein synthesis. These coding regions are found within protein-coding genes making up nearly one-third of the entire genome (Lander *et al.*, 2001; Venter *et al.*, 2001; International Human Genome Sequencing Consortium, 2004).

1.2 Genes: holder of genetic information

The concept of 'gene' has undergone changes in its definition since its first mention as 'inherited cell elements' by Mendel *et al.* (1866). The word 'gene' was subsequently introduced by Johannsen (1909) and quickly became essential in the emerging field of genetics. Nowadays, a definition of a gene in eukaryotic organisms could be: 'a transcription unit, whose expression leads to the production of a functional molecule (RNA or protein)'. Transcription being the first step of gene expression (see 'RNA transcription and maturation process' section). Genes can be broadly classified into two main categories: the first category includes non-coding RNA, such as transfer RNAs (tRNAs), ribosomal RNAs (rRNAs), microRNA (21-25 nucleotides) responsible for controlling expression (Filipowicz *et al.*, 2008; O'Brien *et al.*, 2018) and long RNA molecules (greater than 200 nucleotides) known as lncRNA (Kashi *et al.*, 2016).

The second category comprises protein-coding genes that encode proteins. Protein-coding genes exhibit substantial variation in size across different species. For instance, in humans, the median gene size is approximately 24 kb (Fuchs *et al.*, 2014), whereas in diptera *Drosophila melanogaster*, genes are comparatively shorter with a median size of 2 kb for its 14,000 protein-coding genes (assembly GCF_000001215.4). Additionally, there are pseudogenes, which are remnants of protein-coding genes that have either lost their function over time or are not transcribed. In humans, there are 8,700 pseudogenes annotated (assembly GCF_000001405.38, Mighell *et al.* (2000)).

Protein-coding genes in metazoans follow a specific structure. Each gene begins with a promoter sequence that plays a crucial role in activating the first step of gene expression, its transcription (Fig. 1.2). This region is responsible for attracting RNA polymerase II, the enzyme responsible for initiating its expression.

Following the promoter there is the transcribed region of a gene. This region consist of alternating sequences known as exons and introns. Exons constitute the protein coding region (CDS) and its 5' and 3' untranslated regions, whereas introns are non-coding sequences and represent 25% of our genome (Sakharkar *et al.*, 2004). A recent investigation encompassing 1,700 species reveals that the primary origin of continuous intron formation in eukaryotic genomes stemmed from introners, which are transposable elements capable of creating copies of themself that insert into many genes throughout the

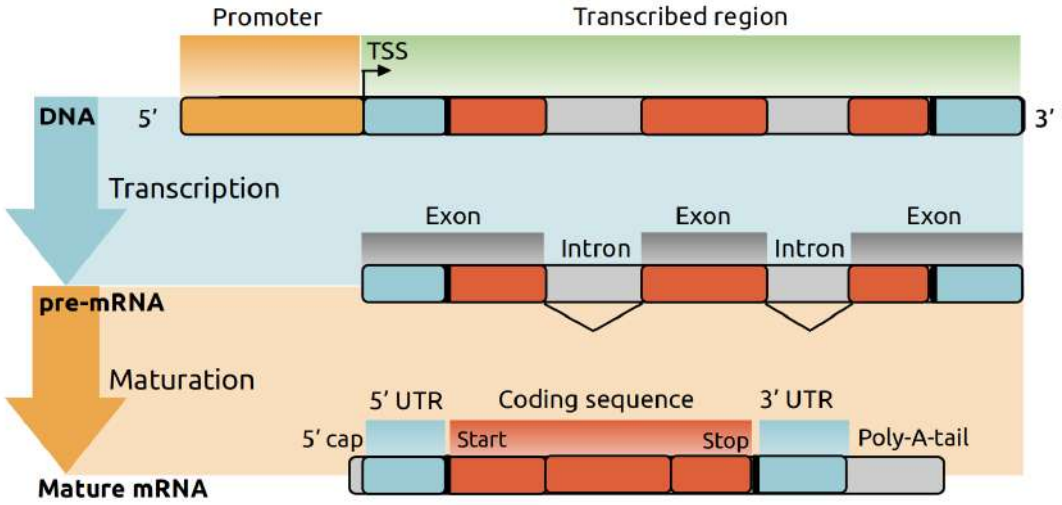


Figure 1.2: Gene structure and expression. *Schematic representation of an eukaryotic gene and its structural components, including the promoter, transcription start site (TSS), untranslated transcribed regions (UTR), coding sequence (CDS), as well as introns and exons. A concise overview of key processes such as transcription of DNA and maturation of pre-mRNA is provided to appreciate the mechanisms underlying gene expression.*

genome, akin to intragenomic parasitic elements (Gozashti *et al.*, 2022). Not all genes contain introns, we called multiexonic genes those with both exons and introns. The fraction of multiexonic genes is high but varies slightly between species, *e.g.* 90% in fly and 95% in *Homo sapiens*. In human, the coding regions of multiexonic genes typically contains an average of 8 introns with a median length of 1,747 bp, and 9 exons with a median length of 131 bp (Piovesan *et al.*, 2019a). These characteristics vary widely between species, in number: 4 exons per gene in dipterans to 10 exons in vertebrates, and in length: on average, the introns measure 1,000 bp in vertebrates, 1,200 bp in mammals, 850 in birds to 80 bp in dipterans.

In the final product of transcription, the coding sequence is flanked by Untranslated Transcribed Regions (UTRs). The 5' UTR region serves as the site where the ribosome enzyme attaches to initiate translation. On the other hand, the 3' UTR region plays a role in translation termination, mRNA stability and localization (Grzybowska *et al.* (2001); Chabanon *et al.* (2004); Mayr (2019); see 'Translation and decoding via the genetic code' section).

Many protein-coding genes have been identified: 6,000 in yeast, 14,000 in *Drosophila melanogaster*, 26,000 in *Arabidopsis thaliana*. Initially, prior to the sequencing of the human genome, estimates suggested that up to 140,000 genes could be present (Fields *et al.*, 1994). Indeed it was expected, due to assumptions and anthropocentric perspectives, that the human being the most complex species of all, would possess a large number of genes. However, the first large-scale comparative analysis of vertebrate genomes revealed a modest count of $\approx 30,000$ genes in our genome (Roest Crollius *et al.*, 2000),

later confirmed by the Genome Sequencing Consortium (Lander *et al.*, 2001; Venter *et al.*, 2001). Following improvement in the assembly and annotation, the current estimation dropped to around 20,000 genes (Clamp *et al.*, 2007; Ezkurdia *et al.*, 2014; Piovesan *et al.*, 2019a). While the notion of organism complexity is difficult to define and subject to anthropocentric bias, its discrepancy with gene count, known as the G-value paradox, continues to be the subject of intense investigation and raises numerous questions (Hahn and Wray, 2002; Straalen *et al.*, 2011; Choi *et al.*, 2020).

1.3 RNA transcription and maturation process

Protein-coding genes are expressed through the transcriptional process, involving the transcription of DNA into premature RNA (pre-mRNA), which subsequently undergoes maturation to yield the mature messenger RNA (mRNA).

1.3.1 Transcription

Transcription of protein-coding genes takes place within the nucleus, where RNA polymerase II synthesizes a ribonucleic acid (RNA) molecule (Nikolov and Burley, 1997). RNA polymerase II's attachment is facilitated by the recognition of specific elements such as the TATA box and/or initiator (Inr; Emami *et al.* (1997)). The TATA box is typically positioned approximately 25 bp upstream of the transcription start site (TSS; Fig. 1.2; Lifton *et al.* (1978)).

During elongation, the RNA polymerase moves in a 5' to 3' direction, transcribing the DNA into a single-stranded RNA molecule. In this RNA molecule, thymine (T) is replaced by uracil (U), while the other nucleotides remain unchanged. The primary distinction between deoxyribonucleic acid (DNA) and ribonucleic acid (RNA) lies in the presence of a hydroxyl group (OH) which is a reason for the relative instability of RNA compared to DNA (Ross, 1995; Wang and Kool, 1995; Fordyce *et al.*, 2013).

Termination of transcription occurs when RNA polymerase II reaches a terminator region signal (Birse *et al.*, 1997; Proudfoot, 2016). At this point, transcription is completed, and the RNA molecule dissociates from the DNA. This RNA molecule is referred to as premature RNA.

1.3.2 Maturation

Premature RNA undergoes three distinct pre-processing steps during transcription in the nucleus. The first step involves the addition of a 7-methylguanosine at the 5' end of the pre-mRNA by a guanylyltransferase, called 5'-cap (Fig. 1.2). Subsequently, this 7mG (7-methylguanosine) cap structure is recognized by the cap binding complex (CBC; Gonatopoulos-Pournatzis and Cowling (2014)). The cap structure bound to CBC is believed to play a crucial role in mRNA stabilization (Beelman and Parker, 1995). Additionally, in the cytoplasm, the cap bound to translation initiation factors to promote

translation by facilitating the interaction between the mRNA and ribosomal subunits (see Translation and decoding via the genetic code).

The second maturation step is splicing. In 1977, research by the Sharp and Roberts labs, as well as Chambon lab, revealed that genes are composed of multiple distinct segments along the DNA molecule (Berget *et al.*, 1977; Breathnach *et al.*, 1977; Breathnach and Chambon, 1981; Berk, 2016; Suran, 2020). Indeed, in eukaryotes, most genes are interrupted by introns, which are removed from the pre-mRNA.

Splicing occurs through two transesterification reactions: first, the hydroxyl group of the donor site reacts with the branch point adenosine nucleotide, and then the donor site reacts with the acceptor site, resulting in the excision of the intron and exon–exon ligation. The resulting intron structure forms a 'lariat structure' (Proudfoot *et al.*, 2002).

This process is carried out by a macromolecular complex, the spliceosome, which includes five small nuclear RNAs (snRNAs) known as U1, U2, U4, U5, and U6, as well as approximately 200 proteins (Lerner *et al.*, 1980; Mount and Wolin, 2015). This spliceosome recognizes specific exon/intron boundaries, 5' and 3' splice sites, the donor site marked by the dinucleotide GU (but also GC; Aebi *et al.* (1987)), and the acceptor site marked by the dinucleotide AG (Breathnach *et al.*, 1978). A minor subclass of introns was further discovered, with AU-AC boundaries, excised by a second spliceosome composed of U11, U12, U4atac, U6atac and U5 snRNAs. Based on the spliceosome pathway that takes in charge their removal, introns are categorized U2-type or U12-type (Sharp and Burge, 1997)

In the last phase of transcription maturation a specific sequence of nucleotides known as a poly(A) tail (Fig. 1.2) is appended to the 3' end of the transcribed RNA molecule at the polyadenylation signal (PAS). This tail consists of approximately 250 adenine nucleotides in mammalian cells and is synthesized by an enzyme called poly(A) polymerase (Xiang and Bartel, 2021). The addition of this poly(A) tail plays a crucial role in enhancing the stability of the RNA molecule by enabling its interaction with the poly(A)-binding protein, thereby protecting it from enzymatic degradation (Preiss, 2013; Nicholson and Pasquinelli, 2019; Passmore and Coller, 2022).

Consequently, the resulting RNA molecule is called mature mRNA. Once fully developed, mature mRNA molecules are transported out of the nucleus to undergo translation. mRNA are much less stable than DNA and eventually degrade, which can serve to modulate gene expression over time after gene transcription (Ross, 1995; Wang and Kool, 1995; Fordyce *et al.*, 2013).

1.4 Alternative splicing

In addition to the discovery of RNA splicing, another yet surprising observation was that not only pre-mRNA contains introns that needed to be excised, but also alternative patterns of splicing could lead to different mature messenger RNAs from a same pre-mRNA molecule. The first example was observed in adenovirus where one pre-mRNA molecule could be spliced at different junctions to produce different mature mRNAs,

with different combination of exons, called alternative splicing (AS) (Berget *et al.*, 1977; Berk, 2016). Three years later alternative splicing was found in immunoglobulin genes of mouse myeloma tumors cells (Early *et al.*, 1980).

Different patterns of alternative splicing have then been described such as exon skipping, where an exon is omitted; intron retention, where an intron is retained; 5' and 3' alternatively spliced site, where different splice sites are selected for splicing; and mutually exclusive exons, where only one of two exons is present (Fig. 1.3; Wang and Burge (2008)).

Soon, the scientific community has tended to draw direct conclusions about the 'raison d'être' of alternative splicing. It has been proposed that AS serves as a mean to enhance the functional diversity of species genomes, thereby providing a straightforward explanation for the G-value paradox previously discussed (see Genes: holder of genetic information; Graveley (2001)). Indeed, if the number of genes alone cannot account for the complexity of an organism, the expansion of the protein repertoire by alternative splicing could be the missing part of the equation. Notably, studies have demonstrated that AS is predominant in humans and primates, which are, by some, considered complex species, where more than 90% of multiexonic protein coding genes undergo alternative splicing (Wang *et al.*, 2008; Pan *et al.*, 2008), compared to 35% in the nematode *Caenorhabditis elegans* and 20% in the fly *Drosophila melanogaster* (Chen *et al.*, 2014).

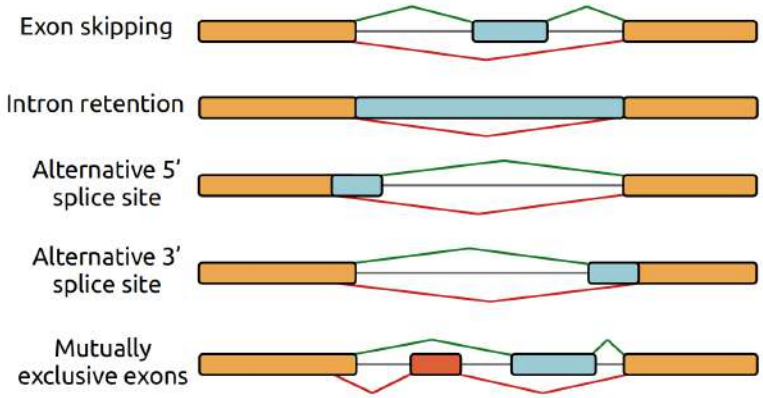


Figure 1.3: Alternative splicing events. *Schematic view of the different pattern of alternative splicing such as exon skipping, where an exon is omitted; intron retention, where an intron is retained; 5' and 3' alternatively spliced site, where different splice sites are selected for splicing; and mutually exclusive exons, where only one of two exons is present. Exons are presented using boxes (unless an intron is retained), introns with lines. Orange exons correspond to constitutive ones.*

Extensive research, observations, and reviews have shown that AS can undoubtedly lead to the generation of distinct functional proteins from a single gene (Blencowe, 2006; Ule and Blencowe, 2019; Wright *et al.*, 2022; Verta and Jacobs, 2022; Singh and Ahi, 2022). Recently, AS has gained significant research focus, particularly in the field of cancer, where its alterations and crucial involvement in tumor proliferation have been extensively investigated (Anczuków and Krainer, 2016; Bonnal *et al.*, 2020; Qi *et al.*, 2020). Thus, AS has emerged as a promising therapeutic target for novel cancer treat-

ments that aim to disrupt oncogenic splicing events or target upstream splicing regulators (Sciarrillo *et al.*, 2020).

However, it is also important to consider that splicing errors by the splicing machinery can contribute to this variability (Pickrell *et al.*, 2010; Rajon and Masel, 2011; Xu and Zhang, 2018; Saudemont *et al.*, 2017). Indeed, the splicing machinery's precision is not infallible, and like any intricate process, it is susceptible to inaccuracies that can lead to the formation of aberrant transcripts. The extent to which erroneous splicing outpaces functional splicing remains a topic of substantial debate within the scientific community (Tress *et al.*, 2017a; Blencowe, 2017; Tress *et al.*, 2017b), especially given the high occurrence of alternative splicing in humans (Pan *et al.*, 2008). My research aims to bring new answers to this ongoing debate, and I explore this topic in the subsequent 'Thesis Objectives' section.

1.5 Translation and decoding via the genetic code

Once the mature mRNA molecule is produced, it is exported to the cytoplasm where it is translated into proteins. This spatial decoupling between transcription/maturation/splicing (in the nucleus) and translation (in the cytoplasm) allows the process of splicing to occur prior to the initiation of protein sequence synthesis.

First, the mRNA is positioned near the endoplasmic reticulum (ER), a cellular structure containing numerous ribosomes (Palade, 1955). These ribosomes are responsible for RNA translation into protein in accordance with the genetic code, which establishes the correspondence between the DNA sequence and the amino acid sequence of a protein (Balis *et al.*, 1958).

Initially, the eukaryotic ribosome's small subunit (40S) and three initiator factor proteins (IF1, IF2, and IF3) forms a preinitiation complex (PIC) to the 5'-cap of mRNAs (Wang *et al.*, 2022). The PIC scans the UTR until it finds a ribosome-binding site (RBS; Kozak (1989)). On this site an initiation complex is formed with the recruitment of the large ribosome subunit (60S; Blanchet and Ranjan (2022)). Then, mRNA sequence is read in triplets of nucleotides known as codons, starting from the initiation codon downstream of the RBS. Each codon corresponds to a transfer RNA (tRNA) molecule. In eukaryotes, tRNA are 90-nucleotide-long molecules produced by RNA polymerase III, and bound to the proper amino acid by aminoacyl-tRNA synthetase enzymes (Sprinzl and Cramer, 1979). tRNAs function by recognizing and binding to the complementary codon through the traditional Watson-Crick base pairing. Additionally, tRNAs undergo post-transcriptional modifications, enabling them to engage in non-Watson-Crick base pairing and unconventional interactions (Percudani, 2001). Notably, adenosine (A) can be modified to inosine (I) through post-transcriptional deamination, allowing for permissive wobble pairing such as I:C, I:U, or I:A (Fig. 1.4). Another common wobble pairing involves G:U/U:G base pairing (Percudani, 2001).

The initiation of translation begins at the start codon 'AUG', where a methionine is attached to the tRNA and positioned on the ribosome (Tamura, 2015). As loaded tRNAs

bind, the ribosome progresses along the mRNA, utilizing three distinct slots: the A-site, P-site, and E-site. The A-site serves as the initial binding site for the tRNA, the P-site holds the tRNA connected to the growing polypeptide chain, and the E-site is the exit site.

While each codon corresponds to an amino acid, there are more codons than there are amino acids. This redundancy in the genetic code implies that multiple codons can encode the same amino acid, called synonymous codons. To study codon usage CU variations between metazoans, GC-content (*i.e.* G+C fraction) at the third position of the codons (GC3) is often used because it is the most synonymous position of the three positions. It has long been observed that synonymous codons usage is uneven across taxa, with a GC3 varying from 20% in some hymenopterans to 60% in mammals and up to 70% in some dipterans. Codon usage also varies within genomes, between genes (Grantham *et al.*, 1980b,a; Gouy and Gautier, 1982; Ikemura, 1985; Parvathy *et al.*, 2022).

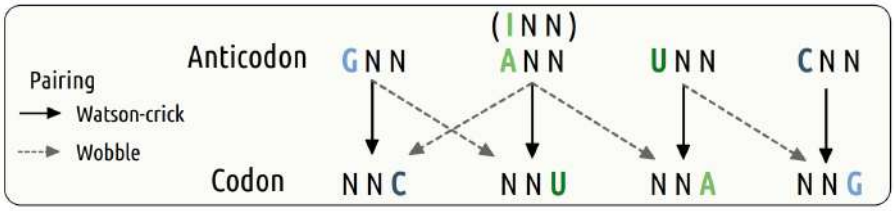


Figure 1.4: Watson-Crick and wobble pairing. *The panel illustrates the various possible pairings: Watson-Crick and wobble pairing (i.e. I:C, I:U, I:A and G:U/U:G).*

Even though modification in the use of synonymous codons do not change the encoded protein, several evidences indicate that they influence gene expression (Hershberg and Petrov, 2008; Plotkin and Kudla, 2011; Martínez *et al.*, 2019; Liu *et al.*, 2021b): transcription regulation in *Neurospora* (Zhou *et al.*, 2016; Zhao *et al.*, 2021) and in human (Fu *et al.*, 2018); translation initiation (Eyre-Walker and Bulmer, 1993; Bhattacharyya *et al.*, 2018; Goodman *et al.*, 2013) and elongation (Sørensen *et al.*, 1989; Boël *et al.*, 2016) in *Escherichia coli*; translation accuracy in *Drosophila melanogaster* (Akashi, 1994), in *Escherichia coli* (Stoletzki and Eyre-Walker, 2007), in yeast, worm, fly, mouse, human (Drummond and Wilke, 2008) and in bacterial taxa (Sun and Zhang, 2022); RNA stability in *Saccharomyces cerevisiae* (Presnyak *et al.*, 2015), *Schizosaccharomyces pombe* (Harigaya and Parker, 2016), *Homo sapiens* (Hia *et al.*, 2019) and in *Escherichia coli* (Kudla *et al.*, 2009); protein folding (Drummond and Wilke, 2008; Buhr *et al.*, 2016; Walsh *et al.*, 2020); RNA splicing in mouse, human (Pagani *et al.*, 2005) and HIV-1 (Takata *et al.*, 2018); RNA toxicity in *Escherichia coli* (Mittal *et al.*, 2018).

Previous studies on model organisms such as *Escherichia coli* (Ikemura, 1981), *Saccharomyces cerevisiae* (Ikemura, 1985) and *Caenorhabditis elegans* (Duret, 2000) have demonstrated that the most frequently used codons are decoded by the most abundant tRNA molecules suggesting a co-adaptation between the tRNA pool and codon usage (Ikemura, 1985). Also, in those species, highly expressed genes, which are likely to be under selection for translation optimization, exhibit a preference for codons associated

with abundant tRNA molecules, which is interpreted as a selection on codons usage (Ikemura, 1981; Percudani *et al.*, 1997; Duret, 2000; Plotkin *et al.*, 2006; Plotkin and Kudla, 2011; Quax *et al.*, 2015). Indeed, to optimize translation the utilization of codons that align with the most prevalent tRNA molecules accelerates the process of locating and attaching ribosome to the appropriate tRNA, thereby diminishing the probability of associating with a non-cognate tRNA (Dana and Tuller, 2014; Quax *et al.*, 2015).

While selection on synonymous codons to optimize translation, called translational selection (TS), has been observed in some model species though rarely in vertebrates (Doherty and McInerney, 2013), we still lack a comprehensive study that embrace numerous metazoans. Also, in human there is still a debate on whether TS is modulating the choice of synonymous codons (Comeron, 2004; Sémon *et al.*, 2006; Doherty and McInerney, 2013; Gingold *et al.*, 2014; Pouyet *et al.*, 2017; Dhindsa *et al.*, 2020). I will provide a more detailed discussion on translational selection in the 'Thesis Objectives' section of my work.

2

The evolutionary forces that shape genomes

Contents

In the preceding section, we delved into the fundamental concepts of genomes, gene structure, and expression, which represent the focal points of investigation in my thesis. The current section aims to provide a description of the evolutionary forces shaping genomes and their nucleotide composition over time, based on several reviews.

We will explore the history of population genetics, investigating the emergence and acceptance of various concepts within the scientific community. Exploring how conceptual ideas and findings emerge over time provides insight and a better understanding of the evolutionary forces that shape genomes architecture, their impact, and how we estimate them today.

Subsequently, we will describe the mechanisms governing DNA changes within individual genomes and populations. Our analysis extends to the intricate processes underlying genetic changes, called mutations.

Lastly, we will delve into the dynamics of DNA variation propagation, specifically the transmission of gene variants, called alleles, within populations. This investigation aims at identifying the primary drivers responsible for the observed evolutionary dynamics, such as the joint product of selection and genetic drift.

2.1 Evolution of Evolutionary biology

I divided the history of evolutionary biology, from its inception to the present day, into four distinct phases spanning the 19th and 21st centuries (Fig. 2.1; Riede (2010)).

2.1.1 Various sub-disciplines (1850-1930)

Jean Baptiste Lamarck was one of the first to introduce the idea of the evolution of life in his book "Philosophie Zoologique" (Lamarck, 1809). Lamarck proposed that individual organisms could modify their organs through use and transmit these changes to their offspring. He illustrated this with the example of giraffes, suggesting that their necks became long because individuals intentionally stretched them to reach leaves. In summary, Lamarck believed that the use of an organ could lead to modifications that could be inherited by the next generation.

Soon after, a young biologist named Charles Darwin set sail on a scientific expedition at the age of 22 aboard the HMS Beagle between 1831 and 1836. During this journey, Darwin observed various species in their natural habitats that eventually led him to publish his renowned book "On the Origin of Species" in 1859, proposing that organisms evolved from a single common ancestor and that these changes were primarily driven by natural selection (see Selection section; Darwin *et al.* (1859)). According to Darwin, spontaneous morphological or physiological variations arise between offspring, making some individual better adapted to their environment. Such individuals are more likely to survive and transmit their advantageous traits to subsequent generations (Nei, 2005).

Contrary to Lamarck, Darwin argued that the elongation of a giraffe's neck was not due to the individual stretching its neck, but rather the result of giraffes with longer necks having a survival advantage over those with shorter necks in reaching food and passing on this trait to future generations through natural selection. Additionally, in 1868, Darwin proposed his theory of heredity through "pangenesis", suggesting that each part of an organism's body emitted small organic particles called gemmules. These gemmules would aggregate in the gonads and contribute heritable information to the gametes (Darwin, 1868). It is now known that this theory was inaccurate, but it is interesting to note that Darwin's pangenesis allowed for the possibility of Lamarckian transmission of acquired characteristics, making him somewhat supportive of Lamarck's ideas while emphasizing the role of natural selection as the main driver of evolutionary changes (Kováč, 2019). Currently, the term "Lamarckism" is often used to refer to the inheritance of acquired characteristics, such as epigenetic inheritance, which remains a topic of controversy in modern evolutionary biology (Burkhardt, 2013).

During the late 19th and early 20th centuries, significant advances were made not only through observations, but also through the application of rigorous protocols and experiments. Charles Darwin, besides his observations, conducted a series of smaller experiments to investigate how species appeared in different geographic locations. His investigations included studying seed germination conditions and the phenomenon of

snails adhering to ducks' feet. However, it was Gregor Johann Mendel, a monk, who introduced the concept of protocols in 1856 by initiating a series of plant hybridization experiments in the monastery gardens. Mendel's goal was to explore how phenotypical traits are inherited across generations. Over a period of seven years, between 1856 and 1863, Mendel cultivated and tested approximately 28,000 plants, with a majority of them being pea plants. Mendel's rigorous plant hybridization experiments and his meticulous observations laid the groundwork for the field of genetics.

Indeed, in 1865, Mendel published a paper in which he proposed that phenotypic traits are passed down to offspring according to the notion of dominant and recessive traits. He also elucidated the phenotypic ratio (9:3:3:1) observed in dihybrid crosses for heterozygous organisms, which is now recognized as the basic law of independent assortment and the law of dominance (Mendel, 1865). Remarkably, Mendel was familiar with Darwin's work, and historians agree that he accepted Darwin's proposition. In a letter, Mendel even proposed a Darwinian scenario for natural selection, using the same German term for "struggle for existence" as found in his copies of Darwin's books (Fairbanks, 2020; Berry and Browne, 2022).

Mendel's groundbreaking work, however, was not fully appreciated until much later. Indeed, it wasn't until the early 1900s that three biologists, namely Hugo de Vries, Carl Correns, and Erich von Tschermak-Seysenegg, independently rediscovered Mendel's laws, giving due recognition to his significant contributions to the understanding of genetics (Keynes and Cox, 2008). This era witnessed remarkable progress in comprehending heredity and evolution, mainly due to the pioneering efforts of eminent scientists. Furthermore, Hugo de Vries, through his experimental study of new varieties of evening primrose in his experimental garden, proposed the "macromutation theory", suggesting that spontaneous variations could occur, leading to new traits in organisms (Vries, 1901).

Then, in 1909, Thomas Morgan worked on variants of Drosophila, particularly focusing on phenotypic variations in their eyes. Initially skeptical of Mendel's work and Darwin's theory, Morgan's investigations confirmed Mendel's observations in the fruit fly *Drosophila melanogaster*, leading to the publication of his seminal work "The Mechanism of Mendelian Heredity" in 1915 (Morgan, 1915). Morgan's findings further led him to suggest that natural selection acts as a mechanism to preserve advantageous variations (mutations) while eliminating deleterious ones. However, he argued that the occurrence of advantageous mutations is the primary driver of evolution, meaning that neutral or deleterious ones play a secondary role (Morgan, 1925; Allen, 1968).

Around the same time, R.A. Fisher authored a book, exploring how morphological variations (mutations) can be passed down through generations based on their impact on the ability to produce offspring (fitness) and the number of individuals in the population (population size), using mathematical modeling (Fisher, 1922; Crow, 2002; Abanda and Xavier, 2012; Charlesworth, 2022).

To gain a comprehensive understanding of evolutionary biology, scientists have faced the challenge of reconciling diverse disciplines, such as evolutionary concepts, biological observations, botanical experiments, and population genetics. The integration of these

diverse fields was necessary to better understand the complexity of evolutionary processes.

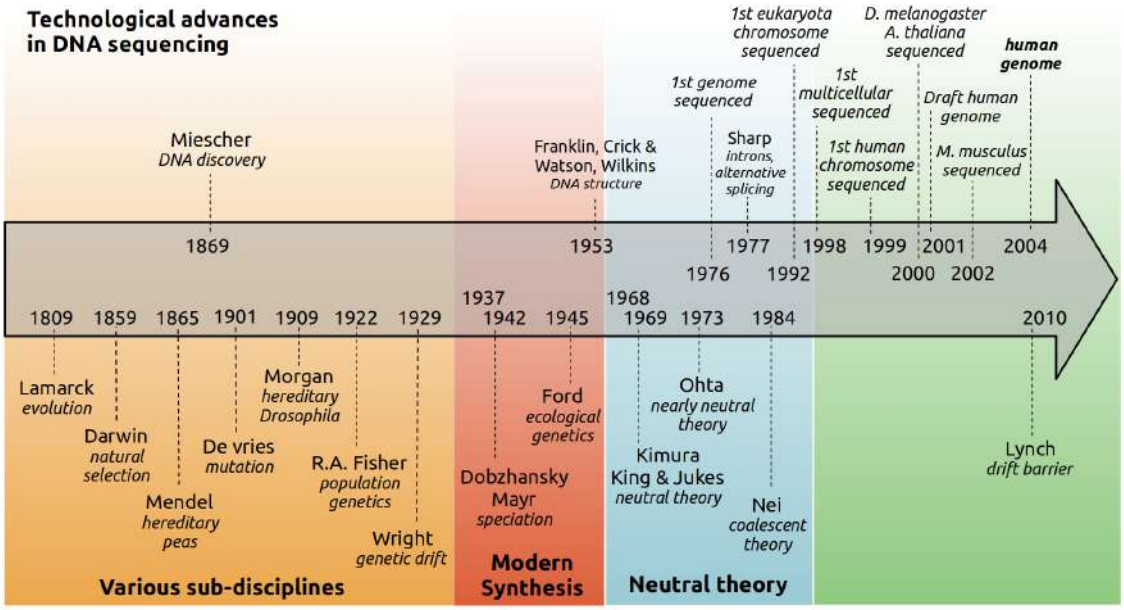


Figure 2.1: Chronology of Evolutionary biology. *Chronology of the different discoveries in evolutionary biology delimiting different periods. Above the timeline are key dates for technological advances associated with DNA sequencing. Meanwhile, the lower section presents key discoveries in the field of evolutionary biology.*

2.1.2 Modern synthesis (1930-1947)

The "modern synthesis" refers to the formulation of evolutionary theory during the early to mid-20th century, which aimed to reconcile classical Darwinian selection theory with the emerging population-oriented perspective of Mendelian genetics, seeking to elucidate the origin of biological diversity. This period witnessed the confluence of various disciplines, resulting in significant contributions and advancements in evolutionary biology.

Various scientists made valuable contributions to this synthesis. Theodosius Dobzhansky, a postdoctoral researcher in Morgan's fruit fly lab, pioneered the application of genetics to natural populations through experimental studies, primarily focusing on *Drosophila pseudoobscura*. In his seminal book, "Genetics and the Origin of Species" published in 1937, Dobzhansky proposed an explanation for the emergence of new species based on the theoretical developments of natural selection as a genetic process: natural mutations constantly arise within populations (Dobzhansky, 1937; Ayala and Fitch, 1997; Barahona and Ayala, 2005). While some mutations can be detrimental under specific conditions, a remarkable portion of these genetic changes have no discernible impact on the organisms' fitness. These neutral mutations persist in different populations and contribute to an unexpectedly vast level of genetic variability, surpassing previous scientific expectations. Dobzhansky's work highlighted the prevalence of neutral mutations in populations. This integration of population genetics with experimental evidence has played a pivotal role in reconciling theoretical concepts with real-world observations.

Fisher, in his seminal work "The Genetical Theory of Natural Selection" (Fisher, 1930; Leigh, 1999), demonstrated mathematically how Mendelian genetics could be reconciled with the concept of evolution through natural selection. Also, E.B. Ford, a pioneering experimental naturalist, played a crucial role in the development of ecological genetics as a scientific discipline. By conducting innovative experiments in nature, he aimed to validate the principles of natural selection. Ford's groundbreaking research focused on wild populations of butterflies and moths, and through close collaboration with R.A. Fisher, he successfully confirmed Fisher's predictions (Fisher and Ford, 1947; Ford, 1945; Baxter *et al.*, 2017).

During this time, Sewall Wright was credited with introducing the term "drift" in his work (Wright, 1929). Initially, he used it as *"the results of a directed process, selection"* but later clarified its definition as "Random drift" (Wright, 1970). The concept of genetic drift acts as a counterbalance to the effect of selection, wherein the chance of a particular variant spreading depends on its selection coefficient and the size of the population (N) (see Genetic drift section; Wright (1931, 1932)). The contributions of both Wright and Fisher laid the foundation for the development of population genetics, where mathematical equations were used to establish connections between natural selection and Mendelian genetics.

An intriguing and contentious scientific debate arose between Fisher and Wright. Despite using different methodologies, their theoretical conclusions for a given problem were congruent. Their discrepancies lay in their interpretations rather than in the mathematical aspects. The focal point of the debate revolved around the significance of genetic drift, which Wright referred to as "random sampling" at that time (Wright, 1951). Ford and Fisher examined color polymorphism frequencies in a population and argued that the population size was too large for these frequency changes to be attributed to drift. Consequently, they suggested that fluctuating selection must be the driving force. As a result, they posited that genetic drift would have minimal impact on phenotypic traits in the vast majority of natural populations (Fisher, 1950; O'Hara, 2005), whereas Wright maintained that fluctuations in population size could offer the greatest chance for evolutionary novelty and significantly accelerate evolution (Bacaër, 2011).

2.1.3 Neutral and Nearly neutral theory (1966-1990)

A significant figure in the fields of evolutionary biology and population genetics during the period from 1966 to 1990 was Motoo Kimura. It was during this era that the genetic support, DNA, was initially uncovered. In 1969, Kimura introduced the neutral theory of molecular evolution, which suggests that the majority of evolutionary changes within and between species are primarily driven by random genetic drift of mutant alleles that have no significant impact on an organism's fitness. This theory posits that the vast majority of mutations are not influenced by natural selection. While earlier scientists, such as Fisher, had mathematically derived aspects of neutral mutation theory, they considered it to be rare (Fisher, 1931). However, Kimura was the first to present

a coherent theory of neutral evolution in 1968, which was independently proposed by King and Jukes in 1969, with a focus on differences within species rather than among species (Kimura, 1968; King and Jukes, 1969).

Kimura and Jukes proposed the non-adaptive theory, suggesting that the majority of mutations are neutral, leading to high substitution rates compared to what one would expect under purifying selective pressures. Therefore, if these modifications persist, it is because they are either neutral or nearly neutral. Over time, as our understanding of genetics advanced, Kimura was able to find evidence supporting his neutral theory of evolution (Kimura, 1991). Notably, he focused on DNA positions that do not affect the translated amino acid due to genetic code redundancy (synonymous positions). His observations revealed that these positions exhibit as much variation as expected under the neutral theory, thereby reinforcing the idea proposed by King and Jukes in 1969 (Kimura, 1977). Additionally, Kimura investigated enzyme (protein) diversity within populations, or polymorphisms, in Drosophila and humans, where multiple forms of the enzyme can coexist, as previously demonstrated by the molecular biologist Richard Lewontin (Lewontin and Hubby, 1966; Harris, 1966; Charlesworth *et al.*, 2016). Kimura suggested that most of these forms are selectively neutral, explaining the observed level of heterozygosity through neutral variations.

This gave rise to a long-standing debate between neutralists and selectionists (Neo-Darwinian proponents). While neo-Darwinian scientists firmly believed that species variations are primarily shaped by natural selection, Kimura opposed this view with his neutral theory. Selectionists justified the high genetic diversity within species by claiming that these polymorphisms are maintained by balancing selection, the fact that different alleles of a genes are effectively maintained by natural selection. Whereas neutralists considered that these variations are simply due to neutral changes, that does not necessarily affect the fitness of an individual (Kimura and Ohta, 1971; Nei, 2005; Lee *et al.*, 2021).

After working and collaborating as a postdoc under Kimura on the neutral theory of evolution, Tomoko Ohta came to the conclusion that the classification of mutations into good, neutral, and harmful was an overly simplistic model insufficient to account for the observed data. Ohta emphasized the significance of nearly neutral mutations, particularly slightly deleterious ones (Ohta, 1973). The dynamics of nearly neutral mutations closely resemble those of neutral mutations unless the selection coefficient's absolute magnitude exceeds the inverse of the number of individuals (population size). Consequently, the population size can influence the number of mutations considered neutral or deleterious (Ohta, 1992).

Another significant contributor to the neutral theory is Masatoshi Nei, who made predictions about the existence of a considerable number of duplicate genes and pseudogenes in organisms based on amino acid substitution rates, gene duplication, and gene inactivation (Nei, 1969, 1984). During the 1960s and 1970s, there was considerable controversy surrounding protein evolution mechanisms and the maintenance of protein diversity. Analysess by Nei and his collaborators of alleles frequency distribution and the relationship between average heterozygosity and protein divergence between species

supported that a substantial portion of protein polymorphism can be explained by the neutral theory (Nei *et al.*, 2010; Zhang and Kumar, 2023).

2.1.4 Drift barrier hypothesis (2010)

Following the nearly neutral theory, Michael Lynch formulated the "drift barrier" hypothesis (Lynch, 2007a, 2010), proposing that the optimization of a trait through natural selection in a specific environment will encounter a theoretical limit. As a trait approaches perfection, the fitness gain of beneficial mutation decreases, and at this barrier, the impact of further beneficial mutations becomes limited in overcoming the influence of random genetic drift. Consequently, each species or genome reaches an equilibrium between the effectiveness of natural selection in promoting advantageous traits and the stochastic effects introduced by random genetic drift. The equilibrium state of a trait is determined by the interplay of both forces, implying that species with smaller population sizes tend to accumulate a genetic burden and be less optimized compared to species with larger population sizes (Fig. 2.2). Therefore, beneficial traits are expected to improve with population size, while deleterious traits should reduce. In practical applications, the effective population size (N_e) is employed rather than the total census population size. The effective population size pertains to a normalized population with stable spatial distribution, sex ratio, and is a measure of the genetic drift intensity as detailed in the subsequent section of this manuscript (see Genetic drift for more details).

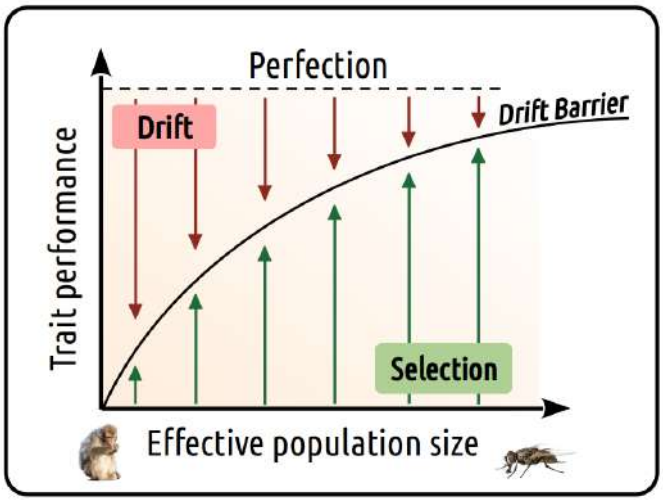


Figure 2.2: The "drift barrier" hypothesis. *Graphic illustrating the "drift barrier" hypothesis according to which the interplay between genetic drift and selection stabilizes a trait performance. Larger population sizes species tend to have better optimized traits compared to smaller population sizes species.*

Lynch presented a compelling argument using mutation rates as an exemple (*i.e.* the number of mutation *per* bp *per* generation) which exhibit considerable variation across species. Selection tends to favor lower mutation rates due to its associated burden of deleterious mutations (Kimura, 1967; Lynch, 2008), and since the power of drift is

inversely proportional to N_e, species with larger N_e are expected to have lower mutation rates as observed in Sung *et al.* (2012).

Furthermore, Michael Lynch proposed that variation in the ability to purge slightly deleterious mutations (*i.e.* variation in N_e) can account for differences in genome architecture among species, such as genome size (Lynch and Conery, 2003; Lefébure *et al.*, 2017; Mérel *et al.*, 2024).

2.2 Emergence of new alleles (mutation)

The mutation theory was first developed by Hugo De Vries in 1901 (Vries, 1901; Allen, 1969) when he studied a group of evening primrose, *Oenothera lamarckiana*, where he found that seeds from this plant produced many new varieties in his experimental garden (13 years of experiments). De Vries labeled these sudden and novel variations as "mutations". However, it is important to note a distinction in terminology from De Vries' work: while he broadly characterized any heritable changes in phenotypic traits as mutations, our focus here is on mutations at the DNA level. These modifications are the primary drivers of genetic diversity, ultimately contributing to morphological alterations (Nei and Nozawa, 2011).

Mutations can be classified into four types: base replacements, involving the replacement of one nucleotide with another; small insertions, which involve the addition of one or several nucleotides; small deletions, resulting in the loss of one or several nucleotides; and larger forms of chromosome structural variations (deletions, inversion, translocation or duplication).

If mutations occur within a coding sequence, their impact on the protein product can lead to three different subcategories of mutations: non-synonymous mutations, where the modification results in a change in the protein product and may even render the protein non-functional; nonsense mutations, in which the mutation causes premature termination due to the formation of a stop codon, mostly leading to an aberrant transcript; and synonymous mutations, occurring when the mutation does not affect the sequence of the protein product (Zia and Moses, 2011; Potapova, 2022).

2.2.1 Insertion-deletion

Small insertion-deletion (Indels) are widely distributed throughout the genome and contribute to both intra and inter species divergence (McGee *et al.*, 2020). These mutations are extensively studied in human genome (Weber *et al.*, 2002; Bhangale *et al.*, 2005; Conrad *et al.*, 2006). Indels can occur during DNA replication, when the strand that is replicated slips, this can lead to the incorporation or deletion of nucleotides. Also, sequences such as transposable elements possess the particularity to replicate within the genome akin to genome parasites, which induce an insertion at other part of the genome (Cai *et al.*, 2022; McClintock, 1950).

A substantial majority of these small Indels (96%) range in length from 1 to 16 bp,

with the largest observed Indel spanning 55 bp, as reported by Mullaney *et al.* (2010). Notably, Indels overlapping coding sequence frequently induce a shift in the reading frame, bearing a high probability of introducing stop codons, thus often resulting in nonsense mutations.

2.2.2 Base replacement

A base replacement occurs when a nucleotide is replaced by another in the genome. Such replacements are categorized as: transitions, *i.e.* involving the exchange between two purines or between two pyrimidines; or transversions, *i.e.* the base replacement between a purine base and a pyrimidine base, or vice versa.

Spontaneous mutations can be caused by external factors such as ionizing radiation, ultraviolet rays, and mutagenic chemicals. These extrinsic agents have the capacity to cause DNA damage, inducing inaccuracies during DNA replication or repair, thus giving rise to mutations (Maki, 2002). Additionally, in the context of dinucleotides CG referred as CpG dinucleotides (Cytosyne Phosphate Guanine), following methylation, C-to-U deamination can happen. In mammals the frequent DNA methylation, linked with gene expression regulation, makes CpGs highly mutable. This phenomenon explains the lower prevalence of CpG dinucleotides in mammals compared to other species (Duncan and Miller, 1980; Brennan *et al.*, 1990).

Mechanisms exist for the repair of these mutations, involving the comparison of the two DNA strands to identify discrepancies. However, post cell division, the repair machinery becomes challenged, as the newly replicated strands are identical correction mutations is hampered (Gao *et al.*, 2017; Cortez, 2019).

If a base replacement occurs within a coding sequence, it will lead to a modification in the codon, yet may not necessarily impact the resulting protein due to the redundancy of the genetic code. It is important to note that not all non-synonymous mutations result in an altered protein function. Specifically, if these base replacements occur outside the protein's active site, they might not significantly affect the protein's function, even if the peptide sequence has been altered. An illustrative example is the mutation that converts a leucine (Leu) codon to an isoleucine (Ile) codon (Sneath, 1966; Miyata *et al.*, 1979; Epstein, 1967). Given the chemical similarity between these two amino acids and their potential interchangeability in proteins, such a mutation might not exert a substantial influence on the protein's structure or function, resulting in what is known as a neutral mutation. Likewise, not all synonymous mutations are inherently neutral. Indeed, the composition of codons can influence translation kinetics, thereby impacting the efficiency of protein synthesis and proper folding (Akashi, 1994; Stoletzki and Eyre-Walker, 2007; Drummond and Wilke, 2008; Plotkin and Kudla, 2011; Yang *et al.*, 2014; Dana and Tuller, 2014; Gorochowski *et al.*, 2015; Quax *et al.*, 2015; Presnyak *et al.*, 2015; Wu *et al.*, 2019).

To calculate the mutation rate *per* base pair *per*, one can enumerate the number of replacements occurring within a particular population over a specific generation. This approach facilitates the assessment of the frequency and pace at which new alleles can

be introduced into a population, consequently contributing to genetic diversity. Mutation rates exhibit significant variation across taxonomic groups and within different genomic regions of a single organism, averaging around 12 x 10^{-9} mutation *per* bp *per* generation in mammalian genomes (Kumar and Subramanian, 2002; Lynch, 2010; Bergeron *et al.*, 2023).

Germ-line mutations arise within gametes or cells that ultimately give rise to gametes. In contrast to somatic mutations, germ-line mutations are heritable and passed down to subsequent generations. Consequently, these mutations contribute to form different version of genes (alleles), and are present in all cell types of future organisms, thus have the potential to propagate within a population.

2.3 The fate of new alleles

alleles in a population are carried by individual genomes and propagate through generations at varying frequency, influenced by diverse forces: selection, genetic drift and biased gene conversion. A new allele within a population reaches fixation when possessed by all individuals, this mutation is then called a substitution.

2.3.1 Selection

The concept of natural selection, introduced by Darwin in is seminal work, underscores the process by which species evolve and adapt. This evolution involves the accumulation of mutations that enhance individual fitness (w). The fitness of a genotype or phenotype is gauged by its ability to produce offspring. This estimation requires to evaluate the reproductive rate, defined as the average number of offspring produced *per* individual, and the survival rate, which represents the percentage of born individuals that reach reproductive maturity. Consequently, for each genotype, the product of the reproductive rate and survival rate is calculated. The relative fitness of each genotype is derived from this product relative to a reference genotype. Fitness values are ≥ 0, with 0 signifying that there is no viable descendant.

Expressed as the selection coefficient (s), the relationship between fitness and selection is given by $s = w - 1$, a measure of the relative strength of selection acting against a genotype. If s assumes a negative value, the allele is deemed deleterious and subjected to counter-selection (or purifying selection). Conversely, a positive s indicates a beneficial allele (Rédei, 2008; Coop, 2020; Akashi, 1999).

The force of natural selection stands as a pivotal determinant in the destiny of novel alleles within populations. It drives the survival and reproductive success of individuals harboring advantageous alleles, thus causing a progressive increase in their prevalence over successive generations. Conversely, alleles that are detrimental to survival or reproductive fitness are either purged or maintained at reduced frequencies.

The effect of counter-selection seems to be amplified in highly expressed genes. Indeed, those genes accumulate non-synonymous substitutions, weakly deleterious, at a

slower rate than less expressed genes (Duret and Mouchiroud, 2000; Rocha and Danchin, 2004; Pagán *et al.*, 2012; Brion *et al.*, 2015) and are more conserved across species (Pál *et al.*, 2001; Geiler-Samerotte *et al.*, 2011; Zhang and Yang, 2015). One hypothesis is that misfolded proteins are toxic in cells and selection act to diminish this toxicity (Yang *et al.*, 2012; Park *et al.*, 2013; Wu *et al.*, 2022; Trucchi *et al.*, 2023). An other hypothesis is based on the fitness cost of deleterious mutations linked with the unnecessary mobilization of metabolic resources and cellular machinery. This fitness cost is more important in highly expressed genes that solicit a lot of resources compared to lowly expressed genes. Consequently, these highly expressed genes tend to be subject to more intense purifying selection, resulting in a more pronounced elimination of deleterious alleles compared to genes with lower expression levels (Saudemont *et al.*, 2017; Nabholz *et al.*, 2012).

2.3.2 Genetic drift

As previously mentioned, genetic drift, a concept introduced by Wright, refers to stochastic fluctuations in allele frequencies within a population across successive generations.

These fluctuations arise due to the inherently random sampling of individuals that reproduce and pass on their alleles to subsequent generations. Notably, the impact of genetic drift is more pronounced in populations of smaller size, where random variations exert a more substantial influence. The intensity of random genetic drift is measured through the concept of the effective population size (N_e). This parameter represents the hypothetical number of individuals within a Hardy-Weinberg population that would yield equivalent patterns of random fluctuations at neutral sites (Husemann *et al.*, 2016; Wang *et al.*, 2016). In a Hardy-Weinberg population, mating occurs randomly (panmixia), as do encounters between gametes (pangamy). Generations do not overlap, implying that individuals from distinct generations cannot reproduce together. Furthermore, no natural selection, mutation, or migration factors are at play within this context (Hardy, 1908; Wright, 1931; Stern, 1943; Felsenstein, 1971; Edwards, 2008).

N_e intervenes in population genetic equations, where the rate of fixation (K) for newly introduced allele within a diploïd population of N_e individuals (*i.e.* allele frequency $p = \frac{1}{2N_e}$) is given by: the number of new mutations in each generation multiplied by the probability of reaching fixation (P^F).

$$K = 2N_e\ \mu\ P^F = 2N_e\ \mu\ \frac{1 - e^{-4N_e sp}}{1 - e^{-4N_e s}} = 2N_e\ \mu\ \frac{1 - e^{-2s}}{1 - e^{-4N_e s}}$$

where s represents the selection coefficient of the allele and μ the mutation rate *per* base pair *per* generation.

For neutral alleles where $s \to 0$, then $K = \mu$.

Thus, for small selection coefficients, when N_e decreases K approaches μ , and alleles behave as if they are effectively neutral (Fig. 2.3).

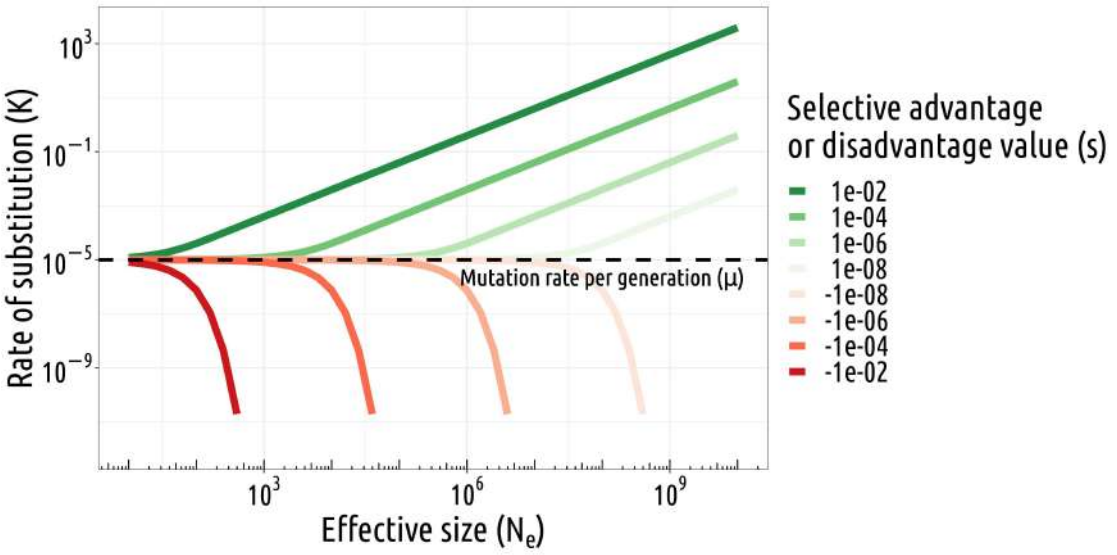


Figure 2.3: Substitution rate for slightly deleterious and advantageous alleles. *In this simulation, based on equations by Lynch (2007b) we observe that for alleles with a slight selection coefficient ($|s| \ll 1$) as the product of effective population size (N_e) and selection coefficient (s) decreases, the fixation rate (K of a particular allele converges towards the neutral rate; Lynch (2007b)).*

In fact, the parameter that matters in determining the ability of selection to promote beneficial mutations or eliminate deleterious mutations is the intensity of selection (s) relative to the power of random genetic drift (N_e), called the population-scaled selection coefficient $S = |4N_e s|$ (Fig. 2.4). If $S \gg 1$ variations in N_e won't affect the fixation probability and selection will be the main force determining the fate of alleles. If the selection coefficient is sufficiently weak relative to drift ($S \ll 1$), alleles behave as if they are effectively neutral leaving no grounds for selection. Between both extremes, changes in N_e will affect the rate of substitutions.

In consequences random genetic drift impact the efficiency of selection in promoting slightly advantageous alleles while suppressing slightly deleterious ones within populations (Fig. 2.5). In low effective population size, slightly deleterious alleles can reach fixation due to the strong stochasticity of genetic drift hindering the effect of purifying selection. This led Lynch to propose the "drift barrier" hypothesis where drift limit the genome optimization by overwhelming the selection (Lynch, 2007a, 2010) (Fig. 2.2).

N_e is generally lower than the census population size (N) (Palstra and Ruzzante, 2008; Palstra and Fraser, 2012). Indeed, variations in population size, difference in sex ratio, and spatial distribution are factors contributing to increase the random drift compare to a Hardy-Weinberg population of the same size (Waples, 2002, 2016). Thus, N_e cannot be estimated by simply counting the number of individuals in a population.

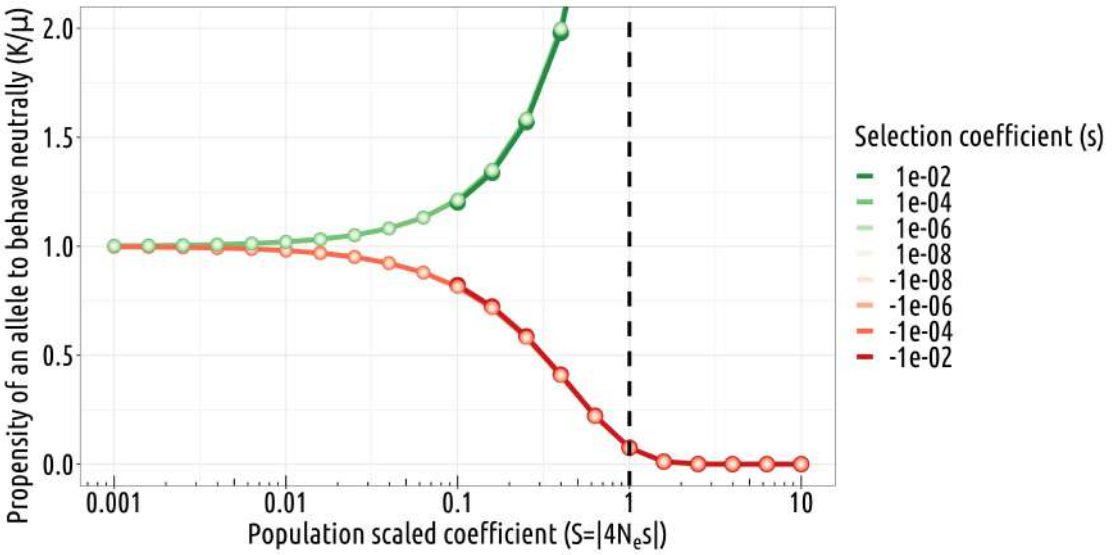


Figure 2.4: Population-scaled selection coefficient impact on allele behavior. *Relation between the population-scaled selection coefficient $S = |4N_e s|$ and the propensity of an allele to be neutral, the ratio of K the substitution rate, and the mutation rate per base pair per generation μ.*

To estimate N_e one can directly measure the intensity of drift by quantifying the allele frequency variations through generations at neutral sites. Because in nature this approach needs a lot of resource other estimates have been proposed from population genetic equations. One such is the genetic diversity. At mutation-drift equilibrium the effective population size (N_e) is directly measured by the degree of genetic diversity (average nucleotide heterozygosity at synonymous sites, π_s) within a diploïd population expected to be equal to $\approx 4N_\mathrm{e}\mu$, μ being the mutation rate *per* base pair *per* generation. Using this equation, estimates of effective population size have been calculated by estimating π_s and μ with base-substitutional mutation rate/site/cell division in several species (Sung *et al.*, 2012; Lynch *et al.*, 2023). With this method, humans have been shown to have a relatively low $N_\mathrm{e} \approx 10^4$, compared to *Carnorhabditis* with 10^7 and Eubacteria reaching 10^8.

Also, random linkage disequilibrium, which measures the dependence between two neutral alleles at different loci, is theoretically associated to N_e and can be use as an estimator (Waples, 2024). Indeed population recombination rate depends on drift intensity as $\rho = 4N_\mathrm{e}r$, where r is the *per*-generation recombination rate (Waples and Do, 2010).

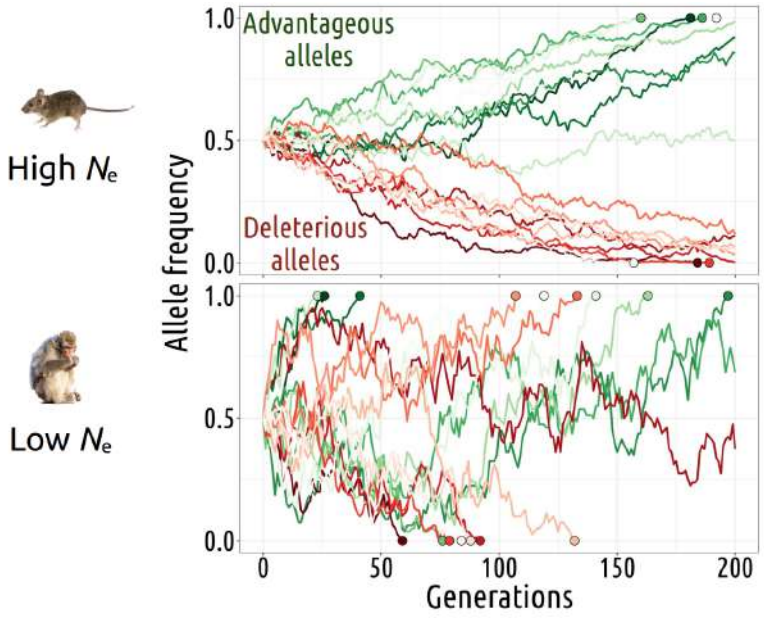


Figure 2.5: Genetic drift and its impact on allele fixations. *Selection tends to promote advantageous alleles while suppressing deleterious ones. But the efficacy of this evolutionary force diminishes with increasing genetic drift intensity, associated with a reduced effective population size. The presented simulations illustrate the variations in frequencies of slightly advantageous (green curves) or deleterious alleles (red curves) across generations, comparing populations with a high effective population size (top) to those with a low effective population size (bottom).*

Genetic diversity and population recombination rate are two measures of short time scale N_e (short-term N_e), which may not reflect the N_e that affected the genome evolution on large time scale (long-term N_e). An additional means of approximating N_e involves the assessment of the magnitude of purifying selection acting on protein sequences, as indicated by the ratio dN/dS (Kryazhimskiy and Plotkin, 2008). The underlying hypothesis is that the rate of synonymous substitutions (dS) quantifies the rate of neutral allele substitutions, while the rate of non-synonymous substitutions (dN) reflects the rate of deleterious allele substitutions (Fig. 2.6). The ratio dN/dS represents the efficacy of selection in suppressing deleterious alleles relative to the influence of genetic drift.

Indeed as mentioned earlier, the probability of fixation for a specific allele with a frequency p within a diploïd population of N_e individuals is given by

$$P^F(p) = \frac{1 - e^{-4N_e sp}}{1 - e^{-4N_e s}}$$

In the case of a diploïd species, a newly introduced allele is expected to have a frequency $p = \frac{1}{2N_e}$, leading to a simplified equation:

$$P^F(\frac{1}{2N_e}) = \frac{1 - e^{-2s}}{1 - e^{-4N_e s}}$$

For small s values ($s \ll 1$), further simplification yields

$$P^F\left(\frac{1}{2N_e}\right) \approx \frac{2s}{1-e^{-4N_e s}}$$

The following equations describe the change in the number of synonymous (dS) and non-synonymous (dN) mutations over generations:

$$dN = \mu \cdot \text{generations} \cdot P^F_{\text{non-synonymous}}$$

$$dS = \mu \cdot \text{generations} \cdot P^F_{\text{synonymous}}$$

where $P^F_{\text{synonymous}}$ indicates fixation probability of nearly neutral alleles because synonymous substitutions $s \to 0$, hence $4N_e s \to 0$.

Thus leading to the simplification $P^F_{\text{synonymous}} = \frac{2s}{4N_e s} = \frac{1}{2N_e}$ (Kimura, 1962), the case where allele behave as neutral.

For non-synonymous substitutions:

$$P^F_{\text{non-synonymous}} \approx \frac{2s_{\text{non-synonymous}}}{1-e^{-4N_e s_{\text{non-synonymous}}}}$$

The ratio dN/dS becomes:

$$\frac{dN}{dS} \approx \frac{4N_e s_{\text{non-synonymous}}}{1-e^{-4N_e s_{\text{non-synonymous}}}}$$

dN/dS appears as a function of $4N_e s$ (Nielsen and Yang, 2003), and if we assumed non-synonymous substitution mostly deleterious because of its impact on protein sequences: $s_{\text{non-synonymous}} < 0$ stable between species, dN/dS is negatively correlated with N_e, provided it is in the range where the population-scaled selection coefficient allows it.

The dN/dS ratio serves as a valuable tool for comparing selective pressures among genes. For example, highly expressed genes tend to have low dN/dS values, indicating greater constraint, in contrast to genes with lower expression levels (Duret and Mouchiroud, 2000; Rocha and Danchin, 2004; Pagán *et al.*, 2012; Brion *et al.*, 2015).

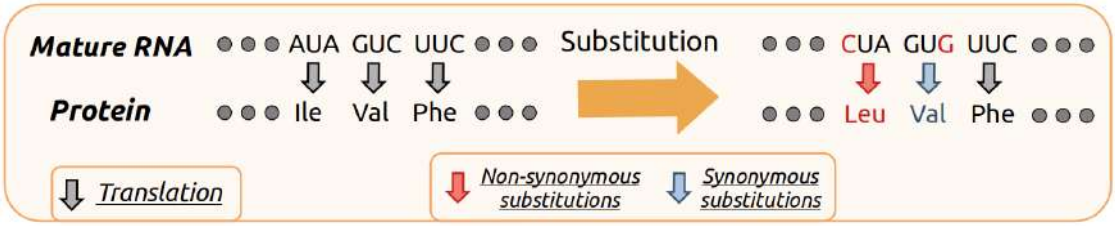


Figure 2.6: Non-synonymous and synonymous substitutions. *Example of one non-synonymous substitution (red) and one synonymous substitution (blue) in a partial sequence of three codons. For the non-synonymous substitution the isoleucine amino acid is replaced by leucine due to a substitution from adenine to cytosine at the first position of the codon. For the synonymous substitution the cytosine at the third position is replaced by a guanine which does not change the translated amino acid.*

Also, various proxies can be employed to examine the relative fluctuations in ef-

fective population size (N_e). One such factor is the examination of life history traits such as longevity, body weight, and body length. This is based on the concept that larger organisms tend to have a reduced number of individuals within their ecological niche, thus impacting the overall population size. And if other parameters such as sex ratio, reproductive mode, spatial distributions are not altered, life history traits variations are expected to be linked to N_e (Waples, 2016; Figuet *et al.*, 2016; Galtier, 2016; Weyna and Romiguier, 2020).

Furthermore, the reproductive system contributes to altering effective population size. For instance, in eusocial species, most individuals are sterile, and only a limited number of female (queens) and males engages in reproduction and transmits their genetic heritage. Consequently, this process reduces the effective population size and subsequently intensifies genetic drift, especially when compared to an equivalent number of individuals within solitary species (Romiguier *et al.*, 2014b).

2.3.3 Biased gene conversion

The last evolutionary force developed in this section is the biased gene conversion. During meiosis, a diploïd cell, characterized by the presence of two homologous chromosomes, undergoes division to produce haploid cells. The pairing of these chromosomes is called genetic recombination, by which genetic exchanges occur between the homologous chromosomes, significantly contributing to the maintenance of genetic diversity within populations. This processes was unveiled by Thomas Morgan's work, but was previously observed in Mendel's plant hybridization.

During recombination, chromosomes harbor a heteroduplex region, wherein the two DNA strands do not possess identical sequences, because they originate from each homologous chromosome. Repair mechanisms are invoked to ensure nucleotide homogeneity on both DNA strands. In many metazoans, this repair process exhibits a noteworthy preference for guanine-cytosine (GC) content, as documented by previous studies (Duret and Arndt, 2008; Duret and Galtier, 2009; Romiguier *et al.*, 2010). This preference for repairing towards GC, when faced with the choice between adenine-thymine (AT) or GC, subsequently leads to an asymmetrical propagation of GC alleles, thereby impacting the distribution of genetic variants. This process, known as GC-biased gene conversion (gBGC), is an evolutionary force similar to selection in the sense that it promotes GC alleles over AT alleles. However, this phenomenon of GC-biased gene conversion is not universally observed across species (Galtier, 2021; Mugal, 2021). Thus, the GC landscape of genomes is highly affected by gBGC in species where this process is observed. Which can ultimately affect the composition of the coding sequences. Furthermore, this force does not necessarily tend to increase the fitness of individuals, and can even go against it.

3

Bioinformatics and Genetics provide a robust framework for investigating genomes evolution

Contents

Since Jean Baptiste Lamarck in 1809 (Lamarck, 1809) molecular evolutionary biology became a scientific discipline regrouping researchers and expanding its results beyond empirical observations and theoretical postulations to encompass protocols, experiments, and mathematical models with genetic as the raw material (see 'The evolutionary forces that shape genomes' chapter). This discipline is dedicated to study the genetic variations among species and within populations with the idea that all species have a shared ancestor. Central to evolutionary biology is the study of evolutionary forces that shape the characteristics of species and populations. Through rigorous examination of these factors, evolutionary biology seeks to gain profound insights into the dynamic processes driving the evolution of life on Earth.

Notably, evolutionary biology has made substantial contributions to diverse domains such as unraveling our historical origins (Cann *et al.*, 1987; Vigilant *et al.*, 1991; Krause *et al.*, 2007; Somel *et al.*, 2011; Callaway, 2021), our cultural development (Pagel, 2013), and even the evolution of languages (Nettle and Harriss, 2003; Levinson and Gray, 2012). Furthermore, the influence of evolutionary biology has now expanded to encompass areas such as pharmaceutical research and biomedical applications, including cancer research (Casás-Selves and DeGregori, 2011; Crespi and Summers, 2005). *'Nothing in*

Biology Makes Sense Except in the Light of Evolution' by Theodosius Dobzhansky (1973), remains profoundly resonant and highlights the pervasive role of evolution in elucidating biological phenomena across various disciplines.

Over the past decades, there has been a remarkable surge in data collection within the field of evolutionary biology. Advancements in methodologies and technological performance have given rise to numerous techniques for generating and analyzing large-scale datasets, significantly enhancing our understanding of evolution and genetic diversity. A considerable forthcoming challenge will be to analyze all these data in a coherent, systematic and reproducible way.

3.1 The burst of genetic data

50 years after the first isolation of DNA by Friedrich Mietscher (Heather and Chain, 2016), Phoebus Levene rightly suggested that DNA was composed of a series of nucleotides (Levene, 1919; Simoni *et al.*, 2002). The decoding of this series has attracted great interest, leading to significant advances in genome sequencing techniques over the course of a few years, which can be classified into three distinct phases presented in the following sections (Hutchison, 2007; Mukhopadhyay, 2009; Ebertz, 2020; Giani *et al.*, 2020).

3.1.1 First-generation sequencing

In 1972, Walter Fiers accomplished the first gene sequencing, where he deciphered the sequence of a gene responsible for encoding a bacteriophage MS2 coat protein (Jou *et al.*, 1972). Then, in a groundbreaking achievement in 1976, Fiers became the pioneer in sequencing a complete genome, that of an RNA-genome bacteriophage (Fig. 2.1). This bacteriophage's genome was relatively small, spanning a total of 5,386 base pairs (Fiers *et al.*, 1976).

