Skoltech

POSITIVE SELECTION IN PARALLEL EVOLUTION

Doctoral Thesis

by

VALENTINA BURSKAIA

Skoltech

Skolkovo Institute of Science and Technology

POSITIVE SELECTION IN PARALLEL EVOLUTION

Doctoral Thesis

by

VALENTINA BURSKAIA

DOCTORAL PROGRAM IN LIFE SCIENCES

Supervisor

Professor Georgii Bazykin

Moscow - 2020

I hereby declare that the work presented in this thesis was carried out by myself at Skolkovo Institute of Science and Technology, Moscow, except where due acknowledgement is made, and has not been submitted for any other degree.

Candidate (Valentina Burskaia)

Supervisor (Prof. Georgii Bazykin)

Abstract

Repeated emergence of similar adaptations is often explained by parallel evolution of underlying genes. Single nucleotide or amino acid substitutions can underlie similar phenotypic traits. Yet when we track simple independent events, it is difficult to distinguish random coincidences (resulting from neutral evolution) from those caused by positive selection. In this work I apply computational methods to investigate parallel molecular evolution in the wildlife and to find cases associated with adaptation.

Firstly, I report a striking pattern of elevated rate of parallel nonsynonymous evolution in amphipods. All 46 species included in the study demonstrate higher rates of parallel nonsynonymous evolution than parallel synonymous evolution. This phenomenon is detectable at the whole-genome level and implies widespread adaptive parallel molecular evolution in this young species flock.

Secondly, I search for genetic signal of convergence in recurrent molecular adaptation to high altitude, migration, diving, wintering and flight in mitochondrial genomes of birds. I develop an approach for detection of repeated coincident changes in genotype and phenotype, indicative of an association between the two. I describe a number of candidate sites involved in recurrent adaptation in NADH dehydrogenase genes; however, the majority of convergence events can be explained by random coincidences without invoking adaptation.

Keywords: parallelism, convergence, speciation, positive selection.

Publications

1. **Burskaia V**, Naumenko S, Schelkunov M, Bedulina D, Neretina T, Kondrashov A, Yampolsky L, Bazykin GA. 2020. Excessive Parallelism in Protein Evolution of Lake Baikal Amphipod Species Flock. Genome Biology and Evolution 12 (9): 1493-1503.

Acknowledgements

I would like to thank my supervisor Georgii Bazykin for the opportunity to study evolutionary genomics and to work in exceptionally supportive environment. Your intellectual leadership and kind attitude made these years productive and challenging. I am also grateful to Alexey Kondrashov for sharing his huge experience and knowledge.

Many thanks to my colleagues: Galina Klink, Anastasia Stolyarova, Kseniia Safina, Alexandra Bezmenova, Nadezhda Potapova, Mikhail Schelkunov, Sergey Naumenko, Olga Vakhrusheva, Sonya Garushyants, Elena Nabieva, Nadezhda Terekhanova, Ivan Kuznetsov, Dmitrii Biba and Lev Yampolsky. Participating in discussions makes me feel alive every time I meet you.

I also want to express my sincere gratitude to Tatiana Neretina, who made all the wet work for this project, and to my great collaborators: Andrey Chaplin, Olga Sigalova, Vsevolod Filaretov and Kirill Konovalov – your help and inspiration are priceless. Special thanks to Ilya Artyushin, who helped me at all stages of the work, and to Ekaterina Demidova and Vera Mukhina, who encouraged me to carry on.

Looking backwards, I want to express my deepest gratitude to Mikhail Gelfand and all the teachers at Moscow Bioinformatics School – for the opportunity to study really fascinating things. I always remember my earlier teachers – Anna Bannikova, Vladimir Lebedev, Viktor Alekseev, Boris Sheftel and Olga Rozhanskaya – your contribution is very important. I also wish to thank the Dynasty Foundation for all the books that brought me here.

I want to thank my family and friends for their endless support. I am especially grateful to my beloved parents, Alexandra Panaiotidi and Oleg Bourskii, who instilled in me a love of natural sciences.

Table of Contents

List of Symbols, Abbreviations

ATP6-ATP8 – genes which code ATP synthase subunits

COX1-COX3 – genes which code Cytochrome c oxidase subunits

CYTB – gene which code Cytochrome b

LCA – last common ancestor

ND genes, or *ND1-ND6, ND4L* – genes which code NADH dehydrogenase subunits

OXPHOS – oxidative phosphorylation

Respiratory complex I – NADH: ubiquinone oxidoreductase, Type I NADH dehydrogenase

SNP – single nucleotide polymorphism

List of Figures

List of Tables

Chapter 1. Introduction

1.1. Relevance and significance of the work

The term “parallel evolution” corresponds to independent evolution of similar traits from a similar ancestral condition. When applied to phenotypic evolution, it usually concerns closely related species and describes, for example, independent appearance of high-altitude adaptations in hummingbirds. A related term – “convergent evolution” – describes a similar phenomenon as parallel evolution, except that the ancestral conditions may be different. For phenotypic evolution, the ancestral state can often not be inferred with confidence, so the distinction between parallel and convergent evolution is somewhat elusive; convergent evolution is usually used to describe independent appearance of similar adaptations when the ancestral state is not known exactly, in particular, in distantly related species, such as evolution of a complex eye in vertebrates and cephalopods.

