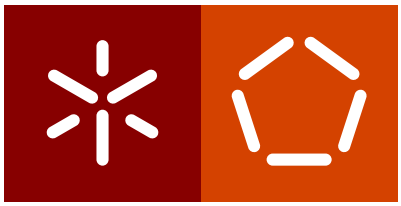
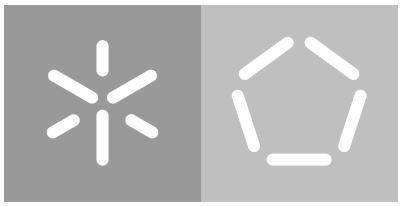


Universidade do Minho
Escola de Engenharia
Departamento de Informática

Inês de Castro Fernandes

Comparison of DNA Sequencing Technologies using Sensory Systems Genes in Bird Genomes

January 2020



Universidade do Minho
Escola de Engenharia
Departamento de Informática

Inês de Castro Fernandes

Comparison of DNA Sequencing Technologies using Sensory Systems Genes in Bird Genomes

Master dissertation
Master Degree in Bioinformatics

Dissertation supervised by
Prof. Agostinho Antunes
Prof. Miguel Rocha

January 2020

DIREITOS DE AUTOR E CONDIÇÕES DE UTILIZAÇÃO DO TRABALHO POR TERCEIROS

ACKNOWLEDGEMENTS

First I would like to thank my supervisor Professor Doctor Agostinho Antunes for giving me the opportunity to work with a team involved in evolutionary genomics and bioinformatics research, an area in which I really enjoyed to work with. I also want to thank for your availability, support, guidance, knowledge transmitted and for constantly encouraging me to give my best in this work.

I am also grateful to my co-supervisor and director of the master's in Bioinformatics, Professor Doctor Miguel Rocha, for always being available to clarify me, and for helping me with my work throughout this process.

I would also like to thank all the members of the Evolutionary Genomics and Bioinformatics Research Team from *Centro Interdisciplinar de Investigação Marinha e Ambiental (CIIMAR)* and *Faculdade de Ciências do Porto (FCUP)*, especially Liliana and Tito for guiding, teaching and helping me in everything I needed during the development of my work. I still have some special words for Liliana, because I don't even know how to thank for her help. A giant thanks for helping me to integrate well in this area, for the constant friendly approach and, essentially, for always having patience for the endless massacre with doubts and questions. Whether in person or by email, she was always available to help me. Without her help, I would have suffered a lot more over the past year. I am really grateful.

A word of thanks also to all my friends. Especially for my library classmates, Sofia, Alexandre and Inês, for the moments of sharing and laughter that they provided me over the last year, during our "snack breaks". No doubt they helped me a lot to not be constantly stressed and anxious. Also to my friends Raquel and Mirz, although we were not often together, our dinners saved me from some moments of despair. And thank you so much for always motivating, supporting and giving me good advice to endure this journey.

To my family, my parents and my brother, I also want to thank you so much for all the help, both personal and financial, that you have been providing me over the years. No doubt, without them none of this would be possible. Thank you so much for always accompanying me in my triumphs but also in my failures.

Finally, but not least, to my two favorite mammals, Blue and Mário, my anti-stress ball and my personal psychologist. Especially to Mário, for the constant support, motivation and love. Without him I don't know how I could have endured all this. There are really no words to describe how much he helped me. Forever grateful.

Thank you all. Very much!

This work was partially supported by the Strategic Funding UID/Multi/04423/2019 through national funds provided by FCT and the European Regional Development Fund (ERDF) in the framework of the program PT2020, by the European Structural and Investment Funds (ESIF) through the Competitiveness and Internationalization Operational Program - COMPETE 2020 and by National Funds through the FCT under the project PTDC/AAG-GLO/6887/2014 (POCI-01-0124-FEDER-016845).

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration.

I further declare that I have fully acknowledged the Code of Ethical Conduct of the University of Minho.

ABSTRACT

Vertebrate sensory systems play a major role in the survival of species, since their fitness and reproduction success depends on their ability to adapt to the surrounding environmental conditions. Consequently, the sub-genomes of protein-coding genes responsible for stimulus perception, are constantly undergoing selective pressures and mutational changes. Birds are a very diverse group of organisms, about which there is still little information regarding the genetic and molecular mechanisms that gave rise to the enormous variability of phenotypes existing among species. However, due to the knowledge that has been acquired about their sensory systems, birds have been considered a good model of study in this area.

The sequencing of genomes, presents a great contribution to this theme and to the understanding of the influence of selective pressure in the modification of sub-genomes. However, a major restriction on this area is related to the quality of sequencing. There are several technologies that allow the sequencing of genomes, although they differ in the method used. One main difference is the size of the DNA reads generated. Technologies that are not so recent (Sanger, Solexa and Illumina), are based on short reads sequencing, on the other hand, more recent technologies (ex: PacBio) are based on long reads. Recently it has been proposed that the sequencing methodology may have a huge influence on the genome assembly process and on the quality of the generated sequence.

Considering this importance, the main purpose of this work was to compare the sequencing efficiency and quality of different technologies and to assess if the PacBio technology presents an "advantage" over the rest. Thus, we selected five species of birds with genomes that have been sequenced through older technologies and with the PacBio technology. From this, several sets of genes from six different sensory systems were extracted, in order to obtain a good representation of each system. The results obtained revealed that in four of the systems, there are significant differences in the quality of sequencing between the "older" technologies and PacBio. Specifically, in PacBio genomes there is a higher quality of sequencing with regard to the fragmentation of genes along the genomes (less fragmented genes) as well as a higher quality in gene integrity (more complete and contiguous genes). In addition, allowed us to corroborate previously conclusions proposed by related studies that suggest that PacBio, and in this way, long read assembly, provides great improvements in genome assembly and gene completeness, as well as improvements in sequencing more complex genome regions.

Keywords: Sequencing Technologies, Sensory Systems, Bird Species, Gene Extraction.

RESUMO

Os sistemas sensoriais dos vertebrados desempenham um papel fulcral para a sobrevivência das espécies, uma vez que o seu sucesso de *fitness* e reprodução depende da sua capacidade de adaptação às condições ambientais envolventes. Assim, os sub-genomas de genes codificantes de proteínas responsáveis por perceção de estímulos, estão constantemente a sofrer pressões seletivas e mutações. As aves, são um grupo de organismos bastante diverso, sobre as quais ainda existe pouca informação relativamente aos mecanismos genéticos e moleculares que deram origem à variabilidade de fenótipos existentes entre as espécies. No entanto, devido ao conhecimento que se tem vindo a adquirir acerca dos seus sistemas sensoriais, têm sido consideradas um bom modelo de estudo nesta área.

A sequenciação do genoma, apresenta-se um contributo para este tema e para a compreensão da influência da pressão seletiva na modificação dos sub-genomas. Porém, uma grande restrição nesta área relaciona-se com a qualidade de sequenciação das várias tecnologias, que diferem no método utilizado. Uma das principais diferenas é o tamanho das *reads* de DNA geradas. Tecnologias menos recentes (Sanger, Solexa e Illumina), baseiam a sequenciação em *short reads*, por outro lado, tecnologias mais recentes (ex: PacBio) baseiam-se nas *long reads*. Tem-se proposto que esta diferença poderá ter uma enorme influência no processo de sequenciação dos genomas.

Considerando esta importância, o principal objetivo foi comparar a eficiência e qualidade de sequenciação de diferentes tecnologias e perceber se a tecnologia PacBio apresenta vantagens perante as restantes. Assim, selecionaram-se cinco espécies de aves com genomas que foram sequenciados através de tecnologias mais antigas e através da tecnologia PacBio. Destes, extraíram-se vários conjuntos de genes pertencentes a seis sistemas sensoriais, com o objetivo de obter uma boa representação de cada sistema. Os resultados revelaram que em quatro dos sistemas, existem diferenças significativas na qualidade de sequenciação entre tecnologias "mais antigas" e PacBio. Especificamente, nos genomas PacBio existe uma maior qualidade em termos de fragmentação dos genes ao longo dos genomas (genes menos fragmentados) assim como uma maior qualidade na integridade dos genes (genes mais completos e contíguos). Mais, permitiu-nos corroborar conclusões anteriormente reportadas por estudos relacionados, as quais sugeriam que PacBio, e assim, *assemblies* através de *long reads*, providencia uma grande melhoria na montagem de genomas e na obtenção de genes completos, assim como melhorias na sequenciação de regiões genómicas mais complexas.

Palavras-chave: Tecnologias de Sequenciação, Sistemas Sensoriais, Espécies de Aves, Extração de Genes.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

A

ASIC Acid-Sensing Ion Channel.

B

BLAST Basic Local Alignment Search Tool.

BSDP Bounded Sparse Dynamic Programming.

C

CDS Coding Sequence.

CIIMAR Centro Interdisciplinar de Investigação Marinha e Ambiental.

CRY Cryptochrome.

D

DBBJ DNA Data Bank of Japan.

DDNTPS Dideoxynucleotides.

DEG/ENAC Degenerin/Epithelial Sodium Channels.

DNA Deoxyribonucleic Acid.

DNTPS Deoxyribonucleotides.

E

EMBL-EBI European Bioinformatics Institute.

ENA European Nucleotide Archive.

ENAC Epithelial Sodium Channels.

F

FAD Flavin Adenine Dinucleotide.

FCUP Faculdade de Ciências do Porto.

G

GPCRS G Protein-Coupled Receptors.

H

HTMRS High-Threshold Mechanoreceptors.

I

IEGS Immediate Early Genes.

ISCA1 Iron-Sulfur-Cluster Assembly Protein.

K

KCNK Potassium Channel Subfamily K.

KCNMA1 Potassium Calcium-Activated Channel Subfamily M alpha 1.

KCNMB1 Potassium Calcium-Activated Channel Subfamily M beta 1.

KCNQ Potassium Voltage-Gated Channel Subfamily Q.

L

LTMRS Low-Threshold Mechanoreceptors.

M

MEGA Molecular Evolutionary Genetics Analysis.

ML Maximum Likelihood.

MOE Main Olfactory Epithelium.

N

NCBI National Center for Biotechnology Information.

NGS Next-Generation Sequencing.

O

ONT Oxford Nanopore Technology.

P

PACBIO Pacific Biosystems.

PIR Protein Information Resource.

R

RNA Ribonucleic Acid.

S

S./I./S. Sanger/Illumina/Solexa.

SBS Sequencing by Synthesis.

SIB Swiss Institute of Bioinformatics.

SLR-SEQ Synthetic Long-Read sequencing.

SMRT Single Molecule Real-Time.

T

TAAR Trace Amine-Associated Receptor.

TAR Taste Receptor.

TAS1R Taste Receptors family 1.

TAS2R Taste Receptors family a.

TES Transposable Elements.

TMC Transmembrane channel-like.

TRP Transient Receptor Potential.

TRPA TRP ankyrin.

TRPC TRP-Canonical.

TRPM TRP melastatin subfamily.

TRPV TRP vanilloid subtype.

U

UNIPARC UniProt Archive.

UNIPROT Universal Protein resource.

UNIPROTKB UniProt Knowledgebase.

UNIREF UniProt Reference Clusters.

Z

ZMWS Zero Mode Waveguides.

1

INTRODUCTION

This dissertation describes the Master's thesis work developed during graduation of Master in Bioinformatics in University of Minho. The practical and experimental context of this work was carried out by resorting the computational resources of FCUP and CIIMAR.

1.1 CONTEXT AND MOTIVATION

Across the world, animal species are exposed to a variety of ecological and physiological challenges. The success of these organisms depends essentially on their ability to adapt to the environment in which they inhabit, to promote advantageous reproductive strategies and, also, to develop mechanisms that keep them healthy [1, 2].

Of the several processes of adaptation that organisms develop, those which contribute to genetic fitness are essential to survival. Such processes are responsible for inducing the molecular dynamic plasticity at the level of protein encoding genes (e.g. gene conversion, gene duplication and positive selection). Genes influencing fitness are often associated with positive selection, since the amino acid replacement mutations may promote the functional improvement of several proteins, which could be involved in important biological mechanisms. Thus, the genome modifications in a species can lead to the emergence of new gene families, gene contraction/expansion and even the gene enhancement, promoting a greater adaptive capacity [1, 2, 3].

Genes that encode proteins belonging to sensory systems are examples of genes that are evolving dynamically, since they play a crucial role in the survival and adaptive capacity of different species of organisms.

Examples of extremely important sensory systems include Chemoreception, namely the *Taste Receptor (TAR)*s and the *Trace Amine-Associated Receptor (TAAR)*s families. TARs are specialized in the detection of chemical components in food resources allowing, in this way, the ingestion of nutrients crucial for survival, as well as the detection of poisonous substances [4]. On the other hand, TAARs are related to the recognition of social cues, since they respond, in addition to trace amines, to biogenic amines, that are able to function as pheromones. This function may have influence in the organisms' behaviour, in what con-

cerns recognition and interaction between individuals and mate choice [5, 6]. In addition to Chemoreception, Magnetoreception, Photoreception and Thermoreception are also systems with preponderant functions, namely: in the orientation and navigation relative to the Earth's magnetic field, having special importance in the survival of migratory birds [7, 8]; in colour vision, which has a great influence on the organisms social interactions, predator avoidance, and even foraging [9]; and maintaining the homeostatic balance of individuals, allowing them to sense, tolerate and adapt to temperature changes in their surroundings; respectively [10]. Finally, but also of great importance are the auditory and tactile systems. Both allow to perceive external stimuli, processing the stimuli received, whether auditory or mechanical, respectively, and transmit them to the central nervous system. In this way it is possible for organisms to successfully perceive the environment around them. They are also fundamental in social interaction between species. Additionally, the auditory system plays a very important role in competition between individuals, choice of partner and avoidance of predators [11, 12]. Furthermore, these two systems are quite complex, and since their receptor molecules participate in a number of biochemical pathways beyond these systems, much remains to be clarified, and more studies are needed, including studies specifically focusing on birds, since much of the work done is focused in mammals.

An important goal for evolutionary biologists is to understand the evolutionary mechanisms by which all the phenotypic variation arose among different species. In fact, understanding how and why closely related organisms are so distinct in highly conserved genes with critical functions is a biologically relevant approach. With enough knowledge in this topic, it may be possible to unveil the evolutionary history of the phenotypic characteristics, at molecular, cellular and even physiological level. In addition, it may also help to understand the adaptability of different species to the environment [1].

Birds are recognized as one of the most diverse groups of all terrestrial vertebrates, and there is already great information about the diversity of their sensory abilities. However, little is known about their evolutionary history and the mechanisms that contribute for their genome size equilibrium, since most evolutionary studies focus only on taxa with high variety of genome sizes, such as, teleost fishes or insects. Thus, there is a need to include birds in this type of studies, and due to the existing information on their sensory capabilities, birds have recently been recognized as a very useful group for this area [13, 14, 15].

Since sensory protein-encoding genes play a central role in the organism's fitness, understanding how ecological conditions have shaped patterns of diversification of, for instance, olfactory and vision abilities, and how the genetic patterns correlate, will allow to understand the evolution of these genes. In this way, these analyses may provide important clues about molecular mechanisms involved in environment adaptation, the role of natural selec-

tion [1], and even provide valuable insights about the evolution and functioning of our own genome [16].

The possibility of characterizing and quantifying adaptation at the molecular level is, thus, one of the main ambitions for evolutionary biology and genomic studies, considering the great potential for revealing, at multiple scales, the evolutionary journey of phenotypic traits. A key aspect to perform these studies is genome sequencing, since it provides even more information in what concerns the adaptive genetic variation and influence of selective pressures on shaping sub-genomes, in case of interest for this study, the sensory systems sub-genomes [17]. Furthermore, in studies of population genetics, genetic data is very valuable, since it will allow establishing phylogenetic relationships with greater precision, understanding population structures, detecting variations among individuals, and even detect signatures of selection [18].

Over the years, technologies for *Deoxyribonucleic Acid (DNA)* sequencing have been developed and, currently, they are grouped into 3 different categories: 1^{st} generation sequencing (Sanger sequencing), 2^{nd} generation sequencing or *Next-Generation Sequencing (NGS)*, and 3^{rd} generation sequencing. These technologies differ in the process used to sequence the DNA, thus directly influencing the speed of which reads are generated, the throughput rate, the error rate, and the length of the generated reads, which in turn will have influence in the sequencing quality. One of the latest technologies, belonging to the 3^{rd} generation, is *Pacific Biosystems (PacBio)* platform. This technology essentially differs from other sequencing technologies in what concerns the length of the generated reads, which are much longer (up to 60 Kb) than the reads of other 1^{st} and 2^{nd} generation technologies. This brings with it a number of advantages, such as, the possibility to locate repetitive sequences through a single read, allows the sequencing of entire transcripts, and also, the real-time detection of biological events [19, 20, 21, 22].

In this way, considering the importance and applications of DNA sequencing and the differences among the sequencing technologies that have been developed, the main interest of this work is to compare the PacBio sequencing against "older" technologies, namely Sanger sequencing, Solexa and Illumina, in order to evaluate if, indeed, they vary in the sequencing quality. For such, we selected a set of birds' genomes, sequenced by 1^{st}, 2^{nd} and 3^{rd} technologies to evaluate the success to extract genes from these genomes, i.e., to evaluate if the sequences were obtained in all their length, with no errors, depending on the technology from each genome was generated.

The molecular markers selected for the extraction belong to six different sensory systems, previously mentioned, since these systems are constantly under variation and adaptation. In some of these systems, such as the chemoreception and the photoreception, the selected genes are well reported and are directly involved in binding/recognition of a molecular target while in others, such as the auditory system, tactile system, and thermoreception,

they operate through complex molecular cascades of recognition, and the receptors also have other molecular functions. The genes belonging to the latter systems mentioned, were chosen based on literature research.

1.2 MAIN AIMS

Since sensory systems are constantly under molecular changes that underlie the evolution of phenotypic traits, they possess an adaptive relevance to differential conditions of environmental selective pressure. In addition, the currently DNA sequencing technologies vary in the sequencing quality, producing different results, that, consequently, have a big impact in a great amount of studies that depend on genetic information. Having this in mind, the main goal of this work is to develop genomic and statistical analyses to understand whether the newer technologies, compared to the older ones, have improved on issues that have been considered major constraints over the years. For such, a set of different techniques will be performed, these being:

- Annotation of multi-gene families involved in sensory systems;
- Analyses of similarities among genes and proteins of different bird species;
- Multi-sequence alignments (all genomes with all genes);
- Gene extraction;
- Statistical analyses for comparison of sequencing technologies quality;
- Analyses of genome synteny of the multi-gene families under study;
- Write thesis with the obtained results from the precedent subjects.

1.3 THESIS OUTLINE

This document includes four more important chapters: the State of the Art, the Materials and Methods, the Results and Discussion, and the Conclusions and Further Work.

In the State of the Art chapter (2), is presented a brief review about molecular and adaptive evolution. Furthermore, in this chapter is presented an analysis in terms of function, structure and protein-coding gene, of the six sensory systems selected for this work. Finally, is presented a brief review about four sequencing technologies (Sanger, Solexa, Illumina and PacBio).

In what concerns the Materials and Methods chapter (3), it is described a succinct analysis of the tools that used during this work (*Molecular Evolutionary Genetics Analysis (MEGA)*

software, Exonerate software, R, Python and Biological Databases). Also, the information obtained and the tasks performed during this work, such as: All the genomes' information that were used (reference, length, number of contigs, sequencing technology, etc), information about all the genes that were used, how the alignments and genes extraction were performed and the statistical tests performed.

In the Results and Discussion chapter (4) we describe all the results of the genes extractions, i.e., all the recorded parameters respecting alignments of the whole set of genes with all genomes. Also presented here are the results of the statistical tests: Normality Tests (Shapiro-Wilk) and Hypothesis Tests (Mann-Whitney U Test). Moreover, we provide hypothetical reasons and explanations for what we obtained. For that, we include a brief description of the main features of birds genomes, an extensive comparison among the results obtained by sequencing based on short reads and long reads, including advantages and disadvantages of the different methods,and finally, a section that presents quite relevant related studies that confirm the conclusions that we achieved in this work.

Lastly, in the Conclusions and Further Work chapter (5), we present the main conclusions achieved with this study, and we present some work that we intend to do, in order to further explore this topic and better understand the reasons for our results.

2

STATE OF THE ART

2.1 MOLECULAR EVOLUTION AND ADAPTATION

Evolution, a concept that refers to changes within and among biological populations, is a constant phenomenon that arises from several events in nature. These events can be at the molecular level such as, gene flow, genetic drift, mutations, genetic conflicts, and, also, natural and sexual selection [23].

Although it is known that different changes in the genome can generate identical phenotypic variations, it is also known that, with regard to the occurrence of mutations (which may be deletion, insertion, substitution or inversion), those that cause changes at the expression and regulation of protein encoding genes, are more likely to generate phenotypic variations and contribute to evolution [24]. These mutations can be considered non-synonymous mutations and synonymous mutations. Non-synonymous mutations are those that change the amino acid sequence and can likely affect protein structure and activity. On the other hand, synonymous mutations do not alter the amino acid sequence. In this way, these mutations are generally detected at the protein level through amino acid substitutions [25]. Nonsynonymous substitutions are, therefore, more likely to participate on phenotypic evolution than synonymous substitutions, as it was expected [24].

Through the analysis of these mutations, the evolutionary pressure effect can be evaluated by the ratio, ω , of non-synonymous substitutions (probably selected) to synonymous substitutions (probably neutral and with no sequence change), ω = dN/dS. This ratio suggests positive, neutral or negative pressures, if it presents a value greater, equal or less than 1, respectively [26].

Moreover, across the worldwide ecosystems, animal species are exposed to a variety of ecological and physiological pressures. This leads to adaptive responses, which often have a genetic basis, and the success of these organisms to survive depends essentially on it. These changes frequently promote advantageous reproductive strategies and, also, favour the development of mechanisms that keep the organisms healthy [1, 2, 23]. In fact, it is noticeable, in animal genomes, the environmental niche specialization [16].

Thus, understanding the evolutionary mechanisms that led to the wide variety of eco-physiological adaptations and phenotypic variation and also, how and why closely related organisms are so distinct at the molecular level, is a biologically relevant approach. Specially in what concerns sensory systems, it allows to uncover the evolutionary history of phenotypic characteristics and provides several clues about the adaptability of organisms to the environment. Furthermore, these studies can promote the perception and knowledge that we have about the evolution and functioning of our own genome [1, 16]. As mentioned before, the DNA sequencing technologies that have been developed until now, are crucial for these studies. The access to genomic information brings the possibility of studying these mechanisms even further and with more accuracy. In this way, it is of extremely importance that these technologies may provide the best results, i.e., genomes assembly of high quality that are able to represent a species [27, 28].

Due to this importance, lately, several studies have been carried out on this subject. In the present study, special attention will be given to differences of gene quality among genomes of different birds species at the level of sensory systems. Since these systems are constantly under environmental pressures experienced for multiple vertebrate species. In the next subsections, we will report and analyse six different sensory systems, as well as a brief description of a set of DNA sequencing technologies that were during the years and, are currently quite resorted to allow these type of studies.

2.2 SENSORY SYSTEMS

2.2.1 *Chemoreception: TARs and TAARs*

Chemoreception or chemosensation is a mechanism widely used by animal species, for the detection of chemicals in the surrounding environment. This mechanism plays a critical role in the survival and fitness of the animal species, since it enables these organisms to communicate with others of the same species, locate food or hosts, detect predators, avoid eating toxic substances, and find fitting mating partners, which may contribute to evolutionary processes like speciation and reproductive isolation [29, 30].

The chemosensory systems (including taste and smell) are responsible for the detection of countless and diverse chemical molecules. Depending on the species, several small molecules and proteins can be perceived through the sense of taste. Regarding smell, organisms are able to detect countless volatile and non-volatile proteins as well as non-volatile hydrocarbons. Consequently, these systems have developed complex and extensive repertories of receptor genes, so that organisms can recognize this huge diversity of chemical structures [29, 17].

In this work, the olfactory perception will not be approached, due to the extensive genetic repertoire and the large amount of gene paralogs. In this way, in what concerns chemoreception, this works will focus on two different chemosensory receptors, the TARs and the TAARs, both of them extremely important for the survival and adaptation of most organism's species.

Taste Receptors

Regarding TARs, these receptors are specialized in the detection of chemical components in food resources, which, in turn will determine the organism's choice of specific food as well as its ingestion. As an example, there are chemical components that cause aversive reactions while others cause appetite reactions, allowing, in this way, the ingestion of nutrients necessary for survival, as well as the detection of poisonous substances, respectively [4].