In 1977, Sanger proposed the dideoxy technique (Sanger *et al.*, 1977a). This technique harnesses chemical analogs of deoxyribonucleotides (dNTPs), the constituent units of DNA strands. By incorporating radiolabeled ddNTPs into a DNA extension reaction at a fraction of the concentration of standard dNTPs, DNA strands of varying lengths are produced, as the incorporation of dideoxy nucleotides during strand elongation leads to premature termination. Through parallel reactions containing individual ddNTP bases and subsequent analysis on polyacrylamide gel lanes, the nucleotide sequence in the original template can be inferred via autoradiography, which reveals a radioactive band at the corresponding gel position.

Frederick Sanger utilized this method to successfully sequence the first DNA genome, comprising 5,375 nucleotides (Sanger *et al.*, 1977b). This sequencing subsequently emerged as the predominant technology for DNA sequencing over the following three decades. Nevertheless, Sanger sequencing technique was labor-intensive and lacks automation (Metzker, 2005; Hutchison, 2007).

During the initial stages of genetics technique development, the primary focus naturally gravitated towards model organisms. In 1992, a first eukaryotic chromosome was fully sequenced, that of the yeast (*i.e. Saccharomyces cerevisiae*). Annotation of this 315-kilobase sequence revealed 182 open reading frames (Oliver *et al.*, 1992). In 1996, with a new sequencing techniques (Roach *et al.*, 1995) its genome was sequenced (Goffeau *et al.*, 1996), followed in 1998 by the genome of the nematode *Caenorhabditis elegans*, a multicellular species (The C. elegans Sequencing Consortium, 1998).

Continuing this trajectory, the year 1999 marked a significant leap with the successful sequencing of the first human chromosome (Fig. 2.1; Dunham *et al.* (1999)). And the turn of the millennium, in 2000, witnessed the sequencing of the genomes of *Drosophila melanogaster* and *Arabidopsis thaliana* (Adams *et al.*, 2000; The Arabidopsis Genome Initiative, 2000).

The publication of the first draft of the *Homo sapiens* genome sequence was produced in 2001 (Lander *et al.*, 2001; Venter *et al.*, 2001). In parallel, during 2002, the genome of *Mus musculus* was sequenced (Waterston *et al.*, 2002). Finally, the collaborative efforts of a consortium project in 2004 achieved an extraordinary accomplishment by publishing the first complete sequence of the human genome (International Human Genome Sequencing Consortium, 2004).

3.1.2 Second-generation sequencing

Commencing in 2005, the landscape of DNA sequencing underwent substantial transformations with the development of Next-generation sequencing (NGS) or second-generation technologies, parallelizing the sequencing of millions of fragments. These technologies are characterized by the need to fragment the genetic material, add adapters, and parallelized both amplification and sequencing.

The major NGS technique is Illumina sequencing, which progressively superseded the conventional capillary sequencing methods (Behjati and Tarpey, 2013; Slatko *et al.*, 2018). This strategy, known for its high-throughput nature, entails breaking down the target DNA or RNA into smaller fragments, which are subsequently linked with specialized adaptors. These adaptors facilitate the binding of the fragments to a solid surface, forming clusters. Through concurrent sequencing-by-synthesis reactions, millions of these clusters are simultaneously sequenced. By incorporating fluorescently labeled nucleotides and capturing their emissions, the sequence can be determined. This technique offers swift and cost-efficient sequencing, rendering it suitable for a wide array of applications, including whole-genome sequencing, transcriptomics.

3.1.3 Third-generation sequencing

Concurrent to Illumina technology, Nanopore technology, although characterized by lower sequencing accuracy, gained prominence due to its capacity to generate considerably longer read lengths, a feature highly valuable in de novo whole-genome sequencing applications (Sevim *et al.*, 2019; Wang *et al.*, 2021). Nanopore sequencing involves passing a

strand of DNA through a nanometer-sized pore, measuring changes in electrical current to determine the DNA sequence without preliminary amplification.

The DNA sequencing landscape has experienced a paradigm shift with the integration of these innovative methodologies. Despite variations in accuracy, these methods, including Nanopore technology, adopted the basic shotgun strategy and embraced parallelization and genome fragmentation to generate templates (Mukhopadhyay, 2009).

Expenses related to genome sequencing have decreased considerably thanks to technical and methodological developments. For example, the first human genome sequenced and assembled in 2001 cost around $2.7 billion. Ten years later generating a 'draft' genome cost $20,000, but today sequencing a human genome costs less than $1,000 with the help of the already existing reference genome (Mullin, 2022; Neville, 2018; Schwarze *et al.*, 2020). Additionally, the enhanced portability and operational ease of sequencing equipment has been facilitated by the ongoing reduction in equipment size. A compelling example of this can be observed in the case of the Oxford Nanopore MinION, which is a representative of third-generation sequencing technology. Functioning akin to a USB device that can be directly linked to a laptop, the MinION offers the advantage of on-site application (Huo *et al.*, 2021). This advances have significantly contributed to a noteworthy escalation in the number of genomes subject to sequencing, a discernible trend evident within the database maintained by the National Center for Biotechnology Information (NCBI, NCBI Resource Coordinators (2018), Fig. 3.1).

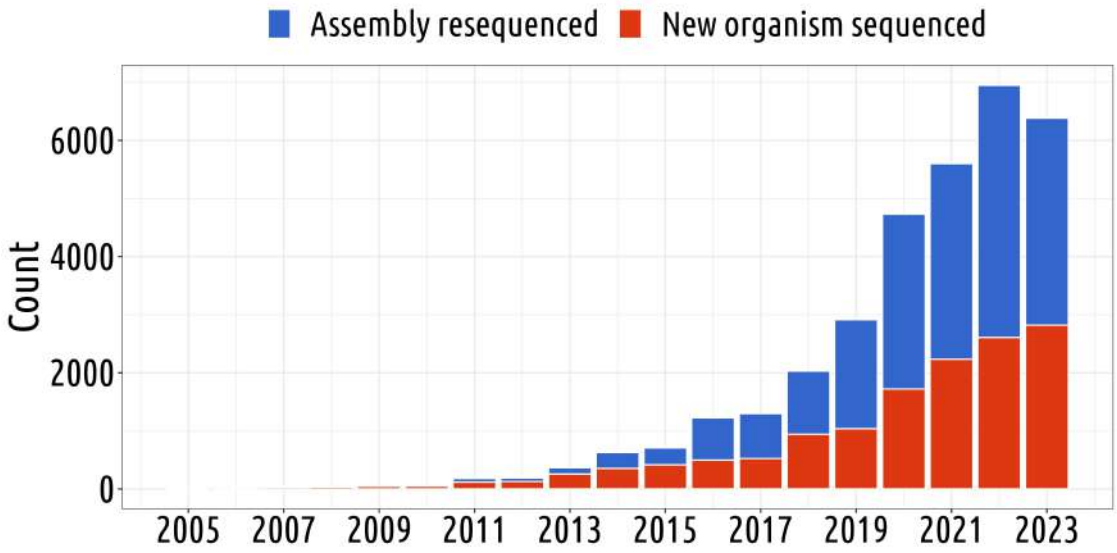


Figure 3.1: Yearly count of Eukaryota sequenced assemblies. *This graph illustrates the annual deposition of Eukaryota genomes at the International Nucleotide Sequence Database Collaboration (INSDC), distinguishing between species deposited for the first time (red) and all assemblies (blue).*

The latest advancements in third-generation sequencing techniques have culminated in long-read sequencing technologies, enabling the generation of reads of 1 to 20 kilo-

base (Marx, 2023; Logsdon *et al.*, 2020).

In addition, dedicated efforts have been directed towards sequencing the genomes of entire populations. This has facilitated the examination of genetic variations at a population level, contributing to our comprehension of the evolutionary forces that underlie these variations. The accumulation of extensive genetic data at this scale has opened novel avenues for investigating genetic diversity, population structure, adaptation, and mechanisms of evolution.

This growing number of genomes and transcriptomes sequenced poses new limitations related to the tremendous amount of data generated, that needs to be stored and archived. Also, new sequencing techniques use algorithms that require more and more resources (GPU, CPU, RAM...). Little by little the sequencing problem becomes now a bioinformatic problem.

3.2 Evolution of Bioinformatics

The term "bioinformatics" was originally introduced by Paulien Hogeweg and Ben Hesper in 1970 to denote the investigation of informational processes within biotic systems (Hesper and Hogeweg, 1970; Hogeweg, 2011). This designation emerged during a period when the volume of data and methodologies available necessitated the computational processing and analysis of substantial datasets that would have been impractical to handle manually. Bioinformatics has increasingly become an essential element in the field of biology. It involves employing computational and statistical methods to analyze, interpret, and manage biological data, particularly genetic and molecular information.

In recent years, significant progress has been made in bioinformatics, driven by enhanced computing power, the development of novel technologies, the improvement of analysis tools and algorithms. These advancements have expedited analyses, increased efficiency, and facilitated the management of ever-growing, intricate datasets. In this context, several algorithms and computer programs have been developed to study evolution.

3.2.1 Sequences alignment

The preceding section has discussed the burst of sequencing methods, followed by the inundation of genetic data in recent years. However, even in the 1970s, scientists had access to data pertaining not to DNA sequences but rather to amino acid sequences of proteins. Notably, at the age of 37, Sanger achieved the sequencing of amino acid chains in bovine insulin, initially unraveling the initial 30 amino acids in chain B (Sanger and Tuppy, 1951a,b), followed by 21 amino acids in chain A (Sanger and Thompson, 1953a,b). After investigation of insulin sequences across multiple species, the next challenge for Sanger was to align these sequences to establish correspondences between positions as they have undergone evolutionary changes. The premise underlying the comparative analysis of sequences from diverse species rested on the notion that conserved regions, retained throughout evolutionary processes, might indicate pivotal elements within the

molecule, such as the 'active center'. Notably, in his alignment Sanger observed that differences were confined to a discrete segment of the molecule (Sanger, 1949; Brown *et al.*, 1955; Harris *et al.*, 1956).

A multitude of methodologies have been devised to facilitate sequence alignments starting with pairwise alignment. The process of pairwise alignment involves aligning two sequences with each other. Among the noteworthy algorithms, the Needleman–Wunsch algorithm, introduced in 1970, stands out as one of the most renowned methods (Needleman and Wunsch, 1970). This algorithm operates by constructing a matrix in which individual cells represent the optimal alignment score for specific subsequences. Through an iterative process, the algorithm populates the matrix, factoring in penalties for gaps and scores for similarity. The final alignment is determined by tracing back along the highest-scoring path within the matrix. It is particularly valuable for aligning closely related sequences of comparable lengths. Derivations of this alignment method have been made, notably with the introduction of the Smith–Waterman algorithm for local alignment (Smith and Waterman, 1981). Local alignment focuses on identifying shorter, highly similar regions within sequences, accommodating gaps and mismatches outside these regions.

The next challenge was to align multiple sequences together, known as Multiple Sequence Alignment (MSA). These methods relies on a guide phylogenetic tree, which is a diagram illustrating the evolutionary descent of various species, organisms, or genes from a common ancestor, to guide the alignment process. This tree is computed using methods such as Neighbor Joining (NJ) based on a genetic distance matrix of the sequences (Saitou and Nei, 1987). Following this tree sequences are successively incorporated to the MSA by pairwise alignment between sequences, a sequence and a consensus sequence or two consensus sequences.

After the creation of sequence databases, such as ACNUC (Gouy *et al.*, 1985) or Genbank (Burks *et al.*, 1991), the need arose to compare a single sequence to a multitude of sequences. It became clear that comparing each pair would be a resource waste. To overcome this difficulty in 1985, the FASTA algorithm was developed, providing rapid search capabilities within protein databases (Lipman and Pearson, 1985). Subsequently, in 1990, it was succeeded by BLAST (Altschul *et al.*, 1990). BLAST's core principle involves identifying short regions of high similarity (local alignments) between the query sequence and sequences within the database. BLAST's algorithm disassemble the query sequence into smaller fragments (k-mers), probing for these fragments within the database, and subsequently extending the matches into more extensive alignments. This methodology empowers BLAST to swiftly recognize regions of similarity without necessitating an exhaustive global alignment spanning the entire sequences' length. This focus on localized alignment makes BLAST an ideal choice for comparing sequences of varying lengths and identifying conserved regions, functional domains, and other biologically significant features. While BLAST predominantly hinges on local alignment principles, it's crucial to acknowledge the existence of alternative sequence alignment tools and algorithms that prioritize global alignment or adopt distinct alignment strategies tailored

to specific bioinformatics objectives.

Around 2008, the emergence of transcriptomic sequencing data and their analyses required the alignment of billions of 100 bp mRNA reads to a reference genome (Mortazavi *et al.*, 2008). To do so, programs such as BOWTIE (Langmead *et al.*, 2009), TopHat2 (Kim *et al.*, 2013) and HISAT (Kim *et al.*, 2019) have been developed. Unlike other alignment methods, these programs rely on a combination of k-mer indexing and graph-based alignment techniques. It creates a graph-like representation of the reference genome, facilitating alignment by accommodating various genomic variations, including single nucleotide polymorphisms (SNPs), insertions, deletions, and splicing events. This approach ensures highly accurate and efficient alignment, particularly for RNA-seq data, where splicing events render local alignment less suitable. The incorporation of k-mer indexing enhances the speed and precision of the alignment process rendering possible gene expression estimation and alternative splicing analyses (Kim *et al.*, 2019).

3.2.2 Collaborative science in the era of big data

Research and collaborative science play a crucial role in the field of genome biology, especially concerning the utilization of massive datasets. The surge in data acquisition, driven by cost efficiencies and technological advancements, has not only spurred cooperative efforts and the exchange of data but has also facilitated large-scale studies. This enables the effective utilization of genetic insights across diverse domains of evolutionary biology.

In response to the arising amount of DNA/RNA sequencing data, instances have created database as an answer to store and use these genetic data. In Europe the European Molecular Biology Laboratory (EMBL) created the EMBL Data Library as a repository for genetic information in 1980 (Hamm and Cameron, 1986). By 1982, this database encompassed 568 sequence entries (Kneale and Kennard, 1984). Subsequently, in 1990 the EMBL Data Library was relocated at the European Bioinformatics Institute (EBI) and renamed EMBL Nucleotide Sequence Database (Rodriguez-Tomé *et al.*, 1996). In response to the advent of Next Generation Sequencing (NGS) data, the European Nucleotide Archive (ENA) emerged in 2008 through the fusion of the EMBL Nucleotide Sequence Database and the former Sequence Read Archive (SRA) (Leinonen *et al.*, 2011a).

In the United States, the GenBank sequence database was established in 1982 (Sayers *et al.*, 2022b). During 1989 to 1992, the GenBank initiative transitioned to the National Center for Biotechnology Information (NCBI; NCBI Resource Coordinators (2018)). Similarly, in 1987, the DNA Data Bank of Japan (DDBJ) pioneered the sole nucleotide sequence data repository in Asia (Tateno and Gojobori, 1997). From their inception, these three databases have maintained a collaborative methods, giving rise to the International Nucleotide Sequence Database Collaboration (INSDC) in 2005 (Karsch-Mizrachi *et al.*, 2012). Daily, the DDBJ/EMBL/GenBank consortium engages in the exchange of data submissions and mutual data sharing (Brunak *et al.*, 2002) (Fig. 3.2). Currently, these databases collectively house data from nearly 14,000 eukaryota genomes, with new submissions pouring in daily (Fig. 3.3). Beyond genome sequences, these

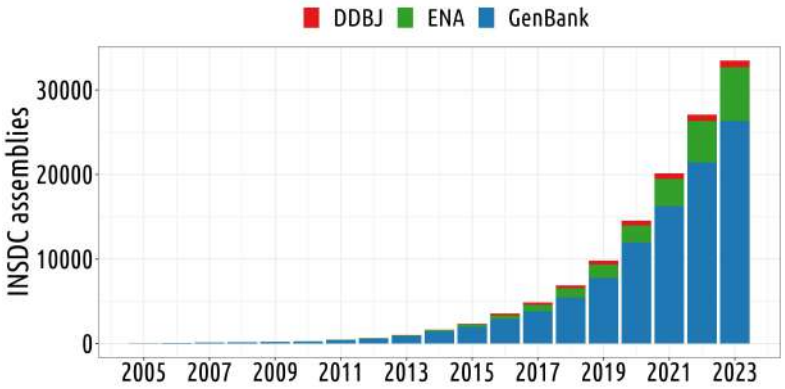


Figure 3.2: Annual INSDC contributions by structures. *Annual deposition of Eukaryota assemblies in distinct partner members of the International Nucleotide Sequence Database Collaboration (INSDC): National Center for Biotechnology Information (NCBI) (depicted in blue), DNA Data Bank of Japan (DDBJ) (depicted in red), and European Nucleotide Archive (ENA) (depicted in green).*

databases host a wealth of information, including annotations, protein and coding sequences. Moreover, they serve as repositories for diverse non-sequence-related data such as species taxonomy information.

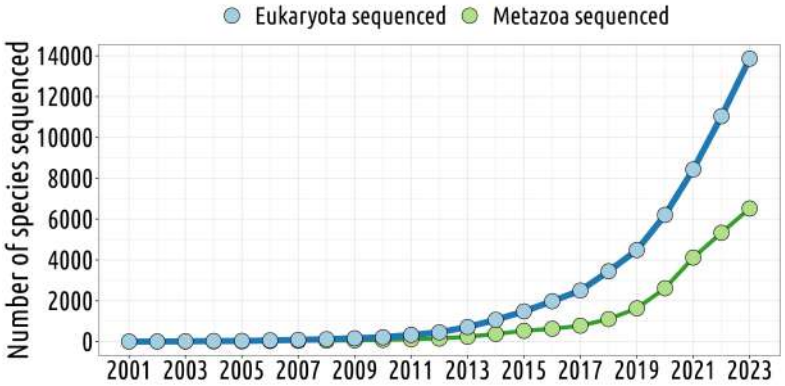


Figure 3.3: Number of species sequenced over time on INSDC. *This figure depicts the number of species' genome sequenced available at the International Nucleotide Sequence Database Collaboration over time. Eukaryota (blue); Metazoa (green).*

Also, the BUSCO annotation program, provides datasets of single-copy orthologous genes for diverse species clades and is derived from the database OrthoDB (Kuznetsov *et al.*, 2023; Manni *et al.*, 2021). OrthoDB database is a comprehensive resource containing single-copy orthologous genes for a wide range of clades. OrthoDB delineates orthologs at key points along the species phylogeny, corresponding to the last common ancestor of the species being studied.

In addition to DNA sequencing data, an abundance of other data sources has given rise to a multitude of specialized databases. For instance, the Human Ageing Genomic Resources integrated the AnAge database in 2013, providing data on life history traits

in animals, encompassing aspects like maximum longevity, body mass, gestation, and sexual maturity (Tacutu *et al.*, 2013, 2018). The Encyclopedia of Life (EOL), hosted by the National Museum of Natural History Smithsonian, became operational in 2014, consolidating information on parameters such as body length, body mass, longevity, and species distribution (Wilson, 2003; Parr *et al.*, 2014). Similarly, the Animal Diversity Web, an online compendium of animal natural history, classification, and conservation biology, has been collaboratively assembled and is maintained by the Museum of Zoology at the University of Michigan (Myers *et al.*, 2023).

The proliferation of databases continues unabated (Fig. 3.4), and a database of molecular biology database is kept updated by Nucleic Acids Research (NAR). It lists the databases that have been described in the annual NAR database issues (Rigden and Fernández, 2018, 2023). The latest Issue contains 178 papers ranging across biology and related fields, and the NAR online Molecular Biology Database Collection presently encompasses an impressive tally of 1,764 databases.

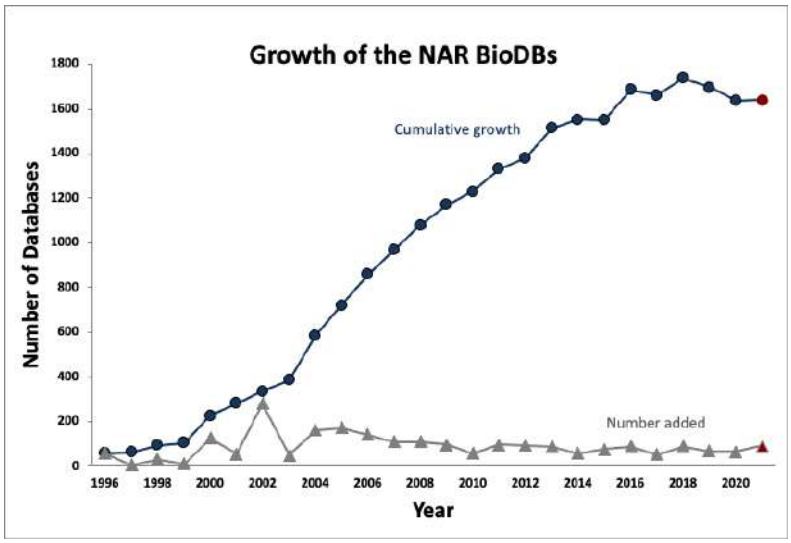


Figure 3.4: Expansion of Nucleic Acids Research Database. *Cumulative growth of the NAR BioDBs described in the annual NAR database issues. Obsoletes and discontinued databases are removed each year.*
Reproduced with permission from Sandra Porter, president of Digital World Biology.

These databases collectively encompass an array of data domains, spanning gene composition, expression patterns, protein sequences, genomic information, cancer studies, plant and metazoan biology, gene orthologs, life history traits, and taxonomy, thereby constituting invaluable resources for researchers across disciplines. However, they inexorably generate such a quantity of data that we no longer know what to do with it. With data resources that are scattered, holed and sometimes annotated inconsistently. They are therefore difficult to reconcile in order to do coherent, reproducible and systemic analyses. A world where everyone has to redo their analyses from the beginning due to the publication of new datasets, new methods, new algorithms, without an integrative structure.

4
Thesis Objectives

Contents

Following Darwin's theory of evolution in 1859, postulating that species adapt to their environment by evolving through natural selection, scientists have been particularly interested in the biological significance of evolutionary changes.

However, with the emergence of sequencing data in 1966, which revealed the true support of evolution (*i.e.* genomes), it became evident that natural selection alone could not account for all changes observed at the molecular level. Instead, alternative theories posited that changes could arise due to stochastic processes, independent of selection.

In recent years, there has been a remarkable increase in genomic data, available for bioinformatic investigations. Notably, metazoan genomes have shown striking complexity and diversity in numerous aspects of their architecture: the genome size (from 43 Gb to 15.3 Mb), the number of protein-coding genes (*e.g.* 20,000 in human, 6,000 in yeast), the genes size (*e.g.* 24 kb in human, 2 kb in flies), the alternative splicing diversity (*e.g.* 90% of genes are subject to AS in human compare to 18% in fly)...

These exciting observations have led researchers to prioritize selection as the primary driver of variations, suggesting adaptive changes. However, some have posited that these changes may be non-adaptive, potentially influenced by an increased of genetic drift. Notably, the "drift barrier" hypothesis predicts that each genome evolves towards an

equilibrium between selection and drift, beyond which further beneficial or deleterious alleles evolve as if they were neutral. Consequently, as the intensity of drift increases, the optimization of genomes decreases.

The amount of data and methodological knowledge, coupled with evolutionary theories, present a unique opportunity to explore the adaptive nature of variations in genomes architecture. By examining the influence of random genetic drift intensity on genome architecture and gene expression, my work aims to determine whether certain genomic features support the "drift barrier", and are, or not, evolving as neutral.

Initially, the development of a robust, reproducible pipeline for systemic analyses of genomes and transcriptomic data is necessary. Indeed, as seen before many data and methods are available but we lack a data resource with comparable analyses across species. I will present in the first section my goal in developing a data resource to capture genomes expression complexity, taking into account their shared evolutionary trajectory, along with effective population size proxies.

This data resource allows us to dive into two highly debated scientific subjects. The first investigation concerns the debate surrounding the adaptative relevance of alternative splicing diversity across metazoans, while the second focuses on genomes base composition to elucidate why translational selection is rarely observed in metazoans.

4.1 Development of an integrated data resource incorporating genomic, transcriptomic, and N_e estimators

At the beginning of my project, developing an analysis to explore the impact of genetic drift on transcriptome complexity in many species was challenging due to the diversity of data available across different studies. Initially, for the exploration of transcriptomic diversity across metazoans, we found that existing databases did not align adequately with our objectives. I will review some of them: Bgee (Bastian *et al.*, 2020), renowned for its extensive compilation of metazoan transcriptomes spanning 52 species, is restricted to vertebrates (N=48 species) lacking representatives of other metazoan clades. Also, it proved unsuitable for our investigation due to its focus on gene expression analysis, and the lack of alternative splicing data we are seeking.

Several databases offer alternative splicing analyses, with four developed by the same group of scientists and shared with the community after my project launched: MeDAS (Li *et al.*, 2020) encompasses 18 metazoans with RNA-seq data from different developmental stages. MetazExp (Liu *et al.*, 2021a) is a comprehensive metazoan database, including 72 non-vertebrates species and a staggering 53,000 uniformly processed RNA-seq samples. FishExp (Tan *et al.*, 2022) contains data on 44 fishes and 26,081 RNA-seq samples. LivestockExp (Liu *et al.*, 2022) focuses on vertebrates, with 14 species and 43,710 RNA-seq samples.

However, while these resources offer a wealth of splicing events and gene expres-

sion data, they are not well suited to address our problematics. Firstly, they are aimed at biologists seeking to analyze alternative splicing patterns on a gene-by-gene manner through web-based queries. This limit, and the large size of the shared data files, complicates cross-species comparisons (the fragmentation across different databases doesn't help). Moreover, they are heterogeneous in the quantity of transcriptomic data between species. For instance, in MetazExp, out of 25,672 RNA-seq samples over 53,000 samples originate from *Drosophila melanogaster*, and in FishExp, out of 21,352 RNA-seq samples over 26,081 samples come from *Danio rerio*. Due to all these limitations, they appeared not well suited for simple analyses aimed at obtaining comparable summary statistics on transcriptomic diversity across a broad spectrum of metazoans.

Thus, we identified the need to develop a comprehensive data resource aimed at facilitating cross-species comparisons of alternative splicing diversity among metazoans. This resource should prioritize simplicity and accessibility, avoiding strong assumptions regarding the functional significance of transcripts variants. It must consist of compressed basic data intended for bioinformaticians, easily available for download. Additionally, the goal includes the development of a user-friendly interface for database exploration, facilitating convenient access to downloadable compressed files.

Another key parameter for my project concerns effective population size proxies, such as life history traits. However, systematically collecting these traits from a wide range of species presents a challenge due to the disparity of the available datasets. For example, the Animal Ageing and Longevity Database (AnAge) (Tacutu *et al.*, 2013) focuses primarily on vertebrates, particularly data on mammals. In contrast, the Encyclopedia of Life (EOL) (Wilson, 2003; Parr *et al.*, 2014) encompasses a broad spectrum of species, with a notable emphasis on invertebrates. The Animal Diversity Web (ADW) (Myers *et al.*, 2023) is a valuable resource, particularly for invertebrate species. Finally, FishBase (Froese and Pauly, 2023) primarily hosts data relating to teleost species. Although AnAge provides comprehensive information on body mass and lifespan, it lacks data on body length. Given our research objectives, no single data resource fits our needs perfectly. We aimed to create a protocol to systematically collect these data from multiple sources.

Also, while the effective population size proxy dN/dS has been estimated for some taxonomic groups in previous studies (Romiguier *et al.*, 2012; Figuet *et al.*, 2016; Lefébure *et al.*, 2017; Bolívar *et al.*, 2019), there is a gap in the literature regarding the analysis of a large data set encompassing a wide range of species. Thus, the first goal of my thesis is to develop a bioinformatic data resource with a controlled pipeline that ensures comparative, reproducible and systemic analyses across diverse species, along with pertinent N_{e} proxies.

4.2 Variations in alternative splicing rates among metazoans: Investigating the impact of drift on splicing errors

As previously outlined, alternative splicing is a prevalent phenomenon in eukaryotes, wherein multiple isoforms are generated from a single gene (Chen *et al.*, 2014). The analyses of transcriptomes from various eukaryotic species showed substantial variation in AS rates across lineages, with the highest rate in primates (Barbosa-Morais *et al.*, 2012; Chen *et al.*, 2014; Mazin *et al.*, 2021). However, The influence of random genetic drift on the diversity of alternative splicing patterns is a subject of significant scientific interest. This inquiry is particularly important as it lays the foundation for addressing another highly contested issue: whether alternative splicing serves an adaptive purpose or primarily constitutes an accumulation of splicing errors. In essence, the fundamental question revolves around whether AS enhances an organism's protein repertoire or predominantly results from the accumulation of erroneous splicing events.

4.2.1 A scientific debate and a lack of evidence

In order to delve into the roots of the ongoing scientific discourse, an in-depth exploration of the relevant literature is imperative. Starting with Brenton Graveley's 2001 review on the alternative splicing diversity, in which he posited that *'It is becoming clear that alternative splicing has an extremely important role in expanding protein diversity and might therefore partially underlie the apparent discrepancy between gene number and organismal complexity' (Graveley, 2001).* Graveley yet acknowledged that *'for the vast majority of alternative splicing events, the functional significance is unknown' (Graveley, 2001).* Nevertheless, this did not deter him from concluding that *'It does not seem possible that the complexity of an organism can be explained by the one gene, one protein hypothesis. Thus, what some consider noise might actually be crucial in facilitating the development of complex organisms from a limited number of genes' (Graveley, 2001).* Graveley's seminal work in 2001 instigated a discourse that endures to the present day (Graveley, 2001).

In 2006, Blencowe's study on alternative splicing mirrored Graveley's uncertainty, as he pondered *'whether we are just observing the tip of the iceberg or whether the majority of important AS events have already been identified' (Blencowe, 2006).* Nevertheless, two years later, Blencowe published a paper in Nature starting by *'Alternative splicing is considered to be a key factor underlying increased cellular and functional complexity in higher eukaryotes' (Pan* et al., *2008).* Postulating a connection between alternative splicing and increased cellular and functional complexity in higher eukaryotes. This transition marked a turning point for many scientists, who began to present their hypotheses as established facts without conclusive evidence.

In 2010, Graveley further emphasized the importance of alternative splicing by stating that *'it is now clear that the 'missing' information is in large part provided by alternative splicing, the process by which multiple different functional messenger RNAs, and there-*

fore proteins, can be synthesized from a single gene' (Nilsen and Graveley, 2010). At this point, we have two respected scientists advocating for a hypothesis without definitive results. Graveley was cognizant of this and, in addressing outstanding questions, mentioned, *'Another crucial question is how many mRNA isoforms are functionally relevant? Teleology suggests that if an isoform exists, it is important [...]. But this idea is hard to prove and is difficult for some to accept' (Nilsen and Graveley, 2010).* This final observation may seem paradoxical because it is not necessary to criticize those who are skeptical. Rather, it underscores that the difficulty in proving the concept is likely the reason for skepticism among some scientists.

The first comparative studies appears in 2011 where researchers investigated the correlation between biological complexity, the number of different cell types, and proteome size (Schad *et al.*, 2011). They demonstrated a link between proteome size and complexity, as well as complexity and the number of alternative splicing events per multi-exon gene. This led them to the conclusion that *'these features suggest that organism complexity increases with increasing functional complexity of gene products' (Schad* et al.*, 2011).* However, their study did not account for the quantity and diversity of RNA extracted. Indeed, low amount of RNA leads to mostly study highly expressed genes because lowly expressed genes won't be observed or rarely. Also, by diversifying the samples under study (*i.e.* tissues, conditions...) a broader spectrum of genes will be detected. Consequently, the repertoire of genes under study exhibited variability due to differences in RNA samples studied.

A particularly intriguing paper by Chen *et al.* (2014) considered the number of cell types as a proxy for complexity and demonstrated a correlation with alternative splicing per gene and the proportion of multi-exon genes. Notably, their study utilized highly divergent species and did not address the issue of phylogenetic inertia, the fact that traits have a shared evolutionary trajectory. For instance, they only examined five mammalian species. Also, they considered Schad *et al.* (2011)'s paper inconclusive due to *'the lack of comparable alternative splicing measures' (Chen* et al.*, 2014).* Interestingly, the authors of (Chen *et al.*, 2014) study were mindful of the concept of the "drift barrier" and the fact that under a non-adaptive model, lowly expressed genes experience lower selective pressure. They noted that if lowly expressed genes exhibit higher levels of splicing, the data could overestimate alternative splicing levels in species with abundant expression data, *i.e.* the quantity of expressed sequenced tags (EST), thereby inflating correlation strength.

In their supplementary materials, Chen *et al.* (2014) presented data showing in lowly expressed genes a small number of ESTs compared to the number of alternative splicing events. Whereas in highly expressed genes they observed a large number of ESTs compared to the number of AS events. However, surprisingly, they seem to have misinterpreted these results, stating, *'contrary to the prediction of the non-adaptive model, we found that more highly expressed genes are also more highly spliced' (Chen* et al.*, 2014).* This assertion seems erroneous as it fails to account for coverage depth (*i.e.* ESTs quantity); the focus should be on the ratio between alternative splicing events and ESTs,

rather than the absolute numbers. This paper is mostly cited for its conclusion on the relationship between alternative splicing and complexity, which seems questionable.

Another line of thought has emerged within the scientific community, positing that the majority of alternative splicing events are likely erroneous (Pickrell *et al.*, 2010; Leoni *et al.*, 2011). It has been proposed that most protein-coding genes predominantly express a single major isoform at significantly higher levels than others (Gonzàlez-Porta *et al.*, 2013; Tress *et al.*, 2017b). These studies suggest that although some minor transcripts may have functional significance, the major isoforms are likely the primary contributors to the proteome.

This discourse took a critical turn in 2017 when a provocative headline emerged, asserting that "Alternative splicing may not be the key to proteome complexity", authored by Tress (Tress *et al.*, 2017a). In his statement, Tress posited that most gene in human have a single major protein isoform, and exons subject to alternative splicing does not appear to be under selective pressure, suggesting that a significant number of predicted alternative transcripts may not even be translated into proteins. In response, Blencowe emphasized the necessity for the development of high-throughput methods to investigate the functions of splice variants, stating, *'an important goal for future studies will be to further develop high-throughput methods for interrogating the functions of splice variants' (Blencowe, 2017)*. He concluded assertively by cautioning that *'In the meantime, one should be mindful of the old aphorism, "absence of evidence is not evidence of absence"' (Blencowe, 2017)*. Finally, Tress countered by stating, *'Researchers tend to make sweeping conclusions about genome-wide roles for alternative splicing when we actually know very little about the detailed functional roles of the vast majority of alternative isoforms, even those that are generated from highly conserved exons' (Tress* et al.*, 2017b)*.

Further evidence challenging the notion of adaptive alternative splicing came to light in 2017 when a negative correlation between gene expression levels and the rates of intron retention and alternative splicing was observed in both humans and paramecium (Saudemont *et al.*, 2017). Chen *et al.* (2014) had previously alluded to this correlation in their supplementary data but had possibly misinterpreted the results. Moreover, in 2014, Blencowe had also observed a similar correlation, albeit without explicitly emphasizing the same interpretation (Braunschweig *et al.*, 2014). He noted that *'we observe that Intron Retention globally impacts gene expression in mammalian cells and tissues by negatively regulating cytoplasmic transcript levels' (Braunschweig* et al.*, 2014)*. It is evident that Blencowe had a belief in the functional importance of alternative splicing, suggesting that the high rates of intron retention might be a causal factor for low gene expression. While he may be correct, he neglected to acknowledge that this outcome aligns with a non-adaptive model. High rates of alternative splicing is not the cause of low gene expression but rather the consequence of reduced selective constraints (Melamud and Moult, 2009).

In summary, please refer to Fig. 4.1 for a chronological overview of the ideas and papers investigated, along with their impact.

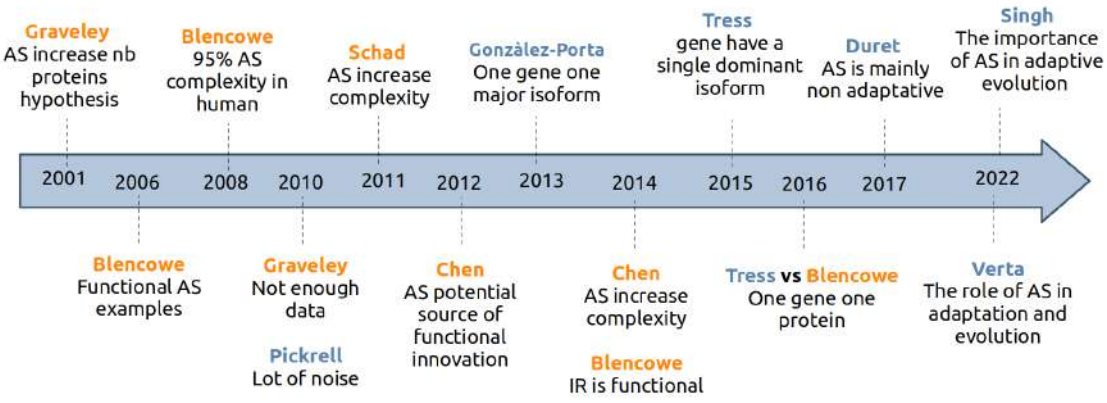


Figure 4.1: Chronology of the literature on the "raidon d'être" of alternative splicing. *Chronology of the main papers taking position in favor of AS primarily increasing proteome diversity (orange) or generating mostly erroneous variants (blue).*

In contemporary literature, the resonance of this far-reaching assertion reverberates, with alternative interpretations in recent publications entitled "the importance of AS in adaptive evolution" (Singh and Ahi, 2022), "The role of alternative splicing in adaptation and evolution" (Verta and Jacobs, 2022). It is not my intent to discredit this work, but rather, to underscore the necessity of addressing the fundamental question of whether alternative splicing (AS) is predominantly adaptive or non-adaptive. It is essential to clarify that our critique does not stem from an assertion of the prior work's inaccuracy but rather from the paucity of concrete findings pertaining to this specific inquiry. However, it is important to note that numerous studies have probed the functionality of AS, and the existence of one observation, functional variants, does not negate the possibility of a substantial number of erroneous variants.

In the aforementioned papers, the study Chen *et al.* (2014) is cited to make the assertion that *'A convincing argument for the importance of alternative splicing in organismal evolution was made by Chen* et al. *(2014)' (Singh and Ahi, 2022).* However, it is prudent to acknowledge that this study may be considered outdated as mentioned earlier. Notably, the authors themselves acknowledge that *'The functional impact of most splice variants on organismal phenotype has been the source of extensive debate as it is largely unknown to what extent different alternative isoforms are translated into functional proteins that can alter phenotypes and hold adaptive importance' (Singh and Ahi, 2022).*

4.2.2 A fresh perspective through the "drift barrier"

In order to address the question of whether alternative splicing (AS) is primarily adaptive or not, we developed a research protocol based on the "drift barrier" hypothesis proposed by Lynch (2007a) (see 'Genetic drift' section).

Population genetics principles posit that the capacity of selection to favor advantageous mutations or eliminate detrimental ones hinges on the strength of selection (s) relative to the influence of random genetic drift, which is characterized by the effec-

tive population size (N_e). When the selection coefficient is considerably weaker than drift ($|N_e s| \ll 1$), alleles behave as if they are effectively neutral. Consequently, random genetic drift imposes an upper limit on selection's ability to impede the fixation of suboptimal alleles (Kimura *et al.*, 1963; Ohta, 1973). This concept, known as the "drift barrier", as introduced by Lynch (2007a), is expected to have repercussions on the efficiency of various cellular processes, including splicing. Therefore, species with lower N_e values are anticipated to be more susceptible to splicing errors compared to species with higher N_e values.

To assess this hypothesis and analyze the impact of genetic drift on alternative splicing patterns, we calculated AS rates in 53 metazoan species, utilizing the initial release of our data resource. These species represent a wide spectrum of N_e values and were selected based on the availability of high-depth transcriptome sequencing data.

Our research was inspired by the findings of Saudemont *et al.* (2017) and Braunschweig *et al.* (2014), as well as the earlier hypothesis put forth by Chen *et al.* (2014). Indeed, under a non-adaptive model, genes expressed at lower levels experience reduced selective pressure, leading to an expectation of greater splicing errors in lowly expressed genes compared to highly expressed ones. We extended this investigation across all the species included in our study to ascertain whether this pattern remains consistent across various species and clades. Finally, we sought to identify functional variant signals, such as the preservation of the reading frame of major isoforms among splicing variants.

4.3 Synonymous codons usage among metazoans

In the early days of deciphering genetic codes, it became evident that the usage of synonymous codons is not uniform; certain synonymous codons are used more frequently than others (Grantham *et al.*, 1980b,a; Ikemura, 1981; Gouy and Gautier, 1982; Sharp *et al.*, 1988; Mouchiroud *et al.*, 1988). This non-uniform utilization of synonymous codons is observed to exhibit considerable variation among different species (Duret and Mouchiroud, 1999). Therefore, the second scientific investigation of my thesis is to elucidate the underlying factors contributing to the variability in synonymous codon usage within animal taxa. In particular, I wanted to test whether the selection on synonymous codons usage for optimizing translation depends on random genetic drift intensity.

4.3.1 Causes of codon usage variations, a long standing debate

The utilization of synonymous codons is under the influence of two distinct but non-exclusive processes: non-adaptive and adaptive mechanisms (Bulmer, 1991; Duret, 2000, 2002; Plotkin and Kudla, 2011; Doherty and McInerney, 2013; Parvathy *et al.*, 2022).

The non-adaptive model posits that genome-wide mutation patterns and factors such as GC-biased gene conversion (gBGC), which is influenced by recombination rates, play a role in shaping synonymous codon usage (Ikemura, 1981; Kanaya *et al.*, 2001; Chen

et al., 2004; Pouyet *et al.*, 2017). These processes affect the entire genome and are regrouped under the term of neutral substitution patterns (NSP), whose variations can be observed in non-coding regions.

The adaptive model suggests the existence of optimal synonymous codons for translation, aiming for rapid, high-fidelity translation. This concept is known as translational selection and is expected to drive the usage of these optimal codons. One key prediction of this model is that optimal codons should correspond to those decoded by the most abundant tRNA, facilitating accelerated translation (Akashi, 1994; Drummond and Wilke, 2008; Hershberg and Petrov, 2008; Morris *et al.*, 2021), while minimizing translation errors (Stoletzki and Eyre-Walker, 2007; Kramer and Farabaugh, 2007; Sun and Zhang, 2022). It is worth noting that, in certain cases, such as the folding of protein structures, there may be a rationale for slowing ribosome translation (Yu *et al.*, 2015; Liu, 2020; Weinberg *et al.*, 2016; Hussmann *et al.*, 2015).

Another key prediction of the adaptive model is that the intensity of translational selection should correlate with gene expression, as highly expressed genes require a larger number of ribosomes for their translation. Consequently, the utilization of suboptimal codons in those genes is anticipated to exert a more pronounced influence on the organism's fitness. Consequently, the codon usage of highly expressed genes is expected to better match the tRNA pool. This correlation has been observed in a variety of organisms, including *Drosophila melanogaster*, *Escherichia coli*, and *Caenorhabditis elegans* (Ikemura, 1981; Sharp *et al.*, 1988; Percudani, 2001; Duret and Mouchiroud, 1999; Duret, 2000).

However, in vertebrates, translational selection in highly expressed genes seems very weak (dos Reis and Wernisch, 2009; Doherty and McInerney, 2013). Indeed, dos Reis and Wernisch (2009) showed that the population-scaled selection coefficient estimated on 9 amino acids, is weak in human and mouse, whereas *Drosophila melanogaster* and *Caenorhabditis elegans* have a higher translational selection. Also, they pointed out that their estimation might be incorrect because they did not take into account the nucleotide heterogeneity along genomes. Methods for quantifying translational selection in these interesting findings may seem circular, because TS is often estimated by considering codons predominant in highly expressed genes as optimal, to finally quantify their degree of predominance. Thus, TS estimators are always positive and may, in fact, be overestimated, because the predominance of these codons could be due to variations in mutational biases along the genome.

A long-standing and controversial question involves the examination of human genome, which has not revealed clear indications of translational selection. Indeed, there is substantial evidence suggesting that non-adaptive processes significantly influence codon usage bias. Notably, the GC3 content, representing the compositional bias at the third position of codons, reflects synonymous codon usage variations and correlates with genome base composition. This suggests that codon usage is affected by process affecting the entire genome not only regions subject to translational selection (Mouchiroud *et al.*, 1988, 1991; Kanaya *et al.*, 2001; Chen *et al.*, 2004; Clay and Bernardi, 2011).

Nonetheless, the debate regarding the impact of translational selection on the human

genome remains highly contested, largely due to variations in non-adaptive processes that are often overlooked in studies. As example, several investigations have reported variations in human codon usage across genes expressed in different tissues or cell types (Vinogradov, 2003; Plotkin *et al.*, 2004; Gingold *et al.*, 2014). Gingold *et al.* (2014) demonstrated significant variations in synonymous codon usage (CU) among genes associated with cellular proliferation and differentiation. Additionally, they observed that the expression of the tRNA pool varies across different cell types, each of which expresses specific sets of genes whose coding sequences may be co-adapted with specific pools of tRNAs. These findings, if valid, suggest a substantial role of translational selection in regulating and determining cell fate.

However, contradictory evidences have been presented suggesting that, despite changes in the tRNA pool between cells, the collective expression of tRNAs with a particular anticodon remains stable throughout development (Schmitt *et al.*, 2014). Another study showed no covariation between tRNA pool and codon usage in contrasting cells undergoing proliferation and differentiation (Rudolph *et al.*, 2016). Furthermore, Pouyet *et al.* (2017) presented evidences that meiotic activity determines CU, as differences in CU between sets of genes reflect disparities in meiotic activity linked to recombination and, consequently, genetic biased gene conversion (gBGC). The prevalence of strong gBGC in mammalian genomes may preclude translational selection from co-adapting the tRNA pools to codon demand.

More recently, Dhindsa *et al.* (2020) identified distinct gene classes employing specific sets of codons, which they interpreted as indicative of translational selection. It is important to note, however, that they did not consider the role of gBGC in their study, despite the well-established preference for GC alleles due to gBGC in humans. Because of this, their results led them to conclude that the transition from optimal to non-optimal codons is less favorable compared to the reverse transition. Nevertheless, their study predominantly focuses on GC to AT transitions, which are heavily influenced by gBGC, and they do not acknowledge this factor in their paper. The debate, therefore, remains unresolved due to the occasional forgetting of non-adaptive models in scientific investigations.

4.3.2 Evaluating translational selection intensity and its relation to drift

In our study, we aim to explore synonymous codon usage across animals and determine the causes, whether adaptive or not, underlying these variations. Additionally, we intend to investigate the impact of random genetic drift on translational selection (TS). Indeed, under the "drift barrier" hypothesis, a strong random genetic drift leaves little room for translational selection to operate.

To provide answers to this long standing debate we propose to systematically analyzed CU and TS in metazoans, based on previous approaches on model species. We possess an extensive resource covering a multitude of species, with which we address various research questions. Our protocol is based on the differentiation of genomic re-

gions affected by non-TS processes, *i.e.* introns, and regions influenced by both non-TS processes and TS, *i.e.* exons.

One primary question that we aim to tackle is about the influence of neutral substitution patterns (NSP) on synonymous codon usage across species. To do so we aimed at systematically compile intronic data from genes to estimate the neutral substitution patterns, considering variables such as the genomic GC content. Concurrently, we intend to collect codon usage patterns within exon sequences.

Using the genes expression obtained from RNA-seq samples, we can investigate another focal point: the variability in TS intensity across diverse species. To investigate this, we analyze the extent to which codons promoted by TS are preferentially utilized in highly expressed genes (Ikemura, 1981, 1985; Duret, 2000). First, it is crucial to identify codons that should be favored by translation, specifically those decoded by the most abundant tRNA molecules. To do so, we propose to systematically quantify the copy numbers of each tRNA gene, a measure highly correlated with tRNA abundance as demonstrated in previous studies on *Homo sapiens* and *Drosophila melanogaster* (Behrens *et al.*, 2021). In cases where this data is not readily available, we plan to use tRNAscan-SE for tRNA annotation (Chan *et al.*, 2021).

One final inquiry concerns the exploration of factors contributing to variations in TS intensity. Within the context of this thesis, we examined factors such as the influence of the "drift barrier" on TS, hence effective population size, utilizing four proxies.

Part II

Studies

5

GTDrift: A resource for exploring the interplay between genetic drift, genomic and transcriptomic characteristics in eukaryotes

The first objective of my thesis is to gather extensive genomic and transcriptomic data encompassing numerous metazoans and assess the extent of variability in random genetic drift among these species. To achieve this goal, we created and submitted for publication GTDrift, a comprehensive data resource that enables explorations of genomic and transcriptomic characteristics alongside proxies of the intensity of genetic drift in individual species. This resource encompasses data for 1,507 eukaryotic species, including 1,414 animals and 93 green plants, and is organized in three components.

The first two components contain approximations of the effective population size, which serve as indicators of the extent of random genetic drift within each species. In the first component, we meticulously investigated public databases to assemble data on life history traits such as longevity, adult body length and body mass for a set of 969 species. The second component includes estimations of the ratio between the rate of non-synonymous substitutions and the rate of synonymous substitutions (dN/dS) in protein-coding sequences for 1,324 species. This ratio provides an estimate of the efficiency of natural selection in purging deleterious substitutions. The third component encompasses various genomic and transcriptomic characteristics. With this component, we aim to facilitate comparative transcriptomics analyses across species, by providing easy-to-use processed data for more than 16,000 RNA-seq samples across 491 species. These data include intron-centered alternative splicing frequencies, gene expression levels and sequencing depth statistics for each species, obtained with a homogeneous analysis protocol.

To enable cross-species comparisons, we provide orthology predictions for conserved single-copy genes based on BUSCO gene sets. To illustrate the possible uses of this database, we identify the most frequently used introns for each gene and we assess how the sequencing depth available for each species affects our power to identify major and minor splice variants.

GTDrift: A resource for exploring the interplay between genetic drift, genomic and transcriptomic characteristics in eukaryotes

Florian Bénitière[1,2], **Laurent Duret**[1], **Anamaria Necsulea**[1]

[1]Laboratoire de Biométrie et Biologie Évolutive, UMR CNRS 5558, Université Claude Bernard Lyon 1, France
[2]Laboratoire d'Ecologie des Hydrosystèmes Naturels et Anthropisés, UMR CNRS 5023, Université Claude Bernard Lyon 1, France

Contents

5.1 Introduction

Genetic drift refers to stochastic fluctuations in allele frequencies within a population across successive generations. These fluctuations arise due to the inherently random sampling of individuals that reproduce and pass on their alleles to subsequent generations (Wright, 1929; Graur and Li, 2000). Population genetics principles state that the ability of natural selection to promote beneficial mutations or eliminate deleterious mutations depends on the intensity of selection (s) relative to the power of genetic drift (defined by the effective population size, N_e): if the selection coefficient is sufficiently weak relative to drift ($|N_e s| < 1$), alleles behave as if they are effectively neutral (Kimura *et al.*, 1963; Ohta, 1973). Thus, random drift sets an upper limit on the efficiency of selection. This limit is called the "drift barrier" (Lynch, 2007a, 2010). Genomes that are subject to intense genetic drift are expected to be less well-optimized compared to those experiencing lower genetic drift. Michael Lynch proposed that variation in the ability to purge slightly deleterious mutations (*i.e.* variation in N_e) can account for differences in genome characteristics among species (Lynch and Conery, 2003). This hypothesis has been empirically validated for multiple genome characteristics and phylogenetic clades. For example, it was shown that the genomes of crustacean species with low N_e values are larger than those of their sister species (Lefébure *et al.*, 2017). Moreover, species with large N_e tend to have a lower mutation rate than species with low N_e, illustrating the notion that natural selection acts to improve replication fidelity, within the constraints defined by random genetic drift (Lynch *et al.*, 2016).

We recently examined the variations in transcriptome complexity across animal species in light of the "drift barrier" hypothesis (Bénitière *et al.*, 2024). In multicellular eukaryotes, the vast majority of genes give rise to multiple isoforms through alternative splicing (Chen *et al.*, 2014). This phenomenon has attracted a great deal of interest since its discovery almost 50 years ago (Berget *et al.*, 1977). Alternative splicing is commonly hypothesized to be adaptive, because it can increase the number of biological functions that are encoded in each genome. Indeed, numerous instances of alternative splicing patterns with beneficial effects have been identified (Mudge *et al.*, 2011; Barbosa-Morais *et al.*, 2012; Merkin *et al.*, 2012; Reyes *et al.*, 2013; Verta and Jacobs, 2022; Singh and Ahi, 2022; Wright *et al.*, 2022). However, these examples represent only a small fraction of all splice variants that are now known, especially given the substantial detection power brought by next-generation RNA sequencing (RNA-seq) techniques. Many of the splice variants that can now be detected with RNA-seq are present at very low frequencies (Gonzàlez-Porta *et al.*, 2013; Tress *et al.*, 2017a) and are poorly conserved during evolution (Barbosa-Morais *et al.*, 2012; Merkin *et al.*, 2012). It was thus hypothesized that they may be the result of errors of the splicing machinery, rather than functional isoforms (Pickrell *et al.*, 2010; Gout *et al.*, 2013; Xu and Zhang, 2014; Saudemont *et al.*, 2017; Xu and Zhang, 2018; Liu and Zhang, 2018b,a; Xu *et al.*, 2019; Xu and Zhang, 2020; Zhang and Xu, 2022). Notably, according to the "drift barrier" hypothesis, one may hypothesize that if alternative splicing (AS) primarily serves functional roles, the

rate of alternative splicing should increase with N_e. Conversely, if AS predominantly involves deleterious processes, its rate should decline with increasing N_e. We applied this reasoning in our previous work (Bénitière *et al.*, 2024), which led us to deduce that AS is predominantly non-functional.

This methodology for exploring the impact of N_e on biological processes holds potential for broader applications. For example, one could examine the functional importance of alternative polyadenylation sites (Xu and Zhang, 2018). Such investigations demand cross-species comparative transcriptomics analyses, a task facilitated by the abundant availability of publicly accessible RNA-seq data. Yet, analysis of transcriptome sequencing data is resource-intensive in terms of time, energy, and computational power. To facilitate future analyses, we provide a comprehensive database that streamlines the process by offering pre-processed data. This dataset includes proxies for effective population size, sets of orthologous single-copy genes, gene expression levels, and intron-centered alternative splicing frequencies, along with phylogenetic trees to control for phylogenetic inertia. These resources have been compiled for 16,000 RNA-seq samples spanning 1,507 multicellular eukaryotic species.

This database, that we name GTDrift, complements other public transcriptomic data resources, such as Bgee (Bastian *et al.*, 2020), which provides gene expression levels for 52 species (Version 15.0.1), but not alternative splicing frequencies. Other databases do provide alternative splicing frequencies. For example, MeDAS (Li *et al.*, 2020) provides AS data for 18 metazoan species, and MetazExp (Liu *et al.*, 2021a) provides data for 72 metazoan species. This latter resource is substantial, including data for $\sim$ 53,000 RNA-seq samples. However, this database favors insects (53 species, with $\sim$ 26 000 RNA-seq samples for *Drosophila melanogaster*) and does not include any representative of the vertebrate clade, for which more computational resources are required because of their large genomes. Our database encompasses a broader phylogenetic distribution of species (Fig. 1), with 93 green plant species, 561 invertebrates and 853 vertebrates. Moreover, while other public databases such as MetazExp are aimed at biologists who want to analyze alternative splicing patterns in a gene-by-gene manner through web queries, in GTDrift we provide all data in flat files, which enable downstream computational analyses. GTDrift is thus mainly aimed at users with some computational skills. Nevertheless, we have created a user-friendly Shiny app to facilitate exploration of the database and species-specific data downloads (Chang *et al.*, 2024).

In GTDrift, we used assemblies and annotations data collected from The National Center for Biotechnology Information (NCBI) (Sayers *et al.*, 2022a), as well as publicly available RNA-seq data to investigate alternative splicing patterns and gene expression profiles. We computed summary statistics across all analyzed RNA-seq samples for each species, which enabled us to determine whether the available sequencing depth is sufficient for the study of alternative splicing. To ensure comparability across species, we annotated Benchmarking Universal Single Copy Orthologs (BUSCO) (Waterhouse *et al.*, 2018) genes in all species and provide phylogenetic trees to control for phylogenetic inertia.

We believe that this tremendous amount of information should be shared with the

scientific community, because it provides the means to investigate the impact of genetic drift on genome and transcriptome architecture, on a broad phylogenetic scale.

5.2 Methods

5.2.1 Species selection

The first criterion for species inclusion in GTDrift is the availability of a genome assembly and annotation in the NCBI database (NCBI Resource Coordinators, 2018; Sayers *et al.*, 2022a), as well as the availability of RNA-seq data in the Short Read Archive (Leinonen *et al.*, 2011b). We included 1,507 multicellular eukaryotic species. This collection encompasses 1,414 animal species as well as 93 species of green plants (Fig. 1). Our Snakemake pipeline can be applied to any species for which genome sequence, genome annotation and RNA-seq data are available, which will enable us to further expand GTDrift in the future(Mölder *et al.*, 2021).

5.2.2 Collecting life history traits

We queried several databases to acquire three specific life history traits, namely: maximum longevity, body mass, and body length. These traits were previously identified as suitable proxies for estimating the effective population size (Romiguier *et al.*, 2014a; Waples, 2016; Figuet *et al.*, 2016; Galtier, 2016; Weyna and Romiguier, 2020). For eusocial species, which live in colonies and have both reproductive and non-breeding individuals, we gather data on the queen of the colony. For solitary species, we collected data for females if available; otherwise, males were considered.

We employed several distinct methodologies to screen the databases. We initially used a manual approach to search across various sources of information, including scientific papers and databases.

We manually searched for information on life history traits from four prominent databases, which encompass diverse taxonomic groups. The Animal Ageing and Longevity Database (AnAge) (Tacutu *et al.*, 2013), is renowned for its comprehensive collection of vertebrates, particularly mammals. The Encyclopedia of Life (EOL) (Wilson, 2003; Parr *et al.*, 2014) encompasses a wide spectrum of species, prominently featuring invertebrates. The Animal Diversity Web (ADW) (Myers *et al.*, 2023), is a particularly rich resource for invertebrates. The FishBase (Froese and Pauly, 2023) predominantly houses data on teleostei species. While AnAge furnishes extensive information regarding body mass and lifespan, it is lacking data pertaining to body length (Fig. 2A,B,C). Furthermore, as previously noted, certain databases are tailored to specific clades. For instance, in comparison to EOL and ADW, AnAge contains relatively fewer records for invertebrates (Fig. 2D,E,F).

We then made efforts to automate the manual search procedures. The primary automated procedure involved the development of a bash script, which utilized the Latin

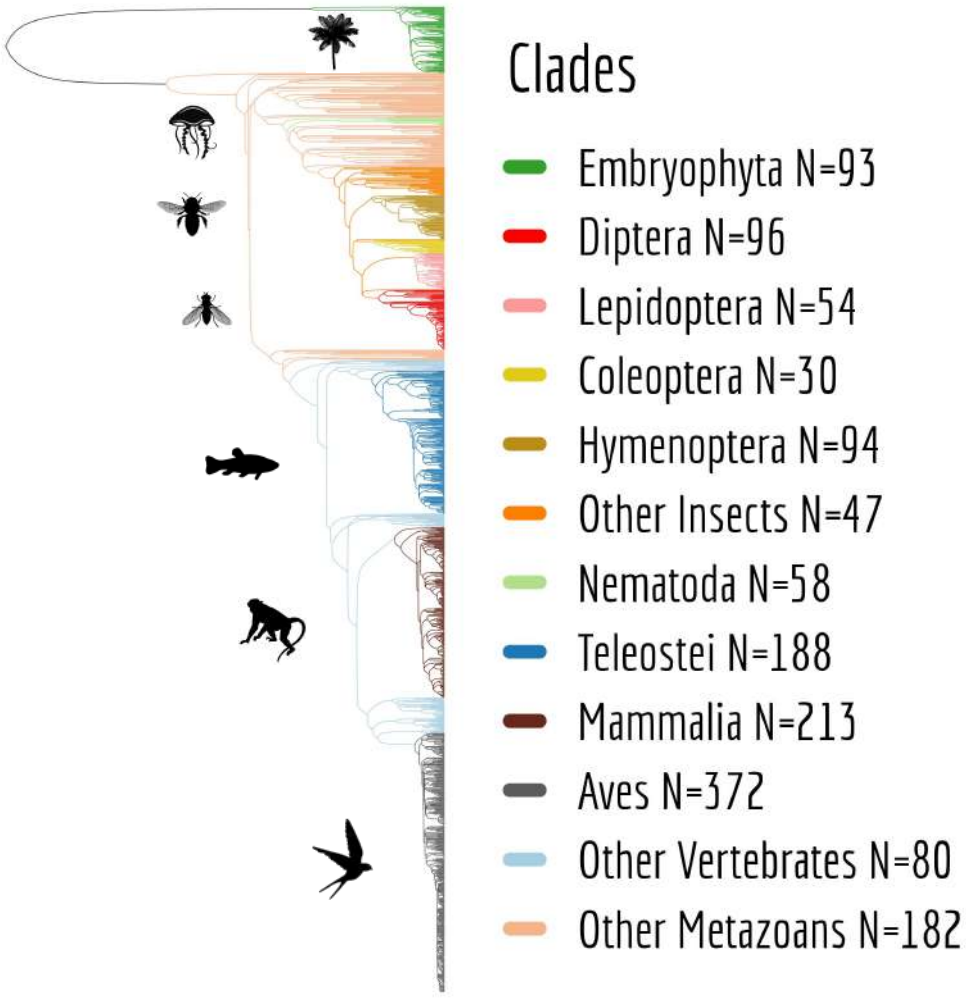


Figure 5.1: Phylogenetic distribution of the species included in the GTDrift database. *The phylogeny was retrieved from TimeTree (Kumar et al., 2022). Not all species studied are present (N=1,221).*

nomenclature of the species to navigate the textual content within the research pages of the 4 databases listed above. The bash script was designed to extract sentences, words, and numerical data in proximity to keywords such as ”longevity”, ”mass”, and ”weight”, serving as indicators of relevant information. Its output was then reformatted through an R script. While this approach proved effective for databases like AnAge, EOL, and FishBase, its applicability to the ADW database was limited due to the manner in which information is embedded within textual paragraphs. Consequently, we employed an alternative method for the ADW database, involving machine learning and Natural Language Processing Question-Answering techniques. We obtained a trained model named ”tinyroberta-squad2” from huggingface.co (noa, 2023). This model was used to answer questions related to specific attributes, such as ‘what is the body length ?’; ‘what is the body mass ?’; ‘what is the longevity ?’. Each question retrieved a pool of 100 potential answers derived from the database’s textual content, ranked by their

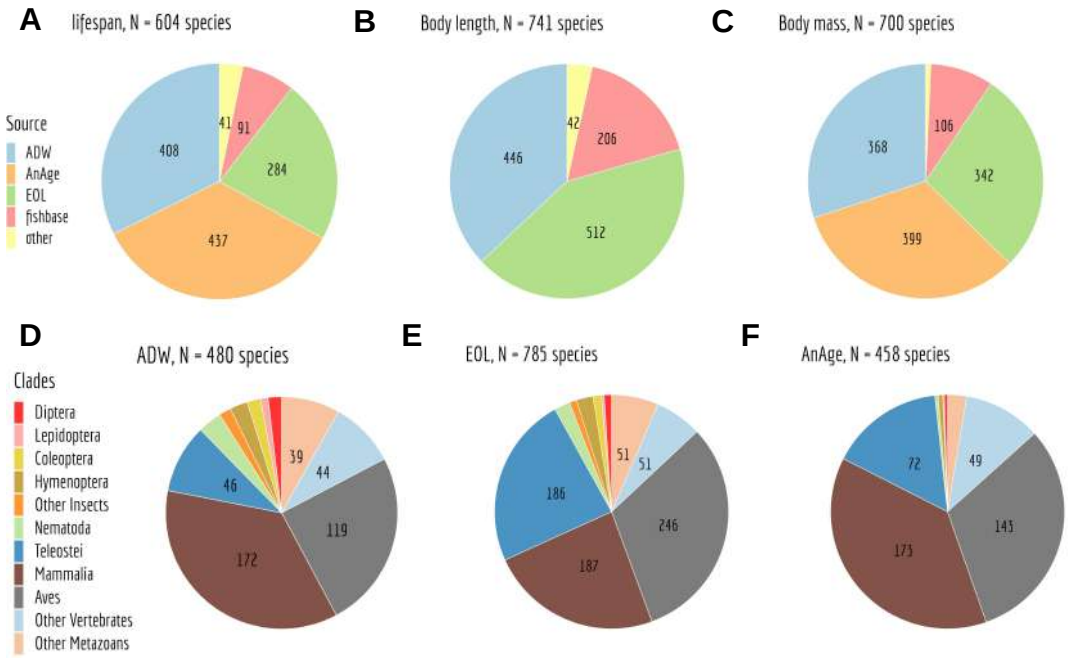


Figure 5.2: Representation of life history traits retrieved from diverse data sources. *Depiction of data origins for lifespan (**A**), body length (**B**), and body mass (**C**). Additionally, distribution of species and their respective clades with at least one recorded life history trait in ADW (**D**), EOL (**E**), and AnAge (**F**).*

predictive scores provided by the model.

We implemented an iterative selection process to identify the highest predicted answer containing relevant units and numeric values. To avoid redundancy, the selected answer was then removed from the text, and the process was repeated up to 10 times. The entire procedure was implemented in a Python script. We processed the script's output to restructure the obtained results.

Discrepancies between the manual approach and the other two methodologies were further re-investigated manually and corrected as needed after a further re-reading of the text. As a result, the curated dataset that we share reflects our highest level of confidence.

In total, our data collection effort resulted in the acquisition of life history traits for 969 metazoan species.

5.2.3 Acquisition of the reference genome sequence and annotations

Using the sra-tools software, we performed an automated identification of the reference genome for each species. Subsequently, we downloaded the annotation data in GFF format, the nucleotide coding sequences in FASTA format, and the peptide sequences in FASTA format from the NCBI database (Sayers *et al.*, 2022a).

5.2.4 dN/dS pipeline

We developed a pipeline to estimate the rate of non-synonymous substitutions divided by the synonymous substitutions rate (dN/dS), representing the frequency at which non-synonymous changes occur relative to synonymous ones. Since non-synonymous substitutions are commonly perceived as errors, dN/dS serves as a measure of the rate of erroneous substitutions per neutral substitution. This ratio is directly dependant of N_e as it is jointly determined by the distribution of selection coefficient of new mutations (s) and the magnitude of genetic drift as defined by N_e (Yang and Nielsen, 1998; Nielsen and Yang, 2003). The transcriptome-wide dN/dS is expected to rise over prolonged periods of small N_e due to the increasing number of slightly deleterious mutations reaching fixation (Ohta, 1992; Galtier, 2016).

Estimating the dN/dS necessitates the annotation of genes shared across all species, their evolutionary history depicted by a phylogenetic tree, and finally a comparative analysis of site evolution to derive the dN/dS ratio.

BUSCO genes identification

We used the BUSCO v.3.1.0 software to identify single-copy orthologous genes within three datasets selected from OrthoDB v9 (Zdobnov *et al.*, 2017): eukaryota (N=303 genes), embryophyta (N=1,440 genes) and metazoan (N=978 genes) sourced from BUSCOv3 (Waterhouse *et al.*, 2018; Seppey *et al.*, 2019; Manni *et al.*, 2021). The search was performed against the longest annotated protein sequences *per* gene within each genome.

Phylogenetic tree reconstruction

Due to the considerable time and resource demands associated with phylogenetic inference for large numbers of species, we employed a strategy in which the analysis was partitioned by clades. On initial releases of the database, which did not encompass all current species, we performed 3 comparable and independent analyses that rely on the three BUSCO datasets, corresponding to the following lineages: eukaryota, embryophyta and metazoa. For each BUSCO dataset, we selected a subset of species that matched the lineage of interest from the available database records at the time of analysis. All of these selected species underwent transcriptomic analyses (see Transcriptomic analyses). We then collected the longest corresponding proteins identified in each species for each BUSCO gene family. We removed proteins for which the amino acid sequence provided with the annotations did not perfectly correspond to the translation of the corresponding coding sequences. We then aligned the resulting sets of protein-coding sequences for each BUSCO gene, using the codon alignment option in PRANK v.170427 (Löytynoja and Goldman, 2008). We translated the codon alignments into protein alignments using the R package seqinr (Charif and Lobry, 2007).

A filter was applied to retain only genes for which enough species have been detected (85% of the analyzed species), reducing the eukaryota set to 126 genes (embryophyta

N=387 genes, metazoa N=731 genes). Then, species were removed from the analysis if they had less than 80% of the studied genes, reducing the number of studied species from 336 to 279 for the eukaryota BUSCO dataset (embryophyta 93 to 80 species, metazoa 293 to 257 species).

To infer the phylogenetic tree rapidly, we sub-sampled the resulting multiple alignments, selecting alignments with the highest number of species (eukaryota N=25 genes, embryophyta N=77 genes, metazoa N=146 genes). We then concatenated these alignments and kept sites that were aligned in most of the analyzed species (see information provided in the supplementary archive for more details). The final alignment for the eukaryota BUSCO dataset included 279 taxa (embryophyta N=80 species, metazoa N=257 species) taxa and 600 sites (embryophyta N=700 sites, metazoa N=3,000 sites). We used RAxML-NG (Kozlov *et al.*, 2019), to infer the species phylogeny on these final alignments. RAxML was set to perform one model *per* gene with a fixed empirical substitution matrix (LG), empirical amino acid frequencies from alignment (F) and 8 discrete GAMMA categories (G8). These parameters were specified in a partition file with one line *per* BUSCO gene multiple alignment. The analysis generated at least 10 starting trees. The best-scoring topology was kept as the final ML tree and 10 bootstrap replicates have been generated.