Parallel and convergent evolution of adaptive phenotypic traits has remained in the focus of interest of evolutionary biologists for over two centuries. Charles Darwin in "The Origin of Species…" (1859) discussed how striking parallel adaptations are, which arise even in quite divergent groups of species in the process of adaptation to similar environments (pages 427-428). He also referred to Lamark, who was the first to draw attention to "analogous" fins of dugongs, whales and fishes back in the 18th century.

Almost at the same time as these ideas originated, the difficulty of distinguishing between independent origin and shared inheritance of similar traits became apparent. Writing about closely related species in "Variation of animals and plants" (1868, pages 348-349), Darwin suggested that widely observed similar traits are a result of ancestral variation. Much later, in 1922, Nikolai Vavilov described “The law of homologous series in variation”, demonstrating the ubiquity of similar variation in multiple groups of closely related species, including almost all known groups of agricultural plants. Similar variation inherited by close species from their ancestors can lead to independent fixation of similar traits. It is preferable to distinguish this scenario from independent origin of the same trait through repeated mutation, but such distinction can be hard.

While the phenomenon of parallel evolution involves nearly all living things and is a widely discussed topic, its genetic mechanisms are largely unknown. In this thesis, I examine the role of single nucleotide or amino acid substitutions in parallel and convergent evolution. Yet not all parallelism and convergence is adaptive. Given a parallel substitution, how do we prove that it is adaptive rather than neutral or, perhaps, deleterious? The most conventional method to prove that parallel (or indeed any) substitutions increase fitness is by detection of signatures of positive selection – a process by which novel advantageous genetic variants fix in population. As evidence of positive selection, one can use elevated ratio of nonsynonymous to synonymous substitutions, or change in amino acid preferences at a site. In addition, if a mutation repeatedly coincides with changes of the phenotype in multiple independent lineages, it is likely to be responsible for this phenotypic effect, and mutations with phenotypic effects that fix

repeatedly in evolution are likely to be adaptive. Here, I use all these approaches to find parallel adaptation acting independently in multiple lineages.

In fact, this study allows to find cases in which similar adaptations can be achieved by a subtle change of protein structure. I apply contemporary methods, and develop new ones, to study parallel adaptations in the two extremes – of very close and very large phylogenetic distances. Examination of phylogenetically close organisms demonstrates to what extent could rise amount of parallel evolution in most favorable conditions (Chapter 3). By contrast, in distantly related species, the phenotypic parallelism is rarely manifested through genetically identical substitutions (Chapter 4).

1.2. Goals and objectives

The goal of this study is to assess the role of single-nucleotide substitutions in adaptive parallel evolution at different interspecific distances.

To meet this goal, I have set forth the following objectives:

1. To estimate the prevalence of single-position adaptive parallel evolution at the whole-genome level in large groups of closely related species of different extent of relatedness.

2. To estimate the contribution of single-position adaptive parallel evolution at large (between-order) phylogenetic distances to repeated parallel origin of similar phenotypes.

1.3. Study system

To study the evolution of close species, I use a transcriptomic dataset of amphipods from the lake Baikal which includes hundreds of recently divergent species, representing a unique example of rapid radiation. Another genome-level dataset of close species comes from the well-known cichlid fish species flock from lake Malawi. As a set of distant species, I use the alignment of vertebrate genomes which represents a dataset with the number of species comparable to that in the amphipod and cichlid datasets. To study parallel phenotypic adaptation, I use a set of mitochondrial genomes of birds that have repeatedly and independently acquired similar phenotypes.

1.4. Implications of this work

It was revealed an unprecedented amount of functionally relevant parallelism in the evolution of lake Baikal amphipods. Prior to this work, inference of parallelism has been largely limited to specific loci, and a genome-wide excess of parallel substitutions has, to our knowledge, never been observed. This finding substantially challenges our understanding of parallel evolution as something limited to special cases, and instead suggests that it may be a prevalent force in genomic evolution.

Broadly speaking, inference of parallel evolution is instrumental in fields of biology as diverse as agricultural and medical genomics.

From one hand, understanding to what extent (and at what interspecific distance) parallel adaptations could be achieved by similar mutations is useful, as it describes stability of adaptive landscape and predictability of mutations function. Each adaptive mutation can maintain its function only in limited range of genomic environment changes. Thus, when model organisms are used for disease or adaptation studies, and some significant mutations in model organisms are found – it is crucial to understand how that mutation will behave in other species, when introduced by gene editing. Thus our findings can help in model organism selection.

From the other hand, in this study I demonstrate the utility of phylogeny-based methods for inference of adaptive mutations. Analysis of parallel evolution under positive selection is a promising method for detection of substitutions responsible for specific traits in multiple distinct lineages. In agricultural studies, genotype-phenotype associations may be mapped by repeated cooccurrence of phenotype changes with genomic changes across the phylogeny. In cancer genomics recurrent mutations of the same gene or site are critical for distinguishing driver from passenger mutations.

Finally, analysis of parallel evolution in phylogenetically close species flocks sheds the light on the mechanisms of speciation. According to some theories, high amount of adaptive parallelisms is a sign of high ancestral polymorphism or hybridization, which facilitates explosive speciation. In times of endangered species diversity our finding of excessive parallel evolution in amphipoda species flock may be useful in many senses.

