Doctoral Dissertation

# A genomic view on the adaptation and diversification of natural populations

SANTIAGO MONTERO-MENDIETA

Advisors: Carles Vilà, Matthew T. Webster



*Programa de Doctorado en Biología Integrada*

A dissertation submitted in fulfillment of the requirements for the degree of Doctor of Philosophy at University of Seville, 2019

**Abstract**

Earlier studies in evolutionary genetics focused on a few model organisms such as fruit flies or mice that are limited when it comes to answering evolutionary and ecological questions. In contrast, genetic studies of natural populations have now become common and can provide a more realistic understanding of how natural selection, genetic drift, mutation, and gene flow shape the patterns of phenotypic and genetic diversity, as well as adaptation and diversification across a range of environmental conditions. This dissertation illustrates how current genomic tools can be effectively used to (1) study the evolutionary history of species along altitudinal gradients, and (2) understand the genetic basis of adaptive phenotypes in populations inhabiting high-altitude habitats. In the first research paper, we compared whole-genome resequencing data of Eastern honey bee (*Apis cerana*) populations from high and low altitudes in southwestern China. We identified several regions of the genome that appeared to have been under positive selection in highland bee populations. Candidate loci in these genomic regions included genes related to reproduction and feeding behavior. In the second paper, we generated a transcriptome reference for the Neotropical frogs of the genus *Oreobates* by sequencing RNA from one individual of the La Paz robber frog (*Oreobates cruralis*). In the third paper, we used that transcriptome to selectively target and enrich ~18,000 genes across species of *Oreobates* collected along the Andean Mountains, in South America. We found that highland species have smaller effective populations and accumulate nonsynonymous mutations faster than species sampled at lower altitudes. These mutations can be targeted by natural selection and contribute to the adaptation and differentiation of populations in mountain environments. In the fourth and final paper, we pointed out that genomics has to be integrated with other sources of evidence to understand evolutionary and ecological processes more deeply than was thought in the past.

**Keywords:** Diversification, Evolution, Honey bee, Neotropical frogs, Natural selection, Population genomics, Phylogenomics, Whole-genome resequencing, Reduced representation sequencing.

**Citation:** Montero-Mendieta, S. (2019). A genomic view on the adaptation and diversification of natural populations. University of Seville. 52 pp.

Para mis padres y hermanos

也献给李菡薏

"*Nothing in biology makes sense*

*except in the light of evolution*"

Theodosius Dobzhansky, 1973

# List of Papers

This dissertation is based on the following papers, which are referred to in the text by their Roman numerals:

I. **Montero-Mendieta**, **S**.*, Tan, K.*, Christmas, M.J., Olsson, A., Vilà, C., Wallberg, A., Webster, M.T. (2018). The genomic basis of adaptation to high-altitude habitats in the Eastern honey bee (*Apis cerana*). *Molecular Ecology*, **28**, 746–760.

II. **Montero-Mendieta**, **S**., Grabherr, M., Lantz, H., De la Riva, I., Leonard, J.A., Webster, M.T., Vilà, C. (2017). A practical guide to build de-novo assemblies for single tissues of non-model organisms: the example of a Neotropical frog. *PeerJ*, **5**, e3702.

III. **Montero-Mendieta**, **S**., De la Riva, I., Leonard, J.A., Webster, M.T., Vilà, C. (in preparation). Phylogenomics and evolutionary history of Neotropical frogs of the genus *Oreobates* along altitudinal gradients.

IV. **Montero-Mendieta**, **S**.* & Dheer, A.* (2019). Digest: Resolving phylogenomic conflicts in characiform fishes. *Evolution*, **73**, 416–418.

* These authors contributed equally.

I also contributed to the following papers during the course of my doctoral studies:

- **Montero-Mendieta**, **S**., Ferrer, J., Hammou, M.A., Dahmani, W., Sanuy, D., Camarasa, S. (2017). Another record or a new taxon? A candidate species of *Chalcides* Laurenti, 1768, in North Africa (Squamata: Sauria: Scincidae). *Herpetozoa*, **29**, 155–161.

- Rodríguez, A., Dugo-Cota, A., **Montero-Mendieta**, **S**., Alonso, R., Vences, M., Vilà, C. (2017). Cryptic within cryptic: genetics, morphometrics, and bioacoustics delimitate a new species of *Eleutherodactylus* (Anura: Eleutherodactylidae) from Eastern Cuba. *Zootaxa*, **4221**, 501–522.

- Vasconcelos, R.*, **Montero-Mendieta**, **S**.*, Simó-Riudalbas, M., Sindaco, R., Santos, X., Fasola, M., Llorente, G.A., Razzetti, E., Carranza, S. (2016). Unexpectedly high levels of cryptic diversity uncovered by a complete DNA barcoding of reptiles of the Socotra Archipelago. *PLoS ONE*, **11**, e0149985.

* These authors contributed equally.

# Declaration

**Carles Vilà**, Research Professor at the Estación Biológica de Doñana (EBD-CSIC), Spain and **Matthew Webster**, Professor at the Department of Medical Biochemistry and Microbiology (IMBIM), Uppsala University, Sweden, advisors of this doctoral dissertation certify that the work has been carried out by **Santiago Montero-Mendieta** and it is suitable to be defended in front of a scientific committee. Advisors assisted in designing, guiding and correcting drafts of the thesis and manuscripts. The contribution of the Ph.D. candidate (S.M.-M.) to each manuscript, the publication status, and the Impact Factor (IF) of the published papers according to the latest available version of the ISI Journal Citation Reports is detailed below:

**Paper I**: S.M.-M. performed the experiments, analyzed the data, prepared figures/tables, and wrote the manuscript. *Molecular Ecology* has an IF of 5.855. This journal is in the first quartile of the areas "Ecology" (13 of 164) and "Evolutionary Biology" (8 of 50).

**Paper II**: S.M.-M. conceived, designed, and performed the experiments, analyzed the data, prepared figures/tables, and wrote the manuscript. *PeerJ* has an IF of 2.353. This journal is in the first quartile (35 of 272) of the area "Agricultural and Biological Sciences (miscellaneous)".

**Paper III**: S.M.-M. designed and performed the experiments, analyzed the data, prepared figures/tables, and wrote the manuscript. The first version of the paper has been finished, and will be submitted for publication soon.

**Paper IV**: S.M.-M. conceived and wrote the manuscript, and prepared figures. *Evolution* has an IF of 3.573. This journal is in the first quartile (40 of 164) of the area "Evolutionary Biology".

Carles Vilà Matthew Webster

# Contents

# Abbreviations

| | |
|---|---|
| BCE | Before Common Era |
| bp | Base pair |
| cDNA | Complementary DNA |
| CDS | Coding sequence |
| DNA | Deoxyribonucleic acid |
| IUCN | International Union for Conservation of Nature |
| MRCA | Most recent common ancestor |
| mRNA | Messenger RNA |
| MSA | Multiple sequence alignment |
| MYA | Million years ago |
| NCBI | National Center for Biotechnology Information |
| NGS | Next-generation sequencing |
| NHGRI | National Human Genome Research Institute |
| PCR | Polymerase chain reaction |
| RNA | Ribonucleic acid |
| RNA-seq | RNA-sequencing |
| SNP | Single nucleotide polymorphism |
| UTR | Untranslated region |
| WGS | Whole-genome sequencing |

# Glossary

1-TO-1 ORTHOLOGS: only one ortholog (i.e., homolog that evolved by a speciation event) copy is found in each species.

ADAPTATION: (1) process by which organisms become fit to their environment; (2) any heritable trait that improves the ability of an individual organism to survive and reproduce.

BACKGROUND SELECTION: loss of genetic diversity at a non-deleterious locus due to negative selection against linked deleterious alleles.

BOTTLENECK: reduction in a population's size leading to a greater genetic drift.

CONVERGENT EVOLUTION: independent development of analogous features in species of different lineages.

COVERAGE: the number of unique sequencing reads that include a given nucleotide in the reconstructed sequence.

DIRECTIONAL SELECTION: selection that favors the fixation of one particular allele in a population.

DISRUPTIVE SELECTION: selection that favors extreme values for a trait over intermediate values.

DIVERSIFICATION: (1) independent accumulation of genetic changes in two or more populations often after reduced gene flow between them; (2) formation of new species from a common ancestor (speciation).

GENE FLOW: transfer of genetic variants from one population to another.

GENETIC DRIFT: random change in allele frequency in finite populations over time.

GENOTYPE: the particular combination of alleles of a locus for a given individual.

HITCHHIKING: increase in the frequency of neutral DNA variants linked to positively selected sites.

HOMOLOGY: existence of shared ancestry between a pair of structures or genes in different taxa.

HOMOPLASY: existence of characters shared by a set of different taxa but not present in their common ancestor.

HORIZONTAL GENE TRANSFER: movement of genetic material between organisms other than by reproduction.

HYBRIDIZATION: process by which individuals from two divergent lineages or species combine their genomes to conceive a new organism with intermediate genetic composition.

INBREEDING: production of offspring from the mating or breeding of individuals or organisms that are closely related genetically.

LOCAL ADAPTATION: traits that produce higher fitness in a given locality than elsewhere.

PHENOTYPE: the observable traits of an organism.

POPULATION STRUCTURE: differences in allele frequencies between subpopulations in a population.

RECOMBINATION: the exchange of genetic material either between multiple chromosomes or between different regions of the same chromosome.

SELECTIVE SWEEP: the reduction or elimination of variation at sites that are physically linked to a site under directional selection.

# 1. Introduction

## 1.1. The study of evolution via DNA

### 1.1.1. Historical context

Evolution, the conception that all living organisms are the result of a series of gradual changes over time from common ancestors, has an ancient origin. Back in the 6th century BCE, the philosopher Anaximander of Miletus speculated that life initially originated in water and later extended to land. He also put forward the idea that humans were born from other kinds of animals, likely fish. In the 4th century BCE, Empedocles suggested that only those organisms "accidentally compounded in a suitable way" had been able to survive. Much later, in the 18th century, the naturalist Jean-Baptiste Lamarck proposed that species slowly adapt to their environment based on the use or disuse of particular organs through generations. But it was not until the mid-19th century that biologists Charles Darwin and Alfred Russel Wallace unveiled the theory of evolution through natural selection, which was presented in detail in Darwin's book 'The Origin of Species' (1859).

Darwin argued that organisms with the combination of traits that better fits the environment have more chances of surviving and producing offspring than those with traits less well fit, thus being 'naturally selected' (Darwin, 1859). He described natural selection as the principle by which any variety of a trait, if useful, is inherited by offspring and spread in a population over time. However, Darwin itself did not exactly know what was being inherited from one generation to the next. Later on, Gregor Mendel through his work on pea plants during the 1860s, concluded that variations in inherited characteristics were controlled by heredity factors, and formulated several laws to explain how traits are passed between generations (Mendel, 1866).

Mendel's heredity factors were eventually recognized as 'genes' after a series of revelations, including the discovery of nucleic acids by Friedrich Miescher in 1869 and their characterization by Rosalind Franklin in 1951, that culminated in the description of the structure of DNA by James Watson

and Francis Crick in 1953. The integration of all these findings and the theoretical work on population genetics by Ronald Fisher, J.B.S. Haldane, and Sewall Wright, together with the contributions of Ernst Mayr, George Simpson, and Theodosius Dobzhansky, helped to unify the research fields of ecology, evolution, and genetics, resulting in the establishment of the modern evolutionary synthesis.

### 1.1.2. DNA sequencing and genomics

In 1956, Arthur Kornberg discovered the DNA polymerase, an enzyme that synthesizes DNA molecules from deoxyribonucleotides. This finding was crucial for the development of one of the first DNA sequencing technologies by Frederick Sanger in the late 1970s. Through Sanger sequencing, scientists had access to the genetic information of life stored in sequences of DNA. The availability of such biological sequences enabled both empirical and theoretical molecular research and provided new sources of evidence to study evolutionary processes such as adaptation and diversification in natural populations. These advances in the field of evolutionary biology were further boosted by the invention of the polymerase chain reaction (PCR) by Kary Mullis in the mid-1980s, which allowed rapidly amplifying specific fragments of DNA across many species.

However, sequencing large genomes using the Sanger method was slow and costly. The sequence of the human genome took 13 years to be completed (1990–2003) at a cost of $2.7 billion (Consortium, 2004). This high investment of time and resources stressed the need for developing technologies capable of sequencing large amounts of DNA at a relatively faster speed and lower price compared to previous methods. In 2005, the debut of the Roche 454 pyrosequencing lead to the next era of sequencing technologies, termed ‘next-generation sequencing’ (NGS) or ‘high-throughput sequencing’ (HTS), which are the main ones used today. These technologies make affordable and less time-consuming to generate millions of short sequences per sample instead of a single sequence based on a pool of molecules as in Sanger sequencing (Mardis, 2008) (**Figure 1**). Using these methods, biologists not only can now achieve lots more data than previously possible but also obtain genomic data of all species and not just classic model organisms (Seehausen et al., 2014).

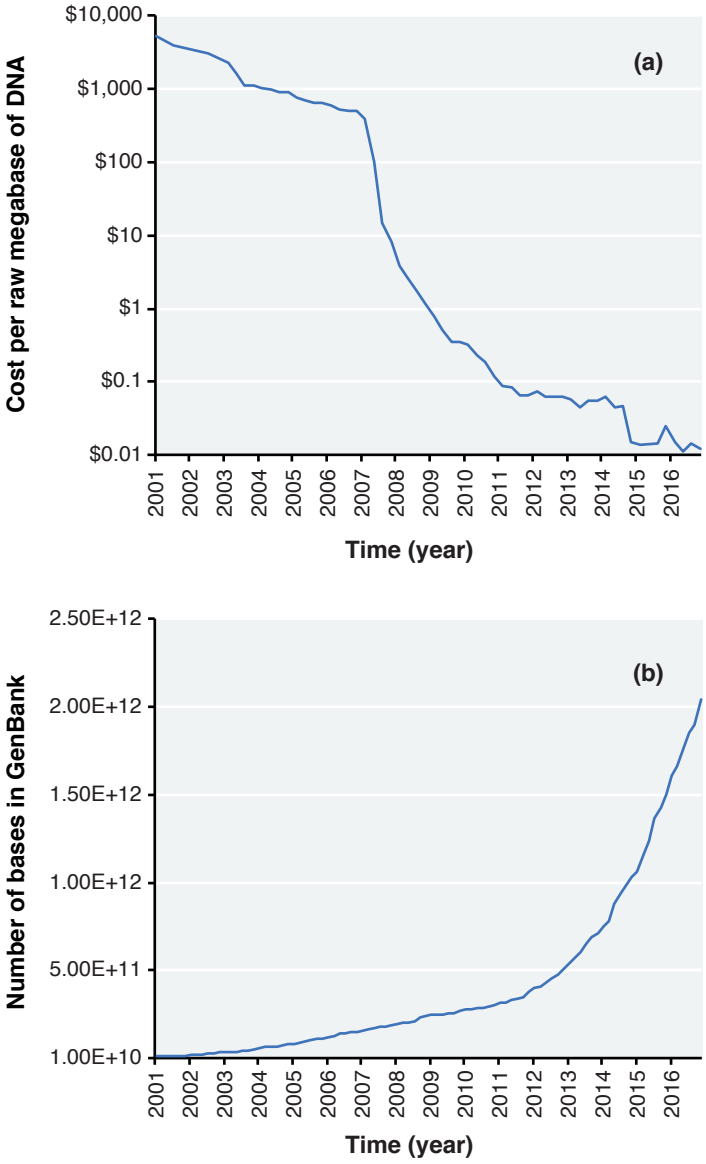


**Figure 1**. The rise of DNA sequencing technologies over time (2001–2016). (a) Reduction of sequencing cost per raw megabase of DNA as estimated by the National Human Genome Research Institute (NHGRI). (b) Growth in the number of base pairs (bp) stored in GenBank.