The phylogenetic trees were rooted using as a reference source the TimeTree phylogeny, which synthesizes data from numerous published studies, despite its incomplete representation of all species (Kumar *et al.*, 2022).

To encompass a broader spectrum of the species included in our latest database release, the one published here, we also reconstructed phylogenetic trees *per* clade. To do this, we divided the full set of metazoan species in 9 groups (Hymenoptera, Diptera grouped with Lepidoptera under the superorder Mecopterida, Nematoda, other insects, Aves, Mammalia, Teleostei, other vertebrates, and finally other invertebrates). We used as a basis for the analysis 73 highly prevalent metazoan BUSCO genes among the 731 genes preselected in the metazoa analysis. We applied the protocol described above to each individual clade. The resulting clade-specific trees were merged using outgroup species as a reference point to construct the complete metazoan phylogenetic tree.

dN/dS computation

We computed dN/dS ratios for BUSCO gene families that were present in at least 85 percent of the species under investigation. We conducted four independent analyses. We first analyzed each of the three BUSCO gene sets: eukaryota (N=126 genes), embryophyta (N=387 genes), metazoa (N=731 genes). We also performed an analysis '*per* clade', as explained above for the phylogenetic tree reconstruction, using the same 731 genes preselected in the metazoa analysis. Codon alignments obtained using PRANK (Löytynoja and Goldman, 2008) served as the basis for this estimation. To manage the computational memory demands during the substitution rate estimation step, we segmented the sequence alignments into clusters. Following the approach recommended

by Bolívar *et al.* (2019), these clusters were defined based on the average GC3 content across species, in order to group genes with similar parameters. We then concatenated the alignments within each group, obtaining alignments that were 200 kb long on average. This process yielded 13 groups for eukaryota (15 for embryophyta and 73 for metazoa). We used bio++ v.3.0.0 libraries (Dutheil and Boussau, 2008; Guéguen *et al.*, 2013; Bolívar *et al.*, 2019) to estimate the dN/dS on each branch of the phylogenetic tree, for each concatenated alignment.

In a first step, we used an homogeneous codon model implemented in bppml to infer the most likely branch lengths, codon frequencies at the root, and substitution model parameters. We used YN98 (F3X4) (Yang and Nielsen, 1998) substitution model, which allows for different nucleotide content dynamics across codon positions. In a second step, we used the MapNH substitution mapping method to count synonymous and non-synonymous substitutions (Dutheil *et al.*, 2012; Guéguen and Duret, 2018). We defined dN as the total number of non-synonymous substitutions divided by the total number of non-synonymous mutational opportunities, both summed across concatenated alignments, for each branch of the phylogenetic tree. Likewise, we defined dS as the total number of synonymous substitutions divided by the total number of synonymous mutational opportunities, both summed across concatenated alignments. The *per*-species dN/dS corresponds to the ratio between dN and dS, on the terminal branches of the phylogenetic tree. We also provide the dN and dS values for each branch within the phylogenetic trees.

For the '*per* clade' approach, the results pertaining to distinct clades were combined in a single table.

5.2.5 Transcriptomic analyses

We developed a pipeline facilitating the detection of alternative splicing events within genes. This process entails the selection of RNA-seq data, subsequent alignment to the reference genome, and the identification of splicing events through the recognition of introns. Utilizing the aligned transcriptomic data, we computed gene expression levels across each sample.

Selection of the RNA-seq samples

To extract RNA-seq data, we queried the Short Read Archive (SRA) database for samples where the library source was 'TRANSCRIPTOMIC' and the library strategy was 'RNA-seq'.

For perfect comparability of transcriptome data among species, we would need to have the same representation of individual tissues, developmental stages *etc.* for each species, with data generated with the same protocol by the same person. However, such data exist only for limited sets of species (*e.g.*, Cardoso-Moreira *et al.* (2019)). Here, we decided not to filter the RNA-seq samples on criteria pertaining to sample origin or experimental protocols, mainly because the relevant information is not always provided in sufficient detail in the SRA database (Leinonen *et al.*, 2011b). Moreover, depending

on the clade, the biological sample of origin can vary from "whole body" in insects, to specific tissues or cell types in mammals. Thus, perfectly comparable sample collections are difficult to obtain across such a broad phylogenetic scale.

Rather than filtering samples on these criteria prior to inclusion in the database, in GTDrift we provide users with the information collected from SRA for all RNA-seq samples. This information includes the library type, the date of extraction and the name of the laboratory that performed the experiment (see Description of the data available in GTDrift).

After evaluating the amount of RNA-seq data that is needed to evaluate global alternative splicing patterns for each species (see below), we decided to include a maximum of 50 RNA-seq samples *per* species in GTDrift. We included more than 50 samples for 150 species (43 embryophyta, 107 metazoa), for which we performed more detailed analyses, considering various tissues or developmental stages.

In the current version of GTDrift, the RNA-seq dataset encompasses a total of 491 distinct species, including 92 plants and 399 animals. (Fig. 3A).

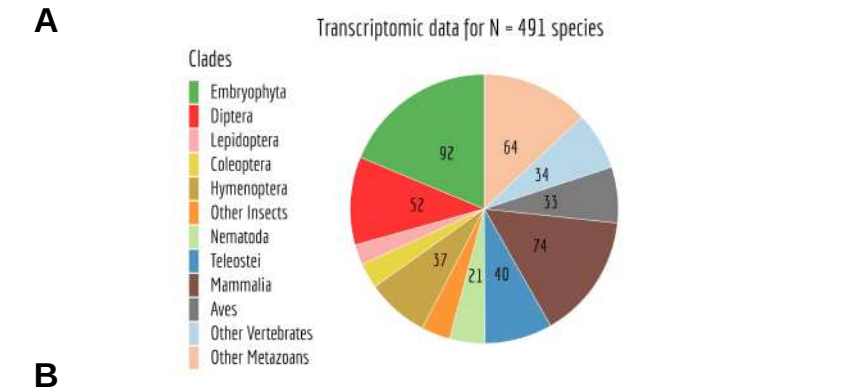
Indexing genomes and aligning RNA-seq data

The RNA-seq alignment phase represents the most time-consuming stage in the pipeline (Fig. 4), and can extend up to one week when utilizing 16 cores for each RNA-seq dataset, particularly for larger genomes such as those of mammals.

For this step, HISAT2 version 2.1.0 was employed to align RNA-seq reads to the respective reference genomes (Kim *et al.*, 2019). To enhance the sensitivity of splice junction detection, we constructed genome indexes incorporating annotated intron and exon coordinates along with genome sequences. The maximum permitted intron length was set at 2,000,000 base pairs. The processed and compressed files generated during this procedure can amass a size exceeding 20 terabytes.

We extracted intron coordinates from the HISAT2 alignments, utilizing a custom Perl script that scanned for CIGAR strings containing "N" characters, which indicate skipped regions in the reference sequence. For intron identification and quantification, we exclusively utilized uniquely mapped reads with a maximum mismatch fraction of 0.02. In the context of new intron identification, we imposed a minimum anchor length (*i.e.*, part of the read that spans each of the two exons flanking a given intron) of 8 base pairs. We then quantified intron splicing frequencies by including aligned reads with a minimum anchor length of 5 base pairs. We retained predicted introns exhibiting GT-AG, GC-AG, or AT-AC splice signals and determined the intron strand based on the splice signal.

Introns were assigned to genes if at least one of their boundaries was within 1 base pair of annotated exon coordinates, combined across all isoforms for each gene. Intron assignments were limited to those that could be unambiguously associated with a single gene. Notably, we differentiated between annotated introns, present in the reference genome annotations, and unannotated introns, identified through RNA-seq data and assigned to previously annotated genes.



C

Variable	Definition
RAS	Relative abundance of a given spliced isoform compared to other overlapping spliced isoforms
RANS	Relative abundance of a given spliced isoform compared to unspliced transcripts overlaping its splice sites
Major-isoform intron	Intron corresponding to a major isoform (*i.e* with RAS > 0.5 and RANS > 0.5)
Minor-isoform intron / SV	Splice variant (SV) = intron corresponding to a minor isoform (*i.e* with RAS ≤ 0.5 or RANS ≤ 0.5)

Figure 5.3: Species with transcriptomic data and alternative splicing estimation. *(cf Fig. 2A Bénitière et al. (2024)) **A**: Taxonomic distribution of the species for which transcriptomic data was included in GTDrift. **B**: Definition of the variables used to compute the relative splicing frequency of a focal intron, compared to splice variants with a common alternative splice boundary (RAS) or compared to the unspliced form (RANS):* N_s*: number of spliced reads corresponding to the precise excision of the focal intron;* N_a*: number of reads corresponding to alternative splice variants relative to this intron (i.e. sharing only one of the two intron boundaries);* N_u*: number of unspliced reads, co-linear with the genomic sequence. **C**: Definitions of the main variables used in this study. The definition of the variables corresponds to the one provided in Bénitière et al. (2024).*

We identified introns situated within protein-coding regions. To do this, for each protein-coding gene, we extracted annotated start and stop codon positions across all annotated isoforms. The minimum start codon and maximum end codon positions were identified, and introns located upstream or downstream of these extreme coordinates were considered as interrupting untranslated regions.

Alternative splicing variables

For each intron, we recorded two key variables: N_s representing the number of reads corresponding to the precise removal of the intron (referred to as spliced reads), and N_a representing the count of reads supporting alternative splicing events (*i.e.* spliced variants sharing only one of the two boundaries of the focal intron). Additionally, we denoted N_u as the count of unspliced reads that align linearly with the genomic sequence and span at

least 10 base pairs on both sides of an exon-intron junction. These definitions are visually clarified in (Fig. 3B,C). Subsequently, we introduced the relative measurement of the target intron's abundance compared to introns with a single alternative splice boundary ($\text{RAS} = \frac{N_s}{N_s + N_a}$), as well as relative to unspliced reads ($\text{RANS} = \frac{N_s}{N_s + \frac{N_u}{2}}$).

To compute these ratios, we required at least 10 reads in their denominators. Thus, we computed the RAS only when $(N_s + N_a) \geq 10$, and the RANS only when $(N_s + \frac{N_u}{2}) \geq 10$. We divided N_u by 2 because unspliced reads that span the two intron boundaries likely refer to the same intron retention event. If these conditions were not met, the resulting values were designated as unavailable (NA). These ratios were computed utilizing data from all available RNA-seq samples, unless explicitly specified (*e.g.* in sub-sampling analyses). Based on these ratios, we divided introns into three categories: major-isoform introns, defined as those introns that have RANS > 0.5 and RAS > 0.5 (these likely correspond to the introns of major isoforms (Gonzàlez-Porta *et al.*, 2013; Tress *et al.*, 2017a; Bénitière *et al.*, 2024)); minor-isoform introns, defined as those introns that have RANS ≤ 0.5 or RAS ≤ 0.5 (these introns are detected in a minority of transcripts); unclassified introns, which do not satisfy the above conditions.

Gene expression estimation

Gene expression levels were computed using Cufflinks version 2.2.1 (Trapnell *et al.*, 2010; Roberts *et al.*, 2011), utilizing the read alignments obtained with HISAT2 for each individual RNA-seq sample. We thus evaluated gene expression levels with the Fragment *Per* Kilobase of exon *per* Million mapped reads (FPKM) method. To determine the representative expression level of each gene, the mean FPKM was calculated across all samples, taking into consideration the sequencing depth of each sample, called 'weighted FPKM'. We used this measure to evaluate the relationship between alternative splicing rates and gene expression levels, within each species.

Estimation of the sequencing depth

We determined for each gene the union of all annotated exon coordinates (termed here exon blocks). Using bedtools v2.25.0 (Quinlan and Hall, 2010), we assessed the read coverage at each position of the exon blocks. The average exonic *per*-base read coverage was subsequently computed for each gene. The sequencing depth of a given sample was evaluated through the median *per*-base read coverage across BUSCO (Benchmarking Universal Single-Copy Orthologs) genes.

5.2.6 Data visualisation using a Shiny app

A Shiny app available at `https://lbbe-shiny.univ-lyon1.fr/ShinyApp-GTDrift/` was deployed to allow users to visualize and compare the summarized data (Chang *et al.*, 2024). Most of the graphics shown in this paper are directly reproducible from the app. In this app, users can also visualize intra-species variables, for example comparing introns

or gene characteristics. Furthermore, a specific tab is dedicated to the investigation of gene structure in relation to the splicing attributes found in the underlying database. Users can also visualize the phylogenetic tree and employ these trees for conducting Phylogenetic Generalized Least Square regression analyses.

The app is organized in several panels or "tabs" in the web page.

The tab 'Inter-species graphics' facilitates the comparison of genome characteristics across different species through graphical representation. Additionally, users have the option to upload their own data in a tab-separated text format, where each species is represented in a separate row, with the variables of interest organized in columns. An example of such a tabular dataset can be found in the repository.

The 'Inter-species Axis' tab explains the variables available in the 'Inter-species graphics' tab.

The 'Intra-species graphics' tab permits the exploration of characteristics within a species, focusing on introns or on genes. Furthermore, users have the ability to retrieve metadata related to BUSCO annotation, gene expression, or intron splicing events (see Methods).

The 'Intra-species Axis' tab describes the variables featured in the 'Intra-species graphics' tab.

Within the 'Gene structure' tab, users can delve into the introns detected in RNA-seq alignments for a specific gene. These introns are color-coded based on various criteria, including their location within the CDS or outside of it, as well as whether they are classified as major or minor-isoform introns (see Methods).

The 'Phylogenetic tree' tab facilitates the examination of phylogenetic trees used for conducting Phylogenetic Generalized Least Squares regression within the 'Inter-species graphics' tab.

5.2.7 Data and code availability

All processed data that we generated and used in this study, as well as the scripts that we used to analyze the data and to generate the figures, are available at the following Zenodo DOI: https://doi.org/10.5281/zenodo.10022493. The database is provided on Zenodo with the DOI: https://doi.org/10.5281/zenodo.10017653. Finally, the Shiny app is available at: https://lbbe-shiny.univ-lyon1.fr/ShinyApp-GTDrift/ and on Zenodo with the DOI: https://doi.org/10.5281/zenodo.10022520.

5.3 Results

5.3.1 Description of the data available in GTDrift

In GTDrift, we provide a manageable number of compressed data tables for each species processed via our pipeline (Fig. 4). Tables are stored in tab-delimited text format, which makes them easy to access for users with experience in bioinformatics. They are user-

friendly because of the simplicity of their contents. To access these tables, users can visit the Zenodo DOI: https://doi.org/10.5281/zenodo.10017653 and select their desired data type. The data can also be easily explored through a web application written in Shiny at https://lbbe-shiny.univ-lyon1.fr/ShinyApp-GTDrift/. Data exploration is thus easily accessible even for users who do not have a background in bioinformatics.

Our database is centered around transcriptomics data. At the time of publication, the database contained over 15,935 RNA-seq samples distributed over 491 embryophytes and metazoans (Fig. 1), providing gene expression and alternative splicing events data. Additionally, we have enriched the database with annotations for orthologous single-copy genes (BUSCO genes) and proxies of effective population size, including the molecular evolutionary rate dN/dS and life history traits such as longevity, body mass, body length. We used similar types of data in our recent publication exploring the relationship between random genetic drift and alternative splicing patterns (Bénitière *et al.*, 2024). However, here we provide considerably more data, for 1,507 species compared to 53 in this publication.

Below, we provide information on the data types that are currently available in GT-Drift.

Life history traits

The table labeled 'life_history_traits.tab' comprises values pertaining to three distinct traits (longevity, body mass, and body weight), for 969 species. This table includes bibliographic references which attribute these values to each species. The species are defined by their scientific names and by the corresponding NCBI taxonomy identifier (taxID).

Protein-coding sequence evolution features

We provide estimates of the representative dN/dS ratio for most species (N=1,324 species after filtering for a sufficient number of annotated orthologous genes). The data are available in the directory 'dNdS'.

We provide the phylogenetic tree of the studied species, with the dN/dS ratios as branch lengths, in the Newick file format. We provide this data separately for the four approaches used to estimate the ratios dN/dS, using the eukaryota, embryophyta or metazoa BUSCO gene sets, or a different gene set for each clade (Methods). In addition, we provide a table comprising the dN and dS values for each terminal branch of the phylogenetic tree, along with the species scientific name and NCBI taxonomy ID, for each of the four approaches.

Gene expression

In the 'Transcriptomic' directory, each species is represented by a dedicated table named 'by_gene_analysis.tab.gz'. This table contains annotated gene coordinates, the mean and median FPKM (Fragments *Per* Kilobase of exon *per* Million mapped reads) across samples. Additionally, the table includes information about RNA-seq read coverage for exonic

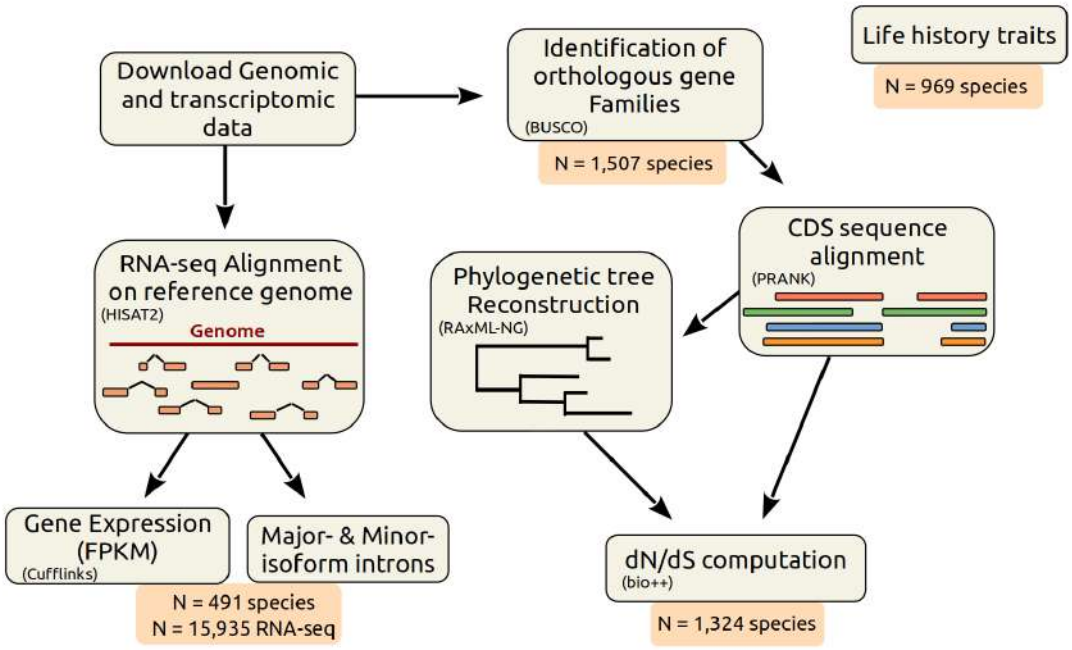


Figure 5.4: Description of the bioinformatic analysis pipeline. *(Adapted from Supplementary Fig. 11 Bénitière et al. (2024)) First, we retrieved genomic sequences and annotations from the NCBI Genomes database. We aligned RNA-seq reads on the corresponding reference genomes with HISAT2. We used these alignments to estimate various variables related to splicing patterns (see Fig. 2), to compute the AS rate, and to estimate gene expression using Cufflinks. To compute the dN/dS ratios, we first identified BUSCO genes with BUSCOv3 and aligned their coding sequences (CDS) using PRANK (codon model). We reconstructed a phylogenetic tree using RAxML-NG. Using bio++, we estimated dN/dS along the phylogenetic tree on concatenated alignments. This pipeline was previously used in Bénitière et al. (2024).*

regions for each gene, including the total read coverage across samples. The individual gene expression data for each RNA-seq experiment can be accessed within the 'RUN' directory. The data are provided in a separate directory for each SRA accession number. The file 'by_gene_db.tab.gz' containing the exon coverage and the FPKM measured for each gene corresponding in line to the previous file 'by_gene_analysis.tab.gz'.

Alternative splicing data

For each species, we provide a summarized table named 'by_intron_analysis.tab.gz', containing for each intron the cumulative counts of spliced reads (N_s), the number of reads supporting alternative splicing of this introns (N_a), and the number of unspliced reads overlapping with this intron (N_u) detected through RNA-seq analysis (Methods). This table contains data combined across all analyzed RNA-seq samples. Detailed information for individual RNA-seq experiments can be found within the 'RUN' directory, in the file 'by_intron_db.tab.gz'. In these files, introns are listed in the same order as in the file 'by_intron_analysis.tab.gz'.

RNA-seq sample description

In the file named 'SRAruninfo.tab', we provide information extracted from the SRA database, for each RNA-seq sample. Depending on the sample, this information can include the library source, the tissue from which the sample is derived, the sex of the sampled individual, the lab that conducted the analysis, the methods used to prepare the library, *etc.*

BUSCO gene identification

In the directory 'BUSCO_annotations', we provide the correspondence between NCBI gene identifiers and BUSCO gene identifiers, determined for three distinct BUSCO datasets: eukaryota, metazoa, and embryophyta.

5.3.2 Data quality validation

Acquiring life history traits

To facilitate the acquisition of life history traits, we have devised and shared a pipeline that uses an automatic screening technique complemented by a machine learning method.

To assess the effectiveness of the automatic screening technique that we used to extract life history traits from various databases, we conducted a comparative analysis, contrasting it with the manual methodology. We also compared it to the machine learning (ML) approach for the ADW database. The screening procedure yielded accurate information with varying false positive rates depending on the source database, as follows: AnAge (98.9% accuracy; 0% false positive), fishbase (100%; 0.2%), EOL (94.5%; 0.18%), and ADW (88%; 5.4%). These results highlight the utility of our screening pipeline for identifying three key life history traits across AnAge, EOL, ADW, and fishbase databases.

For the ADW database, the ML approach exhibited a slight advantage over the screening method, and its results did not completely align with those obtained through the screening approach. Specifically, for life history traits, the ML approach correctly retrieved 89.7% of the results obtained through the manual approach, while introducing a 9.2% false positive rate.

When combining both the ML approach and the screening process, we achieved a 95% accuracy rate in identifying positive cases. However, a 7% error rate persisted in this merged approach.

In GTDrift, we provide data corresponding to a synthesis of the three methodologies including only manually-checked values (Methods).

Estimating the intensity of random drift

As expected, a positive correlation is observed in Fig. 5A,B between the different life history traits, used as indirect predictors of the effective population size (N_e) (Romiguier *et al.*, 2014a; Waples, 2016; Figuet *et al.*, 2016; Galtier, 2016; Weyna and Romiguier,

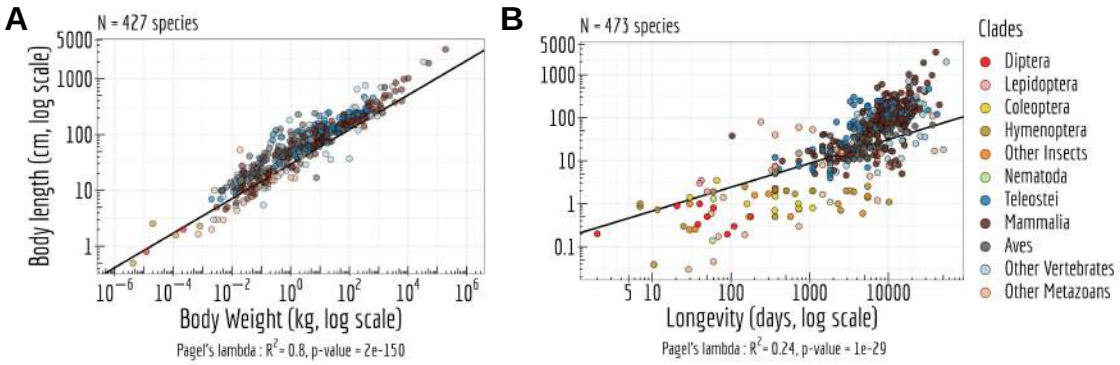


Figure 5.5: N_e proxies. ***A****: Relationship between body length (cm, log scale) and longevity (days, log scale) of the organism. Each dot represents one species (colored by clade).* ***B****: Relationship between body length (cm, log scale) and the body weight (kg, log scale).* ***A,B****: Pagel's lambda model is used to take into account the phylogenetic structure of the data in a regression model.*

2020).

When examining the dN/dS ratio across distinct time scales and using various BUSCO datasets, we consistently observe comparable dN/dS ratios at terminal branches. This uniformity across a range of methodological approaches highlights their concordance (Fig. 6A,B).

Furthermore, the observed dN/dS ratio are significantly correlated with proxies derived from life history traits (Fig. 7A,B) as reported in the literature (Romiguier *et al.*, 2014a; Figuet *et al.*, 2016).

5.3.3 Quality of genome annotations

To assess gene expression levels and alternative splicing patterns, the quality of genome annotations is of paramount importance. We evaluated genome annotation quality by examining the presence of BUSCO genes. We note that the results depend on the BUSCO dataset that is used as a starting point. When using the BUSCO dataset designed for eukaryota, which comprises 303 genes, we have effectively identified nearly all single-copy orthologous genes, and this feature exhibits a high degree of homogeneity across different species (Fig. 8). However, the aves clade demonstrates a deficiency in the number of BUSCO genes compared to the anticipated count based on BUSCO expectations. This is expected given the known genome incompleteness problem for this clade, due to the presence of GC-rich chromosomes (Li *et al.*, 2022).

Because the eukaryota BUSCO gene set is limited, we also performed gene identification for the metazoa and embryophyta BUSCO datasets, leading to substantially larger collections of genes. Specifically, we detected 978 BUSCO genes for the metazoa dataset and 1,440 genes for the embryophyta dataset.

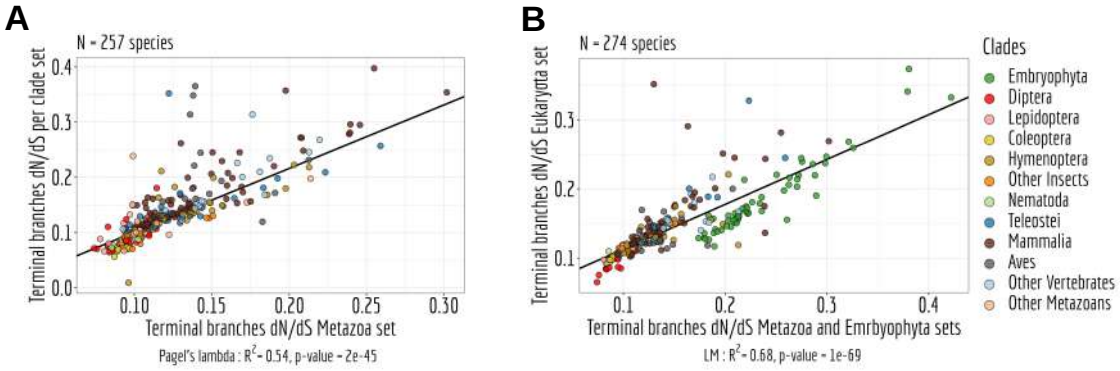


Figure 5.6: Reproducibility of the *dN/dS* ***ratio.*** ***A****: Relation between the dN/dS ratio on terminal branches of the phylogenetic tree of the metazoa set compared to the ones measured in the per clades set.* ***B****: Relation between the dN/dS ratio on terminal branches of the phylogenetic tree of the eukaryota set compared to the ones measured in the embryophyta and the metazoa set.* ***A,B****: LM stands for Linear regression Model and Pagel's lambda model is used to take into account the phylogenetic structure of the data in a regression model.*

5.3.4 Spliced introns classification

A significant body of literature has consistently reported that the majority of genes typically exhibit one predominant isoform (Gonzàlez-Porta *et al.*, 2013; Tress *et al.*, 2017a). This isoform is commonly termed "major isoform". Here, we aimed to assess the influence of sequencing depth on the identification of major-isoform introns, that is, those introns that belong to major isoforms (see Alternative splicing variables). Employing the model organism *Drosophila melanogaster*, we randomly selected between 1 and 20 RNA-seq samples. For each subset of samples, we computed the median read coverage across the exons of BUSCO genes, providing a standardized measure of transcriptome sequencing depth that can be compared across different species. Additionally, we tallied the count of introns falling into various categories (major-isoform introns, minor-isoform introns or unclassified introns - see Methods) for each subset of samples. This entire process was repeated 10 times (Fig. 9A).

As expected, we observed that the number of major-isoform introns that could be identified increased with greater sequencing depth until it reached a threshold of 200 read coverage *per* base (Fig. 9A). Beyond this threshold, no additional major-isoform introns are discernible. Simultaneously, the count of unclassified introns decreased to nearly zero, indicating that introns newly detected above the 200-read coverage threshold predominantly consisted of minor-isoform introns that shared a boundary with a major intron. Indeed, the count of minor-isoform introns continued to rise steadily beyond this point.

We then assessed the proportion of annotated introns that fall within the categories defined above. Our results reveal that the majority of species exhibit well-annotated major-isoform introns, indicating the accuracy of the intron annotation (Fig. 9B). Additionally, as sequencing depth increases, we observed a decreasing fraction of annotated minor-isoform introns. This trend is consistent with expectations, given that higher se-

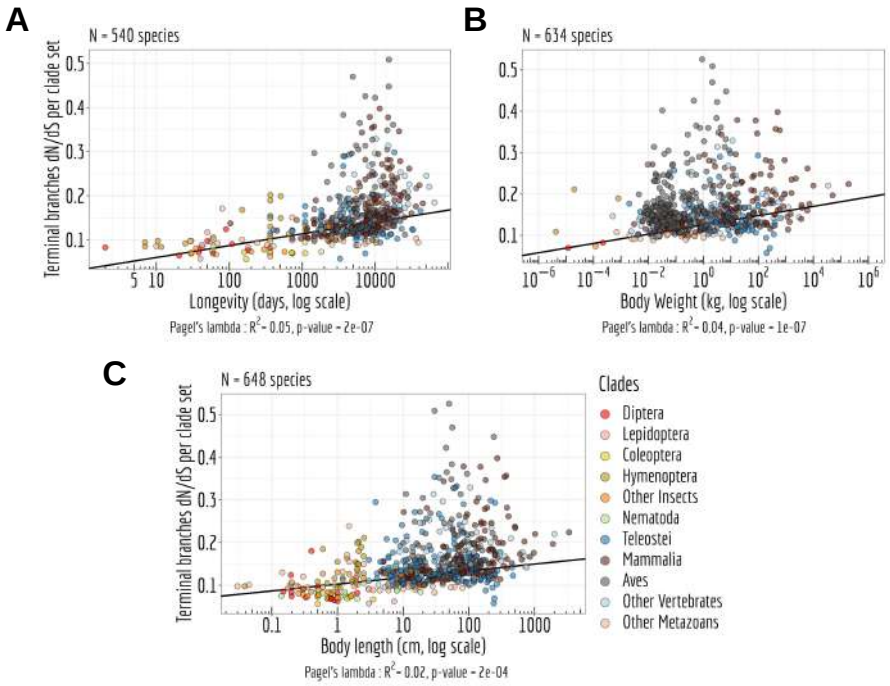


Figure 5.7: Interplay between N_e proxies. *Correlation between the dN/dS ratio on terminal branches of the phylogenetic tree of the per clade set and life history traits: longevity (days, log scale) (**A**), body weight (kg, log scale) (**B**), body length (cm, log scale) (**C**). **A**,**B**,**C**: Pagel's lambda model is used to take into account the phylogenetic structure of the data in a regression model.*

quencing depth expands the pool of rare variants and potential spontaneous errors that may not have been previously observed. It is important to note that there appears to be no inherent limit to this phenomenon, as the intricacies of alternative splicing machinery can give rise to unpredictable errors (Bénitière *et al.*, 2024).

5.4 Discussion

GTDrift is a comprehensive data resource facilitating investigations of genomic and transcriptomic characteristics alongside indicators of genetic drift intensity for distinct species. Notably, this resource offers information on life history traits, including longevity, adult body length, and body mass, for a curated set of 969 species. Additionally, it provides estimates of the ratio between the rate of non-synonymous substitutions over synonymous substitutions (dN/dS) for 1,324 species.

For individual species, intron-centered alternative splicing frequencies, gene expression levels, and sequencing depth statistics have been systematically quantified and shared, encompassing more than 15,935 RNA-seq samples across 491 species. To enable cross-species comparisons, orthology predictions for conserved single-copy genes are provided, based on BUSCO gene sets, encompassing a total of 1,507 eukaryotic species, including 1,414 animals and 93 green plants, along with phylogenetic trees to account for phylogenetic inertia.

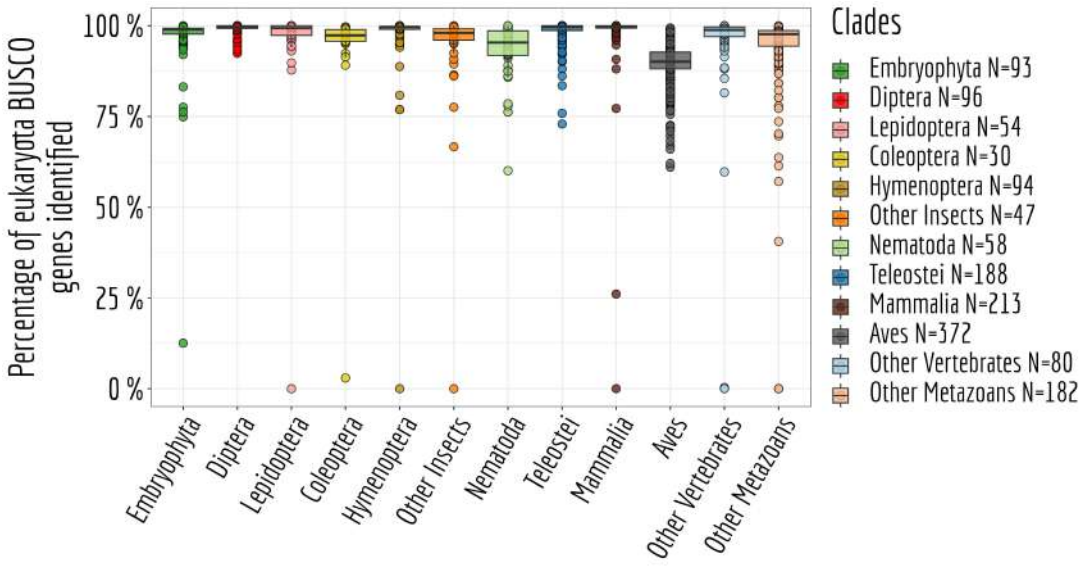


Figure 5.8: BUSCO genes annotation. *Proportion of BUSCO genes, from the BUSCO gene set eukaryota (N=303 genes), identified in each species.*

The number of species per data type varies due to different limitations: availability of life history traits data; completeness of gene annotations for dN/dS calculation; computational resources and availability of RNA-seq samples for transcriptomic analysis (Fig. 4).

These pre-processed data streamlines the work for those interested in investigating the impact of drift on biological processes across a wide range of species. All data are provided in flat files, which enable downstream computational analyses and render GT-Drift mainly aimed at users with some computational skills. Nonetheless, to enhance accessibility, we have developed a user-friendly Shiny app that facilitates database exploration and allows for species-specific data downloads (available at `https://lbbe-shiny.univ-lyon1.fr/ShinyApp-GTDrift/`).

5.4.1 Cautionary considerations in utilizing N_e proxies

Users should bear in mind that the scientific community has yet to establish the most adequate proxies for effective population size. A prominent hypothesis suggests that these proxies are associated with the number of individuals (N). Indeed, species with greater longevity and larger body mass tend to be less abundant within their ecological niche due to resource (mass) and spatial (length, mass) requirements (Damuth, 1981; Nee *et al.*, 1991; White *et al.*, 2007). Therefore, variations in life history traits should correspond to variations in the number of individuals (N), which subsequently impact N_e.

When using the dN/dS ratio as a proxy for N_e, rather than focusing on correlations with the population census, we evaluated the efficiency of natural selection to purge deleterious mutations. This efficiency can be represented as the product of N_e and 's', which denotes the selection coefficient. The extent to which a well-estimated dN/dS

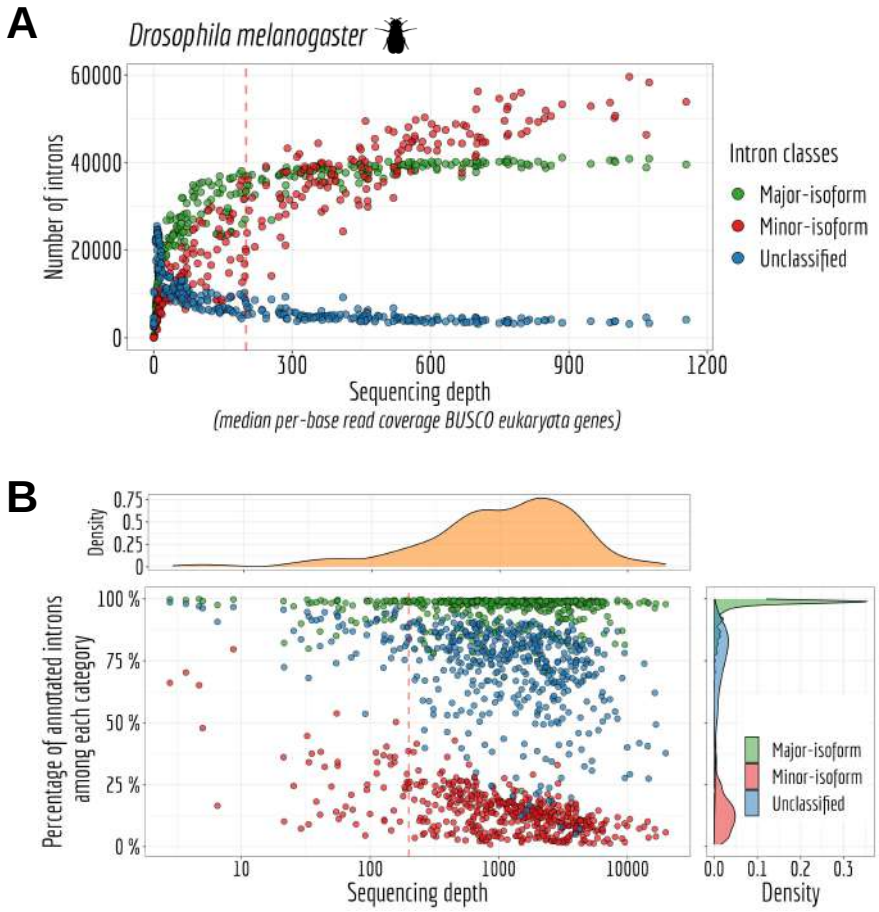


Figure 5.9: Sequencing depth impact on intron classification. ***A****: Number of major (RANS > 0.5 and RAS > 0.5), minor (RANS ≥ 5% or RAS ≥ 5%) and unclassified introns for* Drosophila melanogaster*. The sequencing depth is measured by taking the median per-base read coverage across BUSCO genes from eukaryota gene set.* ***B****: Per species major-isoform introns, minor-isoform introns and undetermined introns* (N_s) ≥ 10*) annotated proportion and sequencing depth measured by taking the median per-base read coverage eukaryota BUSCO genes.*

ratio can be considered as a proxy for N_e remains a subject of debate. Notably, when the rate of synonymous substitutions dS exceeds 1, it indicates a point of saturation where multiple substitutions occur *per* site, rendering dS susceptible to considerable noise due to the challenge of accurately identifying the number of substitutions at given sites. In such cases, the dN component can often still be reliably determined. Given that non-synonymous substitutions have a lower rate compared to synonymous ones, dN reaches a saturation point at a later stage.

Moreover, when the evolutionary time frame is relatively short, characterized by small dS values, the variants under examination are primarily attributed to polymorphism rather than fixed substitutions. In such cases, we are not effectively measuring substitution rates. Consequently, the discussion also revolves around determining a divergence threshold, above which we could assume that dS and dN predominantly represent substitutions, with minimal influence from polymorphism. In this perspec-

tive, the expanding polymorphism data could potentially serve as a means to distinguish between polymorphism and substitutions, offering a more efficient approach to investigate dN/dS (Mugal *et al.*, 2014).

Overall, we found that the various N_e proxies were significantly correlated, even when accounting for the underlying phylogenetic structure. Thus, our dataset, which encompasses information on dN and dS across all branches of the phylogenetic trees, holds the potential to estimate the long-term effective population size (N_e) and its interaction with life history traits over time.

5.4.2 Comparing transcriptomic data

In our study, we have identified BUSCO genes for the eukaryota, metazoa, or embryophyta BUSCO reference gene sets. To ensure meaningful comparisons between species with a sufficient number of detected BUSCO genes, we evaluated the median RNA-seq coverage of these BUSCO genes. As demonstrated in Data quality validation, the median *per*-base read exonic RNA-seq coverage of BUSCO genes is a good indicator of the power to detect alternative splicing patterns. We believe that, for the inclusion of additional species, an examination of the RNA-seq read coverage on BUSCO genes is needed to ensure that we could identify major-isoform introns and analyze alternative splicing patterns.

Additionally, it is essential to assess the completeness of the genome and of the annotation, which can be estimated based on the number of identified BUSCO genes. Some species may have a limited number of well-annotated BUSCO genes, or global gene duplications may result in the presence of two copies of a BUSCO gene, which no longer qualifies as a single copy gene.

Our RNA-seq description table offers users access to information collected from the Sequence Read Archive (SRA) for the RNA-seq datasets under study. This table enables users to filter and select RNA-seq data that align with their specific research needs. Users can tailor their selection based on factors such as sex, tissue, or protocol. Depending on the research question that is asked, it may be important to extract and analyse RNA-seq samples that were generated for the same biological conditions. We provide this information so that GTDrift users are able to filter the data as needed.

To facilitate cross-species comparisons, especially in the context of alternative splicing and gene expression, users can make use of BUSCO gene sets, which should exhibit consistent expression patterns, functionality, and evolutionary constraints across diverse species. However, users should thoroughly validate this assumption and proceed with vigilance.

5.4.3 Conclusion

In conclusion, we are confident that the GTDrift database can be a valuable resource for studies aiming to investigate the relationship between the intensity of genetic drift, genomic and transcriptomic characteristics.

Acknowledgements

We thank Loïc Guille for his contribution to an initial pilot study, Tristan Lefébure for insightful discussions and Laurent Guéguen for his help on dN/dS analyses. Computational analyses were performed using the computing facilities of the CC LBBE/PRABI and the Core Cluster of the Institut Français de Bioinformatique (IFB) (ANR-11-INBS-0013). Silhouette images of taxonomic Families originate from PhyloPic developed and maintained by Mike Keesey available at https://www.phylopic.org/.

Author contributions statement

F.B. conceived the pipeline and conducted the analyses. F.B. and A.N. drafted the manuscript. All authors reviewed the manuscript.

Funding

This work was funded by the French National Research Agency (ANR-20-CE02-0008-01 "NeGA" and ANR-17-CE12-0019-01 "LncEvoSys").

Competing interests

The authors declare no conflicts of interest.

6

Random genetic drift sets an upper limit on mRNA splicing accuracy in metazoans

The second objective of my thesis is to explore the impact of random genetic drift on alternative splicing in metazoans. Indeed, most eukaryotic genes undergo alternative splicing (AS), but the overall functional significance of this process remains a controversial issue. It has been noticed that the complexity of organisms (assayed by the number of distinct cell types) correlates positively with their genome-wide AS rate. This has been interpreted as evidence that AS plays an important role in adaptive evolution by increasing the functional repertoires of genomes.

However, this observation also fits with a totally opposite interpretation: given that 'complex' organisms tend to have small effective population sizes (N_e), they are expected to be more affected by genetic drift, and hence more prone to accumulate deleterious mutations that decrease splicing accuracy. Thus, according to this "drift barrier" theory, the elevated AS rate in complex organisms might simply result from a higher splicing error rate.

To test this hypothesis, based on a pre-release of GTDrift, we analyzed 3,496 transcriptome sequencing samples to quantify AS in 53 metazoan species spanning a wide range of N_e values. Our results led to a published paper where we showed a negative correlation between N_e proxies and the genome-wide AS rates among species, consistent with the drift barrier hypothesis. This pattern is dominated by low abundance isoforms, which represent the vast majority of the splice variant repertoire. We show that these low abundance isoforms are depleted in functional AS events, and most likely correspond to errors. Conversely, the AS rate of abundant isoforms, which are relatively enriched in functional AS events, tends to be lower in more complex species.

All these observations are consistent with the hypothesis that variation in AS rates across metazoans reflects the limits set by drift on the capacity of selection to prevent gene expression errors.

Random genetic drift sets an upper limit on mRNA splicing accuracy in metazoans

Florian Bénitière[1,2], Anamaria Necsulea[1], Laurent Duret[1]

[1]Laboratoire de Biométrie et Biologie Évolutive, UMR CNRS 5558 LBBE, Université Claude Bernard Lyon 1, Villeurbanne, France
[2]Laboratoire d'Ecologie des Hydrosystèmes Naturels et Anthropisés, UMR CNRS 5023 LEHNA, Université Claude Bernard Lyon 1, Villeurbanne, France

Contents

6.1 Introduction

Eukaryotic protein-coding genes are interrupted by introns, which have to be excised from the primary transcript to produce functional mRNAs that can be translated into proteins. The removal of introns from primary transcripts can lead to the production of diverse mRNAs, *via* the differential use of splice sites. This process of alternative splicing (AS) is widespread in eukaryotes (Chen *et al.*, 2014), but its 'raison d'être' (adaptive or not) remains elusive. Numerous studies have shown that some AS events are functional, *i.e.* that they play a beneficial role for the fitness of organisms, either by allowing the production of distinct protein isoforms (Graveley, 2001) or by regulating gene expression post-transcriptionally (McGlincy and Smith, 2008; Hamid and Makeyev, 2014). However, other AS events are undoubtedly not functional. Like any biological machinery, the spliceosome occasionally makes errors, leading to the production of aberrant mRNAs, which represent a waste of resources and are therefore deleterious for the fitness of the organisms (Hsu and Hertel, 2009; Gout *et al.*, 2013). The splicing error rate at a given intron is expected to depend both on the efficiency of the spliceosome and on the intrinsic quality of its splice signals. The information required in cis for the removal of each intron resides in 20 to 40 nucleotide sites, located within the intron or its flanking exons (Lynch, 2006). Besides the two splice sites that are essential for the splicing reaction (almost always GT for the donor and AG for the acceptor), all other signals tolerate some sequence flexibility. Population genetics principles state that the ability of selection to promote beneficial mutations or eliminate deleterious mutations depends on the intensity of selection (s) relative to the power of random genetic drift (defined by the effective population size, N_e): if the selection coefficient is sufficiently weak relative to drift ($|N_e s| \ll 1$), alleles behave as if they are effectively neutral. Thus, random drift sets an upper limit on the capacity of selection to prevent the fixation of alleles that are sub-optimal (Kimura *et al.*, 1963; Ohta, 1973). This so-called "drift barrier" (Lynch, 2007a) is expected to affect the efficiency of all cellular processes, including splicing. Hence, species with low N_e should be more prone to make splicing errors than species with high N_e.

The extent to which AS events correspond to functional isoforms or to errors is a contentious issue (Bhuiyan *et al.*, 2018; Tress *et al.*, 2017b; Blencowe, 2017; Tress *et al.*, 2017a). In humans, the set of transcripts produced by a given gene generally consists of one major transcript (the 'major isoform'), which encodes a functional protein, and of multiple minor isoforms (splice variants), present in relatively low abundance, and whose coding sequence is frequently interrupted by premature termination codons (PTCs) (Tress *et al.*, 2017a; Gonzàlez-Porta *et al.*, 2013). Ultimately, less than 1% of human splice variants lead to the production of a detectable amount of protein (Abascal *et al.*, 2015). Furthermore, comparison with closely related species showed that AS patterns evolve very rapidly (Barbosa-Morais *et al.*, 2012; Merkin *et al.*, 2012) and that alternative splice sites present little evidence of selective constraints (Pickrell *et al.*, 2010). All these observations are consistent with the hypothesis that a vast majority of splice variants observed in human transcriptomes simply correspond to erroneous tran-

scripts (Pickrell *et al.*, 2010). However, some authors argue that a large fraction of AS events might in fact contribute to regulating gene expression. Indeed, PTC-containing splice variants are recognized and degraded by the non-sense mediated decay (NMD) machinery. Thus, AS can be coupled with NMD to modulate gene expression at the post-transcriptional level (McGlincy and Smith, 2008; Hamid and Makeyev, 2014). This AS-NMD regulatory process does not involve the production of proteins and does not necessarily imply strong evolutionary constraints on splice sites. Thus, based on these observations, it is difficult to firmly refute selectionist or non-adaptive models.

The analysis of transcriptomes from various eukaryotic species showed substantial variation in AS rates across lineages, with the highest rate in primates (Barbosa-Morais *et al.*, 2012; Chen *et al.*, 2014; Mazin *et al.*, 2021). Interestingly, the genome-wide average AS level was found to correlate positively with the complexity of organisms (approximated by the number of cell types) (Chen *et al.*, 2014). This correlation was considered as evidence that AS contributed to the evolution of complex organisms by increasing the functional repertoire of their genomes (Chen *et al.*, 2014). This pattern is often presented as an argument supporting the importance of AS in adaptation (Verta and Jacobs, 2022; Singh and Ahi, 2022; Wright *et al.*, 2022). However, this correlation is also compatible with a totally opposite hypothesis. Indeed, eukaryotic species with the highest level of complexity correspond to multi-cellular organisms with relatively large body size, which tend to have small effective population sizes (N_e) (Lynch and Conery, 2003; Figuet *et al.*, 2016). Thus, the higher AS rate observed in 'complex' organisms might simply reflect an increased rate of splicing errors, resulting from the effect of the drift barrier on the quality of splice signals (Bush *et al.*, 2017).

To assess this hypothesis and evaluate the impact of genetic drift on alternative splicing patterns, we quantified AS rates in 53 metazoan species, covering a wide range of N_e values, and for which high-depth transcriptome sequencing data were available. We show that the genome-wide average AS rate correlates negatively with N_e, in agreement with the drift barrier hypothesis. This pattern is mainly driven by low abundance isoforms, which represent the vast majority of splice variants and most likely correspond to errors. Conversely, the AS rate of abundant splice variants, which are enriched in functional AS events, show the opposite trend. These results support the hypothesis that the drift barrier sets an upper limit on the capacity of selection to minimize splicing errors.

6.2 Results

6.2.1 Genomic and transcriptomic data collection

To analyze variation in AS rates across metazoans, we examined a collection of 69 species for which transcriptome sequencing (RNA-seq) data, genome assemblies, and gene annotations were available in public databases. We focused on vertebrates and insects, the two metazoan clades that were the best represented in public databases when we initiated this project. To be able to compare average AS rates across species, we needed

to control for several possible sources of biases. First, given that AS rates vary across genes (Saudemont *et al.*, 2017), we had to analyze a common set of orthologous genes. For this purpose, we extracted from the BUSCO database (Seppey *et al.*, 2019) a reference set of single-copy orthologous genes shared across metazoans (N=978 genes), and searched for their homologues in each species in our dataset. We retained for further analyses those species for which at least 80% of the BUSCO metazoan gene set could be identified (N=67 species; see Materials & Methods). Second, we had to ensure that RNA-seq read coverage was sufficiently high in each species to detect splicing variants. Indeed, to be able to detect AS at a given intron, it is necessary to analyze a minimal number of sequencing reads encompassing this intron (we used a threshold of N=10 reads). To assess the impact of sequencing depth on AS detection, we conducted a pilot analysis with two species (*Homo sapiens* and *Drosophila melanogaster*) for which hundreds of RNA-seq samples are available. This analysis (detailed in Appendix Fig. A.1) revealed that AS rate estimates are very noisy when sequencing depth is limited, but that they converge when sequencing is high enough. We therefore kept for further analysis those species for which the median read coverage across exonic regions of BUSCO genes was above 200 (Appendix Fig. A.1). Our final dataset thus consisted of 53 species (15 vertebrates and 38 insects; Fig. 1A), and of 3,496 RNA-seq samples (66 *per* species on average). In these species, the number of analyzable annotated introns (*i.e.* encompassed by at least 10 reads) among BUSCO genes ranges from 2,032 to 10,981 (which represents 88.6% to 99.6% of their annotated introns; Appendix Tab. A.1). It should be noted that analyzed samples originate from diverse sources; however, they are very homogenous in terms of sequencing technology (99% of RNA-seq samples sequenced with Illumina platforms; refer to `Data10-supp.tab` in the Zenodo data repository).

6.2.2 Proxies for the effective population size (N_e)

Effective population sizes (N_e) can in principle be inferred from levels of genetic polymorphism. However, population genetics data are lacking for most of the species in our dataset. We therefore used two life history traits that were previously proposed as proxies of N_e in metazoans (Waples, 2016; Weyna and Romiguier, 2020; Figuet *et al.*, 2016): body length and longevity (Materials & Methods; Appendix Tab. A.2). An additional proxy for N_e can be obtained by studying the intensity of purifying selection acting on protein sequences, through the dN/dS ratio (Kryazhimskiy and Plotkin, 2008). To evaluate this ratio, we aligned 922 BUSCO genes, reconstructed the phylogenetic tree of the 53 species (Fig. 1A) and computed the dN/dS ratio along each terminal branch (Materials & Methods).

We note that these three proxies provide "inverse" estimates of N_e, meaning that species with high longevity, large body length and/or elevated dN/dS values tend to have low N_e values. As expected, these different proxies of N_e are positively correlated with each other ($p < 1x10^{-3}$, Fig. 1B,C). We note however that these correlations are not very strong. It thus seems likely that none of these proxies provides a perfect

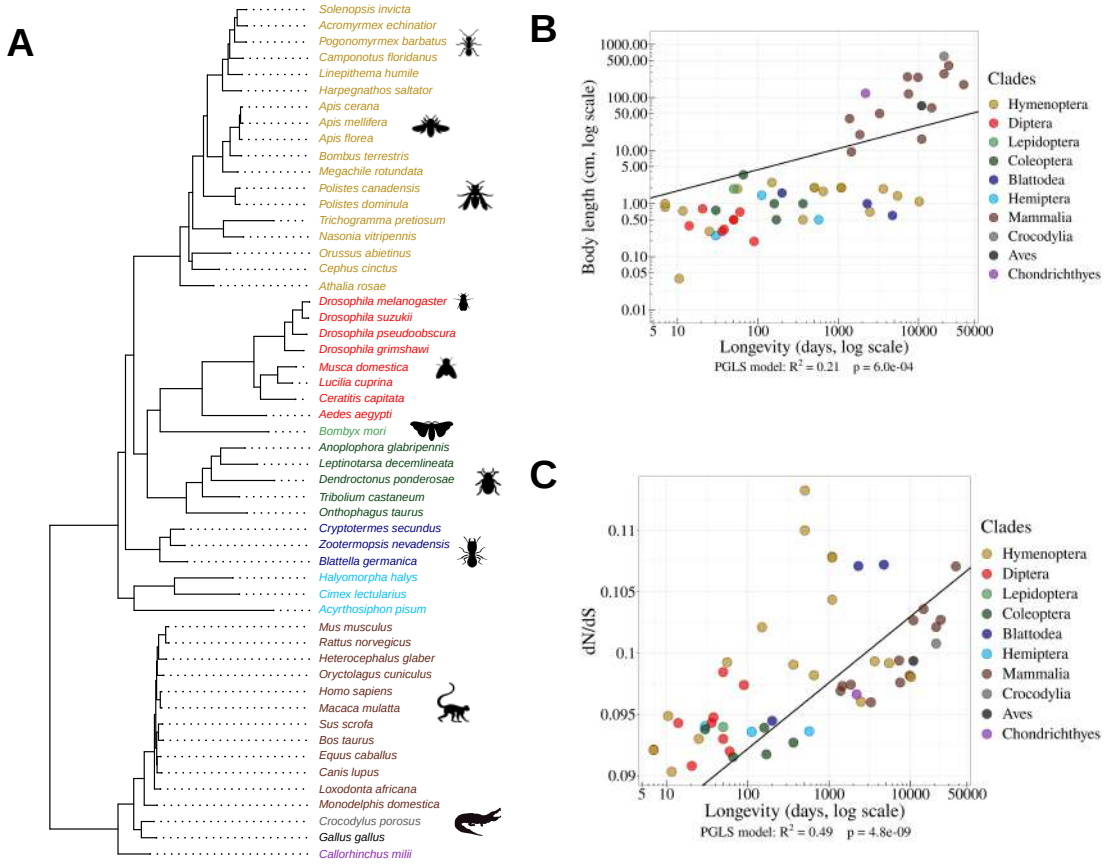


Figure 6.1: Species phylogeny and N_e proxies. ***A****: Phylogenetic tree of the 53 studied species (15 vertebrates and 38 insects).* ***B****: Relationship between body length (cm, log scale) and longevity (days, log scale) of the organism. Each dot represents one species (colored by clade, as in the species tree in panel A).* ***C****: Relationship between longevity (days, log scale) and the dN/dS ratio on terminal branches of the phylogenetic tree (Materials & Methods).* ***B,C****: PGLS stands for Phylogenetic Generalized Least Squared regression, which takes into account phylogenetic inertia (Materials & Methods).*

estimate of N_e. To take phylogenetic inertia into account, all cross-species correlations presented here were computed using Phylogenetic Generalized Least Squared (PGLS) regression (Freckleton *et al.*, 2002).

6.2.3 Alternative splicing rates are negatively correlated with N_e proxies

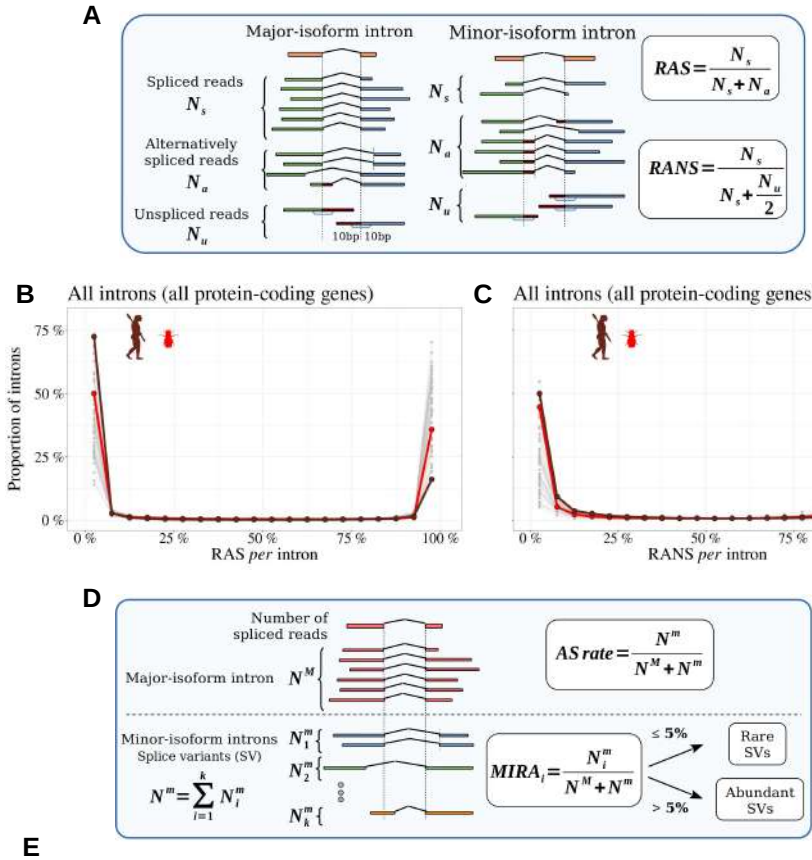
To quantify AS rates, we mapped RNA-seq data of each species on the corresponding reference genome assembly. We detected sequencing reads indicative of a splicing event (hereafter termed 'spliced reads'), and inferred the corresponding intron boundaries. We were thus able to validate the coordinates of annotated introns and to detect new introns, not present in the annotations. For each intron detected in RNA-seq data, we counted the number of spliced reads matching with its two boundaries (N_s) or sharing only one of

its boundaries (N_a), as well as the number of unspliced reads covering its boundaries (N_u) (Fig. 2A). We then computed the relative abundance of this spliced isoform compared to other transcripts with alternative splice boundaries (RAS = $\frac{N_s}{N_s + N_a}$) or compared to unspliced transcripts (RANS = $\frac{N_s}{N_s + \frac{N_u}{2}}$).

To limit measurement noise, we only considered introns for which both RAS and RANS could be computed based on at least 10 reads (Materials & Methods). In all species, both RAS and RANS metrics show clearly bimodal distributions (Fig. 2B,C): the first peak (mode $< 5\%$) corresponds to 'minor-isoform introns', whose splicing occurs only in a minority of transcripts of a given gene, whereas the second one (mode $> 95\%$) corresponds to the introns of major isoforms. It has been previously shown that in humans, for most genes, one single transcript largely dominates over other isoforms (Tress *et al.*, 2017a; Gonzàlez-Porta *et al.*, 2013). Our observations indicate that this pattern is generalized across metazoans. For the rest of our analyses, we computed the rate of alternative splicing with respect to introns of the major isoform. We will hereafter use the term 'splice variant' (SV) to refer to those splicing events that are detected in a minority of transcripts (*i.e.* with RAS ≤ 0.5 or RANS ≤ 0.5; see Fig. 2E for a definition of the main variables used in this study).

We focused our analyses on major-isoform introns interrupting protein-coding regions (*i.e.* we excluded introns located within UTRs, Materials & Methods). In vertebrates, each BUSCO gene contains on average 8.4 major-isoform introns (Appendix Tab. A.1). The intron density is more variable among insect clades, ranging from 2.8 major-isoform introns *per* BUSCO gene in Diptera to 6.1 in Blattodea. As expected, most major-isoform introns have GT/AG splice sites (99.1% on average across species), and only a small fraction have boundaries that do not match the canonical U2-introns splice sites (0.8% GC/AG and 0.1% AT/AC). The fraction of non-canonical splice sites is slightly higher among minor-isoform introns (2.8% GC/AG and 0.3% AT/AC). This might reflect a higher prevalence of U12-type introns but might also be caused by the presence of some false positives in the set of minor-isoform introns. In any case, the difference in splice signal usage between minor and major-isoform introns is small, which indicates that the vast majority of detected minor-isoform introns correspond to *bona fide* splicing events.

The proportion of major-isoform introns for which AS has been detected (*i.e.* with $N_a > 0$) ranges from 16.8% to 95.7% depending on the species (Appendix Tab. A.1). This metric is however not very meaningful because it directly reflects differences in sequencing depth across species (the higher the sequencing effort, the higher the probability to detect a rare SV, Appendix Fig. A.2). To allow a comparison across taxa, we computed the AS rate of introns, normalized by sequencing depth (AS = $\frac{N^m}{N^M + N^m}$, Materials & Methods; Fig. 2D). The average AS rate for BUSCO genes varies by a factor of 5 among species, from 0.8% in *Drosophila grimshawi* (Diptera) to 3.8% in *Megachile rotundata* (Hymenoptera) (3.4% in humans). Interestingly, the average AS rates of BUSCO gene introns are significantly correlated with the three proxies of N_e: species longevity (Fig. 3A), body length and the dN/dS ratio (Supplementary Fig. 3A,B). These correlations are positive, which implies that AS rates tend to increase when N_e decreases. It is noteworthy



Variable	Definition
RAS	Relative abundance of a given spliced isoform compared to other overlapping spliced isoforms
RANS	Relative abundance of a given spliced isoform compared to unspliced transcripts overlaping its splice sites
Major-isoform intron	Intron corresponding to a major isoform (*i.e* with RAS > 0.5 and RANS > 0.5)
Minor-isoform intron / SV	Splice variant (SV) = intron corresponding to a minor isoform (*i.e* with RAS ≤ 0.5 or RANS ≤ 0.5)
AS rate	Alternative splicing rate. For a given major-isoform intron, measures the rate at which SVs occur over its two boundaries
MIRA	Minor-isoform intron relative abundance: abundance of a given SV relative to the whole set of overlapping spliced transcripts
Abundant SV	SV with *MIRA* > 5 %
Rare SV	SV with *MIRA* ≤ 5 %
High-AS major-isoform intron	Major-isoform intron having at least one abundant SV
Low-AS major-isoform intron	Major-isoform intron that does not have any abundant SV

Figure 6.2: Distinguishing major and minor-isoform introns and measuring the rate of alternative splicing.
A*: Definition of the variables used to compute the relative abundance of a spliced isoform compared to other transcripts with alternative splice boundaries (RAS) or compared to unspliced transcripts (RANS):* N_s*: number of spliced reads corresponding to the precise excision of the focal intron;* N_a*: number of reads corresponding to alternative splice variants relative to this intron (i.e. sharing only one of the two intron boundaries);* N_u*: number of unspliced reads, co-linear with the genomic sequence.* ***B,C*** *Histograms representing the distribution of RAS and RANS values (divided into 5% bins), for protein-coding gene introns. Each line represents one species. Two representative species are colored: Drosophila melanogaster (red), Homo sapiens (brown).* ***D****: Description of the variables used to compute the AS rate of a given a major-isoform intron, and the 'minor-isoform intron relative abundance' (MIRA) of each of its splice variants (SVs):* N^M*: number of spliced reads corresponding to the excision of the major-isoform intron;* N_i^m*: number of spliced reads corresponding to the excision of a minor-isoform intron (i);* N^m*: total number of spliced reads corresponding to the excision of minor-isoform introns.* ***E****: Definitions of the main variables used in this study.*

that despite the fact that these proxies are not strongly correlated with each other (Fig. 1B,C), they all show similar relationships with AS rates. It should be stressed that these correlations were estimated using the PGLS method to account for phylogenetic inertia (and they remain significant when analyzing insects and vertebrates separately, Appendix Fig. A.4). Thus, these observations are consistent with the hypothesis that N_e has an impact on the evolution of AS rate.

One limitation of our analyses is that we used heterogeneous sources of transcriptomic data. To obtain enough sequencing depth, we combined for each species many RNA-seq samples, irrespective of their origin (whole body, or specific tissues or organs, in adults or embryos, etc.). It is known that genome-wide average AS rates vary according to tissues or developmental stages (Barbosa-Morais *et al.*, 2012; Mazin *et al.*, 2021), and according to environmental conditions (John *et al.*, 2021). To explore how this might have affected our results, we repeated our analyses using a recently published dataset that aimed to compare transcriptomes across seven organs, sampled at several developmental stages in seven species (six mammals, one bird) (Cardoso-Moreira *et al.*, 2019). In agreement with previous reports (Mazin *et al.*, 2021), our analysis of BUSCO genes revealed substantial differences in AS rates among organs, with consistent patterns of variation across species. For instance, in all species, testes and brain tissues show higher AS rates than liver and kidney (Fig. 3B). However, the variation in AS rate among organs in each species is limited compared to differences between species. Specifically, in an ANOVA analysis performed on the average AS rate across BUSCO gene introns, with the species and the organ of origin as explanatory variables, the species factor explained 89% of the total variance, while the organ factor explained only 9%. Among insects, we found only one species (*Dendroctonus ponderosae*) for which RNA-seq samples were available from multiple tissues. Here again, the variance in AS rate among tissues was limited compared to inter-species variability (Appendix Fig. A.5). Thus, despite the variability that can be introduced by the heterogeneity of RNA-seq samples, the relationship between AS rate and longevity remains detectable among these seven species (Fig. 3B).

6.2.4 Functional vs. non-functional alternative splicing

The negative correlation observed between N_e and alternative splicing rates is consistent with the hypothesis that differences in AS rates across species are driven by variation in the rate of splicing errors (drift barrier model). This does not exclude however that functional splicing variants might also contribute to AS rate variation across species. To evaluate this point, we selected a subset of SVs that are enriched in functional AS events. To do this, we reasoned that selective pressure against the waste of resources should maintain splicing errors at a low rate (as low as permitted by the drift barrier), whereas functional SVs are expected to represent a sizeable fraction of the transcripts expressed by a given gene, at least in some specific conditions (cell type, developmental stage...). Thus, functional SVs are expected to be enriched among abundant SVs compared to rare SVs.