At the level of the sense of taste, there are five different modalities, being this sweet, umami, bitter, salty and sour [31, 32].

The bitter sense of taste usually detects the presence of bitter-tasting chemicals in food, and these are generally poisonous foods, such as insect defensive secretions [33]. It is, therefore, a defence that organisms possess, preventing them from ingesting harmful substances [4].

Contrary to bitter taste, sweet and umami modalities are related to the detection of nutritious foods [33]. More specifically, umami taste is involved in the detection of a few L-amino acids and is therefore related to protein rich diets. On the other hand, sweet taste is triggered by the presence of carbohydrates, which may be mono or disaccharide sugars. Finally, salty and sour tastes detect the presence of sodium (Na) and acids H^+ [34].

Regarding the structure and genetics of these receptors, in most vertebrates they are conserved and are expressed in epithelial clusters of taste sensory cells, known has taste buds [34]. Birds, compared to mammals, have a smaller number of these receptors due to lower saliva secretion, smaller oral cavity and fewer taste buds.

Furthermore, these proteins, in terms of their structure, are *G Protein-Coupled Receptors (GPCRs)* with seven α-helix transmembrane regions (Figure 1) [31].

Also concerning birds, and more specifically chicken (*Gallus gallus*), taste receptors are presented in two families, the *Taste Receptors family 1 (TAS1R)* and the *Taste Receptors family a (TAS2R)*. Wihin the TAS1R there are only two members, TAS1R1 and TAS1R3, with the absence of Homo sapiens ortholog TAS1R2. This family is responsible for umami and sweet taste perception, whereas umami taste receptor is a heterodimer resulting from the interaction of TAS1R1 and TAS1R3 [31, 34]. On the other hand, it is known that the TAS2R, although with only a few genetic studies performed, presents in chicken three members,

TAS2R1, TAS2R2 and TAS2R3, also known as TAS2R40, TAS2R4 and TAS2R7, respectively. This family is responsible for bitter taste perception [34].

Trace Amine-Associated Receptors

Besides TARs, a new other family of GPCRs, belonging to the chemoreceptors group, are the TAARs [35, 6].

Initially, these receptors were thought only to respond to trace amines in the vertebrate's nervous system. Trace amines are a group of endogenous amines with neurotransmitter functions, found at low concentrations in the central nervous system, and this group includes β-phenylethylamine, tryptamine and p-tyramine [5, 36]. However, several recent studies have shown that TAARs not only respond to trace amines but also to classical biogenic amines, due to the structural similarity between these two groups. Biogenic amines such as adrenaline, dopamine, serotonin (5-hydroxytryptamine or 5-HT), noradrenaline, and histamine, are a group of amines that are well known to have hormonal as well as neurotransmitter functions in the central and peripheral nervous system [5]. Furthermore, it has been observed that TAARs genes are selectively expressed in small amounts in *Main Olfactory Epithelium (MOE)*, suggesting that they are also involved in olfactory functions, more specifically, in the detection of volatile biogenic amines [37].

Based on this knowledge, it is believed that TAARs are related to the recognition of social cues, since biogenic amines are able to function as pheromones. This occurs when this group of amines is handled as odorants, inducing as consequence, physiological changes in the organisms [5]. This is a very relevant fact that may provide valuable insights into the pheromone-based behaviour of several species, including the behaviour in terms of recognition and interaction between individuals and mate choice [5, 6, 37].

In what concerns to the genomic field, TAARs genes are single-exon with coding chains of about 1Kb in length (Figure 1). The repertoire of these genes and the number of copies vary widely among vertebrate species [35]. In fact, the TAARs gene family in teleost fishes is quite large and versatile (about 57 intact genes found in zebrafish) compared to tetrapods (for instance, in humans only 5 genes were found intact) [38]. Moreover, regarding to the latter group, the number of TAARs genes found in amphibians and birds is considerably low, with low copy numbers. As for mammals, the number of copies varies widely among species [35].

Concerning to birds, several indications have suggesting that they have a robust sense of smell, and, as mentioned above, this may play a considerable role in the avian interaction among individuals. In addition, considering the absence of the vomeronasal organ in birds, as well as the absence of vomeronasal receptors (responsible for social mediation and communication in mammals), TAARs receptors are of even greater significance for the detection of social cues. In the genome *Gallus gallus*, up to now, three TAARs paralogues

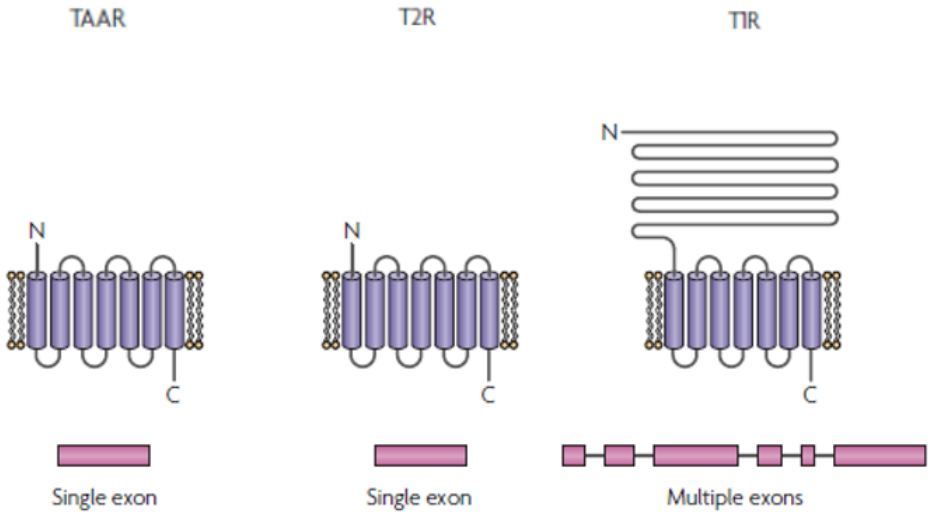


Figure 1.: Representation of TARs and TAARs. TR2 denotes the taste receptor type 2, and T1R denotes the taste receptor type 1. Initially are presented the protein structures of the receptors, in which is possible to understand the seven α-helix transmembrane regions. Below are presented the exon-intron structures of the coding genes. Adapted from [36].

have been found, known as *Taar1*, *Taar2*, and *Taar5*, which are suggested to be orthologs of three mammalian TAARs genes [6].

For the development of this work, all genes belonging to the two types of receptors mentioned above (*Tas1r1*, *Tas1r3*, *Tas2r40*, *Tas2r4* and *Tas2r7*; *Taar1*, *Taar2* and *Taar5*), where approached in order to have a good representation of the chemoreception group.

2.2.2 *Magnetoreception*

Another sensory system that is extremely important for the survival of several animal species, and especially for birds, is magnetoreception. This sensory modality allows organisms to recognize and sense the Earth's magnetic field, that varies across the globe, which, in turn, allows them to gather navigational information [8, 7, 39]. This information can be used as a biological compass, in which the magnetic vector provides directional information, but also acts as a component in the navigational "map", providing position references [39].

In this way, magnetoreception is extremely used for orientation and navigation, having special importance during migrations [7, 8]. Although organisms have other factors available that assist migrations, such as the flow of continental wind currents, the location of rivers, seas and mountains, and even the position of the sun or the moon, magnetoreception still presents a preponderant role [7]. The reason for this is because organisms can resort to

this sensory system whenever such factors are in some way disturbed, i.e., when mechanical, visual and thermal characteristics of the Earth's surface are not propitious to navigation and orientation (for instance, the occurrence of rain, clouds, fog and even changes in the wind patterns) [7, 39].

As mentioned earlier, magnetoreception is widely used for orientation and navigation, although these two terms correspond to two different tasks. Orientation corresponds to the identification of the cardinal directions (North, South, East and West). On the other hand, navigation involves the ability to measure the intensity of the magnetic field, as well as its inclination [40]. As these two different tasks require the measurement of different types of parameters and, although little is known about this sensory system (physiologically, molecularly and genetically), it is expected that there is more than one magnetosensory system for the performance of such tasks. Moreover, it is pointed out that in birds there are about two different types of magnetosensory systems, the Radical Pair Model and the Iron Based Magnetoreception [40, 7, 41].

Radical Pair Model

This model was first proposed by Ritz et al. (2000) [42] and suggests that the avian compass is generated through chemical reactions that occur in the bird's eye, more specifically in the retina. In this way, it is assumed the existence of a *Cryptochrome (CRY)*, a photosensitive flavoprotein, which presents as a cofactor the *Flavin Adenine Dinucleotide (FAD)* molecule. This cofactor molecule is responsible for the absorption of photons with certain energy levels. Therefore, the molecule becomes photoreduced, which, in turn, is oxidized in light independent reactions (FADox form), transferring an electron to a tryptophan (Trp) located nearby. Thus, this molecule undergoes a redox cycle that leads to the generation of a radical pair, a pair of molecules with uncoupled electrons. According to the spins of the uncoupled electrons, this pair of radicals may remain in the singlet state or triplet state (Figure 2) [40, 43].

This pair of radicals remains in interaction for some time, and the intensity and rate of the interaction process is dependent on the orientation and intensity of the magnetic field. The radicals eventually recombine or decay, forming more stable products [40].

From this point, much remains to be studied and the way this mechanism interacts with the avian magnetic perception is still barely perceptible. However, it is anticipated that the reaction that occurs in the retina, between the radical pairs, may have some effects on the sensitivity of the light receptors in the eye. Thus, the effect of the magnetic field on the interaction of the radical pairs may modulate the bird visual sense, producing a field of vision with darker or lighter regions (Figure 3). Therefore, birds can use these light intensity patterns to guide and orient themselves in relation to the magnetic field [42]. In addition, it

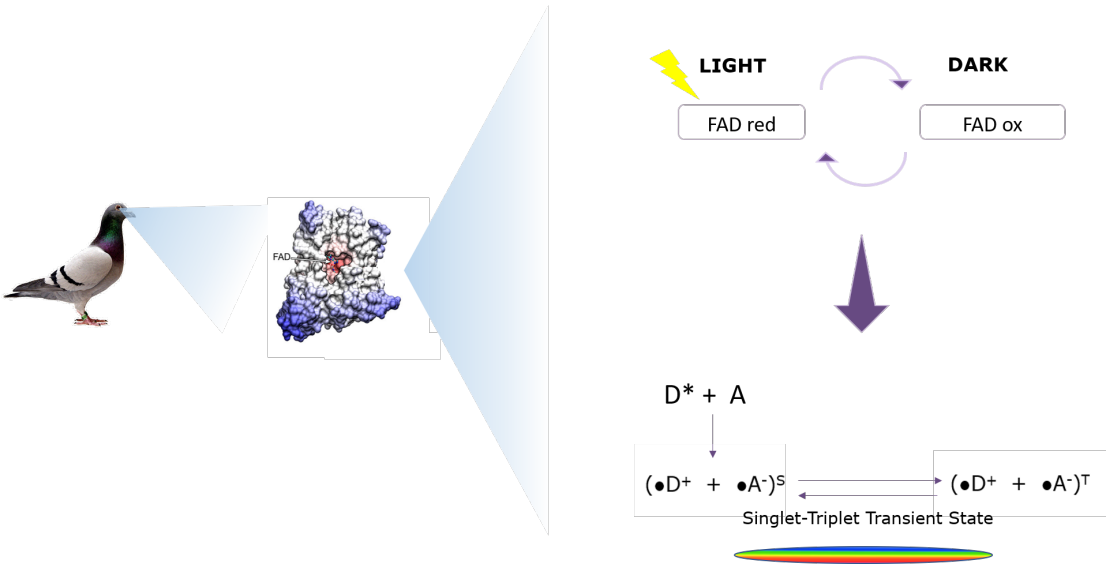


Figure 2.: Schematic representation of the Radical Pair Model. On the left side is presented the cryptochrome (existent in the bird's eye retina) with the anchored FAD molecule. On the right side are the reactions that occur through these molecules. The FAD molecule in the absence of light exists in the oxidized state, however, in the presence of UV and blue light, up to about 500 nm, FAD absorbs these photons and becomes photoreduced (upper right corner scheme). This triggers an electron transfer between FAD and the nearby tryptophan, forming a pair of radical molecules with uncoupled electrons (bottom right corner scheme). This radical pair interacts and alternates between the singlet state and the triplet state according to the surrounding magnetic field. "D" stands for "Electron Donor"; "A" stands for "Electron Acceptor". The scheme in the bottom right corner was based and adapted from [42].

was found that in the dark, birds do not exhibit activity of their magnetic compass, which is a fact that contributes to the proposed model [43].

In what concerns to cryptochrome, a protein found in several animal species (which includes all vertebrates) and in plants, it is suggested that, in addition to being involved in magnetoreception (in light-dependent pathways such as the above example), it is also involved in circadian clock [44, 45, 46, 47, 48, 49]. To date, four different cryptochromes have been found in the birds' eyes. These are CRY1a, CRY1b, CRY2 and CRY4 [44]. In chicken, these proteins have an average size of about 580 amino acids [50].

CRY1 has two isoforms that only differ in the C-termini region, CRY1a and CRY1b. Both isoforms are found in the bird's retina, although in different locations. CRY1a is expressed on the cell membrane of inner and outer segments of UV sensitive photoreceptor cells [51, 52]. Additionally, CRY1a has been detected in its active form in the bird's retina, which meets the light conditions required to provide magnetosensivity [43, 52, 51]. On the other hand, CRY1b shows a higher expression of its m*Ribonucleic Acid (RNA)* in the ganglion cell cytoplasm (cells that belong to the ganglion cell layer, a layer present in the retina), but also, in some amount, in the inner segments of photoreceptors [53, 54]. Moreover, it was

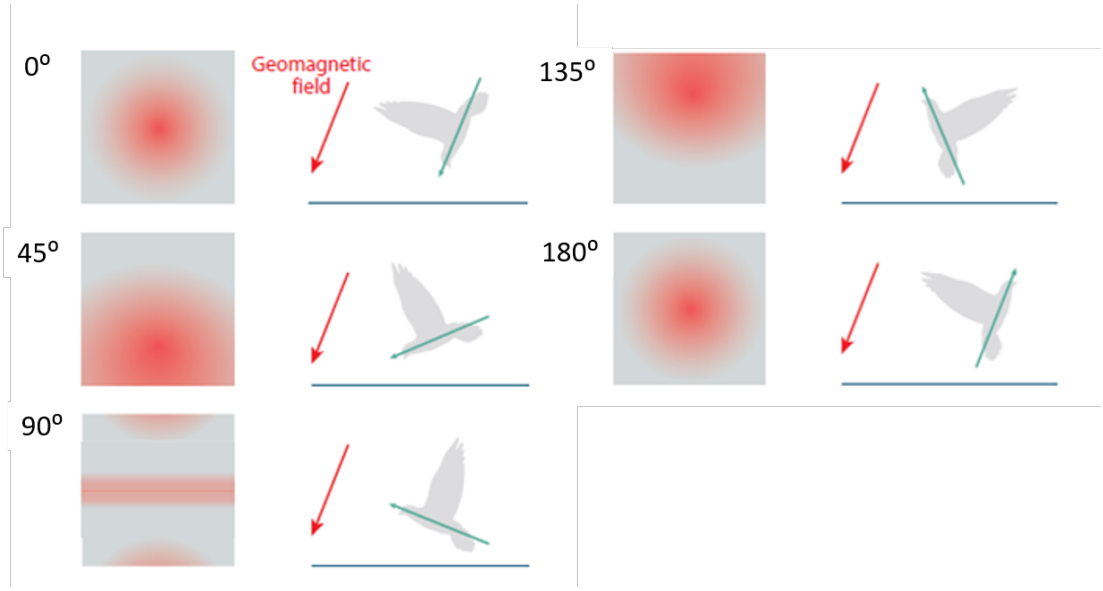


Figure 3.: Birds visual field based on the Radical Pair Model based on the Radical Pair Model. Depending on the inclination and position relative to the surrounding magnetic field, the bird can see darker and lighter zones, allowing it to orient itself. Adapted from [7].

also observed that in night-migrating birds, a higher expression of CRY1b was presented at exactly the time these birds were found active at night [44, 55, 56].

Based on the existing information of both types of CRY1, it is pointed out that, in fact, they theoretically meet the requirements of the radical pair model and that they are indeed magnetoreceptor molecules [45].

Regarding CRY2, this is presented in the retina of several bird species. Specifically, it is expressed in the photoreceptor layer and, also, in small amounts, in the ganglion cell layer. The localization of its expression points that this cryptochrome might perform magnetoreception functions. However, it is also found to be expressed in the nuclear inner membrane of cells. This indicates that CRY2 is probably a clock protein and therefore its designation as a magnetoreceptor still requires some studies [45, 44, 53, 47].

Finally, CRY4 is likewise presented in the retina of several bird species. Its expression was observed in chicken, namely in the cytoplasm of cells of the visual pigment layer, in the ganglion cell layer and also in the inner nuclear layer [45]. Despite the lack of information about CRY4 expression in the retina of migratory birds, it is suggested that this cryptochrome may indeed be involved in magnetoreception due to several factors. One is that this protein has been identified in animal species where magnetic orientation and magnetism-based behaviour are well documented, such as birds, fish and amphibians [46]. In addition, CRY4 has been found to undergo structural modifications in the carboxyl-terminal region in a light-dependent form, which is consistent with the light-dependent magnetic orientation mechanism of birds and the radical pair model [46, 45, 57, 58]. How-

ever, these are very recent observations and information about the behaviour of this cryptochrome as a magnetoreceptor still remains not fully understood [44]. Overall, although with the evidences obtained so far it is suggested that CRY1a is the cryptochrome that best meets the requirements for being considered a magnetoreceptor molecule [44], we decided to include the three genes, *Cry1*, *Cry2* and *Cry4* in the analyses, based on the existing evidences mentioned earlier.

Iron Based Magnetoreception

In the Iron Based models, it is proposed that iron containing molecules play a fundamental role in sensing the surrounding magnetic field. These iron "sensors" are suggested to be composed of biogenic magnetite (Fe_3O_4, iron oxide), initially studied in bacteria, in which the responsible genes for the extraction of iron from the environment and its conversion to magnetite were deduced. However, no orthologs were found in animal species [7, 59].

Biogenic magnetite holds different magnetic properties based on the size of its particles. In sufficiently small particles (between 0.04 μm and 0.12 μm) a permanent magnetic moment is observed. On the other hand, in larger particles the magnetic moments of the domains tend to cancel each other out [7, 44].

It is suggested that the biogenic magnetite can be used in two forms. One of these forms is based on a small piece of iron, with permanent magnetic momentum, that spins in alignment with the Earth's magnetic field (like a compass needle). An alternative form, somewhat less intuitive, suggests the existence of superparamagnetic iron molecules with no permanent magnetic moment that can become magnetized according to the properties of the surrounding magnetic field. These iron molecules will, thus, be responsible for transducing information in respect to the magnetic field [7]. In both models, how this information is transmitted, between specialized cells that carry magnetite particles and neurons, is still poorly understood. These cells should "transform" the alteration in position of the iron molecules and the alteration in magnetization into a change in the electrochemical signals, that will lead to a change in the neuronal activity. It is proposed that, for this to happen, iron particles may change the conformation of intracellular enzymes, of which the product may indirectly induce changes in the neuronal activity. Another alternative is the tension (magnetically generated) that magnetite may exert on the cell membrane, that will lead to a change in the properties of the ion channels [7]. However, as mentioned earlier, a number of studies are still needed to explore magnetite-based mechanisms of magnetoreception, including the formation of iron clusters within cells [39, 44, 7, 60].

ISCA1 (Iron- Sulfur Cluster Assembly)

Although CRY is considered a good candidate to play a role as magnetoreceptor, and, although the structures of some types of cryptochromes are well documented, it alone does

not seem to meet all the requirements and does not explain some observations related to the avian magnetoreception. It is thus proposed that, since the interior of the cell is filled with several molecules, the cryptochrome interacts with other partners within the cell [61, 62].

A possible partner, initially proposed by Qin et al.(2016) [63], is an *Iron-Sulfur-Cluster Assembly Protein (ISCA1)* complex. ISCA1 is considered to be a good candidate for an interaction partner with several cryptochromes, as it has intrinsic magnetic properties, the most important being the ability to bind to iron-sulfur clusters, Fe2S2. This may bring several benefits since the magnetic iron atoms may further enhance the magnetic properties at the ISCA1-CRY interface. Also, the Fe2S2 clusters, known for their participation in electron transfer reactions, may be involved in electron transfer with the CRY. Thus, it is considered that interaction with the ISCA1 complex may further enhance the magnetic orientation and bird's sensitivity to the magnetic field [61].

In a recent study, by Friis et al.(2017) [62], through computational modelling, the structure of ISCA1 was reconstructed and, stability and dynamics analyses of the ISCA1-CRY complex were performed. In this study was found that, contrary to what was expected and initially proposed by Qin et al.(2016) [63], it is unlikely that the interaction and connection between these two protein polymers is essential and robust enough for the electron transfer and, therefore, for improving the magnetic properties of CRY. However, this same study states that the ISCA1-CRY binding was computationally verified and, therefore, do not exclude the possibility that CRY is connected to isolated ISCA1 segments. Friis et al.(2017) [62] also suggest that more studies need to be performed in other types of systems and in several species that have a strong magnetoreception component, in order to verify if this interaction really has advantages in the electron transfer.

In addition, more recently, studies by Kimø et al.(2018) [61] have focused their research on the interaction between CRY4 and ISCA1 and found that, although bonding between these two protein polymers is possible, it is insignificant to improve and facilitate the electron transfer, which excludes the possibility of ISCA1 being a potential partner of CRY4 [61].

Despite these observations, and as much remains to be studied about this possible interaction, it was decided that in this work, the *Isca1* gene would also be included for analysis.

2.2.3 *Thermoreception*

Thermoreception, the ability of organisms to estimate temperature, is a sensory modality with extremely importance for their survival. Temperature is a crucial factor for organisms' homeostasis. In addition, many biological processes are largely temperature dependent. Thus, it has become necessary for organisms to adapt over the course of evolution, to sense and tolerate possible temperature changes in their surrounding environment without compromising their survival [10, 64]. An example of the importance of thermosensa-

tion/thermoreception is the fact that noxious cold and heat trigger negative sensations in the organisms, which causes them to act in a way to avoid such conditions. Thus, if this sensory modality is "damaged" and fails to sense and detect such conditions, it may result in damage of several types of tissues, or even death of the organism [65, 66, 67].

In this way, in order for thermosensation to activates in the organisms' body, it is necessary to initiate the transduction signal that allows thermosensation to occur, and for that there is a set of temperature sensitive ion channels that are responsible for allowing this to happen. These channels belong to the *Transient Receptor Potential (TRP)* channels superfamily, and the members of this family that are temperature sensitive and have functions involved in thermoreception are known as thermoTRP channels. These channels are essentially expressed in the plasma membranes of sensory neurons. They are also calcium permeable and are able to behave as either physical or chemical receptors [10, 68]. Moreover, a major feature of these ion channels is that they are multimodal receptors, i.e., besides being triggered by temperature, they are also activated by other types of physical and chemical stimuli. This means that, consequently, TRP channels are also involved in other types of biological processes. Since these channels are calcium permeable, whenever there is a calcium influx into the cell, will trigger signalling cascades from several calcium dependent reaction networks, and networks can be related to thermoreception or not [10]. In fact, it is known that in addition to temperature, TRP channels can sense pressure, voltage and even osmolarity [65].

As far as thermosensation is concerned, thermoTRP channels are spread over several families of TRP channels. These families include *TRP ankyrin (TRPA)* channels, *TRP vanilloid subtype (TRPV)* channels, *TRP-Canonical (TRPC)* channels, and *TRP melastatin subfamily (TRPM)* channels. For birds, regarding the thermoTRP channels repertoire within the several channel families, they own the following: TRPA1, TRPV1, TRPV2, TRPV4, TRPC5, TRPM2, TRPM3, TRPM5 and TRPM8 [65, 68].