1.5. Personal contribution

All analyses and conceptual work in this thesis were performed by the author, except those explicitly listed below. All the text presented in the thesis was also written by the author, except first and second paragraphs in section 3.2.1: these methods were partially described by Mikhail Shelkunov. The contributions of other coauthors were as follows. For Chapter 3, Sergey Naumenko provided previously published transcriptomes of amphipods, and participated in discussion of methods for inference of parallel evolution. Mikhail Shelkunov constructed the alignment of amphipod orthogroups and reconstructed their phylogeny. Tatiana Neretina resequenced selected regions by Sanger method. Daria Bedulina provided additional unassembled transcriptomic data for population samples of amphipods. Lev Yampolsky consulted on amphipod ecology and evolution. Alexey Kondrashov participated in discussions of the project and contributed to conceptualization of the obtained results. Georgii Bazykin participated in conceiving and conceptualizing the study, supervised the project and edited the papers.

For Chapter 4, Nadezhda Potapova gave the initial impulse for the study of bird parallel evolution and participated in fruitful discussions. Ilya Artushin reconstructed the phylogeny of birds. Kirill Konovalov predicted the native protein structure of respiratory complex I. Georgii Bazykin supervised the project and edited the papers.

Chapter 2. Review of the Literature

The search for molecular causes of parallel evolution is incredibly attractive: it allows to study mechanisms of adaptation and to find associations between phenotype and genotype. However, when you search for parallel adaptation in many genes, it often becomes looking for a needle in a haystack: only a few of parallel mutations are adaptive (if any). Here I discuss the contemporary state of the problem.

2.1. Homoplasy and hemiplasy

When a substitution is observed to appear independently in different lineages, it could be caused by either homoplasy or hemiplasy. The term "homoplasy" unites parallel and convergent evolution, i.e. cases in which similar substitutions result from independent mutations. Parallel mutations imply the origin of the descendant trait state from identical ancestral states, while under convergent evolution, the ancestral states differ. Homoplasy is widely observed in unicellular organisms and viruses due to their huge effective population size, especially under a similar selection pressure (Bailey *et al*. 2015). Experimental bacterial and viral evolution (Woods *et al*. 2006; Baym *et al*. 2016, Bertles *et al*. 2019) allows to trace genomic changes from ancestral state in single organism to identical, but independently evolved derived states. Usually in these experiments strong selection pressure is applied to experimental populations. The bulk of mutations accumulated in these experiments are seemingly functional and have a fitness effect, suggesting that they contribute to adaptation. For example, almost all mutations in protein-coding regions are nonsynonymous. Some studies demonstrate fantastic

repeatability of adaptive changes: in the study by Bertles *et al*. (2019), 62% of high-frequency mutations that occurred in experimental evolution of HIV-1 line were also found in another experimental evolution line.

Another process rich in homoplasy is recurrent adaptation of pathogens to their hosts. For example, Xue *et al*. (2017) describes parallel mutations in the hemagglutinin gene of the influenza virus, in which similar substitutions appear within individual patients, in different patients and even in the global influenza population. Based on parallel evolution detection, recent studies revealed mutations associated with antibiotic resistance in *Neisseria meningitidis* (Collins and Didelot 2018) and *Mycobacterium tuberculosis* (Coll *et al*. 2018). Another evidence of repeatable evolution comes from large hypermutable "populations" of antibodies: Strauli *et al.* (2016) describes convergent antibody repertoire response to influenza vaccine. These examples together indicate that adaptive mutations can arise independently.

Homoplasy in genomes of multicellular organisms is not so widespread, yet it was shown in a considerable number of studies. Karasov *et al.* (2010) describes independent mutations, associated with appearance of insecticide resistance in *Drosophila melanogaster*. Similarly, Kreiner *et al*. (2019) found mutations, associated with independent evolution of herbicide resistance in *Amaranthus tuberculatus*. Lim *et al*. (2019) reports parallel evolution in sites, genes and pathways of high-altitude hummingbirds.

Besides "true" homoplasy which describes fixation of independent mutations, hybridization and incomplete lineage sorting could also lead to a similar pattern of

apparent substitutions. Hahn and Nakleh (2015) call such cases “hemiplasy” – the term unites “any incorrect inference about character-state evolution caused by gene-tree discordance, regardless of the cause of discordance”. Situations when the same trait appears independently as a result of allele sorting are quite widespread and sometimes are driven by positive selection. E.g., independent populations of saltwater sticklebacks repeatedly acquire adaptation to freshwater environments involving similar ancestral alleles (Terekhanova *et al.* 2014). Likewise stick insects repeatedly develop protective coloration based on ancestral polymorphism (Soria-Carrasco *et al.* 2014). Hybridization is another source of "spurious homoplasy". It has been detected in many species flocks and can be adaptive. In particular, hybridization has led to introgression of opsin genes in lake Victoria cichlids (Meier *et al*. 2017). *Cyprinodon* pupfishes developed scale-eater phenotype, based on introgressed loci, in different lakes of San Salvador island (Martin and Feinstein 2014). *Heliconius* butterflies developed similar protective wing coloration due to promiscuous exchange of colour-pattern genes (Dasmahapatra *et al*. 2012). Altogether hemiplasy as well as homoplasy plays a significant role in adaptive evolution.