To assess this prediction, we analyzed the proportion of SVs that preserve the reading

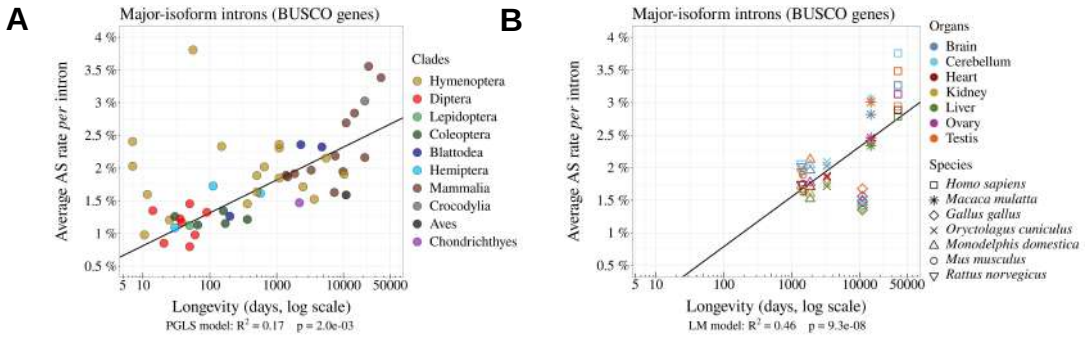


Figure 6.3: The rate of alternative splicing correlates with life history traits across metazoans.
A*: Relationship between the per intron average AS rate of an organism and its longevity (days, log scale).* ***B****: Variation in average AS rate across seven organs (brain, cerebellum, heart, liver, kidney, testis and ovary) among seven vertebrate species (RNA-seq data from Cardoso-Moreira et al. (2019)). AS rates are computed on major-isoform introns from BUSCO genes (Materials & Methods).*

frame according to their abundance relative to the major isoform. For this, we focused on minor-isoform introns that share a boundary with one major-isoform intron and that have their other boundary at less than 30 bp from the major splice site (either in the flanking exon or within the major-isoform intron). We determined whether the distance between the minor-isoform intron boundary and the major-isoform intron boundary was a multiple of 3. We computed the abundance of each minor isoform, relative to the corresponding major isoform, with the following formula:
Minor intron relative abundance $\mathrm{MIRA_i} = \frac{\mathrm{N_i^m}}{\mathrm{N^M} + \mathrm{N^m}}$ (see Fig. 2D).

We divided minor-isoform introns into 5% bins according to their MIRA and computed for each bin the proportion of minor-isoform introns that maintain the reading frame of the major isoform (Fig. 4A). In all species, we observe that this proportion varies according to the abundance of splice variants, with two distinct regimes (Fig. 4A). First, for MIRA values above 5%, the proportion of frame-preserving variants correlates positively with MIRA, reaching up to 60%-70% for the most abundant isoforms. Second, for MIRA values below 1%, the proportion of frame-preserving variants does not covary with MIRA, and fluctuates around 30 to 40%, close to the random expectation (33%). The excess of frame-preserving variants among the most abundant isoforms implies that a substantial fraction of them is under constraint to encode functional protein isoforms. This fraction varies from 0% for MIRA values below 1%, to 50% for isoforms with the highest MIRA values. It should be noted that these estimates correspond to a lower bound, since it is possible that some frame-shifting splice variants are functional. Nevertheless, these observations clearly indicate that the subset of SVs with MIRA values $>$ 5% (hereafter referred to as 'abundant SVs') is strongly enriched in functional isoforms relative to other SVs (MIRA $\leq$ 5%, hereafter referred to as 'rare SVs'). Of note, the subset of rare SVs represents the vast majority of the SV repertoire (from 62.4% to 96.9% depending on the species; Appendix Tab. A.1). Thus, the positive correlation between

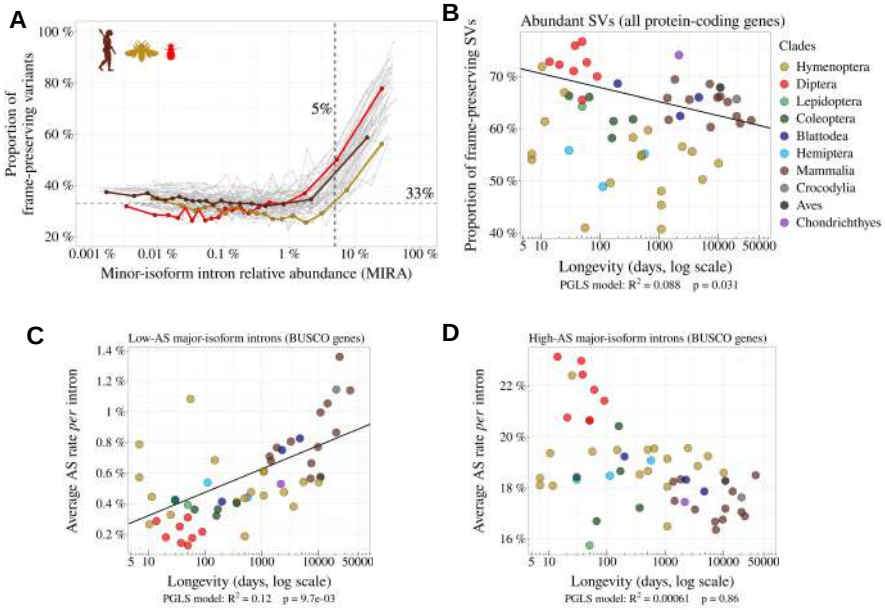


Figure 6.4: Variation in AS rate across metazoans: distinguishing abundant splice variants (enriched in functional variants) from rare splice variants. ***A****: Frame-preserving isoforms are strongly enriched among abundant splice variants (SVs). For each species, SVs were classified into 20 equal-size bins according to their abundance relative to the major isoform (MIRA, see Materials & Methods), and the proportion of frame-preserving SVs was computed for each bin. Each line represents one species. Three representative species are colored: red: Drosophila melanogaster, brown: Homo sapiens, yellow: Apis mellifera. We used a threshold MIRA value of 5% to define 'abundant' vs. 'rare' SVs.* ***B****: Proportion of frame-preserving SVs among abundant SVs across metazoans. Each dot represents one species. All annotated protein-coding genes are used in the analysis.* ***C,D****: Relationship between the average per intron AS rate of an organism and its longevity (days, log scale). Only BUSCO genes are used in the analysis.* ***C****: Low-AS major-isoform introns (i.e. major-isoform introns that do not have any abundant SV),* ***D****: High-AS major-isoform introns (i.e. major-isoform introns having at least one abundant SV).*

AS rate and longevity reported above (Fig. 3A) is mainly driven by the set of introns with a low AS rate (Fig. 4C). Interestingly, introns with high AS rate (enriched in functional SVs) show an opposite trend (Fig. 4D), and they display a lower proportion of frame-preserving SVs in vertebrates than in dipterans (Fig. 4B). This is the opposite of what would have been expected if functional SVs were more prevalent in complex organisms.

6.2.5 Investigating selective pressures on minor splice sites

A complementary approach to assess the functionality of AS events consists in investigating signatures of selective constraints on splice sites. For this, we used polymorphism data from *Drosophila melanogaster* and *Homo sapiens* to measure single-nucleotide polymorphism (SNP) density at major and minor splice sites, considering separately rare and abundant SVs. We focused on the first two and last two bases of each intron (consensus

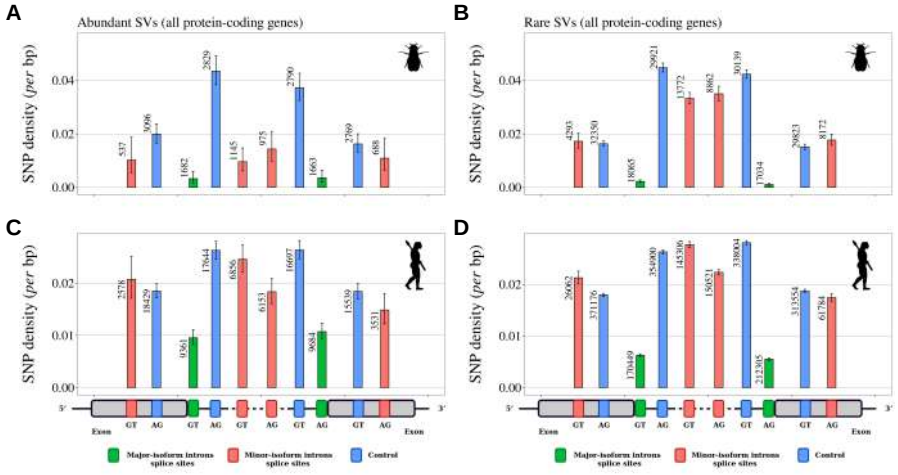


Figure 6.5: Variation in selective constraints on alternative splice signals from rare and abundant SVs. *For each minor-isoform intron sharing one boundary with a major-isoform intron, we measured the SNP density at its minor splice site (red), and at the corresponding major splice site (green). We distinguished minor splice sites that are located in an exon or in an intron of the major isoform. As a control (blue), we selected AG or GT dinucleotides that are unlikely to correspond to alternative splice sites, namely: AG dinucleotides located toward the end of the upstream exon or the beginning of the intron (unlikely to correspond to a genuine acceptor site), and GT dinucleotides located toward the beginning of the downstream exon or the end of the intron (unlikely to correspond to a donor site). To increase the sample size, we analyzed data from all annotated protein-coding genes (and not only the BUSCO gene set). The number of sites studied is shown at the top of each bar. Error bars represent the 95% confidence interval of the proportion of polymorphic sites (proportion test).* ***A,B****: SNP density in Drosophila melanogaster (polymorphism data from 205 inbred lines derived from natural populations, N=3,963,397 SNPs (Huang et al., 2014; Mackay et al., 2012)).* ***C,D****: SNP density in Homo sapiens (polymorphism data from 2,504 individuals, N=80,868,061 SNPs (Auton et al., 2015)). We excluded dinucleotides affected by CpG hypermutability (Materials & Methods, see Appendix Fig. A.6 for CpG sites).* ***A,C****: Abundant SVs (MIRA > 5%).* ***B,D****: Rare SVs (MIRA ≤ 5%).*

sequences GT, AG), which represent the most constrained sites within splice signals. We studied minor-isoform introns that share one splice site with a major-isoform intron and we measured SNP density at the corresponding major and minor splice sites. To account for constraints acting on coding regions, we considered separately minor splice sites that were located in an exon or in an intron of the major isoform. As negative controls, we selected AG or GT dinucleotides that were unlikely to correspond to alternative splice sites (Fig. 5, Materials & Methods). Furthermore, for *Homo sapiens* we controlled for the presence of hypermutable CpG dinucleotides (Tomso and Bell, 2003) (Appendix Fig. A.6, Materials & Methods).

For both species, the lowest SNP density is observed at major splice signals, which reflects the strong selective constraints on these sites (Fig. 5). In *Drosophila melanogaster*, there is also a strong signature of selection on minor splice signals of abundant SVs: both in introns and in exons, the SNP density at minor splice signals of abundant SVs

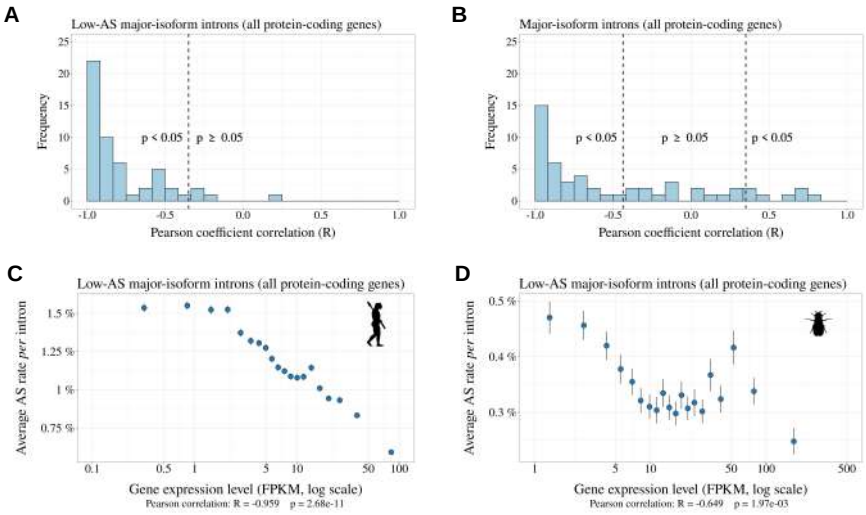


Figure 6.6: Relationship between AS rate and gene expression level. *For each species, we selected major-isoform introns with a sufficient sequencing depth to have a precise measure of their AS rate (*$N_s + N_a \geq 100$*). We divided major-isoform introns into 5% bins according to their gene expression level and computed the correlation between the average AS rate and median expression level across the 20 bins. To increase sample size, these analyses were based on all annotated protein-coding genes (and not only the BUSCO gene set).* ***A****: Distribution of Pearson correlation coefficients (R) between the AS rate and expression level observed in the 53 metazoans. The vertical dashed lines indicates the thresholds under and above which correlations are significant (i.e. p-value < 0.05).* ***B****: Distribution of Pearson correlation coefficients computed on the subsets of low-AS major-isoform introns (i.e. after excluding major-isoform introns with abundant SVs).* ***C,D****: Two representative species illustrating the negative relation between the average AS rate of low-AS major-isoform introns and the expression level of their gene. Error bars represent the standard error of the mean.* ***C****: N=127,599 low-AS major-isoform introns from Homo sapiens,* ***D****: N=31,357 low-AS major-isoform introns from Drosophila melanogaster.*

is much lower than in corresponding controls (from -37% to -74%, Fig. 5A) and than in minor splice signals of rare SVs (from -38% to -71%, Fig. 5B). This observation confirms that abundant SVs are strongly enriched in functional variants compared to rare SVs. In *Homo sapiens*, patterns of SNP density showed little evidence of selective constraints on minor splice sites, irrespective of the abundance of SVs (Fig. 5C,D): minor acceptor splice sites (AG) located within the major-isoform intron show a weak but significant SNP deficit relative to corresponding control sites (p-value < $1x10^{-5}$), but other categories of minor splice sites do not show any sign of selective constraints. The fact that the signature of selection on minor splice signals is much weaker in humans compared to *Drosophila* is indicative of a lower prevalence of functional variants, even among abundant SVs. This observation is therefore in total contradiction with the adaptive hypothesis (more functional alternative splicing in complex organisms).

6.2.6 The splicing rate of rare SVs is negatively correlated with gene expression levels

The above analyses are consistent with the hypothesis that the vast majority of rare SVs correspond to erroneous transcripts, and that changes in N_e contribute to variation in AS rate across taxa by shifting the selection-mutation-drift balance. If true, then this model predicts that the erroneous AS rate should also vary among genes, according to their expression level. Indeed, it has been shown that the selective pressure on splicing accuracy is stronger on highly expressed genes (Saudemont *et al.*, 2017). This reflects the fact that for a given splicing error rate, the waste of resources (both in terms of metabolic cost and of futile mobilization of cellular machineries) increases with gene expression level (Saudemont *et al.*, 2017; Xiong *et al.*, 2017). Thus, the selection-mutation-drift balance should lead to a negative correlation between gene expression level and the rate of splicing errors. To test this prediction, we focused on low-AS major-isoform introns, *i.e.* introns that are unlikely to have functional SVs. For each species, we considered all major-isoform introns with a sufficient sequencing depth to have a precise measure of their AS rate ($N_s + N_a \geq 100$). The selected subset represents 38.1% to 86.7% of major-isoform introns of each species (median=70.9%). Introns were then divided into 20 bins of equal size, according to the expression level of the corresponding genes. For each species, we computed the Pearson correlation between the average AS rate and the average expression level across bins. We observed a negative correlation between AS rates and gene expression levels in 52 out of the 53 species (significant with $p < 0.05$, in 48/53 species; Fig. 6A; two representative examples are shown in Fig. 6C and 6D). This pattern indicates that in almost all metazoan species, genes with a higher expression level have a lower AS rate, consistent with the hypothesis the rate of splicing errors is shaped by the selection-mutation-drift balance. It should be noted that this negative correlation between AS rate and gene expression level is not expected for functional SVs (there is *a priori* no reason why the AS rate of functional SVs should be higher in weakly expressed genes than in highly expressed genes). Interestingly, when we performed this analysis on all introns (including those with abundant SVs, which are enriched in functional variants), then most species (31/53) still showed a negative correlation between AS rate and gene expression level (Fig. 6B), but some species, such as *Drosophila melanogaster* showed the opposite pattern (Appendix Fig. A.7). This probably reflects that fact that, in those species, functional AS events make a significant contribution to the genome-wide average AS rate.

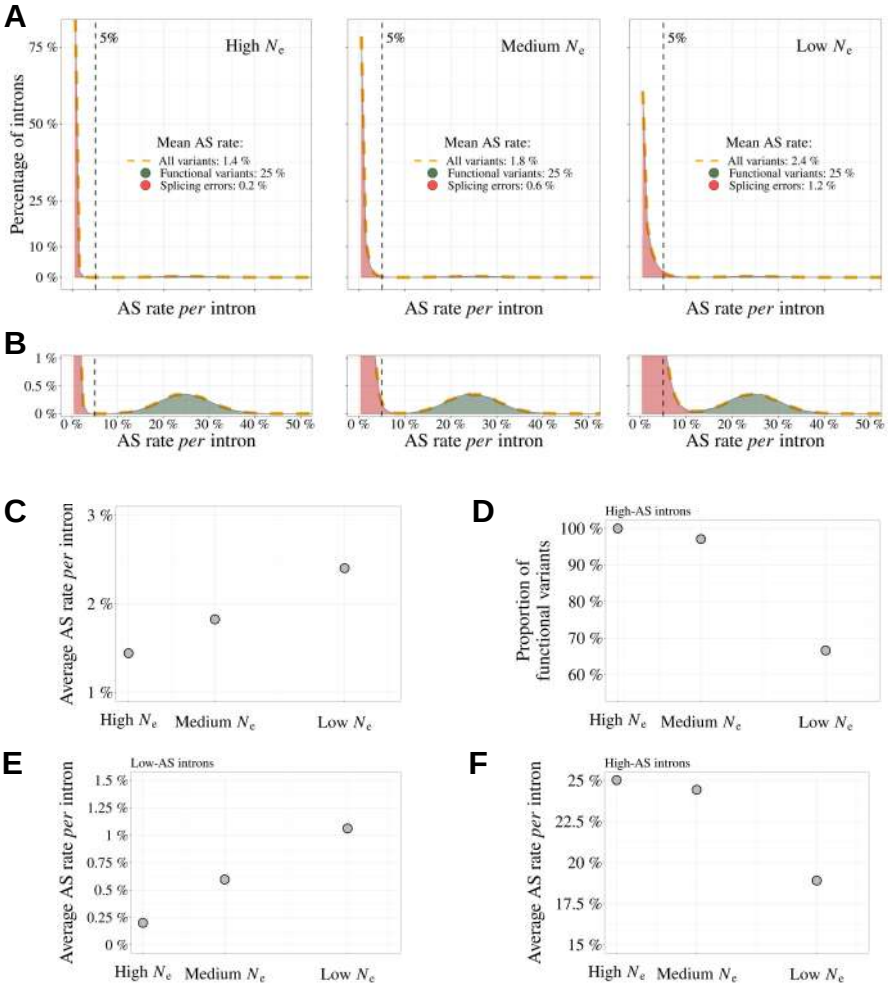


Figure 6.7: Impact of the drift-barrier on the genome-wide AS rate: model predictions. *To illustrate the impact of the drift barrier, we sketched a simple model, with three hypothetical species of different N_e. In this model, the repertoire of SVs consists of a mixture of functional variants and splicing errors. We assumed that in all species, only a small fraction of major-isoform introns (5%) produce functional SVs, but that these variants have a relatively high AS rate (average=25%, standard deviation=5%; see Materials & Methods for details on model settings). Splicing error rates were assumed to be gamma-distributed, with a low mean value. Owing to the drift barrier effect, the mean error rate was set to vary from 0.2% in species of high N_e to 1.2% in species of low N_e (these parameters were chosen to match approximately the AS rates observed in empirical data for rare SVs).* ***A*** *Genome-wide distribution of AS rates in each species (high N_e, medium N_e and low N_e). Each distribution corresponds to a mixture of functional SVs (green) and splicing errors (red).* ***B****: Zoom on the y-axis to better visualize the contribution of functional SVs to the whole distribution: rare SVs ($AS \leq 5\%$) essentially correspond to splicing errors, while abundant SVs ($AS > 5\%$) correspond to a mixture of functional and spurious variants, whose relative proportion depend on N_e. The following panels show how these different distributions, induced by differences in N_e, impact genome-wide AS patterns.* ***C****: Relationship between the average AS rate per major-isoform intron and N_e.* ***D****: Fraction of frame-preserving splice variants among introns with high AS rates vs N_e. Relationship between the average AS rate per intron and N_e, for 'low-AS' major-isoform introns (*MIRA $\leq 5\%$*) (**E**), and for 'high-AS' major-isoform introns (*MIRA $> 5\%$*) (**F**).*

6.3 Discussion

To investigate the factors that drive variation in AS rates across species, we analyzed publicly available RNA-seq data across a large set of 53 species, from diverse metazoan clades, covering a wide range of N_e values. To facilitate comparisons across species, we sought to limit the impact of the among-gene variance in AS rates. For this, we primarily based our analyses on a common set of nearly 1,000 orthologous protein-coding genes (BUSCO gene set). We focused our study on introns located within protein-coding regions, because introns from UTRs or lncRNAs are expected to be subject to different functional constraints. We measured AS rates on introns corresponding to a major isoform. When sequencing depth is limited, the set of introns for which AS can be quantified is biased toward the most highly expressed genes. To avoid this bias, we restricted our study to species for which the median sequencing depth of BUSCO exons was above 200. With this setting, on average 96.9% of BUSCO annotated introns could be analyzed in each species (Appendix Tab. A.1).

We observed a 5-fold variation in the average AS rate of BUSCO introns across species from 0.8% in *Drosophila grimshawi* (Diptera) to 3.8% in *Megachile rotundata* (Hymenoptera)(Fig. 3A). In agreement with previous work, we observed that AS rates tend to be high in vertebrates (average=2.3%), and notably in primates (average=3.1%) (Barbosa-Morais *et al.*, 2012; Chen *et al.*, 2014; Mazin *et al.*, 2021). This observation was previously interpreted as an evidence that AS played an important role in the diversification of the functional repertoire necessary for the development of more complex organisms (Chen *et al.*, 2014). However, this pattern is also compatible with the hypothesis that variation in AS rates across species result from differences in splicing error rates, which are expected to be higher in species with low N_e (Bush *et al.*, 2017). Indeed, consistent with this drift barrier hypothesis, we observed significant correlations between AS rates and proxies of N_e (Fig. 3B, Supplementary Fig. 3A,B).

In their original study, Chen *et al.* (2014) investigated the hypothesis that variation in AS rates across taxa might be driven by variation in N_e. For this, they focused on 12 species, for which they had measured levels of polymorphism at silent sites (π). They found that the correlation between AS rate and the number of cell types (proxy for organismal complexity) remained significant after controlling for π. They therefore concluded that the association between the cellular diversity and alternative splicing was not a by-product of reduced effective population sizes among more complex species. This conclusion was however based on a very small sample of species. More importantly, it assumed that π could be taken as a proxy for N_e. At mutation-drift equilibrium, π is expected to be proportional to $N_e\mu$ (where μ is the mutation rate *per* bp *per* generation). Thus, if μ is constant across taxa, π can be used to estimate variation in N_e. However, the dataset analyzed by Chen *et al.* (2014) included very diverse eukaryotic species, with mutation rates ranging from $1.7\text{x}10^{10}$ mutation *per* bp *per* generation in budding yeast, to $1.1\text{x}10^{8}$ mutation *per* bp *per* generation in humans (Lynch *et al.*, 2016). Hence, at this evolutionary scale, variation in N_e cannot be directly inferred from π without accounting

for variation in μ. Moreover, the drift barrier hypothesis states that the AS rate of a species should reflect the genome-wide burden of slightly deleterious substitutions, which is expected to depend on the intensity of drift over long evolutionary times (*i.e.* long-term N_e). Conversely, π reflects N_e over a short period of time (of the order of N_e generations), and can be strongly affected by recent population bottlenecks (too recent to have substantially impacted the genome-wide deleterious substitution load). The drift barrier hypothesis therefore predicts that the splicing error rate should correlate more strongly with proxies of long-term N_e (such as dN/dS, life history traits, or organismal complexity) than with π. The fact that AS rates remained significantly correlated to cellular diversity after controlling for π (Chen *et al.*, 2014) is therefore not a conclusive argument against the drift barrier hypothesis.

To contrast the two models (drift barrier vs diversification of the functional repertoire in complex organisms), we sought to distinguish functional splice isoforms from erroneous splicing events. Based on the assumption that splicing errors should occur at a low frequency, we split major-isoform introns into two categories, those with abundant SVs (MIRA $> 5\%$), and those without (MIRA $\leq 5\%$). Rare SVs represent the vast majority of the repertoire of splicing isoforms detected in a given transcriptome (from 62.4% to 96.9% according to the species; Appendix Tab. A.1). Two lines of evidence indicate that the small subset of abundant isoforms is strongly enriched in functional transcripts relative to other SVs. First, we observed that in all species, the proportion of SVs that preserve the reading frame is much higher among abundant SVs than among rare SVs (Fig. 4A). Second, the analysis of polymorphism data in *Drosophila* indicates that the average level of purifying selection on alternative splice sites is much stronger for abundant than rare SVs (Fig. 5A,B).

If variation in AS rate across species had been driven by a higher prevalence of functional SVs in more complex organisms, one would have expected the proportion of frame-preserving SVs to be stronger in vertebrates than in insects, in particular for the set of introns with high AS rate (*i.e.* enriched in functional SVs). On the contrary, the highest proportion of frame-preserving SVs is observed in dipterans (Fig. 4B). In fact, the overall higher AS rate of vertebrates (Fig. 3A) is driven by the set of introns with a low AS rate (Fig. 4C), *i.e.* the set of introns in which the prevalence of functional SVs is the lowest. On the contrary, among the set of introns with high AS rate, vertebrates have lower AS rates than insects (Fig. 4D).

These observations are difficult to reconcile with the hypothesis that the higher AS rate in vertebrates results from a higher rate of functional AS. Conversely, these observations fit very well with a model where variation in AS rate across species is entirely driven by variation in the efficacy of selection against splicing errors. To illustrate this model, let us consider three hypothetical species with different N_e, in which a small fraction of major-isoform introns (say 5%) is subject to functional alternative splicing. Let us consider that the distribution of AS rates of functional splicing variants is the same for all species (*i.e.* independent of N_e), with a mean of 25% (and a standard deviation of 5%). In addition, we assume that all major-isoform introns are potentially affected

by splicing errors, with a mean error rate ranging from 0.2% in species of high N_e to 1.2% in species of low N_e, owing to the drift barrier effect (these parameters were set to match approximately the AS rates observed in empirical data for rare SVs). The distributions of AS rate given by this model are presented in Fig. 7A: rare SVs (MIRA $\leq 5\%$) essentially correspond to splicing errors, while abundant SVs (MIRA $> 5\%$) correspond to a mixture of functional and spurious variants, whose relative proportion depend on N_e (Fig. 7B). This simple model makes predictions that match with our observations: we noted a positive correlation between AS rate and longevity (*i.e.* a negative correlation with N_e) for the set of low-AS major-isoform introns (Fig. 4C), but an opposite trend for high-AS major-isoform introns (Fig. 4D), as predicted by the model (Fig. 7D,E). Given that high-AS major-isoform introns represent only a small fraction of major-isoform introns, this model predicts that, overall, AS rates correlate negatively with N_e (Fig. 7), as observed in empirical data (Fig. 3A, Appendix Fig. A.3).

It should be noted that the BUSCO dataset corresponds to genes that are strongly conserved across species, often highly expressed, and hence might not be representative of the entire genome. Notably, AS rates are on average lower in the BUSCO gene set than in other genes, even after accounting for their expression level (Appendix Fig. A.7). However, results remained qualitatively unchanged when we repeated our analyses on the whole set of annotated protein-coding genes for each species: correlations between AS rates and N_e proxies are slightly weaker than on the BUSCO subset, but remain significant (Appendix Fig. A.8).

The model also predicts that the proportion of functional SVs among high-AS major-isoform introns should vary with N_e (Fig. 7C). To assess this point, we measured in each species the enrichment in reading frame-preserving events among abundant SVs compared to rare SVs. As predicted, this estimate of the prevalence of functional SVs tends to decrease with decreasing N_e proxies (*e.g.* Fig. 4B, where N_e is approximated by longevity). However, these correlations are weak, marginally significant after accounting for phylogenetic inertia with only two of the three N_e proxies, and not robust to multiple testing issues (Appendix Fig. A.9). Thus, N_e does not appear to be a strong predictor of the prevalence of functional SVs among high-AS major-isoform introns.

According to the drift-barrier model, the level of splicing errors is expected to decrease with increasing selective pressure. In all above analyses, we considered AS rates measured *per* intron, and not *per* gene. Yet, the trait under selection is the *per*-gene error rate, which depends not only on the error rate *per* intron, but also on the number of introns *per* gene. Given that intron density varies widely across clades (from 2.8 introns *per* gene in diptera to 8.4 introns *per* gene in vertebrates; Appendix Tab. A.1), the correlations reported above between AS rates and N_e may undervalue the predictive power of the drift-barrier model. The RNA-seq datasets that we analyzed consist of short-read sequences, which do not allow a direct quantification of the *per*-gene AS rate. We therefore indirectly estimated the *per*-gene AS rate in each species, based on the *per*-intron AS rate and on the number of introns *per* gene (Materials & Methods). Interestingly, as predicted by the drift-barrier model, N_e proxies correlate more strongly with this estimate of the *per*-gene

AS than with the *per*-intron AS rates (Appendix Fig. A.10).

One other important prediction of the drift barrier model is that splicing error rate should vary not only across species according to N_e, but also among genes, according to their expression level. Indeed, for a given splicing error rate, the waste of resources (and hence the fitness cost) is expected to increase with the level of transcription. Thus, the selective pressure for optimal splice signals is expected to be higher, and hence the error rate to be lower, in highly expressed genes. Consistent with that prediction, we observed a negative correlation between gene expression level and AS rate in low-AS major-isoform introns in all but one species (Fig. 6C).

It should be noted that our analyses suffer from several important limitations. First, the proxies that we considered for N_e are quite noisy (Fig. 1). Second, to maximize the number of species in our analyses, we had to use very heterogeneous sources of RNA (whole-body, specific tissues, or organs, at different life stages, in different sexes, different environmental conditions, etc.). Third, we used short-read sequencing data, which allow the quantification of AS rates for individual introns, but do not provide a direct measure of AS rates *per* gene. Hopefully progress of long-read sequencing technologies will soon allow the comparative analysis of AS rates on full-length transcripts (*e.g.* see Leung *et al.* (2021)). But presently, publicly available long-read transcriptomic data are restricted to a narrow set of model organisms, and their sequencing depth is still too limited to quantify rare splicing events. The fact that we detected significant correlations between AS rate and the three N_e proxies, despite these uncontrolled sources of variability, suggests that we underestimate the effect of N_e on AS rates.

Thus, overall, all observations fit qualitatively well with the predictions of the drift barrier model, according to which most of the variation in AS rate across species reflects differences in splicing error rates. Of course, this model is not in contradiction with the fact, well established, that some AS events play an essential role in various processes. Different criteria can be used to distinguish functional SVs from spurious splicing events. Notably, AS events that are strongly tissue-specific or developmentally dynamic tend to be more conserved across species, which indicates that a substantial fraction of them are evolutionary constrained, and hence functional (Mudge *et al.*, 2011; Barbosa-Morais *et al.*, 2012; Merkin *et al.*, 2012; Reyes *et al.*, 2013). The abundance of a SV is also an important predictor of its functionality. In particular, we observed that in all species, the proportion of frame-preserving events is much higher among abundant SVs than among rare SVs (Fig. 4A). We note however that the threshold that we used to define abundant SVs is somewhat arbitrary. In fact, according to our model, this class of SVs corresponds to a mixture of functional and spurious events, whose relative proportion is expected to depend on N_e (Fig. 7C). Thus, in low-N_e species, even the subset of abundant SVs includes a substantial fraction of errors. This probably explains why, contrarily to *Drosophila*, we do not detect any signature of purifying selection on alternative splice signals in humans, even for abundant SVs (Fig. 5).

In conclusion, all observations fit with the hypothesis that random genetic drift sets an upper limit on the capacity of selection to prevent splicing errors. It should be noted

that this limit on the optimization of genetic systems is expected to affect not only splicing, but all aspects of gene expression. Notably, there is a growing body of evidence that the complexity of transcripts produced by eukaryotic genes (resulting from alternative transcription initiation, polyadenylation, splicing or back-splicing, RNA editing) often does not correspond to fine-tuned adaptations but simply to the accumulation of errors (Pickrell *et al.*, 2010; Saudemont *et al.*, 2017; Xu *et al.*, 2019; Xu and Zhang, 2018; Liu and Zhang, 2018b,a; Xu and Zhang, 2014, 2020; Gout *et al.*, 2013; Zhang and Xu, 2022). It should be noted however that the relationship between the genome-wide error rate and N_e is not expected to be monotonic. Indeed, models predict that in species with very high N_e, selection on each individual gene should favor genotypes that are robust to errors of the gene expression machinery, which in turn, reduces the constraints on the global level of gene expression errors (Rajon and Masel, 2011; Xiong *et al.*, 2017). Thus, paradoxically, species with very large N_e are expected to have gene expression machineries that are more error-prone than species with very small N_e (Rajon and Masel, 2011). This argument was developed by Xiong *et al.* (2017) to account for the fact that transcription error rates had been found to be about 10 times higher in bacteria than in eukaryotes (Traverse and Ochman, 2016; Gout *et al.*, 2013). More recent work indicates that bacterial transcription error rates had been largely overestimated, presumably owing to RNA damages during the preparation of sequencing libraries (Li and Lynch, 2020). Given these uncertainties in the measures of transcription error rates, it seems for now difficult to interpret the differences reported across species. But in any case, it is important to note that it is in principle possible that the drift barrier affects differently the different steps of the gene expression process. It would therefore be important to investigate to which extent each step of gene expression responds (or not) to variation in N_e. As illustrated here by the relationship observed between alternative splicing and N_e, it appears essential to consider the contribution of non-adaptive evolutionary processes when trying to understand the origin of eukaryotic gene expression complexity.

6.4 Materials & Methods

6.4.1 Genomic and transcriptomic data collection

To analyze AS rate variation across metazoans, three types of information are required: transcriptome sequencing (RNA-seq) datasets, genome assemblies, and gene annotations. To obtain this data, we first queried the Short Read Archive database (Leinonen *et al.*, 2011b) to extract publicly available RNA-seq datasets. We also queried the NCBI Genomes database (NCBI Resource Coordinators, 2018) to retrieve genomic sequences and annotations. When this project was initiated, the vast majority of metazoans represented in this database corresponded to vertebrates or insects. We therefore decided to focus our analyses on these two clades (N=69 species).

6.4.2 Identification of orthologous gene families

To be able to compare average AS rates across species, given that AS rates vary among genes (Saudemont *et al.*, 2017), it is necessary to analyze a common set of orthologous genes. We searched for homologues of the BUSCOv3 (Benchmarking Universal Single Copy Orthologs, (Seppey *et al.*, 2019)) metazoan gene subset (N=978 genes) in each of the 69 genomes. To do this, we used the software BUSCO v.3.1.0 to associate BUSCO genes to annotated protein sequences. For each species, BUSCO genes were removed from the analysis if they were associated to more than one annotated gene or to an annotated gene that was associated to more than one BUSCO gene.

6.4.3 RNA-seq data processing and intron identification

We aligned the RNA-seq reads on the corresponding reference genomes with HISAT2 v.2.1.0 (Kim *et al.*, 2019). We built the genome indexes using annotated introns and exons coordinates in addition to genome sequences, to improve splice junction detection sensitivity. The maximum allowed intron length was fixed to 2,000,000 bp. We then extracted intron coordinates from HISAT2 alignments using an in-house perl script that scanned for CIGAR strings containing N, which indicate regions that are skipped from the reference sequence. For intron detection and quantification we used only uniquely mapping reads that had a maximum mismatch ratio of 0.02. We required a minimum anchor length (that is, the number of bases that align on each flanking exon) of 8 bp for intron detection, and of 5 bp for intron quantification. We kept only those predicted introns that had GT-AG, GC-AG or AT-AC splice signals, and we predicted the strand of the introns based on the splice signal.

We assigned an intron to a gene if at least one of the intron boundaries fell within 1 bp of the annotated exon coordinates of the gene, combined across all annotated isoforms. We excluded introns that could not be unambiguously assigned to a single gene. We distinguish annotated introns (which appear as such in the reference genome annotations) and un-annotated introns, which were detected with RNA-seq data and assigned to previously annotated genes.

We further restricted our analyses to introns located within protein-coding regions. To do this, for each protein-coding gene, we extracted the start codons and the stop codons for all annotated isoforms. We then identified the minimum start codon and the maximum end codon positions and we excluded introns that were upstream or downstream of these extreme coordinates.

The alignment process, which is the most time-consuming step in the pipeline (see Appendix Fig. A.11), can take up to one week when using 16 cores *per* RNA-seq for larger genomes, such as mammals. Additionally, the processed compressed files generated during this process can exceed 7 terabytes in size.

6.4.4 Alternative splicing rate definition

For each intron we noted N_s the number of reads corresponding to the precise excision of this intron (spliced reads), and N_a the number of alternatively spliced reads (*i.e.* spliced variant sharing only one of the two intron boundaries). Finally, we note N_u the number of unspliced reads, co-linear with the genomic sequence, and which overlap with at least 10 bp on each side of an exon-intron boundary. These definitions are illustrated in Fig. 2. We then defined the relative abundance of the focal intron compared to introns with one alternative splice boundary ($RAS = \frac{N_s}{N_s + N_a}$), as well as relative to unspliced reads ($RANS = \frac{N_s}{N_s + \frac{N_u}{2}}$).

To compute these ratios we required a minimal number of 10 reads at the denominator. We thus calculated the RAS only if $(N_s + N_a) \geq 10$ and the RANS only if $(N_s + \frac{N_u}{2}) \geq 10$ (We divided N_u by 2 because retention is quantified at two sites, which increases the detection power by a factor of 2). If the criteria were not met, the values were labeled as not available (NA). We computed these ratios using reads from all available RNA-seq samples, unless otherwise specified (for example, in sub-sampling analyses). Based on these ratios we defined three categories of introns: major-isoform introns, defined as those introns that have RANS > 0.5 and RAS > 0.5; minor-isoform introns, defined as those introns that have RANS ≤ 0.5 or RAS ≤ 0.5; unclassified introns, which do not satisfy the above conditions.

We determined the alternative splicing (AS) rate of major-isoform introns using the following formula: $AS = \frac{N^m}{N^M + N^m}$, where N^M is the number of spliced reads corresponding to the excision of the major-isoform intron and N^m is the total number of spliced reads corresponding to the excision of minor-isoform introns sharing a boundary with a major-isoform intron (see Fig. 2)

For minor-isoform introns sharing a boundary with a major-isoform intron, we computed the relative abundance of the minor-isoform intron (i) with respect to the corresponding major-isoform intron, with the following formula:
Minor intron relative abundance $MIRA_i = \frac{N_i^m}{N^M + N^m}$, where N_i^m is the number of spliced reads corresponding to the excision of a minor-isoform intron (i) (see Fig. 2).

We defined the *per*-gene AS rate as the probability to observe at least one alternative splicing event across all the major-isoform introns of a gene. To estimate the per-gene AS rate of a given gene, we assumed that the AS rate is uniform across its major-isoform introns, and that AS events occur independently at each intron. We calculated the AS rate for each gene as the number of spliced reads corresponding to the excision of major-isoform introns, divided by the number of spliced reads corresponding to minor and major-isoform introns ($\frac{\sum N^m}{\sum N^M + N^m}$). The probability for a given gene to produce no splice variant across all its major-isoform introns is thus $p0=(1 - \frac{\sum N^m}{\sum N^M + N^m})^{N_i}$, where N_i is the number of major-isoform introns of the gene. The *per*-gene AS rate (ASg), i.e. the probability to have at least one AS event, is therefore the complement of p0: ASg=1-p0.

6.4.5 Identification of reading frame-preserving splice variants

To determine the proportion of open reading frame-preserving splice variants, we first identified minor-isoform introns that had their minor splice site within a maximum distance of 30 bp from the major splice site (either in the flanking exon or within the major-isoform intron). We chose this length threshold because it is shorter than the size of the smallest introns in metazoans, so that to avoid the possibility of having a skipped exon between the minor and the major splice site (which could induce some ambiguities in the assessment of the reading frame). Among these introns, we considered that frame-preserving variants are those introns for which the distance between the minor-isoform intron boundary and the major-isoform intron boundary was a multiple of 3.

6.4.6 Gene expression level

Gene expression levels were calculated with Cufflinks v2.2.1 (Roberts *et al.*, 2011) based on the read alignments obtained with HISAT2, for each RNA-seq sample individually. We estimated FPKM levels (Fragments *Per* Kilobase of exon *per* Million mapped reads) for each gene.

The overall gene expression of a gene was computed as the average FPKM across samples, weighted by the sequencing depth of each sample. The sequencing depth of a sample is the median *per*-base read coverage across BUSCO genes.

6.4.7 Phylogenetic tree reconstruction

For each of the 978 BUSCO gene families we collected the longest corresponding proteins identified in each species. We removed proteins for which the amino acid sequence provided with the annotations did not perfectly correspond to the translation of the corresponding coding sequences. We then aligned the resulting sets of protein-coding sequences for each BUSCO gene, using the codon alignment option in PRANK v.170427 (Löytynoja and Goldman, 2008). We translated the codon alignments into protein alignments using the R package seqinr (Charif and Lobry, 2007). To infer the phylogenetic tree rapidly, we sub-sampled the resulting multiple alignments (N=461), selecting alignments with the highest number of species (ranging from 49 to 53 species *per* alignment). We then concatenated these alignments and kept sites that were aligned in at least 30 species. We used RAxML-NG v.0.9.0 (Kozlov *et al.*, 2019) to infer the species phylogeny with a final alignment of 53 taxa and 165,648 sites (amino acids). RAxML was set to perform one model *per* gene with fixed empirical substitution matrix (LG), empirical amino acid frequencies from alignment (F) and 8 discrete GAMMA categories (G8), specified in a partition file with one line *per* multiple alignment. The analysis generated 10 starting trees, 5 starting from a random topology and 5 starting from a tree generated by the parsimony-based randomized stepwise addition algorithm. The best-scoring topology was kept as the final ML tree and 10 bootstrap replicates have been generated.

6.4.8 dN/dS computation

We estimated dN/dS ratios for the BUSCO gene families that were present in at least 45 species (N=922 genes), using the codon alignments obtained with PRANK (see above). We divided the 922 sequence alignments into 18 groups, based on their average GC3 content across species, and concatenated the alignments within each group. We thus obtained concatenated alignments that were 209 kb long on average. We used bio++ v.3.0.0 libraries (Guéguen *et al.*, 2013; Dutheil and Boussau, 2008; Bolívar *et al.*, 2019) to estimate the dN/dS on terminal branches of the phylogenetic tree, for each concatenated alignment. We attributed the dN/dS of the terminal branches to the species that corresponds.

In a first step, we used an homogeneous codon model implemented in bppml to infer the most likely branch lengths, codon frequencies at the root, and substitution model parameters. We used YN98 (F3X4) (Yang and Nielsen, 1998) substitution model, which allows for different nucleotide content dynamics across codon positions. In a second step, we used the MapNH substitution mapping method (Guéguen and Duret, 2018) to count synonymous and non-synonymous substitutions (Dutheil *et al.*, 2012). We defined dN as the total number of non-synonymous substitutions divided by the total number of non-synonymous opportunities, both summed across concatenated alignments, for each branch of the phylogenetic tree. Likewise, we defined dS as the total number of synonymous substitutions divided by the total number of synonymous opportunities, both summed across concatenated alignments. The *per*-species dN/dS corresponds to the ratio between dN and dS, on the terminal branches of the phylogenetic tree.

6.4.9 Life history traits

We used various life history traits to approximate the effective population size of each species. For vertebrates species we considered the maximum lifespan (*i.e.* from birth to death) and body length referenced. For insects we took the maximum lifespan and body length of the *imago*. For eusocial insects and the eusocial mammal *Heterocephalus glaber*, the selected values correspond to the queens. The sources from which the lifespan and the body length information was taken are listed in `data/Data9-supp.pdf` in the Zenodo repository (see).

6.4.10 Analyses of sequence polymorphism

We analyzed the distribution of single nucleotide polymorphisms (SNPs) around splice sites in *Drosophila melanogaster* and *Homo sapiens*.

For *Drosophila melanogaster* we used polymorphism data from the *Drosophila* Genetic Reference Panel (DGRP) (Huang *et al.*, 2014; Mackay *et al.*, 2012), from which we extracted 3,963,397 SNPs that were identified from comparisons across 205 inbred lines. We converted the SNP coordinates from the dm3 genome assembly to the dm6 assembly with the liftOver utility (Hinrichs *et al.*, 2006) of the UCSC genome browser, using a whole genome alignment between the two assemblies downloaded from `https://hgdownload.`

soe.ucsc.edu/goldenPath/dm3/liftOver/dm3ToDm6.over.chain.gz.

For *Homo sapiens* we used polymorphism data from the 1000 Genomes project, phase 3 release (Auton *et al.*, 2015). This dataset included 80,868,061 SNPs that were genotyped in 2,504 individuals.

For each minor-isoform intron sharing one boundary with a major-isoform intron, we computed the number of SNPs that occur at their respective splice sites: at their shared boundary, and at the major-isoform intron and minor-isoform introns specific boundaries.

We focused our study on minor-isoform introns that have their specific boundary folding in the exons adjacent to the major-isoform intron or in the major-isoform intron. As a control, for each minor-isoform intron, we searched for one GT and one AG dinucleotides in the interval between 20 and 60 bp with respect to the major splice site, in the neighboring exon and in the major-isoform intron, and computed the number of SNPs that occur on these sites. We searched for control AG dinucleotides in the vicinity of the donor splice site of the major-isoform intron and for GT dinucleotides in the vicinity of its acceptor splice site, to avoid studying sites that might correspond to unidentified minor splice sites. For *Homo sapiens*, we further divided the splice sites and the control dinucleotides into two groups, depending on whether they were subject to CpG hypermutability or not.

6.4.11 Impact of the drift-barrier on genome-wide AS rates: sketched model

To illustrate the impact of the drift barrier, we sketched a simple model, with three hypothetical species of different N_e (low, medium and high N_e). In each species, the repertoire of SVs consists of two categories: functional variants and spurious variants (which result from errors of the splicing machinery). The rate of splicing error was assumed to be low and to depend on N_e, owing to the drift barrier effect. We considered that in all species, only a small fraction of major-isoform introns (5%) produce functional SVs, but that these variants have a relatively high AS rate. The AS rates of functional SVs were modeled by a normal distribution, with a mean of 25% and a standard deviation of 5% (same parameters for the three species). We modeled the distribution of error rates by a gamma distribution, with shape parameter = 1, and with mean values of 0.2%, 0.6% and 1.2% respectively in species of high, medium or low N_e (these parameters were set to match approximately the AS rates observed in empirical data for rare SVs). We then combined the two distributions (functional SVs and splicing errors) to compute the genome-wide average AS rates in each species. We also computed the average AS rate on the subsets of low-AS or high-AS major-isoform introns (*i.e.* with AS rates respectively below or above the threshold AS rate of 5%). Finally, we computed the proportion of frame-preserving SVs among high-AS major-isoform introns, assuming that two thirds of splicing errors induce frameshifts and that all functional SVs preserve the reading frame.

Acknowledgements

We thank Loïc Guille for his contribution to an initial pilot study, Tristan Lefébure for insightful discussions and Laurent Guéguen for his help on dN/dS analyses. Computational analyses were performed using the computing facilities of the CC LBBE/PRABI and the Core Cluster of the Institut Français de Bioinformatique (IFB) (ANR-11-INBS-0013). We thank five anonymous reviewers for their thorough and constructive comments, which were very helpful to improve our manuscript. A preprint version of this article has been peer-reviewed and recommended by PCIEvolBiol (https://doi.org/10.24072/pci.evolbiol.100642).

Funding

This work was funded by the French National Research Agency (ANR-20-CE02-0008-01 "NeGA" and ANR-17-CE12-0019-01 "LncEvoSys").

Conflict of interest disclosure

The authors declare the following non-financial conflict of interest: Laurent Duret is recommender for PCI Evol Biol.

Data and code availability

All processed data that we generated and used in this study, as well as the scripts that we used to analyze the data and to generate the figures, are available on Zenodo DOI: https://doi.org/10.5281/zenodo.7415114.

In particular, the sources of transcriptomic data, genome assemblies and annotations are reported in the Zenodo archive in `data/Data1-supp.tab`. The archive includes several directories, including `figure`, which contains the necessary materials to produce the figures of the manuscript. Rmarkdown scripts located in the `table_supp` directory were used to generate supplementary tables, which are also saved in the same directory. The processed data used to generate figures and conduct analyses are stored in the `data` directory in tab-separated text format.

7

Why is selection for translationally optimal codons so scarce in metazoans? Variation in fitness effects and drift intensity

The third objective of my thesis is to study the variations in synonymous codon usage across metazoans. Synonymous codons do not modify the decoded amino acids. However, studies have shown that the different uses of these synonymous codons are not neutral and have an effect on the phenotype (*i.e.* gene expression, translation *etc.*).

Interestingly, codon usage varies widely among metazoans and within genomes. There are two identified forces responsible for these variations. The first regroup non-adaptive processes such as gBGC or mutational bias, which affect both coding regions (*i.e.* exons) and non-coding regions (*i.e.* introns). Notably, in human, it has been observed that codon usage correlates with the GC content, and also that the GC content of bacterial genomes (from 13% to 75%) is strongly correlated with their CU. These results suggest that non-adaptive processes are at play in determining CU.

The second process that drives codon usage is an adaptive processes, *i.e.* translational selection, which favors the use of codons optimizing the speed and accuracy of translation, thereby affecting coding regions. In particular, this selection tends to promote in highly expressed genes the use of codons that match the tRNA pool, as seen in model species *C. elegans*, *D. melanogaster* and *E. coli*.

We aim to quantify translational selection across 257 metazoans, for which gene expression data are available in GTDrift. Our findings show that translational selection is rare in metazoans and its population-scaled selection coefficient (S) is low. In this range of S values, the "drift barrier" suggests that reducing N_e leads to less efficient selection. Indeed, we observed low TS for low-N_e species. However, large-N_e species show a strong disparity in TS intensity. These variations are not simply explained by variations in mutational biases that could hamper TS. Thus, our results could suggest that the selective advantage in optimizing the translation machinery varies across species.

Why is selection for translationally optimal codons so scarce in metazoans? Variation in fitness effects and drift intensity

Florian Bénitière[1,2], Tristan Lefébure[2], Laurent Duret[1]

[1]Laboratoire de Biométrie et Biologie Évolutive, UMR CNRS 5558 LBBE, Université Claude Bernard Lyon 1, Villeurbanne, France
[2]Laboratoire d'Ecologie des Hydrosystèmes Naturels et Anthropisés, UMR CNRS 5023 LEHNA, Université Claude Bernard Lyon 1, Villeurbanne, France

Contents

7.1 Introduction

Since the early days of DNA sequencing, it has been noticed that the usage of synonymous codons is not random: some synonymous codons are more frequently used than others, and the patterns of synonymous codon usage (SCU) can vary both across species and among genes within a genome (Grantham *et al.*, 1980a). Two types of processes, adaptive or non-adaptive, can contribute to genome-wide patterns of SCU (Sharp *et al.*, 1993). First, neutral substitution patterns (NSPs) vary across taxa and, in some species, can also vary along chromosomes. NSPs are primarily driven by the underlying pattern of mutation, which accounts for 60% of the variance in genome base composition across the tree of life (Long *et al.*, 2018). In addition, in some taxa, NSPs are also strongly affected by GC-biased gene conversion (gBGC), a process associated to homologous recombination that favors the transmission of G:C alleles over A:T alleles (Duret and Galtier, 2009). NSPs affect all genomic compartments (coding or non-coding), and notably have a strong impact on SCU (*e.g.* Pouyet *et al.* (2017); Long *et al.* (2018)). Besides NSPs, SCU can also be affected by selection. Indeed, it has been observed that in some species, SCU varies according to gene expression level (Gouy and Gautier, 1982; Sharp *et al.*, 1986; Duret and Mouchiroud, 1999) and that the synonymous codons that are more frequently used in highly expressed genes correspond to the most abundant tRNAs (Ikemura, 1985; Dong *et al.*, 1996; Moriyama and Powell, 1997; Kanaya *et al.*, 1999; Duret, 2000). This indicates that synonymous codon usage and tRNA content have coevolved in a way that optimizes translation. This co-evolution implies two levels of selection (Bulmer, 1987): 1) selection on the pool of tRNAs to match the relative abundance of different codons in the transcriptome (*i.e.* the codon demand), and 2) selection on the synonymous codon usage of genes to match the pool of tRNAs (classically referred to as "translational selection").

It is generally considered that there are two main benefits of using translationally optimal codons. First, this leads to increase the speed of translation, and hence to reduce the time spent by ribosomes on each mRNA, thereby increasing the pool of free ribosomes available in the cell, which ultimately allows a higher cellular growth rate (Bulmer, 1991). Second, the usage of synonymous codons decoded by the most abundant tRNAs increases the accuracy of translation, and thus reduces the amount of mis-translated proteins that cause an important burden on the cell (Akashi, 1994; Drummond *et al.*, 2006). It is important to note that for both aspects (speed and accuracy of translation), the benefit of using optimal codons is expected to be proportional to gene expression level. Indeed, the higher the expression level of a given gene, the stronger the impact of its translation speed on the pool of free ribosomes, and for a given mis-translation rate, the cost of erroneous protein production (in terms of waste of resources and of direct toxic effect of misfolded proteins) increases directly with expression level. In bacteria, the intensity of translational selection is correlated to the minimal cell division time, which suggests that the selective force for the optimization of SCU is the maximization of cellular growth (Rocha, 2004; Sharp *et al.*, 2005).

It should be noted that besides translational selection, synonymous sites can be sub-

ject to additional levels of selective constraints. For instance, the presence of splice enhancers located within exons skews codon usage near exon-intron boundaries in mammalian genes (Parmley and Hurst, 2007). But this type of selective pressure is site-specific (*i.e.* a particular codon is preferred at a specific site in a given gene), and hence, is not expected to affect the genome-wide pattern of SCU. Similarly, there is evidence that the use of translationally sub-optimal codons can be advantageous at some specific sites to slow-down translation and favor the proper folding of proteins (Buhr *et al.*, 2016; Walsh *et al.*, 2020). But again, this is a local effect, with limited genome-wide impact on SCU.

Interestingly, the intensity of translational selection varies widely across species, not only in unicellular organisms, but also in multicellular eukaryotes (Sharp *et al.*, 2005; Subramanian, 2008; dos Reis and Wernisch, 2009; Galtier *et al.*, 2018). For instance, among animals, early studies on the two main invertebrate model organisms (*Drosophila melanogaster* and the nematode *Caenorhabditis elegans*) showed clear signatures of translational selection (Shields *et al.*, 1988; Duret and Mouchiroud, 1999). Conversely, there is no sign of translational selection in humans (Sémon *et al.*, 2006; Pouyet *et al.*, 2017), despite clear evidence that SCU does affect gene expression in mammals (Kudla *et al.*, 2006; Courel *et al.*, 2019; Wu *et al.*, 2019; Mordstein *et al.*, 2020; Medina-Muñoz *et al.*, 2021). To understand variation in the intensity in translational selection across animals, it is important to refer to basic population genetics principles (Ohta, 1996). Indeed, the SCU in a given genome reflects a balance between selection favoring translationally optimal codons, and the effects of mutation and drift, allowing the fixation of non-optimal codons (Bulmer, 1991). Thus, the frequency of optimal codons is expected to depend on the population-scaled selection coefficient ($S = 4N_e s$), where N_e is the effective population size and s the selection coefficient in favor of translationally optimal codons (Bulmer, 1991; Sharp *et al.*, 2005). Hence, the lack of translational selection in some animal taxa might stem from a small N_e (hereafter referred to as the drift-barrier hypothesis), or from a smaller fitness effect of using translationally optimal codons (*i.e.* lower s).

To explore these hypotheses, several previous studies analyzed variation in the intensity of translational selection across eukaryotes (Subramanian, 2008; dos Reis and Wernisch, 2009; Galtier *et al.*, 2018). These three studies, reported positive correlations between signatures of translational selection and proxies of N_e (Subramanian, 2008; dos Reis and Wernisch, 2009; Galtier *et al.*, 2018). Although this pattern fits qualitatively with the predictions of the drift barrier model, quantitatively, the fit is not so clear. Indeed, dos Reis and Wernisch (2009) estimated S in 10 eukaryotic species, and they reported only a 2-fold difference in S between humans and *D. melanogaster* (respectively $S = 0.5$ and $S = 1.0$), despite a ≈30-fold difference in N_e between the two species (20,000 vs. 600,000; Lynch *et al.* (2023)). According to the authors, this poor fit to the drift barrier model might be due to the fact that their analysis was sensitive to variation in NSP across genes, which might have led to overestimate S in humans (dos Reis and Wernisch, 2009). But, it has also been argued that besides differences in N_e, s is also likely to vary across species, as long-lived organisms, with relatively a slow development, are likely to be less constrained to optimize cell growth than species with a very rapid

development (Subramanian, 2008).

These three studies were based on relatively limited sample sizes (10 to 30 species), and in the end, the causes of the variation in the intensity of translational selection across species remained unclear. To try to go further, we decided here to investigate variation in translational selection intensity across a large dataset of 223 metazoan species, covering a wide range of animal clades. For each species, we predicted the set of optimal codons based on the pool of tRNA genes present in its genome, and we analyzed how the frequency of optimal codons varies with gene expression, controlling for variation in NSP. Based on these variations, we quantified S in each species, and analyzed how it correlates with estimates of N_e or life history traits. Our analyses revealed that overall, few metazoans show clear signs of translational selection. As expected, the highest values of S are observed in species with large N_e, while species with small N_e show little evidence of translational selection. However, overall, N_e appears to be a poor predictor of the intensity of translational selection, which suggests important variation in s across taxa. We discuss several factors that may drive this variation in the fitness effect of optimizing codon usage.

7.2 Results

7.2.1 Non-adaptive processes are the primary drivers of codon usage variations among metazoans

To investigate the factors driving the intensity of translational selection in metazoans, we used the GTDrift database, that compiles genomic and transcriptomic data along with life history traits and proxies of N_e for various eukaryotic species (Bénitière *et al.*, 2024). We initially selected 257 metazoan species available in GTDrift, but we excluded 11 species for which there were not enough transcriptomic data (less than 5,000 genes detected as being expressed). We analyzed patterns of SCU and genomic base composition in the 246 remaining species, covering a wide range of clades (129 vertebrates, 82 insects and 35 other metazoan species; Fig. 1A).

Patterns of SCU can be affected both by translational selection and by NSPs (Sharp *et al.*, 1993). It is possible to distinguish the contribution of NSPs because they affect the base composition of both coding and non-coding regions, whereas translational selection operates only on codons. Thus, if differences in SCU across species are driven by NSPs, then it is expected that they should correlated with variation in the base composition of non-coding regions. And similarly, if intra-genomic variation of SCU in a given species is driven by the heterogeneity of NSPs along its chromosomes, then this should result in a covariation between the codon usage of genes and the base composition of their introns. Owing to the symmetry of the DNA molecule, NSPs generally affect similarly both strands, resulting in an equal proportion of cytosine (C) and guanine (G), as well as an equal proportion of thymine (T) and adenine (A) (Lobry, 1995). Hence, the G+C content provides a good summary statistics of the impact of NSPs on the genomic base

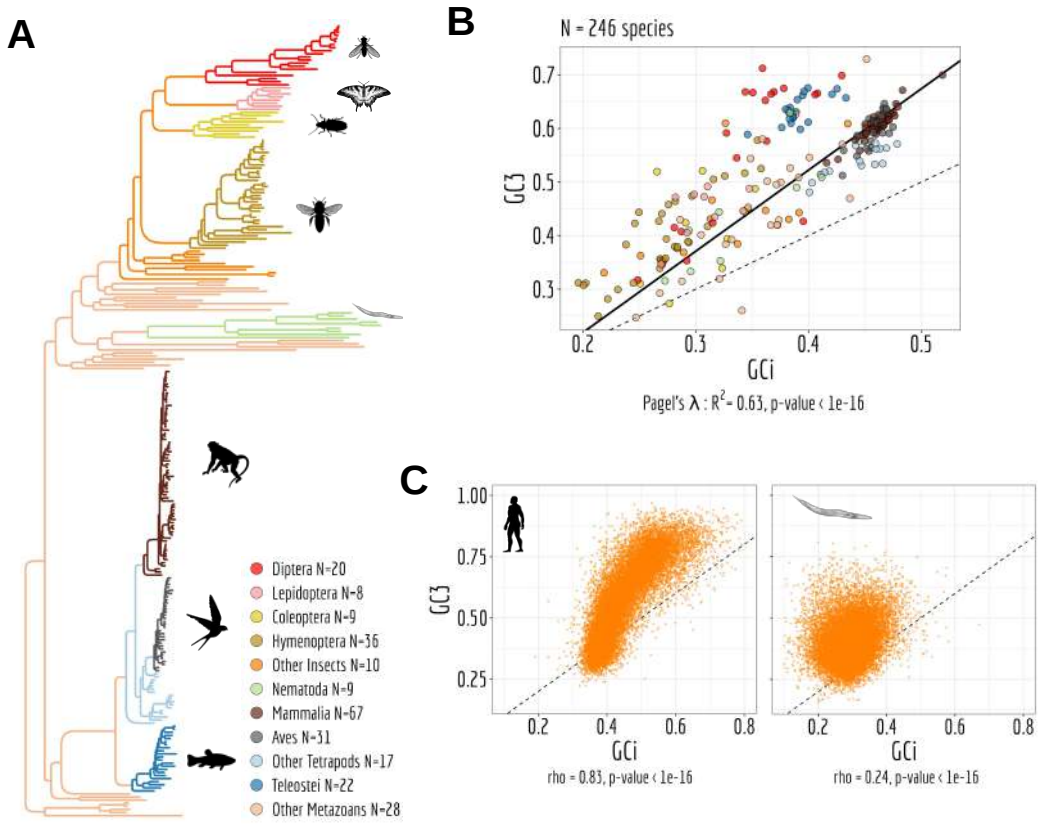


Figure 7.1: Codon usage variations are driven by non-adaptive processes. ***A****: Phylogenetic tree of the 257 studied species.* ***B****: Gene average GC content at the third position of codons (GC3) and the gene average GC in introns (GCi) for each species. Pagel's lambda model is used to take into account the phylogenetic structure of the data in a regression model (black line).* ***C****: Correlation between the gene GC3 and the GCi in Homo sapiens (left) and in Caenorhabditis elegans (right). Spearman's rho and corresponding p-values are displayed under the graph. The dotted lines correspond to x=y.*

composition. Thus, to examine the potential contribution of non-adaptive processes to the observed variations in SCU across the 246 species, we measured their G+C content in introns (GCi) and at the third position of codons (GC3), averaged over all genes. We observed a strong correlation between the average GC3 and the average GCi (Fig. 1B). Our findings suggest that non-adaptive processes, are the primary factor driving the observed variation in codon usage across species. As already noted by Vinogradov (2003) and Amit *et al.* (2012), the relationship between GC3 and GCi is asymmetrical. While introns are predominantly AT-rich (GCi range=0.2 to 0.52), the third position of codons displays a wider range of variation, with GC3 spanning from 0.25 to 0.73. While most species of a clade displayed similar average GC3, dipterans (N=20) exhibit the widest range of GC3 variations (from 0.32 to 0.71).

NSPs can vary within the genome of a given species, and impact codon usage accordingly. In *Homo sapiens*, the *per* gene GC3 and GCi are highly correlated (Spearman's correlation coefficient, rho=0.83, $p<10^{-16}$), whereas this correlation is less pronounced in *Caenorhabditis elegans* (rho=0.24, $p<10^{-16}$; Fig. 1C). Species showing the

strongest intra-genomic variance in codon usage (as assessed by GC3) are the ones with the strongest variance in GCi Supplementary Fig. 1A). These correlations between GC3 and GCi are particularly strong in tetrapods (107/108 species with rho>0.7) and in hymenopterans (25/35 species with rho>0.7) (Supplementary Fig. 1B). The other clades generally show less variance in GCi, and weaker correlations between GC3 and GCi. But overall, 244/246 species (99%) showed a significant positive correlation ($p<0.05$), which indicates that in most species, intra-genomic variation in NSPs somehow contribute to the variance in SCU among genes. Hence, it is important to take this source of variance into account to be able to detect signatures of translational selection within genomes.

7.2.2 tRNA abundance matches proteome requirements

To quantify the intensity of translational selection, we used an approach similar to that of dos Reis and Wernisch (2009) and Sharp *et al.* (2005). This approach is based on the comparison of the frequency of optimal codons between highly and weakly expressed genes, and therefore requires the prior identification of optimal codons. For this, dos Reis and Wernisch focused on the nine amino-acids that are encoded by two codons (duet codons), and predicted the optimal codon of each amino-acid as being the one that is more frequently used in highly expressed genes. One caveat is that if the NSP varies among genes according to their expression level, this may lead to erroneous prediction of codon optimality. Furthermore, this approach does not capture the signal of translational selection from the nine other amino-acids that are encoded by triplet, quartet or sextet codons. To avoid these limitations, we sought here to predict optimal codons based on the tRNA pool. Owing to technical difficulties, there are currently few species for which tRNA abundance has been quantified directly. Behrens *et al.* (2021) recently developed a technique (mim-tRNAseq) that allowed them to measure tRNA abundance in four eukaryotes (Behrens *et al.*, 2021). This study revealed a robust correlation between tRNA abundance and their respective gene copy number, with an adjusted $R^2 > 0.91$ for yeasts (*S. cerevisiae* and *S. pombe*), 0.79 for *Drosophila melanogaster* and 0.62 for *Homo sapiens* (Behrens *et al.*, 2021). These results suggest that tRNA copy numbers are a good predictor of tRNA abundances. To investigate whether the number of tRNA genes could be used as an indirect measure of tRNA abundance across metazoans, we analyzed the co-variation of their tRNA gene repertoires with the amino acid composition of their proteome.

The total number of tRNA gene copies varies widely among clades and species (ranging from an average of 201 tRNA gene copies per genome in hymenopterans to 1,537 copies in teleost fish; Supplementary Fig. 2). However, the relative copy number of distinct isoacceptor tRNA genes is quite conserved among metazoans. There are some rare cases where the gene copy number of a given tRNA has exploded in a given species compared to other genomes (Supplementary Fig. 2). This might reflect the propensity of tRNA genes to become transposable elements. Indeed many SINE retrotransposon families derive from tRNA genes (Sun *et al.*, 2007), and it is therefore possible that some recently evolved SINEs are erroneously annotated as *bona fide* tRNA genes.

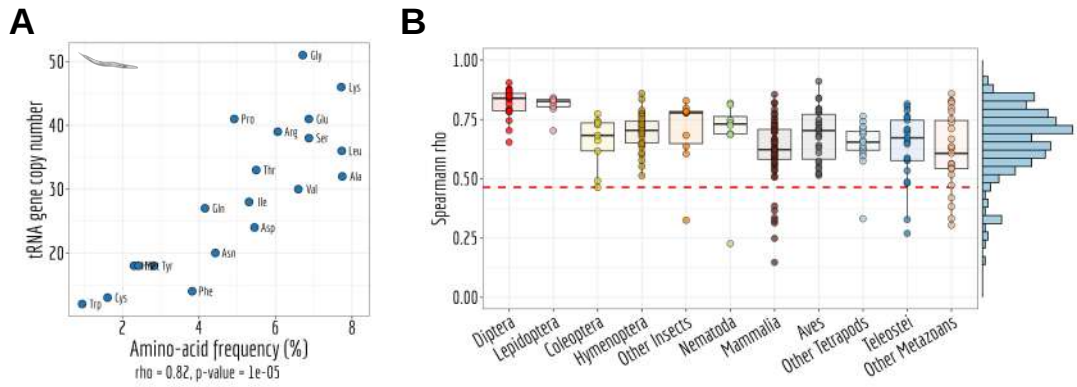


Figure 7.2: The tRNA gene copies number is a good predictor of the transcriptional requirements. ***A****: The relationship between the number of tRNA gene copies per amino acid and the frequency of amino acid weighted by gene expression (FPKM, log scale) in Caenorhabditis elegans. Spearman's rho and corresponding p-value are displayed under the graph.* ***B****: Boxplot illustrating the distribution of Spearman's correlation coefficient (ρ) from Panel A for each species (N=246 species). The red line indicates the threshold above which the p-value is lower than 0.05.*

In both *Drosophila melanogaster* and *Homo sapiens*, we observed a strong correlation between amino acid usage (*i.e.* the frequency of amino acids, weighted by the expression level of genes) and direct measures of tRNA abundance (rho=0.79; Supplementary Fig. 3; Behrens *et al.* (2021)). These results indicate that tRNA abundance matches the amino acid demand. As expected, the amino acid usage of these two species also strongly correlates with their tRNA gene copy numbers (rho=0.78 and 0.68 respectively; Supplementary Fig. 3). As previously reported (Duret, 2000), tRNA gene copy number also correlates with the amino acid demand in *Caenorhabditis elegans* (rho=0.82; Fig. 2A). The same analysis conducted across 246 animal species found a significantly positive Spearmann coefficient (*i.e.* p-value < 0.05) in 93% of the species (Fig. 2B), which indicates that in most of metazoans, the tRNA gene copy number is under constraints to match the amino acid demand. This implies that tRNA abundance is primarily regulated by modulating the copy number of tRNA genes rather than their transcription level. We suspect that the few cases where the number of tRNA genes does not correlate with amino acid usage might be due to annotation errors : some tRNA genes may have been missed (*e.g.* because of gaps in the genome assembly), or conversely, some SINEs or tRNA pseudogenes may have been incorrectly annotated as functional tRNA genes. To ensure that the tRNA gene copy number is a good proxy of the tRNA abundance, we kept in our study only the species for which tRNA gene copy number correlates significantly with amino acid usage (N=230 species). We also excluded 7 species for which the repertoire of annotated tRNA appeared to be incomplete (*i.e.* the cognate tRNAs of certain codons were not found in the genome assembly).