In addition, these different channels are capable of sensing distinct temperature ranges, some of which are specialized in sensing warm/hot temperatures, while others sense cold temperatures. In this way, they are divided into two groups: The hot-sensitive TRP channels and the cold-sensitive TRP channels. Although it is a recent topic and only detailed information for mice and human is known, the hot-sensitive TRP channels group includes TRPV1, which is known to be reactive to hot-painful stimuli, TRPV2, and TRPV4 that responds to warm stimuli. In addition to these, it is believed that the hot-sensitive channels group also includes some TRPM members. Although this group of TRPMs is less studied than the TRPVs, it is known that, at least in a few mammal species, three channels respond to hot and warm stimuli, which are TRPM2, that is non-sensitive to voltage; TRPM3 and TRPM5. In what concerns the cold-sensitive TRP channels group, also a topic that needs to be further studied, TRPM8, which is voltage-dependent is considered to be a member of this

group as well as TRPC5, that has been shown to be sensitive to mild cooling [65, 68, 64, 69]. As for TRPA1, its specificity varies greatly among species. Some studies state that TRPA1 is sensitive to cold while others report that it is sensitive to heat, or even that it is not related to thermoreception, suggesting that this channel would have undergone a number of changes throughout evolution. However, more recent studies, through cloning and characterization, have found that in the chicken system, this channel is heat sensitive. They also found that TRPA1 is sensitive to noxious chemical stimuli [10, 70, 71].

Thus, as mentioned earlier, since thermoTRP channels present several functions beyond thermoreception, many studies claim that in the several vertebrate species existing, they have undergone several episodes of alteration and genetic adaptation throughout evolution. Examples of these events are gene multiplication, gene duplication, gene deletion, and even point mutations. This ensured new abilities to the vertebrate species, such as resistance to acidic environments, resistance to noxious chemicals, etc., which in turn allowed species to ensure their survival by supporting environmental changes that have taken place over time [72].

Thus, given the importance that thermoTRP channels have at the thermoreception level, these were included in this work. All members of the thermoTRP channels group, mentioned throughout the text, existing in birds, were included.

2.2.4 *Photoreception*

Photoreception is also a crucial sensory system for the survival of several species of organisms, and, which consequently, is often influenced by ecological and environmental conditions.

Colour vision, due to the crucial role it plays in organisms' survival, essentially in social interactions, partner choice, predator avoidance, and foraging, is constantly under strong natural and sexual selections [9].

In addition, throughout evolution, so that organisms could assess better environmental signals, for instance, different wavelengths of light, there was a need to improve visual abilities, such as, increasing the capacity for capturing photons and the detection of contrast between objects. These evolutionary advantages are observed specially in the opsin genes, since a great amount of opsin proteins are directly connected with colour vision [73, 74].

In what concerns the avian system, unlike mammals, in which photoreception is hold only by cons, rods and retinal ganglion cells, they also present several extraocular photoreceptors in several types of tissues, for instance in the cerebral tissue and in the pineal gland, which plays a crucial role in the circadian cycle regulation [75, 76].It is important to note that, all the photoreceptors belong to the opsins family.

Opsins, receptor molecules of approximately 350 amino acids in length are the molecular basis for colour vision and membrane-associated GPCRs [73, 9]. The way these proteins function is based on the absorption of the captured photons, converting them into electrochemical signals, which in turn triggers a cascade of visual translation, triggering a neuronal response that will be perceived by the brain, and, therefore, result in perception of light [9, 74].

Opsins can be characterized and phylogenetically divided into five subfamilies. One of these subfamilies are the visual opsins, that includes the rhodopsin, RH1, and the conopsins, responsible for the tetrachromatic vision in birds, RH2, OPN1SW1, OPN2SW (Short Wave Sensitive opsins) and OPN1LW (Long Wave Sensitive opsin). A second subfamily are the melanopsins, which includes two paralogous genes, OPN4m and OPN4x. The remaining three subfamilies belong to the non-visual opsins group, since they are involved in non-image forming reactions to the presence of light. A subfamily belonging to this group is the vertebrate non-visual subfamily, which consists of two types of opsins, the encephalopsin, OPN3, and the teleost multiple tissue opsins, TMT1 and TMT2. Also belonging to the non-visual opsins, is the pineal subfamily which includes the parietopsin, PARE, the parapinopsin, PARA, the pinopsin also known as P-opsin, PIN, and the ancient opsin, VA or OPNVA. Finally, the third subfamily is the photoisomerase group, containing the RGR, RRH and the neuropsin OPN5 [73, 77].

In what concerns the size of the genetic repertoire, opsin genes present a great variation among taxa, and, in addition, the state of conservation and selective pressures observed in these genes also present great differences [9]. In fact, several studies have suggested that multiple molecular processes contributed to the evolution of opsin genes, leading to the development of new phenotypes and adaptation to different conditions. This evolution is thus considered a rather rapid and dynamic process [9].

Studies developed by Escobar et al.(2017) [78] and studies by Steib et al.(2017) [79], in which environmental influences on the evolution of opsins in fish species were studied, revealed that the environment, in which the species inhabits, plays a crucial role in the evolution of the visual system, being that the diversification of opsins would have arisen due to occurrences of gene conversion, gene loss, gene duplication and mutation occurrences.

There are also studies conducted by Borges et al.(2015) [77] in which the selective forces and rates of loss and gain of genes in multiple bird species are assessed. Since birds are fairly visual animals with several adaptations to light conditions, that may give some clues about the phylogenetic evolution of opsin genes [77]. From these studies it was possible to infer that these genes would have been influenced by strong stabilizing selections. They also concluded that the ω-ratio is lower in birds (0.16) than in mammals (0.21), suggesting that the mechanisms of colour discrimination are more stringent in birds than in mammals.

Thus, the fact that the connection between the opsin genotype and phenotype is so well characterized, and since spectral sensitivity changes can be caused by changes in the protein coding sequence and in the level of gene expression, makes opsins and opsin genes a valuable object of study for sensory adaptation analyses [74].

For the development of this work, a set of opsin genes was selected, in order to represent photoreception. Since birds rely on a specialized visual system for their survival, the subfamily of visual opsins (rhodopsin and conopsins) was included in this analysis. The respective genes are *Rh1*, *Rh2*, *Opn1sw1*, *Opn1sw2* and *Opn1lw*. As representation of the pineal subfamily, we included the P-opsin or pinopsin (PIN or, also known as OPNP), since there are only a few studies regarding this opsin. Also, the expression of OPNP is limited to birds and it plays a crucial role in the circadian rhythm regulation [80, 76]. In the development of this work, we refer to P-opsin as OPNP, and it respective gene, *Opnp*. Moreover, we included the ancient opsin (*Opnva*), because studies suggest that this photoreceptor, is located in bird's hypothalamus and that could be involved/mediate the detection of light and day length [81, 75].

2.2.5 *Auditory system*

The auditory system, as the systems previously analysed, is a fundamental system for the survival of all organisms, and with special importance for birds. This system, through vocalization, allows birds a variety of functions ranging from communication (through the transmission of auditive signals that can be simple calls or even very complex sounds), partner choice [82] competition between individuals, avoidance of predators and even foraging [11].

Regarding the structure of the auditory system in birds, despite its small size, it is quite complex and specialized, which allows it to accomplish good sound perception and good auditive performance [83]. It is essentially constituted by three components: The tympanic membrane, the outermost layer surrounding the ear, since there is no external structure beyond this membrane, as observed in mammals; the middle ear and the inner ear. The tympanic membrane is responsible for absorbing sound waves and transmitting them to the fluids found in the membranous labyrinth of the inner ear (endolymph). This transmission of sound waves from one layer to another is possible due to the middle ear which, through the movement of the muscle that is presented there, touches the tympanic membrane and allows it to be carried to the inner ear [84, 85].

Concerning the inner ear, in addition to the labyrinth and the endolymph secretory epithelium, there is also the basilar papilla which is composed of a set of sensory cells that are the main receptors responsible for sensing and perceiving the auditive signals, the hair cells. Each of these particular cells has specific responses and characteristics to different

frequency ranges, and these cells are generally divided into two groups: short hair cells and tall hair cells. Short hair cells are, as their name implies, hold shorter cilia and are found in the thinnest area of the basilar membrane (abneural edge zone), responding to higher frequencies. On the other hand, tall hair cells, which hold cilia with longer size, are located in the thicker area of the membrane (on the neural side). They also respond to low frequencies and appear to connect more closely among each other than short hair cells [85, 86, 87]. Although these facts about the structure of the auditory system are found throughout the several bird species, each species has its own pattern and distribution of hair cells in the inner ear, and, as a result, the species differ in their sensitivity to ranges of frequency [87].

Focusing now on the sounds that birds can perceive, it was found that sounds and vocalizations deriving from the same species (conspecific sounds) origin greater neuronal activity when compared to heterospecific sounds. In addition, the perception of conspecific vocalizations has had strong effects throughout evolution including in speciation, in reproduction and in isolation [88, 89, 11]. Furthermore, studies have reported that the *Zenk* gene is expressed in bird auditory regions and that the number of neurons expressing *Zenk* selectively increases expression of this gene when exposed to conspecific auditory stimuli [90]. *Zenk* is a transcription factor that regulates neuronal plasticity and belongs to the set of genes that are induced in response to external stimuli, the *Immediate Early Genes (IEGs)* [11]. In addition to the *Zenk* gene, it was also found that the *Fos* and *Arc* genes, also belonging to the IEGs group, are expressed in larger amounts in the auditory forebrain of birds when subjected to conspecific auditory stimuli [91, 92]. Although such genes are known to influence species-specific vocalization and perception, the identity of many other genes that are also responsible remains unknown [93].

One other gene known to be related to the auditory experience of heterospecific sounds is the *Chrna3* gene, the α-3 subunit of a member of the nicotinic cholinergic receptors. This gene is related with sensory gating and was differentially expressed in a study, with *Taeniopygia guttata*, when exposed to sounds deriving from another species [94]. This may explain the reduction of birds' neuronal activity when subjected to non-specie-specific sounds, as sensory gating is activated and inhibits the sensory system [11].

Besides genes related to conspecific and heterospecific sounds, there is a set of other genes that are extremely important for the correct functioning of the auditory system, having an important role in the transportation of K^+ ions in the inner ear. Potassium ion is known to be the leading charge carrier in the inner ear of vertebrates. With regards to the bird's auditory system, hair cells depend on the use of this ion's efflux systems since the influx of K^+ into the hair cells triggers a mechanoelectrical transduction in the inner ear, thus activating the auditory system [95, 96]. Moreover, studies developed by Wilms et al.(2016) [97] have shown that in mouse and in chicken, the secretory epithelium in the in-

ner ear has a high conservation of the mechanism responsible for the K^+ secretion into the endolymph. However, although there are similarities about this mechanism between birds and mammals, in what concerns genetic expression of K^+ transporters, the situation is different. Thus, Wilms et al.(2017) [98] found that in birds, particularly in analyses developed using chicken models, occurs the expression of several K^+ efflux systems, which include the *Potassium Voltage-Gated Channel Subfamily Q (KCNQ)*4, *Potassium Calcium-Activated Channel Subfamily M alpha 1 (KCNMA1)* and *Potassium Calcium-Activated Channel Subfamily M beta 1 (KCNMB1)*. *Kcnq4* gene encodes the KCNQ4 channel and *Kcnma1* and *Kcnmb1* genes encode the 1-α and 1-β subunits of the Ca^+ activated potassium channel, respectively. Also, there was an additional expression of the *Kcnq1* gene, which also has efflux functions in the avian hair cell, as well as the expression of the *Nkcc1* gene in support cells, which has been shown to have transportation functions of K^+ and Na^+ along membrane [98].

Thus, given the different sound perception mechanisms of birds, all the previously mentioned genes were used as representative molecular markers of the auditory system.

2.2.6 *Tactile system*

The Tactile System, the last system to be analysed and included in this study, as the other sensory systems, plays a very important role in the survival and development of numerous organisms. This sensory system allows the contact with external objects to be perceived, and also, allows this information to be sent to the central nervous system, where the signal is processed, and recognition of the surrounding environment is performed. Moreover, it is also quite important in social contact among individuals. In this way, the sense of touch is the perception of a variety of mechanical stimuli (from harmless stimuli to noxious chemical stimuli) that come into contact with the skin [12]. Additionally, this is considered the least vulnerable sense among the other sensory systems, however, naturally, it can be damaged under several pathological conditions [99].

In the tactile system the main responsible molecules are the mechanoreceptor neurons, and the vast majority of these cells are located along the several layers of the skin, feeling several mechanical stimuli including vibration, pressure, acceleration and stretch. In addition to detecting these mechanical stimuli, they convert these stimuli into electrical signals and transmit the pertinent information to the central nervous system [12, 100]. These neurons are divided into two main groups: The *Low-Threshold Mechanoreceptors (LTMRs)* and the *High-Threshold Mechanoreceptors (HTMRs)*. The LTMRs are associated with the perception of benign pressure and touch, while HTMRs react to mechanical stimuli that hurt and harm the organism [12, 101].

All of these mechanosensory neurons are known to depend on certain proteins and ion channels to perform these functions, yet little is known about them. This is because these

cells are widely dispersed throughout the skin, making them difficult to collect for molecular studies. In addition to this factor, it is estimated that the number of channel complexes involved in mechanotransduction is low in each cell, which contributes to make their study difficult [102]. Despite such difficulties, some families of channels and proteins that play key roles in this sensory system are already known. The vast majority of these proteins were initially identified in nematode *C. elegans*, and several studies have been conducted with other organisms to corroborate these findings [100].

Currently, about six ion channel families are considered as potential candidates to molecules responsible for mechanotransduction and mechanosensitivity, many of which are present in mechanoreceptors and somatosensory neurons. These families are: *Degenerin/Epithelial Sodium Channels (DEG/ENaC)*, *Acid-Sensing Ion Channel (ASIC)*, Piezo proteins, *Transmembrane channel-like (TMC)*, K^+ channel subfamily, and TRP. The last family mentioned will not be analysed in this context as it shares functions with thermoreception and has already been explored in that section [12, 102, 100]. As regards the first family mentioned herein, the DEG/ENaC superfamily, this includes a set of membrane proteins composed of two transmembrane domains, and coupled to these domains a large extracellular domain. Already known examples of these proteins, belonging to the group of *Epithelial Sodium Channels (ENaC)* are α-ENaC, β-ENaC, and γ-ENaC [102, 100]. Within this superfamily there is a subdivision, in vertebrates, that includes the ASIC. The ASIC subfamily is composed of three members, ASIC1, ASIC2 and ASIC3, with chicken having only ASIC1 and ASIC2. Their structure is very similar to the structure of DEG/ENaC channels and are cation permeable. In addition, they are also sensitive to extracellular acidification and are thus activated by this factor. The role of these three members was essentially studied in mice, where, it was at least found that ASIC2 knockout caused a decrease in the sensitivity of the LTMRs neurons [12, 100, 101]. Regarding the group of Piezo proteins, these, also membrane proteins, have two members, Piezo 1 and Piezo 2, in vertebrates. They have been increasingly recognized as proteins with mechanoreceptor functions [103] as they are expressed in several types of mechanosensitive cells [101]. Finally, the TMC and K^+ channels subfamily groups, besides tactile systems, are both associated with the auditory system, and are said to have mechanotransduction functions. The TMC group has two members, TMC1 and TMC2, two membrane proteins, which have been found to influence mechanotransduction in hair cells [12, 104]. In the case of K^+ channels, some subunits of the *Potassium Channel Subfamily K (KCNK)* members (from K^+ channel domain two-pore family, K2P) that are associated with mechanical gating and expressed in somatosensory neurons are known: KCNK2, KCNK4 and KCNK10 [12].

Thus, given the variety of proteins and ion channels involved in the tactile system with mechanoreceptor functions, all the genes responsible for coding the proteins mentioned

above were included in this work. This way it is possible to obtain a good representation of this system.

2.3 SEQUENCING TECHNOLOGIES

Nucleic acid (DNA or RNA) sequencing refers to a method for determining the specific order and position of the nucleotides belonging to a given DNA or RNA molecule. The order of the nucleotides that constitute these macromolecules is of utmost importance, since it stores all the information that regulates the biochemical and hereditary properties necessary for terrestrial life to be possible [105]. Thus, this allows us to answer fundamental questions of biological research. As far as evolutionary biology is concerned, through analyses of genetic variation of populations, it may be possible to understand several phenomena that may have occurred throughout evolution, for instance, hybridization, adaptive genetic variation, phenotype adaptation, among others [19].

The development of techniques to achieve DNA sequencing began in the year of 1950. Frederick Sanger and colleagues succeeded in sequencing the first protein, the insulin. For such, Sanger fragmented the two chains of the protein, analysed each fragment and overlapped the fragments in order to produce a complete sequence. With this possibility of sequencing, until the year of 1960, numerous proteins have been sequenced, and this has shown that proteins have specific amino acid patterns. It also became possible to understand that each protein sequence varied among species and among individuals [106].

Based on this achievement, and based on the techniques developed so far, it became possible to sequence the RNA molecule for the first time. Examples of sequenced RNA molecules are microbial RNA, transfer RNA, and ribosomal RNA. The fact that RNA does not have two complementary strands, and the fact that RNA molecules are usually smaller than DNA molecules, makes RNA more reachable and easier to handle. And it was a factor that facilitated the sequencing of this molecule [105].

2.3.1 1^{st} *Generation Sequencing*

In 1975, Frederick Sanger developed the first fast DNA sequencing technology (first-generation sequencing), known as Sanger sequencing (with chain-terminating inhibitors). This sequencing technology is based on the principle and biochemistry of DNA replication *in vitro* and involves the use of *Dideoxynucleotides (ddNTPs)*. ddNTPs are chemical analogues to *Deoxyribonucleotides (dNTPs)*, DNA monomers, in which they simply differ in the 3' position carbon. ddNTPs lack the hydroxyl group, which is typically attached to the 3' carbon. This group is crucial for allowing the phosphodiester bonds to form between nucleotides of the same chain (the 3' carbon hydroxyl group of the first nucleotide binds to the 5' carbon

phosphate group of the next nucleotide), which, in turn, allows the DNA extension by the DNA polymerase [107].

The process to make this sequencing possible is quite simple. Multiple copies of the DNA template strand, in the 3 '- 5' direction, are used, as well as multiple copies of the complementary primer to this template strand (primers are short sequences of oligonucleotides required for DNA polymerase to initiate the extension of the complementary strand). In addition, all dNTPs (dATP, dCTP, dGTP, dTTP) are used, as well as the ddNTPs (ddATP, ddCTP, ddGTP, ddTTP), but in smaller amounts. Thus, the process starts, in a very similar way to the DNA replication process, in which DNA polymerase adds to the new strand (which is in the 5'-3' direction) the complementary nucleotides to those of the template strand. Nucleotides chosen by DNA polymerase may be either dNTPs or ddNTPs. Once a ddNTPs is incorporated into the new strand, the extension process stops. This is due, as noted above, to the absence of the hydroxyl group that blocks DNA polymerase from incorporating the next nucleotide into the strand. This process occurs several times and by the end of the process fragments (multiple complementary copies to the template) of several sizes are obtained. Finally, these fragments are placed in the electrophoresis polyacrylamide gel. An autoradiography of the same gel is obtained and the fragments are ordered from the smallest (those that migrated the most into the gel) to the largest (those that migrated less), so it is possible to understand the exact order of the nucleotides of the sequence that is being analysed [107]. For a better understanding, a scheme of this technology is presented in Figure 4.

This method allowed scientists to reach a high per-base accuracy, about 99.999%, and allowed a sequence of approximately 1000 bases to be acquired in a single reading. Furthermore, it was possible, through the Human Genom Project, to sequence the first human genomic sequence [108, 109].

Thus, this sequencing method was widely used for decades, and still is today, in routine sequencing applications, and to validate NGS data. However, with the vast increase of DNA sequencing, there was a need to develop cheaper, faster, larger-scale and more accurate sequencing methods. Moreover, Sanger sequencing presents some imperfections, such as poor sequencing quality in the first 15/40 bases (where the primers bind), only small fragments of DNA can be sequenced, and the sequencing quality starts to deteriorate from the 700/900 bases. Thus, between the years of 2005 and 2010 the NGS methods, also known as second-generation sequencing, started to emerge [105, 19].

2.3.2 2^{nd} *Generation Sequencing*

NGS platforms are based on performing multiple parallel sequencing of small DNA fragments from the same sample. As a result, they brought with them several advantages. Be-

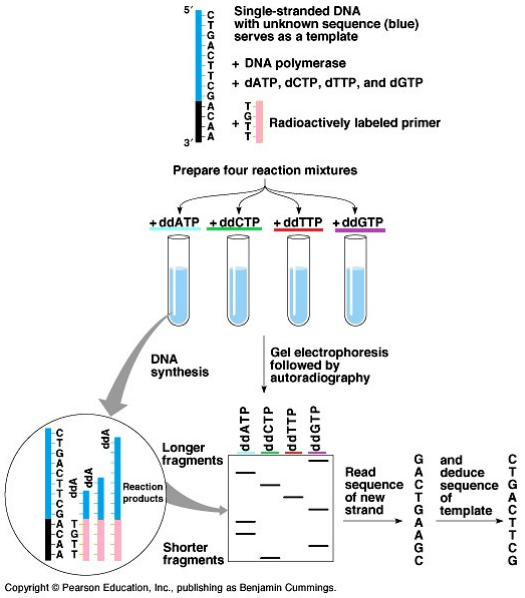


Figure 4.: Schematic representation of Sanger sequencing. This sequencing technology is based on DNA replication *in vitro*. In this process, the DNA polymerase adds complementary nucleotides to the template strand. The nucleotides are randomly chosen by polymerase (it can incorporate either a dNTPs or a ddNTPs). As soon as a ddNTPs is incorporated into the new strand, the extension process stops, forming a specific length fragment. This process is repeated several times, obtaining several different size fragments. The fragments are placed in an electrophoresis gel, and from an autoradiography analysis, these are ordered from the smallest to the largest, allowing to understand the exact order of the nucleotides.Adapted from Pearson Education, Inc., publishing as Benjamin Cummings.

cause they are more accurate, they allowed the data obtained to bring relevant information related to variations in DNA. In addition, higher-throughput sequencing can be performed at lower costs than Sanger sequencing. Furthermore, NGS made possible to completely sequence small genomes in just one day, or even specific genome regions, such as coding genes [109, 110].

Due to these characteristics, NGS methods have a great potential to be applied in several areas. One of these areas is clinical genetics (in human health and disease), as they reveal several types of mutations in the human genome (gene substitution, deletion, insertion, inversion and translocation, deletion of exons or full genes), in contrast to Sanger sequencing that only detects point mutations and small deletions and insertions. In addition, NGS technology can be used to perform *de novo* genome sequencing, i.e., to fully characterize the entire genome of a given species. They may also be applied in the oncology area, assisting in the treatment of several types of cancer due to the possibility of sequencing the cancer subgenome; in microbiology; in expression analysis; in metagenomic; among other areas [111, 110].

In general, NGS are classified in several categories according to the technical details of their sequencing method. These categories include microelectrophoretic methods, sequencing by hybridization, real-time observation of single molecules, and cyclic-array sequencing. In the last category mentioned, there are several platforms available and widely used today, such as 454 pyrosequencing (the first platform available as a commercial product), SOLiD platform, Polonator, Helicos, Solexa, later acquired by Illumina, among others [112]. In this work, the analysis will focus on the Solexa and Illumina platforms, due to their influence on this work, which will be explained later.

Solexa, after platform 454, was one of the first platforms to be commercialised, with which it was possible to sequence the whole Bacteriophage phiX-174 Genome (in the year of 2005), previously also sequenced by Sanger sequencing. With Solexa, it was possible to obtain more sequence data, and a delivering of over 3 million bases from a single run was achieved [113]. Later, in 2007, it was acquired by Illumina, and therefore both have the same sequencing process. Since 2012, Illumina became one of the most widely used sequencing platforms [106].

The sequencing technique of Illumina consists of four basic steps: Sample Preparation, Cluster Generation, Sequencing and Data Analysis (Figure 5). In the Sample Preparation step, the DNA molecule is cleaved into several fragments. These fragments are denatured to produce single stranded fragments, and, at both ends, the adapters are attached. Adapters are chemically synthesized oligonucleotides that can bind to the ends of DNA and RNA molecules, allowing template fragments to remain immobilized on the surface of a flow cell (this is possible since the flow cell already has an amount of oligonucleotides that are complementary to adapters).

This proceeds to the Cluster Generation step. At this stage, DNA polymerase generates, for each immobilized fragment, the complementary strand. The double stranded newly formed is then denatured and the original template is washed away. The remaining immobilized strand folds over, and the adapter region hybridizes to the complementary oligonucleotide present in the flow cell. The polymerase thus generates the complementary strand, and, again, the double strands are denatured. Thus, strands are clonally amplified in a process known as Bridge Amplification. This process occurs simultaneously in several clusters of strands and is repeated over and over again, resulting in a clonal amplification of all fragments, yielding several million dense clusters. The next step comprises deleting all reverse fragments, leaving only the forward fragments bound to the flow cell.