### 1.1.3. NGS applications in evolutionary biology

The accessibility of genome-wide data has undoubtedly been one of the greatest advances in the field of evolutionary biology. Researchers are no longer restricted to sequencing a small number of molecular markers such as microsatellites or amplified fragment length polymorphisms (AFLPs), which do not give a complete picture of the evolutionary processes across the genome. Instead, loci distributed across the whole genome, including protein-coding and putatively neutral sequences, can now be used to detect functional adaptive variants, as well as to improve the understanding of the phylogenetic relationships among organisms (Steiner et al., 2013). One popular example is Darwin's finches on the Galápagos Islands, whose beaks range from small insect-crunchers to large seed-demolishers. Through genomics, researchers not only were able to better understand the evolutionary history of these iconic birds but also to identify genes that lie behind the variation of beak shape and size (Lamichhaney et al., 2015, 2016).

In the current genomic era, the relatively low cost of sequencing new genomes is making a large number of species 'genome-enabled'. At the time of writing this dissertation, roughly 10,000 species have their genome fully or partially sequenced according to NCBI. Among these species, approximately 6,000 are prokaryotes (96% bacteria; 4% archaea) and 4,000 eukaryotes (47% fungi; 34% animals; 11% plants; 8% protists). This large volume of data allows studying homology on a genome-wide scale across a wide range of species, which can reveal previously unappreciated complexity in the evolutionary processes that shaped genomes (Xia, 2011; Rogers & Gibbs, 2014). Moreover, annotated reference genomes are needed to find out how recombination occurs with natural selection (Hoban et al., 2016). In honeybees, for example, recombination plays an important role in the genome's evolution by decreasing the effect of natural selection (Wallberg et al., 2015).

Genomic resources can also boost conservation efforts and management of threatened species like the Iberian Lynx, by identifying loci involved in inbreeding depression using population genomic analyses (Abascal et al., 2016). Population genomics can help to understand evolutionary processes that influence variation across genomes and populations, as it can separate genome-wide effects (e.g., genetic drift, gene flow, and inbreeding) from

locus-specific effects (e.g., natural selection and mutation) (Black et al., 2001). By allowing the identification and filtering of genetic variation under selection, population genomics enables unbiased inference of evolutionary history via accurate estimation of genome-wide, neutral demographic parameters such as effective population size ($N_e$), without the confounding effects of natural selection (Luikart et al., 2003; Funk et al., 2018).

### 1.1.4. Phylogenetics in the genomics era

Understanding the evolutionary history and relationships among individuals or groups of organisms (species and populations) has been a long-standing challenge for biologists. Phylogenetics uses heritable characters to study the relationships of both extinct and current species, which are represented using branching diagrams, referred to as 'trees'. A phylogenetic tree consists of clades, monophyletic groups of species including their most recent common ancestor (MRCA) and its descendants, that can be either nested or mutually exclusive (Castroviejo-Fisher, 2009).

In the dawn of phylogenetics, trees were reconstructed using homologous phenotypic (usually morphological but also behavioral) traits. Later on, with the development of Sanger sequencing and PCR amplification, molecular characters were incorporated into phylogenetic reconstructions. Molecular phylogenetics has traditionally used mitochondrial (e.g., COI, cytochrome b, 16S rRNA) or nuclear (e.g., MC1R, MOS, RAG1) markers to estimate the evolutionary history of species. However, it has been known for decades that the resulting history of single genes or 'gene trees' is not necessarily equivalent to the entire evolutionary history of the species or 'species tree', as different genes can explain different histories (Goodman et al., 1979; Pamilo & Nei, 1988; Takahata, 1989). These gene trees that do not match the underlying species tree form the "anomaly zone" (Degnan & Rosenberg, 2006). In a gene tree, the nodes correspond to coalescent events (i.e., the point where two lineages join in their MRCA), whereas the length of the branches usually represents the number of nucleotide substitutions. In a species tree, internal nodes depict speciation events and its branches reflect the population history between speciations (Mallo & Posada, 2016).

Despite being conceptually different, species and gene trees are expected to be topologically equivalent under many evolutionary scenarios. However, sampling or stochastic errors caused by the finite length of markers used in the inference can disrupt this equivalence and decouple their histories. That is, the smaller the dataset the greater the chance that few homoplastic positions have an impact on the species tree (Delsuc et al., 2005). In addition, discordances between gene and species trees can also result of different evolutionary processes, including convergent evolution, hybridization, gene flow, gene duplication, gene loss, recombination, and incomplete lineage sorting, a phenomenon that occurs when ancestral polymorphisms are incompletely sorted and persist during successive speciation events due to coalescent stochasticity (Maddison, 1997) (**Figure 2**).

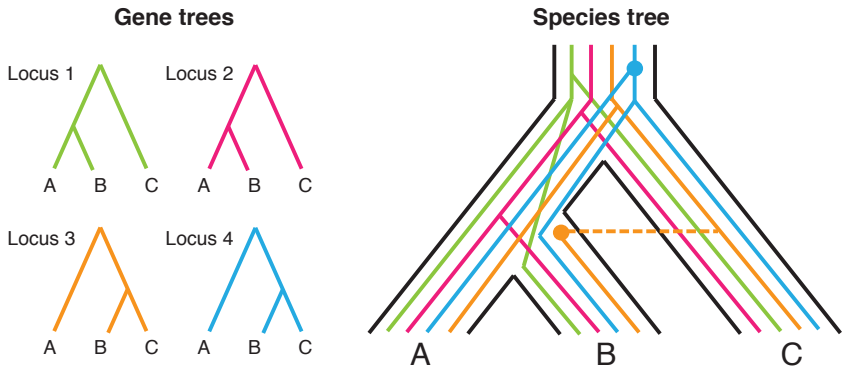


**Figure 2**. Gene trees and species tree conflicts. The species tree of A, B, and C is painted in black. In green (Locus 1) and pink (Locus 2) are two gene trees congruent with the species tree, i.e., with A and B being sister species. In orange (Locus 3), the tree of a gene undergoing gene flow (or horizontal gene transfer) between species B and C. In blue (Locus 4), the tree of a gene undergoing incomplete lineage sorting. Adapted from Marin et al. (2019).

The recent advances in NGS technologies have enabled using large amounts of sequence data generated by broad-scale genome projects to infer trees with many loci, turning the field of phylogenetics into phylogenomics (Rannala & Yang, 2008). Phylogenomics can be defined as the study or reconstruction of the evolutionary relationships between organisms using genomic data (see **1.2.2.2. Phylogenomics**). Because it uses a large number of genetic markers, phylogenomics not only leads to a drastic reduction in

sampling or stochastic errors, but also has the power to unveil heterogeneous signals across different genomic regions, and identify discordances between gene trees and species trees (Delsuc et al., 2005; Bravo et al., 2019). Genomic data has allowed to clarify phylogenetic relationships among many non-model vertebrate lineages (i.e., species in which extensive genomic resources are not available) such as salamanders (Shen et al., 2013), birds (Jarvis et al., 2014), snakes (Ruane et al., 2015), and fishes (Betancur-R. et al., 2019). Such accurate inferences of the tree of life are basic to understand what processes have shaped the evolutionary history of species across a range of environmental conditions.

### 1.1.5. The genomics of local adaptation

The study of the ecological and genetic mechanisms that allow populations to adapt to local environmental conditions offers valuable insights into how species evolved and diversified. Detecting genomic regions and variants involved in local adaptation is also important in conservation to secure that the maximum amount of additive genetic variation is conserved (Allendorf et al., 2010). Studies on local adaption, for instance, can help to ensure that source populations for augmentation of declining populations are not adaptively divergent from the target population, which could lead to outbreeding depression rather than the desired genetic rescue (Edmands, 2007; Frankham et al., 2011; Funk et al., 2018).

During the course of local adaptation, directional selection increases the frequency of genetic variants associated with beneficial phenotypic traits. Additionally, natural selection also raises the frequency of neutral or slightly deleterious alleles around the selected locus through genetic hitchhiking or background selection (Smith & Haigh, 1974; Barton, 2000). This situation called 'selective sweep' results in a reduction of genetic diversity around alleles under selection, which can be used to detect adaptive variants (Cutter & Payseur, 2013) (**Figure 3**).

Nowadays, the gold standard to identify selective sweep regions or adaptive molecular variation is using population genomics. Genome-wide data offers many advantages over sparser sets of genetic markers. One advantage, perhaps the most important, is the potential to detect outlier loci (i.e., genomic

regions that are extremely divergent from the rest of the genome) that are associated with adaptive phenotypes (Pardo-Diaz et al., 2015). These outlier loci are usually characterized by either being highly differentiated (between populations with distinct adaptations) or little differentiated (between populations with similar adaptations) relative to what is expected under the Hardy–Weinberg equilibrium (Luikart et al., 2003).

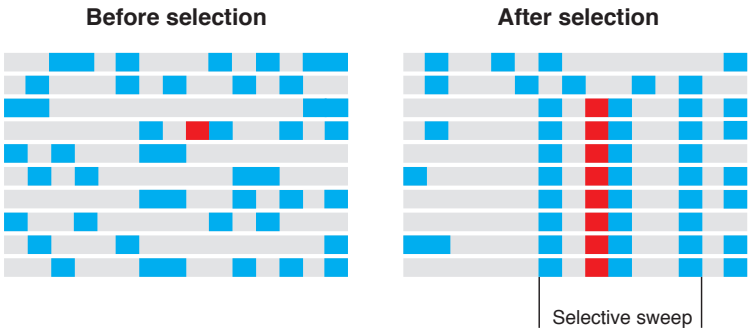


**Figure 3**. A selective sweep. Ancestral alleles are shown in grey and derived (non-ancestral) alleles are shown in blue. Under selection, a new positively-selected allele (red) rises to high frequency and nearby linked alleles 'hitch-hike' along with it.

Several statistical approaches are available to detect natural selection using genomic data, such as genome-wide association studies (GWAS) or selection scans. Furthermore, the sequencing of protein-coding regions enables genome-wide tests for selection based on comparing rates of synonymous and non-synonymous mutations at different sites (Savolainen et al., 2013). However, a major challenge of such studies is to distinguish signals generated by natural selection from the ones caused by genetic drift, population bottlenecks, population structure and other demographic factors (Staubach et al., 2012). The integration of large numbers of putatively neutral and selective markers offer the opportunity to address these issues. A very famous example concerns the adaptation to high-altitude of certain human populations, who seem to have independently acquired a series of physiological adaptations associated with heritable behavioral and genetic changes, to cope with low oxygen levels (Bigham et al., 2010; Peng et al., 2011; Xu et al., 2011; Huerta-Sánchez et al., 2013).

# 1.2. Genomic methods in evolutionary biology

The rise of genomics has been revolutionary in terms of both data collection and bioinformatic analyses. The huge amount of data produced by NGS technologies required developing complex bioinformatics analysis pipelines and innovative storage solutions. Additionally, the data obtained in this way have higher error rates and often shorter read lengths than traditional Sanger sequencing, leading to new methodological challenges (Mardis, 2011; Goodwin et al., 2016). I summarize next some of the most common genomic approaches, with an emphasis on those used in this dissertation to collect and analyze genomic data.

## 1.2.1. Genomic data

### 1.2.1.1. Whole-genome sequencing

Whole-genome sequencing (WGS) is the determination of the complete DNA sequence of an organism's genome using NGS. Unlike Sanger sequencing, which only generates single independent sequences, NGS techniques are massively parallel and one reaction produces hundreds or thousands of independent short sequences of DNA, called 'sequencing reads' (Ekblom & Wolf, 2014). Many NGS platforms are available for WGS which differ in the length of reads, the number of sequences generated in parallel (throughput), runtime, error rate and cost (**Table 1**). Illumina currently accounts for the largest market share with a wide range of instruments that deliver short reads (150–300 bp) with relatively low error rates. Long-range sequencing technologies capable of producing long reads expanding several kilobase pairs (kbp) have flourished recently, such as PacBio and MinION. Their chief advantage lies in their ability to span low-complexity repetitive regions in genome assemblies (Wallberg et al., 2019). However, these technologies have higher error rates than Illumina, thus requiring to be used in conjunction with accurate short reads to produce high-quality reference genomes (Jain et al., 2018).

Regardless of the platform, DNA must be processed and assembled to a library before its sequencing. The workflow for preparing sequencing libraries consists of 4 main steps: (1) extraction or isolation of the DNA; (2) random fragmentation of the DNA and selection of fragments of a predefined length; (3) ligation of technology-appropriate adapter sequences to the ends; and (4) PCR amplification to generate amplicons to be used as sequencing templates (Sambrook et al., 1989).

**Table 1**. Main current NGS platforms.

| Platform | Read length | Throughput | Runtime | Error rate | Cost per Gbp |
|---|---|---|---|---|---|
| SOLiD 5500xl | 75 bp | 180–240 Gb | ~7 days | ~0.1% substitution | $$ |
| *Roche* 454 GS FLX+ | Up to 1 kbp | ~700 Mbp | ~23 h | ~1%, indel | $$$$$ |
| *IonTorrent* S5 (540 chip) | 200 bp | 10–15 Gbp | ~2.5 h | ~1%, indel | $$$ |
| *Illumina* MiSeq v3 | 300 bp* | 13.2–15 Gbp | ~56 h | ~0.1%, substitution | $$$ |
| *Illumina* NextSeq 550 | 150 bp* | 100–120 Gbp | ~29 h | <1%, substitution | $$ |
| *Illumina* HiSeq 2500 | 125 bp* | 0.1–1 Tbp | 1–6 days | ~0.1%, substitution | $$ |
| *Illumina* HiSeq X Ten | 150 bp* | 1.6–1.8 Tbp | <3 days | ~0.1%, substitution | $ |
| *Illumina* NovaSeq S4 | 150 bp* | ~6 Tbp | <2 days | ~0.1%, substitution | $ |
| *Pacific BioSciences* PacBio RS II | ~20 kbp | 0.5–1 Gbp | ~4 days | 10–15% | $$$ |
| *Oxford Nanopore* MinION | Up to 200 kbp | 0.5–1.5 Gbp | ~2 days | 5–10% | $$$$ |

* paired-end sequencing; bp: base pairs; kbp: kilobase pairs; Mbp: megabase pairs; Gbp: gigabase pairs; Tbp: terabase pairs. Adapted from Escoda-Assens (2018).