7.2.3 Definition of putative-optimal codons based on tRNA abundance and wobble-pairing rules

To predict which synonymous codons are optimal for translation, it is first necessary to associate each of the 61 codons to their cognate tRNA. The number of distinct isodecoder tRNAs (*i.e.* distinct anticodons) ranges from 43 to 60 per species (average=47). This implies that 1 to 18 codons cannot be translated through Watson-Crick pairing (WCp), and hence have to be translated via wobble pairing (WBp). We used the rules established by Percudani (Percudani, 2001) to assign each of these codons to their cognate tRNA, allowing for non-standard base pairing with the first nucleotide of the anticodon (Fig. 3A). For example, deamination of adenine in inosine (I) in anticodons ANN makes them permissive to wobble pairing I:C; I:U or I:A. Another common wobble pairing is the G:U/U:G pairing (Percudani, 2001). As an illustration, in human, asparagine is translated by a single tRNA (anticodon GTT) that decodes both AAC (by WCp) and AAT (by G:U WBp). AAT accounts for 48% of asparagine codons, highlighting the significance of wobble pairing. There are 18 amino acids that are encoded by multiple synonymous codons. These amino acids can be classified in two groups: - those whose synonymous codons are translated by at least two distinct isodecoder tRNAs - those for which all synonymous codons are translated by a single isodecoder tRNA

There is some variation in the set of amino acids present in each group, depending on the isodecoder tRNA repertoire of each species. The first group generally corresponds to amino acids encoded by sextet codons (Leu, Arg, Ser), quartets (Val, Gly, Ala, Pro, Thr), triplet (Ile) and NNG/NNA duets (Glu, Gln, Lys). The second group corresponds essentially to the six amino acids encoded by NNC/NNT duets (Phe, Cys, Tyr, Asp, His, Asn) (Fig. 3B).

For each amino acid of the first set, synonymous codons were predicted to be optimal if they were decoded by the isodecoder tRNA with highest gene copy number (*i.e.* predicted to be the most abundant). In case of ex æquo (*i.e.* if all synonymous codons are decoded by isodecoders having the same gene copy number), then the optimal codons of this amino acid were considered as unknown (76 ex æquo cases in total, in 62 species). In the cases where the most abundant tRNA decodes more than one synonymous codon, we considered all of them as potentially optimal (*i.e.* at this stage, we do not make any assumption regarding which of the Watson-Crick pairing or wobble pairing is the most efficient). This first set of putative-optimal codons will hereafter be referred as 'POC1'. The second set of amino acids corresponds to cases where the two synonymous codons (NNC/NNT) are decoded by a single isodecoder (anticodon GNN). There is evidence, based on studies in various eukaryotes, that the wobble pairing GNN:NNU is less efficient than the Watson-Crick pairing GNN:NNC (Stadler and Fire, 2011; Chan *et al.*, 2017; Wang *et al.*, 2017). Consequently, for these amino acids, we defined codons NNC (decoded through WCp) as being the putative-optimal codons 'POC2'.

For the human genome, POC1 have been defined for 13 amino acids and POC2 for 5 amino acids. In contrast, for *Caenorhabditis elegans*, POC1 and POC2 are defined for 12

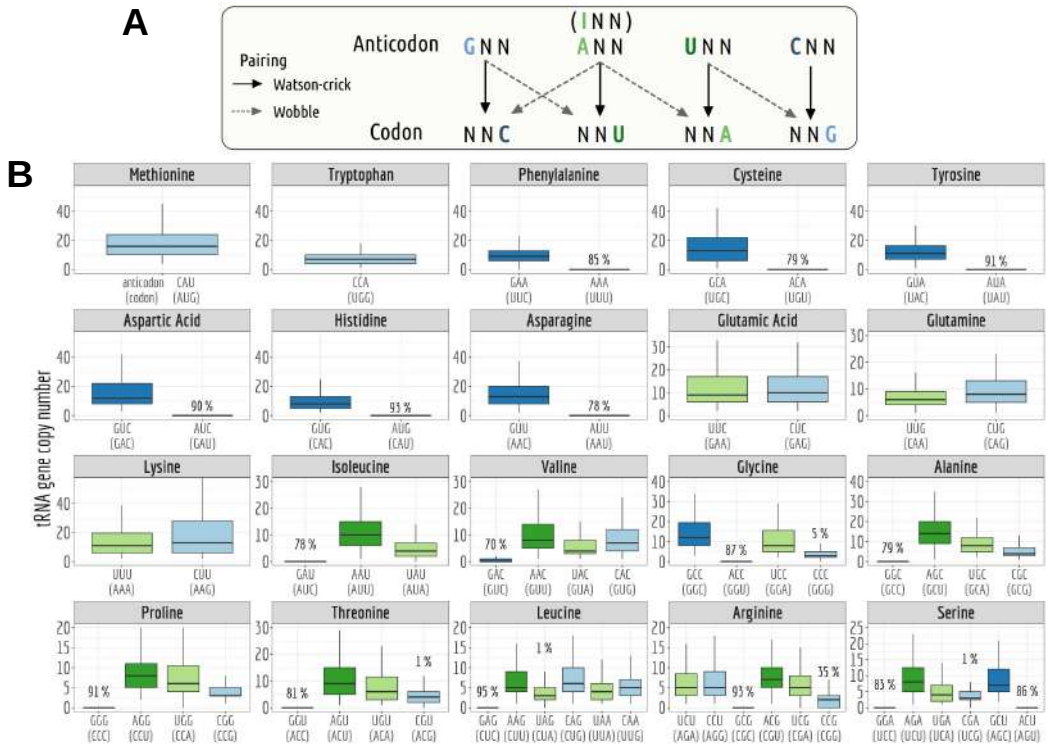


Figure 7.3: Presence-Absence of tRNA defines set of putative-optimal codons. ***A****: Illustration of the various possible pairings: Watson-Crick and wobble pairing.* ***B****: A boxplot illustrating the distribution of tRNA gene copy numbers across 223 species. The percentage of species lacking a tRNA gene copy is also indicated, highlighting the absence of tRNA isodecoder.*

and 6 amino acids, respectively. On average among the 223 species, POC1 are defined for 12.5 amino acids *per* species (ranging from 8 to 17) and POC2 for 5.2 amino acids (ranging from 0 to 6, except *Tyto alba* with 7 POC2, including Ile).

7.2.4 Highly expressed genes are enriched in optimal codons

The intensity of translational selection depends directly on gene expression levels. Given the very wide range of variation of gene expression levels (> 1000 folds), the fitness impact of synonymous codon usage is expected to vary strongly among genes. Hence, a typical feature of genomes subject to translational selection (TS), is that the frequency of optimal codons is particularly high in the most highly expressed genes. Thus, to identify which species are subject to TS, we examined the variations in POC frequency according to gene expression level. To control for possible variations in neutral substitutions pattern, we also analyzed triplet content in introns, referred to as POC-control. It is important to note that the frequency of POC-control is not expected to be equivalent to that of the overall POC frequency due to the differing AT-richness of introns.

For *Homo sapiens* POC frequencies show some slight fluctuations according to gene expression level (Fig. 4A). However, the same weak fluctuations are observed for POC-controls in introns, which implies that a same process, independent of translation effi-

ciency, affects the base composition, both in introns and at synonymous codon positions (Fig. 4A). Indeed, there exists a strong correlation between the GC content of introns and the GC3 content in *Homo sapiens*, along with pronounced variations in GC3 content (Fig. 1C). In contrast, in *Caenorhabditis elegans*, we observed a strong rise in POC frequencies in highly expressed genes, both for POC1 (from 47% to 76%) and for POC2 (from 38% to 70%). These changes in codon usage are not caused by shift in local substitution patterns as we see no similar variation in POC-control (Fig. 4B).

It is important to notice that the non-linear relationship observed between gene expression level and POC frequency is perfectly consistent with the TS model, that assumes that the selection coefficient on synonymous codon usage (s) should increase linearly with gene expression level. Indeed, this model predicts that for lowly expressed genes (such that $S = 4N_e s \ll 1$), the frequency of optimal codons should evolve neutrally, and hence should be independent of expression level. But above the 'nearly-neutral' point (*i.e.* the expression level for which $S \approx 1$), the frequency of optimal codon should strongly increase with expression level (see Appendices for more details on equations). The shape of the POC1 and POC2 curves in Fig. 4B indicates that in *C. elegans*, this 'nearly-neutral' point is reached for a gene expression level of about 50 FPKM. This implies that genes with a lower expression level (which represent 83% of genes in *C. elegans*) are not affected by TS. Of note, the fraction of genes affected by TS is expected to be even more reduced in species with a lower effective population size.

To assess the impact of TS on synonymous codon usage, we measured the difference between the frequency of POCs in the most expressed genes (top 2%), and the frequency of POCs in the 50% lowest expressed genes, controlling for POCs-control variations (see Materials & Methods). This shift in codon usage (denoted ΔPOC^{exp}) was computed for both POC1 and POC2 codons in each of the studied species (N=223 species). ΔPOC_1^{exp} and ΔPOC_2^{exp} are strongly correlated ($R^2 = 46\%$, p-value $< 10^{-16}$), which indicates that the signature of translational selection is effectively captured by both sets of codons (Fig. 4C). For 211 species (95%), the prevalence of POC1 is greater in highly expressed genes compared to other genes (Fig. 4D). Similarly, ΔPOC_2^{exp} is positive for 191 species (86%; Fig. 4D). The highest values of ΔPOC^{exp} were observed in *C. elegans* (+30% for both POC1 and POC2). There are substantial variation across clades: average ΔPOC_1^{exp} and ΔPOC_2^{exp} values are around +14% in Diptera compared to +3% in vertebrates. We obtained very similar results when measuring ΔPOC^{exp} without accounting for POC-control variations (Supplementary Fig. 5). Given that the two sets of codons gave very consistent results, we hereafter considered the whole set of POCs regrouping both POC1 and POC2 .

7.2.5 Highly constrained amino acids are enriched in optimals codons

Synonymous codon usage is expected to be under selection not only for its impact on the speed of translation, but also on the accuracy of translation. For both traits, selective

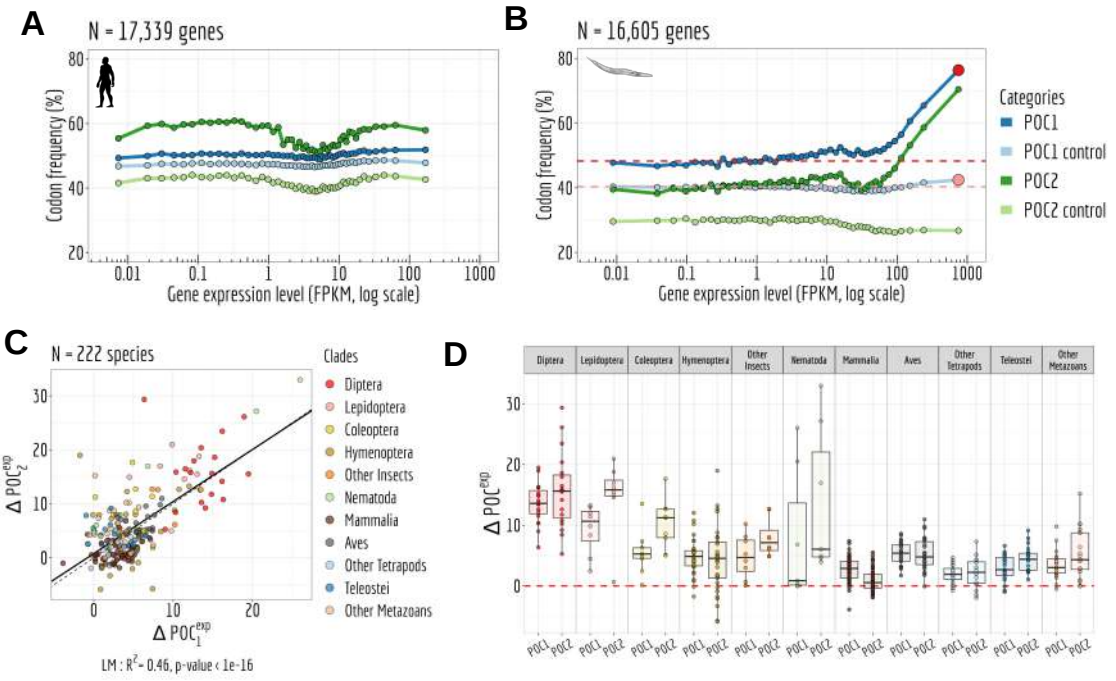


Figure 7.4: Differences in usage of putative-optimal codon between highly- and lowly-expressed genes. ***A,B****: Variation in the proportion of POC within coding sequences (POC1: dark blue; POC2: dark green) according to gene expression level. To control for variations in neutral substitution patterns, we analyzed the frequency of corresponding triplets within introns (POC1 control: light blue; POC2 control: light green). Each point represents a 2% bin of genes, with the red point at the end of each POC1 curve denoting the 2% most highly expressed genes. The red lines indicate the average POC1 proportions observed in the 50% least expressed genes (FPKM, log scale).* ***A*** *represents Homo sapiens , and* ***B*** *represents Caenorabditis elegans.* ***C****: Relation between the variations in POC1 and POC2 frequency with expression. Calculated as the difference between POC frequency in the 2% most highly expressed genes and the 50% least expressed genes. To this variations we removed variations measured on control.* ***D****: Boxplot illustrating the differences, for each species, between the TS intensity measured on POC1 and POC2.*

constraints are expected to vary among genes, according to their expression level. One specific feature of selection for translation accuracy is that the strength of selection is also expected to vary among sites within a protein: selection on translation accuracy should be stronger at sites that are essential for the structure or function of the protein. To test this prediction, we analyzed within-gene variation in POC usage according to the level of constraint on amino acid sites. For this, we focused on a set of 976 orthologous genes, present in single copy in most metazoan genomes (BUSCO genes; Waterhouse *et al.* (2018)). For each protein of a given species, we classified its sites into four groups of equal size, according to their level of conservation across 293 metazoans (see Materials & Methods), and then measured the shift in POC frequency between its 25% most conserved sites and its 25% least conserved sites. Finally, we computed the average of these shift values over all proteins of this species (noted ΔPOC^{cons}). Given that the shift is computed within each gene, ΔPOC^{cons} measures variation in codon usage across sites that inherently have the same expression level. One difficulty however is that in

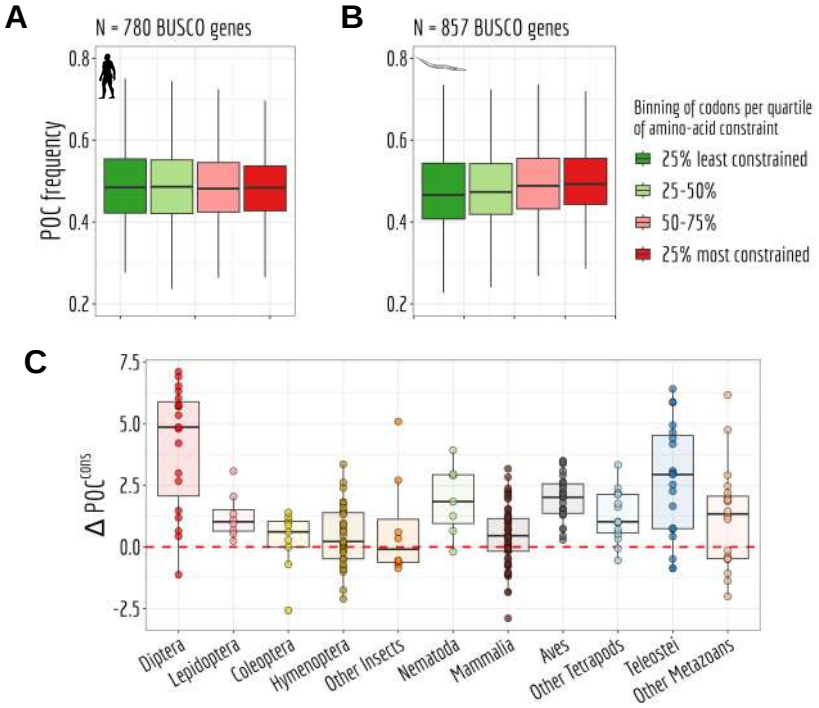


Figure 7.5: Most highly conserved regions exhibit a preference for using POCs. ***A,B****: Investigation of POCs frequency by dividing genes into four constraint groups of equal size, based on the gap proportion of gene alignments across all species. The frequency of POCs was calculated for each gene within each constraint group. A boxplot is shown, with darker green indicating the least constrained sites and darker red indicating the most constrained sites.* ***A*** *represents Homo sapiens, and* ***B*** *represents Caenorabditis elegans.* ***C****: Distributions depicting the average differences per species between the frequency of POCs in highly constrained sites and unconstrained sites of given genes.*

tetrapods, many genes contain a GC-rich CpG island at their 5' end (Deaton and Bird, 2011). The presence of a CpG island affects the base composition of the beginning of genes, up to about 1 kb, as illustrated by the analysis of the intronic GC content (Supplementary Fig. 7). This results in differences in codon usage between the first exon and the rest of the coding region. Given that the N-termini of proteins evolve faster than their center (Bricout *et al.*, 2023), this causes a spurious association between codon usage and variation in amino acid constraints along proteins. To avoid this bias, we measured ΔPOC^{cons} in tetrapods only on codons located beyond 1 kb of the start codon (in genomic coordinates). In other clades, the base composition of introns shows little variation along genes (Supplementary Fig. 7), and hence ΔPOC^{cons} was measured on the entire coding region. In *C. elegans*, the frequency of POC increased significantly between the least constrained and most constrained sites within proteins (from 48.5% on average to average 51.2%), whereas no variation was observed in humans (Fig. 5A,B). Overall, ΔPOC^{cons} is positive in 75% of species (refer to Fig. 5C). As for ΔPOC^{exp} , ΔPOC^{cons} shows substantial variation across clades, and is maximal for Diptera.

7.2.6 Selection favors optimal codons in highly expressed genes of *Drosophila melanogaster*

To further assess whether POCs are under selection in Diptera, we investigated patterns of polymorphism and substitution in *Drosophila*, based on a multiple genome alignment of three closely related species (*D. melanogaster, D. simulans, D. erecta*) and on single nucleotide polymorphism (SNP) data from 205 *D. melanogaster* individuals. We inferred the ancestral and derived state at each substitution or SNP, so that to distinguish synonymous changes corresponding to POC to non-POC mutations (PO>nPO) vs non-POC to POC mutations (nPO>PO) (see Materials & Methods; Fig. 6A). To control for possible variation in local mutation patterns, we conducted a parallel analysis on triplets in intronic regions. In coding sequences (CDS), we identified 44,288 nPO>PO synonymous SNPs and 139,256 PO>nPO synonymous SNPs, 26,770 nPO>PO synonymous substitutions and 81,666 PO>nPO synonymous substitutions. In introns, we observed 187,321 nPO>PO SNPs and 260,366 PO>nPO SNPs, 56,916 nPO>PO substitutions and 77,686 PO>nPO substitutions.

We observed that the rate of nPO>PO changes (number of nPO>PO changes/number of non-POC codons) increases with increasing gene expression level, while the rate of PO>nPO changes (number of PO>nPO changes/number of POC codons) decreased, both for SNPs and for substitutions(Fig. 6B,D). Importantly, this trends is specific to coding regions, and is not observed for the corresponding triplets in introns (Fig. 6C,E). These observations are consistent with the hypothesis that selection favors mutations leading to the incorporation of translationnally optimal codons in genes with high expression level.

7.2.7 Weak relationship between the strength of translational selection and the effective population size

According to standard population genetic models of translational selection (Bulmer (1991); Sharp *et al.* (2005); dos Reis and Wernisch (2009); Appendices), the difference in codon usage between highly and weakly expressed genes is expected to be directly linked to the population-scaled selection coefficient in favor of optimal synonymous codons ($S = 4N_e s$). Indeed, considering that synonymous codon usage evolves neutrally in lowly expressed genes, then S in highly expressed genes can be expressed as:

$$S^{hx} = ln(\frac{FOP^{hx}}{1 - FOP^{hx}}) - ln(\frac{FOP^{lx}}{1 - FOP^{lx}}) \tag{7.1}$$

where FOP^{hx} and FOP^{lx} are the observed frequencies of optimal codons in highly and lowly expressed genes respectively (Sharp *et al.*, 2005; dos Reis and Wernisch, 2009). It should be noted however that this equation holds true only if underlying mutation patterns (and possibly gBGC) do not vary with gene expression level (Sharp *et al.*, 2005; dos Reis and Wernisch, 2009). We used the above equation to estimate S^{hx} in each species, based on the observed POC frequencies in the top 2% most highly expressed

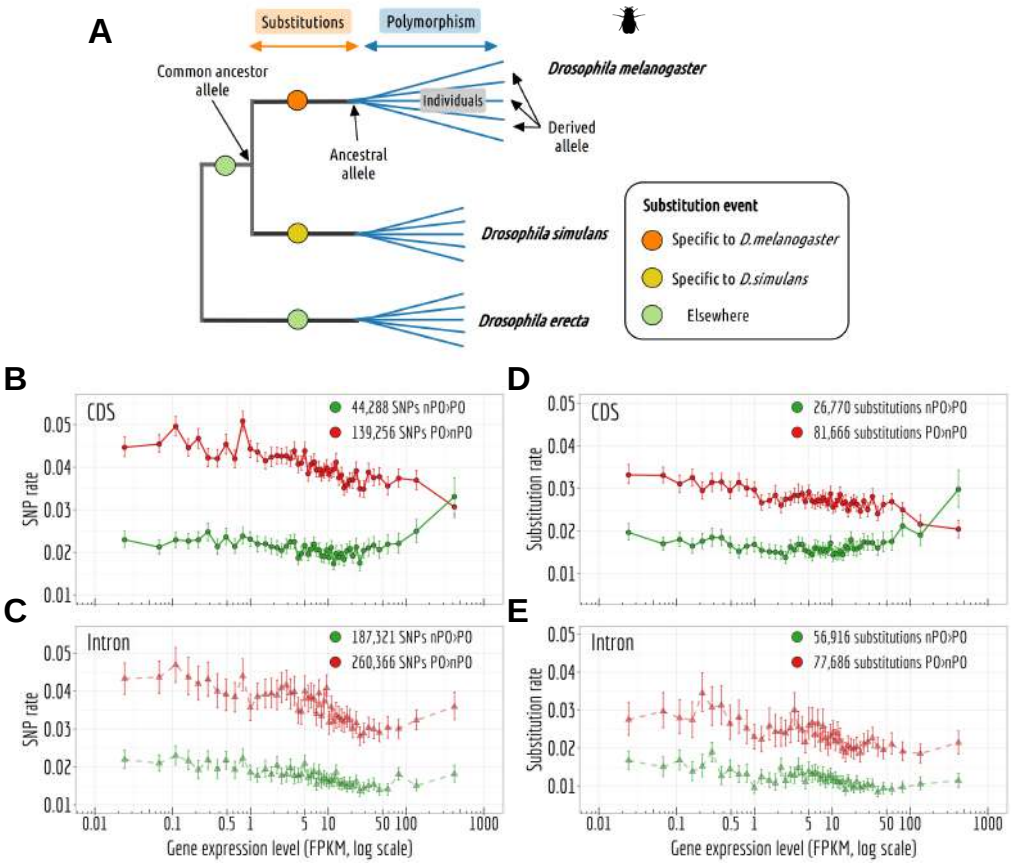


Figure 7.6: Selective pressure on non-POCs to POCs mutations. ***A****: Schematic representation of the method used to identify SNPs and substitutions in Drosophila melanogaster.* ***B,C****: Rate variations of SNPs non-POC towards POC (green) and POC towards non-POC (red) with gene expression in CDS (****B****) and in intronic control (****C****).* ***D,E****: Rate variations of substitutions non-POC towards POC (green) and POC towards non-POC (red) with gene expression in CDS (****D****) and in intronic control (****E****). Error bars represent the 2.5th and 97.5th percentiles of values obtained from 100 simulations using a binomial distribution, following the same site structure and substitution rate (see Materials & Methods).*

genes, compared to the 50% least expressed. The choice of this latter threshold is based on the observation that in species with clear signature of translational selection, POC frequencies show little variation in genes below the median expression level (Fig. 4B; Supplementary Fig. 9).

If constraints on synonymous codon usage are similar across species (*i.e.* if s^{hx} is constant), then S^{hx} is expected to vary linearly with the effective population size ($S^{hx} = 4N_e s^{hx}$). To test this prediction, we sought to estimate N_e for each species. Lynch and colleagues recently compiled a list of species for which the germline mutation rate (μ) and the level of neutral diversity (π_s) have been measured, and hence for which it is possible to infer the effective population size ($N_e = \pi_s/4\mu$) (Lynch *et al.*, 2023). This list included 24 species of our data set, and in addition allowed us to get a proxy of N_e for 17 species, for which species from the same genus were available. To explore the relationship between S^{hx} and N_e in more species, we also used three indirect proxies

(longevity, body length and the dN/dS ratio) that correlate with the effective population size (Supplementary Fig. 10).

Among the 223 species analyzed, the strongest intensity of selection is observed in the nematodes *C. elegans* (S^{hx} = 1.3) and *C. nigoni* (S^{hx} = 1.0; Fig. 7). Dipters also show relatively strong values of S^{hx} (mean = 0.63±0.16 sd), followed by lepidopters (mean S^{hx} = 0.41±0.15 sd). In vertebrates, signals of translational selection are weak (mean = 0.15±0.12 sd), but nevertheless, S^{hx} are on average significantly non null (Student's t-Test, p-value < 10^{-16}). Our estimates match with those previously published for *C. elegans*, *Drosophila melanogaster*, human and mice (dos Reis and Wernisch, 2009). As predicted by the drift barrier model, the species with the strongest signs of translational selection all show a relatively short lifespan, low body mass and low dN/dS (Fig. 7B,D), *i.e.* traits associated to organisms with large N_e. Conversely species with traits associated to low N_e all show low S^{hx}. However, the correlations between S^{hx} and N_e proxies are weak, and significant for only two of them (longevity and dN/dS) (Fig. 7B,D). The weakness of these correlation might be due to the fact that these traits are only indirect proxies of N_e. However, even for the few species for which it is possible to get more direct estimates of N_e (based on π_s and μ), the correlation between S^{hx} and N_e remains weak (Fig. 7A).

7.2.8 In species subject to translational selection, the tRNA pool evolves in response to changes in neutral substitution patterns

The above analyses show that for most metazoan species, translational selection is very weak, and hence that their synonymous codon usage is essentially shaped by neutral neutral substitution patterns (NSP). Interestingly, even in species with clear signal of translational selection, codon usage appears to be influenced by variations in NSP. Notably, Diptera and Lepidoptera span a wide range of GC-content in non-coding regions (genome-wide average GCi ranging from 0.25 to 0.43), that strongly correlates with their average GC3 (from 0.32 to 0.71; Fig. 1B). Yet, most of Diptera and Lepidoptera show a strong translational selection compared to other metazoans (26 out of 27 species with $S^{hx} > 0.3$, the only exception being *Eumeta japonica*, with S^{hx} = 0.09). This raises the question of how the tRNA pool evolved in these species in response to NSP changes. To investigate this point, we focused our analyses on the 26 Diptera and Lepidoptera species with a strong signal of translational selection.

In this dataset, we observed that the decoding of 11 NNA/NNG synonymous codon pairs (Glu, Gln, Lys, Val, Ala, Pro, Thr, Ser, both CTA/CTG and TTA/TTG pairs of Leu, and the AGA/AGG 'duet' of Arg) never involves wobble pairing: the two complementary isodecoder tRNAs (anticodons UNN and CNN, respectively) are systematically present altogether in their genome (Supplementary Fig. 12). Thus, for each of these 11 pairs, we identified the 'preferred' isodecoder tRNA (*i.e.* the one with the highest gene copy number) in each species. We observed that the proportion of the 11 preferred tRNAs

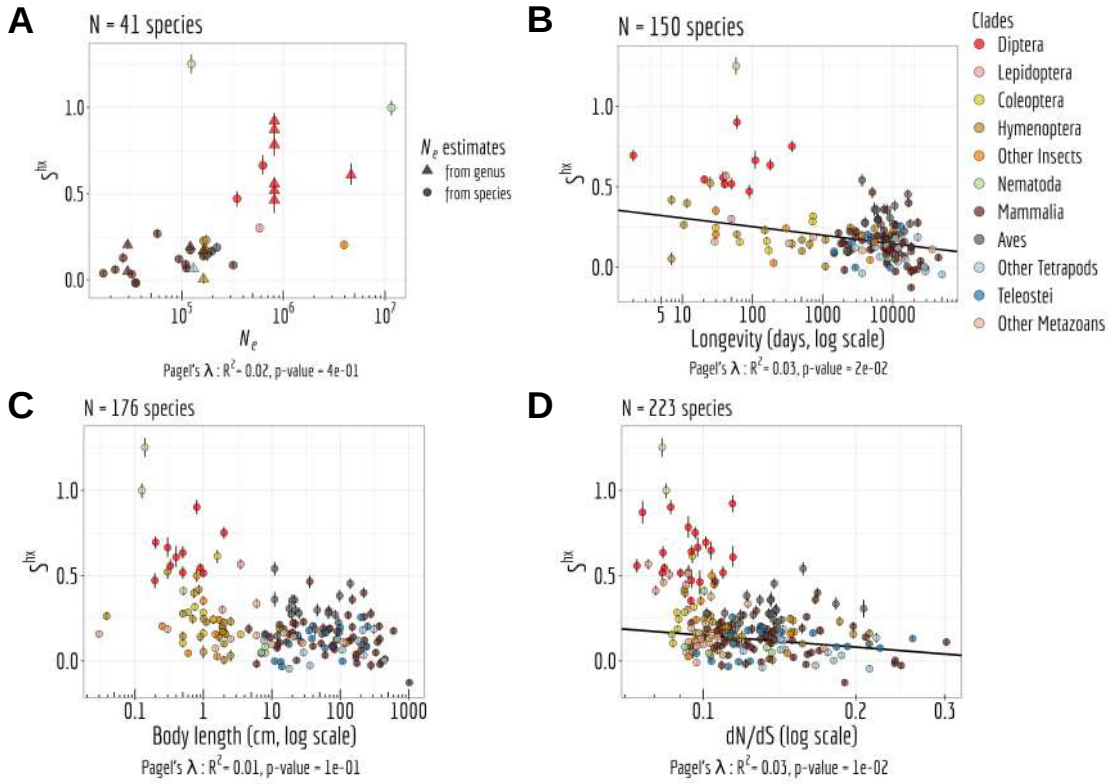


Figure 7.7: Relationship between N_e and translational selection intensity (S). *Relationship between the population-scaled selection coefficient (S) and N_e (**A**), longevity (days, log scale; **B**), body length (cm, log scale; **C**), dN/dS (log scale; **D**). The translational selection intensity S is measured on the top 2% most highly expressed genes (S^{hx}). Pagel's lambda model is used to take into account the phylogenetic structure of the data in a regression model (the regression line is displayed in black when the correlation is significant). Error bars represent the 2.5th and 97.5th percentiles of S values obtained from 1000 draws with replacement among the top 2% most highly expressed genes, and the 50% least expressed.*

having a CNN anticodon in a given species correlates positively with the GCi of species (Fig. 8A). This implies that tRNA gene copy number co-evolved in response to changes in NSP, consistent with the hypothesis that tRNA abundance is under selective pressure to match the demand in synonymous codon usage. For the two other synonymous codon pairs CGG/CGA of Arg and GGG/GGA of Gly, the CNN-tRNA is absent in 67% of cases, where UNN-tRNA decodes both codons NNA/NNG (Supplementary Fig. 12).

NNT/NNC synonymous codon pairs (N=16) are generally decoded by a single isodecoder tRNA (Fig. 3B, Supplementary Fig. 2). Among dipters and lepidopters, this is the case for 94% of the 416 NNT/NNC synonymous codons pairs analyzed (16 pairs × 26 species; Supplementary Fig. 12). In such cases, shifts in NSP cannot be compensated for by a change in the relative abundance of isodecoder tRNAs. Nevertheless, the affinity of tRNAs for their cognate codons can be changed by post-transcriptional modifications, and hence might evolve in response to the demand. To investigate whether such changes occur, it is necessary to identify which of the two codons is best decoded by this unique isodecoder tRNA. For this, we relied on the fact that in species that are subject to trans-

lational selection, codons that are more efficiently decoded show a higher prevalence in highly expressed genes compared to lowly expressed ones (these codons will hereafter be referred to as preferred codons). Two sets of NNT/NNC synonymous codons pairs can be distinguished: the 7 pairs corresponding to the amino acids with duet codons (Phe, Cys, Tyr, Asp, His, Asn and the AGT/AGC 'duet' of Ser), and the 9 pairs from amino acids with triplet (Ile) or quartet codons (Val, Gly, Ala, Pro, Thr, Leu, Arg, Ser). For NNT/NNC duets, when a single tRNA is present (95% of cases), it is always the GNN-tRNA, and in 99% of cases it is the NNC codon, decoded through Watson-Crick pairing that is preferred. For 8 of the 9 other pairs, when a single tRNA is present (94% of cases), it is always the ANN-tRNA, the only exception being Gly (GNN-tRNA). For Gly, the GGT codon, decoded via wobble pairing, is preferred to the GGC codon in 84% of species. For the other pairs (decoded by ANN-tRNA) there is more variability: the NNC codon (wobble pairing) is preferred in 79% of species, whereas the NNT codon (watson-crick pairing) is preferred in the others. These observations indicate that when a single tRNA is present for two codons, it is not systematically the one with watson-crick pairing that is the most efficiently translated. Furthermore, although the NNC codon tends to be preferred to the NNT codon (except for Gly), there are some variation across species, notably for those decoded by a ANN-tRNA (Supplementary Fig. 11). We computed in each species the proportion of NNC preferred codons among NNT/NNC synonymous codon pairs decoded by a single tRNA. Interestingly, the species showing the highest proportion of NNT preferred codons are the ones with the lowest genomic GC content (Fig. 8B). Thus, it appears that the relative affinity of ANN-tRNAs for the NNT or NNC codon can evolve in response to the demand.

These observations suggest a straightforward model to explain variation in the set of optimal synonymous codons across species (Fig. 8C). First, variation in mutational patterns or in the intensity of gBGC can lead to changes in the base composition of genomes, thereby directly shifting the codon usage of genes. Given that translational selection is a weak force, most genes are affected by this shift. This results in a change in the codon demand, and hence induces a selective pressure to change the pool of tRNA (both in terms of abundance and of affinity for their cognate codons). In turn, translational selection will modify the codon usage of highly expressed genes to match the new set of tRNAs, thereby reinforcing the selection on the tRNA pool to match the codon demand, and resulting to a new co-adaptation between the tRNA pool and codon usage.

7.2.9 Weak translational selection in species with large intra-genomic variability in neutral substitution patterns

An implicit assumption of the above model is that all genes of a given genome are affected by similar neutral substitution patterns. There is evidence however that some genomes are subject to heterogenous neutral substitution patterns. Notably, in mammals and birds, variation in recombination rates along chromosomes induce a strong heterogeneity in GC-content, driven by gBGC (Duret and Galtier, 2009). This process accounts for

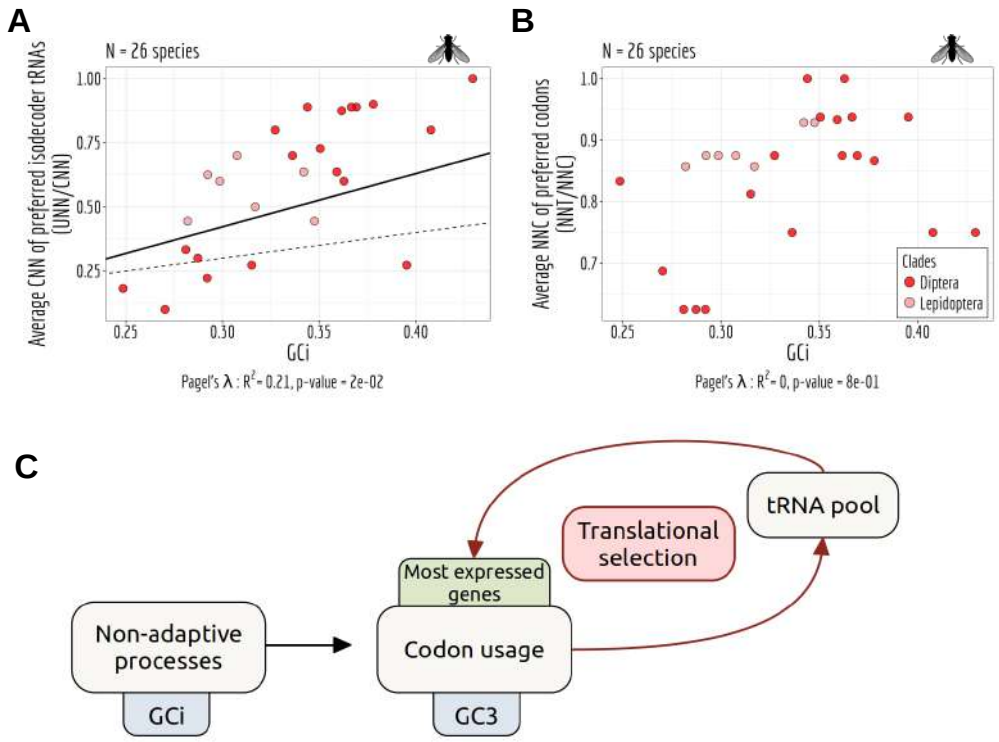


Figure 7.8: Genomic substitution pattern shapes the tRNA pool. ***A****: Relashionship between the per species CNN fraction of preferred isodecoder tRNAs (corresponding to the most abundant tRNAs) among 11 NNA/NNG synonymous codon pairs and the gene average GC in introns (GCi), for Diptera (dark red) and Lepidoptera (light red).* ***B****: Relashionship between the per species proportion of NNC preferred codons (the most overused codons in highly expressed genes compare to lowly expressed genes) among NNT/NNC synonymous codon pairs decoded by a single tRNA, along with the GCi.* ***A,B****: Pagel's lambda model is used to take into account the phylogenetic structure of the data in a regression model (black line if significant).* ***C****: Hypothetical schemes explaining how synonymous codon usage can be shaped conjointly by translational selection and by neutral substitution patterns.*

70% of the variance in synonymous codon usage among human genes (Pouyet *et al.*, 2017). Thus, in these species, the synonymous codon usage of a given gene essentially depends on the base composition of the genomic region where it resides, as shown by the strong correlation observed between GC3 and GCi across human genes (Fig. 1C). It is important to notice that these regional variations in GC-content affect all genes, even those that are widely expressed. To illustrate this point, we analyzed the codon usage of 2,249 human housekeeping genes (defined as genes that are in the top 20% most highly expressed genes in at least 75% of tissues). The distribution of GC3 in housekeeping genes shows a very strong heterogeneity (GC3 ranging from 25% to 95%; Fig. 9B), as strong as in the entire gene set (Fig. 9A). Housekeeping genes are involved in basal function and have to be expressed at high level in most cell types. This implies that in any given cell, there is a strong heterogeneity of the codon demand. Such a situation is predicted to hinder the co-adaptation between the tRNA pool and codon usage: any increase in the abundance

of a given tRNA (say decoding a GC-ending codon) is expected to be beneficial for the translation of GC-rich genes, but detrimental for the translation of the GC-poor ones (and vice versa for a tRNA decoding an AU-ending codon). Hence, the selective pressure imposed by the codon demand is expected to maintain a balanced tRNA pool, able to decode GC-rich genes as well as GC-poor genes. In turn, the presence of a balanced tRNA pool should reduce the difference in translational efficiency between synonymous codons, and hence is expected to decrease the intensity of translational selection. Thus, genomes that are subject to heterogenous neutral substitution patterns are expected to be less subject to translational selection. To test this prediction, we analyzed the relationship between the intensity of translational selection (S^{hx}) and the intra-genomic heterogeneity in base composition (assessed by the variance in GCi across genes). We observed that all species with a strong signal of translational selection show a very small variance in GCi, while species with a high variance in GCi show relatively low S^{hx} (Fig. 9C). This is consistent with the hypothesis that intra-genomic heterogeneity in base composition precludes translational selection. However, species with a high variance in GCi mainly correspond to three clades (Mammal, Aves, Hymenoptera) that have relatively small N_e, and hence it is difficult to disentangle the impact of intra-genomic heterogeneity in base composition from that of drift (Fig. 9D). In any case, even though intra-genomic heterogeneity in base composition might explain the weakness of translational selection in some species, there must be some other factors that affect the intensity of translational selection. Indeed, among insect species predicted to have a N_e similar to that of dipters, many show a low S^{hx} despite a small variance in GCi (Fig. 9E,F).

7.3 Discussion

7.3.1 Predicting translationally optimal codons

Patterns of SCU vary widely across metazoan species, and are strongly correlated to the base composition of their non-coding regions (Fig. 1A). This implies that variation in codon usage across species are primarily shaped by differences in genome-wide neutral substitution patterns (driven by the underlying mutation pattern, by gBGC or both). NSPs vary not only across species, but also along chromosomes, and in some clades, such as tetrapods or hymenopters, this intra-genomic heterogeneity of NSPs is a major determinant of the variance in SCU among genes (Supplementary Fig. 1B). The fact that genome-wide patterns of SCU are strongly affected by NSPs does not exclude that it can also be shaped by a selective pressure favoring the use of translationally optimal codons. Indeed, in dipters and lepidopters, which both show clear evidence of translational selection, we observed that the tRNA pool evolves in response to changes in genome-wide NSPs (Fig. 8). Thus, variation in NSPs can lead to shifts in the translation apparatus, and thereby drive the evolution of SCU, not only in weakly expressed genes where codon usage is effectively neutral, but also in genes under strong translational selection. In other words, selective and non-adaptive models are not mutually

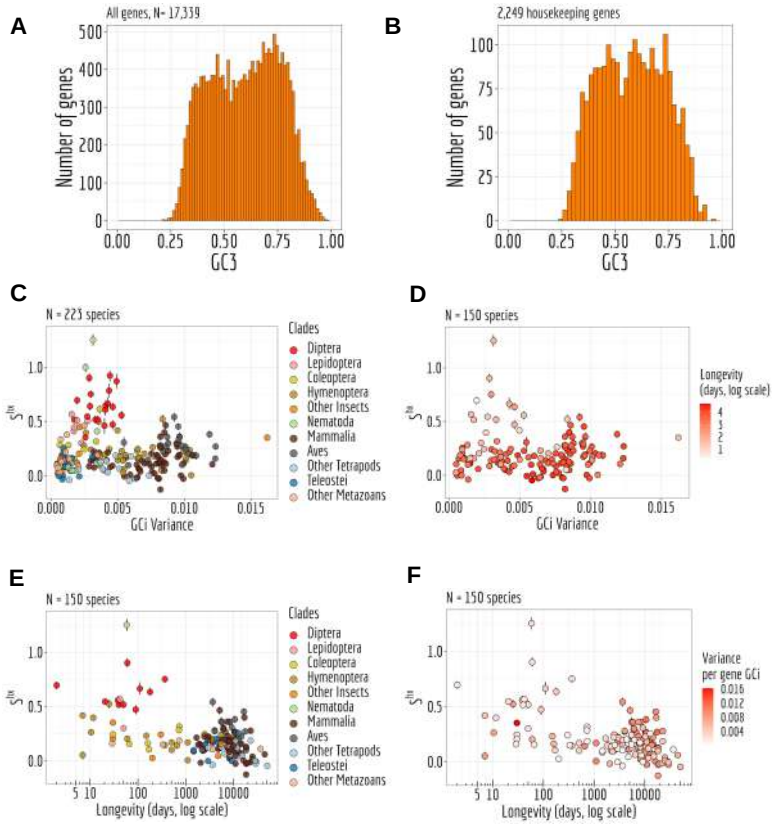


Figure 7.9: Large intra-genomic variability in neutral substitution patterns impact on translational selection. ***A****: Distribution of GC content at the third position of codons (GC3) across human genes.* ***B****: Distribution of GC3 across human housekeeping genes (identified based on gene expression data from 27 healthy tissues, extracted from (Pouyet* et al.*, 2017).* ***C,D****: Relation between translational selection intensity S and the gene GCi variance.* ***D****: Species are colored with a longevity gradient (log scale).* ***E,F****: Relation between translational selection intensity S and longevity (days, log scale).* ***F****: Species are colored with a GC intron variance gradient (log scale).*

exclusive, but it is important to take NSPs into account to be able to detect signatures of translational selection within genomes.

To quantify the intensity of translational selection in metazoans, we used a method based on standard population genetics equations, that infers the population-scaled selection coefficient ($S = 4N_e s$) from the difference in optimal codon frequency between highly expressed genes and weakly expressed genes (Sharp *et al.*, 2005; dos Reis and Wernisch, 2009). This method first requires to identify the set of optimal codons in each species. To predict optimal codons, previous studies generally searched for codons whose frequency increases with gene expression level (*e.g.* Duret and Mouchiroud (1999); dos Reis and Wernisch (2009)). One caveat, is that in some species, NSP varies with gene expression (Pouyet *et al.*, 2017), which may therefore lead to errors in the inference of optimal codons. Furthermore, in that situation, the method would systematically overestimate S for codons that are favored by NSPs in highly expressed genes. To limit this bias, we sought to predict optimal codons from the tRNA pool available in each species.

For this, we estimated the abundance of each tRNA based on its gene copy number in the genome. In most species, we observed a strong correlation between the number of iso-acceptor tRNA gene copies and the frequency of their cognate amino-acid in the proteome (Fig. 3B). These strong correlations are consistent with the fact that cellular tRNA abundance is highly constrained to match the amino-acid demand, and indicate that tRNA gene copy number is a good proxy to infer tRNA abundance, in agreement with previous experimental evidence from a limited set of species (Behrens *et al.*, 2021). Based on our estimates of the tRNA pool, we predicted two sets of putative optimal codons (POCs): for amino acids for which more than one iso-decoder tRNA is available, optimal synonymous codons were defined as those decoded by the most abundant tRNA (POC1); for amino acids encoded by NNC/NNU duet codons and with one single iso-decoder tRNA (GNN), the NNC codons were predicted to be optimal (POC2), based on previous studies showing that the wobble pairing GNN:NNU was less efficient than the Watson-Crick pairing GNN:NNC (Stadler and Fire (2011); Chan *et al.* (2017); Fig. 3B).

Several lines of evidence indicate that our predictions of translationally optimal codons are accurate. First, our sets of POCs are consistent with previous predictions: we identified 25 POCs in *C. elegans* and 27 in *D. melanogaster*, while respectively 26 and 25 optimal codons had been inferred based on difference in codon usage between highly and lowly expressed genes (Duret and Mouchiroud, 1999), of which 88.4% and 88.0% match with our POCs. Furthermore, the analysis of substitution patterns and polymorphism in *Drosophila melanogaster* confirmed that selection favors POC alleles over non-POC alleles in highly expressed genes (Fig. 6). Finally, we observed that although the definition of POC1 and POC2 relies on very different principles, the two sets of codons show very similar signatures of translational selection (Fig. 4C).

7.3.2 Variation in the intensity of selection in favor of translationally optimal codons across metazoans

For each species, we measured the frequency of optimal codons (combining POC1 and POC2) in highly expressed genes (top 2%), to estimate the population-scaled selection coefficient in favor of translationally optimal codons ($S = 4N_e s$), using weakly expressed genes as a reference to account for the NSP (Sharp *et al.*, 2005). Across the 223 species, the highest values of S are observed in Caenorhabditis nematodes ($S = 1.25$ in *C. elegans* and $S = 1.00$ in *C. nigoni*). We also found a clear signal of translational selection in diptera (mean $S = 0.63$, N=19 species), and to a lesser extent in lepidoptera (mean $S = 0.41$, N=8 species). Overall, estimates of S are weaker in other clades (Fig. 7). The weakness of translational selection in vertebrates (mean $S = 0.16$, N=100 species) was a priori expected given that these organisms tend to have relatively small N_e. But besides Caenorhabditis and dipters, our dataset included 83 invertebrate species covering a wide range of clades (58 other insects, 12 other Ecdysozoa, 6 Spiralia, 4 Cnidaria, 3 Deuterostomia). What is surprising is that all these species show S values that are lower than the average of dipters. This implies that the high values of S observed

in Caenorhabditis and in dipters represent exceptions rather than the rule, and that translational selection is weak in most metazoan lineages.

If the selection coefficient in favor of optimal codons (s) was constant across metazoans, then S should scale linearly with N_e. To test this prediction, we used silent-site polymorphism and germline mutation rate data (Lynch *et al.*, 2023) to estimate the effective population size ($N_e = \pi_s/4\mu$, hereafter noted $N_e^{\pi\mu}$) in 41 species. As expected S tends to increase with $N_e^{\pi\mu}$, but the correlation is not significant after accounting for phylogenetic inertia (Fig. 7A). The weakness of the correlation might be due to the fact that these two parameters evolve on different time scales: $N_e^{\pi\mu}$ is indicative of the recent effective population size (on the order of N_e generations) and hence can change quite rapidly compared to S, that is estimated from the codon composition of genomes, resulting from a long-term accumulation of substitutions. This can explain why *C. nigoni* and *C. elegans* display similar values of S, despite a 75-fold difference in $N_e^{\pi\mu}$ (respectively $N_e^{\pi\mu} = 9.4 \times 10^6$ and $N_e^{\pi\mu} = 1.2 \times 10^5$). This difference in $N_e^{\pi\mu}$ is due to the fact that *C. nigoni* is an outcrossing species, like most other Caenorhabditis species, while the *C. elegans* lineage evolved towards selfing hermaphroditism (Li *et al.*, 2014; Vielle *et al.*, 2016). This transition in reproductive mode is recent, and hence the SCU of *C. elegans* still retains the signature of strong translational selection inherited from its outcrossing ancestors. Thus, we can predict that the SCU of *C. elegans* is not at selection-mutation-drift equilibrium (which could be tested by analyzing synonymous polymorphism).

To further test the relationship between S and N_e, we considered three parameters (longevity, body length and dN/dS), that are all correlated with $N_e^{\pi\mu}$ (Supplementary Fig. 10), but that are expected to reflect N_e over a longer time scale. A further interest of these proxies is that they can be estimated on much larger datasets (150 to 223 species). But here again we obtained similar results: S tends to increase with N_e, but correlations are weak, marginally significant after accounting for phylogeny (Fig. 7B,C and D). The weakness of the correlation is mainly due to the fact that some species have a low S, despite life-history traits or dN/dS values indicative of a high N_e.

Not only the correlations between N_e and S are weak, but also the range of variation in S appears to be quite limited compared to what would be expected given the variance in N_e. For instance, the mean value of $N_e^{\pi\mu}$ is about 15 times higher in diptera than in mammalia (based on respectively 6 and 41 species for which $N_e^{\pi\mu}$ can be estimated; Lynch *et al.* (2023)). Yet, the mean value of S is only 5.3 times higher in diptera (mean $S = 0.63$, N=19 species) than in mammals (mean $S = 0.12$, N=65 species). Thus, the difference in S between diptera and mammals is smaller than what would be expected if S scaled linearly with N_e.

One possible explanation for this discrepancy is that S is overestimated in mammals. As discussed by dos Reis and Wernisch (2009), the estimate of S is based on the assumption that NSPs are constant across genes, *i.e.* that the difference in optimal codon frequency between highly and weakly expressed genes is entirely due to translational selection. In reality, in some species, NSPs vary with gene expression (*e.g.* in humans Pouyet *et al.* (2017)). To try to account for these variations, we measured the differences

in POC frequency between highly and weakly expressed genes, controlling for differences in the corresponding triplet frequencies in introns. It is however possible that the base composition of introns is not a perfect predictor of NSPs, notably because introns are affected by indels and transposable elements, which are not allowed in coding regions. This is well illustrated by POC2 codons in humans, whose frequency clearly covaries with their non-coding controls, but with a wider amplitude in exons than in introns (Fig. 4A).

An alternative hypothesis is that the discrepancy might result from a strong heterogeneity in the fitness effect of synonymous mutations. Indeed, the analysis of synonymous polymorphism in *D. melanogaster* indicated that a majority of codons are under weak selection in favor of translationally optimal codon ($|N_e s| \approx 1$), but that a small fraction (10%-20%) are under strong selection ($|N_e s| > 10$; Machado *et al.* (2020). With a N_e value 15 times lower, the first class of codons is expected to evolve neutrally in mammals. But the second class of codons would still appear under effective translational selection, which might explain the small but non-null value of S measured in mammals.

One last unexpected observation is that many species predicted to have a high N_e (based on their LHTs or dN/dS) show very weak S (Fig. 7B,C and D). In some species, this could be explained by the heterogeneity of NSPs along chromosomes, inducing a strong variance in SCU that precludes a co-adaptation of the tRNA pool. This might be the case notably for some hymenopters, which, like tetrapods, are subject to gBGC (Wallberg *et al.*, 2015) and present a very strong heterogeneity in NSPs (Fig. 9C). However, our dataset also includes some species with small S values, despite a high N_e proxy and homogenous NSPs. So finally, we are left with the conclusion that variation in S across metazoans are not driven simply by the drift barrier and by gBGC, but that they are also probably due to variation in s, the selection coefficient in favor of translationally optimal codons. There is evidence that in unicellular organisms, the selective force for the optimization of SCU is the maximization of cellular growth (Rocha, 2004; Sharp *et al.*, 2005). It is possible that the selective pressure on cellular growth also vary across metazoans. Most Caenorhabditis species grow in ephemeral environments (rotting vegetation) and hence have been selected for their capacity to proliferate very rapidly (Cutter, 2015). Manthey *et al.* (2024) recently quantified growth rates in 33 insects. The only dipter present in their dataset (*Lucilia sericata*) is the species that presented the highest growth rate, 12 times higher than the average growth rate of other holometabole insects (N=10) and 52 times higher than the average growth rate of hemimetabole insects (N=22). If the *Lucilia sericata* is representative of other dipters, this might explain why translational selection is particularly strong in that clade compared to other insects. It is noteworthy that the two invertebrate species that have been historically used as model organisms (*D. melanogaster* and *C. elegans*) both belong to the very rare metazoan clades with clear evidence of translational selection. This might reflect the fact that they have been chosen a model organisms for the very reason that they can grow very fast in the lab.

7.4 Materials & Methods

7.4.1 Gene expression and data collection

The reference genome assemblies and genome annotations were acquired from the National Center for Biotechnology Information (NCBI; Sayers *et al.* (2022a). We obtained gene expression data for 257 metazoan species from GTDrift (Bénitière *et al.*, 2024), where Fragment *Per* Kilobase of exon *per* Million mapped reads (FPKM) was estimated over thousands of RNA-seq samples using cufflinks. For each species we considered the *per*-gene median FPKM values across all analyzed RNA-seq samples. Additionally, a phylogenetic tree was retrieved from GTDrift to account for phylogenetic inertia (Bénitière *et al.*, 2024).

7.4.2 tRNAscan-SE annotation

If for a given species tRNA genes copies were previously annotated and so present in the NCBI annotation file, we took these annotations into account (N=44 species). Other wise (N=213 species) we annotated tRNA gene copies using the program tRNAscan-SE 2.0.12 (Nov 2022), with the -E option specifically designed for eukaryotic tRNA identification search (Chan *et al.*, 2021). To keep in the study functional gene copies we retained those with a score exceeding 55, threshold based on Chan *et al.* (2021) analysis. Thus, for each of these copies we obtained the decoded codon and the translated amino acid.

7.4.3 Codon usage

For each species in our study, we conducted a detailed assessment of codon usage across the longest annotated coding sequences (CDS) of each expressed gene within our dataset. This analysis was paralleled by an examination of the occurrences of nucleotide triplets within intron regions. It is important to note that our analysis deliberately excluded the acceptor and donor splice sites to avoid skewing the results with these highly conserved motifs.

7.4.4 Site constraint

Multiple gene alignments of 976 BUSCO genes and 293 species were collected to study site constraints from the metazoa dataset alignment of GTDrift repository (Bénitière *et al.*, 2024). For each gene, we determined the proportion of gaps at each site across the alignment. This information guided the *per*-species segmentation of genes into bins, with each bin representing 25% of a gene sequence. Our examination of tetrapods focused on sites located beyond 1,000 base pairs downstream from the start codon.

7.4.5 SNPs analysis

We used polymorphism data from the *Drosophila* Genetic Reference Panel 2 (DGRP2) (Mackay *et al.*, 2012; Huang *et al.*, 2014), where polymorphic sites have been identified from comparisons across 205 inbred lines of *Drosophila melanogaster*, downloaded from `http://dgrp2.gnets.ncsu.edu/data/website/dgrp2.vcf`. We converted the single nucleotide polymorphism (SNP) coordinates from the dm3 genome assembly to the dm6 assembly, with the liftOver utility (Hinrichs *et al.*, 2006) of the UCSC genome browser, using a whole genome alignment between the two assemblies downloaded from `https://hgdownload.soe.ucsc.edu/goldenPath/dm3/liftOver/dm3ToDm6.over.chain.gz`. We kept in the study 3,738,302 biallelic SNPs for which more than 181 individuals have been genotyped.

We then identified two sister species *Drosophila simulans* and *Drosophila erecta* that we aligned against *Drosophila melanogaster* genome using liftOver utility (Hinrichs *et al.*, 2006). We removed from the analysis genes located in regions where the multiple alignment was of poor quality (Supplementary Fig. 8. We used the program est-sfs release 2.04 (Keightley and Jackson, 2018) to polarize SNPs, *i.e.* to identify the ancestral allele and the derived allele.

For the longest annotated coding sequence of each expressed gene we were able to identify the ancestral and derived codons for each SNP. The same approach was applied for intron regions by studying nucleotide triplets.

To determine the confidence interval for each data point (Fig. 6), we employed a simulation approach. We simulated a sample with a similar structure in terms of the number of sites *per* gene and the SNP rate, using a binomial distribution. We calculated the average SNP rate for each simulation, repeating the process 100 times. Afterward, we utilized the 2.5th and 97.5th percentiles of these mean values to establish the error bars.

7.4.6 Substitutions analysis

Based on the multiple genome alignment of *Drosophila simulans*, *Drosophila erecta* and *Drosophila melanogaster* previously described, we identified the non polymorphic sites where a substitution occurred. To do so, we studied sites for which we were able to determine the ancestral and substituted allele by considering the most parsimonious scenario : if reference alleles of $D.erecta = D.melanogaster$ or $D.simulans = D.melanogaster$ there is no substitution; if reference alleles of $D.erecta = D.simulans$ but differ from $D.melanogaster$ then there is a substitution and the ancestral allele is the one observed on $D.erecta$ and $D.simulans$.

We identified a total of 1,759,664 substitutions, and were able for each codon containing at least one substitution to determine its ancestral and substituted state. The same approach was applied for intron regions by studying nucleotide triplets. This protocol was executed on the longest annotated coding sequence of each expressed gene.

7.4.7 Effective population size estimates

We retrieved proxies for the effective population size from the GTDrift data resource (Bénitière *et al.*, 2024), which included life history traits such as body length, longevity, and the ratio of non-synonymous to synonymous substitutions rate (dN/dS). It is expected that the genome-wide dN/dS ratio increases during prolonged periods of small N_e, attributed to the fixation of slightly deleterious mutations (Ohta, 1992; Galtier, 2016). To enhance the dataset, we supplemented the effective population size proxies with body mass data extracted from Lynch *et al.* (2023) for 45 species, encompassing 26 species within our dataset and 19 for which species from the same genus were available.

Furthermore, from Lynch *et al.* (2023), we obtained direct estimates of N_e by deriving the effective population size ($N_e = \pi_s/4\mu$) using the germline mutation rate (μ) and the level of neutral diversity (π_s) for 45 species, comprising 27 species within our dataset and 18 for which species from the same genus were available. Additionally, we expanded our dataset with the N_e estimate for *C. nigoni* by including $\pi_s = 0.06$ (Asher Cutter, personal communication) and $\mu = 1.3 \times 10^{-9}$ (Denver *et al.* (2012); assuming a similar mutation rate as in *C. briggsae*). We calculated $N_e = \pi_s/4\mu \approx 1.1 \times 10^{-7}$.

Acknowledgements

Computational analyses were performed using the computing facilities of the CC LBBE/PRABI and the Core Cluster of the Institut Français de Bioinformatique (IFB) (ANR-11-INBS-0013). Silhouette images of taxonomic Families originate from PhyloPic developed and maintained by Mike Keesey available at https://www.phylopic.org/.

Author contributions statement

F.B. conceived the pipeline and conducted the analyses. F.B. drafted the manuscript. All authors reviewed the manuscript.

Funding

This work was funded by the French National Research Agency (ANR-20-CE02-0008-01 "NeGA").

Competing interests

The authors declare no conflicts of interest.

Part III

Discussion & Perspectives

8

Discussion & Perspectives

Contents

8.1 Summary of main results

During my thesis I analyzed transcriptomic and genomic data, organized in a data resource including almost 16,000 RNA-seq samples and 1,507 species along with proxies of the random genetic drift intensity. These information have been used to study how random genetic drift affects alternative splicing and translational selection across metazoans. I summarize in the following sections the main findings of my thesis.

8.1.1 Alternative splicing, a genetic burden limited by drift

In the first scientific study we investigated the alternative splicing products, alternative variants, and their functional significance across several metazoans. We have developed protocols to tackle the question from different angles. The main one was to use the "drift barrier" hypothesis, according to which biological processes within a genome will be optimized up to the limit imposed by genetic drift. Indeed, to rephrase it briefly, Lynch postulated that the genomes of species with small effective population size would

be subject to more intense genetic drift compared to species with high effective population size, thus reducing the effectiveness of selection to purge slightly deleterious mutations in small N_e species.

Through the estimation of the *per* intron alternative splicing rates across 53 species, our results demonstrate a negative correlation between alternative splicing and effective population size. This relationship is robust to phylogenetic inertia and the quantity of transcriptomic data analyzed. Thus, the increase in the rate of alternative splicing between species (from 0.8% to 3.8%) mainly reflects the increase in the intensity of genetic drift, and corresponds to transcription errors whose quantity is modulated by drift.

In a second protocol we identified two categories of introns, rare splice variants (SVs) representing the vast majority of the repertoire of splicing isoforms (from 62.4% to 96.9%) and abundant SVs. We observed that abundant SVs have a strong signal of functionality, indeed up to 70% are frame preserving compared to 33% in rare variants, a rate expected if the splice site is randomly selected on the gene. Also, the AS rate measured on rare SVs is strongly related to N_e, as expected under the "drift barrier" hypothesis, which states that errors should increase with decreasing N_e. This relationship does not hold for the AS rate measured on abundant SVs, which are supposed to contain a large proportion of functional transcripts.

Another line of research consisted of studying splice sites constraints, by comparing those of the spliced variants and those of the main isoforms in *Drosophila melanogaster* and *Homo sapiens*. Our results show that, in *Homo sapiens*, the splice sites of the main isoform are constrained, but the spliced variants do not present any particular constraint compared to the control regions. Whereas in *Drosophila melanogaster*, there is selection on splice sites of the most abundant SVs. These observations also support the hypothesis that AS products are predominantly non-functional and therefore not under selective constraints, except for abundant SVs in some species, such as *Drosophila melanogaster*.

Finally, we investigated whether low-expressed genes have more rare variants than high-expressed genes, as we expect them to be purged more efficiently into the latter category if they arise primarily from splicing errors. For most species, the rate of rare variants decreases with gene expression accordingly to our predictions.

All in all, our first study reveals that AS mainly reflects erroneous transcripts which rate is controlled by the intensity of random genetic drift in metazoans.

8.1.2 Translational selection is rare in metazoans: variations in drift and fitness

In our second scientific study we analyzed codon usage variations among metazoans, focusing our analyses on translational selection which promotes codons optimizing translation process. Our first observation was that inter-species variations in codon usage are strongly influenced processes impacting both coding and non-coding sequences, called neutral substitution patterns (NSP).

Subsequently, in each species, we identified the set of codons decoded by the most

abundant tRNAs, that we called putative-optimal codons (POCs), predicted to be codons promoted by TS. Interestingly, highly expressed genes are enriched in POCs compared to low expressed genes in most studied species. This enrichment reaches +26% in *Caenorhabditis elegans*, +14% in flies, and a mere +3% in vertebrates. We further showed that constrained sites of a gene tend to overuse POCs compared to less constrained sites. Additionally, analyses on substitution patterns and polymorphism in *Drosophila melanogaster* reveal that non-POCs towards POCs substitutions are favored in highly expressed genes compared to lowly expressed genes. These analyses strongly suggest a selection to promote the use of codons that match the tRNA pool.

Then, we investigated, for species for which TS is effective, how the tRNA pool responds to variations in neutral substitution patterns (NSP). This question is particularly interesting in Diptera and Lepidoptera because of the strong TS signal that coexists with large variations in NSP across species. We demonstrated that the translation machinery is co-adapting to the NSP changes by modulating both the tRNA abundance and the tRNA affinity for a particular codon.

Overall, our results show that TS is scarce in metazoans, with a small population-scaled selection coefficient (*i.e.* $S < 1$), and that species where NSP is detected correspond to species with large N_e. In this range of S values, the "drift barrier" suggests that N_e must be large for selection to be efficient and promote codons optimizing translation. Indeed, in small N_e population TS is barely observed in our dataset. However, while TS is observed only in large N_e species, some large N_e species also show no TS signal.

Finally, we investigated how TS can become ineffective due to heterogeneous neutral substitution patterns. It appears that species with heterogeneous NSP do not present NSP signal, and that TS is only observed in species with homogeneous NSP. However, some species with a large N_e and a homogeneous NSP do not exhibit TS signal, such as some hymenopterans. These results lead us to hypothesize that the selective advantage in optimizing the translation machinery is not the same for all species.

8.2 Discussion

I explore the possible consequences of this thesis on other scientific questions by first presenting how the "drift barrier" hypothesis can be useful for deciphering what is adaptive or not. Then, by presenting how our results could be interesting in applied scientific subjects.

Also, in the following sections, I discuss how we/I, as scientists, can work to bring compelling reproducible data to the community. I will delve into the accessibility and reproducibility of the data in research with the tools available to the bioinformaticians today, on which I devoted a lot of time to provide all the information necessary for the reproduction of our articles, data and results.

8.2.1 The "drift barrier", an attractive framework

In biology it is common to study biological processes as if they were adaptive. But we know that the non-adaptive forces cannot be systematically ruled out, and need careful consideration (Lynch, 2007a). In population genetics, the "drift barrier" hypothesis is one of the most attractive concept to examine non-adaptive vs adaptive model. Theoretically slightly deleterious/advantageous mutations with $|N_e\ s| \ll 1$ propagate in the population as if they are neutral. Thus, if s is constant a decrease in N_e implies that more and more slightly deleterious mutations behave neutral and thus have a greater chance to reach fixation in a species. With the same reasoning, advantageous mutations will behave as neutral and will have less chance of reaching fixation than in large-N_e population.

This observation led Lynch to propose that biological processes, as they approach optimality, will encounter a barrier beyond which any further optimization will be hampered by drift (Lynch *et al.*, 2016). Indeed, for a trait close to optimality, new beneficial mutations are supposed to have diminished fitness advantages, decreasing s, and will behave as neutral.

The question is whether this could be observed in nature: are there features of the genome that actually accumulate a slightly deleterious burden, or purge that burden, due to the change in N_e? Does N_e alone determine the level of optimization of biological processes?

First, Lynch observed that the mutation rate per generation was linked to the N_e (Lynch, 2010; Sung *et al.*, 2012; Bergeron *et al.*, 2023). Thus, he concluded that selection operates to minimize the mutation rate, with an efficiency limited by random genetic drift. The genome size of Asellid isopods has also been shown to increase as long-term N_e decrease, due to an accumulation of repeated elements (Lefébure *et al.*, 2017). However, in some other metazoan clades the predictions are not observed (Whitney and Garland, 2010; Roddy *et al.*, 2021; Marino *et al.*, 2024).

During this thesis we showed that AS is correlated with genetic drift intensity, supporting the idea that selection tends to optimize a low rate of AS, but that drift keeps it quite high for species with small N_e. These observations, combined with others, led us to conclude that AS products are primarily errors in low-N_e species. However, in Chapter 7, we showed that in metazoans, N_e might be responsible for variations of TS intensity but is not the only factor. These results suggest that if our measure of N_e is correct, and genomes have reached equilibrium, other parameters than N_e are at play on translational selection. For instance, the fitness landscape of optimizing translational machinery may differ across species, *i.e.* fast growing species (Manthey *et al.*, 2024) could have better interest to optimize translation than species with slow growth rate (Rocha, 2004).

Overall, by acknowledging that both selection (s) and drift (N_e) are at play, the "drift barrier" provides an interesting framework to ask whether biological processes, or genomic traits, are actually adaptive. For small population-scaled selection coefficient, our studies show that there are cases where the "drift barrier" hypothesis makes it possible to explain why genomic characteristics vary, and how. As such, non-adaptive evolution of

certain aspects of genome architecture cannot be overlooked. With this in mind, human species must be studied with extreme caution, especially since biologists tend to draw sweeping conclusions about the extreme complexity of our genome, when in reality, we are part of the species which exhibit the greatest random genetic drift, making us more vulnerable to the accumulation of genetic burden.