Above-mentioned studies show particular mechanisms of similar genotype appearance: it is possible in some fortunate conditions. However, most data keep secret of identical sequence origin. The first signal of possible homo- or hemiplasy often comes from phylogenetic reconstructions. Homoplasy can affect few genes and few positions in these genes. Therefore, it can sometimes be detected in single-gene phylogenetic reconstructions (Li *et al*. 2008, Castoe *et al*. 2009). Hemiplasy can cause unresolved phylogenies, even when phylogenetic reconstructions are based on full genomes. Thus,

Hahn and Nakhleh (2015) show that three alternative phylogenies of three basal bat clades are almost equally probable. The study is based on thousands of genes: it means that unresolved phylogeny is a result of a specific evolutionary process, but not an artifact of tree reconstruction methodology. Similar conflicting tree topologies were obtained in genome-based studies of bird deep phylogeny (Jarvis *et al*. 2014), cichlid phylogenies of lakes Malawi and Victoria (Malinsky *et al*. 2018, Meier *et al*. 2017) and many others. Thus, hemiplasy as well as homoplasy is hidden in many phylogenies, although only some cases are caused by adaptive evolution.

2.2. Adaptive and neutral parallel evolution

Not all parallelisms are adaptive. Recurrent substitutions are typically classified into two types: foreground and background (Rey *et al.* 2018).

Foreground substitutions are associated with the convergent phenotype. That association can be a consequence of parallel positive selection, or it can emerge after relaxation of selection in recurrent cases of regressive evolution.

Background substitutions may occasionally form a pattern of changes that are almost indistinguishable from those of foreground substitutions. This could happen by chance alone, as there is only a limited number of different nucleotides and amino acids. Additionally, the rates of substitutions can be biased compared to expectation by two processes. First, the rate of a particular mutation can be unexpectedly high or low. Second, once a mutation arises, its change in frequency can have a preferential direction – either upward (e.g., for advantageous mutations or for mutations favored by biased

gene conversion) or downward (e.g., for deleterious mutations or for mutations disfavored by biased gene conversion). The former process is referred to as mutation bias, while the latter, as fixation bias. Either process can increase the probability of changes between some amino acids and cause repeatability of corresponding changes. The most typical example of a mutation bias is the CpG hypermutability, whereby a certain type of nucleotide substitution, a C->T change in the CpG context, occurs at a much higher frequency than the same mutation in other contexts. An example of a fixation bias is the GC-biased gene conversion (Pessia *et al*. 2012), whereby the high-GC-content nucleotide tracts are favored by the process of gene conversion and thus increase their population frequency. These effects can be amplified by demography; for example, population bottlenecks can potentially lead to apparent convergence, as fixation of random mutations is more probable in small populations, and some random mutations are more probable than others (Lefebure *et al.* 2017). Furthermore, background substitutions can be favored by selection not directly but as a result of their epistatic interactions with adaptive ones (Kryazhimskiy *et al*. 2014; Storz 2016). Finally, the set of potential adaptative substitutions is shaped by adaptive constraints: if only a few amino acids are permitted in a particular position, they will change from one to another, and not accounting for this in the null-model can lead to spurious inference of adaptive convergence (Klink and Bazykin 2017; Klink *et al*. 2017).

A wide variety of approaches have been developed for detection of parallel evolution and distinguishing between foreground (adaptive) and background (non-

adaptive) convergent substitutions. The methods differ in many respects, including the definition of parallel evolution.

2.3. Parallel evolution at individual sites

Castoe *et al*. (2009) and Parker *et al*. (2013) called a site convergent if it supports a phylogeny where branches with independently derived states are grouped together (site-specific likelihood support or ΔSSLS method). Thomas and Hahn (2015) argue that ΔSSLS is a method for indirect assessment which does not measure convergence itself: aberrant phylogeny could be caused by many factors other than convergent substitutions. Another weakness of ΔSSLS is the problem with a proper design of a null model: simulations hardly can imitate the real evolution process, so we cannot say for sure how many sites with "wrong" topology should be found in the background (Castoe *et al*. 2009). So, the best null model could be obtained from the real data. For example, Zou and Zhang (2015a) demonstrated that in the Parker *et al.* (2013) study, ΔSSLS gave erroneous evidence of genome-wide convergence in echolocating mammals. Zou and Zhang (2015a) calculated ΔSSLS for a pair of species consisting of one echolocating (bat and one non-echolocating (cow), and found similar number of "convergent" genes as in the original study of convergence between an echolocating bat and an echolocating dolphin, undermining the conclusions of the original study.

Another indirect method for inference of a genotype change associated with a trait change is estimation of the dn/ds levels in branches with phenotype change. Comparison of dn/ds levels between background and foreground branches allows to detect sites that

experience pressure of positive selection at different linages simultaneously (Kosakovsky Pond 2005). Yet this approach does not allow to distinguish between convergent and divergent evolution at the chosen branches.

To estimate convergent evolution directly, the most straightforward solution would be counting all mutations giving rise to the same amino acid or nucleotide state that happened at the same branch as the phenotype change. This strategy was used by Thomas and Hahn (2015) and by Collins and Didelot (2018). Yet for highly divergent species, a change to same amino acid could be too strong a restriction. Therefore, another useful definition of convergence was used by Rey *et al*. (2018). They call convergent all substitutions that appeared at the same branch as the phenotype change, and lead to a change in amino acid preferences at this position.