### 1.2.1.2. Reduced representation sequencing

For non-model organisms with large genomes in which WGS is still not cost-effective, multiple approaches to reduce genome complexity and enrich selected regions have been recently developed (Turner et al., 2009). Such 'reduced representation' methods make possible to sequence and genotype hundreds if not thousands of homologous loci for multiple individuals across

different populations from one or many species (Zhang et al., 2019). The resulting genome-wide data, either as sequences or SNPs, can then be compared to address questions in an evolutionary and ecological context (Rokas & Abbot, 2009). This enables genomic studies in species that lack a reference genome, such as most amphibians (Gregory et al., 2007).

Reduced representation sequencing strategies include: (1) restriction-digest based methods such as RADseq (Davey & Blaxter, 2010) or genotyping by sequencing (GBS) (Elshire et al., 2011); (2) RNA-sequencing (RNA-seq); and (3) sequence capture or target-enrichment methods based on molecular probes (also called 'baits') or ultra-conserved elements (UCEs) (Mamanova et al., 2010; Faircloth et al., 2012). In the sequence capture approach, genomic libraries are hybridized with probes that are complementary to target regions (Bi et al., 2012; Faircloth et al., 2012; Lemmon et al., 2012).

One widely-used target region is the transcriptome, the entire collection of the mRNA molecules (i.e., transcripts) expressed from the genes of an organism. These transcripts dynamically change in response to genetic, environmental or physiological factors, such as the development stage or the tissue (Domazet-Lošo & Tautz, 2010; Melé et al., 2015). Using RNA-seq, the RNA in the cells is reverse-transcribed to cDNA and sequenced, thus providing a snapshot of the set of expressed genes and their abundances in an individual at a given time (Wang et al., 2009). RNA-seq has proven to be a powerful tool for cost-effectively identifying molecular markers, particularly in species that lack a reference sequence (Hirsch et al., 2014).

In the first research paper of this dissertation (**Paper I**), my colleagues and I performed whole-genome resequencing of Eastern honey bees (*Apis cerana*) using the Illumina HiSeq 2500 platform. Moreover, in that study, we also used SOLiD data downloaded from NCBI. In the second paper (**Paper II**), we generated a transcriptome reference for the Neotropical frogs of the genus *Oreobates* by sequencing RNA from one individual of the La Paz robber frog (*Oreobates cruralis*). In the third paper (**Paper III**), we used that transcriptome to selectively target and enrich ~18,000 genes via sequence capture across all species of *Oreobates*, and some other species of closely-related genera (**Figure 4**). All the sequencing in these two papers was also carried out on Illumina HiSeq 2500 instruments.

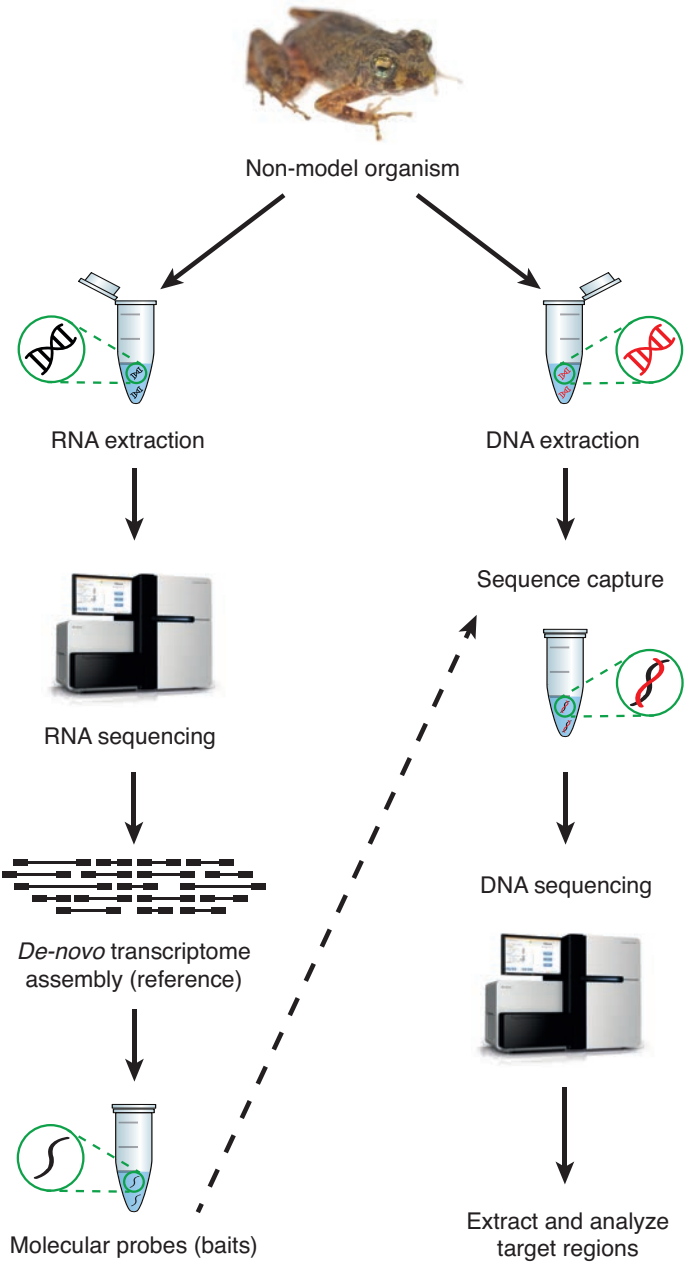


**Figure 4**. Reduced representation pipeline used in this thesis to extract and analyze target regions of interest in a non-model group of frogs.

## 1.2.2. Bioinformatics analyses

### 1.2.2.1. Transcriptome assembly

Transcriptome assembly consists of aligning and merging short RNA-seq reads in order to reconstruct the original sequence of the expressed transcripts. By sequencing a large number of such reads, a large fraction of the transcriptome is covered by overlapping sequences that can be assembled into longer regions. In species without a reference genome, or if only a fragmented draft genome is available, it is possible to build up the transcriptome using a *de-novo* assembly program such as Trinity (Grabherr et al., 2011) that does not require any prior knowledge of the genome.

However, *de-novo* transcriptome assembly is a challenging task as it usually results in fragmented transcripts and not truly complete sequences (Conesa et al., 2016). Particularly in eukaryotes, reconstructing the transcriptome is not trivial due to alternative spliced transcripts sharing sequence content, which makes not always possible to uniquely assign a read to a transcript (Sibbesen, 2016). It is also problematic to discriminate between transcript variants expressed several times from paralog genes (i.e., homolog genes separated by gene duplication events) (Vijay et al., 2013). These issues can be minimized by performing *in silico* normalization of the data, which removes redundant transcripts without impacting the assembly (Brown et al., 2012).

The quality of the resulting assembly or transcriptome will depend on its accuracy, completeness, and contiguity (Moreton et al., 2016). Commonly used metrics to assess the assembly's quality when there is no close reference are based on checking the number of contiguous sequences (i.e., 'contigs') produced from the assembly, the contig length, the mean transcript length, the N50 value (i.e., the minimum contig length needed to cover 50% of the genome), the proportion of reads that can be mapped back to the assembled transcripts (Zhao et al., 2011), and the number of assembled transcripts that appear to be full-length (or nearly full-length) when compared to public databases such as the UniProtKB/Swiss-Prot (Grabherr et al., 2011).

### 1.2.2.2. Phylogenomics

After the genomic sequences have been obtained, by whole-genome sequencing, target capture or any other method, phylogenomic reconstructions traditionally involve extensive processing of the raw data. Most current phylogenomic projects share similar analytical designs that undergo minor modifications depending on the specific methodological choices, the type of data, and the hypotheses to test. These common steps are illustrated and described below (**Figure 5**).

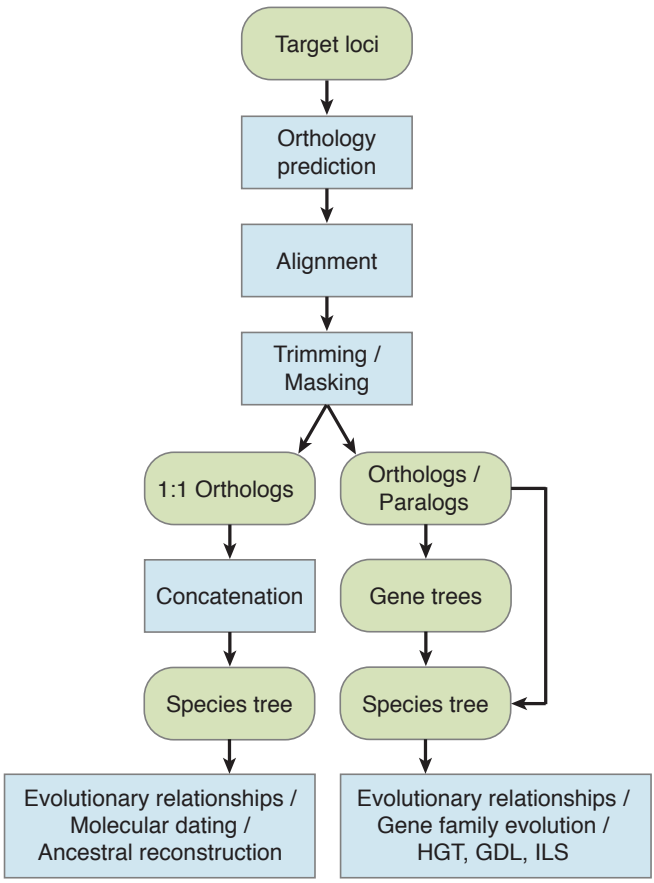


**Figure 5**. The phylogenomics flow chart. Data is represented by green rounded rectangles, and blue rectangles represent analyses. HGT: horizontal gene transfer, GDL: gene duplication and loss, ILS: incomplete lineage sorting.

The first essential step in the phylogenomic flow chart involves identification of homologous DNA sequences across taxa, e.g., by comparing the data to the probes used for target-enrichment or by matching it to pre-existing data from public databases (e.g., GenBank) using local alignment or profile-based search methods like BLAST (Altschul et al., 1990) and HMMER (Eddy, 1995). These data can be further filtered using an orthology prediction tool such as the program OrthoFinder (Emms & Kelly, 2015) to separate orthologous genes (i.e., homologs that evolved by speciation events) from paralogs (**Figure 6**). Current phylogenomic studies are usually restricted to 1-to-1 orthologs as their evolutionary history most likely reflects the evolution of the species.

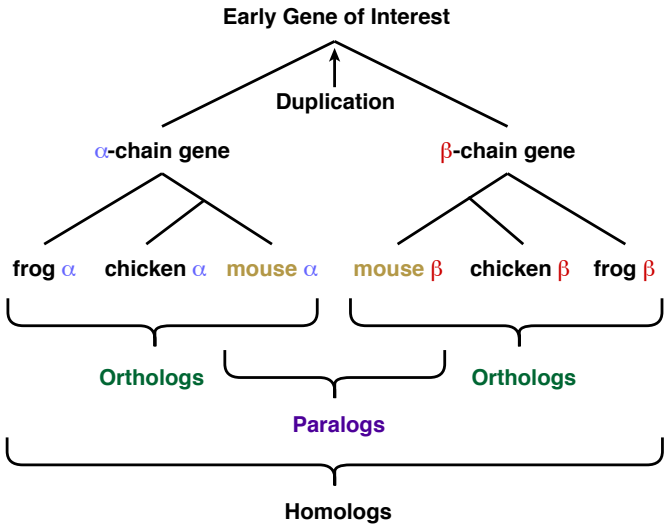


**Figure 6**. Diagram of homology subtypes showing orthologs and paralogs. A duplication event produces two copies of an early gene of interest (α and β). Two speciation events give rise to three species (frog, chicken, and mouse) that have both gene copies. Genes α and β are paralogs within species, but orthologs across species.

In the second step, orthologous genes are aligned at the nucleotide level (adding gaps within sequences that present insertions or deletions) using a multiple sequence alignment (MSA) method, such as MAFFT (Katoh et al., 2002) or PASTA (Mirarab et al., 2015). Alignment regions that present a large number of such gaps are arguably not reliable for phylogenetic analyses and they can be trimmed or masked using so-called trimming methods

(e.g., trimAl; Capella-Gutiérrez et al., 2009). The resulting MSA can next be translated and adjusted to the correct reading frame to identify true orthologs and separate them from pseudogenes. The quality of the resulting alignment can be assessed by summary statistics including, for example, the percent of missing data in the MSA, or the proportion of variable and parsimony-informative sites, which can be calculated using AMAS (Borowiec, 2016).

One important consideration before reconstructing gene and species trees is deciding upon appropriate thresholds for the inclusion of loci and taxa with missing data (Hosner et al., 2016). In addition to missing nucleotides derived from biological processes such as insertions and deletions, phylogenomic matrices with thousands of loci add another layer of complexity to the inference process due to: (1) the variable sequence yield among sample libraries leading to missing data within loci (**Figure 7a**); and (2) the stochasticity inherent in collecting data, where not all loci are detected in all taxa (**Figure 7b**).

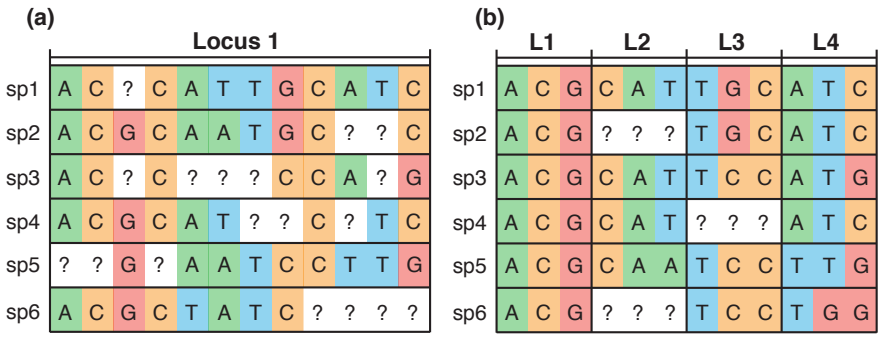


**Figure 7**. Multiple sequence alignments (MSA) of 6 species (sp1, sp2, sp3, sp4, sp5, and sp6) representing the different missing data types. (a) Missing data in a single-locus (Locus 1). (b) Incomplete taxon coverage in a concatenated MSA of 4 loci (L1, L2, L3, L4). Adapted from Mallo (2017).