8.2.2 The limit of the "drift barrier" approach

While we showed that drift impacts some fundamental processes that are not under strong selection (*i.e.* small population-scaled coefficient), it is not clear if this test would be appropriate for other traits under stronger selection. If there is a causal relationship between N_e and a trait, it seems relevant to interpret what is the adaptive significance of a trait (increasing or decreasing) providing an indication of its biological functionality, which could ideally be complemented by other indicators.

However, if there is no relationship, interpretation is very difficult and requires extreme caution. Indeed, we can invoke different reasons to explain this absence of relationship: the trait is not at equilibrium selection/drift; the drift proxy is noisy; the N_e used is not relevant for this trait selection/drift balance (*i.e.* short-term vs long-term N_e); the fitness landscape varies (*i.e.* not the same interest to optimize a trait in each species). Also, the non-existence of the relationship between drift and a trait variations may simply be a true observation due to the fact that N_e varies in a range that does not apply to the "drift barrier" either because this trait is subject to strong selection (*i.e.* $S \gg 1$), or because this trait evolve neutrally (*i.e.* $S \ll 0.01$).

We must be careful not to reproduce the same cognitive biases that we criticized previously. This means that we should not over-interpret our results, nor indirectly force expected correlations, but keep in mind that inconclusive results are still results. It is encouraging to observe in the literature that we accept that the hypothesis may not work. For example the most notable variation in genome architecture is genomes size, and this has recently been shown to not support the "drift barrier" hypothesis (Roddy *et al.*, 2021; Marino *et al.*, 2024).

Unfortunately, to test this attractive "drift barrier" hypothesis, we only have the combination of N_e and s that biology on earth offers us. We are in a laboratory where the possibilities for variations of N_e and s are limited, and where many other parameters, that we cannot control, change.

8.2.3 Potential consequences for future research

As mentioned in the introduction, many scientists consider that the primary purpose of alternative splicing is to increase the functional repertoire of genomes, particularly ours (Graveley, 2001; Black, 2003; Pan *et al.*, 2008; Nilsen and Graveley, 2010; Blencowe, 2017). These far-reaching conclusions have already permeated the scientific community without clear evidence, as it can be read in many recent papers that AS *'contributes to the majority of protein diversity'* (Jiang and Chen, 2020; Verta and Jacobs, 2022; Singh

and Ahi, 2022; Manuel *et al.*, 2023), with some still pointing out that there is controversy over this (Pozo *et al.*, 2021; Wright *et al.*, 2022; Singh and Ahi, 2022; Manuel *et al.*, 2023).

These sloppy conclusions have reached pharmaceutical studies, particularly in oncology where AS is widely studied for its implication in tumor development (Venables, 2004; Kalnina *et al.*, 2005; Srebrow and Kornblihtt, 2006; David and Manley, 2010; Huang *et al.*, 2020; Qi *et al.*, 2020; Sciarrillo *et al.*, 2020). In some studies all variants are considered as functional, and disruption in AS events is interpreted as a loss of function (Schmitz *et al.*, 2020; Cummings *et al.*, 2020). But it might be important to keep in mind that most AS events are actually irrelevant, functionally speaking, and taking this into account could help improve protocols and avoid misinterpreting its results. Our work (Bénitière *et al.*, 2024) complements studies that investigate the relative proportion of functional product of AS, concluding that in human most AS variants are errors and the "one gene many proteins" hypothesis corresponds to rare cases (Pickrell *et al.*, 2010; Gonzàlez-Porta *et al.*, 2013; Tress *et al.*, 2017b,a; Saudemont *et al.*, 2017). Also, we identified a set of variants that seems to be functionally relevant in most species, *i.e.* the abundant spliced variants. These results appear useful for prioritizing further investigations in more applied research, aimed at studying how AS modulates phenotypes (Verta and Jacobs, 2022; Singh and Ahi, 2022), diseases (Scotti and Swanson, 2016), drug development (Ren *et al.*, 2021) *etc.*

Our second analysis aligns with papers identifying no, or negligible, translational signals in humans or other vertebrates (Mouchiroud *et al.*, 1988; Kanaya *et al.*, 2001; Duret, 2002; Pouyet *et al.*, 2017). Indeed, we searched for translational selection in 250 metazoans and found it to be negligible in vertebrate species, contrary to the findings of other papers (Chamary *et al.*, 2006; Gingold *et al.*, 2014; Dhindsa *et al.*, 2020). These articles often lack negative control or they are misinterpreted. For example Gingold *et al.* (2014) observed that gene sets belonging to different functional categories have a different codon usage, which they interpreted as selection on the translation program for cell proliferation and differentiation. But in fact Pouyet *et al.* (2017) showed that these differences are linked to recombination, a process impacting both coding and non-coding regions, thus unrelated to the translation process. This underscores the necessity to have neutral control when searching for adaptive traits, especially in this controversial case. Interestingly we showed that we can predict a set of codons optimized for translation based on the tRNA pool. Thereby, just because codon usage in primates is not optimized does not mean it cannot be optimized. This latter statement is particularly interesting for a biological field of genome recoding. Indeed, many scientists are working on the incorporation of synonymous mutations to improve cellular properties (Singh *et al.*, 2021), or therapeutic strategies to prevent viral diseases (Martínez *et al.*, 2019). A striking example is the development of the mRNA vaccine encoding the SARS-CoV-2 Spike during the pandemic. The synthesis of this mRNA requires choosing which synonymous codons to use in order to optimize immunogenicity (Giménez-Roig *et al.*, 2021; Lai *et al.*, 2023; Zhang *et al.*, 2023). Our work, sheds light on how recoding can be prioritized, by preferentially targeting putative optimal codons, decoded by the most abundant tRNA, and by taking into account wobble pairing. Our claims are mainly based on dipterans

where codon usage appears to be much more driven by translational selection than in our genome. However, other species exhibit a non-null population-scaled selection coefficient, meaning that we might indeed, with this protocol, capture codons optimizing translation.

8.2.4 Data accessibility

To convince, researchers need to share as much data used in their paper as they can. Even if the methods and the results presented in a scientific paper are peer reviewed, it is not rare to realise that data are not always shared, or can be erroneous in comparison to what is published. This may be due to the human cost of revising a paper which does not allow a researcher for more time to check this kind of details systematically, but also because many paper don't linger to share there data (Dance, 2023). In my little experience, I have often encountered this kind of limitation when I wanted to collect data from a paper, which undoubtedly led one to be skeptical of the article itself, as it couldn't even get its hands on the most basic data, like that used in the charts. In this regard, this discredits the message given in a scientific article. An article could share all data, in order to be independent of the author, who may no longer work in the field. Due to technical advances in machine learning and computing, it might be reasonable to expect that in the near future scientific journals will come up with an automated method/pipeline to at least check whether all the numbers/graphic of an article are reproducible from data provided by authors (Schulz *et al.*, 2022).

This leads me to discuss how researchers can share persistent data with available online archives, such as Zenodo. I myself used Zenodo to share a larger amount of additional data, and provide everything necessary for reproducibility. Zenodo is free and was built and developed ten years ago by researchers to promote Open Science and Open Data as part of the OpenAIRE project. It allows researchers to share data to which a DOI is attributed for each change made to the repository. Thus, one can track the version of the scientific paper, linked to the version of the Zenodo archive. In these repositories can be shared many tools that have been and are developed to enable more reproducible research. Meaning that processed files and results can be reproduced based on the source data using the same program and the same version.

8.2.5 Reproducibility

Multiple tools can be used and nested together. For example, the pipeline used in a bioinformatic analysis can be described in a snakemake file (Köster and Rahmann, 2012). Snakemake is a Python based workflow management program with which a bioinformatician describes the different step, program, codes used to produce each file, resulting in a tree structure of the pipeline. Then, by mentioning what output a researcher expects, the snakemake will identify a chain of jobs to be executed, parallelize them if possible, and re-execute those that are obsolete due to corrupted output files. Snakemake can be used on clusters, composed of high-performance resources, which provide a powerfull means for large scale study. Other similar programs than Snakemake are used for workflow

management such as Nextflow (Di Tommaso *et al.*, 2017), or Galaxy (noa, 2022), but Snakemake seems to be the prominent one recently (Cokelaer *et al.*, 2023).

Such workflow management programs can use other informatics tools: compartmentalized micro environment, such as conda, docker or singularity. In conda a user can create an environment, similar to micro virtual machine (VM), that can be shared to others in order to run scripts and program in the same framework. For my usage conda was at some point too slow due too a lot of different environments on my computer resources, so I switches to an alternative solution by using container in which are nested program with the required environment. This container can be load from an image spontaneously to run an analysis. Two main program are used, singularity and docker. Contrary to conda that is dependent of python, images are built at the OS level, which allow more reproducibility power and are easier to share.

Another layer can be used to appreciate all the changes that have been made in a repository or a pipeline. Indeed, eventually the bioinformatician can provide Zenodo archive with different versions, but also it can be accompanied or linked to a Git repository which is a web versioning tool. This means that each change to the codes can be traced back, commented and documented. Git allows the user to tag certain states of the Git repository, which can then be easily collected by Zenodo to be stored in an archive.

All these developed tools represent an excellent opportunity for science to be more reproducible than ever. The limits will still be to define the level of detail of the data to be provided; confidentiality clauses, which may limit sharing; and the time required to restart the analysis. But this is still an opportunity for readers to get crazy details about the data production process and the data relating to the direct figure of the paper.

Hopefully these tools will be maintained, and journals themselves will provide these services to allow the maintenance of data relating to the article they publish and for which, in one way or another, they have the responsibility in order to maintain appropriate ethics.

8.3 Perspectives

It seems clear that some minds are hard to convince, even with the growing numbers of evidences rejecting hypotheses. This has been the case for the debate over alternative splicing products, as noted previously. And it is the case for codons optimizing translation in humans, as articles often discuss its existence but are often misguided and ignore non-adaptive hypotheses. Thus, it is our/my responsibility to approach the questions with rigorous and comprehensive protocols, and to describe my observations as they are, which will potentially yield convincing arguments.

I will delve into new scientific analyses that can be conducted to assure us and the community that our findings are robust and should be considered, as new problematics arise. Also, I will try to offer new avenues for studying variation in genome architecture and its relationship with random genetic drift.

8.3.1 Elucidating alternative splicing role

One of the most debated topic on which I was working on was alternative splicing supposed to be mostly non-functional in humans. To me, the emerging field of third generation sequencing appears to be an opportunity to incorporate long RNA molecules into study, which may ultimately provide access to the full mRNA molecule (Logsdon *et al.*, 2020). Indeed, in our work, because we studied short reads that limit us in the detection of more than one intron *per* read, we made a strong assumption that alternative variants are independent from one intron to another. With such long mRNA molecules it might be interesting to examine the dependency between intron variants. Additionally, because what we care about is the *per*-gene alternative splicing rate, accessing the entire mRNA molecule could allow us to improve the estimation and not making it on the hypothesis that introns variant are independent. This technique could help us incorporating other alternative splicing events such as intron retention. In my work, the use of short reads (100 bp) has limited me to detect full-length intron retention because in humans, for example, they are larger than 1 kb. Nonetheless, I attempted to capture intron retention by measuring unspliced reads at splice sites. However, this estimate was strongly influenced by RNA sequencing protocols and noise due to pre-mRNA. With a complete mRNA molecule, one can examine intron retention in mature mRNA (*i.e.* mRNA exhibiting a poly(A) tail or splicing events).

The limitation of using third generation sequencing has been its large proportion of errors, 1 in every 10 bases (8–15%), compared to illumina short reads sequencing (1% error rate) (Morisse *et al.*, 2021). A lot of work is being invested to improve the precision of sequencing and correction programs (Luo *et al.*, 2022). Recent works obtain a quality score of Q20 (1% error rate) with some reaching Q30 (0.1% error rate). However, in our case, to quantify AS we need to map the reads to the genome, and because the mapping is robust to some sequencing errors, Q20 is already more than sufficient for our purpose. Thus, long-reads sequencing is a more than interesting opportunity.

Another study that might be interesting could be to do a meta-analysis on paper results. Indeed, papers showing no evidence of functional AS are more likely to not be published than those showing satisfactory results. Perhaps with the help of machine learning and word processing, estimating how many papers show functional AS variants could be done in a near future. Similarly, a survey of laboratory studies of AS may be relevant in determining how many studies searching for functional AS have been inconclusive. These results could provide more perspective and open more dialogue on the subject.

8.3.2 Digging in why some species do translational selection

Although translational selection is rare in metazoans, my results did not really capture the biological reason why it varies in large N_e species. First, it seems interesting to focus the study on species where we observed translational selection, in order to unravel what determines the variations in TS intensity in these species. To me, it appears that the

Diptera clade is a good candidate, as it is a well studied clade, with numerous species (*i.e.* to date I identified at least 95 species for which genomes are available; Appendix C Fig. 1A), with a wide variation in genomic GC-content across species (Appendix C Fig. 1D). Since RNA-seq samples are not available for all dipterans, I suggest using *Drosophila melanogaster* as a reference species and assigning its gene expression level to corresponding genes in other species. Indeed, the gene expression appears to be conserved between species for homologous genes (*i.e.* reciprocal blast hits; Appendix C Fig. 1B,C). With this in hands, it seems affordable to replicate our previous estimate of the population-scaled selection coefficient (S). This could reassure us that translational machinery varies depending on the genome base composition (Appendix C Fig. 1F). Furthermore, by using N_e proxies, we could observe whether N_e is the main drivers of the TS intensity or not in this clade.

Because the "drift barrier" hypothesis made prediction for genome characteristics reaching selection/drift equilibrium, it seems interesting to test whether these genomes are indeed at translational selection equilibrium. To study if there is an enrichment/diminishment of POC in a genome, one can study the number of POC to non-POC substitutions compared to non-POC to POC substitutions.

Finally, if the growth rate of a species is a parameter having an impact on translational selection, we could consider capturing its level. This could be done either qualitatively, *e.g.* using hemimetabolous, *i.e.* slow growth rate, versus holometabolous, *i.e.* rapid growth rate. Or quantitavely by estimating the relative growth rate (RGR), which is the rate of growth *per* unit time relative to the size.

Also, if our prediction are correct, and that N_e explain some of the variations of TS, it should be of interest to study species with less intense genetic drift, maybe outside of the metazoans range, looking at unicellular eukaryotes for example (Lynch *et al.*, 2023).

8.3.3 Estimating N_e

Studying the impact of random genetic drift on genome characteristics is challenging, and the data I used were not a perfect fit for this study, which could undermine the confidence in the findings. For me, one of the most still debated knowledge is the measure of N_e, *i.e.* the genetic drift intensity, which often seems abstract (Waples, 2022). This sometimes complicates the interpretation and the messages of the papers studying the effect of N_e on genomes evolution. Indeed, there are still many assumptions regarding the measurement of N_e.

In my study I used four indirect proxies, that are far from perfect. Notably, the three life history traits (LHT) are proxies of the census size, which in small N populations, such as mammals is expected to be correlated with N_e. However, if other parameters change (*e.g.* the reproductive mode or the sex ratio), then we don't know is we can predict N_e variations based on the life history traits (*i.e.* body mass, body length, longevity). Investigating how changes in these parameters affect N_e and LHT could clarify how to use this proxies.

Also, the dN/dS is expected to be related to $4N_e s$, but it is based on some assumptions. The first one is that synonymous codons are neutral but we know that they are not. Thus, we may underestimate the N_e with dN/dS proxy in species where synonymous codons are selected. Nevertheless, it seems that in metazoans synonymous codons are mostly driven by non-adaptive processes due to their strong relation with GC-content. However, dN/dS is also used to detect positive selection, thus one assumption is that it is sufficiently rare to posit that non-synonymous substitutions are mostly deleterious. In GTDrift we discussed about the limits regarding the impact of polymorphism and saturation on this estimator. With the arising amount of genetic data it will be soon available if not already, to have polymorphism in populations ($\pi_s = 4N_e\ \mu$), which with the specific mutation rate, will give direct measure of short-term N_e (Lynch *et al.*, 2023). At this point it may be relevant to understand how dN/dS fluctuate with π_s, and how the polymorphism affects part of the dN/dS that should be estimated only on substitutions.

Soon the project NeGA should produce data answering the limitation of the imperfect estimate of N_e to investigate its impact on genome architecture. Indeed, the project will bring together dozens of pairs of closely related species with a decrease in effective population size. Five biological models will be proposed, including Asellidae isopods, passerine birds, Drosophila, swallowtail butterflies and ants. In each pair there has been a shift in the ecological niche which is followed by a decrease in effective population size, *e.g.* in Asellidae isopods there are isopods living in surface and ground water. The subterranean species is expected to have a reduced effective population size (Lefébure *et al.*, 2017). For each pair it will be interesting to study the evolution of the AS rate. If our hypothesis is correct, we should expect for each pair an increase in the AS rate in species with a reduction in effective population size. The question remains if the equilibrium between drift and selection for alternative splicing rate/error will be reached in a large time scale or not.

Another parameter to predict N_e could involve estimating population density using collaborative databases such as Global Biodiversity Information Facility (GBIF) to collect public observations. I tried to initiate a project aimed at gathering such data, but they appeared to be very heterogeneous and mostly focused on birds. Moreover, even within the bird dataset it seems that this may be biased towards public knowledge of birds. Thanks to machine learning, it might be possible to collect images and automatically annotate them, thereby offering more unbiased data.

Overall, conducting an integrative study utilizing the maximum number of available N_e proxies could help determine the reliability of each and the conditions under which we can have confidence in them. This presents an exciting opportunity to predict the absolute effective population size (*i.e.* the number of individual in a Hardy-Weinberg population that would yield equivalent patterns of random fluctuations at neutral sites).

8.3.4 How N_e impact genome architecture

During my thesis, I investigated various genomic characteristics. However, there is potential for further exploration in the near future utilizing the resources offered by GTDrift

(Bénitière *et al.*, 2024). A comparative analysis could shed light on the impact of N_e on variation in size and number of the major isoform introns. I rapidly observed from GTDrift web interface that in hymenopterans the median intron length seems impacted by N_e variations (Appendix C Fig. 2B), but on large scale study among vertebrates the intron length variations does not seem to be affected by N_e and is conserved within clades (*i.e.* 850 bp for birds, 300 bp for fishes and 1,200 bp for mammals; Appendix C Fig. 2A). Additionally, studying genome size and other dominant genome architecture parameters would be valuable avenues of investigation. Although it has been shown not to be affected by N_e at the metazoan scale (Whitney and Garland, 2010; Roddy *et al.*, 2021; Marino *et al.*, 2024), it appears that *Hymenoptera* genome size might be explained by N_e variations (Appendix C Fig. 2D).

Also, I would like to pursue the investigation of drift impact on genomes architecture in embryophytes, as some species are available in GTDrift data resource. But I've had trouble getting N_e proxies for these species. The ratio dN/dS can provide one but as said before, it's reassuring to have several estimate of N_e, and for plants, what should be use is not well established. The periods of flowering and survival of plants could be a line of inquiry: *i.e.* annual, multi-annual, biennial plants, *etc.* This will then allows us to re-investigate our previous observations in a new clades. Nevertheless, new problematics could be faced because AS is very different in plants and metazoans, *e.g.* intron retention is enriched in plants (up to 56%; Reddy *et al.* (2012, 2013)) while exon skipping is most prominent in humans (58%).

8.3.5 Environmental cost of research

Now more than ever, it is imperative to be aware of the environmental impact of our research. During my work in the laboratory, significant computing resources were used, particularly for the alignment of RNA-seq samples on genomes, which could span a week using 16 cores. Additionally, resource-intensive analyses were conducted on the clusters, including phylogenetic tree inference, dN/dS estimation, gene expression profiling, systematic analyses across multiple species *etc.*

In total, my research represented 3,189,232 hours of CPU usage on the computing facilities of CC LBBE/PRABI and the Core Cluster of the Institut Français de Bioinformatique (IFB). By estimating at 1,260 kgs the construction and transport of a 16-core server which has a 7-year lifespan (source: Eco-info `https://ferme.yeswiki.net/Empreinte/?PagePrincipale`), and by considering that each core consumes 23.9 W per hour, with a corresponding carbon emission of 79 gCO2/kWh (source: Agence de la transition écologique), the carbon footprint is approximately 3 gCO2/h (1.1 gCO2/h from transport and construction, and 1.9 gCO2/h from electricity). My research was therefore responsible for the emission of 10 tons of CO2, equivalent to 3 Paris - New York round-trip, or 19 Paris - Nice round-trip by plane (based on estimates by Ayoun (2021)).

This highlights the importance of meticulously building protocols before running lots of unnecessary calculations, optimizing scripts, and checking for errors. Addition-

ally, sharing data allows others to use it without the need for redundant analyses, and publishing negative results helps identify ineffective methodologies. Above all, it seems important to find a balance between robustness of our analyzes and their energy cost. In the field of bioinformatics, given the vast availability of genomic data and the ease of running calculations, the temptation to conduct repetitive analyses on large quantities of data is omnipresent, one click away. Indeed, at some point what is working is not our brain anymore, but, day and night, our computers.

This was one of the main motivations for me to share and publish my data with as much information as possible so that others could replicate the analysis and understand exactly what was done, without having to re-run heavy analyses. In the near future, perhaps research funding agencies will require that a project's carbon emissions be estimated as is done for budgeting.

8.3.6 Accessibility

Lastly, I would like to present in a few words my point of view, which is obviously questionable, about the accessibility not only to the data pertaining to analysis, but the science itself. On this matter, trying to develop a model/protocol that is as simple as possible to address a problem seems to have a better chance of convincing a wide audience. One way to find a compromise on qualitative, ethical science, accessible through both reproducibility and scientific knowledge, could be to focus science more on methods than on results. Meaning defining a primary question; addressing this question with a peer-validated protocol to produce and interpret results. This could avoid extreme observation biases that can be encountered in bioinformatics due to the possibility of changing the protocol and rerunning entire pipeline within a few days in unintentional search for satisfactory results. These published methods/articles could potentially show negative and positive results, be shorter and therefore more accessible. As I observed during my thesis, it is typically from this perspective that alternative splicing seems to be put forward as being mainly functional. Indeed, people mainly publish positive results which overshadow the negative results that many might observe.

Adopting this 'method' approach could have the privilege of scientifically recognizing methods that did not work, not making the same mistakes, ultimately reducing unnecessary carbon emissions, and perhaps being published and revised more quickly to ultimately be more satisfying.

Appendices

A

Supplementary data and figures Chapter 6

Random genetic drift sets an upper limit on mRNA splicing accuracy in metazoans

Figures

Supplementary Table 1: Description of the main features of the samples analyzed in this study.

	Clade	Number of RNA-seq samples	Sequencing depth (*per*-base read)[a]	Number of annotated introns	Number of analyzable introns[b]	Average number of introns *per* BUSCO gene	Fraction of major-isoform introns alternatively spliced[c]	Average AS rate among BUSCO introns	Fraction of rare SVs[d]
Vertebrates									
Callorhinchus milii	Chondrichthyes	11	1068	7700	7467	8.0	0.491	1.47 %	0.831
Gallus gallus	Aves	217	9657	8741	8621	8.4	0.854	1.59 %	0.958
Crocodylus porosus	Crocodylia	12	1819	7867	7668	8.5	0.817	3.02 %	0.908
Monodelphis domestica	Mammalia	269	11371	8538	8407	8.5	0.915	1.91 %	0.957
Heterocephalus glaber	Mammalia	54	2072	9409	9324	8.6	0.803	2.69 %	0.914
Macaca mulatta	Mammalia	177	5571	9328	9261	8.6	0.908	2.84 %	0.948
Oryctolagus cuniculus	Mammalia	338	15503	8036	7885	8.4	0.950	1.97 %	0.969
Rattus norvegicus	Mammalia	362	16611	8469	8196	8.5	0.953	1.89 %	0.965
Mus musculus	Mammalia	317	12245	9327	9080	8.4	0.937	1.87 %	0.958
Bos taurus	Mammalia	26	710	9046	8926	8.5	0.511	1.63 %	0.856
Loxodonta africana	Mammalia	23	3667	9000	8652	8.3	0.896	3.55 %	0.938
Sus scrofa	Mammalia	55	910	8982	8798	8.5	0.644	1.95 %	0.886
Canis lupus	Mammalia	5	348	9279	8628	8.2	0.436	2.18 %	0.764
Homo sapiens	Mammalia	313	10269	11122	10981	8.4	0.957	3.38 %	0.949
Equus caballus	Mammalia	19	998	9190	9072	8.5	0.658	2.16 %	0.884
Insects									
Bombyx mori	Lepidoptera	14	459	5001	4681	5.3	0.393	1.12 %	0.835
Athalia rosae	Hymenoptera	6	359	4772	4701	4.8	0.348	1.6 %	0.782
Cephus cinctus	Hymenoptera	17	2566	5035	5016	4.7	0.744	2.4 %	0.907
Orussus abietinus	Hymenoptera	2	197	4801	4664	4.7	0.370	2.03 %	0.763
Nasonia vitripennis	Hymenoptera	114	4871	4273	4158	4.5	0.648	1.21 %	0.913
Trichogramma pretiosum	Hymenoptera	4	350	3794	3734	4.4	0.268	0.98 %	0.782
Harpegnathos saltator	Hymenoptera	166	1888	4745	4711	4.7	0.565	2.02 %	0.886
Linepithema humile	Hymenoptera	23	1476	4726	4615	4.8	0.570	1.45 %	0.882
Camponotus floridanus	Hymenoptera	37	449	4596	4546	4.7	0.358	1.52 %	0.761
Pogonomyrmex barbatus	Hymenoptera	39	1388	4678	4440	4.5	0.579	1.91 %	0.866
Polistes canadensis	Hymenoptera	14	440	4665	4562	4.8	0.424	1.88 %	0.834
Polistes dominula	Hymenoptera	12	218	4698	4161	4.3	0.180	1.63 %	0.624
Solenopsis invicta	Hymenoptera	23	436	4516	4394	4.6	0.430	1.71 %	0.807
Acromyrmex echinatior	Hymenoptera	42	1470	4716	4638	4.7	0.529	2.15 %	0.835
Megachile rotundata	Hymenoptera	108	3400	5120	5086	4.8	0.898	3.81 %	0.927
Apis mellifera	Hymenoptera	40	1777	4939	4897	4.9	0.673	2.3 %	0.892
Apis florea	Hymenoptera	4	503	4881	4332	4.4	0.318	1.85 %	0.711
Apis cerana	Hymenoptera	12	1401	4508	4439	4.6	0.578	2.36 %	0.839
Bombus terrestris	Hymenoptera	33	2648	4857	4683	4.7	0.763	2.33 %	0.922
Acyrthosiphon pisum	Hemiptera	35	3163	4918	4844	6.0	0.709	1.09 %	0.933
Cimex lectularius	Hemiptera	10	462	5640	5588	6.3	0.431	1.61 %	0.838
Halyomorpha halys	Hemiptera	6	1460	5715	5676	6.5	0.591	1.73 %	0.885
Aedes aegypti	Diptera	27	2469	2369	2290	2.6	0.514	1.35 %	0.870
Drosophila grimshawi	Diptera	30	256	2190	2032	2.7	0.168	0.8 %	0.726
Drosophila pseudoobscura	Diptera	32	3628	2312	2244	2.6	0.433	1.32 %	0.871
Drosophila melanogaster	Diptera	129	4542	2414	2390	2.7	0.551	1.22 %	0.909
Drosophila suzukii	Diptera	23	1979	2187	2052	2.6	0.287	1.17 %	0.810
Ceratitis capitata	Diptera	29	1168	3067	3015	3.3	0.418	1.45 %	0.860
Lucilia cuprina	Diptera	23	2446	2566	2405	2.8	0.268	0.85 %	0.823
Musca domestica	Diptera	12	1056	2545	2401	2.9	0.254	0.98 %	0.795
Onthophagus taurus	Coleoptera	53	644	2836	2753	3.2	0.377	1.34 %	0.810
Tribolium castaneum	Coleoptera	14	2618	3333	3225	3.6	0.556	1.15 %	0.881
Dendroctonus ponderosae	Coleoptera	30	2262	4370	4269	4.9	0.505	1.26 %	0.882
Anoplophora glabripennis	Coleoptera	20	325	3764	3567	4.1	0.299	1.13 %	0.781
Leptinotarsa decemlineata	Coleoptera	21	2071	3372	3132	3.8	0.512	1.21 %	0.883
Blattella germanica	Blattodea	30	943	4911	4454	5.4	0.423	1.26 %	0.827
Cryptotermes secundus	Blattodea	11	481	6471	6391	6.4	0.573	2.32 %	0.832
Zootermopsis nevadensis	Blattodea	53	3944	6727	6613	6.4	0.802	2.36 %	0.927

[a] Median *per*-base read coverage computed on BUSCO gene exons

[b] Number of analyzable introns (*i.e.* with $N_a + N_a \geq 10$) among BUSCO genes

[c] Proportion of major-isoform introns for which alternative splicing has been detected (*i.e.* with $N_a > 0$) among BUSCO genes

[d] Fraction of rare spliced variants introns (*i.e.* with MIRA $\leq 5\%$) among all protein-coding genes

Table A.1

Supplementary Table 2: Longevity and body lenth across the 53 metazoans studied.

	Clade	Longevity (Days)	Body length (cm)
Vertebrates			
Callorhinchus milii	Chondrichthyes	2190	120.00
Gallus gallus	Aves	10950	70.00
Crocodylus porosus	Crocodylia	20805	600.00
Homo sapiens	Mammalia	36500	175.00
Loxodonta africana	Mammalia	23725	400.00
Equus caballus	Mammalia	20805	280.00
Macaca mulatta	Mammalia	14600	64.00
Heterocephalus glaber	Mammalia	10950	16.50
Sus scrofa	Mammalia	9855	240.00
Canis lupus	Mammalia	7519	117.00
Bos taurus	Mammalia	7300	245.00
Oryctolagus cuniculus	Mammalia	3285	50.00
Monodelphis domestica	Mammalia	1862	20.00
Mus musculus	Mammalia	1460	9.50
Rattus norvegicus	Mammalia	1387	40.00
Insects			
Bombyx mori	Lepidoptera	50	1.90
Pogonomyrmex barbatus	Hymenoptera	10220	1.10
Acromyrmex echinatior	Hymenoptera	5475	1.40
Camponotus floridanus	Hymenoptera	3650	1.90
Solenopsis invicta	Hymenoptera	2482	0.70
Apis mellifera	Hymenoptera	1095	2.00
Apis florea	Hymenoptera	1095	2.00
Apis cerana	Hymenoptera	1095	2.00
Harpegnathos saltator	Hymenoptera	653	1.70
Polistes canadensis	Hymenoptera	506	2.00
Polistes dominula	Hymenoptera	506	2.00
Linepithema humile	Hymenoptera	365	0.50
Bombus terrestris	Hymenoptera	150	2.50
Megachile rotundata	Hymenoptera	56	1.90
Nasonia vitripennis	Hymenoptera	25	0.30
Athalia rosae	Hymenoptera	12	0.73
Trichogramma pretiosum	Hymenoptera	10	0.04
Cephus cinctus	Hymenoptera	7	0.86
Orussus abietinus	Hymenoptera	7	1.00
Cimex lectularius	Hemiptera	572	0.50
Halyomorpha halys	Hemiptera	112	1.44
Acyrthosiphon pisum	Hemiptera	30	0.25
Drosophila pseudoobscura	Diptera	90	0.20
Musca domestica	Diptera	60	0.70
Drosophila grimshawi	Diptera	50	0.50
Ceratitis capitata	Diptera	50	0.50
Drosophila suzukii	Diptera	38	0.33
Drosophila melanogaster	Diptera	36	0.30
Lucilia cuprina	Diptera	21	0.80
Aedes aegypti	Diptera	14	0.38
Leptinotarsa decemlineata	Coleoptera	365	1.00
Tribolium castaneum	Coleoptera	170	0.50
Onthophagus taurus	Coleoptera	160	1.00
Anoplophora glabripennis	Coleoptera	66	3.50
Dendroctonus ponderosae	Coleoptera	30	0.75
Cryptotermes secundus	Blattodea	4745	0.60
Zootermopsis nevadensis	Blattodea	2300	1.00
Blattella germanica	Blattodea	200	1.59

* The sources from which the lifespan and the body length information was taken are listed in Data9supp.pdf in the Zenodo data repository (see Data and code availability).

Table A.2

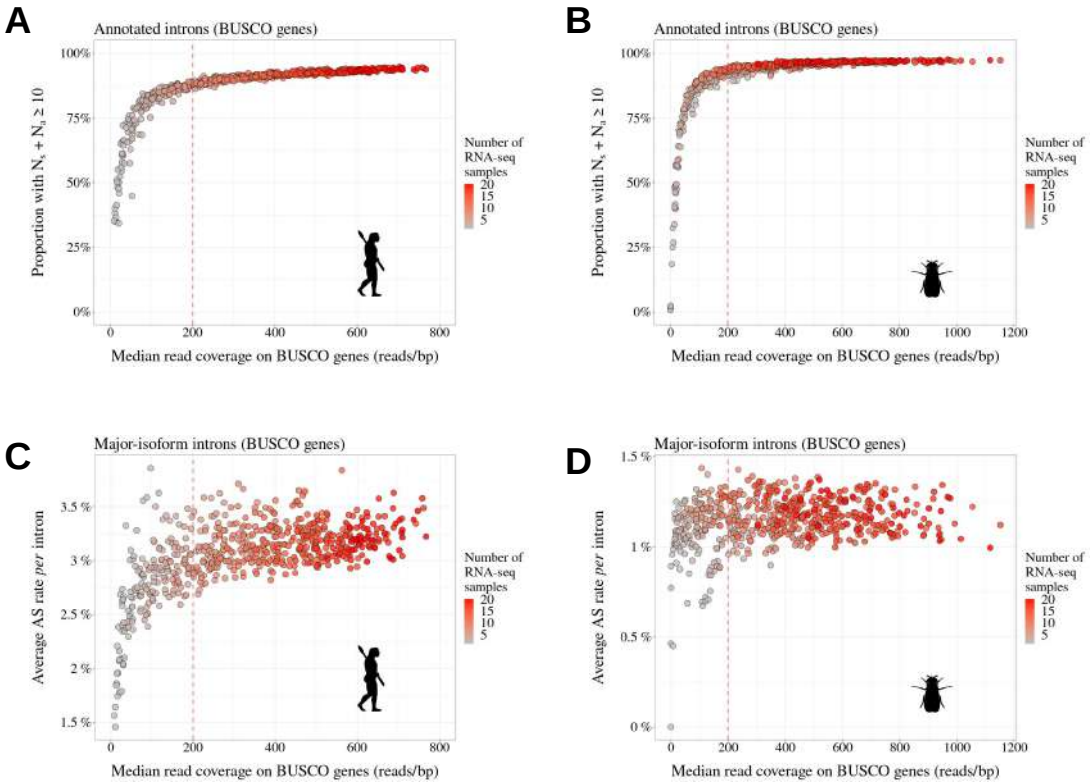


Figure A.1: Transcriptome sequencing depth affects intron detection power and AS rate estimates. *To assess the impact of sequencing depth on AS detection, we conducted a pilot analysis with two species (**A,C**: Homo sapiens and **B,D**: Drosophila melanogaster) for which hundreds of RNA-seq samples are available (Supplementary Tab. 1; refer to Data10-supp.tab in the Zenodo data repository). We randomly drew 1 to 20 RNA-seq samples and, for each draw, we computed the median read coverage across BUSCO gene exons (to get a measure of transcriptome sequencing depth that is comparable across species). We also computed for each draw the average AS rate and the fraction of introns supported by at least 10 RNA-seq reads, out of all introns annotated for BUSCO genes (Materials & Methods). We repeated this procedure 30 times. As expected, the fraction of BUSCO introns that are supported by at least 10 reads (i.e.* $N_s + N_a \geq 10$*) increases with sequencing depth (**A,B**). More importantly, we observed that when sequencing depth is limited, the mean AS rate of BUSCO introns is very variable across draws (**C,D**). However, AS rate estimates converge when sequencing depth exceeds 200. We therefore kept for further analysis those species for which the median read coverage across exonic regions of BUSCO genes was above this threshold.*

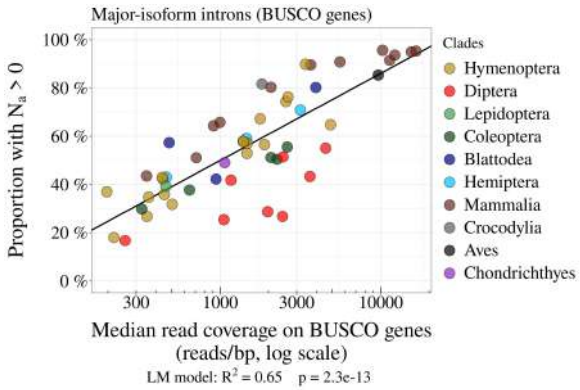


Figure A.2: The power to detect AS events is positively correlated with transcriptome sequencing depth. *Relationship between the proportion of major-isoform introns that have at least one read corresponding to splice variants (i.e.* $N_a > 0$*; see Fig. 2), and the median per-base read coverage computed on BUSCO gene exons, across metazoans. Each dot represents one species, colored by taxonomic clade.*

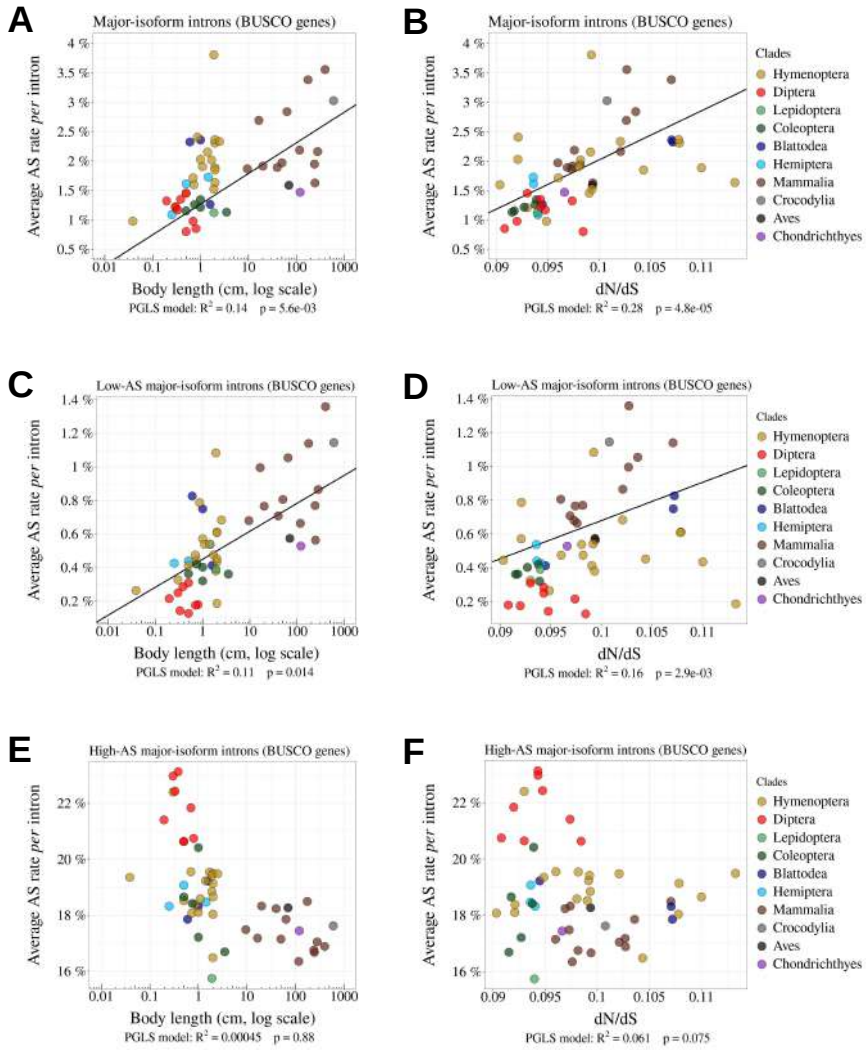


Figure A.3: Relationship between AS rates and other N_e proxies. ***A,B****: Correlation between the average AS rate per intron and the body length of each species (cm, log scale) (****A****) or the dN/dS ratio on terminal branches of the phylogenetic tree (****B****).* ***C,D,E,F****: Relationship between the average AS rate per intron and the body length (cm, log scale) (****C,E****) or the dN/dS ratio (****D,F****).* ***C,D****: Low-AS major-isoform introns (i.e. major-isoform introns that do not have any abundant SV).* ***E,F****: High-AS major-isoform introns (i.e. major-isoform introns having at least one abundant SV). Only BUSCO genes were used in the analysis.*

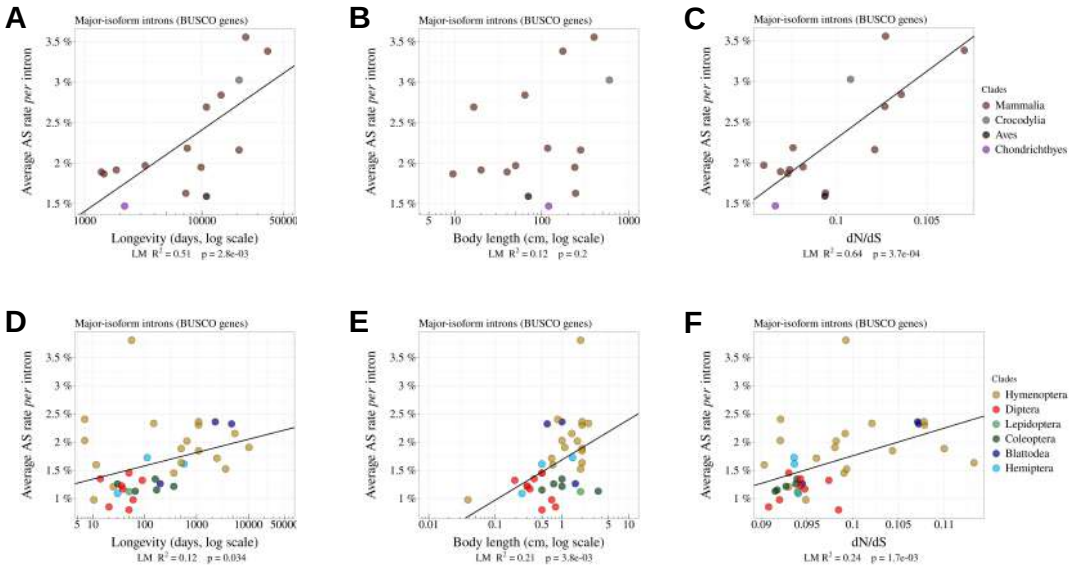


Figure A.4: The rate of alternative splicing correlates with life history traits in both vertebrates and insects. *Correlation between the average AS rate per intron and longevity of each species (days, log scale) (**A,B**), body length (cm, log scale) (**B,E**), or the dN/dS ratio on terminal branches of the phylogenetic tree (**C,F**). In vertebrates (**A,B,C**) and insects **C,D,E**). Only the BUSCO genes were included in the analysis.*

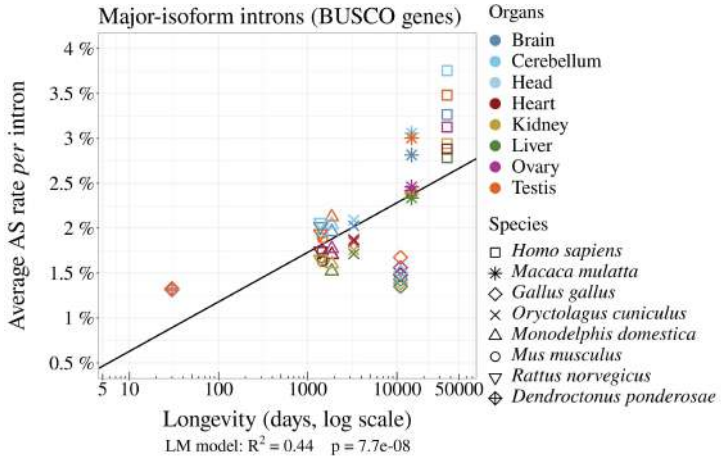


Figure A.5: The variation in AS rates between species is not explained by organ differences. *Variation in average AS rate across seven organs (brain, cerebellum, heart, liver, kidney, testis, and ovary) among seven vertebrate species (RNA-seq data from Cardoso-Moreira et al. (2019)) and across three organs (ovary, testis, and head) for one insect (Dendroctonus ponderosae, Coleoptera). AS rates were computed for the major-isoform introns from BUSCO genes (Materials & Methods).*

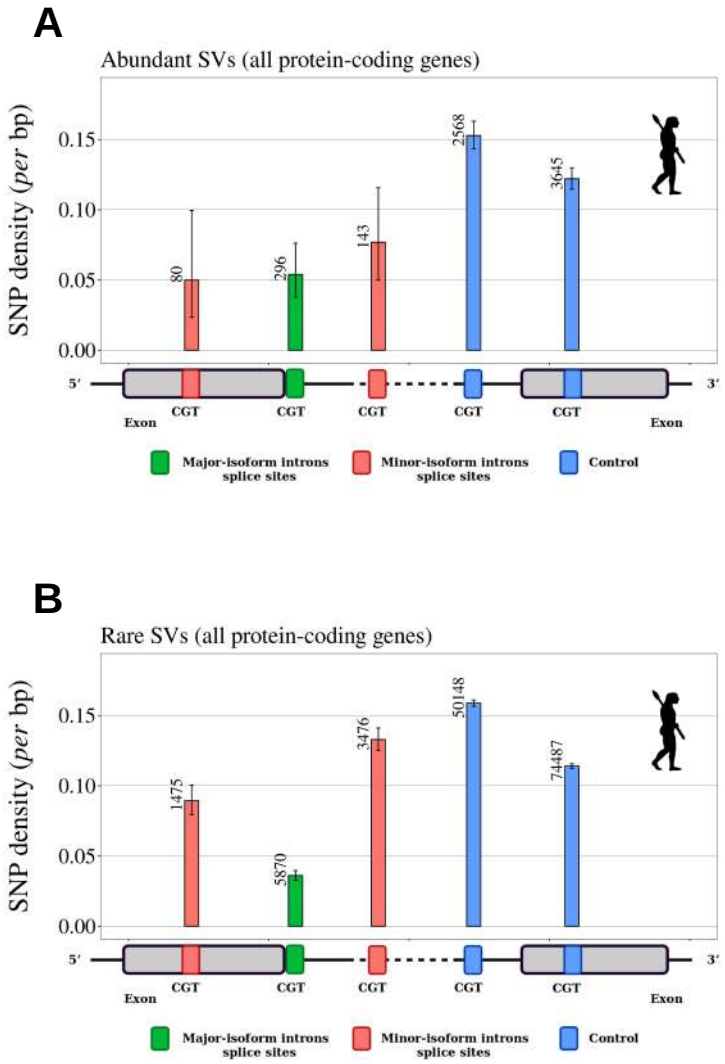


Figure A.6: SNP density in human splice signals, for dinucleotides affected by CpG hypermutability. *Density of SNPs on splice signals for major-isoform introns and for SVs that have their minor splice site within the adjacent exon or in the major-isoform intron. The number of introns studied is shown at the top of each bar.* ***A,B****: SNP data from the human 1000 Genomes project (Auton et al., 2015). We included only dinucleotides affected by CpG hypermutability (Materials & Methods). Error bars represent the 95% confidence interval of the proportion of polymorphic sites (proportion test).* ***A****: Abundant SVs (MIRA > 5%).* ***B****: Rare SVs (MIRA ≤ 5%). green: major splice sites; red: minor splice sites; blue: control dinucleotides.*

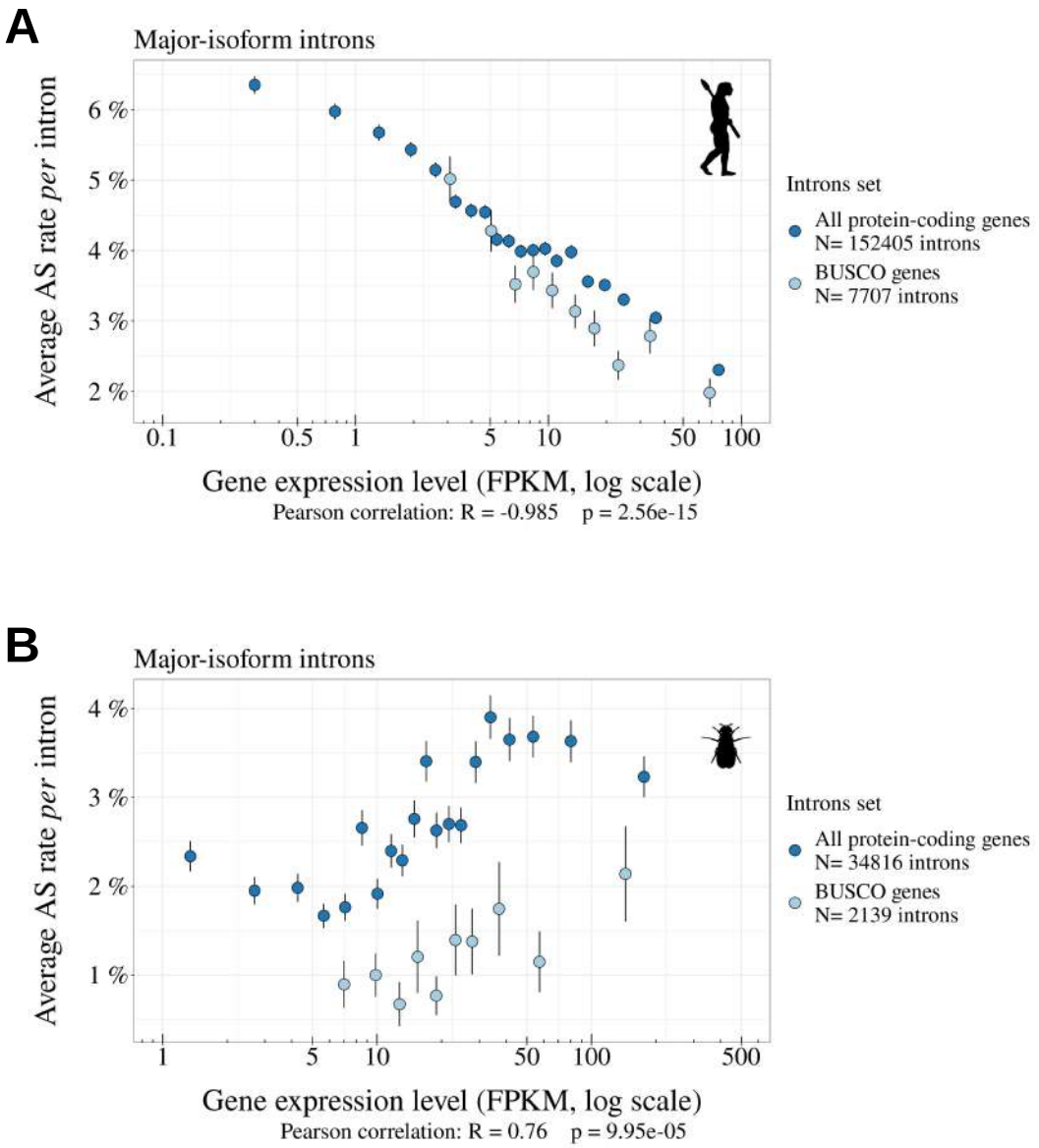


Figure A.7: Correlations between gene expression levels and AS rates differ among species. ***A,B****: Relationship between the average AS rate of major-isoform introns (with* $N_s + N_a \geq 100$*, see Fig. 2) and the expression levels of the corresponding genes (FPKM, log scale). We divided major-isoform introns into 5% bins according to the expression level of the corresponding genes and computed for each bin the average AS rate and the median expression level. Error bars represent the standard error of the mean.* ***A****: Homo sapiens,* ***B****: Drosophila melanogaster. This analysis was performed on all protein-coding genes (blue) and BUSCO genes (light blue). Pearson correlation presented here was computed on protein-coding genes.*

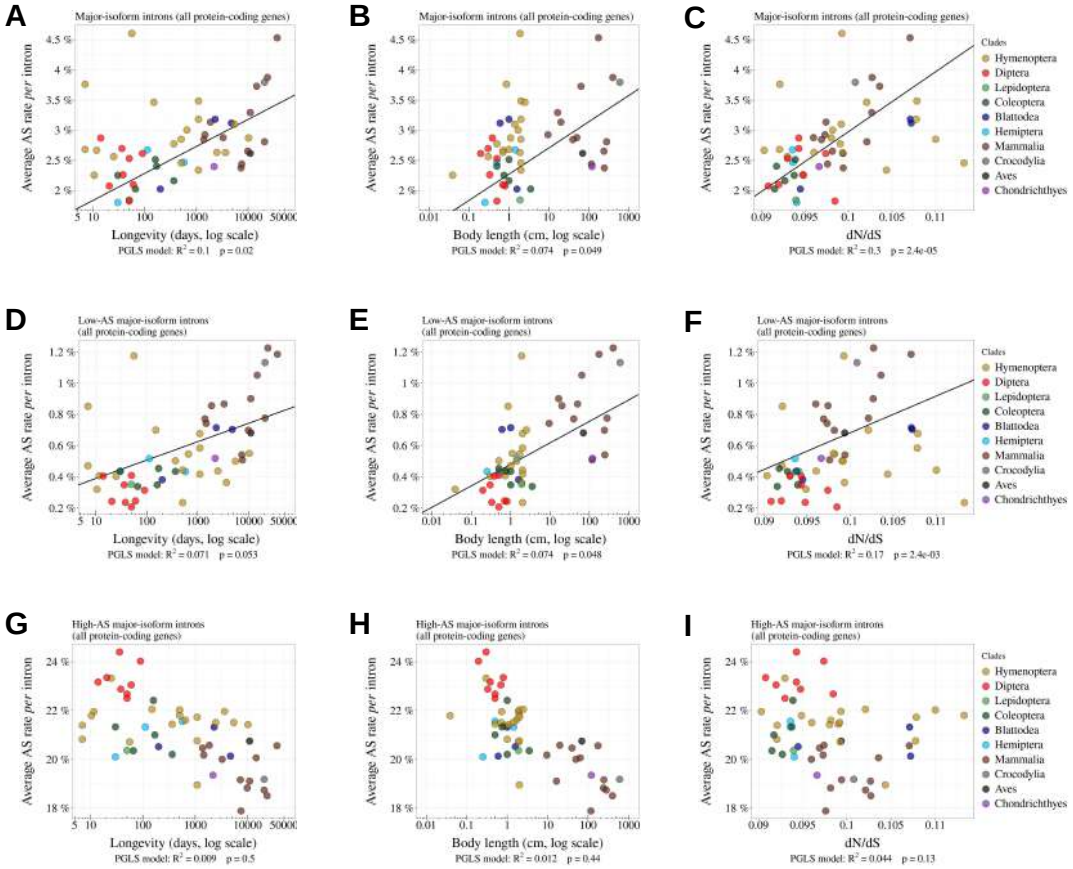


Figure A.8: Relationship between AS rates and N_e proxies, for all major-isoform introns, low-AS major-isoform introns (i.e. major-isoform introns that do not have any abundant spliced variants) and high-AS major-isoform introns (i.e. major-isoform introns having at least one abundant spliced variants). *Relationship between the average AS rate of all major-isoform introns (**A,B,C**) or low-AS major-isoform introns (**D,E,F**) or high-AS major-isoform introns (**G,H,I**) and longevity (days, log scale) (**A,D,G**) or body length (cm, log scale) (**B,E,H**) or the dN/dS ratio (**C,F,I**).*

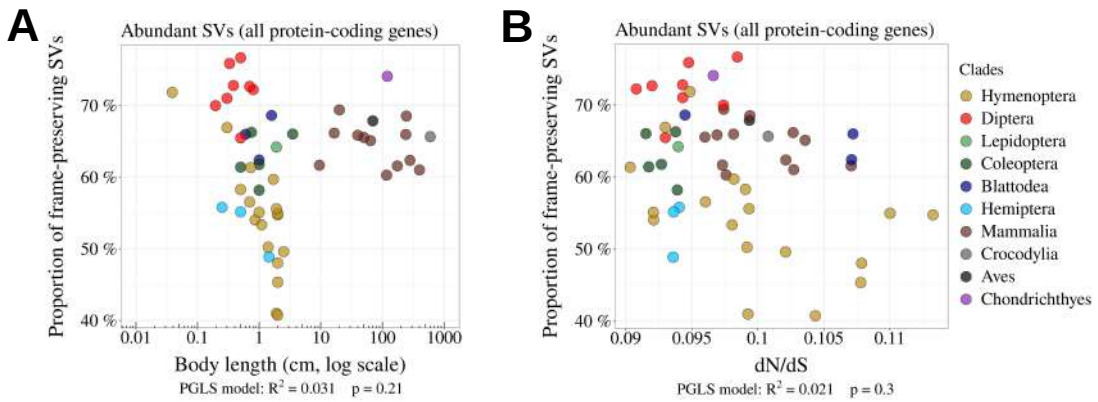


Figure A.9: Relationship between the proportion of frame-preserving SVs and N_e proxies. ***A,B****: Relationship between the proportion of frame-preserving SVs among abundant SVs, and the body length (cm, log scale) of the organism (**A**) or the dN/dS ratio (**B**). Each dot represents one species. All protein-coding genes were used in the analysis.*

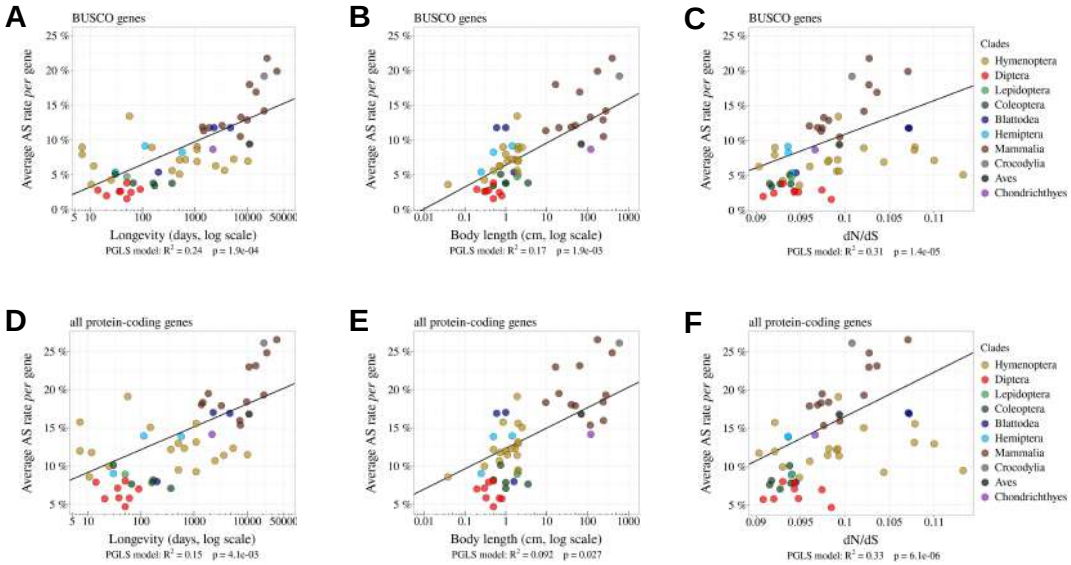


Figure A.10: The per-gene AS rate is negatively correlated with N_e. *Relationship between per-gene average AS rates and N_e proxies. We use as inverse N_e proxies the longevity (days, log scale) (**A,D**) or the body length (cm, log scale) (**B,E**) or the dN/dS ratio (**C,F**). The analysis was done on BUSCO genes (**A,B,C**) and on all protein-coding genes (**D,E,F**).*

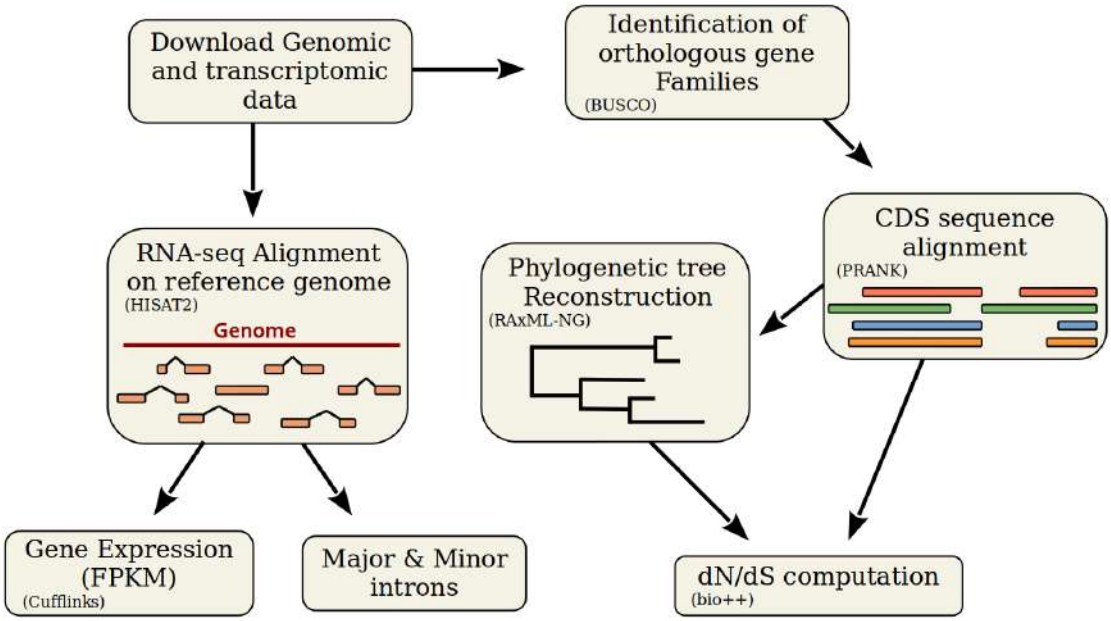


Figure A.11: Description of the bioinformatic analyses pipeline. *First, we retrieved genomic sequences and annotations from the NCBI Genomes database. We aligned RNA-seq reads with HISAT2 on the corresponding reference genomes, to analyze various variables (see Fig. 2), to compute the AS rate, and to estimate gene expression using Cufflinks. To compute dN/dS, we first identified BUSCO genes with BUSCOv3 and aligned their coding sequences (CDS) using PRANK (codon model). We reconstructed a phylogenetic tree using RAxML-NG with 461 multiple alignments. Using bio++, we estimated dN/dS along the phylogenetic tree on concatenated alignments.*

B

Supplementary data and figures Chapter 7

Why is selection for translationally optimal codons so scarce in metazoans?
Variation in fitness effects and drift intensity

Figures

Estimating the strength of selection on synonymous codon usage using population genetics.

The frequency of optimal codons (FOP) reflects the balance between the optimal to non-optimal codons synonymous substitution rate (K_{on}) and the non-optimal to optimal codons synonymous substitution rate (K_{no}):

$$\text{Non-optimal codon} \underset{K_{on}}{\overset{K_{no}}{\rightleftharpoons}} \text{Optimal codon}$$

Substitution rates depend on the corresponding mutation rates (μ_{no}, μ_{on}) and fixation probabilities (P_{no}, P_{on}): $K_{no} = 2N_e\mu_{no}P_{no}$ and $K_{on} = 2N_e\mu_{on}P_{on}$, where N_e is the effective population size.

Fixation probabilities are given by:

$$P_{no} = \frac{1 - e^{-4N_e f_0 s}}{1 - e^{-4N_e s}} = \frac{1 - e^{-2s}}{1 - e^{-4N_e s}} \overset{s \to 0}{=} \frac{2s}{1 - e^{-4N_e s}} \text{ similarly } P_{on} \overset{s \to 0}{=} \frac{-2s}{1 - e^{4N_e s}}$$

Where s is the selection coefficient in favor of optimal codons and f_0 the allele frequency of a new arrival mutation ($f_0 = 1/2N_e$).