2.4. Parallel evolution at gene level

Parallelism at individual sites is often hard to detect: best hits are "killed" by multiple test correction if whole genomes are analyzed. Therefore, many classical examples of parallel adaptation, like high-altitude adaptation in bird hemoglobin (Natarajan *et al*. 2016), would have never been detected if the study was based on a huge alignment. On the other hand, whole-genome site-based screening for parallel evolution often leads to false positive findings (Parker *et al*. 2013, Foote *et al*. 2015). So often a better way for search for parallel adaptation is usage of gene-based approaches. One way is to measure the changes in evolutionary rate at those branches where the convergent phenotype appears (Partha *et al*. 2019). The main idea here is that a decrease in

evolutionary rate is caused by stronger selective constraints, while increased evolutionary rate can be caused either by relaxed selective constraints or by adaptation. With this approach, Chikina *et al*. (2016) discovered genes with parallel acceleration and deceleration of evolutionary rates in marine mammals.

A separate group of studies is focused on phenotype loss followed by gene inactivation. These studies are specialized on detection of gene losses and specific inactivation signatures. Hiller *et al*. (2012) demonstrated independent inactivation of *Gulo* gene in primates, guinea pigs, and some bats (this gene is responsible for synthesis of vitamin C). Meredith *et al*. (2011, 2014) found gene inactivation in connection with loss of enamel in birds and other toothless animals, and Liu *et al*. (2019) demonstrated convergent degeneration of olfactory receptor gene repertoire in marine mammals.

2.5. Parallel evolution at whole-genome level

While at individual sites or genes it is possible to find signal of parallel adaptation, at whole genome level, we see the cumulative effect of all types of parallelisms, both adaptive and neutral, and the neutral effect can be expected to predominate. Since 1968, when Kimura proposed the neutral theory of molecular evolution, discussions about the proportion of adaptive mutations in divergent evolution have become a popular topic. Smith and Eyre-Walker (2002) estimated that 45% of all amino-acid substitutions have been fixed by natural selection in Drosophila. According to the estimates by Sawyer *et al*. (2003), 94% of fixed mutations in Drosophila are beneficial. The amount of positively selected mutations should be closely related with the

number of parallel adaptive evolution at whole-genome level. Yet, to claim adaptation, the level of parallelism needs to exceed that expected neutrally. At the whole-genome level, scans for parallel nonsynonymous substitutions revealed almost no evidence for higher-than-neutral levels of molecular parallelism (Bazykin *et al*. 2007; Foote *et al*. 2015; Thomas and Hahn 2015; Zou and Zhang 2015a, 2015b). In those studies that demonstrate a higher than neutral level of parallel evolution at the whole-genome level, the effect can be explained by constraints of negative selection. Thus, Rokas and Carroll (2008) showed that across 8 clades protein sequences underwent twice as many homoplastic substitutions than what was expected by neutral processes alone. They explain these findings by a combination of negative selection constraints and parallel positive selection, and it is difficult to distinguish between those forces.

The frequency of parallel amino acid substitutions between the two lineages (relative to the neutral expectation) tends to decrease with genetic distance between them (Conte *et al*. 2012; Usmanova *et al*. 2015; Zou and Zhang 2015b). Although there is a temptation to explain this effect by parallel evolution driven by positive selection, this pattern can actually be caused by many factors. Probably the first of them is the change in constraints set by negative selection with phylogenetic distance (Povolotskaya and Kondrashov 2010, Klink *et al*. 2017).

Recently, as amount of sequenced data grows dramatically, more complex understanding of parallel evolution patterns appears. Here I examine reproducibility of adaptation at different phylogenetic distances and at different scales: from single position

to the whole genome. Modern methods allow us to look inside the machinery of evolution and to discover its patterns.

Chapter 3. Excessive Parallelism in Protein Evolution of Lake Baikal Amphipod Species Flock

3.1. Introduction

Adaptive parallel evolution of closely related species is widespread and may affect single positions, genes, or entire pathways. For example, Natarajan *et al*. (2016) demonstrated, that in close groups of birds similar substitutions in hemoglobin cause similar effect of enhanced oxygen affinity, while in distant species those substitutions are ineffective. It is quite natural to expect adaptive parallel evolution in close species: in same adaptive landscapes similar mutations will likely cause similar effects.

However, at the whole-genome level most parallel substitutions are not caused by positive selection. Many previous studies demonstrated, that the closer two species are, the higher is frequency of parallel amino acid substitutions between the two lineages (relative to the neutral expectation) at the whole-genome level (Conte *et al*. 2012; Usmanova *et al*. 2015; Zou and Zhang 2015b). Yet that excess of parallel amino acid substitutions in close species can be explained by change in constraints set by negative selection (Povolotskaya and Kondrashov 2010, Klink *et al*. 2017). Still there stays a possibility that adaptive parallel evolution can prevail over neutral parallel evolution even at the whole-genome level, especially in closely related species. In this chapter, I examine a group of closely related amphipoda species flock and cichlidae species flock for presence of genome-wide signature of adaptive parallel evolution, which is based on single-nucleotide parallelism.