As a result of MSAs with large amounts of missing data or with incomplete taxon coverage, sequences can become very distant from each other, and so the inferred trees may exhibit very long branches (Darriba et al., 2016). These very long branches due to missing data tend to erroneously group, and also with other long branches that could be correct (fast-evolving lineages), regardless of their true evolutionary relationships. This, which might poten-

tially compromise phylogenetic accuracy, is a phenomenon known as 'long-branch attraction' (Felsenstein, 1978; Wiens, 2006).

Inference methods to estimate phylogenies use molecular evolution models (i.e., substitution models) as part of their likelihood calculation (Arenas 2015). This step can be carried out for each MSA with programs such as jModelTest (Darriba et al., 2012). However, in phylogenomics, the most complex model, e.g., the generalized time-reversible (GTR) for DNA data, is often blindly chosen as overparameterization is less of a problem with large sequence datasets (Lemmon & Moriarty, 2004). Concatenated MSAs add even more complexity to the analyses since the alignments can be partitioned in sub-datasets with different substitution models or the same model but different parameters. Nonetheless, PartitionFinder (Lanfear et al., 2017) can be used to estimate the optimal partition scheme among all combinations of pre-specified data blocks and find the best-fit model for each of them. Along with MSAs, substitution models are the input data used to estimate evolutionary (gene and species) trees.

There is a wide variety of approaches and programs to infer gene trees (classical phylogenetic methods), and their description and comparison is not part of the research aims of this dissertation. However, the most popular programs for each strategy include: Neighbor-Joining (Saitou & Nei, 1987) and FastTree (Price et al., 2010) for distance-based clustering methods; PAUP (Swofford, 2002) for Maximum Parsimony (MP); RAxML (Stamatakis, 2014) for Maximum Likelihood (ML); and BEAST (Drummond et al., 2012) for Bayesian Inference (BI).

Methods for reconstructing species trees can be classified according to their input data, in three main categories: (1) Supermatrix or concatenation approaches, that depend on joining all single-locus MSAs into a multi-locus MSA and use a classical phylogenetic method (i.e., distance methods, MP, ML or BI) to infer the species tree; (2) Supertree approaches, consisting of two steps: (2a) estimating gene trees independently for each locus using classical phylogenetic methods, and (2b) combining the resulting gene trees into a single species tree summary methods such as ASTRAL (Mirarab et al., 2014) or ASTRID (Vachaspati & Warnow, 2015); and (3) full-data approaches, that directly estimate the species tree based on the sequence data, such as SVDquartets (Chifman & Kubatko, 2014).

Both species and gene trees constitute hypotheses of the evolutionary history of a sample of organisms or molecular sequences. These hardly ever constitute results by themselves, and thus they must be used as tools to answer scientific questions. Apart from providing insights into evolutionary relationships, phylogenomic research can also be used to (1) predict gene functions based on sequence similarity (Eisen & Fraser, 2003); (2) identify gene family evolution events such as gene duplication and loss (Hellmuth et al., 2015); or (3) discover horizontal gene transfers, hybridization events, and other forms of reticulated evolution (Whitaker et al., 2009). Regardless of the goal, understanding and keeping in mind the assumptions and known biases of the different methods, is crucial to any phylogenomic study. Performing several iterations of a phylogenomic pipeline (e.g., modifying certain parameters or methods) is basic to test the consistency of the results.

#### 1.2.2.3. Genome scans

An important goal in nearly all genomic studies about local adaptation is to identify loci related to adaptive phenotypes (Pardo-Diaz et al., 2015). When there is no prior knowledge regarding the genetic basis of the phenotype of interest, the challenging task of narrowing down genomic regions of interest is usually done using reverse genetic (or genome scan) approaches (Stinchcombe & Hoekstra, 2008). Unlike forward genetic approaches such as quantitative trait locus (QTL) mapping or GWAS that require information about phenotypes, these methods do not need quantifying phenotypic traits and are based on genome-wide screening of loci to detect footprints of selection (Storz, 2005).

Within a single population, genome scans of nucleotide diversity ($\pi$) (i.e., the average number of nucleotide differences per site for a group of DNA sequences), Tajima's $D$ (i.e., the difference between two sequences obtained comparing the mean number of pairwise differences with the number of segregating sites), and patterns of linkage disequilibrium (i.e., non-random association of alleles at different loci) are commonly used to detect signatures of selection. Between populations, the fixation index ($F_{ST}$) (i.e., a ratio of the average number of differences between pairs of chromosomes sampled within a population with the average number of sampled between different populations) is frequently used to identify loci showing genetic differ-

entiation. More specifically, by sampling a large number of single nucleotide polymorphisms (SNPs) throughout the genome, loci that have been affected by disruptive selection can be separated from neutral loci and identified as outliers in the $F_{ST}$ analysis (Martins et al., 2016). Therefore, it becomes possible to study both processes that affect the entire genome (genetic drift, gene flow, demographic changes, etc.) and adaptive processes that only affect outlier genes.

Other statistics developed more recently combine individual SNP frequency estimates with haplotype structure (e.g., XP-EHH; Sabeti et al., 2007), population history (e.g., PBS; Yi et al., 2010) or environmental variables (e.g., genotype-environment association analysis; Stucki et al., 2016) to quantify genetic differentiation and detect outlier loci between populations. The results of all these scans are most often presented as Manhattan plots that highlight genomic regions enriched for signals of selection (**Figure 8**).

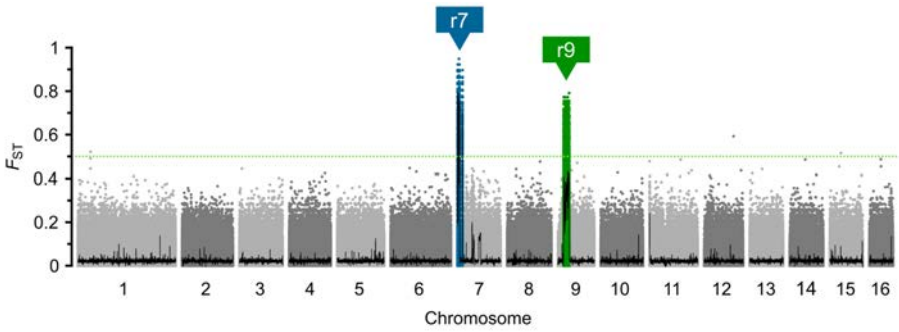


**Figure 8**. An illustration of a Manhattan plot based on allele frequency differences ($F_{ST}$) depicting two genomic regions (r7 and r9) highly divergent between highland and lowland populations of *Apis mellifera* in East Africa. Adapted from Wallberg et al. (2017).

## 1.3. Understanding the evolutionary history of natural populations with genomics

The spatial and temporal patterns of distribution of the genetic diversity of individuals, populations, and species are driven by several factors. Large-scale patterns are mainly determined by the biogeographical history of species, whereas finer-scale patterns are usually the result of ecological factors such as climate, resource availability, and competition (Wiens & Graham, 2005). Populations and species that are adapted to particular environmental conditions not only differ in their ecology but also show phenotypic and genetic differences compared to their congeners. Geographical barriers to gene flow such as mountains and rivers can also increase differentiation. However, we are yet to get a complete understanding of the impact of such barriers at the genetic level (Hewitt, 2001). With the advances in the field of genomics, it has become possible to characterize the genetic variation among organisms living at different habitats. This has given rise to further questions such as: (1) Is local adaptation caused by single genes or by multiple genes? Which genes are these? (2) What is the extent of gene flow during the speciation process? (3) Are rates and patterns of molecular evolution influenced by the environment?

In this thesis, we used genomic data from both genome-enabled and non-model organisms to study how mountains shaped adaptation and diversification in two different animal systems across the tree of life. In the first paper (**Paper I**), we examined the genetic differences between Eastern honey bees (*A. cerana*) populations inhabiting high-altitude and low-altitude environments in southwestern China. In the third paper (**Paper III**), we conducted phylogenomic analyses using genomic data derived from the transcriptome assembled at **Paper II** to explore the evolutionary history and diversification of *Oreobates* species collected from different altitudes in northwestern South America.

### 1.3.1. Honey bees as genome-enabled organisms

Honey bees or genus *Apis* (Insecta: Apidae) comprise a small group of 10 out of the roughly 20,000 known species of bees in the world (Michener, 2000; Arias & Sheppard, 2005). Honey bees are crucial to agriculture and food production worldwide due to their roles as pollinators (Aizen & Harder, 2009). Additionally, honey, beeswax, pollen, and other honey bee products have long been appreciated as foods and pharmaceuticals by humans. Ancient rock paintings in Bicorp (Spain) evidence that people have harvested honey for at least 7000–8000 years (Crane, 1999) (**Figure 9**).

**Figure 9**. The 'Man of Bicorp' holding onto lianas to gather honey from a beehive as depicted on an 8000-year-old cave painting near Valencia, Spain.

According to previous studies, honey bees probably originated in Asia, where 9 out of 10 species occur, and later expanded into Europe and Africa (Han et al., 2012). There is substantial phenotypic variation among honey bee species, including small dwarf bees (e.g., *Apis florea*) and large giant bees (e.g., *Apis dorsata*) that build open nests, as well as cavity-nesting intermediate-sized bees (Hepburn & Radloff, 2011). The two most important species for commercial beekeeping belong to the latter type: the Western or European honey bee (*Apis mellifera*) and the Eastern or Asiatic honey bee (*A. cerana*) (**Figure 10**). Both species inhabit extensive non-overlapping native ranges and have adapted to a wide range of environmental conditions

(Koetz, 2013). Humans have introduced Western honey bees to all continents except Antarctica, whereas the movements of Eastern honey bees have been less extensive but have led to consider it invasive species in areas such as Australia (Koetz, 2013).

**Figure 10**. *Apis cerana* worker collecting pollen on a red *Passiflora sp.* (Passifloraceae) in Yunnan, China. Photo credit: © Nicolas Vereecken.

Uncovering the genomic basis of local adaptation is a key goal in evolutionary genetics (Hoban et al., 2016). High-altitude populations are of particular interest as they are generally closely related to nearby lowland populations but possess unique phenotypic adaptations. Also, if populations from multiple highland areas exist, it is possible to examine whether evolution tends to follow the same path to adaptation, involving similar genetic variants, genes or pathways. Interestingly, highland individuals of both *A. mellifera* and *A. cerana* have distinct morphological traits compared to lowland conspecifics (Tan et al., 2003; Tan & Ling-juan, 2008; Gruber et al., 2013), which likely represent an adaptation for these habitats. Physiological and behavioral differences may also be involved in this local adaptation. Recent studies have shown that two chromosomal inversions govern adaptation to montane habitats in *A. mellifera* (Wallberg et al., 2017; Christmas et al., 2019) (see **Fig-**

**ure 8**). However, the genomic basis of adaptation to highland areas in *A. cerana* is unknown.

Populations of *A. cerana* have decreased in Asia since the early 20th century due to changes in local agriculture practices and the introduction of *A. mellifera* (Theisen-Jones & Bienefeld, 2016). Therefore, finding out how populations of this species are adapted to particular environmental conditions can help design conservation plans to manage their colonies given current and future challenges (Parker et al., 2010). Moreover, honey bees present relatively small genomes (~200 Mb), which makes them a suitable study system for population genomic studies. While biology and genomics of *A. mellifera* have been extensively studied since the sequencing of its genome in 2006 (Weinstock et al., 2006), our insights into *A. cerana* are comparatively reduced. Now that the draft genome of *A. cerana* is available (Park et al., 2015), there is a great opportunity to unravel the genomic basis of local adaptation to high-altitude habitats across different honey bee species.

### 1.3.2. Neotropical frogs as non-model organisms

The Neotropical region, which includes South America, Central America, the Caribbean islands, and southern North America, is a biodiversity hotspot that harbors the highest number of anuran species in the world (Fouquet et al., 2007). Sadly, it is also the region where amphibians and particularly frogs are suffering the most severe population declines and extinctions according to the IUCN (Scheele et al., 2019). Many factors such as habitat loss, altered climatic conditions, invasive species and, infectious diseases contribute to these declines (Wake & Vredenburg, 2008; O'Hanlon et al., 2018).

*Oreobates* (Jiménez de la Espada, 1872) (Anura: Craugastoridae) (**Figure 11**) is a genus of Neotropical frogs that at present consists of 25 species (Frost, 2019). Several species are threatened by habitat loss and some species are extremely rare, known only from one or two individuals in single localities (e.g., *Oreobates yanucu* and *Oreobates zongoensis*) (Köhler & Padial, 2016). A remarkable particularity of the genus *Oreobates* is that it is distributed across a wide range of habitats and altitudes in South America (**Figure 12**). Many species inhabit montane forests, cloud forests, and puna

grasslands of the Andes in Bolivia and Peru, others occur along the adjacent lowlands and some can be found even farther east in dry Atlantic environments (Padial & De la Riva, 2005; Padial et al., 2012; Teixeira et al., 2012). This provides an excellent opportunity to understand the implications of altitudinal gradients on the evolutionary history of this group of frogs.

**Figure 11**. *Oreobates quixensis* from the Amazon Cusco, Madre de Dios, Peru. Photo credit: © Ignacio De la Riva.

The number of species of *Oreobates* is likely underestimated as many areas of the Andes and the Amazon remain poorly surveyed (Köhler & Padial, 2016). Increasing fieldwork efforts combined with genetic, morphometric and bioacoustic tools, have contributed to the description of 7 new species since 2012 and 15 species in the last two decades (Vaz-Silva et al., 2018). Yet, their origin and diversification is unclear as previous phylogenetic studies were based on short fragments of mitochondrial and nuclear genes that recovered different relationships between some species (Padial et al., 2008, 2012, 2014; Hedges et al., 2008; Pyron & Wiens, 2011; Pereyra et al., 2014; Köhler & Padial, 2016; Jetz & Pyron, 2018; Vaz-Silva et al., 2018). Moreover, there are no common morphological synapomorphies in *Oreobates*

which makes studying their relationships even more challenging (Padial et al., 2012).