After this stage is completed, the Sequencing step follows. In it, dNTPs labelled with different fluorescence tones are added, which complementarily bind to the nucleotides of each strand in each cluster. However, in each cycle, only one nucleotide is incorporated, and after this addition, the clusters are excited by laser, allowing each cluster to emit fluorescence, and thus allowing the identification of the nucleotide that was incorporated. This

process is known as *Sequencing by Synthesis (SBS)*. This cycle is repeated several times, all clusters being sequenced in a massively parallel process, until the entire sequence is known.

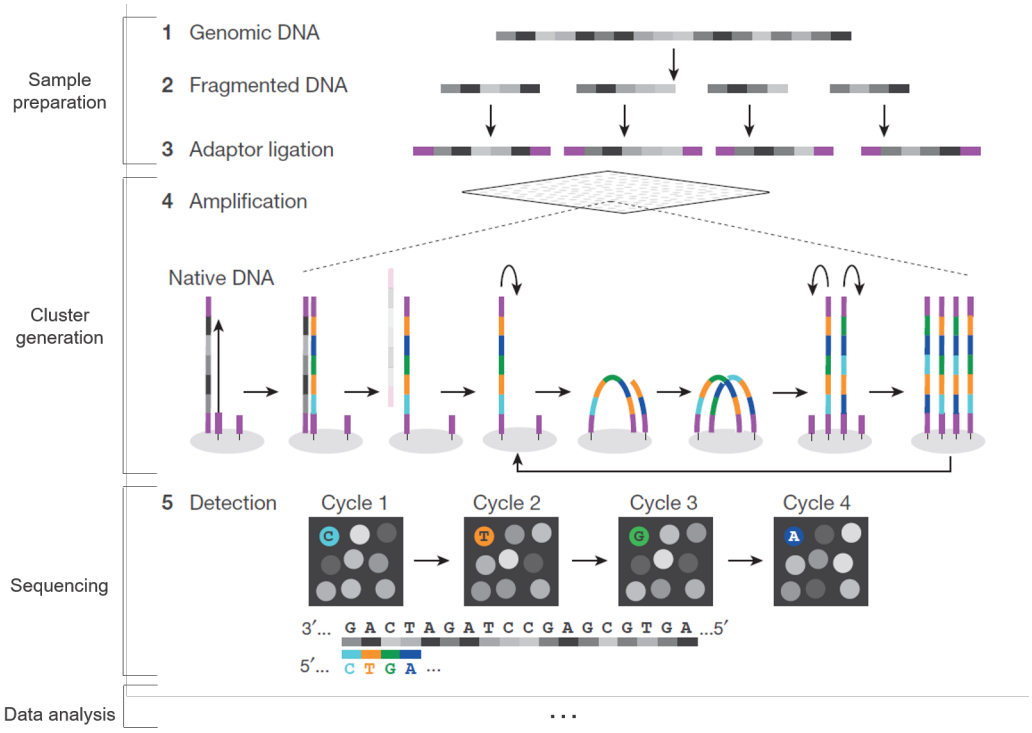


Figure 5.: Schematic representation of Illumina sequencing technology. This NGS sequencing platform includes four main steps, known as "Sample preparation", in which the template DNA is cleaved in several fragments and these are immobilized in the flow cell; the "Cluster generation" in which occurs a clonal amplifications of all the fragments; the "Sequencing" step, in which the clusters of fragments are sequenced in a massive parallel process. This cycle is repeated several times until all the sequence is known. The nucleotides are identified due to the fluorescence emitted when they are incorporated into the new strand of DNA. Finally, the last step, not shown in this figure that corresponds to the data analyses that includes bioinformatic analyses necessary for the alignment and comparison of the fragments. Adapted from [106].

Finally, in the last step, Data Analysis, the obtained reads, each representing each fragment, are clustered based on similarity, and, ultimately, aligned and compared using bioinformatic software [114].

Despite the numerous advantages that 2^{nd} generation platforms have brought, and despite their high throughput, they still present some disadvantages compared to Sanger sequencing. Examples of such disadvantages are the production of smaller sequences (for instance, 70-300 bp in Illumina), some copying errors during amplification, loss of information, and a higher error rate (accuracy >99,5%, while Sanger sequencing presents an accuracy of about 99,99%) [19, 106]. In this way, as an attempt to combine the 2^{nd} genera-

tion speed and high throughput with 1^{st} generation accuracy, and to produce longer reads, the 3^{rd} generation sequencing technologies were developed.

2.3.3 3^{rd} *Generation Sequencing*

Approximately between the years of 2011 and 2014 the 3^{rd} generation technologies came up. In contrast to the previously developed techniques, these allowed to produce reads of unprecedent length, an average of 20 kb. This factor played a key role in increasing the ease of genome assembly, especially of unknown genomes (*de novo* assembly). This was indeed important in the oncology area, specifically on the cancer behaviour studies, since possible new structural rearrangements in the genome may be related to this disease. Thus, this is now possible to analyse due to the *de novo* assembly [19, 115, 108].

Currently, there are two main methods used by 3^{rd} generation technologies. These are, Single Molecule Real-Time sequencing, *Single Molecule Real-Time (SMRT)*-seq, which is used by PacBio and *Oxford Nanopore Technology (ONT)*; and *Synthetic Long-Read sequencing (SLR-seq)*, used by Illumina synthetic long reads and 10x Genomics. Apart from these two methods differing in the throughput, the error rates and also the cost, the main difference between them is based on how the long reads are produced. In SMRT-seq methods long reads are produced from single DNA molecules. In SLR-seq methods, long reads are computationally assembled from small reads obtained from the same molecule [19, 116, 115]. Since in this work, regarding third generation technologies, only genomes obtained through the PacBio technique were analysed, and since this is currently the most widely used long reads platform, only this technique will be explored in here.

The procedure of this sequencing technique is based on the use of the DNA template, known as SMRTbell, which, by attaching the adapter to the ends of the DNA molecule, produces a single stranded circular DNA molecule. This DNA is placed on a SMRT cell that has arrays of nanophotonic chambers (small wells), known as *Zero Mode Waveguides (ZMWs)*, in which DNA polymerase is ready to initiate polymerization. In addition to DNA polymerase, phospholinked nucleotides, which are fluorescently labelled on the terminal phosphate, are also present. Each nucleotide (dATP, dTTP, dCTP or dGTP) is labelled with different colour. In this way, the DNA migrates to the ZMWs and the sequencing process begins. During this process, whenever DNA polymerase adds a new nucleotide to the growing chain, the fluorescent label goes into an excited state, thus emitting a pulse of light that is captured by a sensitive detector. After/during this nucleotide incorporation, the phosphate bond is broken, and therefore the fluorescent label is cleaved, leaving behind a completely natural chain (Figure 6). This process is repeated countless times, while the emitted signals are being recorded, which allows, in the end, to reconstruct the whole sequence. Since DNA polymerase is able to incorporate 10 or more bases per second,

and since hundreds of ZMWs can be working at the same time, thousands of continuous incorporations occur simultaneously, resulting in a high speed (the all process takes less than one day) and in longer reads (from 10 kb to a maximum of 60 kb) [21, 117, 118].

Thus, PacBio platform presents several advantages in the DNA sequencing area. The generated long reads allow to locate repetitive sequences through a single read, which in turn allows to exclude particular ambiguities in what concerns to genomic element sizes and even their positions. Likewise, they allow the sequencing of entire transcripts, which in turn improves metagenomic studies, and, as mentioned earlier, play a key role in *de novo* sequencing and *de novo* assembly. Furthermore, PacBio shows an absence of GC bias (extreme tolerance to GC content) and also allows real-time detection of biological events. Since it is possible to observe the polymerase activity in real time, any modification in the DNA causes a change in the enzyme kinetics, i.e., slowing or increasing speed. Kinetic patterns are specific to each modification, thus, it is possible to understand what specific modification is being detected [115, 111, 108].

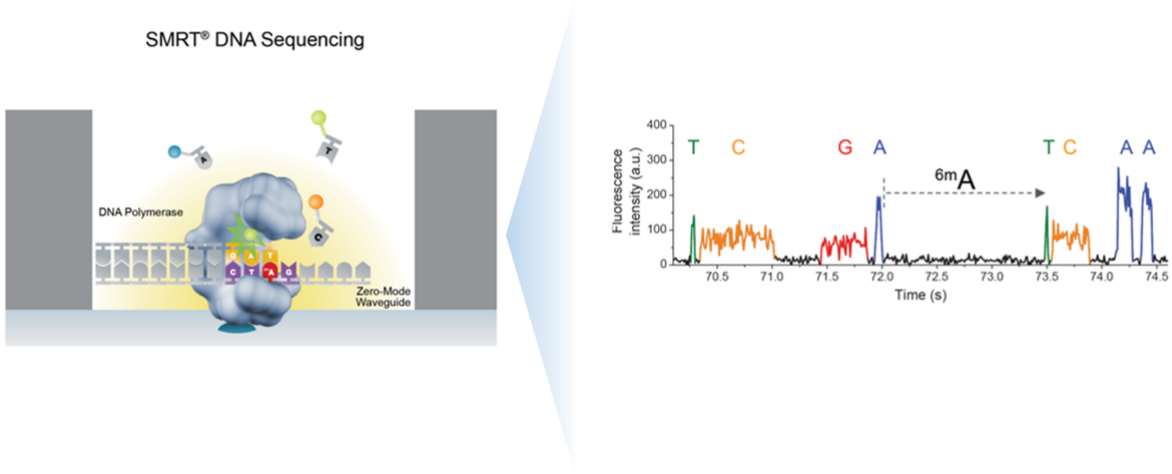


Figure 6.: Schematic representation of PacBio technology. This SMRT DNA sequencing technology is based on small wells, ZMWs, in which the DNA is immobilized and replicated by DNA polymerase. The nucleotides used to produce the new strand are labelled with different fluorescent colours, and each time a new nucleotide is incorporated, a specific peak of fluorescence is detected, and, in this way, the sequence is revealed. This can be monitored in real time, and when the polymerase finds a DNA modification, in this case 6mA, this can be noticed immediately (through measuring the time variation between base incorporations). Adapted from Pacific Biosciences [117], https://s3.amazonaws.com/files.pacb.com/png/basemod_benefits_lg.png.

Despite all the promising features, PacBio sequencing has some imperfections when compared to the other existing techniques. In spite of the quickness of this technique to sequence complete genomes, when compared to NGS methods, it still cannot provide the distinctive high throughput of these methods. For instance, of 150,000 ZMWs wells, only from 35,000 to 70,000 can produce successful reads. One reason for this is due to possible polymerase anchorage errors into the well. In this way, one SMRT cell will be able to generate between 0.5 to 1 billion bases, whereas, the Illumina HiSeq 2500 method can generate up to 167 billion bases per day. Moreover, the PacBio technique presents a higher cost than other techniques and its error rate is much higher than any 2^{nd} generation technique, from 15% to 20%. However, as the errors are randomly distributed in the continuous long reads, this error rate could be reduced, and 99.99% accuracy could be achieved with 50x coverage [19, 20, 21]. Thus, despite the disadvantages, all the PacBio features have numerous applications and no other platform is capable of produce this type of results.

Based on the differences among all the sequencing technologies that have been developed to date, in this work, some of these technologies will be compared, in order to evaluate their quality in selected analysis, such as, gene extraction. The bird species were selected based on the available genomes sequenced to date (Table 1, in the chapter 3 of Materials and Methods). In this way, for each species a genome sequenced using 1^{st} or 2^{nd} generation techniques and also a genome sequenced via PacBio were obtained. Additionally, in order to evaluate the sequencing quality and the success to extract genes from this genomes, six types of molecular biomarkers from different sensory systems were selected. This information can be found in Table 2 in the chapter 3 of Materials and Methods.

3

MATERIALS AND METHODS

3.1 TOOLS AND SOFTWARE USED

According to the objectives proposed for this work, to compare and evaluate the quality of different DNA sequencing technologies, bioinformatics and statistical analyses were performed. These included sequence analyses, similarity searches of genes, alignments execution and searches of gene families among genomes of different species. As a more final analysis, statistical analyses were conducted, in order to understand if there are significant differences in the quality of extraction of different genes from different genomes (sequenced from different technologies).Thus, during this work, to execute this set of analyses with significant biological meaning, several software tools were used.

Information contained in biological databases was crucial for this work. It was assessed considering the following web sites and services: Ensemble [119], *National Center for Biotechnology Information (NCBI)* [120],GenBank [121] and UniProt [50].

To perform the alignments, the sequences analyses, and to proceed with the quality evaluation of the different technologies, a set of tools were of great convenience for this study. These tools include the Exonerate software [122], the MEGA software [123], the R software [124] and Python programming language [125]. All the databases, tools and services are summarized in the diagram presented in Figure 7 and will be described below.

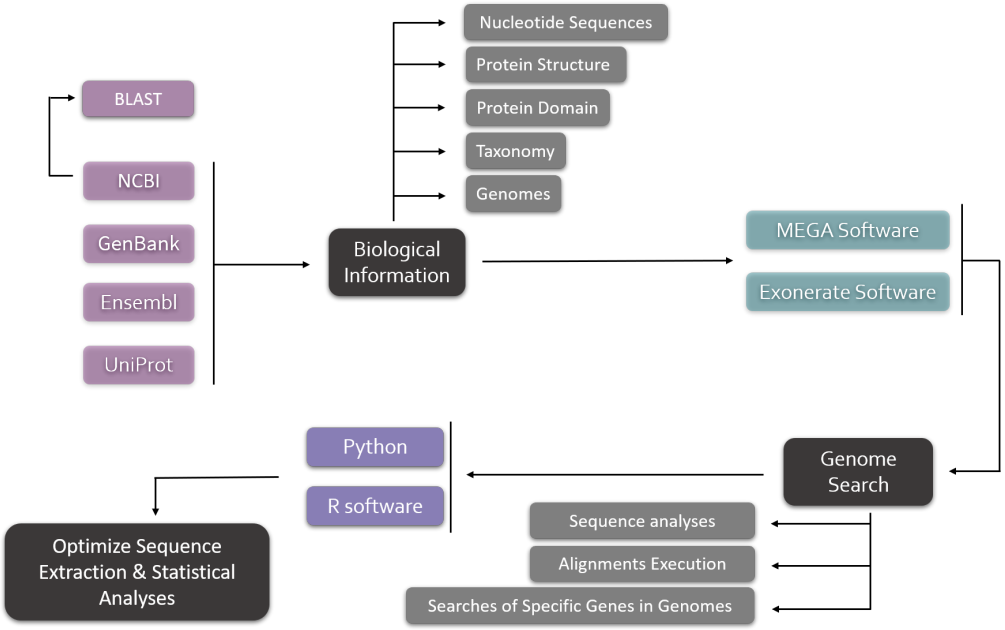


Figure 7.: Representation of the several databases, programs and programming language explored during the development of this work. In this diagram, with the colors light purple, blue and dark purple are represented the tools that were used. In black color is presented, in general, the applicability that these tools had for carrying out the work. In addition, specific and important tasks performed in this work are in light gray, for which the tools were a great advantage. Noted that, with regard to *Basic Local Alignment Search Tool (BLAST)*, in light purple, this tool was widely used throughout the work, through the NCBI computational resources.

3.1.1 *Genome browsers and Biological Databases*

To access the diverse genomic information across the several existent species, the Ensemble project provides a genome browser (`http://www.ensembl.org`) with distinct genome resources and information [119]. It includes comprehensive information for the most accessed genomes: mouse, rat, zebrafish, chicken and human [119]. The resources available in Ensemble include gene annotations, comparative genomics, including alignments and homology, and also gene-level phylogenetic trees [126]. This project collaborates directly with several bioinformatics databases, including *Universal Protein resource (UniProt)* [50] and NCBI [126].

The NCBI web site and available services and databases will also be crucial for the development of this work. This site provides multiple computational resources for the analysis of biological data, for instance, BLAST [127] sequence analysis, and it also provides access to several types of biological data, such as taxonomy, protein structure and domain information and genomes. In addition, it provides access to scientific publications and data from

NCBI's databases such as PubMed and GenBank. All resources are available at the NCBI home page, www.ncbi.nlm.nih.gov [120].

Regarding GenBank, this is a publicly available database, built by NCBI, that provides access to several nucleotide sequences of approximately 300,000 described species. This comprehensive database provides up-to-date DNA sequence information, in which all sequences are obtained by individual laboratories and submitted to this database. It also has the advantage of ensuring worldwide coverage since it exchanges data daily with the *DNA Data Bank of Japan (DBBJ)* and *European Nucleotide Archive (ENA)* [121].

Finally, UniProt, also a publicly available database, that collaborates with other databases, such as *European Bioinformatics Institute (EMBL-EBI)*, *Swiss Institute of Bioinformatics (SIB)* and *Protein Information Resource (PIR)* provides several types of information regarding proteins. UniProt is divided into three databases: *UniProt Knowledgebase (UniProtKB)*, *UniProt Reference Clusters (UniRef)*, *UniProt Archive (UniParc)*, of which UniProtKB was of special importance. This database includes information of numerous annotated proteins, from numerous different species, and it is possible to access a wide range of information about them, including the description, function, structure, amino acid sequence, taxonomy and related publications [50]. Of all this information, the proteins description and the sequence of amino acids was what really helped a lot in the execution of this work.

3.1.2 *Genome Search and Phylogenetic Analyses: Exonerate and MEGA*

One program that proved to be extremely important in the development of this work was the Exonerate software. This program, which can be used on the command line in Linux, is well known for performing alignments, allowing the execution of spliced alignments and performing several types of genome search. Spliced alignment is an alignment method that allows, in addition to finding exonic regions in the genome, to also detect introns that exist between those regions, that is, it allows to find multiexonic genes [128].

When performing alignments, exonerate can execute and produce gapped or ungapped alignments, and accepts both nucleotide and protein sequences as queries, however, the most used approach is based on the use of protein sequences as queries, as it was performed in this work. This is due to the fact that the alignment between a DNA sequence and a protein sequence includes some advantages in obtaining sequence information. It allows, through the use of a homologous protein sequence of that specific region of DNA, to detect possible errors in the DNA sequence, since protein sequences are generally more conserved than nucleotide sequences. In addition, it brings the possibility of quickly locating protein-coding genes and detecting the exon-intron organization of these same genes, through spliced alignment, with a very reliable prediction of its structure [129]. Additionally, exonerate incorporates a *Bounded Sparse Dynamic Programming (BSDP)*, a heuristic

approach. Unlike the exhaustive alignment algorithms, which are much slower, but guarantee an optimal solution to the problem, the heuristic method is much faster, but does not guarantee the best solution (the best optimal alignment). However, in the case of this software, BSDP allows a novel, fast and more accurate heuristic for aligning sequences. Furthermore, this program, although by default performs the heuristic approach, it allows the execution of exhaustive alignments (Smith-Whaterman local alignment), if the user wants to achieve a maximum sensitivity [122]. Exonerate is freely available for download at https://www.ebi.ac.uk/about/vertebrate-genomics/software/exonerate.

To study the phylogenetic relationships among different species and to search through DNA and protein sequences, the MEGA software was used, a user-friendly bioinformatics tool, which allows to understand and infer evolutionary relationships and patterns of evolution. This is a desktop application designed to perform comparative analyses of homologous nucleotide or amino acid sequences, among several species, or even sequences of multigenic families [123]. To do so, it applies statistical methods in the course of phylogenetic analyses, which include *Maximum Likelihood (ML)* methods [130], to estimate the selective divergences among sequences.

Of the several methods available in this tool, which the user can choose based on those that will be most appropriate to the study, these include sequence alignment, evolutionary hypothesis testing, estimation of sequence divergence rate, phylogenetic tree reconstructions, and web-based collection of sequence data.

The user is able to explore input data through modules that allow to edit and calculate statistical metrics. In addition, it also integrates functions, with which, it is possible to select specific species, genes, codon positions and even domains, for the analyses execution [123].

In what concerns the results obtained from the evolutionary analyses can also be explored in several ways, since MEGA contains visual explorers, which may be quite relevant for the results preparation for scientific publications. These are the Tree Explorer and the Distance Matrix Explorer. The Tree Explorer provides multiple representations of the generated phylogenetic tree. It also presents sub-trees, for a better analysis of relationships between sequences to be possible. Likewise, it builds consensus trees, estimates ancestral states of nucleotide or protein sequences, in each node of the tree, and it allows these trees to be exported in formats compatible with the Newick format [131], which in turn can be used in other evolutionary programs [123].

The second one, the Distance Matrix Explorer, presents standard error estimates of the respective pairwise distances calculated between sequences, and has the ability to identify pairs of sequences for which pairwise distances could not be calculated (due to the absence of common positions along the alignment). The latter fact is advantageous for the analysis process of phylogenetically fairly distanced sequences, since in this way these are easily identified [123].

Lastly, this software was quite relevant in the development of this work, considering all the properties previously mentioned. It is freely available in http://www.megasoftware.net (in several versions).

3.1.3 *Statistical analysis tool: R*

For the development of statistical analysis in this work, a well-known tool had a special use, the R software. This program is very useful since it allows the user to manipulate and visualize the data. In terms of data manipulation, R provides tools to perform statistical computation, including a wide variety of statistical tests, linear and non-linear modelling, analysis of several metrics, clustering and, also very important when it comes to visualizing results, high-quality graphical representations. In addition, it is constantly updated and enriched with a huge variety of packages, allowing the user to perform numerous functions of their choice. This program is available for download at https://www.r-project.org/ and can be run on all different operating systems [132, 133].

3.1.4 *Programming language: Python*

Python is a high-level programming language, with a very simple, easily readable and very flexible syntax. This programming language is known to be a very productive language, as it has the advantage of being quite useful for scripting and for the rapid development of programs and applications. Its interpreter, as well as its vast library are freely available on the Python Web site, https://www.python.org/. The applicability and utility that this programming language presents in different areas extends, in addition to the development of web tools, to the scientific area where it has become especially popular, namely in the field of Bioinformatics [cite https://docs.python.org/3/tutorial/index.html; [134].

In bioinformatics, a particular package that Python presents is of special importance, the Biopython package. This package, available at http://www.biopython.org, was specially developed to ease the use of Python for bioinformatics, by generating high-quality, reusable modules and classes. Biopython provides a programmatic interface for sequence manipulation and includes many others important features. It has the skill to parse several file formats (FASTA, GenBank, Blast output, SwissProt and others) into usable Python structures, it has access to online services, like NCBI, and interfaces with programs (ClustalW and Standalone Blast). In addition, it presents several tools to handle biological sequences, including tools for transcription, translation and weight calculations [135].

Thus, presenting these great advantages for bioinformatics studies, and to ease some tasks throughout the work, python was considered a quite useful tool.

3.2 GENOME COMPILATION

For the development of this work, as previously mentioned in section 2.3, genomes of six different bird species were selected, being these species *Anas platyrhynchos* (Mallard), *Calypte anna* (Anna's Hummingbird), *Gallus gallus* (Chicken), *Taeniopygia guttata* (Zebra finch) and *Strigops habroptila* (Kakapo, a parrot).

For each species, a genome sequenced through older technologies, 1^{st} or 2^{nd} generations, (Sanger sequencing, Illumina or Solexa) and a genome sequenced through a 3^{rd} generation technology (PacBio) were obtained. The goal of this task was to allow the comparison of the quality of sequencing between the different technologies, i.e., to understand if, indeed, the PacBio technique (since it produces long reads) presents advantages in what concerns the extraction of specific genes.

All genomes were obtained from the NCBI Sequence Set Browser, `https://www.ncbi.nlm.nih.gov/Traces/wgs/`, in the FASTA format [136]. All the available information regarding each of these genomes was also analysed and recorded (Table 1).

3.3 MOLECULAR MARKERS SELECTION AND QUERY SEQUENCES OBTAINING

Once the genomes were selected and obtained, a collection of molecular markers belonging to several sensory systems was selected. These systems include Chemoreception, Magnetoreception, Photoreception, Thermoreception, Auditory system and Tactile system. The representative genes of each sensory system are presented in Table 2. In some of these systems, such as the chemoreception and the photoreception, selected genes are directly involved in binding/recognition of a molecular target while in others, such as the auditory system, tactile system, and thermoreception, they operate through complex molecular cascades of recognition and the receptors also have other molecular functions.