To compare the abundance of parallel substitutions driven by selection with neutral expectations, one can extend the conventional dN/dS-type approach, namely, estimate the relative rates of nonsynonymous and synonymous parallel evolution. Under purely neutral evolution, a site that has experienced a substitution between a pair of species can only be neutral, and if the fitness landscape is invariant, an identical substitution at this site between other two species is also expected to occur at the neutral rate. Deviation from this expectation could occur due to weak selection preventing substitutions within one of the pairs of species, or due to changing amino acid preferences between pairs (Bazykin *et al*. 2007); either trend will cause parallel nonsynonymous substitutions to be less frequent than the synonymous control. Conversely, nonsynonymous parallel substitutions could be more frequent than synonymous ones if the parallel evolution is adaptive, that is, driven by positive selection.

A genome-wide analysis of three quartets of species of vertebrates, insects, and fungi has shown that the rate of parallelism at nonsynonymous sites is lower than that at synonymous sites. This has been interpreted as evidence for weak negative selection and/or change in single-position fitness landscape between species (Bazykin *et al*. 2007), consistent with findings using other logic (Bazykin 2015).

Here, I apply this approach across many genome comparisons, asking how the amount of parallel amino acid evolution depends on the divergence level between the considered species. For this purpose, I use recently published data set of transcriptomes of the species flock of closely related baikalian amphipods (Naumenko *et al*. 2017), as well as two other data sets: cichlid fishes from the lake Malawi species flock, and a group

of more distantly related vertebrates. In the absence of adaptation, the rate of parallel nonsynonymous to synonymous evolution should be less or equal to 1. Strikingly, in the amphipod data set, I find nonsynonymous parallel substitutions to be more frequent than synonymous ones, suggesting prevalent selection in favor of the same derived variants in different species. A similar, although weaker, effect was found in closely related cichlid fish data set. In amphipods, within species polymorphism at sites of past parallel nonsynonymous substitutions is low, indicating that these substitutions were driven by positive selection. By contrast, in a data set of distantly related vertebrates, the rate of nonsynonymous parallel substitutions is lower than that of synonymous ones, consistent with prevalent negative selection.

3.2. Methods

3.2.1. Divergence data

The three datasets were analyzed as follows. First, I used the transcriptomic sequences of closely related amphipod species from Lake Baikal (Naumenko *et al*. 2017). Of the 67 species analyzed in that work, I picked the 46 species for which the sequenced sample was based on exactly one individual. Orthologous groups of genes were calculated with OrthoMCL 2.0.9 with the inflation parameter set to 1.5 (Li 2003). If a particular species carried multiple paralogous sequences of a gene, this species was excluded from the analysis of this gene. Codon-aware alignments for orthogroups were obtained with TranslatorX (Abascal *et al*. 2010) using the Muscle method (Edgar 2004).

Poorly aligned sequences were detected and removed from the alignments using the following rule:

1) A column in an alignment was considered "good" if it carried the same nucleotide in at least 50% of species;

2) Sequences for which fewer than 50% positions were "good" were removed from the alignment.

This exclusion process was performed using TrimAl 1.4 (Capella-Gutierrez *et al*. 2009). It resulted in 4366 orthologous groups of genes. Alignments for all genes were concatenated, and a phylogenetic tree was reconstructed using RAxML 8.1.20 (Stamatakis 2014) with GTR+Gamma model, 20 starting maximum parsimony trees and 100 bootstrap analysis pseudoreplicates. As mutations in the third positions of codons are often synonymous, the third positions of codons accumulate substitutions quicker than the first two. Therefore, partitioning was used, with separate substitution models for the first two and for the third codon positions. The obtained tree (Figure 3.3A) was similar to that obtained previously (Naumenko *et al*. 2017).

For the cichlid species flock from lake Malawi, exon alignments were extracted from genomic data of 62 species each mapped onto the assembly of the *Maylandia zebra* (Boulenger, 1899) (assembly ID = MetZeb1.1_prescreen) (Malinsky *et al*. 2018). Using this annotation (available at the Cambridge cichlid browser), I picked the longest isoform of each gene, for a total of 15318 transcripts. As the phylogenetic tree for this set of species, I used the Maximum Clade Credibility phylogenetic tree from the original paper (Malinsky *et al*. 2018).

For the analysis of highly divergent vertebrates, I used the exon alignments of 100 species to the hg19 human genome. All together, 21520 gene alignments were used. These data were fetched from the UCSC database (Karolchik *et al*. 2007) together with the corresponding tree.

3.2.2. Calculation of P statistic for a species quartet

To compare the amount of amino acid-level parallelism to that expected neutrally while controlling for the heterogeneity of mutation rates between genomic sites, and to make these values comparable between groups of species, I used the approach developed previously (Bazykin *et al*. 2007). In brief, I rely on quartets of species with a particular phylogenetic relationship, namely, composed of two clades ('evolutionary paths') each involving two species (Figure 3.4A); and consider positions orthologous between these four species. A difference between the two species of a pair implies that a substitution has occurred in at least one of the two lineages leading to these species from their LCA, even though the direction of this substitution can be unknown. I use one species pair ('path I') to identify the sites that had experienced such a substitution and ask if the same substitution has occurred in parallel in the other species pair ('path II'). I assume parsimony, i.e., that at most one substitution has occurred between the two species within a path; violation of this assumption may lead to underestimation of the rate of parallelism. Pooling across different sites with substitutions in path I, it's possible to infer the rate of parallelism by measuring the fraction of sites at which the same substitution has also occurred between the species of path II.