At equilibrium, the frequency of optimal codons is given by: $FOP = \frac{K_{no}}{K_{on}+K_{no}}$ which can be written as:

$$FOP = \frac{2N_e\mu_{no}P_{no}}{2N_e\mu_{no}P_{no} + 2N_e\mu_{on}P_{on}} = \frac{\mu_{no}P_{no}}{\mu_{no}P_{no} + \mu_{on}P_{on}} = \frac{\frac{\mu_{no}}{\mu_{on}}\frac{2s}{1-e^{-4N_e s}}}{\frac{\mu_{no}}{\mu_{on}}\frac{2s}{1-e^{-4N_e s}} + \frac{-2s}{1-e^{4N_e s}}}$$

Let us note lambda, the ratio of mutation rates: $\lambda = \frac{\mu_{no}}{\mu_{on}}$

$$FOP = \frac{\lambda}{\lambda + \frac{-(1-e^{-4N_e s})}{1-e^{4N_e s}}}$$

$$\frac{1}{FOP} = 1 + \frac{1}{\lambda} \times \frac{-(1 - e^{-4N_e s})}{1 - e^{4N_e s}}$$

$$\frac{1}{FOP} - 1 = \frac{-(1 - e^{-4N_e s})}{1 - e^{4N_e s}} \times \frac{1}{\lambda}$$

$$\frac{1 - FOP}{FOP} \times \lambda = \frac{-(1 - e^{-4N_e s})}{1 - e^{4N_e s}}$$

$$\frac{FOP}{1 - FOP} \times \frac{1}{\lambda} = \frac{1 - e^{4N_e s}}{-(1 - e^{-4N_e s})} \tag{B.1}$$

With the following simplification:

$$\frac{1-e^{4N_e s}}{-(1-e^{-4N_e s})} = \frac{1-e^{4N_e s}}{-(1-\frac{1}{e^{4N_e s}})} = \frac{e^{4N_e s}\times(1-e^{4N_e s})}{1-e^{4N_e s}} = e^{4N_e s}$$

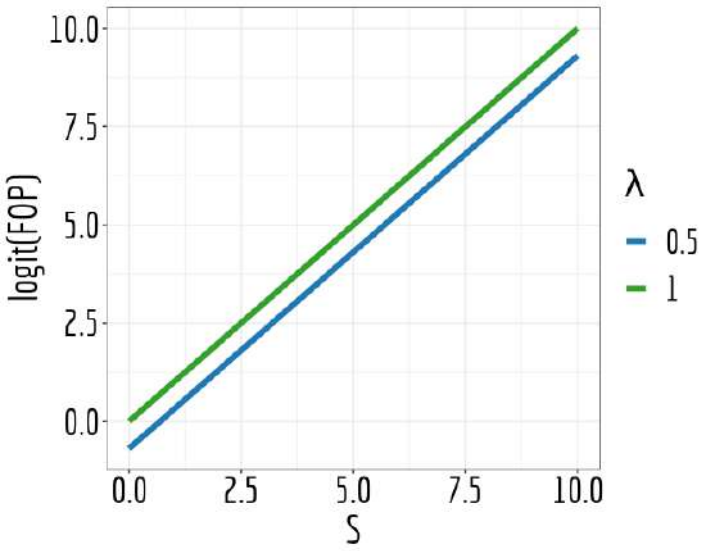
$$(1) \rightarrow \frac{FOP}{1-FOP}\times\frac{1}{\lambda} = e^{4N_e s}$$

Thus, the population-scaled selection coefficient ($S = 4N_e s$) is given by:

$$S = log(\frac{FOP}{1-FOP}) - log(\lambda) = logit(FOP) - log(\lambda)$$

Hence, we expect a linear correlation between logit(FOP) and S:

$$logit(FOP) = S + log(\lambda)$$

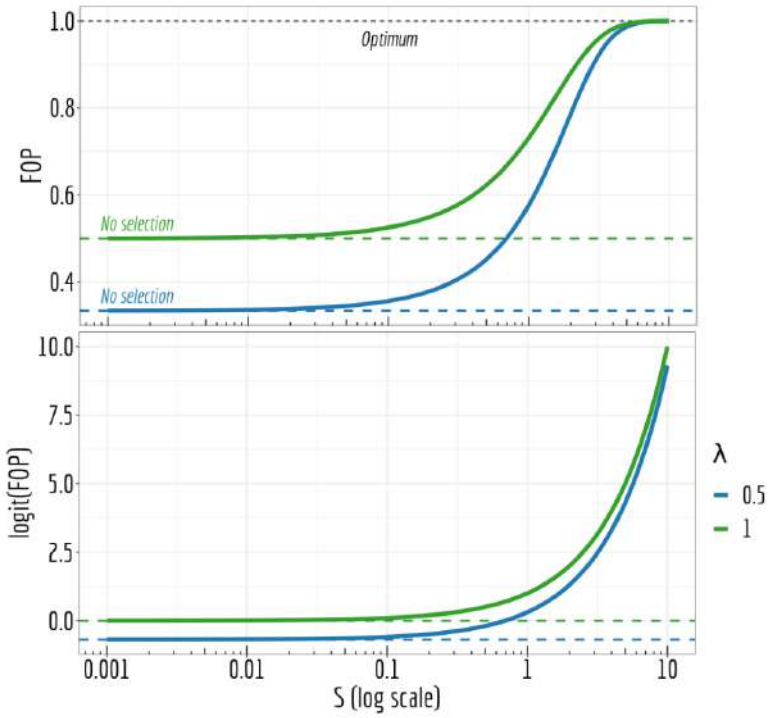


If for weakly-expressed genes there is no selection, implied by the non-variation of FOP with gene expression, $S^{lx} \approx 0$:

$$logit(FOP^{lx}) = 0 + log(\lambda)$$

$$logit(FOP^{hx}) = S^{hx} + log(\lambda)$$

$$S^{hx} = logit(FOP^{hx}) - logit(FOP^{lx})$$



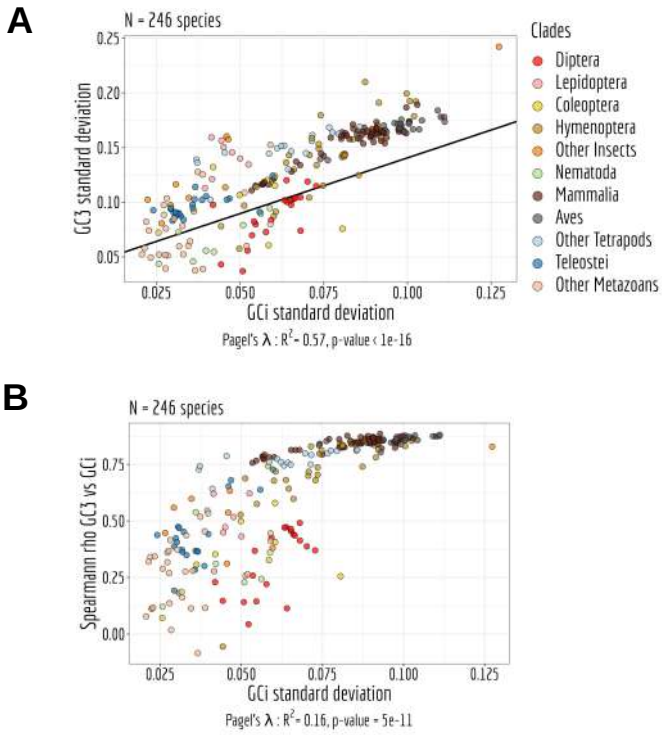


Figure B.1: Intra-species codon usage variations. ***A****: Relationship between the standard deviation of the per gene GC at the third position (GC3) and the GC in introns (GCi).* ***B****: Relationship between the Spearman coefficient (rho) reflecting the correlation between GC3 of genes and GC content in introns within a specific species, and the standard deviation of the per gene GC in introns.*

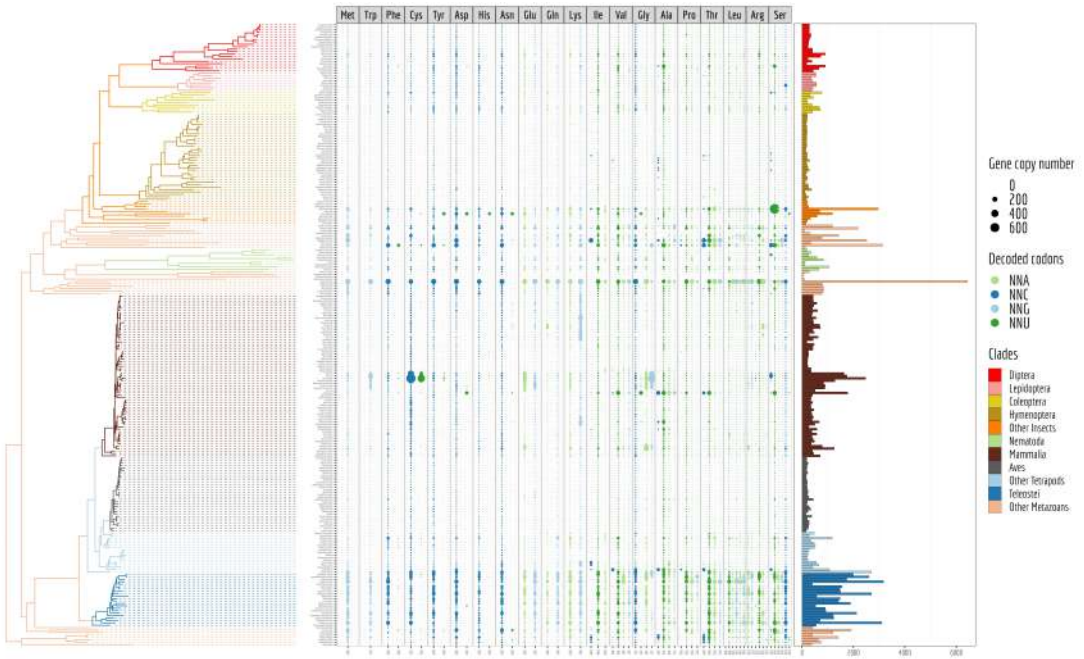


Figure B.2: tRNA gene copy number. *For the 257 studied species from left to right: phylogenetic tree, number of tRNA gene copy number per amino acid per codons and total number of tRNA gene copy.*

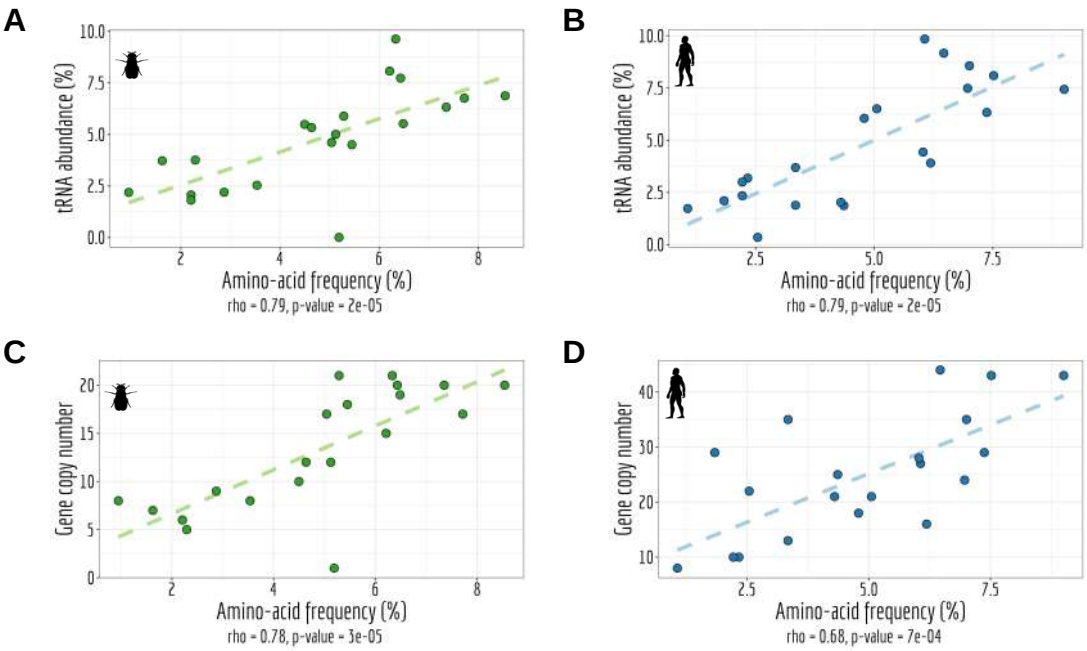


Figure B.3: tRNA abundance proxies. ***A,B****: Relationship between the tRNA abundance measured by Behrens* et al. *(2021) and the frequency of amino acid weighted by gene expression (FPKM).* ***C,D****: Relationship between tRNA gene copy number and the frequency of amino acid weighted by gene expression (FPKM). Left: Drosophila melanogaster; Right: Homo sapiens*

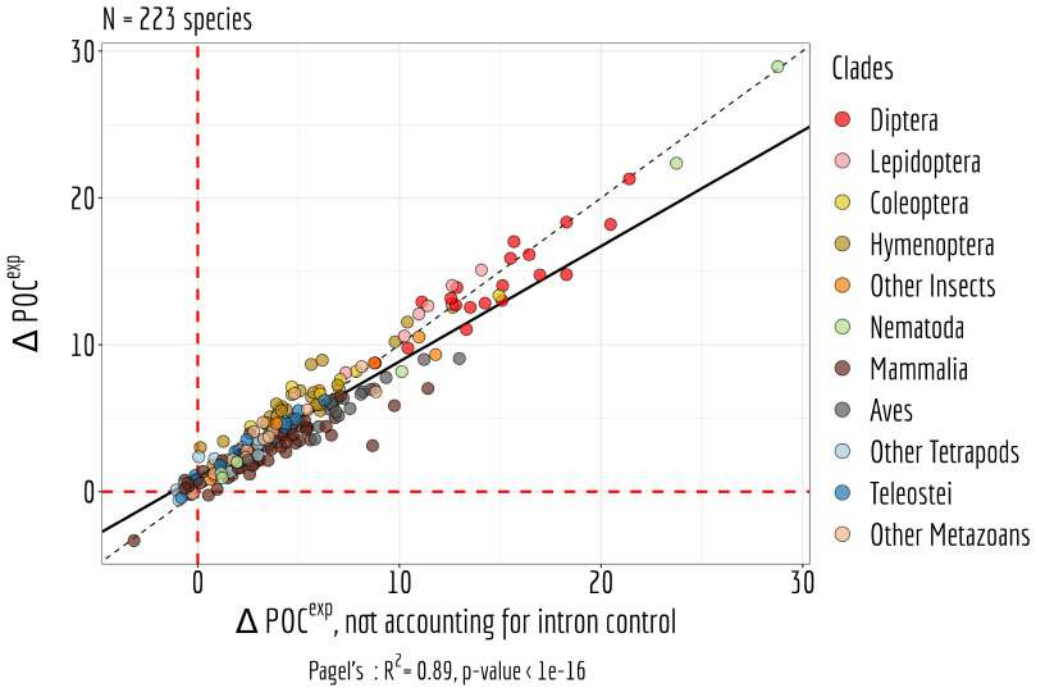


Figure B.4: Counting for intronic background do not change the signal of translational selection. *The X-axis represents variations in POCs frequency, calculated as the difference between POC frequency in the 2% most highly expressed genes and the 50% least expressed genes. The Y-axis depicts the refined X-axis values by eliminating variations arising from non-adaptive processes, such as the difference in POC-control frequency between the 2% most highly expressed genes and the 50% least expressed genes. The black line represents the pagel's lambda model, and the dotted line represents x=y.*

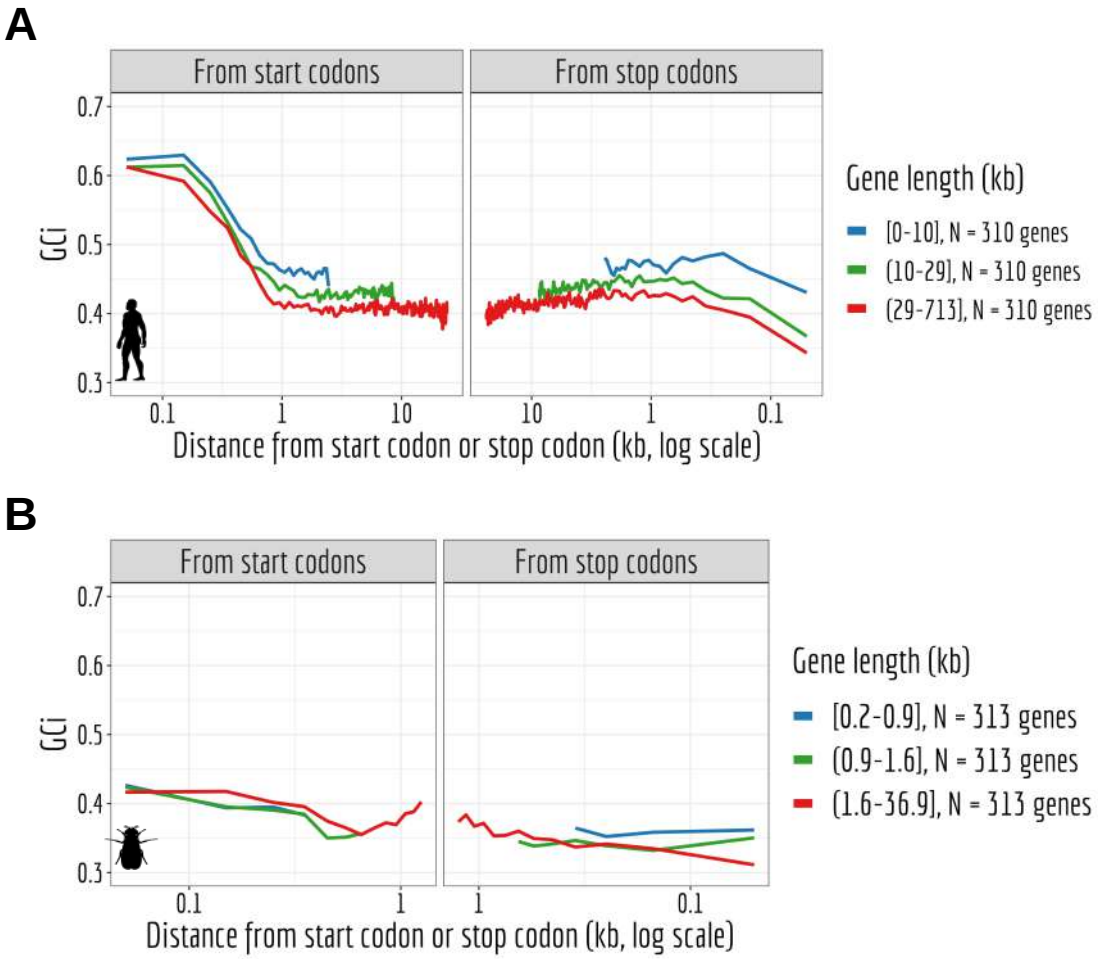


Figure B.5: Non homogenous GC composition along genes. ***A,B****: Measured of the GC composition in introns using 100 bp windows from the start codon and the stop codon (kb, log scale). Equal group of genes have been formed regarding their length, represented by distinct color groups.* ***A****: Homo sapiens;* ***B****: Drosophila melanogaster*

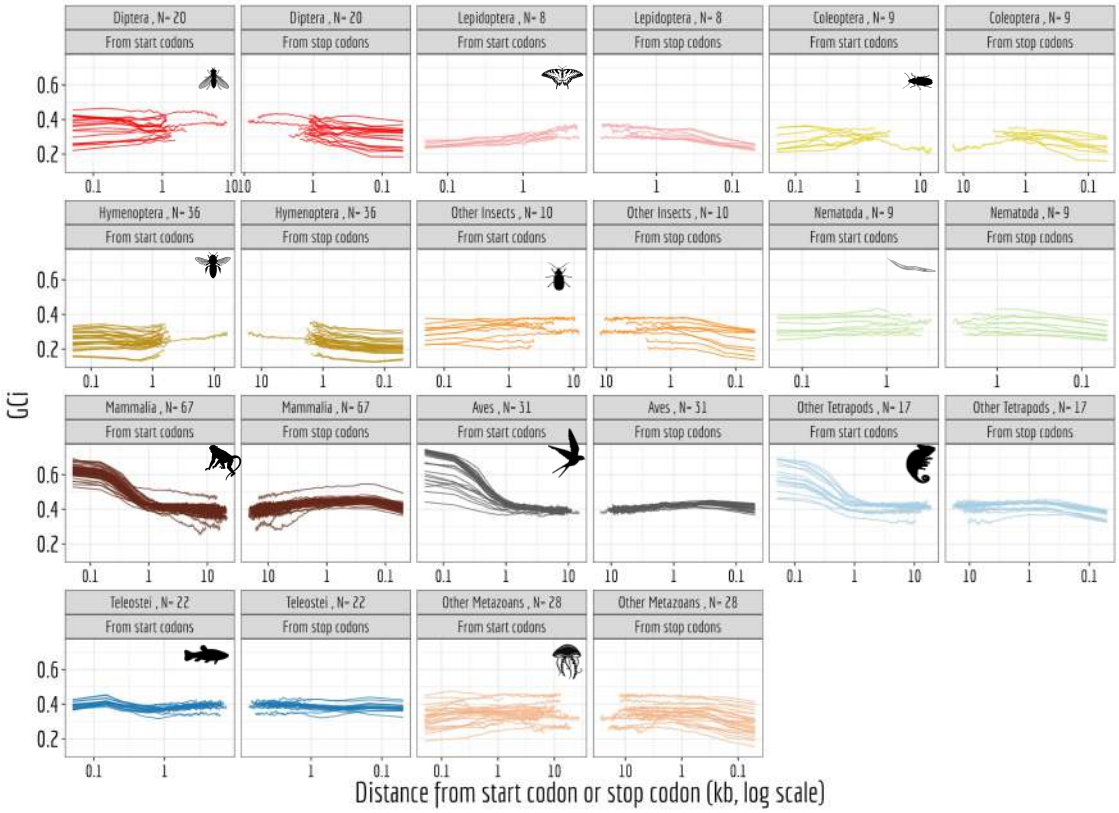


Figure B.6: Non homogenous GC composition along genes for 11 clades. *Measured of the GC composition in introns using 100 bp windows from the start codon and the stop codon (kb, log scale). Each clade of the study is represented by color, one line represents one species.*

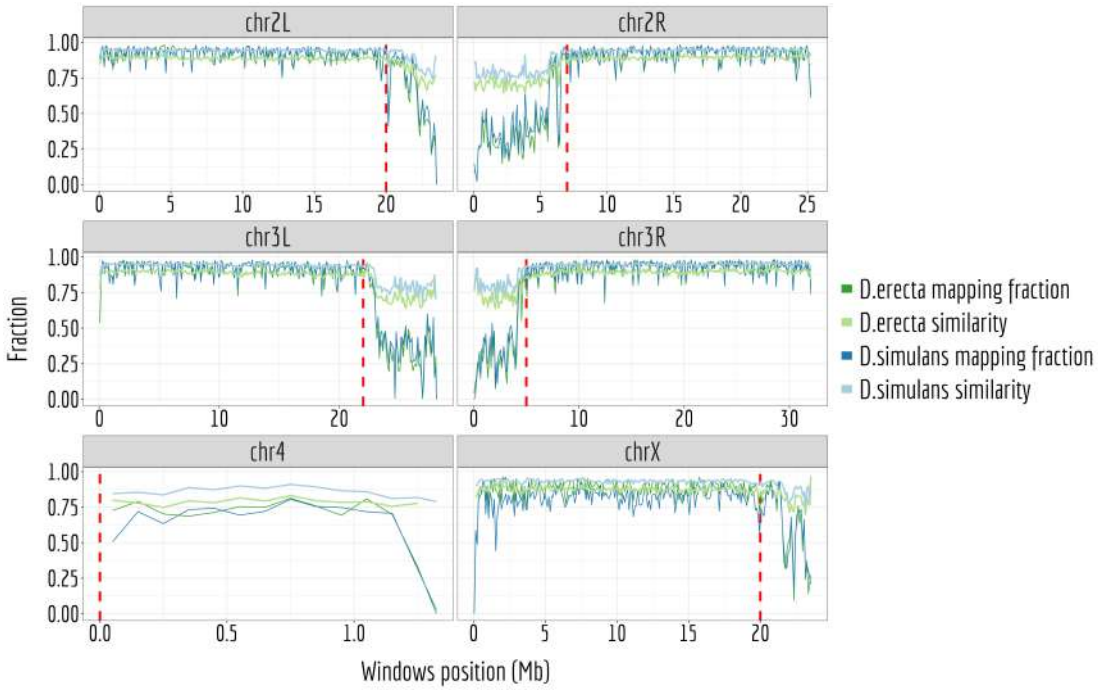


Figure B.7: Multiple genome alignment quality of Drosophila simulans and Drosophila erecta on Drosophila melanogaster. *On the genome of Drosophila melanogaster we quantify the fraction of sites mapped to the two other genomes, and the similarity of these sites. The red dotted lines represent the threshold above or below which the quality of the multiple genome alignment is poor.*

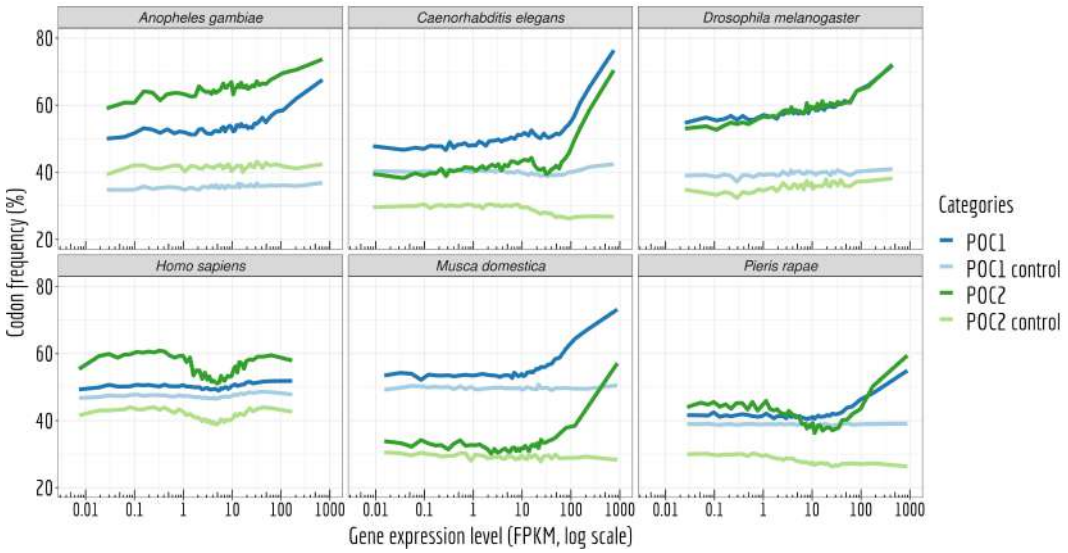


Figure B.8: Differences in usage of putative-optimal codon between highly- and lowly-expressed genes in 6 species. *Variation in the proportion of POC within coding sequences (POC1: dark blue; POC2: dark green) according to gene expression level. To control for variations in neutral substitution patterns, we analyzed the frequency of corresponding triplets within introns (POC1 control: light blue; POC2 control: light green). Each point represents a 2% bin of genes, with the red point at the end of each POC1 curve denoting the 2% most highly expressed genes. The red lines indicate the average POC1 proportions observed in the 50% least expressed genes (FPKM, log scale).*

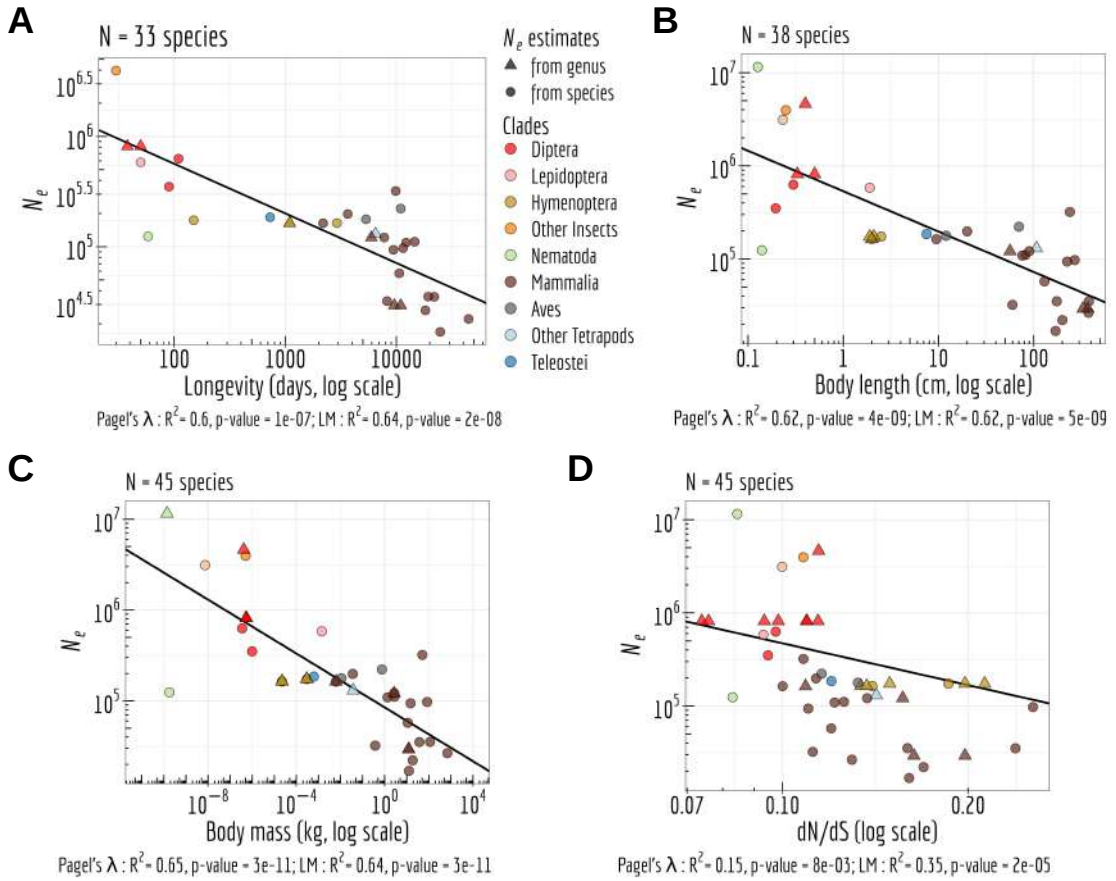


Figure B.9: Relationship between N_e and its proxies. *Relationship between N_e and the longevity (days, log scale; **A**), body length (cm, log scale; **B**), body mass (kg, log scale; **C**), dN/dS (log scale; **D**). Pagel's lambda model is used to take into account the phylogenetic structure of the data in a regression model (the regression line is displayed in black when the correlation is significant).*

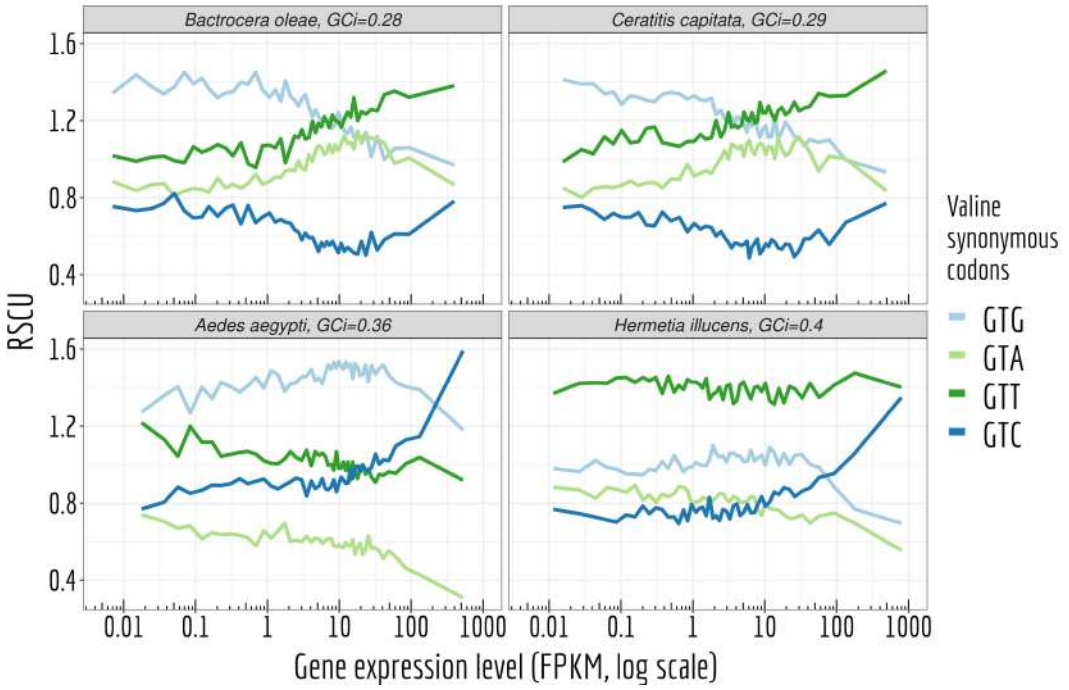


Figure B.10: Valine synonymous codons usage variations with expression among 4 species. *Relationship between the relative synonymous codon usage (RSCU) of valine synonymous codons (GTG/GTA/GTT/GTC) and gene expression in four species with different GC richness.*

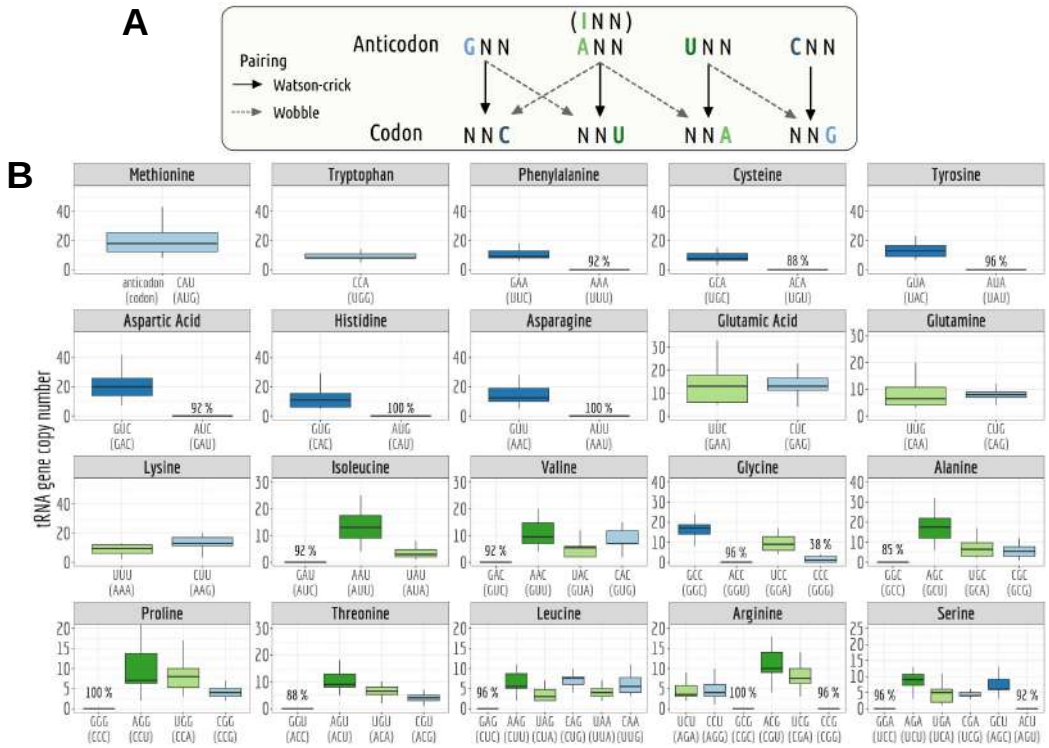


Figure B.11: Presence-Absence of tRNA defines set of putative-optimal codons for species subject to translational selection. ***A****: Illustration of the various possible pairings: Watson-Crick and wobble pairing.* ***B****: A boxplot illustrating the distribution of tRNA gene copy numbers across 26 species subject to translational selection (Lepidoptera and Diptera). The percentage of species lacking a tRNA gene copy is also indicated, highlighting the absence of tRNA isodecoder.*

C
Preliminary results

Figures

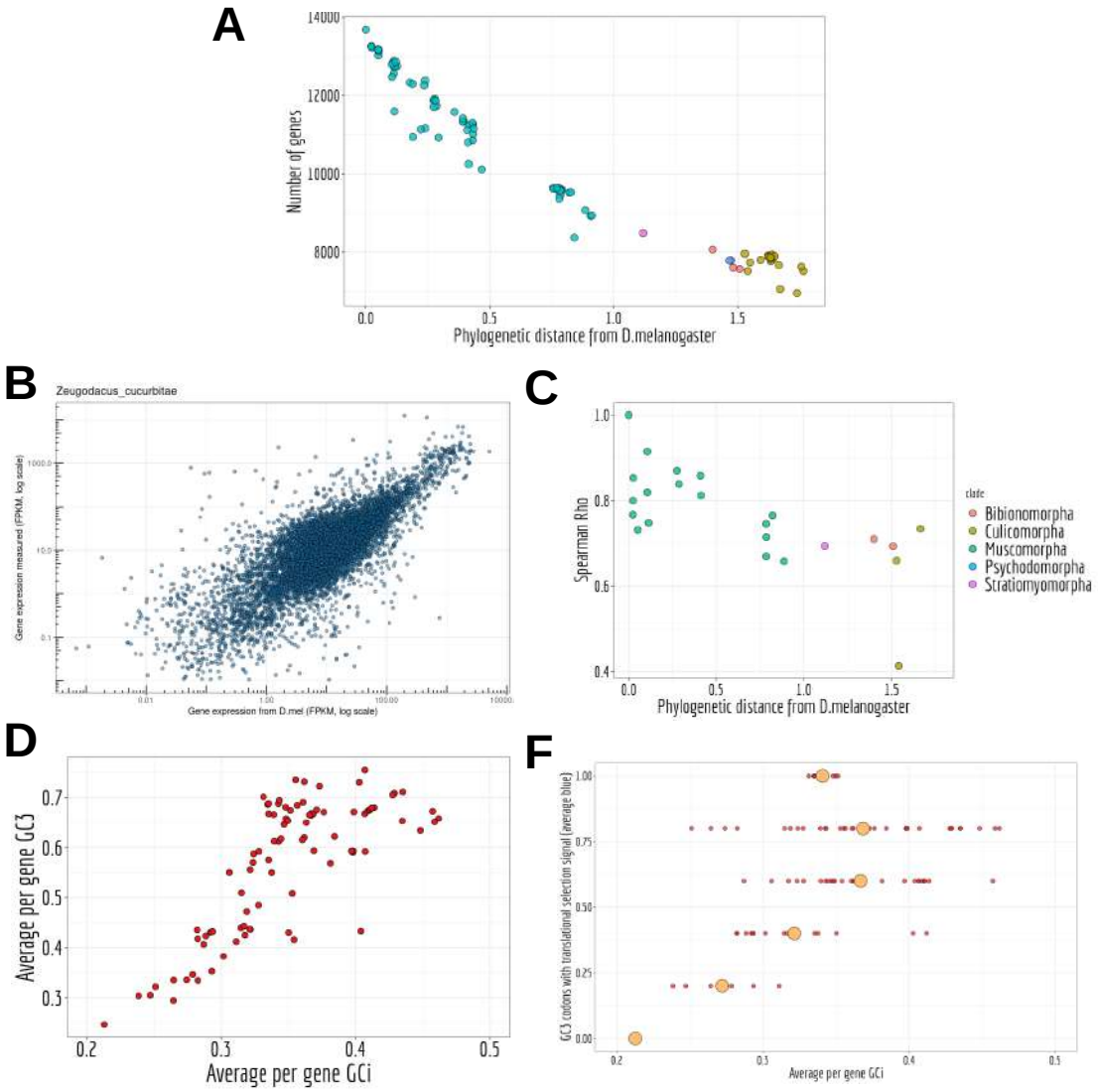


Figure C.1: Translational selection in Diptera. ***A****: Number of genes with reciprocal blast hits with D. melanogaster for 95 dipterans, and phylogenetic distance from D. melanogaster.* ***B****: Gene expression measured in Z. cucurbitae compared to the gene expression measured in D. melanogaster for genes with reciprocal blast hits.* ***C****: Spearmann coefficient (rho), corresponding to the graphic in* ***B****, for 22 species for which gene expression data were available. X-axis is the phylogenetic distance from D. melanogaster.* ***D****: Relationship between GC3 and GCi for the 95 dipterans studied.* ***F****: Relationship between the GC3 of codons optimizing translation and GCi for the 95 dipterans studied.*

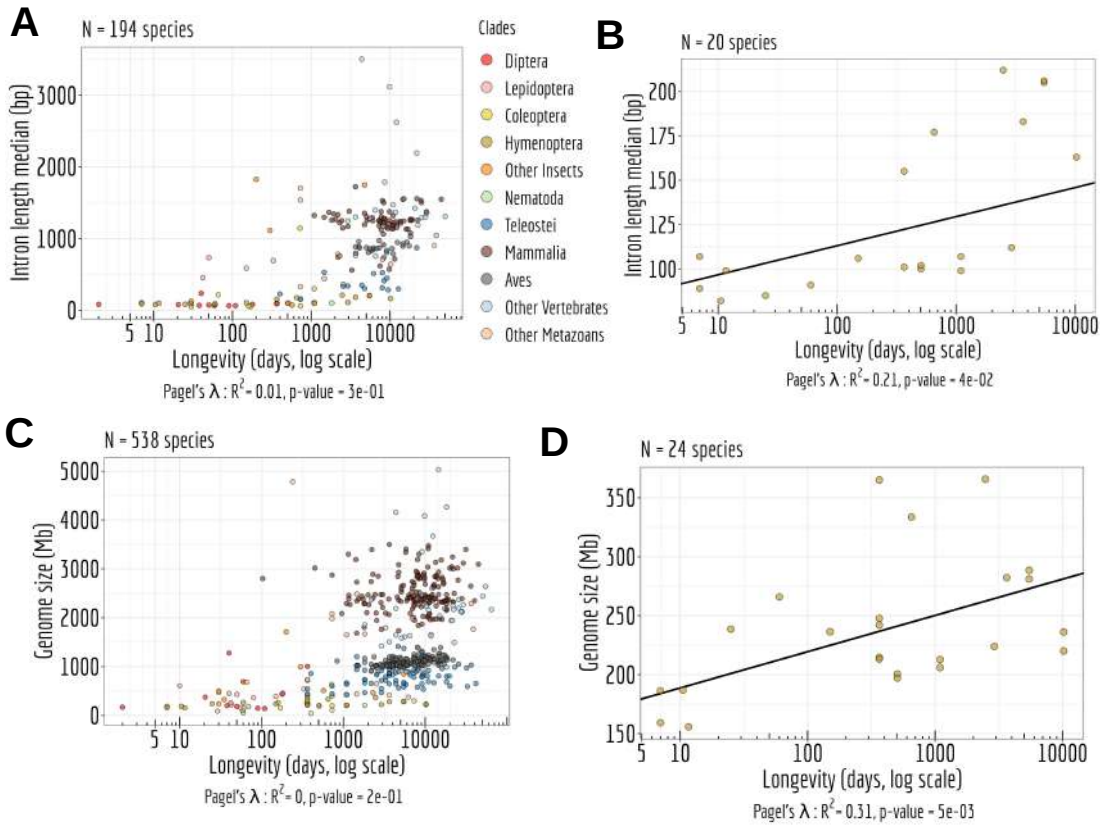


Figure C.2: Impact of N_e on introns length and genomes size. ***A****: Relationship between median intron length and longevity (days, log scale) per species from GTDrift.* ***B****: Relationship between median intron length and longevity (days, log scale) focused on hymenopterans.* ***C****: Relationship between genomes size (Mb) and longevity (days, log scale).* ***D****: Relationship between genomes size (Mb) and longevity (days, log scale) focused on hymenopterans. Pagel's lambda model is used to take into account the phylogenetic structure of the data in a regression model (black line if significant).*

Bibliography

2022. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2022 update. *Nucleic Acids Research*, 50(W1): W345–W351. *Cited at page 135*

2023. Hugging Face – The AI community building the future. *Cited at page 55*

Abanda, N. and Xavier, R.F. 2012. *Régulation des bio-agresseurs dans les cultures associées de blé dur et de pois : impact de la diversité végétale sur la démographie des pucerons du pois*. thesis, Toulouse 3. Publication Title: http://www.theses.fr. *Cited at page 15*

Abascal, F. et al 2015. Alternatively Spliced Homologous Exons Have Ancient Origins and Are Highly Expressed at the Protein Level. *PLOS Computational Biology*, 11(6): e1004325. Publisher: Public Library of Science. *Cited at page 76*

Adams, M.D. et al 2000. The Genome Sequence of Drosophila melanogaster. *Science*, 287(5461): 2185–2195. Publisher: American Association for the Advancement of Science. *Cited at page 31*

Aebi, M., Hornig, H. and Weissmann, C. 1987. 5′ cleavage site in eukaryotic pre-mRNA splicing is determined by the overall 5′ splice region, not by the conserved 5′ GU. *Cell*, 50(2): 237–246. Publisher: Elsevier. *Cited at page 8*

Akashi, H. 1994. Synonymous codon usage in Drosophila melanogaster: natural selection and translational accuracy. *Genetics*, 136(3): 927–935. *Cited at pages 11, 21, 46, 102*

Akashi, H. 1999. Within- and between-species DNA sequence variation and the 'footprint' of natural selection. *Gene*, 238(1): 39–51. *Cited at page 22*

Allen, G.E. 1968. Thomas Hunt Morgan and the problem of natural selection. *Journal of the History of Biology*, 1(1): 113–139. *Cited at page 15*

Allen, G.E. 1969. Hugo De Vries and the Reception of the "Mutation Theory". *Journal of the History of Biology*, 2(1): 55–87. Publisher: Springer. *Cited at page 20*

Altschul, S.F. et al 1990. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3): 403–410. *Cited at page 34*

Amit, M. et al 2012. Differential GC Content between Exons and Introns Establishes Distinct Strategies of Splice-Site Recognition. *Cell Reports*, 1(5): 543–556. *Cited at page 105*

Anczuków, O. and Krainer, A.R. 2016. Splicing-factor alterations in cancers. *RNA*, 22(9): 1285–1301. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab. *Cited at page 9*

Auton, A. et al 2015. A global reference for human genetic variation. *Nature*, 526(7571): 68–74. Number: 7571 Publisher: Nature Publishing Group. *Cited at pages 85, 98, 149*

Ayala, F.J. and Fitch, W.M. 1997. Genetics and the origin of species: An introduction. *Proceedings of the National Academy of Sciences*, 94(15): 7691–7697. Publisher: Proceedings of the National Academy of Sciences. *Cited at page 16*

Ayoun, L. 2021. Quelle est l'empreinte carbone d'un vol en avion ? *Cited at page 139*

Bacaër, N. 2011. Wright and random genetic drift (1931). In N. Bacaër, editor, *A Short History of Mathematical Population Dynamics*, pages 105–109. Springer, London. *Cited at page 17*

Balis, M.E. et al 1958. Role of the Ribonucleoprotein Particle in Protein Synthesis and the Effects of Growth Hormone. *Journal of Biological Chemistry*, 233(5): 1152–1155. *Cited at page 10*

Barahona, A. and Ayala, F.J. 2005. Theodosius Dobzhansky's Role in the Emergence and Institutionalization of Genetics in Mexico. *Genetics*, 170(3): 981–987. *Cited at page 16*

Barbosa-Morais, N.L. et al 2012. The evolutionary landscape of alternative splicing in vertebrate species. *Science (New York, N.Y.)*, 338(6114): 1587–1593. *Cited at pages 41, 52, 76, 77, 82, 89, 92*

Bastian, F.B. et al 2020. The Bgee suite: integrated curated expression atlas and comparative transcriptomics in animals. *Nucleic Acids Research*, 49(D1): D831–D847. *Cited at pages 39, 53*

Baxter, S.W. et al 2017. EB Ford revisited: assessing the long-term stability of wing-spot patterns and population genetic structure of the meadow brown butterfly on the Isles of Scilly. *Heredity*, 118(4): 322–329. Number: 4 Publisher: Nature Publishing Group. *Cited at page 17*

Beelman, C.A. and Parker, R. 1995. Degradation of mRNA in eukaryotes. *Cell*, 81(2): 179–183. *Cited at page 7*

Behjati, S. and Tarpey, P.S. 2013. What is next generation sequencing? *Archives of Disease in Childhood. Education and Practice Edition*, 98(6): 236–238. *Cited at page 31*

Behrens, A., Rodschinka, G. and Nedialkova, D.D. 2021. High-resolution quantitative profiling of tRNA abundance and modification status in eukaryotes by mim-tRNAseq. *Molecular Cell*, 81(8): 1802–1815.e7. Publisher: Elsevier. *Cited at pages 48, 106, 107, 121, 159*

Bergeron, L.A. et al 2023. Evolution of the germline mutation rate across vertebrates. *Nature*, 615(7951): 285–291. Number: 7951 Publisher: Nature Publishing Group. *Cited at pages 22, 131*

Berget, S.M., Moore, C. and Sharp, P.A. 1977. Spliced segments at the 5' terminus of adenovirus 2 late mRNA. *Proceedings of the National Academy of Sciences*, 74(8): 3171–3175. Publisher: Proceedings of the National Academy of Sciences. *Cited at pages 8, 9, 52*

Berk, A.J. 2016. Discovery of RNA splicing and genes in pieces. *Proceedings of the National Academy of Sciences*, 113(4): 801–805. Publisher: Proceedings of the National Academy of Sciences. *Cited at pages 8, 9*

Berry, A. and Browne, J. 2022. Mendel and Darwin. *Proceedings of the National Academy of Sciences*, 119(30): e2122144119. Publisher: Proceedings of the National Academy of Sciences. *Cited at page 15*

Bhangale, T.R. et al 2005. Comprehensive identification and characterization of diallelic insertion–deletion polymorphisms in 330 human candidate genes. *Human Molecular Genetics*, 14(1): 59–69. *Cited at page 20*

Bhattacharyya, S. et al 2018. Accessibility of the Shine-Dalgarno Sequence Dictates N-Terminal Codon Bias in E. coli. *Molecular Cell*, 70(5): 894–905.e5. *Cited at page 11*

Bhuiyan, S.A. et al 2018. Systematic evaluation of isoform function in literature reports of alternative splicing. *BMC Genomics*, 19(1): 637. *Cited at page 76*

Birse, C.E. et al 1997. Transcriptional termination signals for RNA polymerase II in fission yeast. *The EMBO Journal*, 16(12): 3633–3643. *Cited at page 7*

Black, D.L. 2003. Mechanisms of alternative pre-messenger RNA splicing. *Annual Review of Biochemistry*, 72: 291–336. *Cited at page 132*

Blanchet, S. and Ranjan, N. 2022. Translation Phases in Eukaryotes. In *Ribosome Biogenesis: Methods and Protocols [Internet]*. Humana. *Cited at page 10*

Blencowe, B.J. 2006. Alternative Splicing: New Insights from Global Analyses. *Cell*, 126(1): 37–47. Publisher: Elsevier. *Cited at pages 9, 41*

Blencowe, B.J. 2017. The Relationship between Alternative Splicing and Proteomic Complexity. *Trends in Biochemical Sciences*, 42(6): 407–408. Publisher: Elsevier. *Cited at pages 10, 43, 76, 132*

Bolívar, P. et al 2019. GC-biased gene conversion conceals the prediction of the nearly neutral theory in avian genomes. *Genome Biology*, 20(1): 5. *Cited at pages 40, 59, 97*

Bonnal, S.C., López-Oreja, I. and Valcárcel, J. 2020. Roles and mechanisms of alternative splicing in cancer — implications for care. *Nature Reviews Clinical Oncology*, 17(8): 457–474. Number: 8 Publisher: Nature Publishing Group. *Cited at page 9*

Boël, G. et al 2016. Codon influence on protein expression in E. coli correlates with mRNA levels. *Nature*, 529(7586): 358–363. *Cited at page 11*

Braunschweig, U. et al 2014. Widespread intron retention in mammals functionally tunes transcriptomes. *Genome Research*, 24(11): 1774–1786. *Cited at pages 43, 45*

Breathnach, R. and Chambon, P. 1981. Organization and Expression of Eucaryotic Split Genes Coding for Proteins. *Annual Review of Biochemistry*, 50(1): 349–383. _eprint: https://doi.org/10.1146/annurev.bi.50.070181.002025. *Cited at page 8*

Breathnach, R., Mandel, J.L. and Chambon, P. 1977. Ovalbumin gene is split in chicken DNA. *Nature*, 270(5635): 314–319. Number: 5635 Publisher: Nature Publishing Group. *Cited at page 8*

Breathnach, R. et al 1978. Ovalbumin gene: evidence for a leader sequence in mRNA and DNA sequences at the exon-intron boundaries. *Proceedings of the National Academy of Sciences*, 75(10): 4853–4857. Publisher: Proceedings of the National Academy of Sciences. *Cited at page 8*

Brennan, S.O. et al 1990. Hypermutability of CpG dinucleotides in the propeptide-encoding sequence of the human albumin gene. *Proceedings of the National Academy of Sciences*, 87(10): 3909–3913. *Cited at page 21*

Bricout, R. et al 2023. Evolution is not Uniform Along Coding Sequences. *Molecular Biology and Evolution*, 40(3): msad042. *Cited at page 112*

Brion, C. et al 2015. Evolution of intraspecific transcriptomic landscapes in yeasts. *Nucleic Acids Research*, 43(9): 4558–4568. *Cited at pages 23, 27*

Brown, H., Sanger, F. and Kitai, R. 1955. The structure of pig and sheep insulins. *Biochemical Journal*, 60(4): 556–565. *Cited at page 34*

Brunak, S. et al 2002. Nucleotide Sequence Database Policies. *Science*, 298(5597): 1333–1333. Publisher: American Association for the Advancement of Science. *Cited at page 35*

Buhr, F. et al 2016. SYNONYMOUS CODONS DIRECT CO-TRANSLATIONAL FOLDING TOWARDS DIFFERENT PROTEIN CONFORMATIONS. *Molecular cell*, 61(3): 341–351. *Cited at pages 11, 103*

Bulmer, M. 1987. Coevolution of codon usage and transfer RNA abundance. *Nature*, 325(6106): 728–730. Publisher: Nature Publishing Group. *Cited at page 102*

Bulmer, M. 1991. The Selection-Mutation-Drift Theory of Synonymous Codon Usage. *Genetics*, 129(3): 897–907. *Cited at pages 45, 102, 103, 113*

Burkhardt, R.W. 2013. Lamarck, Evolution, and the Inheritance of Acquired Characters. *Genetics*, 194(4): 793–805. *Cited at page 14*

Burks, C. et al 1991. GenBank. *Nucleic Acids Research*, 19 Suppl(Suppl): 2221–2225. *Cited at page 34*

Bush, S.J. et al 2017. Alternative splicing and the evolution of phenotypic novelty. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 372(1713): 20150474. Publisher: Royal Society. *Cited at pages 77, 89*

Bénitière, F., Duret, L. and Necsulea, A. 2024. GTDrift: A resource for exploring the interplay between genetic drift, genomic and transcriptomic characteristics in eukaryotes. Pages: 2024.01.23.576799 Section: New Results. *Cited at pages 104, 124, 126, 139*

Bénitìere, F., Necsulea, A. and Duret, L. 2024. Random genetic drift sets an upper limit on mRNA splicing accuracy in metazoans. *eLife*, 13. Publisher: eLife Sciences Publications Limited. *Cited at pages 52, 53, 61, 62, 64, 65, 69, 133*

Cai, X. et al 2022. Transposable element insertion: a hidden major source of domesticated phenotypic variation in Brassica rapa. *Plant Biotechnology Journal*, 20(7): 1298–1310. *Cited at page 20*

Callaway, E. 2021. Oldest DNA from a Homo sapiens reveals surprisingly recent Neanderthal ancestry. *Nature*, 592(7854): 339–339. Bandiera_abtest: a Cg_type: News Number: 7854 Publisher: Nature Publishing Group Subject_term: Genomics, Palaeontology, Anthropology. *Cited at page 29*

Cann, R.L., Stoneking, M. and Wilson, A.C. 1987. Mitochondrial DNA and human evolution. *Nature*, 325(6099): 31–36. Number: 6099 Publisher: Nature Publishing Group. *Cited at page 29*

Cardoso-Moreira, M. et al 2019. Gene expression across mammalian organ development. *Nature*, 571(7766): 505–509. *Cited at pages 59, 82, 83, 148*

Casás-Selves, M. and DeGregori, J. 2011. How cancer shapes evolution, and how evolution shapes cancer. *Evolution*, 4(4): 624–634. *Cited at page 29*

Chabanon, H., Mickleburgh, I. and Hesketh, J. 2004. Zipcodes and postage stamps: mRNA localisation signals and their trans-acting binding proteins. *Briefings in Functional Genomics & Proteomics*, 3(3): 240–256. *Cited at page 6*

Chamary, J.V., Parmley, J.L. and Hurst, L.D. 2006. Hearing silence: non-neutral evolution at synonymous sites in mammals. *Nature Reviews Genetics*, 7(2): 98–108. *Cited at page 133*

Chan, P. et al 2021. tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. *Nucleic Acids Research*, 49(16): 9077–9096. *Cited at pages 48, 124*

Chan, S. et al 2017. Frequent GU wobble pairings reduce translation efficiency in Plasmodium falciparum. *Scientific Reports*, 7(1): 723. Number: 1 Publisher: Nature Publishing Group. *Cited at pages 108, 121*

Chang, W. et al 2024. *shiny: Web Application Framework for R.* *Cited at pages 53, 62*

Charif, D. and Lobry, J.R. 2007. SeqinR 1.0-2: A Contributed Package to the R Project for Statistical Computing Devoted to Biological Sequences Retrieval and Analysis. In U. Bastolla, M. Porto, H. E. Roman, and M. Vendruscolo, editors, *Structural Approaches to Sequence Evolution: Molecules, Networks, Populations*, Biological and Medical Physics, Biomedical Engineering, pages 207–232. Springer, Berlin, Heidelberg. *Cited at pages 57, 96*

Charlesworth, B. 2022. Fisher's historic 1922 paper On the dominance ratio. *Genetics*, 220(3): iyac006. *Cited at page 15*

Charlesworth, B. et al 2016. Hubby and Lewontin on Protein Variation in Natural Populations: When Molecular Genetics Came to the Rescue of Population Genetics. *Genetics*, 203(4): 1497–1503. *Cited at page 18*

Chen, L. et al 2014. Correcting for Differential Transcript Coverage Reveals a Strong Relationship between Alternative Splicing and Organism Complexity. *Molecular Biology and Evolution*, 31(6): 1402–1413. *Cited at pages 9, 41, 42, 43, 44, 45, 52, 76, 77, 89, 90*

Chen, S.L. et al 2004. Codon usage between genomes is constrained by genome-wide mutational processes. *Proceedings of the National Academy of Sciences*, 101(10): 3480–3485. Publisher: Proceedings of the National Academy of Sciences. *Cited at pages 45, 46*

Choi, I.Y., Kwon, E.C. and Kim, N.S. 2020. The C- and G-value paradox with polyploidy, repeatomes, introns, phenomes and cell economy. *Genes & Genomics*, 42(7): 699–714. *Cited at page 7*

Clamp, M. et al 2007. Distinguishing protein-coding and noncoding genes in the human genome. *Proceedings of the National Academy of Sciences*, 104(49): 19428–19433. Publisher: Proceedings of the National Academy of Sciences. *Cited at page 7*

Clay, O.K. and Bernardi, G. 2011. GC3 of Genes Can Be Used as a Proxy for Isochore Base Composition: A Reply to Elhaik et al. *Molecular Biology and Evolution*, 28(1): 21–23. *Cited at page 46*

Cokelaer, T., Cohen-Boulakia, S. and Lemoine, F. 2023. Reprohackathons: promoting reproducibility in bioinformatics through training. *Bioinformatics*, 39(Supplement_1): i11–i20. *Cited at page 135*

Comeron, J.M. 2004. Selective and mutational patterns associated with gene expression in humans: influences on synonymous composition and intron presence. *Genetics*, 167(3): 1293–1304. *Cited at page 12*

Conrad, D.F. et al 2006. A high-resolution survey of deletion polymorphism in the human genome. *Nature Genetics*, 38(1): 75–81. Number: 1 Publisher: Nature Publishing Group. *Cited at page 20*

Coop, G. 2020. *Population and quantitative genetics*. *Cited at page 22*

Cortez, D. 2019. Replication-coupled DNA Repair. *Molecular cell*, 74(5): 866–876. *Cited at page 21*

Courel, M. et al 2019. GC content shapes mRNA storage and decay in human cells. *eLife*, 8: e49708. Publisher: eLife Sciences Publications, Ltd. *Cited at page 103*

Crespi, B. and Summers, K. 2005. Evolutionary biology of cancer. *Trends in Ecology & Evolution*, 20(10): 545–552. *Cited at page 29*

Crow, J.F. 2002. Perspective: Here's to Fisher, additive genetic variance, and the fundamental theorem of natural selection. *Evolution; International Journal of Organic Evolution*, 56(7): 1313–1316. *Cited at page 15*

Cummings, B.B. et al 2020. Transcript expression-aware annotation improves rare variant interpretation. *Nature*, 581(7809): 452–458. Number: 7809 Publisher: Nature Publishing Group. *Cited at page 133*

Cutter, A.D. 2015. Caenorhabditis evolution in the wild. *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology*, 37(9): 983–995. *Cited at page 123*

Damuth, J. 1981. Population density and body size in mammals. *Nature*, 290(5808): 699–700. *Cited at page 70*

Dana, A. and Tuller, T. 2014. The effect of tRNA levels on decoding times of mRNA codons. *Nucleic Acids Research*, 42(14): 9171–9181. *Cited at pages 12, 21*

Dance, A. 2023. Stop the peer-review treadmill. I want to get off. *Nature*, 614(7948): 581–583. Bandiera_abtest: a Cg_type: Career Feature Number: 7948 Publisher: Nature Publishing Group Subject_term: Careers, Peer review, Publishing, Research management, Lab life. *Cited at page 134*

Darwin, C...A.d.t. 1868. *The variation of animals and plants under domestication. Volume 2 / by Charles Darwin,...* *Cited at page 14*

Darwin, C. et al 1859. *On the origin of species by means of natural selection, or, The preservation of favoured races in the struggle for life*. John Murray, Albemarle Street, London. Pages: 1-564. *Cited at page 14*

David, C.J. and Manley, J.L. 2010. Alternative pre-mRNA splicing regulation in cancer: pathways and programs unhinged. *Genes & Development*, 24(21): 2343–2364. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab. *Cited at page 133*

Deaton, A.M. and Bird, A. 2011. CpG islands and the regulation of transcription. *Genes & Development*, 25(10): 1010–1022. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab. *Cited at page 112*

Denver, D.R. et al 2012. Variation in Base-Substitution Mutation in Experimental and Natural Lineages of Caenorhabditis Nematodes. *Genome Biology and Evolution*, 4(4): 513–522. *Cited at page 126*

Dhindsa, R.S. et al 2020. Natural Selection Shapes Codon Usage in the Human Genome. *American Journal of Human Genetics*, 107(1): 83–95. *Cited at pages 12, 47, 133*

Di Tommaso, P. et al 2017. Nextflow enables reproducible computational workflows. *Nature Biotechnology*, 35(4): 316–319. *Cited at page 135*

Dobzhansky, T. 1937. *Genetics and the Origin of Species*. Columbia University Press. Google-Books-ID: ICpMAAAAMAAJ. *Cited at page 16*

Doherty, A. and McInerney, J.O. 2013. Translational Selection Frequently Overcomes Genetic Drift in Shaping Synonymous Codon Usage Patterns in Vertebrates. *Molecular Biology and Evolution*, 30(10): 2263–2267. *Cited at pages 12, 45, 46*

Dong, H., Nilsson, L. and Kurland, C.G. 1996. Co-variation of tRNA Abundance and Codon Usage inEscherichia coliat Different Growth Rates. *Journal of Molecular Biology*, 260(5): 649–663. *Cited at page 102*

dos Reis, M. and Wernisch, L. 2009. Estimating Translational Selection in Eukaryotic Genomes. *Molecular Biology and Evolution*, 26(2): 451–461. *Cited at pages 46, 103, 106, 113, 115, 120, 122*

Drummond, D.A. and Wilke, C.O. 2008. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell*, 134(2): 341–352. *Cited at pages 11, 21, 46*

Drummond, D.A., Raval, A. and Wilke, C.O. 2006. A Single Determinant Dominates the Rate of Yeast Protein Evolution. *Molecular Biology and Evolution*, 23(2): 327–337. *Cited at page 102*

Duncan, B.K. and Miller, J.H. 1980. Mutagenic deamination of cytosine residues in DNA. *Nature*, 287(5782): 560–561. Number: 5782 Publisher: Nature Publishing Group. *Cited at page 21*

Dunham, I. et al 1999. The DNA sequence of human chromosome 22. *Nature*, 402: 489–495. ADS Bibcode: 1999Natur.402..489D. *Cited at page 31*

Duret, L. 2000. tRNA gene number and codon usage in the C. elegans genome are co-adapted for optimal translation of highly expressed genes. *Trends in Genetics*, 16(7): 287–289. Publisher: Elsevier. *Cited at pages 11, 12, 45, 46, 48, 102, 107*

Duret, L. 2002. Evolution of synonymous codon usage in metazoans. *Current Opinion in Genetics & Development*, 12(6): 640–649. *Cited at pages 45, 133*

Duret, L. and Arndt, P.F. 2008. The impact of recombination on nucleotide substitutions in the human genome. *PLoS genetics*, 4(5): e1000071. *Cited at page 28*

Duret, L. and Galtier, N. 2009. Biased gene conversion and the evolution of mammalian genomic landscapes. *Annual Review of Genomics and Human Genetics*, 10: 285–311. *Cited at pages 28, 102, 117*

Duret, L. and Mouchiroud, D. 1999. Expression pattern and, surprisingly, gene length shape codon usage in Caenorhabditis, Drosophila, and Arabidopsis. *Proceedings of the National Academy of Sciences of the United States of America*, 96(8): 4482–4487. *Cited at pages 45, 46, 102, 103, 120, 121*

Duret, L. and Mouchiroud, D. 2000. Determinants of Substitution Rates in Mammalian Genes: Expression Pattern Affects Selection Intensity but Not Mutation Rate. *Molecular Biology and Evolution*, 17(1): 68–070. *Cited at pages 23, 27*

Dutheil, J. and Boussau, B. 2008. Non-homogeneous models of sequence evolution in the Bio++ suite of libraries and programs. *BMC Evolutionary Biology*, 8(1): 255. *Cited at pages 59, 97*

Dutheil, J.Y. et al 2012. Efficient selection of branch-specific models of sequence evolution. *Molecular Biology and Evolution*, 29(7): 1861–1874. *Cited at pages 59, 97*

Early, P. et al 1980. Two mRNAs can be produced from a single immunoglobulin μ gene by alternative RNA processing pathways. *Cell*, 20(2): 313–319. *Cited at page 9*

Ebertz, D.A. 2020. A Journey Through The History Of DNA Sequencing. *Cited at page 30*

Edwards, A.W.F. 2008. G. H. Hardy (1908) and Hardy–Weinberg Equilibrium. *Genetics*, 179(3): 1143–1150. *Cited at page 23*

Emami, K.H., Jain, A. and Smale, S.T. 1997. Mechanism of synergy between TATA and initiator: synergistic binding of TFIID following a putative TFIIA-induced isomerization. *Genes & Development*, 11(22): 3007–3019. *Cited at page 7*

Epstein, C.J. 1967. Non-randomness of Ammo-acid Changes in the Evolution of Homologous Proteins. *Nature*, 215(5099): 355–359. Number: 5099 Publisher: Nature Publishing Group. *Cited at page 21*

Eyre-Walker, A. and Bulmer, M. 1993. Reduced synonymous substitution rate at the start of enterobacterial genes. *Nucleic Acids Research*, 21(19): 4599–4603. *Cited at page 11*

Ezkurdia, I. et al 2014. Multiple evidence strands suggest that there may be as few as 19 000 human protein-coding genes. *Human Molecular Genetics*, 23(22): 5866–5878. *Cited at page 7*

Fairbanks, D.J. 2020. Mendel and Darwin: untangling a persistent enigma. *Heredity*, 124(2): 263–273. Number: 2 Publisher: Nature Publishing Group. *Cited at page 15*

Felsenstein, J. 1971. Inbreeding and Variance Effective Numbers in Populations with Overlapping Generations. *Genetics*, 68(4): 581–597. *Cited at page 23*

Fields, C. et al 1994. How many genes in the human genome? *Nature Genetics*, 7(3): 345–346. Number: 3 Publisher: Nature Publishing Group. *Cited at page 6*

Fiers, W. et al 1976. Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. *Nature*, 260(5551): 500–507. Number: 5551 Publisher: Nature Publishing Group. *Cited at page 30*

Figuet, E. et al 2016. Life History Traits, Protein Evolution, and the Nearly Neutral Theory in Amniotes. *Molecular Biology and Evolution*, 33(6): 1517–1527. *Cited at pages 28, 40, 54, 66, 67, 77, 78*

Filipowicz, W., Bhattacharyya, S.N. and Sonenberg, N. 2008. Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? *Nature Reviews Genetics*, 9(2): 102–114. Number: 2 Publisher: Nature Publishing Group. *Cited at page 5*

Fisher, R.A. 1922. 024: On the Dominance Ratio. Accepted: 2006-11-17T01:40:58Z. *Cited at page 15*

Fisher, R.A. 1930. *The genetical theory of natural selection.* Clarendon Press, Oxford. Open Library ID: OL7084333M. *Cited at page 17*

Fisher, R.A. 1931. XVII.—The Distribution of Gene Ratios for Rare Mutations. *Proceedings of the Royal Society of Edinburgh*, 50: 204–219. Publisher: Royal Society of Edinburgh Scotland Foundation. *Cited at page 17*

Fisher, R.A. 1950. THE "SEWALL WRIGHT EFFECT". *Cited at page 17*

Fisher, R.A. and Ford, E.B. 1947. The spread of a gene in natural conditions in a colony of the moth Panaxia dominula L. *Heredity*, 1(2): 143–174. Number: 2 Publisher: Nature Publishing Group. *Cited at page 17*

Ford, E.B. 1945. *Butterflies. New Naturalist No 1*. Collins, London, first edition edition. *Cited at page 17*

Fordyce, S.L. et al 2013. Long-term RNA persistence in postmortem contexts. *Investigative Genetics*, 4: 7. *Cited at pages 7, 8*

Franklin, R.E. and Gosling, R.G. 1953. Molecular Configuration in Sodium Thymonucleate. *Nature*, 171(4356): 740–741. Number: 4356 Publisher: Nature Publishing Group. *Cited at page 4*

Freckleton, R., Harvey, P. and Pagel, M. 2002. Phylogenetic Analysis and Comparative Data: A Test and Review of Evidence. *The American naturalist*, 160: 712–26. *Cited at page 79*

Froese, R. and Pauly, D. 2023. FishBase. Published: World Wide Web electronic publication. *Cited at pages 40, 54*

Fu, J. et al 2018. Codon usage regulates human KRAS expression at both transcriptional and translational levels. *Journal of Biological Chemistry*, 293(46): 17929–17940. *Cited at page 11*

Fuchs, G. et al 2014. 4sUDRB-seq: measuring genomewide transcriptional elongation rates and initiation frequencies within cells. *Genome Biology*, 15(5): R69. *Cited at page 5*

Galtier, N. 2016. Adaptive Protein Evolution in Animals and the Effective Population Size Hypothesis. *PLOS Genetics*, 12(1): e1005774. *Cited at pages 28, 54, 57, 66, 126*

Galtier, N. 2021. Fine-scale quantification of GC-biased gene conversion intensity in mammals. *Peer Community Journal*, 1. *Cited at page 28*

Galtier, N. et al 2018. Codon Usage Bias in Animals: Disentangling the Effects of Natural Selection, Effective Population Size, and GC-Biased Gene Conversion. *Molecular Biology and Evolution*, 35(5): 1092–1103. *Cited at page 103*

Gao, Y. et al 2017. Mechanisms of Post-Replication DNA Repair. *Genes*, 8(2): 64. *Cited at page 21*

Geiler-Samerotte, K.A. et al 2011. Misfolded proteins impose a dosage-dependent fitness cost and trigger a cytosolic unfolded protein response in yeast. *Proceedings of the National Academy of Sciences*, 108(2): 680–685. Publisher: Proceedings of the National Academy of Sciences. *Cited at page 23*

Giani, A.M. et al 2020. Long walk to genomics: History and current approaches to genome sequencing and assembly. *Computational and Structural Biotechnology Journal*, 18: 9–19. *Cited at page 30*

Giménez-Roig, J. et al 2021. Codon Usage and Adenovirus Fitness: Implications for Vaccine Development. *Frontiers in Microbiology*, 12: 633946. *Cited at page 133*

Gingold, H. et al 2014. A Dual Program for Translation Regulation in Cellular Proliferation and Differentiation. *Cell*, 158(6): 1281–1292. Publisher: Elsevier. *Cited at pages 12, 47, 133*

Goffeau, A. et al 1996. Life with 6000 Genes. *Science*, 274(5287): 546–567. Publisher: American Association for the Advancement of Science. *Cited at page 31*

Gonatopoulos-Pournatzis, T. and Cowling, V. 2014. Cap-binding complex (CBC). *Biochemical Journal*, 457(Pt 2): 231–242. *Cited at page 7*

Gonzàlez-Porta, M. et al 2013. Transcriptome analysis of human tissues and cell lines reveals one dominant transcript per gene. *Genome Biology*, 14(7): 1–11. Number: 7 Publisher: BioMed Central. *Cited at pages 43, 52, 62, 68, 76, 80, 133*

Goodman, D.B., Church, G.M. and Kosuri, S. 2013. Causes and effects of N-terminal codon bias in bacterial genes. *Science (New York, N.Y.)*, 342(6157): 475–479. *Cited at page 11*

Gorochowski, T.E. et al 2015. Trade-offs between tRNA abundance and mRNA secondary structure support smoothing of translation elongation rate. *Nucleic Acids Research*, 43(6): 3022–3032. *Cited at page 21*

Gout, J.F. et al 2013. Large-scale detection of in vivo transcription errors. *Proceedings of the National Academy of Sciences*, 110(46): 18584–18589. Publisher: Proceedings of the National Academy of Sciences. *Cited at pages 52, 76, 93*

Gouy, M. and Gautier, C. 1982. Codon usage in bacteria: correlation with gene expressivity. *Nucleic Acids Research*, 10(22): 7055–7074. *Cited at pages 11, 45, 102*