For each set of genes previously chosen the procedure was the same. Initially, it was extracted from UniProt [50] (UniProtKB) database all the annotated amino acidic sequences of interest from *Gallus gallus* (since this is a model organism, representative of birds). Therefore, in order to ensure that we had obtained all available paralog in *Gallus gallus*, a web tBLASTn [127, 137] was performed in this species, in the NCBI database. New nucleotide sequences (only concerning the *Coding Sequence (CDS)* region) of paralogs available in NCBI databases were collected and translated to protein sequences throughout MEGA software version 5.2. All the obtained sequences where renamed as "query sequences".

Table 1.: Information of all genomes used in this study (size of each genome (in bp), name and number of contigs (#Contigs), assembly date, coverage, and the DNA sequencing technology from which genomes were generated (referred as Technology).

Genome	Species	Length	Contigs	#Contigs	Assembly Date	Coverage	Technology
RHJV01	*Anas platyrhynchos*	1,126,159,488 bp	RHJV01000001 -RHJV01002149	2.149	February 2018	50,0x	PacBio
ADON01	*Anas platyrhynchos*	1,069,956,150 bp	ADON01000001 - ADON01227448	227.448	———	60x	Solexa
MUGM01	*Calypte anna*	2,021,121,536 bp	MUGM01000001 -MUGM01001076	5.971	May 2016	55.0x	PacBio
JJRV01	*Calypte anna*	1,067,027,607 bp	JJRV01000001 -JJRV01124820	124.820	2014	110x	Illumina HiSeq
MUGN01	*Taeniopygia guttata*	1,982,686,095 bp	MUGN01001160 -MUGN01003347	3.347	June 2016	100.0x	PacBio
ABQF01	*Taeniopygia guttata*	1,222,847,868 bp	ABQF01000001 -ABQF01124805	124.805	Jully 2008	5.5x	Sanger
AADN05	*Gallus gallus*	1,054,617,308 bp	AADN05000001 -AADN05001583	1.583	———	82x	PacBio
PDMY01	*Gallus gallus*	1,021,022,236 bp	PDMY01000001 -PDMY01001822	1.822	June 2019	243.8x	Illumina HiSeq
RXXE01	*Strigops habroptila*	1,165,639,803 bp	RXXE01000001 -RXXE01000100	100	September 2018	76.1x	PacBio

Table 2.: Set of selected genes for analysis, from each sensory system. Abreviattions: In CHEMORECEPTION: TASR: Taste Receptors, TAAR: Trace Amine-Associated Receptors; In MAGNETORECEPTION: CRY: Cryptochrome, ISCA: Iron-sulfur Cluster Assembly; In PHOTORECEPTION: RH: Rhodopsin, LW: Long Wave Sensitive opsin, SWS: Short WaveSensitive opsin; In THERMORECEPTION: TRPV: TRP vanilloid subtype, TRPM: TRP melastatin subfamily, TRPA: TRP ankyrin, TRPC: TRP-Canonical; In AUDITORY SYSTEM: KCNQ: Potassium Voltage-Gated Channel Subfamily Q, KCNM: Potassium Calcium-Activated Channel Subfamily M; In TACTILE SYSTEM: ASIC: Acid-Sensing Ion Channel, KCNK: Potassium Channel Subfamily K, DEG/ENaC: Degenerin/Epithelial Sodium Channels, TMC: Transmembrane channel-like.

Sensory System	**Genes**
Chemoreceptors	TASR: TAS1R1, TAS1R3; TAS2R4, TAS2R40, TAS2R7; TAAR: TAAR1, TAAR2, TAAR5
Magnetoreceptors	CRY: CRY1, CRY2, CRY4 ISCA1
Thermoreceptors	TRPV: TRPV1, TRPV2, TRPV4 TRPM: TRPM2,TRPM3, TRPM5, TRPM8 TRPA1 TRPC5
Photoreceptors	SWS: OPN1SW, OPN2SW RH: RH1, RH2 OPNVA, OPNP, OPN1LW
Auditory System	ZENK FOS ARC CHRNA3 K+ channels: KCNQ4, KCNQ1, KCNMB1, KCNMA1, NKCC1
Tactile System	ASIC: ASIC1, ASIC2 K+ channels: KCNK2, KCNK4, KCNK10 PIEZO: PIEZO1, PIEZO2 DEG/ENaC: SCNN1A, SCNN1B, SCNN1G TMC: TMC1, TMC2

3.4 GENOME AND SEQUENCES ALIGNMENT

Once the previous steps completed, the queries were aligned with the several genomes. To execute this task, the Exonerate software was used, a generic tool that allows alignment and comparison of sequences [122]. This tool is only available for the Unix/Linux operating system, with a command line interface. The command used to perform the alignments is as follows:

exonerate –model protein2genome -q query.fas -t 1.fsa_nt –showcigar yes –showquerygff yes –ryo ”>%ti(%tab-%tae)\n%tas\n”>2.out

As mentioned previously, the query sequences were aligned with the genomes, hence the existence of the terms “model protein2genome” in the command used. The term “query.fas” corresponds to each set of amino acid sequences from each sensory system. In addition, the term ”1.fsa” represents the genome file, with which sequences were aligned. Finally, the term ”2.out” corresponds to the name that would be assigned to the output file. Due to the large amount of data in genome files, this alignment process was a very time-consuming step.

3.5 OUTPUT ANALYSIS

Once the alignments were performed, we proceeded to the analysis of the output obtained and then, for the extraction of the sequences. The output files contain all the possible alignment results, and it is up to the user to select those with the best values (good values of raw score), in order to extract the gene sequences (belonging to the genome under analysis), that are more similar, or that present a higher homology, to the query sequences. For a better understanding, an example of an alignment result obtained in an output file is shown in Figure 8.

To extract the sequences that presented the best alignment result, it was crucial to consider several parameters, which are presented in Table 3, with their respective meaning and relevance to the analysis. In addition to this parameters, Exonerate software also provides graphical information through particular symbols found between the aligned sequences, as it is also shown in Figure 8. These symbols ”|”, ”!”, ”:”, ”.” and ” ” (blank space), inform about the degree of homology between the query and target sequences. For instance, the more homologous, the greater the number of symbols ”|” found throughout the alignment.

After the output analysis was completed, the nucleotide sequences from the best alignment results were selected. Each of these sequences was saved in a correspondent file. Since Exonerate software generates output files with some difficulty to handle and manipulate se-

Figure 8.: Output example of the Exonerate software. "Query: TAAR2_Gallus_gallus" means that the alignment corresponds to the specific receptor TAAR2. The name underlined, "AADN05000003.1", denotes for the contig in which the alignment was executed. The "Model: protein2genome" means that the alignment was produced between a protein sequence and a DNA sequence. The "Raw score" corresponds to the alignment score, the "Query range" corresponds to the query segment that aligned with the genome, and the "Target range" corresponds to the genome segment that aligned with the query sequence. The symbol "|" that is found between the aligned sequences determines a high degree of homology. The blue arrow represents the direction of the alignment, and the suspension points represent the remainder alignment.

quences (as it is shown in Figure 8), to optimize the sequence extraction a small and simple Python [125] script was developed (this script is presented in the Listings section A, Figure 12). In this simple program, the aim was to remove everything that did not correspond to the nucleotide sequence, i.e., symbols, extra spaces, numbers, and amino acid abbreviations were deleted. What this program does is, initially, starting at the line corresponding to the alignment, it reads the first 4 lines of the file and records the 4^{th} line, since, as can be seen in Figure 8, the 4^{th} line contains the characters corresponding to the nucleotide sequence. After this first registration, instead of reading the 4^{th} line, the program reads the file every

Table 3.: Parameters used to evaluate the alignment results quality, produced by the Exonerate software. Also, is presented the meaning and relevance of each parameter.

Parameter	Meaning/ Relevance
Raw score	Alignment Score. It is influenced by the homology between the genome and the query sequence and also by the size of the sequences.
Query range	Query sequence length that was engaged in the alignment.
Target range	Genome region in which the query sequence was aligned with success (i.e., in which homology was found between the genome and the query).
Contig	Number or reference of the contig in which alignment was performed.
"***" or "#" presence	*** represents the presence of a STOP codon. # represents the insertion of a nucleotide (when present means sequence pseudogenization).

5 lines, because, it is in these lines that the characters corresponding to the nucleotide sequence are found. After the entire file has been processed, that is, after all lines with the characters corresponding to the nucleotides have been collected, it is necessary to remove everything that is not a nucleotide character. In this step, the program saves only uppercase characters, and thus numbers, spaces and other symbols are excluded. Finally, it creates a new file with all nucleotide characters, sorted in the same way as in the original file, with the original file name and ending with ".fas", allowing it to be used in several programs, namely in this work, in MEGA software version 5.2

Finally, having the sequences files ready and analysed, it was recorded the number of fragments with which each sequence was extracted. The number of nucleotides of each sequence as well the number of artefacts present (presence of "***" or "#" in the middle of the sequence) were also recorded using MEGA software.

3.6 STATISTICAL ANALYSIS

The statistical analysis was performed to determine if there are significant differences between the different sequencing methods, regarding the quality of gene extraction. For this, in each sensory system, sequence data obtained from the "oldest" genomes (sequenced through Solexa, Illumina or Sanger) were compared with sequence data obtained from the "most recent" genomes (PacBio) in three variables: Number of fragments, gene integrity

percentage, and number of artefacts (Figure 9). Each gene sequence was used as if it was a replica of a sample, in this case the sample corresponds to the group of genes of a given sensory system, belonging to several different bird genomes, sequenced by a given sequencing technology.

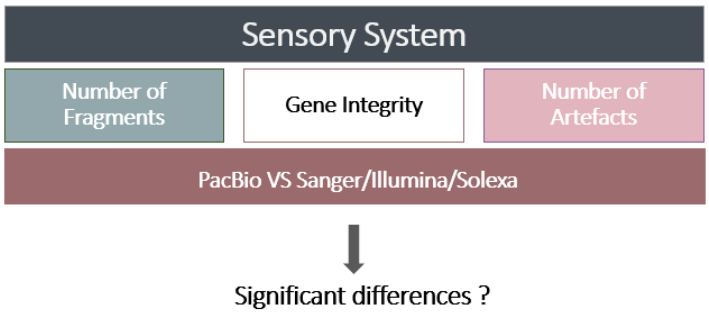


Figure 9.: Schematic representation of the statistical analysis. In each of the six sensory systems, the quality differences between the sequencing technologies, always comparing a more recent technology (PacBio) against an older technology (Sanger, Solexa or Illumina), were evaluated in 3 components: "Number of fragments", "Gene Integrity", "Number of Artefacts". For each component/variable we searched for significant differences, using nonparametric tests.

In this analysis, the number of fragments corresponds to the number of segments in which the gene was fragmented throughout the genome (i.e., number of fragments with which each sequence was extracted), whereas only 1 fragment means that the gene was found all at once, indicating a possible good result.

The integrity percentage of each gene corresponds to the ratio between nucleotide number of the obtained sequence and the nucleotide number of the query sequence.

Finally, the number of artefacts corresponds to the number of "#" (insertions) or "***" present throughout the sequences. If there are no such occurrences in the sequences obtained, it is a promising hint, meaning that the sequence has been extracted from a good alignment.

Thus, we resorted to the R software [124] in order to perform several hypothesis tests. Before the tests were performed, all data was filtered, and normality tests were performed (Shapiro-Wilks Test). In all cases, the normality tests confirmed that the samples did not have a normal distribution and nonparametric statistical tests (Mann-Whitney U Test) were performed to compare two independent groups (a 95% confidence interval was defined for these tests).

To carry on with the statistical tests, for each variable, the null hypothesis and alternative hypothesis were formulated:

Ho: There are no significant differences in the number of fragments, integrity of genes or number of artefacts, between PacBio and Solexa/Illumina/Sanger.

H1: There are significant differences in the number of fragments, integrity of genes or number of artefacts between PacBio and Solexa/Illumina/Sanger.

After performing the statistical tests, in order to obtain a better visualization of the data, again using the R software [124], boxplot graphics were developed for each sensory system. The graphics were developed for each variable under analysis, which allowed a more intuitive observation of the differences among the sequencing technologies.

4

RESULTS AND DISCUSSION

In this work, different sequencing technologies were compared in terms of sequencing quality. A 3^{rd} generation technology, PacBio, known by its advantage in genome sequencing, due to the assembly efficiency produced by its long-reads, was compared with 1^{st} and 2^{nd} technologies, such as Sanger sequencing, Illumina and Solexa, which use a short-read approach. For such, we extracted a set of several genes, belonging to six sensory systems, from five different bird species genomes. For each species a genome sequenced by a 1^{st} or 2^{nd} generation technique was used, as well as a genome sequenced via PacBio, and the quality of the genes sequences was assessed. In the next sections the main results and differences in the sequencing quality among technologies are presented and discussed.

4.1 EXONERATE OUTPUT ANALYSIS: IMPACT ON GENE EXTRACTION

Through the Exonerate software it was possible to generate the alignments, in a protein to genome model, between the collected genomes and all the query sequences (protein sequences) of the several sensory systems initially defined. Thus, in order to extract those sequences from the genomes, it was necessary to analyse the match with the best alignment score. In this analysis it was necessary to consider several parameters, all of them with great importance and influence on the obtained results. As mentioned in the chapter 3, of Materials and Methods, in section 3.5, these parameters include "Raw score", "Query range", "Target range", "Contig" and "presence of "***" and "#"".

In this way, one of the first parameters to consider is the "Raw score". This parameter corresponds to the alignment score and its value is directly influenced by the homology between the genome and the query sequence, and by the length of the query sequence. In this way, it is expected that, the larger the sequence size, the greater the raw score value. Normally, good (or even great) alignments, in which query sequences of about 300 amino acids are used, display a raw score value of approximately 1000. On the other hand, with larger queries, about 800 amino acids, the raw score presents a value of 3000. For instance, it is expected that a query sequence belonging to the *Gallus gallus* species, when aligned against the *Gallus gallus* genome, to present a raw score value very close to the "optimal"

value, as mentioned earlier. It should also be noted that since all the query sequences obtained belong to the species *Gallus gallus*, when we performed the alignments in other bird species genomes, the raw score values slightly decreased, due to the phylogenetic distance effect.

Besides the "Raw score", the "Query range" parameter is also of great importance. Since the size of all query sequences is already known from the beginning, it is possible to understand with this parameter whether if the complete sequence was found in the genome or not. That is, the value of "Query range" that is presented as in the example of Figure 8, in section 3.5 of Materials and Methods (chapter 3), (0 ->343, for protein sequence of *Gallus gallus*, TAAR2), it means that the complete sequence was found, from the beginning, starting at position 0, until its last amino acid, at position 343. Thus, considering this important feature, the main goal was always to select the most complete sequences.

Additionally, the parameters "Target range" and "Contig" also present a great influence in sequences selection and extraction. Each contig is a contiguous segment of the genome that resulted from the assembly of overlapping reads at the time of the genome sequencing. When the contigs are ordered and linked with other contigs it is produced, a representation of the genome [27]. Thus, it is important to know in which section of the genome the alignment occurred, and for that purpose, the Exonerate software presents, in each output, the "name" and number of the contig in which the alignment occurred. The "Target range" also represents the section of the genome, in a specific contig, where the alignment was found. In addition, what makes these two parameters quite important is that the Exonerate software may produces several alignments, with good scores, for the same region if it were given similar query sequences (for example, different paralogs of the same gene family). In these specific cases, since query sequences belong to the same family, is expected that they share high homology, so the region of the genome where the best alignment was found may be quite similar for these sequences. Thus, during this work, whenever a possible good alignment was found, it became necessary to pay attention to the contig where the alignment was performed, as well as the target range, in order to avoid extraction of duplicated regions.

Finally, the presence of "***" and "#" was also of great importance in the choice of the alignment and, consequently, the extraction of the gene's sequences. In what concerns the presence of "***", if it is presented at the end of the sequence it is a good indicator as it represents a STOP codon and, in this way, the sequence is completed until its last residue. On the other hand, the presence of this same symbol in the middle of the sequence indicates that the sequence is interrupted by a STOP codon. Therefore, the resulting protein will be early truncated and not functional. As for the presence of "#", this indicates nucleotide insertions that disrupt the reading frame as originate the pseudogenization of the sequence.

Another occurrence that was occasionally detected was the fragmentation of genes throughout the genomes. In some cases, only part of the gene sequence could be found, meaning that the remaining sequence was present in another region. This made the extraction of some genes a little more time consuming and difficult, and, once again, it was necessary to pay special attention to the Contig number and the "Target range" parameters.

Taking all these parameters into account, the sequences were extracted. The data was collected, and some important characteristics of every single gene, found in each genome, were recorded. One of these characteristics include the number of fragments in which the gene was found throughout the genome. It is expected that the higher the quality of the genome produced by the sequencing technology, less fragmented will be the genes found. It is, therefore, supposed and expected that in a reference quality assembly, capable of a good representation of an entire genome of a given species, it is possible to find a complete gene in only one fragment. Another very important feature recorded here includes the integrity of the gene (ratio between nucleotide number of the obtained gene sequence and the nucleotide number of the query sequence). As well as the number of fragments with which the gene is found, the integrity of the gene (its contiguity and completeness) is of great importance in determining the sequencing quality of the several technologies. In addition, the phylogenetic distance between the genome of the species from which query sequences have been collected and the genome of the species with which these sequences are being aligned, is a factor likely to have some influence. For instance, in the case of this study, sequences that present a high homology with the query sequence are expected to have quite high percentages of integrity since its length in bp should not differ much from the query length. On the other hand, in species that are phylogenetically distant from the *Gallus gallus* species, the gene sequences obtained may have differences in the number of bp.

Lastly, we recorded the number of ambiguities, i.e., the existence of "N", which means the absence of nucleotides to align with that specific region, and the number of artefacts (presence of "***" and "#"), as explained before. All the data is shown in Table 4.

4.2 Evaluation of extracted genes: General analysis

Generally, it is observable (Table 4) that in each of the several bird species, the results of gene extraction were somewhat consistent between both genomes sequencing techniques (PacBio and Sanger/ Solexa/Illumina).

Regarding the presence or absence of genes, in some species, it was not possible to find certain genes. For example, searches inside TAS2R of *Calypte anna* did not reveal the *Tas2r40* and *Tas2r7* genes. The absence of these two genes was found in both *Calypte anna* genomes MUGM01 (sequenced via PacBio technology) and JJRV0 (sequenced via Illumina technol-

Table 4.: Representation of the main characteristics related to all genes of the 6 sensory systems, extracted from the several genomes. All genes represented here are those that showed the best alignment values. On the left side of the table are presented the genomes that were used. For the same species we analysed a genome sequenced via PacBio (3^{rd} generation) and a genome sequenced using an older technology (1^{st} or 2^{nd} generation). In parameters field each abbreviation means as follows: No. Frag.: Number of Fragments; Gene Int.: Gene Integrity; No. Amb.: Number of ambiguities;. Also, the "-" symbol presented in several fields means absence/nonexistence. In addition, in the field of artefacts, "0" corresponds to the absence of artefacts, and "number#" corresponds to the number of artefacts found. In the field of gene integrity, the displayed value corresponds to the ratio between the nucleotide number of the obtained gene sequence and the nucleotide number of the query sequence.

Species	Genome Reference	Technology	Parameters	Chemoreceptors								Magnetoreceptors			
				TASR								CRY			
				TAS1R			TAS2R								
				TAS1R1	TAS1R3	TAS2R4	TAS2R40	TAS2R7	TAAR1	TAAR2	TAAR5	CRY1	CRY2	CRY4	ISCA1
Anas platyrhynchos	RHJV01	PacBio	No. Frag.	1	1	1	1	1	1	1	1	1	1	1	1
			Gene Int.	1.008	1.000	0.990	0.971	0.957	1.000	0.997	1.000	1	0.9754	1	1
			No. Amb.	-	-	-	-	-	-	-	-	-	-	-	-
			Artefacts	0	0	0	0	0	0	0	0	0	2#	0	0
Anas platyrhynchos	ADON01	Solexa	No. Frag.	1	1	1	1	2	1	1	1	1	1	2	2
			Gene Int.	0.507	1.000	0.990	0.971	0.957	1.000	0.997	1.000	0.916	0.762	0.998	0.662
			No. Amb.	-	-	-	-	-	-	-	-	-	-	-	-
			Artefacts	0	0	0	0	0	0	0	0	0	0	0	0
Calypte anna	MUGM01	PacBio	No. Frag.	2	1	1	-	-	1	1	1	1	1	-	1
			Gene Int.	0.946	1.000	0.984	-	-	1.003	0.997	0.994	0.998	0.995	-	1
			No. Amb.	-	-	-	-	-	-	-	-	-	-	-	-
			Artefacts	0	0	0	-	-	0	0	0	0	0	-	0
Calypte anna	JJRV01	Illumina	No. Frag.	2	1	1	-	-	1	1	1	1	1	1	1
			Gene Int.	0.669	1.000	0.984	-	-	1.003	0.997	0.994	0.915	0.775	0.679	0.792
			No. Amb.	-	-	-	-	-	-	-	-	-	-	-	-
			Artefacts	0	0	0	-	-	0	0	0	0	0	0	0
Taeniopygia guttata	MUGN01	PacBio	No. Frag.	1	1	-	1	-	1	-	-	1	1	1	1
			Gene Int.	0.992	0.987	-	0.968	-	1.003	-	-	0.998	0.995	0.998	1
			No. Amb.	-	-	-	-	-	-	-	-	-	-	-	-
			Artefacts	3#	0	-	0	-	0	-	-	0	0	0	0
Taeniopygia guttata	ABQF01	Sanger	No. Frag.	1	1	-	1	-	1	-	-	1	1	1	1
			Gene Int.	0.792	0.965	-	0.968	-	1.003	-	-	0.926	0.925	0.998	1
			No. Amb.	-	-	-	-	-	-	-	-	-	-	-	-
			Artefacts	0	3#	-	0	-	0	-	-	0	0	0	0
Gallus gallus	AADN05	PacBio	No. Frag.	1	1	1	1	1	1	1	1	1	1	1	1
			Gene Int.	0.999	1.000	1.000	0.997	1.000	0.999	1.000	0.999	1	1	1	1
			No. Amb.	-	-	-	-	-	-	-	-	-	-	-	-
			Artefacts	0	0	0	0	0	2#	0	2#	0	0	0	0
Gallus gallus	PDMY01	Illumina	No. Frag.	1	1	1	1	1	1	1	1	1	1	1	1
			Gene Int.	1.000	1.000	1.000	0.997	1.000	1.000	1.000	0.999	1	1	1	1
			No. Amb.	-	-	-	-	-	-	-	-	-	-	-	-
			Artefacts	4#	0	0	0	0	0	0	0	0	0	0	0
Strigops habroptila	RXXE01	PacBio	No. Frag.	1	1	1	-	1	-	1	1	1	1	1	1
			Gene Int.	0.983	0.998	0.987	-	0.953	-	0.994	0.991	0.998	0.995	1	1
			No. Amb.	-	-	1N	-	-	-	-	-	-	-	-	-
			Artefacts	5#	0	0	-	0	-	0	0	0	0	0	0

Table 4.: Continued.