Phylogenetic inferences based on genomic data have the potential to resolve the evolutionary history of species. However, genomic data of amphibians are still relatively scarce compared to other vertebrates, due in part to their large and highly repetitive genomes that make the genome assembly challenging (Koepfli et al., 2015; Funk et al., 2018; Liedtke et al., 2018). Only eight amphibian species have reference genomes available to date: the model organisms *Xenopus tropicalis* and *Xenopus laevis* (Hellsten et al., 2010; Session et al., 2016), the Tibetan frog *Nanorana parkeri* (Sun et al., 2015), the American bullfrog *Rana catesbeiana* (Hammond et al., 2017), the African bullfrog *Pyxicephalus adspersus* (Denton et al., 2018), the cane toad *Rhinella marina* (Edwards et al., 2018), the strawberry dart-poison frog *Oophaga pumilio* (Rogers et al., 2018), and the axolotl *Ambystoma mexicanum* (Nowoshilow et al., 2018)

Given the lack of a reference genome for *Oreobates*, reduced representation sequencing approaches, such as sequence capture, can be successfully used to obtain large phylogenomic datasets. These approaches not only can help to understand the evolutionary history and diversification of *Oreobates* species along altitudinal gradients but also guide conservation priorities for endangered amphibians in the Neotropical region.

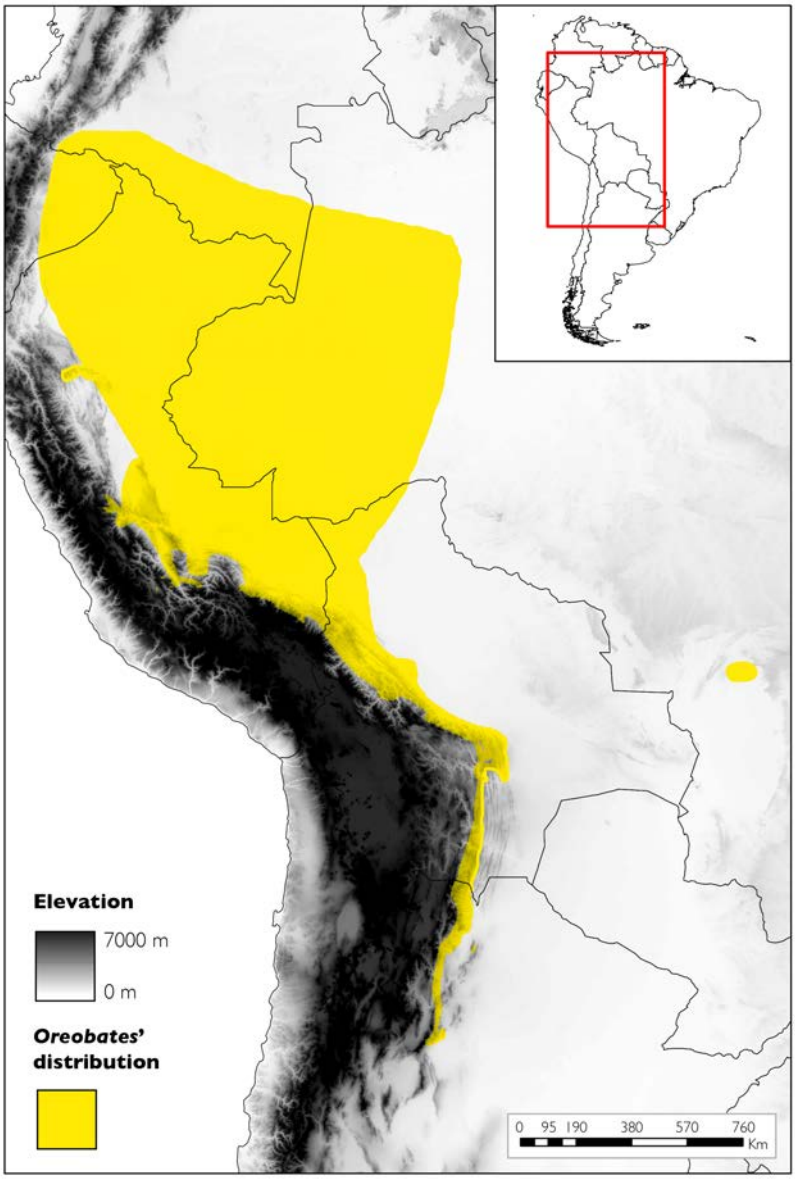


**Figure 12**. Approximate distribution of the genus *Oreobates* in South America as currently estimated by the International Union for Conservation of Nature (IUCN).

# 2. Research aims

## 2.1. General aims

The general aim of this dissertation was to use genomic approaches to (1) study the evolutionary history of species from mountain regions, and (2) understand the genetic basis of adaptive phenotypes in populations inhabiting high-altitude environments. To achieve these goals, my colleagues and I applied both whole-genome resequencing and reduced representation sequencing methods combined with comparative genomics and population genomic analyses. We characterized the genetic variation in two different animal systems to understand which evolutionary and ecological processes promote divergence along montane gradients.

## 2.2. Specific aims

- To investigate molecular variation between highland and lowland populations of the Eastern honey bee (*A. cerana*) and determine which regions in the genome are involved in the process of local adaptation to high-altitude habitats (**Paper I**).

- To characterize the transcriptome of the La Paz robber frog (*O. cruralis*) and identify and functionally annotate a large number of expressed genes for the development of molecular markers in other species of *Oreobates* and closely-related genera (**Paper II**).

- To study the phylogenetic relationships among frogs of the genus *Oreobates* using genomic data and clarify their evolution and diversification along altitudinal gradients (**Paper III**).

- To highlight the integration of genomics with other established methods in ecology and evolution (**Paper IV**).

# 3. Summaries of results

## 3.1. Paper I

**The genomic basis of adaptation to high-altitude habitats in the Eastern honey bee (*Apis cerana*)**

In this article, we explored the genomic differences between highland and lowland populations of *A. cerana* using a combination of statistical approaches to scan the genome based on relative divergence, haplotypes length, and relative branch length. We identified several genomic regions under positive selection that had high haplotype homozygosity specifically in highland bees and were strongly enriched for coding sequences. This indicates that these regions have been subjected to recent selection in high-altitude habitats. In contrast to previous findings in *A. mellifera*, we did not detect evidence of structural rearrangements between high and low altitude populations. This suggests that inversions do not always contribute to highland adaptation in honey bees and also that very different pathways could be involved in adapting to similar environments. Our analyses indicated that multiple loci likely contributed to high-altitude adaptation in populations of *A. cerana* from China.

Despite highland individuals of *A. cerana* being larger, darker and having longer body hair compared to lowland bees, we did not find genes with known functions in morphology or pigmentation separating highland and lowland populations. Such genes were not observed in *A. mellifera* either. Instead, we identified signals of selection in genes with potential function in feeding behavior and olfactory learning, indicating that behavioral differences may be more important for highland adaptation than morphological differences. Genes with similar functions were also identified in mountain populations of *A. mellifera*. The ability to find and memorize the location of food resources in high-altitude environments seems essential for highland honey bee populations.

Even though local adaptation to high-altitude environments in *A. cerana* has a different genomic basis compared to *A. mellifera*, and involves selection at multiple loci instead of structural rearrangements, it is possible that genes with similar functions are being selected in both species.

## 3.2. Paper II

### A practical guide to build de-novo assemblies for single tissues of non-model organisms: the example of a Neotropical frog

Previous genetic studies in the Neotropical frogs of the genus *Oreobates* were limited to few genes and genomic resources were lacking. An important step in facilitating genomic studies in non-model organisms is obtaining a reference genome. However, generating and assembling such references in amphibians is difficult due to their large and highly repetitive nature. Transcriptomes, which include exons and UTRs, only represent a small portion of the genome and thus are easier to assemble than whole genomes.

In this paper, we extracted and sequenced RNA from different tissues of an individual of the species *O. cruralis*. We assembled the resulting reads into a transcriptome, which included thousands of putative genes (i.e., 'unigenes'), and provided a functional characterization of the genome. Our study described in detail the procedure taken from RNA extraction to transcriptome assembly and annotation. Moreover, we presented a pipeline aimed to help inexperienced users to follow the same steps. The use of BLAST searches, whereby the resultant contigs produced from the assembly were matched to sequence data in the Swiss-Prot database, meant that functional annotations were obtained for a large proportion of the transcriptome.

Our results showed a large number of genes related to the immune system and defense mechanisms in the transcriptome of *O. cruralis*. We speculated that the presence of these genes could be associated with the elevated temperature and relative humidity of tropical rainforests, where amphibian pathogens are common. This study was essential to develop the next article of this dissertation (**Paper III**) in that it provided the reference genomic resources required for the targeted gene sequencing approach used in that work.

## 3.3. Paper III

**Phylogenomics and evolutionary history of Neotropical frogs of the genus *Oreobates* along altitudinal gradients**

In this manuscript, we studied the evolutionary history and diversification of the frogs of the genus *Oreobates* along altitudinal gradients using ~18,000 unigenes obtained from targeted capture sequencing.

By comparing phylogenetic trees and networks reconstructed from multiple phylogenomic datasets, we recovered two main clades with different distribution in the *Oreobates* phylogeny: one encompassing species from Argentina and Bolivia, which are found in lowland habitats (the Amazon Basin, dry forests, inter-Andean dry valleys, humid forests, and the yungas below 1,700 m); and another that comprises species from Colombia and Peru, which some inhabit lowlands but others inhabit high-altitude environments (cloud forests, elfin forests and puna grasslands up to 3,800 m).

Our phylogenomic examination revealed conflicting phylogenetic signals across different genomic regions (gene trees) in both highland and lowland species. However, most of the contrasting signals only affected intra-specific relationships. Among species from lowland habitats, taxonomic problems were observed between some species that could be due to misidentification or extensive gene flow. Although network analyses showed that secondary contact between some species has likely taken place across lowland taxa, evidence from other sources is needed to fully address these questions.

Finally, we found that highland species have smaller effective populations and thus stronger genetic drift compared to species sampled at lower altitudes. This results in weaker purifying selection and a more likely fixation of slightly deleterious mutations. Our data suggest that the small effective population size in highland species of *Oreobates* could promote the relative accumulation of functional divergence between lineages, thus facilitating speciation.

## 3.4. Paper IV

### Digest: Resolving phylogenomic conflicts in characiform fishes

This article is a short communication corresponding to the article by: Betancur-R.R., Arcila, D., Vari, R.P., Hughes, L.C., Oliveira, C., Sabaj, M.H., and Ortí, G. (2019), "Phylogenomic incongruence, hypothesis testing, and taxonomic sampling: The monophyly of characiform fishes" (*Evolution*. 73-2: 329–345).

The aim of this study was two-fold: first, improve accessibility of the research to evolutionary biologists outside of the original article's subfield of specialization; and second, provide added value by (1) placing the research in a broader context than described in the original paper; (2) adding figures illustrating the methods and findings; and (3) offering new insights on other implications of the research, as well as avenues of further study.

The original study applied a recently developed hypothesis-testing approach based on comparing top-ranking gene trees to resolve phylogenomic conflicts in characiform fishes. Even though massive parallel sequencing and genomic analyses are powerful new tools for understanding the evolutionary history of species, we argued that the best research is integrative, combining new genomic technologies with evidence from multiple sources such as field observations, controlled laboratory experiments, and modeling.

Although this article is not a full contribution, I included it in my dissertation because the idea of integrating genomics with other established methods fits nicely in the overall discussion of my research.

# 4. General discussion

A fundamental aspect in modern evolutionary biology refers to the study of how natural populations and species adapt and diversify across a wide range of environmental conditions. However, progress in such studies has traditionally been hampered by the absence of a genomic perspective and the lack of sufficient genetic markers that could allow assessing selective changes across many different regions of the genome. The technological innovations brought by the 'omics' revolution in the last two decades have enabled sequencing enormous amounts of DNA or RNA to address questions in new, creative, and more powerful ways than previously possible with traditional Sanger sequencing (Funk et al., 2018). In this thesis, my colleagues and I used the power of genomics to gain knowledge on the evolutionary history of natural populations as well as their adaptation and diversification.

Specifically, we studied the evolutionary and ecological processes that promoted divergence along altitudinal gradients in two different animal systems: Eastern honey bees (*A. cerana*) and frogs of the genus *Oreobates*, collected from high and low altitudes in southwestern China and northwestern South America, respectively. These two animal groups differ quite markedly not only in their life-history traits but also in the extent of available genetic resources for each of them. Eastern honey bees are genome-enabled, which offers a framework to assess genome-wide changes and to identify genes putatively relevant in adaptive diversification. We found evidence that natural selection has led to local adaptation in high elevation populations. By contrast, the frogs of the genus *Oreobates* lack of reference genomic resources, and although the capture approaches developed here allowed us to identify what processes have shaped the evolutionary history of species at different altitudes, it was difficult to reach a comparable level of resolution as with the honey bees. Nevertheless, the study of frogs suggested that the effect of genetic drift could be greater in highland species compared to species from lowland environments due to habitat fragmentation. Consequently, purifying selection could have been weaker and promoted the fixation of nonsynonymous mutations in highland species of *Oreo-*

*bates*. These mutations can be targeted by natural selection and contribute to the adaptation and differentiation of populations in mountain environments.

In **Paper I**, the availability of an annotated reference genome allowed us to characterize the genetic architecture for high-altitude adaptations in the Eastern honey bee. Our results on the genomic regions that show strong differentiation between highland and lowland populations of *A. cerana* were consistent with a scenario in which natural selection in montane environments has affected multiple loci across the whole genome. Candidate genes within these regions include some that have key roles in reproduction and foraging behavior in honey bees and other insects. These genes and the associated adaptive variants identified in our study may have important implications for managing Eastern honey bee colonies (Lozier & Zayed, 2017). For example, they can be of interest for marker-assisted breeding programs to select lineages with an increased foraging activity and more offspring per generation. As a result, colonies of *A. cerana* in Asia could become more productive and so more suitable for commercial beekeeping, offering a competitive advantage over imported colonies of European honey bee, *A. mellifera*. The latter species has been introduced in Asia since the early 20$^{th}$ century to meet higher production demands, but it is less well adapted to the local environment and more sensitive to pathogens than the Eastern honey bee (Theisen-Jones & Bienefeld, 2016).

Further work including functional experiments would be needed to fully address the relevance of these genes in honey bees and the effects of the variants under selection. Gene knockouts or knockdowns in a model organism such as *Drosophila melanogaster* could serve as a proxy to understand gene expression. Here, we could test whether outlier genes identified in **Paper I** display different expression levels through a series of environmental manipulations. This could help us understand if some populations are better at up-regulating or down-regulating certain genes in response to high-altitude environmental conditions such as lower temperatures, higher UV radiation, or lower partial pressure of oxygen.