Gouy, M. et al 1985. ACNUC–a portable retrieval system for nucleic acid sequence databases: logical and physical designs and usage. *Computer applications in the biosciences: CABIOS*, 1(3): 167–172. *Cited at page 34*

Gozashti, L. et al 2022. Transposable elements drive intron gain in diverse eukaryotes. *Proceedings of the National Academy of Sciences*, 119(48): e2209766119. Publisher: Proceedings of the National Academy of Sciences. *Cited at page 6*

Grantham, R. et al 1980a. Codon catalog usage and the genome hypothesis. *Nucleic Acids Research*, 8(1): r49–r62. *Cited at pages 11, 45, 102*

Grantham, R., Gautier, C. and Gouy, M. 1980b. Codon frequencies in 119 individual genes confirm consistent choices of degenerate bases according to genome type. *Nucleic Acids Research*, 8(9): 1893–1912. *Cited at pages 11, 45*

Graur, D. and Li, W.H.L. 2000. *Fundamentals of Molecular Evolution*. Oxford University Press, Oxford, New York, second edition. *Cited at page 52*

Graveley, B.R. 2001. Alternative splicing: increasing diversity in the proteomic world. *Trends in Genetics*, 17(2): 100–107. *Cited at pages 9, 41, 76, 132*

Grzybowska, E.A., Wilczynska, A. and Siedlecki, J.A. 2001. Regulatory Functions of 3′UTRs. *Biochemical and Biophysical Research Communications*, 288(2): 291–295. *Cited at page 6*

Guéguen, L. and Duret, L. 2018. Unbiased Estimate of Synonymous and Nonsynonymous Substitution Rates with Nonstationary Base Composition. *Molecular Biology and Evolution*, 35(3): 734–742. *Cited at pages 59, 97*

Guéguen, L. et al 2013. Bio++: efficient extensible libraries and tools for computational molecular evolution. *Molecular Biology and Evolution*, 30(8): 1745–1750. *Cited at pages 59, 97*

Hahn, M.W. and Wray, G.A. 2002. The g-value paradox. *Evolution and Development*, 4(2): 73–75. *Cited at page 7*

Hamid, F.M. and Makeyev, E.V. 2014. Emerging functions of alternative splicing coupled with nonsense-mediated decay. *Biochemical Society Transactions*, 42(4): 1168–1173. *Cited at pages 76, 77*

Hamm, G.H. and Cameron, G.N. 1986. The EMBL data library. *Nucleic Acids Research*, 14(1): 5–9. *Cited at page 35*

Hardy, G.H. 1908. *A course of pure mathematics*. University Press, Cambridge, [Eng.], first edition edition. Open Library ID: OL6113675M. *Cited at page 23*

Harigaya, Y. and Parker, R. 2016. Analysis of the association between codon optimality and mRNA stability in Schizosaccharomyces pombe. *BMC Genomics*, 17(1): 895. *Cited at page 11*

Harris, H. 1966. Enzyme polymorphisms in man. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, 164(995): 298–310. *Cited at page 18*

Harris, J.I., Sanger, F. and Naughton, M.A. 1956. Species differences in insulin. *Archives of Biochemistry and Biophysics*, 65(1): 427–438. *Cited at page 34*

Heather, J.M. and Chain, B. 2016. The sequence of sequencers: The history of sequencing DNA. *Genomics*, 107(1): 1–8. *Cited at page 30*

Hershberg, R. and Petrov, D.A. 2008. Selection on codon bias. *Annual Review of Genetics*, 42: 287–299. *Cited at pages 11, 46*

Hesper, B. and Hogeweg, P. 1970. Bioinformatica: een werkconcept. *Kameleon*, 1(6): 28–29. *Cited at page 33*

Hia, F. et al 2019. Codon bias confers stability to human mRNAs. *EMBO Reports*, 20(11): e48220. *Cited at page 11*

Hinrichs, A.S. et al 2006. The UCSC Genome Browser Database: update 2006. *Nucleic Acids Research*, 34(Database issue): D590–D598. *Cited at pages 97, 125*

Hogeweg, P. 2011. The Roots of Bioinformatics in Theoretical Biology. *PLOS Computational Biology*, 7(3): e1002021. Publisher: Public Library of Science. *Cited at page 33*

Hotaling, S., Kelley, J.L. and Frandsen, P.B. 2021. Toward a genome sequence for every animal: Where are we now? *Proceedings of the National Academy of Sciences*, 118(52): e2109019118. *Cited at page 4*

Hsu, S.N. and Hertel, K.J. 2009. Spliceosomes walk the line: splicing errors and their impact on cellular function. *RNA biology*, 6(5): 526–530. *Cited at page 76*

Huang, A.Z., Delaidelli, A. and Sorensen, P.H. 2020. RNA modifications in brain tumorigenesis. *Acta Neuropathologica Communications*, 8(1): 64. *Cited at page 133*

Huang, W. et al 2014. Natural variation in genome architecture among 205 Drosophila melanogaster Genetic Reference Panel lines. *Genome Research*, 24(7): 1193–1208. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab. *Cited at pages 85, 97, 125*

Huo, W. et al 2021. Miniaturized DNA Sequencers for Personal Use: Unreachable Dreams or Achievable Goals. *Frontiers in Nanotechnology*, 3. *Cited at page 32*

Husemann, M. et al 2016. Effective population size in ecology and evolution. *Heredity*, 117(4): 191–192. *Cited at page 23*

Hussmann, J.A. et al 2015. Understanding Biases in Ribosome Profiling Experiments Reveals Signatures of Translation Dynamics in Yeast. *PLOS Genetics*, 11(12): e1005732. Publisher: Public Library of Science. *Cited at page 46*

Hutchison, III, C.A. 2007. DNA sequencing: bench to bedside and beyond †. *Nucleic Acids Research*, 35(18): 6227–6237. *Cited at page 30*

Ikemura, T. 1981. Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the E. coli translational system. *Journal of Molecular Biology*, 151(3): 389–409. *Cited at pages 11, 12, 45, 46, 48*

Ikemura, T. 1985. Codon usage and tRNA content in unicellular and multicellular organisms. *Molecular Biology and Evolution*, 2(1): 13–34. *Cited at pages 11, 48, 102*

International Human Genome Sequencing Consortium 2004. Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011): 931–945. Number: 7011 Publisher: Nature Publishing Group. *Cited at pages 5, 31*

Jiang, W. and Chen, L. 2020. Alternative splicing: Human disease and quantitative analysis from high-throughput sequencing. *Computational and Structural Biotechnology Journal*, 19: 183–195. *Cited at page 132*

Johannsen, W. 1909. *Elemente der exakten erblichkeitslehre. Deutsche wesentlich erweiterte ausgabe in fünfundzwanzig vorlesungen*. G. Fischer, Jena. Pages: 1-530. *Cited at page 5*

John, S., Olas, J.J. and Mueller-Roeber, B. 2021. Regulation of alternative splicing in response to temperature variation in plants. *Journal of Experimental Botany*, 72(18): 6150–6163. *Cited at page 82*

Jou, W.M. et al 1972. Nucleotide Sequence of the Gene Coding for the Bacteriophage MS2 Coat Protein. *Nature*, 237(5350): 82–88. Number: 5350 Publisher: Nature Publishing Group. *Cited at page 30*

Kalnina, Z. et al 2005. Alterations of pre-mRNA splicing in cancer. *Genes, Chromosomes & Cancer*, 42(4): 342–357. *Cited at page 133*

Kanaya, S. et al 1999. Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of *Bacillus subtilis* tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis. *Gene*, 238(1): 143–155. *Cited at page 102*

Kanaya, S. et al 2001. Codon Usage and tRNA Genes in Eukaryotes: Correlation of Codon Usage Diversity with Translation Efficiency and with CG-Dinucleotide Usage as Assessed by Multivariate Analysis. *Journal of Molecular Evolution*, 53(4-5): 290–298. *Cited at pages 45, 46, 133*

Karsch-Mizrachi, I. et al 2012. The International Nucleotide Sequence Database Collaboration. *Nucleic Acids Research*, 40(D1): D33–D37. *Cited at page 35*

Kashi, K. et al 2016. Discovery and functional analysis of lncRNAs: Methodologies to investigate an uncharacterized transcriptome. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, 1859(1): 3–15. *Cited at page 5*

Keightley, P.D. and Jackson, B.C. 2018. Inferring the Probability of the Derived *vs.* the Ancestral Allelic State at a Polymorphic Site. *Genetics*, 209(3): 897–906. *Cited at page 125*

Keynes, M. and Cox, T. 2008. William Bateson, the rediscoverer of Mendel. *Journal of the Royal Society of Medicine*, 101(3): 104. *Cited at page 15*

Kim, D. et al 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology*, 14(4): 1–13. Number: 4 Publisher: BioMed Central. *Cited at page 35*

Kim, D. et al 2019. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology*, 37(8): 907–915. Number: 8 Publisher: Nature Publishing Group. *Cited at pages 35, 60, 94*

Kimura, M. 1962. On the Probability of Fixation of Mutant Genes in a Population. *Genetics*, 47(6): 713–719. *Cited at page 27*

Kimura, M. 1967. On the evolutionary adjustment of spontaneous mutation rates. *Genetics Research*, 9(1): 23–34. Publisher: Cambridge University Press. *Cited at page 19*

Kimura, M. 1968. Evolutionary Rate at the Molecular Level. *Nature*, 217(5129): 624–626. Number: 5129 Publisher: Nature Publishing Group. *Cited at page 18*

Kimura, M. 1977. Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. *Nature*, 267(5608): 275–276. Number: 5608 Publisher: Nature Publishing Group. *Cited at page 18*

Kimura, M. 1991. Recent development of the neutral theory viewed from the Wrightian tradition of theoretical population genetics. *Proceedings of the National Academy of Sciences*, 88(14): 5969–5973. *Cited at page 18*

Kimura, M. and Ohta, T. 1971. Protein Polymorphism as a Phase of Molecular Evolution. *Nature*, 229(5285): 467–469. *Cited at page 18*

Kimura, M., Maruyama, T. and Crow, J.F. 1963. The Mutation Load in Small Populations. *Genetics*, 48(10): 1303–1312. *Cited at pages 45, 52, 76*

King, J.L. and Jukes, T.H. 1969. Non-Darwinian Evolution. *Science*, 164(3881): 788–798. Publisher: American Association for the Advancement of Science. *Cited at page 18*

Kneale, G.G. and Kennard, O. 1984. The EMBL nucleotide sequence data library. *Biochemical Society Transactions*, 12(6): 1011–1014. *Cited at page 35*

Kováč, L. 2019. Lamarck and Darwin revisited. *EMBO Reports*, 20(4): e47922. *Cited at page 14*

Kozak, M. 1989. The scanning model for translation: an update. *The Journal of cell biology*, 108(2): 229–241. *Cited at page 10*

Kozlov, A.M. et al 2019. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics*, 35(21): 4453–4455. *Cited at pages 58, 96*

Kramer, E.B. and Farabaugh, P.J. 2007. The frequency of translational misreading errors in E. coli is largely determined by tRNA competition. *RNA*, 13(1): 87–96. *Cited at page 46*

Krause, J. et al 2007. The Derived FOXP2 Variant of Modern Humans Was Shared with Neandertals. *Current Biology*, 17(21): 1908–1912. *Cited at page 29*

Kryazhimskiy, S. and Plotkin, J.B. 2008. The Population Genetics of dN/dS. *PLoS Genetics*, 4(12). *Cited at pages 26, 78*

Kudla, G. et al 2006. High Guanine and Cytosine Content Increases mRNA Levels in Mammalian Cells. *PLOS Biology*, 4(6): e180. Publisher: Public Library of Science. *Cited at page 103*

Kudla, G. et al 2009. Coding-sequence determinants of gene expression in Escherichia coli. *Science (New York, N.Y.)*, 324(5924): 255–258. *Cited at page 11*

Kumar, S. and Subramanian, S. 2002. Mutation rates in mammalian genomes. *Proceedings of the National Academy of Sciences*, 99(2): 803–808. Publisher: Proceedings of the National Academy of Sciences. *Cited at page 22*

Kumar, S. et al 2022. TimeTree 5: An Expanded Resource for Species Divergence Times. *Molecular Biology and Evolution*, 39(8): msac174. *Cited at pages 55, 58*

Kuznetsov, D. et al 2023. OrthoDB v11: annotation of orthologs in the widest sampling of organismal diversity. *Nucleic Acids Research*, 51(D1): D445–D451. *Cited at page 36*

Köster, J. and Rahmann, S. 2012. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics*, 28(19): 2520–2522. *Cited at page 134*

Lai, C.J. et al 2023. Viral codon optimization on SARS-CoV-2 Spike boosts immunity in the development of COVID-19 mRNA vaccines. *Journal of Medical Virology*, 95(10): e29183. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jmv.29183. *Cited at page 133*

Lamarck, J.B.d.M.d...A.d.t. 1809. *Philosophie zoologique, ou Exposition des considérations relatives à l'histoire naturelle des animaux. Tome 1 / ... par J.-B.-P.-A. Lamarck,...* *Cited at pages 14, 29*

Lander, E.S. et al 2001. Initial sequencing and analysis of the human genome. *Nature*, 409(6822): 860–921. Number: 6822 Publisher: Nature Publishing Group. *Cited at pages 5, 7, 31*

Langmead, B. et al 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, 10(3): 1–10. Number: 3 Publisher: BioMed Central. *Cited at page 35*

Lee, D. et al 2021. Balancing selection maintains hyper-divergent haplotypes in Caenorhabditis elegans. *Nature Ecology & Evolution*, 5(6): 794–807. Number: 6 Publisher: Nature Publishing Group. *Cited at page 18*

Lefébure, T. et al 2017. Less effective selection leads to larger genomes. *Genome Research*, 27(6): 1016–1028. *Cited at pages 20, 40, 52, 131, 138*

Leigh, E.G. 1999. The modern synthesis, Ronald Fisher and creationism. *Trends in Ecology & Evolution*, 14(12): 495–498. *Cited at page 17*

Leinonen, R. et al 2011a. The European Nucleotide Archive. *Nucleic Acids Research*, 39(suppl_1): D28–D31. *Cited at page 35*

Leinonen, R., Sugawara, H. and Shumway, M. 2011b. The Sequence Read Archive. *Nucleic Acids Research*, 39(Database issue): D19–D21. *Cited at pages 54, 59, 93*

Leoni, G. et al 2011. Coding potential of the products of alternative splicing in human. *Genome Biology*, 12(1): R9. *Cited at page 43*

Lerner, M.R. et al 1980. Are snRNPs involved in splicing? *Nature*, 283(5743): 220–224. Number: 5743 Publisher: Nature Publishing Group. *Cited at page 8*

Leung, S.K. et al 2021. Full-length transcript sequencing of human and mouse cerebral cortex identifies widespread isoform diversity and alternative splicing. *Cell Reports*, 37(7): 110022. *Cited at page 92*

Levene, P.A. 1919. THE STRUCTURE OF YEAST NUCLEIC ACID: IV. AMMONIA HYDROLYSIS. *Journal of Biological Chemistry*, 40(2): 415–424. *Cited at page 30*

Levinson, S.C. and Gray, R.D. 2012. Tools from evolutionary biology shed new light on the diversification of languages. *Trends in Cognitive Sciences*, 16(3): 167–173. *Cited at page 29*

Lewontin, R.C. and Hubby, J.L. 1966. A MOLECULAR APPROACH TO THE STUDY OF GENIC HETEROZYGOSITY IN NATURAL POPULATIONS. II. AMOUNT OF VARIATION AND DEGREE OF HETEROZYGOSITY IN NATURAL POPULATIONS OF DROSOPHILA PSEUDOOBSCURA. *Genetics*, 54(2): 595–609. *Cited at page 18*

Li, M. et al 2022. De Novo Assembly of 20 Chicken Genomes Reveals the Undetectable Phenomenon for Thousands of Core Genes on Microchromosomes and Subtelomeric Regions. *Molecular Biology and Evolution*, 39(4): msac066. *Cited at page 67*

Li, S. et al 2014. Specialist versus generalist life histories and nucleotide diversity in Caenorhabditis nematodes. *Proceedings of the Royal Society B: Biological Sciences*, 281(1777): 20132858. Publisher: Royal Society. *Cited at page 122*

Li, W. and Lynch, M. 2020. Universally high transcript error rates in bacteria. *eLife*, 9: e54898. Publisher: eLife Sciences Publications, Ltd. *Cited at page 93*

Li, Z. et al 2020. MeDAS: a Metazoan Developmental Alternative Splicing database. *Nucleic Acids Research*, 49(D1): D144–D150. *Cited at pages 39, 53*

Lifton, R.P. et al 1978. The Organization of the Histone Genes in Drosophila melanogaster: Functional and Evolutionary Implications. *Cold Spring Harbor Symposia on Quantitative Biology*, 42: 1047–1051. Publisher: Cold Spring Harbor Laboratory Press. *Cited at page 7*

Lipman, D.J. and Pearson, W.R. 1985. Rapid and Sensitive Protein Similarity Searches. *Science*, 227(4693): 1435–1441. Publisher: American Association for the Advancement of Science. *Cited at page 34*

Liu, J. et al 2021a. MetazExp: a database for gene expression and alternative splicing profiles and their analyses based on 53 615 public RNA-seq samples in 72 metazoan species. *Nucleic Acids Research*, 50(D1): D1046–D1054. *Cited at pages 39, 53*

Liu, J. et al 2022. A web-based database server using 43,710 public RNA-seq samples for the analysis of gene expression and alternative splicing in livestock animals. *BMC Genomics*, 23(1): 706. *Cited at page 39*

Liu, Y. 2020. A code within the genetic code: codon usage regulates co-translational protein folding. *Cell Communication and Signaling*, 18(1): 145. *Cited at page 46*

Liu, Y., Yang, Q. and Zhao, F. 2021b. Synonymous but not Silent: The Codon Usage Code for Gene Expression and Protein Folding. *Annual review of biochemistry*, 90: 375–401. *Cited at page 11*

Liu, Z. and Zhang, J. 2018a. Human C-to-U Coding RNA Editing Is Largely Nonadaptive. *Molecular Biology and Evolution*, 35(4): 963–969. *Cited at pages 52, 93*

Liu, Z. and Zhang, J. 2018b. Most m6A RNA Modifications in Protein-Coding Regions Are Evolutionarily Unconserved and Likely Nonfunctional. *Molecular Biology and Evolution*, 35(3): 666–675. *Cited at pages 52, 93*

Lobry, J.R. 1995. Properties of a general model of DNA evolution under no-strand-bias conditions. *Journal of Molecular Evolution*, 40(3): 326–330. *Cited at page 104*

Logsdon, G.A., Vollger, M.R. and Eichler, E.E. 2020. Long-read human genome sequencing and its applications. *Nature Reviews Genetics*, 21(10): 597–614. Number: 10 Publisher: Nature Publishing Group. *Cited at pages 33, 136*

Long, H. et al 2018. Evolutionary determinants of genome-wide nucleotide composition. *Nature Ecology & Evolution*, 2(2): 237–240. *Cited at page 102*

Luo, X., Kang, X. and Schönhuth, A. 2022. VeChat: correcting errors in long reads using variation graphs. *Nature Communications*, 13(1): 6657. Number: 1 Publisher: Nature Publishing Group. *Cited at page 136*

Lynch, M. 2006. The Origins of Eukaryotic Gene Structure. *Molecular Biology and Evolution*, 23(2): 450–468. *Cited at page 76*

Lynch, M. 2007a. The frailty of adaptive hypotheses for the origins of organismal complexity. *Proceedings of the National Academy of Sciences*, 104(suppl_1): 8597–8604. Publisher: Proceedings of the National Academy of Sciences. *Cited at pages 19, 24, 44, 45, 52, 76, 131*

Lynch, M. 2007b. *The Origins of Genome Architecture*. Sinauer Associates Inc, Sunderland, 1st edition edition. *Cited at page 24*

Lynch, M. 2008. The cellular, developmental and population-genetic determinants of mutation-rate evolution. *Genetics*, 180(2): 933–943. *Cited at page 19*

Lynch, M. 2010. Evolution of the mutation rate. *Trends in genetics : TIG*, 26(8): 345–352. *Cited at pages 19, 22, 24, 52, 131*

Lynch, M. and Conery, J.S. 2003. The origins of genome complexity. *Science (New York, N.Y.)*, 302(5649): 1401–1404. *Cited at pages 20, 52, 77*

Lynch, M. et al 2016. Genetic drift, selection and the evolution of the mutation rate. *Nature Reviews Genetics*, 17(11): 704–714. Number: 11 Publisher: Nature Publishing Group. *Cited at pages 52, 89, 131*

Lynch, M. et al 2023. The divergence of mutation rates and spectra across the Tree of Life. *EMBO reports*, 24(10): e57561. Publisher: John Wiley & Sons, Ltd. *Cited at pages 25, 103, 114, 122, 126, 137, 138*

Löytynoja, A. and Goldman, N. 2008. Phylogeny-Aware Gap Placement Prevents Errors in Sequence Alignment and Evolutionary Analysis. *Science*, 320(5883): 1632–1635. Publisher: American Association for the Advancement of Science. *Cited at pages 57, 58, 96*

Machado, H.E., Lawrie, D.S. and Petrov, D.A. 2020. Pervasive Strong Selection at the Level of Codon Usage Bias in *Drosophila melanogaster*. *Genetics*, 214(2): 511–528. *Cited at page 123*

Mackay, T.F.C. et al 2012. The Drosophila melanogaster Genetic Reference Panel. *Nature*, 482(7384): 173–178. Number: 7384 Publisher: Nature Publishing Group. *Cited at pages 85, 97, 125*

Maki, H. 2002. Origins of Spontaneous Mutations: Specificity and Directionality of Base-Substitution, Frameshift, and Sequence-Substitution Mutageneses. *Annual Review of Genetics*, 36(1): 279–303. *Cited at page 21*

Manni, M. et al 2021. BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Molecular Biology and Evolution*, 38(10): 4647–4654. *Cited at pages 36, 57*

Manthey, C. et al 2024. Rapid growth and the evolution of complete metamorphosis in insects. Pages: 2024.02.12.579885 Section: New Results. *Cited at pages 123, 131*

Manuel, J.M. et al 2023. Re-evaluating the impact of alternative RNA splicing on proteomic diversity. *Frontiers in Genetics*, 14. *Cited at page 133*

Marino, A. et al 2024. Effective population size does not explain long-term variation in genome size and transposable element content in animals. Pages: 2024.02.26.582137 Section: New Results. *Cited at pages 131, 132, 139*

Martínez, M.A. et al 2019. Synonymous genome recoding: a tool to explore microbial biology and new therapeutic strategies. *Nucleic Acids Research*, 47(20): 10506–10519. *Cited at pages 11, 133*

Marx, V. 2023. Method of the year: long-read sequencing. *Nature Methods*, 20(1): 6–11. Number: 1 Publisher: Nature Publishing Group. *Cited at page 33*

Mayr, C. 2019. What Are 3′ UTRs Doing? *Cold Spring Harbor Perspectives in Biology*, 11(10): a034728. *Cited at page 6*

Mazin, P.V. et al 2021. Alternative splicing during mammalian organ development. *Nature Genetics*, 53(6): 925–934. Number: 6 Publisher: Nature Publishing Group. *Cited at pages 41, 77, 82, 89*

McClintock, B. 1950. The origin and behavior of mutable loci in maize. *Proceedings of the National Academy of Sciences*, 36(6): 344–355. Publisher: Proceedings of the National Academy of Sciences. *Cited at page 20*

McGee, M.D. et al 2020. The ecological and genomic basis of explosive adaptive radiation. *Nature*, 586(7827): 75–79. Number: 7827 Publisher: Nature Publishing Group. *Cited at page 20*

McGlincy, N.J. and Smith, C.W.J. 2008. Alternative splicing resulting in nonsense-mediated mRNA decay: what is the meaning of nonsense? *Trends in Biochemical Sciences*, 33(8): 385–393. *Cited at pages 76, 77*

Medina-Muñoz, S.G. et al 2021. Crosstalk between codon optimality and cis-regulatory elements dictates mRNA stability. *Genome Biology*, 22(1): 1–23. Number: 1 Publisher: BioMed Central. *Cited at page 103*

Melamud, E. and Moult, J. 2009. Stochastic noise in splicing machinery. *Nucleic Acids Research*, 37(14): 4873–4886. *Cited at page 43*

Mendel, G. 1865. Versuche über Pflanzen-Hybriden. *Verhandlungen des naturforschenden Vereines in Brünn*, Bd.4 (1865-1866): 3–47. *Cited at page 15*

Mendel, G., Punnett, R.C. and Burndy Library, d.D. 1866. *Versuche über Pflanzen-Hybriden*. Brünn : Im Verlage des Vereines. *Cited at page 5*

Merkin, J. et al 2012. Evolutionary Dynamics of Gene and Isoform Regulation in Mammalian Tissues. *Science*, 338(6114): 1593–1599. *Cited at pages 52, 76, 92*

Metzker, M.L. 2005. Emerging technologies in DNA sequencing. *Genome Research*, 15(12): 1767–1776. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab. *Cited at page 30*

Meyer, A. et al 2021. Giant lungfish genome elucidates the conquest of land by vertebrates. *Nature*, 590(7845): 284–289. Number: 7845 Publisher: Nature Publishing Group. *Cited at page 4*

Mighell, A. et al 2000. Vertebrate pseudogenes. *FEBS Letters*, 468(2-3): 109–114. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1016/S0014-5793%2800%2901199-6. *Cited at page 5*

Mittal, P. et al 2018. Codon usage influences fitness through RNA toxicity. *Proceedings of the National Academy of Sciences*, 115(34): 8639–8644. Publisher: Proceedings of the National Academy of Sciences. *Cited at page 11*

Miyata, T., Miyazawa, S. and Yasunaga, T. 1979. Two types of amino acid substitutions in protein evolution. *Journal of Molecular Evolution*, 12(3): 219–236. *Cited at page 21*

Mordstein, C. et al 2020. Codon Usage and Splicing Jointly Influence mRNA Localization. *Cell Systems*, 10(4): 351–362.e8. Publisher: Elsevier. *Cited at page 103*

Morgan, T.H. 1915. *The mechanism of Mendelian heredity*. Holt, New York. Pages: 1-288. *Cited at page 15*

Morgan, T.H. 1925. *Evolution and genetics, 2nd ed.* Evolution and genetics, 2nd ed. Princeton University Press, Princeton, NJ, US. Pages: ix, 211. *Cited at page 15*

Morisse, P. et al 2021. Scalable long read self-correction and assembly polishing with multiple sequence alignment. *Scientific Reports*, 11(1): 761. Number: 1 Publisher: Nature Publishing Group. *Cited at page 136*

Moriyama, E.N. and Powell, J.R. 1997. Codon Usage Bias and tRNA Abundance in Drosophila. *Journal of Molecular Evolution*, 45(5): 514–523. *Cited at page 102*

Morris, C., Cluet, D. and Ricci, E.P. 2021. Ribosome dynamics and mRNA turnover, a complex relationship under constant cellular scrutiny. *Wiley Interdisciplinary Reviews. RNA*, 12(6): e1658. *Cited at page 46*

Mortazavi, A. et al 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, 5(7): 621–628. Number: 7 Publisher: Nature Publishing Group. *Cited at page 35*

Mouchiroud, D., Gautier, C. and Bernardi, G. 1988. The compositional distribution of coding sequences and DNA molecules in humans and murids. *Journal of Molecular Evolution*, 27(4): 311–320. *Cited at pages 45, 46, 133*

Mouchiroud, D. et al 1991. The distribution of genes in the human genome. *Gene*, 100: 181–187. *Cited at page 46*

Mount, S.M. and Wolin, S.L. 2015. Recognizing the 35th anniversary of the proposal that snRNPs are involved in splicing. *Molecular Biology of the Cell*, 26(20): 3557–3560. *Cited at page 8*

Mudge, J.M. et al 2011. The Origins, Evolution, and Functional Potential of Alternative Splicing in Vertebrates. *Molecular Biology and Evolution*, 28(10): 2949–2959. *Cited at pages 52, 92*

Mugal, C.F. 2021. A systematic approach to the study of GC-biased gene conversion in mammals. *Peer Community in Genomics*, 1: 100012. Company: Peer Community in Genomics Distributor: Peer Community in Genomics Institution: Peer Community in Genomics Label: Peer Community in Genomics Publisher: Peer Community In. *Cited at page 28*

Mugal, C.F., Wolf, J.B. and Kaj, I. 2014. Why Time Matters: Codon Evolution and the Temporal Dynamics of dN/dS. *Molecular Biology and Evolution*, 31(1): 212–231. *Cited at page 72*

Mukhopadhyay, R. 2009. DNA sequencers: the next generation. *Analytical Chemistry*, 81(5): 1736–1740. Publisher: American Chemical Society. *Cited at pages 30, 32*

Mullaney, J.M. et al 2010. Small insertions and deletions (INDELs) in human genomes. *Human Molecular Genetics*, 19(R2): R131–R136. *Cited at page 21*

Mullin, E. 2022. The Era of Fast, Cheap Genome Sequencing Is Here. *Wired*. Section: tags. *Cited at page 32*

Myers, P. et al 2023. The Animal Diversity Web (online). URL https://animaldiversity.org. Accessed on 8/24/2023. *Cited at pages 37, 40, 54*

Mérel, V. et al 2024. Relaxed purifying selection is associated with an accumulation of transposable elements in flies. preprint, Evolutionary Biology. *Cited at page 20*

Mölder, F. et al 2021. Sustainable data analysis with Snakemake. Technical Report 10:33, F1000Research. Type: article. *Cited at page 54*

Nabholz, B., Ellegren, H. and Wolf, J. 2012. High Levels of Gene Expression Explain the Strong Evolutionary Constraint of Mitochondrial Protein-Coding Genes. *Molecular biology and evolution*. *Cited at page 23*

NCBI Resource Coordinators 2018. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, 46(D1): D8–D13. *Cited at pages 32, 35, 54, 93*

Nee, S. et al 1991. The relationship between abundance and body size in British birds. *Nature*, 351(6324): 312–313. Number: 6324 Publisher: Nature Publishing Group. *Cited at page 70*

Needleman, S.B. and Wunsch, C.D. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3): 443–453. *Cited at page 34*

Nei, M. 1969. Gene Duplication and Nucleotide Substitution in Evolution. *Nature*, 221(5175): 40–42. Number: 5175 Publisher: Nature Publishing Group. *Cited at page 18*

Nei, M. 1984. Genetic Polymorphism and Neomutationism. In G. S. Mani, editor, *Evolutionary Dynamics of Genetic Diversity*, Lecture Notes in Biomathematics, pages 214–241, Berlin, Heidelberg. Springer. *Cited at page 18*

Nei, M. 2005. Selectionism and Neutralism in Molecular Evolution. *Molecular biology and evolution*, 22(12): 2318–2342. *Cited at pages 14, 18*

Nei, M. and Nozawa, M. 2011. Roles of Mutation and Selection in Speciation: From Hugo de Vries to the Modern Genomic Era. *Genome Biology and Evolution*, 3: 812–829. *Cited at page 20*

Nei, M., Suzuki, Y. and Nozawa, M. 2010. The neutral theory of molecular evolution in the genomic era. *Annual Review of Genomics and Human Genetics*, 11: 265–289. *Cited at page 19*

Nettle, D. and Harriss, L. 2003. Genetic and linguistic affinities between human populations in Eurasia and West Africa. *Human Biology*, 75(3): 331–344. *Cited at page 29*

Neville, S. 2018. Cheaper DNA sequencing unlocks secrets of rare diseases. *Financial Times*. *Cited at page 32*

Nicholson, A.L. and Pasquinelli, A.E. 2019. Tales of Detailed Poly(A) Tails. *Trends in cell biology*, 29(3): 191–200. *Cited at page 8*

Nielsen, R. and Yang, Z. 2003. Estimating the Distribution of Selection Coefficients from Phylogenetic Data with Applications to Mitochondrial and Viral DNA. *Molecular Biology and Evolution*, 20(8): 1231–1239. *Cited at pages 27, 57*

Nikolov, D. and Burley, S. 1997. RNA polymerase II transcription initiation: A structural view. *Proceedings of the National Academy of Sciences*, 94(1): 15–22. Publisher: Proceedings of the National Academy of Sciences. *Cited at page 7*

Nilsen, T.W. and Graveley, B.R. 2010. Expansion of the eukaryotic proteome by alternative splicing. *Nature*, 463(7280): 457–463. *Cited at pages 42, 132*

O'Brien, J. et al 2018. Overview of MicroRNA Biogenesis, Mechanisms of Actions, and Circulation. *Frontiers in Endocrinology*, 9. *Cited at page 5*

Ohta, T. 1973. Slightly Deleterious Mutant Substitutions in Evolution. *Nature*, 246(5428): 96–98. Number: 5428 Publisher: Nature Publishing Group. *Cited at pages 18, 45, 52, 76*

Ohta, T. 1992. The Nearly Neutral Theory of Molecular Evolution. *Annual Review of Ecology and Systematics*, 23: 263–286. Publisher: Annual Reviews. *Cited at pages 18, 57, 126*

Ohta, T. 1996. The neutral theory is dead. The current significance and standing of neutral and nearly neutral theories. *BioEssays*, 18(8): 673–677. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/bies.950180811. *Cited at page 103*

Oliver, S.G. et al 1992. The complete DNA sequence of yeast chromosome III. *Nature*, 357(6373): 38–46. Number: 6373 Publisher: Nature Publishing Group. *Cited at page 31*

O'Hara, R. 2005. Comparing the effects of genetic drift and fluctuating selection on genotype frequency changes in the scarlet tiger moth. *Proceedings of the Royal Society B: Biological Sciences*, 272(1559): 211–217. *Cited at page 17*

Pagani, F., Raponi, M. and Baralle, F.E. 2005. Synonymous mutations in CFTR exon 12 affect splicing and are not neutral in evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 102(18): 6368–6372. *Cited at page 11*

Pagel, M. 2013. *Wired for Culture – Origins of the Human Social Mind*. W. W. Norton & Company, New York, NY, reprint édition edition. *Cited at page 29*

Pagán, I., Holmes, E.C. and Simon-Loriere, E. 2012. Level of Gene Expression Is a Major Determinant of Protein Evolution in the Viral Order Mononegavirales. *Journal of Virology*, 86(9): 5253. Publisher: American Society for Microbiology (ASM). *Cited at pages 23, 27*

Palade, G.E. 1955. A SMALL PARTICULATE COMPONENT OF THE CYTOPLASM. *The Journal of Biophysical and Biochemical Cytology*, 1(1): 59–68. *Cited at page 10*

Palstra, F.P. and Fraser, D.J. 2012. Effective/census population size ratio estimation: a compendium and appraisal. *Ecology and Evolution*, 2(9): 2357–2365. *Cited at page 24*

Palstra, F.P. and Ruzzante, D.E. 2008. Genetic estimates of contemporary effective population size: what can they tell us about the importance of genetic stochasticity for wild population persistence? *Molecular Ecology*, 17(15): 3428–3447. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1365-294X.2008.03842.x. *Cited at page 24*

Pan, Q. et al 2008. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature Genetics*, 40(12): 1413–1415. *Cited at pages 9, 10, 41, 132*

Park, C. et al 2013. Differential requirements for mRNA folding partially explain why highly expressed proteins evolve slowly. *Proceedings of the National Academy of Sciences*, 110(8): E678–E686. Publisher: Proceedings of the National Academy of Sciences. *Cited at page 23*

Parmley, J.L. and Hurst, L.D. 2007. Exonic Splicing Regulatory Elements Skew Synonymous Codon Usage near Intron-exon Boundaries in Mammals. *Molecular Biology and Evolution*, 24(8): 1600–1603. *Cited at page 103*

Parr, C.S. et al 2014. The Encyclopedia of Life v2: Providing Global Access to Knowledge About Life on Earth. *Biodiversity Data Journal*, 2: e1079. Publisher: Pensoft Publishers. *Cited at pages 37, 40, 54*

Parvathy, S.T., Udayasuriyan, V. and Bhadana, V. 2022. Codon usage bias. *Molecular Biology Reports*, 49(1): 539–565. *Cited at pages 11, 45*

Passmore, L.A. and Coller, J. 2022. Roles of mRNA poly(A) tails in regulation of eukaryotic gene expression. *Nature reviews. Molecular cell biology*, 23(2): 93–106. *Cited at page 8*

Percudani, R. 2001. Restricted wobble rules for eukaryotic genomes. *Trends in Genetics*, 17(3): 133–135. Publisher: Elsevier. *Cited at pages 10, 46, 108*

Percudani, R., Pavesi, A. and Ottonello, S. 1997. Transfer RNA gene redundancy and translational selection in Saccharomyces cerevisiae11Edited by J. Karn. *Journal of Molecular Biology*, 268(2): 322–330. *Cited at page 12*

Pickrell, J.K. et al 2010. Noisy Splicing Drives mRNA Isoform Diversity in Human Cells. *PLOS Genetics*, 6(12): e1001236. Publisher: Public Library of Science. *Cited at pages 10, 43, 52, 76, 77, 93, 133*

Piovesan, A. et al 2019a. Human protein-coding genes and gene feature statistics in 2019. *BMC Research Notes*, 12(1): 315. *Cited at pages 6, 7*

Piovesan, A. et al 2019b. On the length, weight and GC content of the human genome. *BMC Research Notes*, 12: 106. *Cited at page 4*

Plotkin, J.B. and Kudla, G. 2011. Synonymous but not the same: the causes and consequences of codon bias. *Nature Reviews Genetics*, 12(1): 32–42. Number: 1 Publisher: Nature Publishing Group. *Cited at pages 11, 12, 21, 45*

Plotkin, J.B., Robins, H. and Levine, A.J. 2004. Tissue-specific codon usage and the expression of human genes. *Proceedings of the National Academy of Sciences*, 101(34): 12588–12591. Publisher: Proceedings of the National Academy of Sciences. *Cited at page 47*

Plotkin, J.B. et al 2006. Codon Usage and Selection on Proteins. *Journal of Molecular Evolution*, 63(5): 635–653. *Cited at page 12*

Potapova, N.A. 2022. Nonsense Mutations in Eukaryotes. *Biochemistry (Moscow)*, 87(5): 400–412. *Cited at page 20*

Pouyet, F. et al 2017. Recombination, meiotic expression and human codon usage. *eLife*, 6: e27344. Publisher: eLife Sciences Publications, Ltd. *Cited at pages 12, 46, 47, 102, 103, 118, 120, 122, 133*

Pozo, F. et al 2021. Assessing the functional relevance of splice isoforms. *NAR Genomics and Bioinformatics*, 3(2): lqab044. *Cited at page 133*

Preiss, T. 2013. The End in Sight: Poly(A), Translation and mRNA Stability in Eukaryotes. In *Madame Curie Bioscience Database [Internet]*. Landes Bioscience. *Cited at page 8*

Presnyak, V. et al 2015. Codon Optimality Is a Major Determinant of mRNA Stability. *Cell*, 160(6): 1111–1124. *Cited at pages 11, 21*

Proudfoot, N.J. 2016. Transcriptional termination in mammals: Stopping the RNA polymerase II juggernaut. *Science*, 352(6291): aad9926. Publisher: American Association for the Advancement of Science. *Cited at page 7*

Proudfoot, N.J., Furger, A. and Dye, M.J. 2002. Integrating mRNA Processing with Transcription. *Cell*, 108(4): 501–512. Publisher: Elsevier. *Cited at page 8*

Pál, C., Papp, B. and Hurst, L.D. 2001. Highly Expressed Genes in Yeast Evolve Slowly. *Genetics*, 158(2): 927–931. *Cited at page 23*

Qi, F. et al 2020. Significance of alternative splicing in cancer cells. *Chinese Medical Journal*, 133(2): 221–228. *Cited at pages 9, 133*

Quax, T.E.F. et al 2015. Codon Bias as a Means to Fine-Tune Gene Expression. *Molecular Cell*, 59(2): 149–161. Publisher: Elsevier. *Cited at pages 12, 21*

Quinlan, A.R. and Hall, I.M. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics (Oxford, England)*, 26(6): 841–842. *Cited at page 62*

Rajon, E. and Masel, J. 2011. Evolution of molecular error rates and the consequences for evolvability. *Proceedings of the National Academy of Sciences of the United States of America*, 108(3): 1082–1087. *Cited at pages 10, 93*

Reddy, A. et al 2012. Deciphering the Plant Splicing Code: Experimental and Computational Approaches for Predicting Alternative Splicing and Splicing Regulatory Elements. *Frontiers in Plant Science*, 3. *Cited at page 139*

Reddy, A.S. et al 2013. Complexity of the Alternative Splicing Landscape in Plants. *The Plant Cell*, 25(10): 3657–3683. *Cited at page 139*

Ren, P. et al 2021. Alternative Splicing: A New Cause and Potential Therapeutic Target in Autoimmune Disease. *Frontiers in Immunology*, 12. *Cited at page 133*

Reyes, A. et al 2013. Drift and conservation of differential exon usage across tissues in primate species. *Proceedings of the National Academy of Sciences*, 110(38): 15377–15382. Publisher: Proceedings of the National Academy of Sciences. *Cited at pages 52, 92*

Riede, F. 2010. Why isn't archaeology (more) Darwinian? A historical perspective. *Journal of Evolutionary Psychology*, 8: 183–204. *Cited at page 14*

Rigden, D.J. and Fernández, X.M. 2018. The 2018 Nucleic Acids Research database issue and the online molecular biology database collection. *Nucleic Acids Research*, 46(D1): D1–D7. *Cited at page 37*

Rigden, D.J. and Fernández, X.M. 2023. The 2023 Nucleic Acids Research Database Issue and the online molecular biology database collection. *Nucleic Acids Research*, 51(D1): D1–D8. *Cited at page 37*

Roach, J.C. et al 1995. Pairwise end sequencing: a unified approach to genomic mapping and sequencing. *Genomics*, 26(2): 345–353. *Cited at page 31*

Roberts, A. et al 2011. Identification of novel transcripts in annotated genomes using RNA-Seq. *Bioinformatics*, 27(17): 2325–2329. *Cited at pages 62, 96*

Rocha, E.P. 2004. Codon usage bias from tRNA's point of view: Redundancy, specialization, and efficient decoding for translation optimization. *Genome Research*, 14(11): 2279–2286. *Cited at pages 102, 123, 131*

Rocha, E.P.C. and Danchin, A. 2004. An Analysis of Determinants of Amino Acids Substitution Rates in Bacterial Proteins. *Molecular Biology and Evolution*, 21(1): 108–116. *Cited at pages 23, 27*

Roddy, A.B., Alvarez-Ponce, D. and Roy, S.W. 2021. Mammals with Small Populations Do Not Exhibit Larger Genomes. *Molecular Biology and Evolution*, 38(9): 3737. Publisher: Oxford University Press. *Cited at pages 131, 132, 139*

Rodriguez-Tomé, P. et al 1996. The European Bioinformatics Institute (EBI) databases. *Nucleic Acids Research*, 24(1): 6–12. *Cited at page 35*

Roest Crollius, H. et al 2000. Estimate of human gene number provided by genome-wide analysis using Tetraodon nigroviridis DNA sequence. *Nature Genetics*, 25(2): 235–238. Publisher: Nature Publishing Group. *Cited at page 6*

Romiguier, J. et al 2010. Contrasting GC-content dynamics across 33 mammalian genomes: Relationship with life-history traits and chromosome sizes. *Genome Research*, 20(8): 1001–1009. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab. *Cited at page 28*

Romiguier, J. et al 2012. Fast and Robust Characterization of Time-Heterogeneous Sequence Evolutionary Processes Using Substitution Mapping. *PLOS ONE*, 7(3): e33852. *Cited at page 40*

Romiguier, J. et al 2014a. Comparative population genomics in animals uncovers the determinants of genetic diversity. *Nature*, 515(7526): 261–263. *Cited at pages 54, 66, 67*

Romiguier, J. et al 2014b. Population genomics of eusocial insects: the costs of a vertebrate-like effective population size. *Journal of Evolutionary Biology*, 27(3): 593–603. *Cited at page 28*

Ross, J. 1995. mRNA stability in mammalian cells. *Microbiological Reviews*, 59(3): 423–450. *Cited at pages 7, 8*

Rudolph, K.L.M. et al 2016. Codon-Driven Translational Efficiency Is Stable across Diverse Mammalian Cell States. *PLOS Genetics*, 12(5): e1006024. Publisher: Public Library of Science. *Cited at page 47*

Rédei, G.P. 2008. Selection Coefficient. In *Encyclopedia of Genetics, Genomics, Proteomics and Informatics*, pages 1775–1775. Springer Netherlands, Dordrecht. *Cited at page 22*

Saitou, N. and Nei, M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4(4): 406–425. *Cited at page 34*

Sakharkar, M.K., Chow, V.T.K. and Kangueane, P. 2004. Distributions of exons and introns in the human genome. *In Silico Biology*, 4(4): 387–393. *Cited at page 5*

Sanger, F. 1949. Species Differences in Insulins. *Nature*, 164(4169): 529–529. Number: 4169 Publisher: Nature Publishing Group. *Cited at page 34*

Sanger, F. and Thompson, E.O.P. 1953a. The amino-acid sequence in the glycyl chain of insulin. 1. The identification of lower peptides from partial hydrolysates. *Biochemical Journal*, 53(3): 353–366. *Cited at page 33*

Sanger, F. and Thompson, E.O.P. 1953b. The amino-acid sequence in the glycyl chain of insulin. 2. The investigation of peptides from enzymic hydrolysates. *Biochemical Journal*, 53(3): 366–374. *Cited at page 33*

Sanger, F. and Tuppy, H. 1951a. The amino-acid sequence in the phenylalanyl chain of insulin. 1. The identification of lower peptides from partial hydrolysates. *Biochemical Journal*, 49(4): 463–481. *Cited at page 33*

Sanger, F. and Tuppy, H. 1951b. The amino-acid sequence in the phenylalanyl chain of insulin. 2. The investigation of peptides from enzymic hydrolysates. *Biochemical Journal*, 49(4): 481–490. *Cited at page 33*

Sanger, F., Nicklen, S. and Coulson, A.R. 1977a. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, 74(12): 5463–5467. *Cited at page 30*

Sanger, F. et al 1977b. Nucleotide sequence of bacteriophage phi X174 DNA. *Nature*, 265(5596): 687–695. *Cited at page 30*

Saudemont, B. et al 2017. The fitness cost of mis-splicing is the main determinant of alternative splicing patterns. *Genome Biology*, 18. *Cited at pages 10, 23, 43, 45, 52, 78, 87, 93, 94, 133*

Sayers, E.W. et al 2022a. Database resources of the national center for biotechnology information. *Nucleic Acids Research*, 50(D1): D20–D26.
Cited at pages 53, 54, 56, 124

Sayers, E.W. et al 2022b. GenBank. *Nucleic Acids Research*, 50(D1): D161–D164.
Cited at page 35

Schad, E., Tompa, P. and Hegyi, H. 2011. The relationship between proteome size, structural disorder and organism complexity. *Genome Biology*, 12(12): R120.
Cited at page 42

Schmitt, B.M. et al 2014. High-resolution mapping of transcriptional dynamics across tissue development reveals a stable mRNA-tRNA interface. *Genome Research*, 24(11): 1797–1807. *Cited at page 47*

Schmitz, U. et al 2020. Widespread Aberrant Alternative Splicing despite Molecular Remission in Chronic Myeloid Leukaemia Patients. *Cancers*, 12(12): 3738.
Cited at page 133

Schulz, R. et al 2022. Is the future of peer review automated? *BMC Research Notes*, 15(1): 203. *Cited at page 134*

Schwarze, K. et al 2020. The complete costs of genome sequencing: a microcosting study in cancer and rare diseases from a single center in the United Kingdom. *Genetics in Medicine*, 22(1): 85–94. Number: 1 Publisher: Nature Publishing Group.
Cited at page 32

Sciarrillo, R. et al 2020. The role of alternative splicing in cancer: From oncogenesis to drug resistance. *Drug Resistance Updates*, 53: 100728. *Cited at pages 10, 133*

Scotti, M.M. and Swanson, M.S. 2016. RNA mis-splicing in disease. *Nature Reviews Genetics*, 17(1): 19–32. Number: 1 Publisher: Nature Publishing Group.
Cited at page 133

Seppey, M., Manni, M. and Zdobnov, E.M. 2019. BUSCO: Assessing Genome Assembly and Annotation Completeness. *Methods in Molecular Biology (Clifton, N.J.)*, 1962: 227–245. *Cited at pages 57, 78, 94*

Sevim, V. et al 2019. Shotgun metagenome data of a defined mock community using Oxford Nanopore, PacBio and Illumina technologies. *Scientific Data*, 6(1): 285. Number: 1 Publisher: Nature Publishing Group. *Cited at page 31*

Sharp, P.A. and Burge, C.B. 1997. Classification of Introns: U2-Type or U12-Type. *Cell*, 91(7): 875–879. Publisher: Elsevier. *Cited at page 8*

Sharp, P.M., Tuohy, T.M. and Mosurski, K.R. 1986. Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. *Nucleic Acids Research*, 14(13): 5125–5143. *Cited at page 102*

Sharp, P.M. et al 1988. Codon usage patterns in Escherichia coli, Bacillus subtilis, Saccharomyces cerevisiae, Schizosaccharomyces pombe, Drosophila melanogaster and Homo sapiens; a review of the considerable within-species diversity. *Nucleic Acids Research*, 16(17): 8207–8211. *Cited at pages 45, 46*

Sharp, P.M. et al 1993. Codon usage: mutational bias, translational selection, or both? *Biochemical Society Transactions*, 21(4): 835–841. *Cited at pages 102, 104*

Sharp, P.M. et al 2005. Variation in the strength of selected codon usage bias among bacteria. *Nucleic Acids Research*, 33(4): 1141–1153. *Cited at pages 102, 103, 106, 113, 120, 121, 123*

Shields, D.C. et al 1988. "Silent" sites in Drosophila genes are not neutral: evidence of selection among synonymous codons. *Molecular Biology and Evolution*, 5(6): 704–716. *Cited at page 103*

Simoni, R.D., Hill, R.L. and Vaughan, M. 2002. The Structure of Nucleic Acids and Many Other Natural Products: Phoebus Aaron Levene. *Journal of Biological Chemistry*, 277(22): e11–e12. *Cited at page 30*

Singh, P. and Ahi, E.P. 2022. The importance of alternative splicing in adaptive evolution. *Molecular Ecology*, 31(7): 1928–1938. Publisher: John Wiley & Sons, Ltd. *Cited at pages 9, 44, 52, 77, 132, 133*

Singh, T. et al 2021. Genome recoding strategies to improve cellular properties: mechanisms and advances. *aBIOTECH*, 2(1): 79–95. *Cited at page 133*

Slatko, B.E., Gardner, A.F. and Ausubel, F.M. 2018. Overview of Next Generation Sequencing Technologies. *Current protocols in molecular biology*, 122(1): e59. *Cited at page 31*

Slyusarev, G.S. et al 2020. Extreme Genome and Nervous System Streamlining in the Invertebrate Parasite Intoshia variabili. *Current biology: CB*, 30(7): 1292–1298.e3. *Cited at page 4*

Smith, T.F. and Waterman, M.S. 1981. Identification of common molecular subsequences. *Journal of Molecular Biology*, 147(1): 195–197. *Cited at page 34*

Sneath, P.H.A. 1966. Relations between chemical structure and biological activity in peptides. *Journal of Theoretical Biology*, 12(2): 157–195. *Cited at page 21*

Somel, M. et al 2011. MicroRNA-Driven Developmental Remodeling in the Brain Distinguishes Humans from Other Primates. *PLOS Biology*, 9(12): e1001214. Publisher: Public Library of Science. *Cited at page 29*

Sprinzl, M. and Cramer, F. 1979. The -C-C-A end of tRNA and its role in protein biosynthesis. *Progress in Nucleic Acid Research and Molecular Biology*, 22: 1–69. *Cited at page 10*

Srebrow, A. and Kornblihtt, A.R. 2006. The connection between splicing and cancer. *Journal of Cell Science*, 119(Pt 13): 2635–2641. *Cited at page 133*

Stadler, M. and Fire, A. 2011. Wobble base-pairing slows in vivo translation elongation in metazoans. *RNA*, 17(12): 2063–2073. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab. *Cited at pages 108, 121*

Stern, C. 1943. The Hardy-Weinberg Law. *Science*, 97(2510): 137–138. Publisher: American Association for the Advancement of Science. *Cited at page 23*

Stoletzki, N. and Eyre-Walker, A. 2007. Synonymous codon usage in Escherichia coli: selection for translational accuracy. *Molecular Biology and Evolution*, 24(2): 374–381. *Cited at pages 11, 21, 46*

Straalen, N.M.v. et al 2011. *An Introduction to Ecological Genomics*. Oxford University Press, Oxford, New York, second edition, second edition edition. *Cited at page 7*

Subramanian, S. 2008. Nearly Neutrality and the Evolution of Codon Usage Bias in Eukaryotic Genomes. *Genetics*, 178(4): 2429–2432. *Cited at pages 103, 104*

Sun, F.J. et al 2007. Common evolutionary trends for SINE RNA structures. *Trends in Genetics*, 23(1): 26–33. Publisher: Elsevier. *Cited at page 106*

Sun, M. and Zhang, J. 2022. Preferred synonymous codons are translated more accurately: Proteomic evidence, among-species variation, and mechanistic basis. *Science Advances*, 8(27): eabl9812. Publisher: American Association for the Advancement of Science. *Cited at pages 11, 46*

Sung, W. et al 2012. Drift-barrier hypothesis and mutation-rate evolution. *Proceedings of the National Academy of Sciences*, 109(45): 18488–18492. Publisher: Proceedings of the National Academy of Sciences. *Cited at pages 20, 25, 131*

Suran, M. 2020. Finding the tail end: The discovery of RNA splicing. *Proceedings of the National Academy of Sciences*, 117(4): 1829–1832. Publisher: Proceedings of the National Academy of Sciences. *Cited at page 8*

Sémon, M., Lobry, J.R. and Duret, L. 2006. No evidence for tissue-specific adaptation of synonymous codon usage in humans. *Molecular Biology and Evolution*, 23(3): 523–529. *Cited at pages 12, 103*

Sørensen, M.A., Kurland, C.G. and Pedersen, S. 1989. Codon usage determines translation rate in Escherichia coli. *Journal of Molecular Biology*, 207(2): 365–377. *Cited at page 11*

Tacutu, R. et al 2013. Human Ageing Genomic Resources: integrated databases and tools for the biology and genetics of ageing. *Nucleic Acids Research*, 41(Database issue): D1027–1033. *Cited at pages 37, 40, 54*

Tacutu, R. et al 2018. Human Ageing Genomic Resources: new and updated databases. *Nucleic Acids Research*, 46(D1): D1083–D1090. *Cited at page 37*

Takata, M.A. et al 2018. Global synonymous mutagenesis identifies cis-acting RNA elements that regulate HIV-1 splicing and replication. *PLoS Pathogens*, 14(1): e1006824. *Cited at page 11*

Tamura, K. 2015. Origins and Early Evolution of the tRNA Molecule. *Life*, 5(4): 1687–1699. *Cited at page 10*

Tan, S. et al 2022. FishExp: A comprehensive database and analysis platform for gene expression and alternative splicing of fish species. *Computational and Structural Biotechnology Journal*, 20: 3676–3684. *Cited at page 39*

Tateno, Y. and Gojobori, T. 1997. DNA Data Bank of Japan in the age of information biology. *Nucleic Acids Research*, 25(1): 14–17. *Cited at page 35*

The Arabidopsis Genome Initiative 2000. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature*, 408(6814): 796–815. Number: 6814 Publisher: Nature Publishing Group. *Cited at page 31*

The C. elegans Sequencing Consortium 1998. Genome Sequence of the Nematode C. elegans: A Platform for Investigating Biology. *Science*, 282(5396): 2012–2018. Publisher: American Association for the Advancement of Science. *Cited at page 31*

Tikhonenkov, D.V. et al 2020. Insights into the origin of metazoan multicellularity from predatory unicellular relatives of animals. *BMC Biology*, 18(1): 39. *Cited at page 3*

Tomso, D.J. and Bell, D.A. 2003. Sequence Context at Human Single Nucleotide Polymorphisms: Overrepresentation of CpG Dinucleotide at Polymorphic Sites and Suppression of Variation in CpG Islands. *Journal of Molecular Biology*, 327(2): 303–308. *Cited at page 85*

Trapnell, C. et al 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, 28(5): 511–515. Number: 5 Publisher: Nature Publishing Group. *Cited at page 62*

Traverse, C.C. and Ochman, H. 2016. From the Cover: Conserved rates and patterns of transcription errors across bacterial growth states and lifestyles. *Proceedings of the National Academy of Sciences of the United States of America*, 113(12): 3311. Publisher: National Academy of Sciences. *Cited at page 93*

Tress, M.L., Abascal, F. and Valencia, A. 2017a. Alternative Splicing May Not Be the Key to Proteome Complexity. *Trends in Biochemical Sciences*, 42(2): 98–110. *Cited at pages 10, 43, 52, 62, 68, 76, 80, 133*

Tress, M.L., Abascal, F. and Valencia, A. 2017b. Most Alternative Isoforms Are Not Functionally Important. *Trends in biochemical sciences*, 42(6): 408–410. *Cited at pages 10, 43, 76, 133*

Trucchi, E. et al 2023. Gene expression is the main driver of purifying selection in large penguin populations. Pages: 2023.08.08.552445 Section: New Results. *Cited at page 23*

Ule, J. and Blencowe, B.J. 2019. Alternative Splicing Regulatory Networks: Functions, Mechanisms, and Evolution. *Molecular Cell*, 76(2): 329–345. Publisher: Elsevier. *Cited at page 9*

Venables, J.P. 2004. Aberrant and Alternative Splicing in Cancer. *Cancer Research*, 64(21): 7647–7654. *Cited at page 133*

Venter, J.C. et al 2001. The sequence of the human genome. *Science (New York, N.Y.)*, 291(5507): 1304–1351. *Cited at pages 5, 7, 31*

Verta, J.P. and Jacobs, A. 2022. The role of alternative splicing in adaptation and evolution. *Trends in Ecology & Evolution*, 37(4): 299–308. Publisher: Elsevier. *Cited at pages 9, 44, 52, 77, 132, 133*

Vielle, A. et al 2016. Convergent evolution of sperm gigantism and the developmental origins of sperm size variability in Caenorhabditis nematodes. *Evolution*, 70(11): 2485–2503. *Cited at page 122*

Vigilant, L. et al 1991. African Populations and the Evolution of Human Mitochondrial DNA. *Science*, 253(5027): 1503–1507. Publisher: American Association for the Advancement of Science. *Cited at page 29*

Vinogradov, A.E. 2003. DNA helix: the importance of being GC-rich. *Nucleic Acids Research*, 31(7): 1838–1844. *Cited at pages 47, 105*

Vries, H.d. 1901. *Die mutationstheorie. Versuche und beobachtungen über die entstehung von arten im pflanzenreich*. Leipzig, Veit & comp. *Cited at pages 15, 20*

Wallberg, A., Glémin, S. and Webster, M.T. 2015. Extreme Recombination Frequencies Shape Genome Variation and Evolution in the Honeybee, Apis mellifera. *PLOS Genetics*, 11(4): e1005189. Publisher: Public Library of Science. *Cited at page 123*

Walsh, I.M. et al 2020. Synonymous codon substitutions perturb cotranslational protein folding in vivo and impair cell fitness. *Proceedings of the National Academy of Sciences of the United States of America*, 117(7): 3528–3534. *Cited at pages 11, 103*

Wang, E.T. et al 2008. Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456(7221): 470–476. *Cited at page 9*

Wang, H., McManus, J. and Kingsford, C. 2017. Accurate Recovery of Ribosome Positions Reveals Slow Translation of Wobble-Pairing Codons in Yeast. *Journal of Computational Biology*, 24(6): 486–500. Publisher: Mary Ann Liebert, Inc., publishers. *Cited at page 108*

Wang, J., Santiago, E. and Caballero, A. 2016. Prediction and estimation of effective population size. *Heredity*, 117(4): 193–206. Number: 4 Publisher: Nature Publishing Group. *Cited at page 23*

Wang, J. et al 2022. Rapid 40S scanning and its regulation by mRNA structure during eukaryotic translation initiation. *Cell*, 185(24): 4474–4487.e17. *Cited at page 10*

Wang, S. and Kool, E.T. 1995. Origins of the large differences in stability of DNA and RNA helices: C-5 methyl and 2'-hydroxyl effects. *Biochemistry*, 34(12): 4125–4132. *Cited at pages 7, 8*

Wang, Y. et al 2021. Nanopore sequencing technology, bioinformatics and applications. *Nature Biotechnology*, 39(11): 1348–1365. Number: 11 Publisher: Nature Publishing Group. *Cited at page 31*

Wang, Z. and Burge, C.B. 2008. Splicing regulation: From a parts list of regulatory elements to an integrated splicing code. *RNA*, 14(5): 802–813. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab. *Cited at page 9*

Waples, R.S. 2002. Effective Size of Fluctuating Salmon Populations. *Genetics*, 161(2): 783–791. *Cited at page 24*

Waples, R.S. 2016. Life-history traits and effective population size in species with overlapping generations revisited: the importance of adult mortality. *Heredity*, 117(4): 241–250. *Cited at pages 24, 28, 54, 66, 78*

Waples, R.S. 2022. What Is Ne, Anyway? *The Journal of Heredity*, 113(4): 371–379. *Cited at page 137*

Waples, R.S. 2024. Practical application of the linkage disequilibrium method for estimating contemporary effective population size: A review. *Molecular Ecology Resources*, 24(1): e13879. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1755-0998.13879. *Cited at page 25*

Waples, R.S. and Do, C. 2010. Linkage disequilibrium estimates of contemporary Ne using highly variable genetic markers: a largely untapped resource for applied

conservation and evolution. *Evolutionary Applications*, 3(3): 244–262. *Cited at page 25*

Waterhouse, R.M. et al 2018. BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics. *Molecular Biology and Evolution*, 35(3): 543–548. *Cited at pages 53, 57, 111*

Waterston, R.H. et al 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420: 520–562. ADS Bibcode: 2002Natur.420..520W. *Cited at page 31*

Watson, J.D. and Crick, F.H.C. 1953. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature*, 171(4356): 737–738. Number: 4356 Publisher: Nature Publishing Group. *Cited at page 4*

Weber, J.L. et al 2002. Human Diallelic Insertion/Deletion Polymorphisms. *The American Journal of Human Genetics*, 71(4): 854–862. *Cited at page 20*

Weinberg, D.E. et al 2016. Improved ribosome-footprint and mRNA measurements provide insights into dynamics and regulation of yeast translation. *Cell reports*, 14(7): 1787–1799. *Cited at page 46*

Weyna, A. and Romiguier, J. 2020. Relaxation of purifying selection suggests low effective population size in eusocial Hymenoptera and solitary pollinating bees. *bioRxiv*, page 2020.04.14.038893. Publisher: Cold Spring Harbor Laboratory Section: New Results. *Cited at pages 28, 54, 66, 78*

White, E.P. et al 2007. Relationships between body size and abundance in ecology. *Trends in Ecology & Evolution*, 22(6): 323–330. *Cited at page 70*

Whitney, K.D. and Garland, T. 2010. Did Genetic Drift Drive Increases in Genome Complexity? *PLoS Genetics*, 6(8): e1001080. *Cited at pages 131, 139*

Wilkins, M.H.F., Stokes, A.R. and Wilson, H.R. 1953. Molecular Structure of Nucleic Acids: Molecular Structure of Deoxypentose Nucleic Acids. *Nature*, 171(4356): 738–740. Number: 4356 Publisher: Nature Publishing Group. *Cited at page 4*

Wilson, E.O. 2003. The encyclopedia of life. *Trends in Ecology & Evolution*, 18(2): 77–80. *Cited at pages 37, 40, 54*

Wright, C.J., Smith, C.W.J. and Jiggins, C.D. 2022. Alternative splicing as a source of phenotypic diversity. *Nature Reviews Genetics*, 23(11): 697–710. Number: 11 Publisher: Nature Publishing Group. *Cited at pages 9, 52, 77, 133*

Wright, S. 1929. The Evolution of Dominance. *The American Naturalist*, 63(689): 556–561. Publisher: University of Chicago Press, American Society of Naturalists. *Cited at pages 17, 52*

Wright, S. 1931. Evolution in Mendelian Populations. *Genetics*, 16(2): 97–159. Publisher: Genetics Section: INVESTIGATIONS. *Cited at pages 17, 23*

Wright, S. 1951. Fisher and Ford on "the Sewall Wright Effect". *American Scientist*, 39(3): 452–479. Publisher: Sigma Xi, The Scientific Research Society. *Cited at page 17*

Wright, S. 1970. Random Drift and the Shifting Balance Theory of Evolution. In K.-i. Kojima, editor, *Mathematical Topics in Population Genetics*, Biomathematics, pages 1–31. Springer, Berlin, Heidelberg. *Cited at page 17*

Wright, S.a.o. 1932. The roles of mutation, inbreeding, crossbreeding, and selection in evolution. *Cited at page 17*

Wu, Q. et al 2019. Translation affects mRNA stability in a codon-dependent manner in human cells. *eLife*, 8: e45396. Publisher: eLife Sciences Publications, Ltd. *Cited at pages 21, 103*

Wu, Z. et al 2022. Expression level is a major modifier of the fitness landscape of a protein coding gene. *Nature Ecology & Evolution*, 6(1): 103–115. Number: 1 Publisher: Nature Publishing Group. *Cited at page 23*

Xiang, K. and Bartel, D.P. 2021. The molecular basis of coupling between poly(A)-tail length and translational efficiency. *eLife*, 10: e66493. Publisher: eLife Sciences Publications, Ltd. *Cited at page 8*

Xiong, K. et al 2017. Drift Barriers to Quality Control When Genes Are Expressed at Different Levels. *Genetics*, 205(1): 397–407. *Cited at pages 87, 93*

Xu, C. and Zhang, J. 2018. Alternative polyadenylation of mammalian transcripts is generally deleterious, not adaptive. *Cell systems*, 6(6): 734–742.e4. *Cited at pages 10, 52, 53, 93*

Xu, C. and Zhang, J. 2020. A different perspective on alternative cleavage and polyadenylation. *Nature Reviews Genetics*, 21(1): 63–63. Number: 1 Publisher: Nature Publishing Group. *Cited at pages 52, 93*

Xu, C., Park, J.K. and Zhang, J. 2019. Evidence that alternative transcriptional initiation is largely nonadaptive. *PLoS Biology*, 17(3): e3000197. *Cited at pages 52, 93*

Xu, G. and Zhang, J. 2014. Human coding RNA editing is generally nonadaptive. *Proceedings of the National Academy of Sciences*, 111(10): 3769–3774. Publisher: Proceedings of the National Academy of Sciences. *Cited at pages 52, 93*

Yang, J.R. et al 2012. Protein misinteraction avoidance causes highly expressed proteins to evolve slowly. *Proceedings of the National Academy of Sciences*, 109(14): E831–E840. Publisher: Proceedings of the National Academy of Sciences. *Cited at page 23*

Yang, J.R., Chen, X. and Zhang, J. 2014. Codon-by-codon modulation of translational speed and accuracy via mRNA folding. *PLoS biology*, 12(7): e1001910. *Cited at page 21*

Yang, Z. and Nielsen, R. 1998. Synonymous and nonsynonymous rate variation in nuclear genes of mammals. *Journal of Molecular Evolution*, 46(4): 409–418. *Cited at pages 57, 59, 97*

Yu, C.H. et al 2015. Codon Usage Influences the Local Rate of Translation Elongation to Regulate Co-translational Protein Folding. *Molecular Cell*, 59(5): 744–754. *Cited at page 46*

Zdobnov, E.M. et al 2017. OrthoDB v9.1: cataloging evolutionary and functional annotations for animal, fungal, plant, archaeal, bacterial and viral orthologs. *Nucleic Acids Research*, 45(D1): D744–D749. *Cited at page 57*

Zhang, H. et al 2023. Algorithm for optimized mRNA design improves stability and immunogenicity. *Nature*, 621(7978): 396–403. *Cited at page 133*

Zhang, J. and Kumar, S. 2023. Masatoshi Nei (1931–2023). *Nature Ecology & Evolution*, pages 1–2. Publisher: Nature Publishing Group. *Cited at page 19*

Zhang, J. and Xu, C. 2022. Gene product diversity: adaptive or not? *Trends in Genetics*, 38(11): 1112–1122. *Cited at pages 52, 93*

Zhang, J. and Yang, J.R. 2015. Determinants of the rate of protein sequence evolution. *Nature Reviews Genetics*, 16(7): 409–420. Number: 7 Publisher: Nature Publishing Group. *Cited at page 23*

Zhao, F. et al 2021. Genome-wide role of codon usage on transcription and identification of potential regulators. *Proceedings of the National Academy of Sciences*, 118(6): e2022590118. Publisher: Proceedings of the National Academy of Sciences. *Cited at page 11*

Zhou, Z. et al 2016. Codon usage is an important determinant of gene expression levels largely through its effects on transcription. *Proceedings of the National Academy of Sciences of the United States of America*, 113(41): E6117–E6125. *Cited at page 11*

Zia, A. and Moses, A.M. 2011. Ranking insertion, deletion and nonsense mutations based on their effect on genetic information. *BMC Bioinformatics*, 12: 299. *Cited at page 20*