Species	Genome Reference	Technology	Parameters	Thermoreceptors									Photoreceptors						
				TRPV			TRPM									RHO		SWS	
				TRPV1	TRPV2	TRPV4	TRPM2	TRPM3	TRPM5	TRPM8	TRPA1	TRPC5	OPNVA	OPNP	OPN1LW	RH1	RH2	OPN1SW	OPN2SW
Anas platyrhynchos	RHJV01	PacBio	No. Frag.	1	1	-	1	2	1	1	1	1	1	1	-	1	1	-	-
			Gene Int.	1.000	1.000	-	1.000	0.734	1.002	1.000	1.000	0.798	0.997	0.957	-	1.000	1.001	-	-
			No. Amb.	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
			Artefacts	0	0	-	0	0	0	0	0	0	0	0	-	0	0	-	-
Anas platyrhynchos	ADON01	Solexa	No. Frag.	1	1	2	2	2	2	2	2	2	1	1	-	1	1	-	-
			Gene Int.	1.000	1.000	0.593	0.997	0.734	1.002	0.974	0.807	1.000	0.997	0.634	-	1.000	0.993	-	-
			No. Amb.	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
			Artefacts	0	0	4#	0	0	0	0	4#	0	0	0	-	0	0	-	-
Calypte anna	MUGM01	PacBio	No. Frag.	1	1	1	1	1	1	1	1	1	1	1	-	1	1	-	-
			Gene Int.	0.999	0.999	0.970	1.000	0.999	1.004	1.000	1.000	1.000	0.997	0.997	-	1.000	1.001	-	-
			No. Amb.	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
			Artefacts	0	0	2#	2#	0	0	0	0	0	0	0	-	0	0	-	-
Calypte anna	JJRV01	Illumina	No. Frag.	1	1	1	2	2	2	1	1	3	1	1	2	1	1	-	-
			Gene Int.	0.999	0.841	0.638	1.000	0.982	0.841	1.000	1.000	0.993	0.997	0.997	0.521	1.000	1.001	-	-
			No. Amb.	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
			Artefacts	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-	-
Taeniopygia guttata	MUGN01	PacBio	No. Frag.	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
			Gene Int.	1.001	0.997	0.984	0.999	1.000	1.002	0.999	1.000	1.000	0.994	0.994	0.986	1.000	1.001	0.994	0.994
			No. Amb.	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
			Artefacts	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
Taeniopygia guttata	ABQF01	Sanger	No. Frag.	1	1	1	2	4	2	1	1	1	1	1	-	1	1	2	-
			Gene Int.	0.459	0.589	0.390	0.825	0.953	0.902	0.984	1.000	1.000	0.995	0.994	-	1.000	0.995	0.994	-
			No. Amb.	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
			Artefacts	0	0	-	0	0	0	0	0	0	1	0	-	0	0	0	-
Gallus gallus	AADN05	PacBio	No. Frag.	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	-
			Gene Int.	1.000	1.000	0.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.001	0.646	-
			No. Amb.	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
			Artefacts	0	0	4#	0	0	0	0	0	0	0	0	0	0	0	0	-
Gallus gallus	PDMY01	Illumina	No. Frag.	1	1	1	1	1	1	1	1	1	1	1	-	1	1	-	-
			Gene Int.	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	-	1.000	1.001	-	-
			No. Amb.	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
			Artefacts	0	0	0	0	0	0	0	0	0	0	0	-	0	0	-	-
Strigops habroptila	RXXE01	PacBio	No. Frag.	1	1	1	1	1	1	1	1	1	1	1	-	1	1	1	-
			Gene Int.	1.000	0.994	1.000	1.000	1.000	1.003	1.000	1.000	1.000	0.997	0.986	-	1.000	1.001	1.000	-
			No. Amb.	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
			Artefacts	0	-	0	0	0	0	0	0	0	0	0	-	0	0	0	-

Table 4.: Continued.

Species	Genome Reference	Technology	Parameters	Auditory System									Tactile System										
								K+ Channels					ASIC		K+ Channels			PIEZO		ENaC Channels			TMC	
				ZENK	FOS	ARC	CHRNA3	KCNQ4	KCNQ1	KCNMB1	KCNMA1	NKCC1	ASIC1	ASIC2	KCNK2	KCNK4	KCNK10	PIEZO1	PIEZO2	SCNN1A	SCNN1B	SCNN1G	TMC1	TMC2
Anas platyrhynchos	RHJV01	PacBio	No. Frag.	1	2	1	1	2	1	1	1	1	1	2	1	1	1	2	1	1	1	1	1
			Gene Int.	1.000	1.002	1.000	1.000	0.997	0.984	1.000	1.000	1.000	1.000	0.990	0.998	0.957	1.000	0.986	1.001	0.992	0.979	0.998	0.988	0.818
			No. Amb.	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
			Artefacts	0	2#	0	0	0	0	0	0	0	0	0	0	0	0	0	2#	0	0	0	3#	0
Anas platyrhynchos	ADON01	Solexa	No. Frag.	1	4	1	1	10	3	1	13	4	4	2	2	1	2	16	3	1	1	1	-	4
			Gene Int.	0.853	0.886	1.000	0.865	0.803	0.582	1.000	0.982	0.860	0.975	0.636	0.932	0.957	0.970	0.633	0.917	0.992	0.974	0.837	-	0.997
			No. Amb.	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
			Artefacts	0	0	0	0	0	0	0	1 C.STOP	1 C.STOP	0	0	1 C. STOP	0	0	14#	0	0	4#	0	-	0
Calypte anna	MUGM01	PacBio	No. Frag.	1	1	1	1	1	1	1	1	1	1	2	1	1	1	-	1	1	1	1	1	1
			Gene Int.	1.000	0.999	1.000	1.000	0.996	0.911	1.000	0.576	0.949	1.001	1.000	1.000	0.960	1.002	-	1.001	0.989	0.973	1.002	0.927	0.992
			No. Amb.	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
			Artefacts	0	2#	5#	0	0	2#	0	1#	0	1#	0	0	0	0	-	0	0	0	0	4#	1#
Calypte anna	JJRV01	Illumina	No. Frag.	1	2	1	1	1	-	1	3	2	1	-	2	1	2	4	3	1	1	1	-	1
			Gene Int.	0.853	0.899	1.000	1.000	0.882	-	1.000	0.770	0.884	0.720	-	0.665	0.675	0.972	0.941	0.942	0.989	0.973	0.939	-	0.992
			No. Amb.	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
			Artefacts	0	0	0	0	2#	-	0	0	0	0	-	0	0	0	0	3	0	0	0	-	0
Taeniopygia guttata	MUGN01	PacBio	No. Frag.	1	1	1	1	1	1	1	1	1	1	2	1	1	1	3	1	1	1	1	1	1
			Gene Int.	1.000	1.009	1.000	1.000	1.001	0.922	1.000	1.000	0.979	1.001	1.000	1.000	0.960	1.000	0.979	1.000	0.989	0.750	0.992	0.927	0.999
			No. Amb.	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
			Artefacts	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Taeniopygia guttata	ABQF01	Sanger	No. Frag.	1	1	1	2	2	2	2	4	3	2	3	2	3	1	4	4	1	2	2	3	1
			Gene Int.	1.000	1.008	1.000	1.000	0.797	0.738	1.000	0.985	0.979	0.953	1.012	0.738	0.960	1.000	0.359	0.985	0.989	0.959	0.705	0.927	0.999
			No. Amb.	-	-	-	-	-	-	-		-	-	-	-	-	-	-	-	-	-	-	-	-
			Artefacts	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1#	0	0	2#	0	0
Gallus gallus	AADN05	PacBio	No. Frag.	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1
			Gene Int.	1.000	1.000	1.000	1.000	0.974	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.992	1.000	1.000	1.000	1.000
			No. Amb.	-	-	-	-	-	-	-	-	-	-	-	-	-		-	-	-	-	-	-	-
			Artefacts	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2#	0	0	0	0
Gallus gallus	PDMY01	Illumina	No. Frag.	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1
			Gene Int.	1.000	0.998	1.000	1.000	0.970	0.997	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.992	1.000	1.000	1.000	1.000
			No. Amb.	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
			Artefacts	0	2#	0	0	0	2#	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Strigops habroptila	RXXE01	PacBio	No. Frag.	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1
			Gene Int.	0.996	1.000	1.000	1.000	1.000	0.937	1.000	1.000	0.949	1.000	1.000	1.000	0.960	1.000	0.344	1.000	0.989	0.990	1.002	0.992	0.999
			No. Amb.	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
			Artefacts	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4#	0

ogy), and similar situations are also found in other species genomes and other sensory systems (e.g.: In photoreceptors of *Anas platyrhynchos* and in TAARs family of *Taeniopygia guttata*). These facts allow us to propose/assume that genomes, regardless of the technology with which they were sequenced, are relatively in agreement with the biological information they present.

Additionally, there are, however, punctual cases that genomes of the same species differ in the number of genes found. One such case is the *Calypte anna* species, in three sensory systems: magnetoreception, photoreception and tactile system. In terms of magnetoreception and photoreception, it is observed that the PacBio genome, MUGM01, is at a "disadvantage" with Illumina's JJRV01 genome, since the *Cry4* and *Opn1lw* genes could not be found and extracted in the MUGM01 genome. On the other hand, in the tactile system, the situation is reversed, in which the *Asic2* and *Tmc1* genes are missing in the JJRV01 genome. Also, in *Anas platyrhynchos* there are differences in the thermoreception, where there is the absence of the *Trpv4* gene in the PacBio genome, RHJV01. In this species, there are also differences in the tactile system, in which the *Tmc1* gene is absent in the genome obtained by Solexa technology, ADON01. Apart from these, another case is the *Taeniopygia guttata* species, in the photoreception, in which the genes *Opn1lw* and *Opn2sw* were not found in the genome generated by Sanger sequencing, ABQF01.

Despite such occurrences, it appears that these genomes seem to share some common features. In both *Calypte anna* and *Anas platyrhynchos*, in the tactile system it was not found the same gene, *Tmc1*, which may be related to a phylogenetic approach, due to the evolutionary proximity of these species (Figure 10).

Moreover, this absence is only found in genomes sequenced by 2^{nd} generation technologies, Illumina and Solexa, respectively, which is a fact that demonstrates the advantage of PacBio sequencing.

Calypte anna and *Taeniopygia guttata* are also found to share a similarity in the photoreception system, although they are phylogenetically distant. These do not have the *Opn1lw* gene, however, as mentioned earlier, this gene was not found in the more recent *C. anna* genome, whereas in *T. guttata* the gene was not found in the older genome, which seems not be related either to the phylogenetic level or to the sequencing technology.

With regard the sensory systems, a fact also more noticeable than in other sensory systems is concerned with the auditory and tactile systems. It is observed in the genomes sequenced by Sanger, Solexa and Illumina, a greater tendency for the genes of these two systems to be much more fragmented and, although not so evident, the integrity of the obtained genes is lower than in the genomes obtained by PacBio. These facts, especially with regard to the number of fragments, may be essentially related to the genome sequencing technology. Sanger, Solexa, and Illumina technologies base their sequencing process on short reads, which leads to difficulties during genome assembly. An example of one prob-

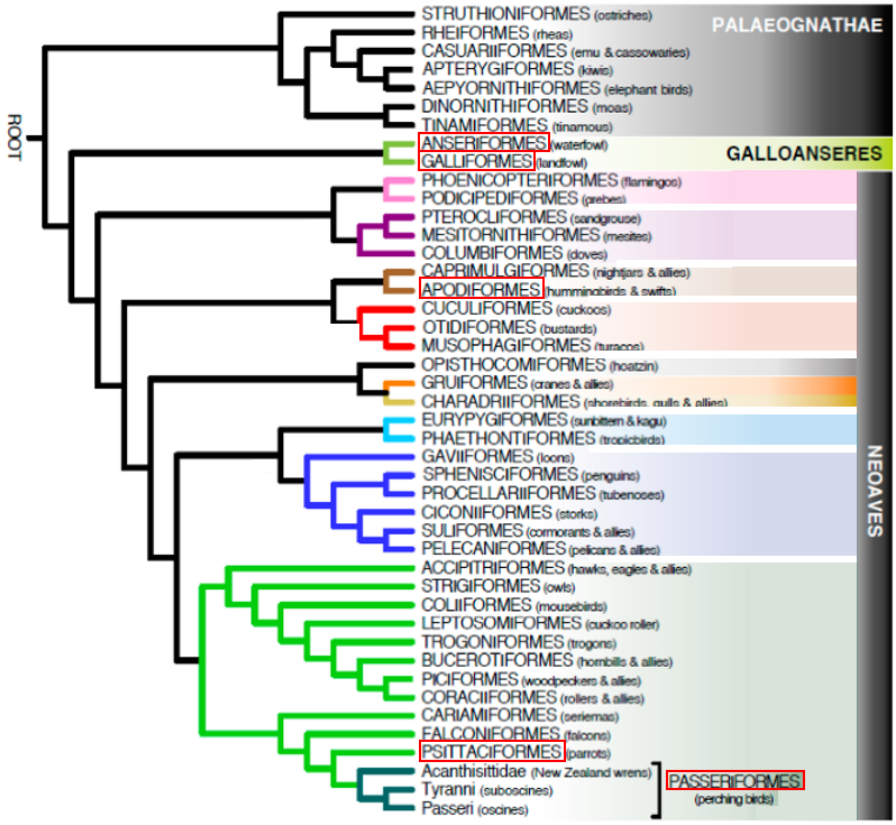


Figure 10.: Phylogenetic representation of birds species by its order. Surrounded in red are the orders to which the species used in this work belong: *Anas platyrhynchos* species belongs to the Anseriformes order; *Calypte anna* species belongs to Apodiformes order; *Taeniopygia guttata* belongs to Passeriformes order; *Gallus gallus* belongs to Galliformes order; *Strigops habroptila* belongs to Psittaciformes order. This image was adapted from [138]

lem concerns the existence of repetitive regions in the genome, which through short reads is very difficult to reconstruct them correctly, which creates ambiguities in the alignment and in the genome assembly, leading to rather fragmented assemblies. Sequencing using short reads is also associated with a higher number of substitution errors across sequences [116, 139, 140].

Another important and somewhat controversial fact observed here is in the *Gallus gallus* species as regards the sensory system of photoreception. It would be expected in this species that, regardless of sequencing technology and sensory system, since the query sequence and the genomes with which the alignments were produced both belong to *Gallus gallus* species, in principle, all genes would be fully extracted and with a relatively good quality. However, what happens in photoreception is the absence of certain genes. In the PDMY01 genome obtained by Illumina, it was not possible to extract three ospins genes: *Opn1lw*, *Opn1sw* and *Opn2sw*. In the other hand, in the AADN05 genome only *Opn2sw* could not be obtained, however, it appears that *Opn1sw* has some "flaws": It is fragmented

and quite incomplete. There are, thus, three reasons why this might have happened. The first reason concerns the whole process of genome sequencing. It is known that generally there may be errors throughout the process, either during the library preparation process or during assembly itself, which may lead to certain sequences not being reported in the genome. Also, possible contamination of the assembly with foreign species contigs, can inducing false positives due to horizontal gene transfer events [141]. In addition, as already mentioned in section 2.3 of Sequencing Technologies, older sequencing technologies generally generate more sequencing flaws, and this is confirmed in this example because, despite the flaws, PacBio produces better results since it was possible to extract two genes unlike through Illumina. This fact confirms, once again, the advantage that long reads assemblies have in genome sequencing over short reads. The second reason comprises the process of assembling these genomes, in which reference genomes may have been used. These reference genomes could already have intrinsic flaws in the region where the sequences are concerned. Consequently, genomes sequenced based on this reference inevitably acquired the same flaw [142]. Finally, a third reason concerns the process of query sequences obtaining. These sequences used as query may be reported because of work done that specifically aimed to amplify a particular region, or even through sequencing of transcriptome regions, which greatly simplifies the process of sequences obtaining [141].Thus they might have been successful in obtaining those specific gene sequences, however, when talking about sequencing a complete genome, these sequences may eventually get lost. Still related to photoreception, it was found that in all species, it was not possible to extract some of these same genes that were not also found in *Gallus gallus*. This is especially true in the *Sanger/Illumina/Solexa (S./I./S.)* genomes, once again highlighting the advantage of long reads based sequencing.

With regard to the presence of artefacts and ambiguities, another important general observation is that there does not appear to be a tendency for the Sanger, Solexa or Illumina genomes to include more artefacts than the PacBio genomes, as would be expected at first view. It is observed throughout Table 4 that the presence of artefacts is in a homogeneous distribution. Additionally, all extracted genes, regardless of genomes and species, present no ambiguities in their sequences, with the exception of *Strigops habroptila*, which, in the *Tas2r4* gene (in the chemoreceptors group), presents an ambiguity after its STOP codon. However, since it is after the STOP codon, it does not seem to have much influence on the rest of the sequence.

Thus, although we were able at first glance to notice several differences and particularities between genomes and sensory systems, in order to understand if there really are significant differences among all data and to discuss the results accurately, statistical analyses were performed.

A final point that should be mentioned is the presence of the *Strigops habroptila* genome. This was not included in the statistical analysis as only one genome of this species could be obtained (through PacBio technology). This genome was incorporated in this work with the objective of "validation" of the sequencing technology. In fact, the results in this genome were found to be quite similar to the PacBio results of the remaining species, which allows us to "conclude" a homogeneity of methodology/results.

4.3 Statistical analysis: significant differences between PacBio and Sanger/Illumina/Solexa

Before proceeding with any specific test, the normality tests were performed, in order to understand if it would be necessary to use parametric or nonparametric tests. Thus, the Shapiro-Wilk normality test was used to assess whether the various sets of samples had a normal distribution of data. The null and alternative hypotheses are, respectively, to the following:

-Data follows a normal distribution.
-Data does not follow a normal distribution.

The normality test was applied to all data sets of all sensory systems. In addition, a 95% confidence interval was defined for this test, which means a significance level (α) of 0.05. The normality test results are shown in Table 5.

Thus, according to the results obtained in the normality tests, it was found that in all cases the p-value was lower than 0.05, or even in some cases, due to the great similarity between the data, it was not possible to perform the Shapiro-Wilk test and, therefore, rejected the null hypothesis. Thus, a non-normal distribution of data was assumed. However, in an attempt to approximate the distribution of data to a normal distribution, we resorted to the linearization of the data, however, this had no influence on its distribution. This was probably due to the size of the samples (some samples with less than 30 replicas, other just slightly larger than 30), since the sample size has a big influence on its own distribution. Generally, small samples are associated with non-normal distributions, because it does not allow the frequency distribution to result in a normal curve [143].

Therefore, taking on a non-normal distribution of data, nonparametric hypothesis tests were performed. Thus, in each sensory system, PacBio technology was compared with the other technologies, S./I./S. at three different levels (or variables) by the nonparametric Mann-Whitney U test. With this test two independent samples are compared, and it is verified whether there are significant differences between the distributions of the two samples. Note that the samples are considered independent since the genes (here considered as

Table 5.: Representation of results and p-values obtained in the normality tests, Shapiro-Wilk Test. These tests were performed on all sensory systems at 3 different levels (Number of fragments, Integrity of genes and Number of artefacts), in both PacBio samples and S./I./S. samples. Fields with the symbol "——" indicate that the normality test could not be performed due to the similarity among the data, and therefore, a non-normal distribution was accepted. "N.N.D." is an abbreviation for Non-Normal Distribution.

	Chemoreceptors	**Magnetoreceptors**	**Photoreceptors**	**Thermoreceptors**	**Auditory System**	**Tactile System**
Number of fragments	N.N.D.	N.N.D.	N.N.D.	N.N.D.	N.N.D.	N.N.D.
P-value:	PacBio: 1.214e-10 S./I./S.: 7.518e-10	PacBio: —— S./I./S.: 7.525e-07	PacBio: —— S./I./S.: 7.572e-08	PacBio: 6.472e-13 S./I./S.: 1.747e-07	PacBio: 9.691e-07 S./I./S.: 1.676e-09	PacBio: 3.621e-12 S./I./S.: 1.142e-11
Integrity of genes	N.N.D.	N.N.D.	N.N.D.	N.N.D.	N.N.D.	N.N.D.
P-value:	PacBio: 7.921e-06 S./I./S.: 9.802e-09	PacBio: 2.396e-06 S./I./S.: 0.007829	PacBio: 2.334e-06 S./I./S.: 1.274e-07	PacBio: 1.016e-11 S./I./S.: 9.333e-08	PacBio: 1.229e-11 S./I./S.: 1.619e-06	PacBio: 1.314e-11 S./I./S.: 4.45e-09
Number of artefacts	N.N.D.	N.N.D.	N.N.D.	N.N.D.	N.N.D.	N.N.D.
P-value:	PacBio: 8.719e-10 S./I./S.: 8.786e-10	PacBio: 9.834e-08 S./I./S.: ——	PacBio: 1.057e-08 S./I./S.: 1.057e-08	PacBio: 1.345e-11 S./I./S.: 3.394e-12	PacBio: 7.688e-11 S./I./S.: 1.376e-10	PacBio: 2.473e-12 S./I./S.: 1.437e-13

sample replicas) from PacBio and S./I./S. genomes were obviously extracted from genomes generated from different sequencing technologies. As performed in the normality tests, a 95% confidence interval (α of 0.05) was also defined here and the null and alternative hypotheses were formulated, respectively, at each level:

- There are no significant differences in the number of fragments / integrity of genes / number of artefacts between PacBio and S./I./S..
- There are significant differences in the number of fragments / integrity of genes / number of artefacts between PacBio and S./I./S..

The results obtained are presented in Table 6. In addition to obtaining these data, boxplot graphs were also generated for each system, in order to ease the visualization of the results (Figure 11).

From the p-values obtained (Table 6), it is observable that not all sensory systems present significant differences in the quality of gene extraction between sequencing technologies (p-values lower than 0.05). Moreover, none of the sensory systems analysed here show significant differences regarding the number of artefacts present in the sequences, which is in agreement with the general analysis previously performed.

It is, however, important to mention and remember that, although there are no significant differences in some sensory systems, by analysing Table 4, as described above, it is noteworthy that, despite not having statistical significance, in most of the genes PacBio technology has produced better quality sequences (more complete and less fragmented sequences), as expected. It is also observable from Figure 11-d that only the photoreception system has remained fairly consistent and homogeneous across different types of sequencing technologies. On the other hand, although there is no significant differences in the chemoreception sensory system, there is a slight difference between PacBio and S./I./S. data in gene integrity, in which S./I./S. data present greater variability among itself, with values from 0.507 to 1.000.

Regarding the existence of significant differences, these are present in four sensory systems, in magnetoreception, at the gene integrity level, and in thermoreception, in the auditory system and in the tactile system, all in the same type of variables: number of fragments and gene integrity. It is also observed that in these same cases, observing the boxplots of Figure 11-c, Figure 11-e, Figure 11-f, all S./I./S. results vary considerably more than the PacBio values.

Concerning, specifically the number of fragments with which the genes were obtained, in all systems that showed significant differences, it is observed that the PacBio results essentially are around the value 1. On the other hand, in the results of S./I./S., the values vary and deviate greatly from 1, reaching, in the case of the tactile system, to 15 fragments

Table 6.: Representation of results and p-values obtained through the hypothesis test, Mann-Whitney U Test. Hypothesis tests were performed on all sensory systems, in the same way as were the normality tests. Also listed here, the abbreviations "N.S.D." and "S.D." mean absence or presence of significant differences, respectively, between sequencing technologies, respectively.