On the other hand, the cost and complexity of generating a high quality, fully assembled and annotated genome is still prohibitive in most cases. The transcriptome that we assembled in **Paper II**, allowed obtaining genome-wide data via 'sequence capture' for the frogs of the genus *Oreobates* in the

absence of a reference genome at a relatively reduced cost. Although a transcriptome does not give any information on the location of genes in a genome, it does provide an abundance of data on which genes are being transcribed and, therefore, potentially of functional importance. Nonetheless, one potential pitfall of this procedure is that transcriptome annotation relied upon matching the assembled contigs to putatively orthologous genes of known function in species included in the Swiss-Prot database. Depending on the stringency of the similarity threshold required between sequences to perform that match, divergent orthologs may be missed (if the threshold is too strict) or paralogous genes recovered (if the threshold is too relaxed). This can result in the incorrect annotation of gene sequences. Also, any genes that do not have orthologs in other species will be unnoticed and so ignored, despite their potential functional importance. This implies a limitation on the results that could be obtained but, fortunately, has had no impact on the results presented in **Paper III** because we did not draw links between genotype and phenotype.

One way to filter out hidden paralogs from our initial set of ~18,000 unigenes, would have been to use the program BUSCO (Simão et al., 2015) to search for the presence or absence of conserved orthologs in the 'tetrapoda-odb9' database that represents a collection of ~4,000 single-copy Tetrapoda orthologs. This would have greatly reduced the number of initial unigenes as well as the experimental and computational costs of this project. Restricting our capture to these fewer genes would have reduced the effort required to later separate potentially paralog sequences, while at the same time would have allowed the sequence capture of more individuals with a similar cost. Although decreasing the number of markers could have caused a loss of phylogenetic resolution, our results also show that a reduced dataset consisting of just ~160 genes was able to reconstruct relatively similar phylogenetic relationships compared to those obtained with the entire data. Future phylogenomic studies in *Oreobates* could be restricted to those genes.

Despite the presence of potential paralogous genes, our phylogenomic analyses provided robust support for the phylogenetic relationships and evolutionary history of *Oreobates*. Similarly, previous studies have shown that genome-wide data sets are sufficient to generate fully resolved phylogenetic trees, even in the presence of horizontal gene transfer (Hellmuth et al., 2015). We detected phylogenetic discordances between species from low-

land habitats. The use of large phylogenomic datasets demonstrated that such discordances occur across many different genomic regions or gene trees, and thus incomplete lineage sorting can be discounted as an explanation. Instead, species trees and phylogenetic networks reconstructed using different approaches indicated that the conflicts among lowland species could be due to taxonomic misidentification and/or secondary contact between distinct evolutionary lineages. Consequently, our findings stress the need for a taxonomic reassessment to refine the phylogenetic status of some species of *Oreobates* that remain problematic, and also suggest that gene flow across divergent lineages may have contributed to the diversification of this genus.

The sampling efforts of this thesis were enough for the questions we set out to address, but the limited knowledge and number of samples available for *Oreobates* restricted the reach of the conclusions in **Paper III**. Here, a wider taxon sampling, including individuals from all currently recognized species, would have provided a more complete picture of the evolutionary history of this group of Neotropical frogs. However, that is not an easy task because some species of *Oreobates* have been discovered very recently (e.g., *O. antrum*; Vaz-Silva et al., 2018) and others are extremely difficult to find, such as *O. yanucu* and *O. zongoensis* which are known from single specimens only collected more than 15 years ago despite their type localities being relatively well surveyed (Köhler & Padial, 2016). Similarly, a wider population sampling, particularly for species encompassing large geographic areas, would be very important to evaluate the possible existence of cryptic species or to improve the understanding of the relationship between lowland taxa and their intraspecific diversity. Even though our study was based on specimens deposited at the Museo Nacional de Ciencias Naturales (MNCN-CSIC) in Madrid, which probably has the best collection in the world for these frogs due to the sampling efforts of our collaborators during the last two decades, more extensive sampling will benefit from future collaborations with other science museums.

Despite the limitations derived from the reduced sampling, we have shown that genomics can greatly expand our knowledge about natural populations by allowing us to characterize genetic variation across a much larger proportion of the genome compared to traditional phylogenetic methods. Importantly, some populational inferences have been possible using very few

individuals per population. This implies that robust inferences can be obtained minimizing the impact over natural populations. Given the power of these estimates, the emphasis for future research would be the careful selection of specimens that could represent the existing diversity without implying very large sample sizes per population. This means that genomic approaches will be particularly useful in combination with hypothesis-driven field surveys that do not necessarily imply large sample sizes but carefully selected sampling points. In this sense, genomics can help reduce the environmental impact of population genetic studies.

As scientists, we are sometimes captivated by the power of new technologies and tend to dismiss "old" scientific approaches. However, we must keep in mind that genomics represents just a single line of evidence to understand evolutionary and ecological processes determining biodiversity. As we briefly pointed out in **Paper IV**, research in evolutionary biology must be integrative, because any single line of evidence may not accurately reflect the evolutionary history and relationships of species (e.g., due to morphological convergence, gene tree/species tree discordance, etc.) (Padial et al., 2010; Schlick-Steiner et al., 2010; Derkarabetian & Hedin, 2014). Field observations could allow us solving the taxonomic problems seen among frogs of the genus *Oreobates* in **Paper III**; whereas our study of local adaptation in honey bees (**Paper I**) would require reciprocal transplant experiments to confirm whether the putatively adaptive differences identified through the genome scans actually result in greater fitness in the local environment.

Therefore, only by combining new genomic technology with other independent lines of evidence such as phenotypic, behavioral and ecological data, as well as controlled experiments and modeling, we will be able to fully understand how natural populations and species adapt and diversify across a wide range of environmental conditions.

# 5. Conclusions

- Genomics has enormous yet largely untapped potential to advance understanding of evolution, ecology, and behavior, as well as to improve the conservation of natural populations.

- Comparison of genome sequences in genome-enabled organisms allows the genomic characterization of natural populations and the detection of candidate loci involved in the process of local adaptation.

- Reduced representation sequencing strategies, such as the transcriptome-based exon capture approach used in this thesis, are an excellent alternative to whole-genome sequencing for non-model organisms lacking in genomic resources.

- Evolutionary processes promoting divergence along altitudinal gradients between populations of Eastern honey bee (*A. cerana*) are related to natural selection, whereas in frogs of the genus *Oreobates* the emphasis is on changes in the rate of evolution promoted by neutral demographic parameters.

- Several genomic regions show strong differentiation between highland and lowland populations of *A. cerana* from southwestern China. These regions are biased towards coding sequences and contain a higher proportion of nonsynonymous mutations compared to the rest of the genome. They also tend to have high haplotype homozygosity in the highland bees, indicating selective sweeps in these populations.

- Positive selection has led to local adaptation in populations of *A. cerana* inhabiting high-altitude habitats by targeting multiple loci across the whole genome. Candidate loci include genes related to reproduction and foraging behavior in honey bees and other insects, which may play a crucial role under the particular environmental conditions offered by mountain regions.

- The genus *Oreobates* has two geographically separated lineages distributed along altitudinal gradients in northwestern South America: one encompassing northern species found in lowland environments; and another that includes southern species. Within the latter lineage, a subsequent split separates a clade that includes mountain species. The colonization of lowland habitats from high altitudes is a difficult event but has independently taken place two times.

- Species of *Oreobates* inhabiting mountain environments have lower effective populations (i.e., lower heterozygosity) due to habitat fragmentation. This causes purifying selection to be less efficient at removing slightly deleterious mutations due to strong genetic drift, resulting in an increased ratio of $d_N$ (nonsynonymous substitution rate) to $d_S$ (synonymous substitution rate) compared to species sampled at lower altitudes. The accumulation of functional variation may have promoted differentiation of highland taxa.

- Large phylogenomic datasets have unveiled taxonomic problems between species of *Oreobates* from lowland habitats, and also show that gene flow across divergent lineages (i.e., secondary contact) may have contributed to the diversification of lowland taxa.

- Although high-throughput sequencing and genomic analyses are powerful new tools for understanding the biology of natural populations, the best research is integrative, combining new genomic technology with tried-and-true approaches such as field observations, controlled experiments, and modeling.

# 6. Acknowledgements

I carried out this dissertation at the Doñana Biological Station (EBD-CSIC), Seville, Spain, and at the Department of Medical Biochemistry and Microbiology (IMBIM) at Uppsala University, Uppsala, Sweden. In both places, many people have helped me to arrive at this beautiful, important and fulfilling moment that is to finish my Ph.D. Despite those who know me will probably agree that I am not very good at these sorts of things, I would like to use this opportunity to thank everyone I had the pleasure to meet and work with during the course of my doctoral studies. None of this work would have been possible without help from all of you.

To **Carles**, your trust in taking me on as a student without actually knowing me has changed my life and I am forever thankful to you. There are no enough words to explain how grateful I am to get an opportunity to work with you. You have been nothing but supportive throughout, providing invaluable advice, critique, and direction. You have also ensured the financial support was available for my project, as well as enabling me to attend conferences and workshops. Apart from being an outstanding scientist, you are also a great human being. I appreciate you always cared about my health and I sincerely thank you for letting me work at my own pace. Although you have a large number of personnel working with you on multiple projects, you always had the time to help me (even when you were on vacation with a poor internet connection). I think I could not have had a better supervisor than you. Thank you for all your support and I look forward to continuing to work with you into the future.

To **Jennifer**, I am very thankful for all the hours you invested in guiding the experimental part of my research. You have taught me a whole new set of genomic skills in the lab. Thanks for your unwavering help despite the innumerable trouble that my frogs have caused through these years. I also appreciate the use of English during lab meetings. My ability to express my thoughts in English has surely improved during the last years, and I think that is very important for my scientific career. Thank you (and **Carles**) for all the wonderful Christmas lunches and BBQs.

To **Matt W**, I am incredibly proud of having you as my co-supervisor. Thanks for your patience (particularly with the Spanish paperwork), endless knowledge, and for always providing really interesting and clever comments on my manuscripts. The research excellence and friendly environment in your group made me particularly looking forward to coming back to Uppsala every summer. Many thanks to **Anna**, **Andreas**, **Julia**, and **Matt C**, for being the best colleagues during my time in Sweden. Working with all of you has been inspirational, intellectually challenging and motivating at the same time. Beyond the office, I enjoyed a lot going to bowling, kart racing and BBQing with you. I hope we can do that again someday in the future.

To **Ignacio**, without your generosity and expertise in the frogs of the genus *Oreobates*, this thesis would have been impossible. Thanks not only for kindly sharing with me numerous samples of your exceptional collection (thanks also to **Bea** for getting them ready), but also thank you (and **Padial**) for traveling to Seville to be part of my Ph.D. committee and provide insightful discussions and suggestions.

To all the present members of the CONSEVOL group (**Álvaro**, **Andres**, **Anna**, **Arlo**, **Carlos**, **Iker**, **Inés**, **Isa**, **Sara**), it has been a pleasure working with you all. Particular thanks to **Andres**, **Inés**, **Isa**, and **Sara**, for all your feedback in early drafts of this thesis. Special mention to **Irene**, I am extremely grateful for your invaluable work in the lab. Many thanks also to **Arlo**, **Giovanni**, **Mar**, and **Miguel**, I appreciate you all convinced me several times to get off the computer and go out hiking.

Thank you to the Doñana Biological Station and the University of Seville, for providing me with the facilities, infrastructure, and support required for getting through a postgraduate degree. Thanks also to the Spanish Government (Ministerio de Economía y Competitividad) and the CSIC (Consejo Superior de Investigaciones Científicas) for funding my research in various ways. Many thanks to **José Manuel** for his tutorship and support with all the paperwork. I especially want to thank the technical support received from **Ana** (labwork), **Antonio Jesús** (administration), **Arturo** (genomic servers), **David** (GIS mapping), and **Jesús** (IT). Thanks to **Soraya** for keeping my workplace tidy every day. I am also in debt to "**J**", thank you for inviting me to play board games so many times.

Many thanks to Uppsala University and to the people I met in Uppsala, you have made these years diverse and constructive. Thanks to **Henrik** and all the staff at SciLifeLab, UPPMAX and sequencing platforms for being extremely helpful. Thanks to **Fan**, **Ginger**, **Iris**, **Jonas**, **Matteo**, and **Sangeet**, for being always so nice to me. Outside the work, I am particularly thankful to my friend **Xinzhu**, it was so nice hanging out with you. Thank you not only for sending me a postcard from the U.S. but most importantly for inspiring and encouraging me through these years. I also want to thank all the people who played "Badminton on Sundays" with me, it was good fun.

My thanks also to the numerous people that answered all my technical questions in Biostars, Github, Researchgate, etc. I also would like to thank all the instructors at the various workshops I attended in Český Krumlov, Rome, and Valencia. Many thanks to **Iker I**, **Lisa**, **Marina**, **Rosa**, **Tamara**, **Tomáš**, and **Toni**, for their feedback on phylogenomic aspects. My most sincere gratitude to **Daniel**, **Michael**, and **Valentina**, for caring about me and bringing me medicines when I got sick in one of the workshops.

Para mi familia, sé que estar a cientos de kilómetros de distancia no es fácil, pero saber que estáis ahí para lo que necesite me da mucha fuerza. Gracias por todo el amor incondicional y el apoyo que siempre me brindáis. Muchas gracias a mi **madre** y a mi **padre** no solo por enseñarme a respetar y amar la naturaleza, sino también por contribuir a que sea la persona que soy hoy en día. También agradecer a mis **hermanos** por su apoyo. Ya sabéis que no se me dan muy bien estas cosas, pero sin duda todos me habéis animado a que hiciera lo que más me gustaba en cada momento y por ello estoy hoy aquí acabando de escribir esta tesis doctoral. Así que os doy las gracias.

最后，最重要的是，如果没有李菡薏这一切不可能完成。我不知该如何形容你对我有多重要。感谢你每一次不辞辛劳、跨越距离来到我身边陪伴我，更要感谢你与我共享最困难、最幸福的时刻。我还想感谢你在博士论文期间给予我的耐心和巨大支持。我是如此幸运能遇到你，我也希望我们有一个幸福的未来。

# 7. References


Abascal, F., Corvelo, A., Cruz, F., Villanueva-Cañas, J.L., Vlasova, A., … Godoy, J.A. (2016). Extreme genomic erosion after recurrent demographic bottlenecks in the highly endangered Iberian lynx. *Genome Biology*, **17**, 251.

Aizen, M.A. & Harder, L.D. (2009). The global stock of domesticated honey bees is growing slower than agricultural demand for pollination. *Current Biology*, **19**, 915–918.

Allendorf, F.W., Hohenlohe, P.A., Luikart, G. (2010). Genomics and the future of conservation genetics. *Nature Reviews Genetics*, **11**, 697–709.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J. (1990). Basic local alignment search tool. *Journal of molecular biology*, **215**, 403–410.