More precisely, at a site with a difference between the two species of path I, four possible patterns can be observed in a quartet: (i) ((A,B)(A,B)), (ii) ((A,B)(B,A)), (iii) ((A,B)(A,A)) and (iv) ((A,B)(B,B)). Here, in each of the four cases, the first bracket represents the two species belonging to path I, while the second bracket represents the two species belonging to path II; identical letters signify identical nucleotides. The cases of mutation into a different (non-A, non-B) nucleotide in path II were not considered. Like in Bazykin *et al*. 2007, I also exclude invariant sites ((A,A)(A,A)) and ((B,B)(B,B)); considering those sites would make our estimated additionally dependent on the rate of substitutions in path II, and harder to interpret.

Patterns (i) and (ii) correspond to parallel evolution. As a scaleless estimate of the rate of parallelism for a given category of sites and substitutions, I measure the fraction of sites, among those with a substitution in path I, that also experienced a substitution in path II:

$$d_P = \frac{N\big((A,B)(A,B)\big) + N\big((A,B)(B,A)\big)}{N\big((A,B)(A,B)\big) + N\big((A,B)(B,A)\big) + N\big((A,B)(A,A)\big) + N\big((A,B)(B,B)\big)}$$

where *N* corresponds to the number of sites with the corresponding pattern.

d_P can be calculated for any class of sites and/or substitutions. All four-fold degenerate sites were used to estimate the rate of synonymous parallel substitutions dS_P, while all non-degenerate sites were used to estimate the rate of nonsynonymous parallel substitutions dN_P. To account for the differences in mutation rates between the six single-

nucleotide mutation types, namely, A↔C, A↔T, A↔G, C↔T, C↔G and T↔G, I calculated dS_P and dN_P for each such mutation type separately.

Finally, I use the synonymous sites to estimate the level of parallelism expected neutrally, and calculate $P = \frac{dN_P}{dS_P}$, i.e., the ratio of the rates of parallel nonsynonymous and synonymous substitutions. For Figure 3.4B, the values of *P* were averaged across the six mutation types. *P*=1 implies that the substitutions that occurred in path I also occur in path II at the neutral rate; while deviations from 1 imply selection on parallel nonsynonymous substitutions. The *P* test uses normalization, which should not cause *P*>1 when nonsynonymous substitutions are neutral (even when they are limited by constraints).

3.2.3. Comparing the numbers of parallel and divergent substitutions

To visualize the relationships between parallel and divergent evolution in amphipods (Figure 3.3B,C), I compared, for each quartet, the number of sites with the identical (((A,B)(A,B)) or ((A,B)(B,A))) and different (((A,B)(A,C)), ((A,B)(C,A)), ((A,B)(B,C)), ((A,B)(C,B)), or ((A,B)(C,D)), where C and D are distinct non-A, non-B nucleotides) nucleotide substitutions between species in Path I and in Path II. That analysis was performed independently for 6 mutation types (AC, AT, AG, CT, CG and TG). That comparison of parallel and divergent substitutions is used in the work to demonstrate prevalence of parallel nonsynonymous evolution rate over parallel synonymous evolution rate, which can emerge because of negative selection constraints or due to parallel adaptive evolution.

3.2.4. P test insensitivity to negative selection constraints

In *P* test (3.2.2. section) and in more typical comparison of convergent and divergent substitutions (3.2.3 section) different normalization approaches are used. Both are based on analysis of 6 kinds of substitutions. However, only *P* test is insensitive to excessive amount of parallel evolution, caused by negative selection constraints. Here I demonstrate that *P* test never gives values higher than one under purely negative selection, applying it on two substitution matrices (Figure 3.1A,B). One of these matrices models evolution in neutral nucleotide (four-fold degenerate) sites (Figure 3.1A). Another one models evolution under negative selection constraints in non-degenerate coding sites: strong matrix asymmetry imitates amino acid site, in which only switches between two amino acids are permitted (Figure 3.1B).

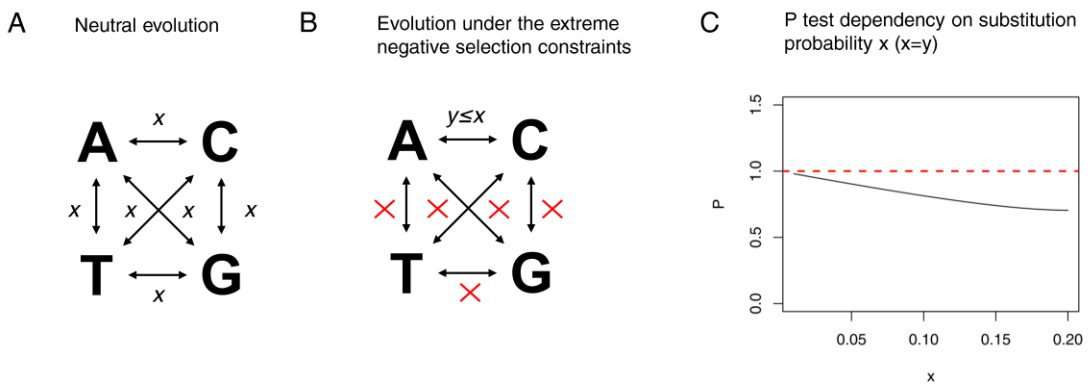


Fig. 3.1: P test under purely negative selection.