	Chemoreceptors	**Magnetoreceptors**	**Photoreceptors**	**Thermoreceptors**	**Auditory System**	**Tactile System**
Number of fragments	N.S.D.	N.S.D.	N.S.D.	S.D.	S.D.	S.D.
Statistical test used	Mann-Whitney U Test	Mann-Whitney U Test	Mann-Whitney U Test	Mann-Whitney U Test	Mann-Whitney U Test	Mann-Whitney U Test
P-value:	0.5714	0.1641	0.163	0.0001732	0.000504	0.0002914
Integrity of genes	N.S.D.	S.D.	N.S.D.	S.D.	S.D.	S.D.
Statistical test used	Mann-Whitney U Test	Mann-Whitney U Test	Mann-Whitney U Test	Mann-Whitney U Test	Mann-Whitney U Test	Mann-Whitney U Test
P-value:	1	0.008145	0.6971	0.009715	0.007545	0.001233
Number of artefacts	N.S.D.	N.S.D.	N.S.D.	N.S.D.	N.S.D.	N.S.D.
Statistical test used	Mann-Whitney U Test	Mann-Whitney U Test	Mann-Whitney U Test	Mann-Whitney U Test	Mann-Whitney U Test	Mann-Whitney U Test
P-value:	0.9671	0.3506	1	0.6931	0.9461	0.9728

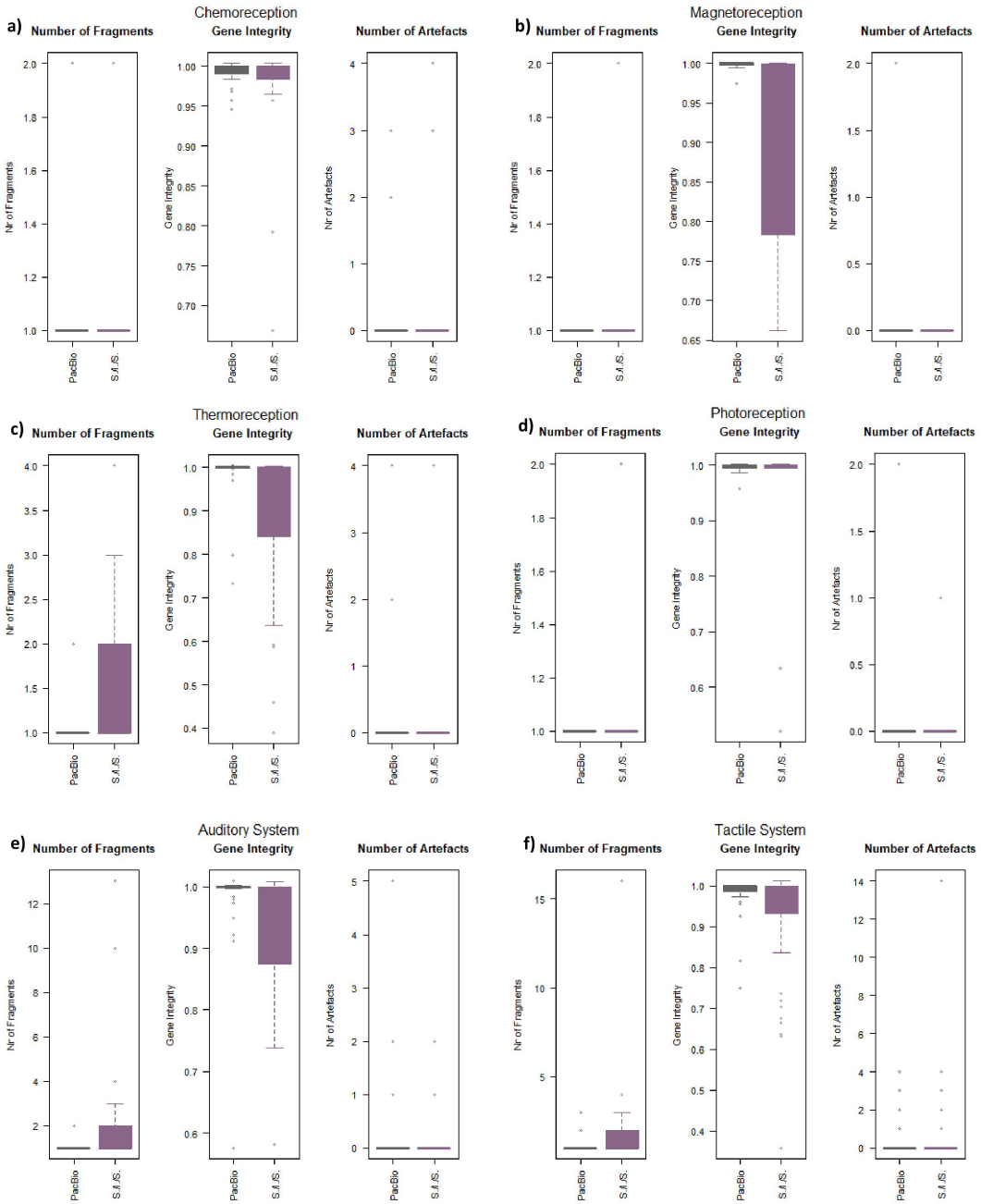


Figure 11.: Boxplots representing data variation of the six sensory systems studied: Chemoreception, Magnetoreception, Thermoreception, Photoreception, Auditory System and Tactile System; in the three different variables: Number of Fragments, Gene Integrity, Number of Artefacts. Each boxplot presents data from PacBio technology and S./I./S. technologies. The set of boxplots from each sensory system has a corresponding letter, and therefore, the sensory systems are organized from "a" to "f".

in a single gene. This fact is quite important to confirm the significant improvement that PacBio sequencing technique brings to the extraction of complete genes. Due to the large size of the reads, from the average value of 10 Kbp to 60 Kbp, the assembly process becomes much more effective. The approach commonly used for assembling long reads is known as overlap-layout-consensus assembly, which consists of aligning all reads, and recording this information in graphs (overlap-graphs). In these graphs the nodes correspond to the reads sequences, while the edges correspond to the connection between the sequences, due to alignment (overlap). This information is organized, with which the position of all sequences is determined according to the alignments. Finally, the resulting alignments generate the consensus sequences that will give rise to contigs [116, 144].

On the other hand, in 1^{st} and 2^{nd} generation technologies, the length of reads is considerably smaller, which in turn makes assembly difficult. In addition, these short-read based technologies typically use a different approach, the k-mers assembly. K-mers are produced by fragmenting the sequence reads, which makes them even smaller in length. The assembly process of these sequences is also based on graphs that record overlap of reads, in this case, de Brujin graphs, and contigs are formed by reading the graph. The problem with this type of approach is that genomes can include many repetitive sequences, and if these sequences are larger than the produced k-mers, it will become very difficult to reconstruct all fragments, which leads to generation of highly fragmented assemblies. In addition, the situation becomes even more complex, when repetitive sequences align in more than one region of the genome (known as multi-reads). This may negatively influence the execution of analyses that depend on regions next to the multi-reads, since these regions were probably misassembled [140].

The case of Sanger sequencing differs slightly from Solexa and Illumina, as the reads produced are slightly larger than those produced by Illumina and Solexa and are considered of intermediate size. In addition, Sanger sequencing uses the overlap-layout-consensus assembly approach and, therefore, is not so prone to errors.

Additionally, another short-read consequence may be related to low rates of assembly contiguity, and thus the number of disconnected contigs. In this way, several genes, which may even be longer than the length of one contig, may be fragmented into different unordered contigs, and therefore the orientation and order of these genes will remain unknown and may even, for example, induce wrong counting of the number of genes in that specific genome. This becomes even more difficult in eukaryotic genomes, as is the case, where genes are more spaced apart and include introns in their structure [141, 28].

Thus, long read sequences, as already confirmed in several studies, allow complete genes to be obtained, rather than just fragmented sequences with little biological information [116, 145].

As for the other component that showed significant differences, gene integrity, this shows to be quite similar to the number of fragments. It also has values very close to 1.0 in the PacBio genomes, that is, the sequences of the obtained genes are quite complete, and their length does not differ much from the query length. In the case of S./I./S. values, quite variable values are observed, and in the case of magnetoreception this variation is more noticeable. It is important to note that although the results of fragment number and gene integrity are inevitably related, in the case of this sensory system, there is a difference. Although the number of fragments is similar between PacBio and S./I./S., with relatively "good" values, the gene integrity values do not show the same in the S./I./S. technologies. This indicates, once again, the improvement that PacBio brings, not only for annotation of complete genes, but also for reducing the number of errors during sequencing. Since PacBio uses SMRT sequencing, it does not resort to the process of DNA amplification when preparing the library, which, in the case of 2^{nd} generation technologies, namely Illumina, proves to induce substitution errors [139]. In fact, short reads assemblers can sometimes introduce errors in gene sequences, namely erroneous repetitive sequences, into regions that are not really repetitive. This fact, together with other types of errors already mentioned, may lead to erroneous predictions about the sequences of the genes and proteins encoded by them [146]. Additionally, still regarding errors, PacBio provides uniform coverage of the entire genome, which does not induce GC bias unlike other technologies [117].

Besides all that was described above, there are several characteristics that eukaryotic genomes have that make the sequencing process difficult, regardless of the technology used. In addition to the well-known ones, such as repetitive sequences and *Transposable Elements (TEs)*, other variants such as deletions, insertions, GC-rich regions (which are a challenge in avian genomes due to the high amount of existing GC-rich genes), duplicated genes (paralogs), polymorphic genes and rearrangements in the genome (inversions and translocations), including structural variations in large chromosome regions, lead to increased difficulty when assembling the genome [147, 140, 148]. However, through long reads assembling, this type of more complex regions of the genome and gene sequences, can be more easily resolved. In addition, the fact that the genome comes from a long reads' assembly may contain less gaps. It also allows contigs to be more easily sorted and, in turn, genes to be more easily annotated, with complete sequences sorted in the correct form, which was corroborated with this study.

Moreover, this work has proven that despite the characteristics that define birds genomes, PacBio still presents advantages over 1^{st} and 2^{nd} generation technologies. Bird genomes are known to be smaller than the other amniota genomes and less complex compared to other vertebrates. They also present a more conserved genetic structure (including chromosome level, with a high degree of synteny across several divergent species), fewer repetitive elements and less abundance of TEs than other vertebrates [149, 142]. Furthermore, as far as

protein-coding genes are concerned, they are smaller than most mammalian genes, due to the shorter length of introns and intergenic regions they present [149].

For these reasons, we could be led to deduce that the type of technology would not have much influence on the quality of sequencing, but the fact is that despite being more accessible genomes, they still present challenges in the world of DNA sequencing. Indeed, bird microsomes and macrosomes continue to present challenges as long reads are unable to cover these regions, longer than 3Mb in length and between 30 to 250Mb in length, respectively [149]. Moreover, regions close to centromeres and telemores (heterochromatic regions) remain difficult to sequence and assemble correctly due to the high percentage of repetitive elements (almost exclusively by tandem repeats) that is found in those regions [150].

4.4 Related evidences from other studies

There are published studies that confirm what was developed and stated in this paper. One of these studies by Korlach et al.(2017) [146], investigated the quality of the PacBio genomes of *Taenipygia guttata*, zebra finch, and *Calypte anna*, Anna's hummingbird.

In the study, in short, they searched for a set of genes of interest in order to assess their completeness and contiguity in relation to genomes of these species sequenced by older technologies. In the set of genes was included the *Egr1* gene, also used in this work as representative of the tactile system (referred in this work as *Zenk* gene).

With regard specifically to the *Egr1* gene, they found that in the two species, in the "oldest" genome, there was no promoter region of that gene, and in addition, the promoter region was located in a GC rich region. They also found in these genomes that the gene included some gaps in its structure. After analysis of the PacBio genomes, they concluded that long reads allowed assembly gaps to be eliminated, and found that the gap in Anna's hummingbird's oldest genome encompassed the first full exon, two thirds of the first intron, and the promoter region.

Moreover, through further analysis, they concluded, as in this study, that the long-read assembly has proven to provide several improvements in genome and gene completeness and contiguity, solving the problems associated with the genes of interest.

They also concluded that PacBio long reads are capable of filling gaps in particularly problematic regions (high GC content), are able to overcome and eliminate erroneous tandem duplications near gaps, eliminate poor quality sequences and thus prevent incorrect gene annotations and predictions to be made.

Another study by Beauclair et al.(2019) [151] investigated the difficulty of sequencing genes with high GC content (typical characteristic of the avian genome) and for such, they used GC rich genes with quite long introns. Additionally, the non-coding regions of these

genes also had some complexity, such as repetitive sequences, both tandem and inverted, as well as motifs capable of altering the conformation of their structure (non-B DNA).

In the study, the main objective was to compare Illumina and PacBio technology. What they found from results was that PacBio was far more effective in sequencing and that Illumina could not even represent the non-coding regions. They concluded that PacBio was even able to sequence the regions with high GC content, above 60%, however, with some difficulty, due, according to what they point out, to the G_4 structures found in these GC-rich regions. They also pointed out that this kind of structures will be one of the main reasons for the absence of certain structures when sequencing.

5

CONCLUSIONS AND FURTHER WORK

Birds are a model organism of great importance in a wide range of investigations, with a strong impact on genetic studies, due to the characteristics that their genomes present. The information obtained from genetic studies will allow to answer about the evolutionary history of these organisms, establishing genotype-phenotype relationships more easily, understanding regulatory mechanisms at the molecular level, how natural selection works at the level of several regions of the genome, comprehend the changes that can occur in the genome throughout evolution and adaptation, and promoting the study of gene families [142, 149].

As already mentioned, the development of sequencing technologies has a strong role in enabling such studies to happen. The development of new techniques and algorithms that are more efficient, less expensive, and allow the sequencing and assembly of genomes, is an area in constant update, which will allow to solve in the future many of the restrictions that we face today [19].

In this work, we focused on comparing the quality of sequencing between 1^{st} and 2^{nd} generation technologies, which include an approach based on the production of short reads, with a 3^{rd} generation technology, which is based on the production of long reads, PacBio. This evaluation was carried out using a set of genes belonging to several sensory systems. Since these types of genes are subject to constant environmental pressures, adaptation and evolution, they become a very interesting group to test the quality of the sequencing produced by these varied sequencing technologies. What we concluded with the development of this work was that PacBio presented better results in most of the sensory systems analysed, essentially at the level of the fragmentation of genes along the genomes, as well as at the level of their integrity. Even though a higher error rate is related with this technology, reported up to 20% [116], what was observed here was that, in fact, PacBio presented better quality of sequencing, in what concerns to the parameters evaluated in this work, when compared to Sanger, Illumina and Solexa.

Furthermore, we realized that during the development of this work, a new genome with strong interest in this study was made available in the Sequence Set Browser of the NCBI. The genome belongs to the zebra finch *Taeniopygia guttata* (reference: VOHI01), also se-

quenced using PacBio technology and, as expected, the sequencing parameters are even better, since the bases are practically the same (1,115,340,858 bp), but have less contigs/-less fragments (204 contigs). However, unlike both *T. guttata* genomes used here, this new genome corresponds to a female, which is the heterogametic sex. This fact may be of great interest for work in the future, since we have the possibility to compare and understand whether PacBio can detect differences in the sequences of several genes in organisms of the same species, but of different sex. Also, it is important to mention that birds are different from mammals in what concerns the sex chromosomes. While mammals possess the XY system, in birds the heterogametic sex (ZW) belongs to females while the homogametic (ZZ) sex belongs to males [152].

In addition to the inclusion of this new genome, in the future we may also include genomes of more bird species, include new sets of genes from other sensory systems in the search and observe whether the "behaviour" of PacBio remains.

Additionally, in order to deepen this research and better understand the reasons for certain results, in the future we will be able to develop synteny analyses. Through these analyses we will be able to look for genes that may be flanking genes of interest, to understand if a particular gene is conserved over a group of species, to understand the orientation of the genes and to conclude several modification events that may have occurred during evolution (for example insertions of other genes and inversions).

BIBLIOGRAPHY

[1] Tibisay Escalona, Cameron J Weadick, and Agostinho Antunes. Adaptive patterns of mitogenome evolution are associated with the loss of shell scutes in turtles. *Molecular biology and evolution*, 34(10):2522–2536, 2017.

[2] Liliana Silva and Agostinho Antunes. Vomeronasal receptors in vertebrates and the evolution of pheromone detection. *Annual review of animal biosciences*, 5:353–370, 2017.

[3] Imran Khan, Zhikai Yang, Emanuel Maldonado, Cai Li, Guojie Zhang, M Thomas P Gilbert, Erich D Jarvis, Stephen J O'brien, Warren E Johnson, and Agostinho Antunes. Olfactory receptor subgenomes linked with broad ecological adaptations in sauropsida. *Molecular biology and evolution*, 32(11):2832–2843, 2015.

[4] Kai Wang and Huabin Zhao. Birds generally carry a small repertoire of bitter taste receptor genes. *Genome biology and evolution*, 7(9):2705–2715, 2015.

[5] Yasuyuki Hashiguchi and Mutsumi Nishida. Evolution of trace amine–associated receptor (taar) gene family in vertebrates: lineage-specific expansions and degradations of a second class of vertebrate chemosensory receptors expressed in the olfactory epithelium. *Molecular biology and evolution*, 24(9):2099–2107, 2007.

[6] Jakob C Mueller, Silke Steiger, Andrew E Fidler, and Bart Kempenaers. Biogenic trace amine–associated receptors (taars) are encoded in avian genomes: Evidence and possible implications. *Journal of heredity*, 99(2):174–176, 2008.

[7] Benjamin L Clites and Jonathan T Pierce. Identifying cellular and molecular mechanisms for magnetosensation. *Annual review of neuroscience*, 40:231–250, 2017.

[8] Roswitha Wiltschko and Wolfgang Wiltschko. Magnetoreception. In *Sensing in nature*, pages 126–141. Springer, 2012.

[9] Gregory L Owens and Diana J Rennison. Evolutionary ecology of opsin gene sequence, expression and repertoire. *Molecular ecology*, 26(5):1207–1210, 2017.

[10] Shigeru Saito and Makoto Tominaga. Functional diversity and evolutionary dynamics of thermotrp channels. *Cell Calcium*, 57(3):214–221, 2015.

[11] Matthew IM Louder, Shelby Lawson, Kathleen S Lynch, Christopher N Balakrishnan, and Mark E Hauber. Neural mechanisms of auditory species recognition in birds. *Biological Reviews*, 94(5):1619–1635, 2019.

[12] Yann Roudaut, Aurélie Lonigro, Bertrand Coste, Jizhe Hao, Patrick Delmas, and Marcel Crest. Touch sense: functional organization and molecular determinants of mechanosensitive receptors. *Channels*, 6(4):234–245, 2012.

[13] Aurélie Kapusta, Alexander Suh, and Cédric Feschotte. Dynamics of genome size evolution in birds and mammals. *Proceedings of the National Academy of Sciences*, 114(8):E1460–E1469, 2017.

[14] Santiago Claramunt and Joel Cracraft. A new time tree reveals earth history's imprint on the evolution of modern birds. *Science advances*, 1(11):e1501005, 2015.

[15] Douglas Richard Wylie, Cristian Gutiérrez-Ibáñez, and Andrew Iwaniuk. Integrating brain, behavior, and phylogeny to understand the evolution of sensory systems in birds. *Frontiers in neuroscience*, 9:281, 2015.

[16] Sara Hayden, Michaël Bekaert, Tess A Crider, Stefano Mariani, William J Murphy, and Emma C Teeling. Ecological adaptation determines functional mammalian olfactory subgenomes. *Genome research*, 20(1):1–9, 2010.

[17] Andreas Keller and Leslie B Vosshall. Better smelling through genetics: mammalian odor perception. *Current opinion in neurobiology*, 18(4):364–369, 2008.

[18] C Alex Buerkle and Zachariah Gompert. Population genomics based on low coverage sequencing: how low should we go? *Molecular ecology*, 22(11):3028–3035, 2013.

[19] Angela P Fuentes-Pardo and Daniel E Ruzzante. Whole-genome sequencing approaches for conservation biology: Advantages, limitations and practical recommendations. *Molecular ecology*, 26(20):5369–5406, 2017.

[20] Konstantin Berlin, Sergey Koren, Chen-Shan Chin, James P Drake, Jane M Landolin, and Adam M Phillippy. Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nature biotechnology*, 33(6):623, 2015.

[21] Anthony Rhoads and Kin Fai Au. Pacbio sequencing and its applications. *Genomics, proteomics & bioinformatics*, 13(5):278–289, 2015.

[22] Medhat Mahmoud, Marek Zywicki, Tomasz Twardowski, and Wojciech M Karlowski. Efficiency of pacbio long read correction by 2nd generation illumina sequencing. *Genomics*, 111(1):43–49, 2019.

[23] David N Reznick and Cameron K Ghalambor. The population ecology of contemporary adaptations: what empirical studies reveal about the conditions that promote adaptive evolution. In *Microevolution Rate, Pattern, Process*, pages 183–198. Springer, 2001.

[24] David L Stern and Virginie Orgogozo. The loci of evolution: how predictable is genetic evolution? *Evolution*, 62(9):2155–2177, 2008.

[25] Nicolas Galtier. Adaptive protein evolution in animals and the effective population size hypothesis. *PLoS genetics*, 12(1):e1005774, 2016.

[26] Emanuel Maldonado, Kartik Sunagar, Daniela Almeida, Vitor Vasconcelos, and Agostinho Antunes. Impact_s: integrated multiprogram platform to analyze and combine tests of selection. *PloS one*, 9(10):e96243, 2014.

[27] Anthony JF Griffiths, Susan R Wessler, Richard C Lewontin, William M Gelbart, David T Suzuki, Jeffrey H Miller, et al. *An introduction to genetic analysis*, chapter 12, pages 395–428. Macmillan, 2005.

[28] Toni Gabaldón and Tyler S Alioto. Whole-genome sequencing recommendations. In *Field Guidelines for Genetic Experimental Designs in High-Throughput Sequencing*, pages 13–41. Springer, 2016.

[29] Peter Mombaerts. Genes and ligands for odorant, vomeronasal and taste receptors. *Nature Reviews Neuroscience*, 5(4):263, 2004.

[30] Filipe G Vieira and Julio Rozas. Comparative genomics of the odorant-binding and chemosensory protein gene families across the arthropoda: origin and evolutionary history of the chemosensory system. *Genome Biology and Evolution*, 3:476–490, 2011.

[31] Ping Feng and Shichu Liang. Molecular evolution of umami/sweet taste receptor genes in reptiles. *PeerJ*, 6:e5570, 2018.

[32] Bernd Lindemann. Taste reception. *Physiological reviews*, 76(3):719–766, 1996.

[33] Shira L Cheled-Shoval, Maik Behrens, Wolfgang Meyerhof, Masha Y Niv, and Zehava Uni. Perinatal administration of a bitter tastant influences gene expression in chicken palate and duodenum. *Journal of agricultural and food chemistry*, 62(52):12512–12520, 2014.

[34] E Roura, Maude W Baldwin, and KC Klasing. The avian taste system: Potential implications in poultry nutrition. *Animal Feed Science and Technology*, 180(1-4):1–9, 2013.

[35] Y Hashiguchi. The origin and evolution of the trace amine-associated receptor family in vertebrates. In *Trace Amines and Neurological Disorders*, pages 45–62. Elsevier, 2016.

[36] Masatoshi Nei, Yoshihito Niimura, and Masafumi Nozawa. The evolution of animal chemosensory receptor gene repertoires: roles of chance and necessity. *Nature Reviews Genetics*, 9(12):951–963, 2008.

[37] Stephen D Liberles and Linda B Buck. A second class of chemosensory receptors in the olfactory epithelium. *Nature*, 442(7103):645–650, 2006.

[38] David EI Gloriam, Thóra K Bjarnadóttir, Yi-Lin Yan, John H Postlethwait, Helgi B Schiöth, and Robert Fredriksson. The repertoire of trace amine g-protein-coupled receptors: large expansion in zebrafish. *Molecular phylogenetics and evolution*, 35(2):470–482, 2005.

[39] Wolfgang Wiltschko and Roswitha Wiltschko. Magnetic orientation and magnetoreception in birds and other animals. *Journal of Comparative Physiology A*, 191(8):675–693, 2005.

[40] DA Kishkinev and NS Chernetsov. Magnetoreception systems in birds: a review of current research. *Biology Bulletin Reviews*, 5(1):46–62, 2015.

[41] Jing-Jing Xu, Ying-Chao Zhang, Jian-Qi Wu, Wei-Hong Wang, Yue Li, Gui-Jun Wan, Fa-Jun Chen, Gregory A Sword, and Wei-Dong Pan. Molecular characterization, spatial-temporal expression and magnetic response patterns of iron-sulfur cluster assembly1 (isca1) in the rice planthopper, nilaparvata lugens. *Insect science*, 26(3):413–423, 2019.

[42] Thorsten Ritz, Salih Adem, and Klaus Schulten. A model for photoreceptor-based magnetoreception in birds. *Biophysical journal*, 78(2):707–718, 2000.

[43] Roswitha Wiltschko, Margaret Ahmad, Christine Nießner, Dennis Gehring, and Wolfgang Wiltschko. Light-dependent magnetoreception in birds: the crucial step occurs in the dark. *Journal of The Royal Society Interface*, 13(118):20151010, 2016.

[44] Roswitha Wiltschko and Wolfgang Wiltschko. Magnetoreception in birds. *Journal of the Royal Society Interface*, 16(158):20190295, 2019.

[45] Atticus Pinzon-Rodriguez, Staffan Bensch, and Rachel Muheim. Expression patterns of cryptochrome genes in avian retina suggest involvement of cry4 in light-dependent magnetoreception. *Journal of The Royal Society Interface*, 15(140):20180058, 2018.

[46] Anja Günther, Angelika Einwich, Emil Sjulstok, Regina Feederle, Petra Bolte, Karl-Wilhelm Koch, Ilia A Solov'yov, and Henrik Mouritsen. Double-cone localization and seasonal expression pattern suggest a role in magnetoreception for european robin cryptochrome 4. *Current Biology*, 28(2):211–223, 2018.

[47] Andrea Möller, Sven Sagasser, Wolfgang Wiltschko, and Bernd Schierwater. Retinal cryptochrome in a migratory passerine bird: a possible transducer for the avian magnetic compass. *Naturwissenschaften*, 91(12):585–588, 2004.

[48] Aziz Sancar. Structure and function of dna photolyase and cryptochrome blue-light photoreceptors. *Chemical reviews*, 103(6):2203–2238, 2003.

[49] Miriam Liedvogel, Kiminori Maeda, Kevin Henbest, Erik Schleicher, Thomas Simon, Christiane R Timmel, PJ Hore, and Henrik Mouritsen. Chemical magnetoreception: bird cryptochrome 1a is excited by blue light and forms long-lived radical-pairs. *PloS one*, 2(10), 2007.

[50] UniProt Consortium. Uniprot: the universal protein knowledgebase. *Nucleic acids research*, 45(D1):D158–D169, 2016.