Arenas, M. (2015). Trends in substitution models of molecular evolution. *Frontiers in Genetics*, **6**, 319.

Arias, M.C. & Sheppard, W.S. (2005). Phylogenetic relationships of honey bees (Hymenoptera: Apinae: Apini) inferred from nuclear and mitochondrial DNA sequence data. *Molecular Phylogenetics and Evolution*, **37**, 25–35.

Barton, N.H. (2000). Genetic hitchhiking. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **355**, 1553–1562.

Betancur-R, R., Arcila, D., Vari, R.P., Hughes, L.C., Oliveira, C., … Ortí, G. (2019). Phylogenomic incongruence, hypothesis testing, and taxonomic sampling: the monophyly of characiform fishes. *Evolution*, **73**, 329–345.

Bi, K., Vanderpool, D., Singhal, S., Linderoth, T., Moritz, C., Good, J.M. (2012). Transcriptome-based exon capture enables highly cost-effective comparative genomic data collection at moderate evolutionary scales. *BMC Genomics*, **13**, 403.

Bigham, A., Bauchet, M., Pinto, D., Mao, X., Akey, J.M., … Shriver, M.D. (2010). Identifying signatures of natural selection in Tibetan and Andean populations using dense genome scan data. *PLoS Genetics*, **6**, e1001116.

Black, W.C., Baer, C.F., Antolin, M.F., DuTeau, N.M. (2001). Population genomics: genome-wide sampling of insect populations. *Annual Review of Entomology*, **46**, 441–469.

Borowiec, M.L. (2016). AMAS: a fast tool for alignment manipulation and computing of summary statistics. *PeerJ*, **4**, e1660.

Bravo, G.A., Antonelli, A., Bacon, C.D., Bartoszek, K., Blom, M.P.K., … Edwards, S.V. (2019). Embracing heterogeneity: coalescing the Tree of Life and the future of phylogenomics. *PeerJ*, **7**, e6399.

Brown, C.T., Howe, A., Zhang, Q., Pyrkosz, A.B., Brom, T.H. (2012). A reference-free algorithm for computational normalization of shotgun sequencing data. *arXiv*. 2019; https://arxiv.org/abs/1203.4802

Capella-Gutiérrez, S., Silla-Martínez, J.M., Gabaldón, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, **25**, 1972–1973.

Castroviejo-Fisher, S. (2009). Species limits, and evolutionary history of glassfrogs. Uppsala University.

Chifman, J. & Kubatko, L. (2014). Quartet inference from SNP data under the coalescent model. *Bioinformatics*, **30**, 3317–3324.

Christmas, M.J., Wallberg, A., Bunikis, I., Olsson, A., Wallerman, O., … Webster, M.T. (2019). Chromosomal inversions associated with environmental adaptation in honeybees. *Molecular Ecology*, **28**, 1358–1374.

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., … Mortazavi, A. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biology*, **17**, 13–19.

Consortium (2004). Finishing the euchromatic sequence of the human genome. *Nature*, **431**, 931–945.

Crane, E. (1999). The World history of beekeeping and honey hunting. Taylor & Francis.

Cutter, A.D. & Payseur, B.A. (2013). Genomic signatures of selection at linked sites: unifying the disparity among species. *Nature Reviews Genetics*, **14**, 262–274.

Darriba, D., Taboada, G.L., Doallo, R., Posada, D. (2012). JModelTest 2: more models, new heuristics and parallel computing. *Nature Methods*, **9**, 772.

Darriba, D., Weiß, M., Stamatakis, A. (2016). Prediction of missing sequences and branch lengths in phylogenomic data. *Bioinformatics*, **32**, 1331–1337.

Darwin, C. (1859). On the origin of species by means of natural selection, or preservation of favoured races in the struggle for life. John Murray, London.

Davey, J.L. & Blaxter, M.W. (2010). RADseq: next-generation population genetics. *Briefings in Functional Genomics*, **9**, 416–423.

Degnan, J.H. & Rosenberg, N.A. (2006). Discordance of species trees with their most likely gene trees. *PLoS Genetics*, **2**, e68.

Delsuc, F., Brinkmann, H., Philippe, H. (2005). Phylogenomics and the reconstruction of the tree of life. *Nature Reviews Genetics*, **6**, 361–375.

Denton, R.D., Kudra, R.S., Malcom, J.W., Du Preez, L., Malone, J.H. (2018). The African Bullfrog (*Pyxicephalus adspersus*) genome unites the two ancestral ingredients for making vertebrate sex chromosomes. *bioRxiv*. 2018; https://doi.org/10.1101/329847.

Derkarabetian, S. & Hedin, M. (2014). Integrative taxonomy and species delimitation in harvestmen: a revision of the western North American genus Sclerobunus (Opiliones: Laniatores: Travunioidea). *PLoS ONE*, **9**, e104982.

Domazet-Lošo, T. & Tautz, D. (2010). A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns. *Nature*, **468**, 815–818.

Drummond, A.J., Suchard, M.A., Xie, D., Rambaut, A. (2012). Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Molecular Biology and Evolution*, **29**, 1969–1973.

Eddy, S.R. (1995). Multiple alignment using hidden Markov models. *Proceedings. International Conference on Intelligent Systems for Molecular Biology*, **3**, 114–120.

Edmands, S. (2007). Between a rock and a hard place: evaluating the relative risks of inbreeding and outbreeding for conservation and management. *Molecular Ecology*, **16**, 463–475.

Edwards, R.J., Tuipulotu, D.E., Amos, T.G., O'Meally, D., Richardson, M.F., … White, P.A. (2018). Draft genome assembly of the invasive cane toad, *Rhinella marina*. *GigaScience*, **7**, 1–13.

Eisen, J.A. & Fraser, C.M. (2003). Phylogenomics: intersection of evolution and genomics. *Science*, **300**, 1706–1707.

Ekblom, R. & Wolf, J.B.W. (2014). A field guide to whole-genome sequencing, assembly and annotation. *Evolutionary Applications*, **7**, 1026–1042.

Elshire, R.J., Glaubitz, J.C., Sun, Q., Poland, J.A., Kawamoto, K., … Mitchell, S.E. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE*, **6**, e19379.

Emms, D.M. & Kelly, S. (2015). OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology*, **16**, 157.

Escoda-Assens, L. (2018). Applications of next-generation sequencing in conservation genomics: kinship analysis and dispersal patterns. Universitat de Barcelona.

Faircloth, B.C., McCormack, J.E., Crawford, .N.G., Harvey, M.G., Brumfield, R.T., Glenn, T.C. (2012). Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. *Systematic Biology*, **61**, 717–726.

Felsenstein, J. (1978). Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Biology*, **27**, 401–410.

Fouquet, A., Gilles, A., Vences, M., Marty, C., Blanc, M., Gemmell, N.J. (2007). Underestimation of species richness in Neotropical frogs revealed by mtDNA analyses. *PLoS ONE*, **2**, e1109.

Frankham, R., Ballou, J.D., Eldridge, M.D., Lacy, R.C., Ralls, K., … Fenster, C.B. (2011). Predicting the probability of outbreeding depression. *Conservation Biology*, **25**, 465–475.

Frost, D.R. (2019). Amphibian Species of the World: an online reference. Version 6.0 Accessed on 22–08–2019. Electronic database accessible at http://research.amnh.org/herpetology/amphibia/index.html. American Museum of Natural History, New York, USA.

Funk, W.C., Zamudio, K.R., Crawford, A.J. (2018). Advancing understanding of amphibian evolution, ecology, behavior, and conservation with massively parallel sequencing. *Population Genomics. Springer, Cham*, 1–44.

Goodman, M., Czelusniak, J., Moore, G.W., Romero-Herrera, A.E., Matsuda, G. (1979). Fitting the gene lineage into its species lineage, a parsimony strategy illustrated by cladograms constructed from globin sequences. *Systematic Zoology*, **28**, 132–163.

Goodwin, S., McPherson, J.D., McCombie, W.R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, **17**, 333–351.

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., … Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, **29**, 644–652.

Gregory, T.R., Nicol, J.A., Tamm, H., Kullman, B., Kullman, K., … Bennett, M.D. (2007). Eukaryotic genome size databases. *Nucleic Acids Research*, **35**, D332–8.

Gruber, K., Schöning, C., Otte, M., Kinuthia, W., Hasselmann, M. (2013). Distinct subspecies or phenotypic plasticity? Genetic and morphological differentiation of mountain honey bees in East Africa. *Ecology and Evolution*, **3**, 3204–3218.

Hammond, S.A., Warren, R.L., Vandervalk, B.P., Kucuk, E., Khan, H., … Birol, I. (2017). The North American bullfrog draft genome provides insight into hormonal regulation of long noncoding RNA. *Nature Communications*, **8**, 1433.

Han, F., Wallberg, A., Webster, M.T. (2012). From where did the western honeybee (*Apis mellifera*) originate? *Ecology and Evolution*, **2**, 1949–1957.

Hedges, S.B., Duellman, W.E., Heinicke, M.P. (2008). New World direct-developing frogs (Anura: Terrarana): Molecular phylogeny, classification, biogeography, and conservation. *Zootaxa*, **1737**, 1–182.

Hellmuth, M., Wieseke, N., Lechner, M., Lenhof, H., Middendorf, M., Stadler, P.F. (2015). Phylogenomics with paralogs. *Proceedings of the National Academy of Sciences of the United States of America*, **112**, 2058–2063.

Hellsten, U., Harland, R.M., Gilchrist, M.J., Hendrix, D., Jurka, J., … Rokhsar, D.S. (2010). The genome of the Western clawed frog *Xenopus tropicalis*. *Science*, **328**, 633–636.

Hepburn, H.R. & Radloff, S.E. (2011). Honeybees of Asia. Springer-Verlag Berlin Heidelberg.

Hewitt, G.M. (2001). Speciation, hybrid zones and phylogeography - Or seeing genes in space and time. *Molecular Ecology*, **10**, 537–549.

Hirsch, C.D., Evans, J., Buell, C.R., Hirsch, C.N. (2014). Reduced representation approaches to interrogate genome diversity in large repetitive plant genomes. *Briefings in Functional Genomics and Proteomics*, **13**, 257–267.

Hoban, S., Kelley, J.L., Lotterhos, K.E., Antolin, M.F., Bradburd, G., … Whitlock, M.C. (2016). Finding the genomic basis of local adaptation: pitfalls, practical solutions, and future directions. *The American Naturalist*, **188**, 379–397.

Hosner, P.A., Faircloth, B.C., Glenn, T.C., Braun, E.L., Kimball, R.T. (2016). Avoiding missing data biases in phylogenomic inference: an empirical study in the landfowl (Aves: Galliformes). *Molecular Biology and Evolution*, **33**, 1110–1125.

Huerta-Sánchez, E., Degiorgio, M., Pagani, L., Tarekegn, A., Ekong, R., … Nielsen, R. (2013). Genetic signatures reveal high-altitude adaptation in a set of ethiopian populations. *Molecular biology and evolution*, **30**, 1877–1888.

Jain, M., Koren, S., Miga, K.H., Quick, J., Rand, A.C., … Loose, M. (2018). Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nature Biotechnology*, **36**, 338–345.

Jarvis, E.D., Mirarab, S., Aberer, A.J., Li, B., Houde, P., … Zhang, G. (2014). Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science*, **346**, 1320–1331.

Jetz, W. & Pyron, R.A. (2018). The interplay of past diversification and evolutionary isolation with present imperilment across the amphibian tree of life. *Nature Ecology and Evolution*, **2**, 850–858.

Katoh, K., Misawa, K., Kuma, K., Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic acids research*, **30**, 3059–3066.

Koepfli, K.-P., Paten, B., O'Brien, S.J. (2015). The genome 10K project: a way forward. *Annual Review of Animal Biosciences*, **30**, 3059–3066.

Koetz, A. (2013). Ecology, behaviour and control of *Apis cerana* with a focus on relevance to the Australian incursion. *Insects*, **4**, 558–592.

Köhler, J. & Padial, J.M. (2016). Description and phylogenetic position of a new (singleton) species of *Oreobates* Jiménez De La Espada, 1872 (Anura: Craugastoridae) from the yungas of Cochabamba, Bolivia. *Annals of Carnegie Museum*, **84**, 23–38.

Lamichhaney, S., Berglund, J., Almén, M.S., Maqbool, K., Grabherr, M., … Andersson, L. (2015). Evolution of Darwin's finches and their beaks revealed by genome sequencing. *Nature*, **518**, 371–375.

Lamichhaney, S., Han, F., Berglund, J., Wang, C., Almén, M.S., … Andersson, L. (2016). A beak size locus in Darwin´s finches facilitates character displacement during a drought. *Science*, **352**, 470–474.

Lanfear, R., Frandsen, P.B., Wright, A.M., Senfeld, T., Calcott, B. (2017). Partitionfinder 2: new methods for selecting partitioned models of evolution for molecular and morphological phylogenetic analyses. *Molecular Biology and Evolution*, **34**, 772–773.

Lemmon, A.R., Emme, S.A., Lemmon, E.M. (2012). Anchored hybrid enrichment for massively high-throughput phylogenomics. *Systematic Biology*, **61**, 727–744.

Lemmon, A.R. & Moriarty, E.C. (2004). The importance of proper model assumption in bayesian phylogenetics. *Systematic Biology*, **53**, 265–277.

Liedtke, H.C., Gower, D.J., Wilkinson, M., Gomez-Mestre, I. (2018). Macroevolutionary shift in the size of amphibian genomes and the role of life history and climate. *Nature Ecology & Evolution*, **2**, 1792–1799.

Lozier, J.D. & Zayed, A. (2017). Bee conservation in the age of genomics. *Conservation Genetics*, **18**, 713–729.

Luikart, G., England, P.R., Tallmon, D., Jordan, S., Taberlet, P. (2003). The power and promise of population genomics: From genotyping to genome typing. *Nature Reviews Genetics*, **4**, 981–994.

Maddison, W.P. (1997). Gene trees in species trees. *Systematic Biology*, **46**, 523–536.

Mallo, D. (2017). Evaluation of phylogenomic methods for species tree estimation. University of Vigo.

Mallo, D. & Posada, D. (2016). Multilocus inference of species trees and DNA barcoding. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **371**, 20150335.

Mamanova, L., Coffey, A.J., Scott, C.E., Kozarewa, I., Turner, E.H., … Turner, D.J. (2010). Target-enrichment strategies for next-generation sequencing. *Nature Methods*, **7**, 111–118.

Mardis, E.R. (2008). The impact of next-generation sequencing technology on genetics. *Trends in Genetics*, **24**, 133–141.

Mardis, E.R. (2011). A decade's perspective on DNA sequencing technology. *Nature*, **470**, 198–203.