[51] Christine Nießner, Susanne Denzau, Leo Peichl, Wolfgang Wiltschko, and Roswitha Wiltschko. Magnetoreception in birds: I. immunohistochemical studies concerning the cryptochrome cycle. *Journal of Experimental Biology*, 217(23):4221–4224, 2014.

[52] Christine Nießner, Susanne Denzau, Julia Christina Gross, Leo Peichl, Hans-Joachim Bischof, Gerta Fleissner, Wolfgang Wiltschko, and Roswitha Wiltschko. Avian ultraviolet/violet cones identified as probable magnetoreceptors. *PloS one*, 6(5), 2011.

[53] Henrik Mouritsen, Ulrike Janssen-Bienhold, Miriam Liedvogel, Gesa Feenders, Julia Stalleicken, Petra Dirks, and Reto Weiler. Cryptochromes and neuronal-activity markers colocalize in the retina of migratory birds during magnetic orientation. *Proceedings of the National Academy of Sciences*, 101(39):14294–14299, 2004.

[54] Petra Bolte, Florian Bleibaum, Angelika Einwich, Anja Günther, Miriam Liedvogel, Dominik Heyers, Anne Depping, Lars Wöhlbrand, Ralf Rabus, Ulrike Janssen-Bienhold, et al. Localisation of the putative magnetoreceptive protein cryptochrome 1b in the retinae of migratory birds and homing pigeons. *PLoS One*, 11(3), 2016.

[55] Christine Nießner, Julia Christina Gross, Susanne Denzau, Leo Peichl, Gerta Fleissner, Wolfgang Wiltschko, and Roswitha Wiltschko. Seasonally changing cryptochrome 1b expression in the retinal ganglion cells of a migrating passerine bird. *PloS one*, 11(3), 2016.

[56] Leonida Fusani, Cristiano Bertolucci, Elena Frigato, and Augusto Foà. Cryptochrome expression in the eye of migratory birds depends on their migratory status. *Journal of Experimental Biology*, 217(6):918–923, 2014.

[57] Hiromasa Mitsui, Toshinori Maeda, Chiaki Yamaguchi, Yusuke Tsuji, Ryuji Watari, Yoko Kubo, Keiko Okano, and Toshiyuki Okano. Overexpression in yeast, photocycle, and in vitro structural change of an avian putative magnetoreceptor cryptochrome4. *Biochemistry*, 54(10):1908–1917, 2015.

[58] Ryuji Watari, Chiaki Yamaguchi, Wataru Zemba, Yoko Kubo, Keiko Okano, and Toshiyuki Okano. Light-dependent structural change of chicken retinal cryptochrome4. *Journal of Biological Chemistry*, 287(51):42634–42641, 2012.

[59] Shannon E Greene and Arash Komeili. Biogenesis and subcellular organization of the magnetosome organelles of magnetotactic bacteria. *Current opinion in cell biology*, 24(4):490–495, 2012.

[60] Nathaniel B Edelman, Tanja Fritz, Simon Nimpf, Paul Pichler, Mattias Lauwers, Robert W Hickman, Artemis Papadaki-Anastasopoulou, Lyubov Ushakova, Thomas Heuser, Guenter P Resch, et al. No evidence for intracellular magnetite in putative vertebrate magnetoreceptors identified by magnetic screening. *Proceedings of the National Academy of Sciences*, 112(1):262–267, 2015.

[61] Sarafina M Kimø, Ida Friis, and Ilia A Solov'yov. Atomistic insights into cryptochrome interprotein interactions. *Biophysical journal*, 115(4):616–628, 2018.

[62] Ida Friis, Emil Sjulstok, and Ilia A Solov'yov. Computational reconstruction reveals a candidate magnetic biocompass to be likely irrelevant for magnetoreception. *Scientific reports*, 7(1):1–12, 2017.

[63] Siying Qin, Hang Yin, Celi Yang, Yunfeng Dou, Zhongmin Liu, Peng Zhang, He Yu, Yulong Huang, Jing Feng, Junfeng Hao, et al. A magnetic protein biocompass. *Nature materials*, 15(2):217–226, 2016.

[64] J Antonio Lamas, Lola Rueda-Ruzafa, and Salvador Herrera-Pérez. Ion channels and thermosensitivity: Trp, trek, or both? *International journal of molecular sciences*, 20(10):2371, 2019.

[65] Karen Castillo, Ignacio Diaz-Franulic, Jonathan Canan, Fernando Gonzalez-Nilo, and Ramon Latorre. Thermally activated trp channels: molecular sensors for temperature detection. *Physical biology*, 15(2):021001, 2018.

[66] A Yamamoto, K Takahashi, S Saito, M Tominaga, and T Ohta. Two different avian cold-sensitive sensory neurons: Transient receptor potential melastatin 8 (trpm8)-dependent and-independent activation mechanisms. *Neuropharmacology*, 111:130–141, 2016.

[67] Michael Bandell, Lindsey J Macpherson, and Ardem Patapoutian. From chills to chilis: mechanisms for thermosensation and chemesthesis via thermotrps. *Current opinion in neurobiology*, 17(4):490–497, 2007.

[68] Ardem Patapoutian, Andrea M Peier, Gina M Story, and Veena Viswanath. Thermotrp channels and beyond: mechanisms of temperature sensation. *Nature Reviews Neuroscience*, 4(7):529–539, 2003.

[69] Miriam García-Ávila and León D Islas. What is new about mild temperature sensing? a review of recent findings. *Temperature*, 6(2):132–141, 2019.

[70] Shigeru Saito, Nagako Banzawa, Naomi Fukuta, Claire T Saito, Kenji Takahashi, Toshiaki Imagawa, Toshio Ohta, and Makoto Tominaga. Heat and noxious chemical sensor, chicken trpa1, as a target of bird repellents and identification of its structural determinants by multispecies functional comparison. *Molecular biology and evolution*, 31(3):708–722, 2014.

[71] Yuji Karashima, Karel Talavera, Wouter Everaerts, Annelies Janssens, Kelvin Y Kwan, Rudi Vennekens, Bernd Nilius, and Thomas Voets. Trpa1 acts as a cold sensor in vitro and in vivo. *Proceedings of the National Academy of Sciences*, 106(4):1273–1278, 2009.

[72] Arijit Ghosh, Navneet Kaur, Abhishek Kumar, and Chandan Goswami. Why individual thermo sensation and pain perception varies? clue of disruptive mutations in trpvs from 2504 human genome data. *Channels*, 10(5):339–345, 2016.

[73] Jorge L Pérez-Moreno, Danielle M DeLeo, Ferran Palero, and Heather D Bracken-Grissom. Phylogenetic annotation and genomic architecture of opsin genes in crustacea. *Hydrobiologia*, pages 1–17, 2018.

[74] Diana J Rennison, Gregory L Owens, Nancy Heckman, Dolph Schluter, and Thor Veen. Rapid adaptive evolution of colour vision in the threespine stickleback radiation. *Proc. R. Soc. B*, 283(1830):20160242, 2016.

[75] Wayne IL Davies, Michael Turton, Stuart N Peirson, Brian K Follett, Stephanie Halford, Jose M Garcia-Fernandez, Peter J Sharp, Mark W Hankins, and Russell G Foster. Vertebrate ancient opsin photopigment spectra and the avian photoperiodic response. *Biology letters*, 8(2):291–294, 2012.

[76] Davide M Dominoni. The effects of light pollution on biological rhythms of birds: an integrated, mechanistic perspective. *Journal of Ornithology*, 156(1):409–418, 2015.

[77] Rui Borges, Imran Khan, Warren E Johnson, M Thomas P Gilbert, Guojie Zhang, Erich D Jarvis, Stephen J O'Brien, and Agostinho Antunes. Gene loss, adaptive evolution and the co-evolution of plumage coloration genes with opsins in birds. *BMC genomics*, 16(1):751, 2015.

[78] Daniel Escobar-Camacho, Erica Ramos, Cesar Martins, and Karen L Carleton. The opsin genes of amazonian cichlids. *Molecular ecology*, 26(5):1343–1356, 2017.

[79] Sara M Stieb, Fabio Cortesi, Lorenz Sueess, Karen L Carleton, Walter Salzburger, and N Justin Marshall. Why uv vision and red vision are important for damselfish (pomacentridae): structural and expression variation in opsin genes. *Molecular ecology*, 26(5):1323–1342, 2017.

[80] Marcela Petrusewicz-Kosińska, Barbara Przybylska-Gornowicz, Magdalena Prusik, Natalia Ziółkowska, and Bogdan Lewczuk. Pinopsin and photoreception in the pineal organ of the domestic turkey during post-hatching development. *Micron*, 126:102749, 2019.

[81] Stephanie Halford, Susana S Pires, Michael Turton, Lei Zheng, Irene González-Menéndez, Wayne L Davies, Stuart N Peirson, José M García-Fernández, Mark W Hankins, and Russell G Foster. Va opsin-based photoreceptors in the hypothalamus of birds. *Current Biology*, 19(16):1396–1402, 2009.

[82] Donald E Kroodsma and Bruce E Byers. The function (s) of bird song. *American Zoologist*, 31(2):318–328, 1991.

[83] Robert J Dooling, Bernard Lohr, and Micheal L Dent. Hearing in birds and reptiles. In *Comparative hearing: Birds and reptiles*, pages 308–359. Springer, 2000.

[84] James C Saunders. The role of hair cell regeneration in an avian model of inner ear injury and repair from acoustic trauma. *ILAR journal*, 51(4):326–337, 2010.

[85] Robert J Dooling and Sandra H Blumenrath. Avian sound perception in noise. In *Animal communication and noise*, pages 229–250. Springer, 2013.

[86] Otto Gleich and Geoffrey A Manley. The hearing organ of birds and crocodilia. In *Comparative hearing: Birds and reptiles*, pages 70–138. Springer, 2000.

[87] Robert C Beason. Through a bird's eye–exploring avian sensory perception. 2003.

[88] Matthew IM Louder, Henning U Voss, Thomas J Manna, Sophia S Carryl, Sarah E London, Christopher N Balakrishnan, and Mark E Hauber. Shared neural substrates for song discrimination in parental and parasitic songbirds. *Neuroscience letters*, 622:49–54, 2016.

[89] Christine R Lattin, Frank A Stabile, and Richard E Carson. Estradiol modulates neural response to conspecific and heterospecific song in female house sparrows: an in vivo positron emission tomography study. *PloS one*, 12(8), 2017.

[90] Claudio V Mello and David F Clayton. Song-induced zenk gene expression in auditory pathways of songbird brain and its relation to the song control system. *Journal of Neuroscience*, 14(11):6652–6666, 1994.

[91] David J Bailey and Juli Wade. Differential expression of the immediate early genes fos and zenk following auditory stimulation in the juvenile male and female zebra finch. *Molecular Brain Research*, 116(1-2):147–154, 2003.

[92] Tarciso AF Velho, Raphael Pinaud, Paulo V Rodrigues, and Claudio V Mello. Co-induction of activity-dependent genes in songbirds. *European Journal of Neuroscience*, 22(7):1667–1678, 2005.

[93] David Wheatcroft and Anna Qvarnström. A blueprint for vocal learning: auditory predispositions from brains to genomes. *Biology letters*, 11(8):20150155, 2015.

[94] Matthew IM Louder, Mark E Hauber, and Christopher N Balakrishnan. Early social experience alters transcriptomic responses to species-specific song stimuli in female songbirds. *Behavioural Brain Research*, 347:69–76, 2018.

[95] Philine Wangemann. K+ cycling and its regulation in the cochlea and the vestibular labyrinth. *Audiology and Neurotology*, 7(4):199–205, 2002.

[96] Hiroshi Hibino and Yoshihisa Kurachi. Molecular and physiological bases of the k+ circulation in the mammalian inner ear. *Physiology*, 21(5):336–345, 2006.

[97] Viviane Wilms, Christine Köppl, Chris Söffgen, Anna-Maria Hartmann, and Hans Gerd Nothwang. Molecular bases of k+ secretory cells in the inner ear: shared and distinct features between birds and mammals. *Scientific reports*, 6:34203, 2016.

[98] Viviane Wilms, Chris Söffgen, and Hans Gerd Nothwang. Differences in molecular mechanisms of k+ clearance in the auditory sensory epithelium of birds and mammals. *Journal of Experimental Biology*, 220(15):2701–2705, 2017.

[99] Katsuyuki Moriwaki and Osafumi Yuge. Topographical features of cutaneous tactile hypoesthetic and hyperesthetic abnormalities in chronic pain. *Pain*, 81(1-2):1–6, 1999.

[100] Shana L Geffeney and Miriam B Goodman. How we feel: ion channel partnerships that detect mechanical inputs and give rise to touch and pain perception. *Neuron*, 74(4):609–619, 2012.

[101] Mila I Svechtarova, Irene Buzzacchera, B Jelle Toebes, Jan Lauko, Nicoleta Anton, and Christopher J Wilson. Sensor devices inspired by the five senses: A review. *Electroanalysis*, 28(6):1201–1241, 2016.

[102] Martin Chalfie. Neurosensory mechanotransduction. *Nature reviews Molecular cell biology*, 10(1):44–52, 2009.

[103] Bertrand Coste, Bailong Xiao, Jose S Santos, Ruhma Syeda, Jörg Grandl, Kathryn S Spencer, Sung Eun Kim, Manuela Schmidt, Jayanti Mathur, Adrienne E Dubin, et al. Piezo proteins are pore-forming subunits of mechanically activated channels. *Nature*, 483(7388):176–181, 2012.

[104] Yoshiyuki Kawashima, Gwenaëlle SG Géléoc, Kiyoto Kurima, Valentina Labay, Andrea Lelli, Yukako Asai, Tomoko Makishima, Doris K Wu, Charles C Della Santina, Jeffrey R Holt, et al. Mechanotransduction in mouse inner ear hair cells requires transmembrane channel–like genes. *The Journal of clinical investigation*, 121(12):4796–4809, 2011.

[105] James M Heather and Benjamin Chain. The sequence of sequencers: The history of sequencing dna. *Genomics*, 107(1):1–8, 2016.

[106] Jay Shendure, Shankar Balasubramanian, George M Church, Walter Gilbert, Jane Rogers, Jeffery A Schloss, and Robert H Waterston. Dna sequencing at 40: past, present and future. *Nature*, 550(7676):345–353, 2017.

[107] Frederick Sanger, Steven Nicklen, and Alan R Coulson. Dna sequencing with chain-terminating inhibitors. *Proceedings of the national academy of sciences*, 74(12):5463–5467, 1977.

[108] Alice McCarthy. Third generation dna sequencing: pacific biosciences' single molecule real time technology. *Chemistry & biology*, 17(7):675–676, 2010.

[109] Ayman Grada and Kate Weinbrecht. Next-generation sequencing: methodology and application. *The Journal of investigative dermatology*, 133(8):e11, 2013.

[110] Sam Behjati and Patrick S Tarpey. What is next generation sequencing? *Archives of Disease in Childhood-Education Education Practice*, 98:236–238, 2013.

[111] HPJ Buermans and JT Den Dunnen. Next generation sequencing technology: advances and applications. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, 1842(10):1932–1941, 2014.

[112] Jay Shendure and Hanlee Ji. Next-generation dna sequencing. *Nature biotechnology*, 26(10):1135, 2008.

[113] Illumina history of sequencing by synthesis. https://www.illumina.com/science/technology/next-generation-sequencing/illumina-sequencing-history.html.

[114] Illumina Inc. Illumina sequencing technology: highest data accuracy, simple workflow, and a broad range of applications., 2010.

[115] Sara Goodwin, John D McPherson, and W Richard McCombie. Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, 17(6):333, 2016.

[116] Christoph Bleidorn. Third generation sequencing: technology and its potential impact on evolutionary biodiversity research. *Systematics and biodiversity*, 14(1):1–8, 2016.

[117] PACBIO smrt sequencing. https://www.pacb.com/smrt-science/smrt-sequencing/.

[118] John Eid, Adrian Fehr, Jeremy Gray, Khai Luong, John Lyle, Geoff Otto, Paul Peluso, David Rank, Primo Baybayan, Brad Bettman, et al. Real-time dna sequencing from single polymerase molecules. *Science*, 323(5910):133–138, 2009.

[119] Paul Flicek, M Ridwan Amode, Daniel Barrell, Kathryn Beal, Simon Brent, Denise Carvalho-Silva, Peter Clapham, Guy Coates, Susan Fairley, Stephen Fitzgerald, et al. Ensembl 2012. *Nucleic acids research*, 40(D1):D84–D90, 2011.

[120] David L Wheeler, Tanya Barrett, Dennis A Benson, Stephen H Bryant, Kathi Canese, Vyacheslav Chetvernin, Deanna M Church, Michael DiCuccio, Ron Edgar, Scott Federhen, et al. Database resources of the national center for biotechnology information. *Nucleic acids research*, 35(suppl_1):D5–D12, 2006.

[121] Howard S Bilofsky and Burks Christian. The genbank® genetic sequence data bank. *Nucleic acids research*, 16(5):1861–1863, 1988.

[122] Guy St C Slater and Ewan Birney. Automated generation of heuristics for biological sequence comparison. *BMC bioinformatics*, 6(1):31, 2005.

[123] Sudhir Kumar, Masatoshi Nei, Joel Dudley, and Koichiro Tamura. Mega: a biologist-centric software for evolutionary analysis of dna and protein sequences. *Briefings in bioinformatics*, 9(4):299–306, 2008.

[124] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014.

[125] Python Core Team. *Python: A dynamic, open source programming language*. Python Software Foundation, 2015.

[126] Daniel R Zerbino, Premanand Achuthan, Wasiu Akanni, M Ridwan Amode, Daniel Barrell, Jyothish Bhai, Konstantinos Billis, Carla Cummins, Astrid Gall, Carlos García Girón, et al. Ensembl 2018. *Nucleic acids research*, 46(D1):D754–D761, 2017.

[127] Stephen F Altschul, Warren Gish, Webb Miller, Eugene W Myers, and David J Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410, 1990.

[128] Hiroaki Iwata and Osamu Gotoh. Benchmarking spliced alignment programs including spaln2, an extended version of spaln that incorporates additional species-specific features. *Nucleic acids research*, 40(20):e161–e161, 2012.

[129] Osamu Gotoh. Direct mapping and alignment of protein sequences onto genomic sequence. *Bioinformatics*, 24(21):2438–2444, 2008.

[130] Lucien Le Cam. Maximum likelihood: an introduction. *International Statistical Review/Revue Internationale de Statistique*, pages 153–171, 1990.

[131] Marchal-BIOI team. Newick Tree file format. http://bioinformatics.intec.ugent.be/MotifSuite/treeformat.php, 2015.

[132] William N Venables, David M Smith, R Development Core Team, et al. An introduction to r, 2009.

[133] What is R? introduction to r. https://www.r-project.org/about.html. Accessed: 2019-01-25.

[134] John M Zelle. *Python programming: an introduction to computer science*. Franklin, Beedle & Associates, Inc., 2004.

[135] Jeff Chang, Brad Chapman, Iddo Friedberg, Thomas Hamelryck, Michiel De Hoon, Peter Cock, Tiago Antao, and Eric Talevich. Biopython tutorial and cookbook. *Update*, pages 15–19, 2010.

[136] Shonda A Leonard, Timothy G Littlejohn, and Andreas D Baxevanis. Common file formats. *Current protocols in bioinformatics*, 16(1):A–1B, 2006.

[137] Scott McGinnis and Thomas L Madden. Blast: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic acids research*, 32(suppl_2):W20–W25, 2004.

[138] Rebecca T Kimball, Carl H Oliveros, Ning Wang, Noor D White, F Keith Barker, Daniel J Field, Daniel T Ksepka, R Terry Chesser, Robert G Moyle, Michael J Braun, et al. A phylogenomic supertree of birds. *Diversity*, 11(7):109, 2019.

[139] Merly Escalona, Sara Rocha, and David Posada. A comparison of tools for the simulation of genomic next-generation sequencing data. *Nature Reviews Genetics*, 17(8):459, 2016.

[140] Todd J Treangen and Steven L Salzberg. Repetitive dna and next-generation sequencing: computational challenges and solutions. *Nature Reviews Genetics*, 13(1):36–46, 2012.

[141] Steven L Salzberg. Next-generation genome annotation: we still struggle to get it right, 2019.

[142] Daren C Card, Drew R Schield, Jacobo Reyes-Velasco, Matthew K Fujita, Audra L Andrew, Sara J Oyler-McCance, Jennifer A Fike, Diana F Tomback, Robert P Ruggiero, and Todd A Castoe. Two low coverage bird genomes and a comparison of reference-guided versus de novo genome assemblies. *PloS one*, 9(9), 2014.

[143] Jogikalmat Krithikadatta. Normal distribution. *Journal of conservative dentistry: JCD*, 17(1):96, 2014.

[144] Jason R Miller, Sergey Koren, and Granger Sutton. Assembly algorithms for next-generation sequencing data. *Genomics*, 95(6):315–327, 2010.

[145] Suparna Mitra, Max Schubach, and Daniel H Huson. Short clones or long clones? a simulation study on the use of paired reads in metagenomics. *BMC bioinformatics*, 11(1):S12, 2010.

[146] Jonas Korlach, Gregory Gedman, Sarah B Kingan, Chen-Shan Chin, Jason T Howard, Jean-Nicolas Audet, Lindsey Cantin, and Erich D Jarvis. De novo pacbio long-read and phased avian genome assemblies correct and add to reference genes generated with intermediate and short reads. *GigaScience*, 6(10):gix085, 2017.

[147] Rachel L Goldfeder, James R Priest, Justin M Zook, Megan E Grove, Daryl Waggott, Matthew T Wheeler, Marc Salit, and Euan A Ashley. Medical implications of technical accuracy in genome sequencing. *Genome medicine*, 8(1):24, 2016.

[148] Tomáš Hron, Petr Pajer, Jan Pačes, Petr Bartněk, and Daniel Elleder. Hidden genes in birds. *Genome biology*, 16(1):164, 2015.

[149] Josefin Stiller and Guojie Zhang. Comparative phylogenomics, a stepping stone for bird biodiversity studies. *Diversity*, 11(7):115, 2019.

[150] Indrajit Nanda, David Schrama, Wolfgang Feichtinger, Thomas Haaf, Manfred Schartl, and Michael Schmid. Distribution of telomeric (ttaggg) n sequences in avian chromosomes. *Chromosoma*, 111(4):215–227, 2002.

[151] Linda Beauclair, Christelle Ramé, Peter Arensburger, Benoît Piégu, Florian Guillou, Joëlle Dupont, and Yves Bigot. Sequence properties of certain gc rich avian genes, their origins and absence from genome assemblies: case studies. *BMC genomics*, 20(1):1–16, 2019.

[152] Zongji Wang, Jilin Zhang, Wei Yang, Na An, Pei Zhang, Guojie Zhang, and Qi Zhou. Temporal genomic evolution of bird sex chromosomes. *BMC evolutionary biology*, 14(1):250, 2014.

A

LISTINGS

```
import os
path = os.getcwd() #Save directory

for filename in os.listdir(path): #for all files in the directory
    if filename.endswith("cut.txt"):#if the file~ name end in "cut.txt"
        name = filename #save the name of the file
        f = open(name, 'r', newline = '\n') #open that  file
        i = 0
        j = 4
        fin = []
        for line in f: #runs each line of the file
            i += 1
            if i == j : #corresponds to the DNA sequence
                fin.append(line) #save that specific line
                j += 5 #go to the next line with DNA sequence
        f.close()

        upper=[] #list for all upper letters
        for el in fin: #for each element of the list with the DNA sequences
            for i in el: #for each character in each string (DNA sequence)
                if i.isupper(): #if that is an upper case character
                    upper.append(i) #save in the upper list
        result = ''.join(upper) #put all characters together in a single string

        num = name.count('_') #count the number of "_" of the file original name
        final_name= name.split('_') #split the original name by "_" and save it
        title = ''
        for i in range(0, num):
            title += final_name[i]
            if i < num-1:
                title+='_'

        title1 = title + '.fas' # it uses the title created before and adds the ".fas" extension,
        so that is created a FASTA format file

        j = open(title1, 'w') #create the file with the title1

        j.write('>' + title + '\n') #1st line will have the title of the file,
        so that is possible to have the file in FASTA format.Ex: >TAAR2_Calypte_anna_JJRV01

        j.write(result) #write the string with all the upper case
        characters (all nucleotides of the DNA sequence) saved before
        j.close()
```

Figure 12.: Script developed in Python programming language to extract, from the Exonerate software output files, only what corresponded to the DNA sequences. With this script we were able to rapidly creating a new file for each gene, with its nucleotide sequence, in FASTA format.