Marin, J., Achaz, G., Crombach, A., Lambert, A. (2019). The genomic view of diversification. *bioRxiv*. 2018; https://doi.org/10.1101/413427

Martins, H., Caye, K., Luu, K., Blum, M.G.B., François, O. (2016). Identifying outlier loci in admixed and in continuous populations using ancestral population differentiation statistics. *Molecular ecology*, **25**, 5029–5042.

Melé, M., Ferreira, P.G., Reverter, F., DeLuca, D.S., Monlong, J., … Guigó, R. (2015). Human genomics. The human transcriptome across tissues and individuals. *Science*, **348**, 660–665.

Mendel, G. (1866). Versuche über Pflanzenhybriden. Verhandlungen des naturforschenden Vereines in Brünn. *Journal of the Royal Horticultural Society*, **4**, 3–47.

Michener, C.D. (2000). The bees of the world. Johns Hopkins Press, Baltimore, Maryland, USA.

Mirarab, S., Nguyen, N., Guo, S., Wang, L.S., Kim, J., Warnow, T. (2015). PASTA: ultra-large multiple sequence alignment for nucleotide and amino-acid sequences. *Journal of Computational Biology*, **22**, 377–386.

Mirarab, S., Reaz, R., Bayzid, M.S., Zimmermann, T., Swenson, M.S., Warnow, T. (2014). ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics*, **30**, 541–548.

Moreton, J., Izquierdo, A., Emes, R.D. (2016). Assembly, assessment, and availability of *de novo* generated eukaryotic transcriptomes. *Frontiers in Genetics*, **6**, 361.

Nowoshilow, S., Schloissnig, S., Fei, J.F., Dahl, A., Pang, A.W.C., … Myers, E.W. (2018). The axolotl genome and the evolution of key tissue formation regulators. *Nature*, **554**, 50–55.

O'Hanlon, S.J., Rieux, A., Farrer, R.A., Rosa, G.M., Waldman, B., … Fisher, M.C. (2018). Recent Asian origin of chytrid fungi causing global amphibian declines. *Science*, **360**, 621–627.

Padial, J.M., Chaparro, J.C., Castroviejo-Fisher, S., Guayasamin, J.M., Lehr, E., … De la Riva I. (2012). A revision of species diversity in the Neotropical genus *Oreobates* (Anura: Strabomantidae), with the description of three new species from the Amazonian slopes of the Andes. *American Museum Novitates*, **3752**, 1–55.

Padial, J.M., Chaparro, J.C., De la Riva, I. (2008). Systematics of *Oreobates* and the Eleutherodactylusdiscoidalis species group (Amphibia, Anura), based on two mitochondrial DNA genes and external morphology. *Zoological Journal of the Linnean Society*, **152**, 737–773.

Padial, J.M., Grant, T., Frost, D.R. (2014). Molecular systematics of terraranas (Anura: Brachycephaloidea) with an assessment of the effects of alignment and optimality criteria. *Zootaxa*, **3825**, 1–132.

Padial, J.M. & De la Riva, I. (2005). Rediscovery, redescription, and advertisement call of *Eleutherodactylus heterodactylus* (Miranda Ribeiro, 1937). (Anura: Leptodactylidae), and notes on other *Eleutherodactylus*. *Journal of Herpetology*, **39**, 372–379.

Padial, J.M., Miralles, A., De la Riva, I., Vences, M. (2010). The integrative future of taxonomy. *Frontiers in Zoology*, **7**, 16.

Pamilo, P., Nei, M. (1988). Relationships between gene trees and species trees. *Molecular biology and evolution*, **5**, 568–583.

Pardo-Diaz, C., Salazar, C., Jiggins, C.D. (2015). Towards the identification of the loci of adaptive evolution. *Methods in Ecology and Evolution*, **6**, 445–464.

Park, D., Jung, J.W., Choi, B.S., Jayakodi, M., Lee, J., … Kwon, H.W. (2015). Uncovering the novel characteristics of Asian honey bee, *Apis cerana*, by whole genome sequencing. *BMC genomics*, **16**, 1.

Parker, R., Melathopoulos, A.P., White, R., Pernal, S.F., Guarna, M.M., Foster, L.J. (2010). Ecological adaptation of diverse honey bee (*Apis mellifera*) populations. *PLoS ONE*, **5**, e11096.

Peng, Y., Yang, Z., Zhang, H., Cui, C., Qi, X., … Su, B. (2011). Genetic variations in Tibetan populations and high-altitude adaptation at the Himalayas. *Molecular Biology and Evolution*, **28**, 1075–1081.

Pereyra, M.M.O., Cardozo, D.E. D.E., Baldo, J., Baldo, D. (2014). Description and phylogenetic position of a new species of *Oreobates* (Anura: Craugastoridae) from Northwestern Argentina. *Herpetologica*, **70**, 211–227.

Price, M.N., Dehal, P.S., Arkin, A.P. (2010). FastTree 2 - Approximately maximum-likelihood trees for large alignments. *PLoS ONE*, **5**, e9490.

Pyron, R.A. & Wiens, J.J. (2011). A large-scale phylogeny of Amphibia including over 2800 species, and a revised classification of extant frogs, salamanders, and caecilians. *Molecular Phylogenetics and Evolution*, **61**, 543–583.

Rannala, B. & Yang, Z. (2008). Phylogenetic Inference Using Whole Genomes. *Annual Review of Genomics and Human Genetics*, **9**, 217–231.

Rogers, J. & Gibbs, R.A. (2014). Comparative primate genomics: Emerging patterns of genome content and dynamics. *Nature Reviews Genetics*, **15**, 347–359.

Rogers, R.L., Zhou, L., Chu, C., Márquez, R., Corl, A., … Nielsen, R. (2018). Genomic takeover by transposable elements in the strawberry poison frog. *Molecular Biology and Evolution*, **35**, 2913–2927.

Rokas, A. & Abbot, P. (2009). Harnessing genomics for evolutionary insights. *Trends in Ecology and Evolution*, **24**, 192–200.

Ruane, S., Raxworthy, C.J., Lemmon, A.R., Lemmon, E.M., Burbrink, F.T. (2015). Comparing species tree estimation with large anchored phylogenomic and small Sanger-sequenced molecular datasets: an empirical study on Malagasy pseudoxyrhophiine snakes. *BMC Evolutionary Biology*, **15**, 221.

Sabeti, P.C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., … Stewart, J. (2007). Genome-wide detection and characterization of positive selection in human populations. *Nature*, **449**, 913–918.

Saitou, N. & Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution*, **4**, 406–425.

Sambrook, J., Fritsch, E.F., Maniatis, T. (1989). Molecular cloning: a laboratory manual. Cold Spring Harbor Laboratory Press.
Savolainen, O., Lascoux, M., Merilä, J. (2013). Ecological genomics of local adaptation. *Nature Reviews Genetics*, **14**, 807–820.
Scheele, B.C., Pasmans, F., Skerratt, L.F., Berger, L., Martel, A., … Canessa, S. (2019). Amphibian fungal panzootic causes catastrophic and ongoing loss of biodiversity. *Science*, **363**, 1459–1463.
Schlick-Steiner, B.C., Steiner, F.M., Seifert, B., Stauffer, C., Christian, E., Crozier, R.H. (2010). Integrative taxonomy: a multisource approach to exploring biodiversity. *Annual Review of Entomology*, **55**, 421–438.
Seehausen, O., Butlin, R.K., Keller, I., Wagner, C.E., Boughman, J.W., … Widmer, A. (2014). Genomics and the origin of species. *Nature Reviews Genetics*, **15**, 176–192.
Session, A.M., Uno, Y., Kwon, T., Chapman, J.A., Toyoda, A., … Rokhsar, D.S. (2016). Genome evolution in the allotetraploid frog *Xenopus laevis*. *Nature*, **538**, 336–343.
Shen, X.X., Liang, D., Feng, Y.J., Chen, M.Y., Zhang, P. (2013). A versatile and highly efficient toolkit including 102 nuclear markers for vertebrate phylogenomics, tested by resolving the higher level relationships of the caudata. *Molecular Biology and Evolution*, **30**, 2235–2248.
Sibbesen, J.A. (2016). Probabilistic transcriptome assembly and variant graph genotyping. University of Copenhagen.
Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., Zdobnov, E.M. (2015). BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, **31**, 3210–3212.
Smith, J.M. & Haigh, J. (1974). The hitch-hiking effect of a favourable gene. *Genetical Research*, **23**, 23–35.
Stamatakis, A. (2014). RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, **30**, 1312–1313.
Staubach, F., Lorenc, A., Messer, P.W., Tang, K., Petrov, D.A., Tautz, D. (2012). Genome patterns of selection and introgression of haplotypes in natural populations of the house mouse (*Mus musculus*). *PLoS Genetics*, **8**, e1002891.
Steiner, C.C., Putnam, A.S., Hoeck, P.E.A., Ryder, O.A. (2013). Conservation genomics of threatened animal species. *Annual Review of Animal Biosciences*, **1**, 261–281.
Stinchcombe, J.R. & Hoekstra, H.E. (2008). Combining population genomics and quantitative genetics: Finding the genes underlying ecologically important traits. *Heredity*, **100**, 158–170.
Storz, J.F. (2005). Using genome scans of DNA polymorphism to infer adaptive population divergence. *Molecular Ecology*, **14**, 671–688.

Stucki, S., Orozco-terWengel, P., Forester, B.R., Duruz, S., Colli, L., … Joost, S. (2016). High performance computation of landscape genomic models including local indicators of spatial association. *Molecular Ecology Resources*, **17**, 1072–1089.

Sun, Y.B., Xiong, Z.J., Xiang, X.Y., Liu, S.P., Zhou, W.W., … Zhang, Y.P. (2015). Whole-genome sequence of the Tibetan frog *Nanorana parkeri* and the comparative evolution of tetrapod genomes. *Proceedings of the National Academy of Sciences*, **112**, E1257–E1262.

Swofford, D.L. (2002). PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4. Sinauer Associates.

Takahata, N. (1989). Gene genealogy in three related populations: consistency probability between gene and population trees. *Genetics*, **122**, 957–66.

Tan, K., Fuchs, S., Koeniger, N., Ruiguang, Z. (2003). Morphological characterization of *Apis cerana* in the Yunnan province of China. *Apidologie*, **34**, 553–561.

Tan, K. & Ling-juan, L. (2008). Morphology and taxonomy of *Apis cerana* in Deqin. *Journal of Yunnan Agricultural University*, **23**, 230–232.

Teixeira, M., Amaro, R.C., Recoder, R.S., De Sena, M.A., Rodrigues, M.T. (2012). A relict new species of *Oreobates* (Anura, Strabomantidae) from the seasonally dry tropical forests of Minas Gerais, Brazil, and its implication to the biogeography of the genus and that of South American dry forests. *Zootaxa*, **3158**, 37–52.

Theisen-Jones, H. & Bienefeld, K. (2016). The Asian honey bee (*Apis cerana*) is significantly in decline. *Bee World*, **93**, 90–97.

Turner, E.H., Ng, S.B., Nickerson, D.A., Shendure, J. (2009). Methods for Genomic Partitioning. *Annual Review of Genomics and Human Genetics*, **10**, 263–284.

Vachaspati, P. & Warnow, T. (2015). ASTRID: Accurate species TRees from internode distances. *BMC Genomics*, **16**, S3.

Vaz-Silva, W., Maciel, N.M., De Andrade, S.P., Amaro, R.C. (2018). A new cryptic species of *Oreobates* (Anura: Craugastoridae) from the seasonally dry tropical forest of central Brazil. *Zootaxa*, **4441**, 089–108.

Vijay, N., Poelstra, J.W., Künstner, A., Wolf, J.B.W. (2013). Challenges and strategies in transcriptome assembly and differential gene expression quantification. A comprehensive in silico assessment of RNA-seq experiments. *Molecular Ecology*, **22**, 620–634.

Wake, D.B. & Vredenburg, V.T. (2008). Are we in the midst of the sixth mass extinction? A view from the world of amphibians. *Proceedings of the National Academy of Sciences*, **105**, 11466–11473.

Wallberg, A., Bunikis, I., Pettersson, O.V., Mosbech, M.B., Childers, A.K., … Webster, M.T. (2019). A hybrid de novo genome assembly of

the honeybee, *Apis mellifera*, with chromosome-length scaffolds. *BMC Genomics*, **20**, 275.
Wallberg, A., Glémin, S., Webster, M.T. (2015). Extreme recombination frequencies shape genome variation and evolution in the honeybee, *Apis mellifera*. *PLoS Genetics*, **11**, 1–27.
Wallberg, A., Schöning, C., Webster, M.T., Hasselmann, M. (2017). Two extended haplotype blocks are associated with adaptation to high altitude habitats in East African honey bees. *PLoS Genetics*, **13**, 1–30.
Wang, Z., Gerstein, M., Snyder, M. (2009). RNA-seq: a revolutionary tool for transcriptomics. *Nature reviews. Genetics*, **10**, 57–63.
Weinstock, G.M., Robinson, G.E., Gibbs, R.A., Worley, K.C., Evans, J.D., … Rita, W. (2006). Insights into social insects from the genome of the honeybee *Apis mellifera*. *Nature*, **443**, 931–949.
Whitaker, J.W., McConkey, G.A., Westhead, D.R. (2009). The transferome of metabolic genes explored: Analysis of the horizontal transfer of enzyme encoding genes in unicellular eukaryotes. *Genome Biology*, **10**, R36.
Wiens, J.J. (2006). Missing data and the design of phylogenetic analyses. *Journal of Biomedical Informatics*, **39**, 34–42.
Wiens, J.J. & Graham, C.H. (2005). Niche conservatism: Integrating evolution, ecology, and conservation biology. *Annual Review of Ecology, Evolution, and Systematics*, **36**, 519–539.
Xia, X. (2011). Comparative Genomics. *Handbook of Statistical Bioinformatics*, Springer Handbooks of Computational Statistics, 567–600.
Xu, S., Li, S., Yang, Y., Tan, J., Lou, H., … Jin, L. (2011). A genome-wide search for signals of high-altitude adaptation in Tibetans. *Molecular Biology and Evolution*, **28**, 1003–1011.
Yi, X., Liang, Y., Huerta-Sanchez, E., Jin, X., Cuo, Z.X., … Wang, J. (2010). Sequencing of 50 human exomes reveals adaptation to high altitude. *Science,* **329**, 75–78.
Zhang, F., Ding, Y., Zhu, C. (2019). Phylogenomics from low-coverage whole-genome sequencing. *Methods in Ecology and Evolution*, **10**, 507–517.
Zhao, Q.Y., Wang, Y., Kong, Y.M., Luo, D., Li, X., Hao, P. (2011). Optimizing *de novo* transcriptome assembly from short-read RNA-Seq data: a comparative study. *BMC bioinformatics,* **12**, S2.