Each kind of substitutions (AC, AT, AG, CT, CG, TG) in path I was considered separately. I suppose that terminal branch lengths in path I and path II are same, and substitution probabilities do not change along the tree. I also do not consider mutations in inner branches, as if both of them are very short.

Let us take as x probability of neutral substitution between any 2 of 4 nucleotides (A,C,T,G) in time t. In nonsynonymous sites probability of substitution between any nucleotide pair (let it be y) is typically 10 times lower, than in synonymous sites. As a maximum, it could be equal to the probability of neutral substitutions ($y \leq x$).

Now, when probabilities of substitutions are defined, let us consider the normalization approach, applied to the AC substitution type. In *P* test parallel substitutions of AC type (A,C)(A,C) are normalized by number of parallel and divergent substitutions of same kind: (A,C)(A,C), (A,C)(A,A) and (A,C)(C,C). Ancestral state of all 4 species could be A, C, T or G. First I estimate probabilities of different patterns in neutral model (Figure 3.1A), when each substitution is equally probable.

If ancestral state was A, probability of AC mutation in path I is $2x(1- 3x)$: x for probability of substitution from A to C in first species, (1-3x) – for probability that in second species substitution from A to C, T or G have not happened, 2 – for two substitution combinations, which give AC in path I. Probability of AC substitution in path II from ancestral state A is similar, and finally probability of obtaining parallel pattern (A,C)(A,C) is $(2x(1- 3x))^2$. For the normalization it is needed to calculate the probability of patterns (A,C)(A,A) and (A,C)(C,C). Probability of pattern (A,C)(A,A) is equal to probability that AC substitution happened in path I, which is $2x(1- 3x)$, multiplied by probability that no other substitutions happened in path II. Probability that no other substitutions occurred in path II is $(1-3x)^2$, where (1- 3x) is for probability that in particular branch no substitutions happened. Finally, probability of (A,C)(A,A) pattern is $2x(1- 3x)^3$. Probability of pattern (A,C)(C,C) is equal to probability that AC substitution

happened in path I, which is $2x(1- 3x)$, multiplied by probability that two substitutions to C happened in path II, which is x^2. Finally, it is $2x^3(1- 3x)$.

If ancestral state of four species was C, calculations are same.

If ancestral state was T, probability of AC pattern in path I is $2x^2$: x^2 for two independent mutations from T to A or C, 2 – for two possible combinations. Probability of AC substitution in path II from ancestral state T is same, and finally probability of obtaining parallel pattern (A,C)(A,C) from ancestral state T is $4x^4$. Probability of obtaining (A,C)(A,A) pattern is probability of AC pattern in path I, which is x^2, multiplied by probability of two mutations to A in path II (which is x^2). Finally, probability of (A,C)(A,A) pattern is $2x^4$. Probability of obtaining (A,C)(C,C) pattern is $2x^4$ too.

If ancestral state of 4 species was G, calculations are same.

Finally, if we suppose, that frequencies of nucleotides in the last common ancestor is similar, $dSp = 2((2x(1- 3x))^2 + 4x^4) / 2(((2x(1- 3x))^2 + 4x^4 + (2x(1- 3x)^3) + 2x^4 + 2x^3(1- 3x) + 2x^4))$.

Probabilities of different substitution patterns in model with negative selection constraints (Figure 3.1B) are calculated similarly, yet ancestral states T and G are not considered, as substitution probabilities from T and G to A and C are taken as equal to zero. Mutations to T and G nucleotides are also absent.

If ancestral state was A, probability of AC mutation in path I is 2y(1- y): y for probability of substitution from A to C in first species, (1-y) – for probability that in second species substitution from A to C have not happened, 2 – for two substitution combinations, which give AC in path I. Probability of AC substitution in path II from ancestral state A is similar, and finally probability of obtaining parallel pattern (A,C)(A,C) is $(2y(1- y))^2$. For normalization it is needed to calculate the probability of patterns (A,C)(A,A) and (A,C)(C,C). Probability of pattern (A,C)(A,A) is equal to probability that AC substitution happened in path I, which is 2y(1- y), multiplied by probability that no other substitutions happened in path II. Probability that no other substitutions occurred in path II is $(1-y)^2$. Finally, probability of (A,C)(A,A) pattern is $2y(1- y)^3$. Probability of pattern (A,C)(C,C) is equal to probability that AC substitution happened in path I, which is 2y(1- y), multiplied by probability that two substitutions to C happened in path II, which is y^2. Finally, it is $2y^3(1- y)$.

If ancestral state was C, calculations are same. Ancestral states T and G are not considered, as in asymmetric matrix mutations from T and G are forbidden.

Finally, $dNp = 2((2y(1- y))^2) / 2(((2y(1- y))^2 + 2y(1- y)^3 + 2y^3(1- y)))$.

In simplified form, if we take y=x to obtain highest possible values of *P*, equation for *P* looks as follows: $P = dNp /dSp = ((8 x^3 - 16 x^2 + 7 x - 1) (x - 1)) / (10 x^2 - 6 x + 1)$. That equation is defined at the interval from x=0 to x=0.33 and at all that interval it gives values lower than one. Due to simplified evolutionary model, most realistic values of *P* are given for low substitution rate x (Figure 3.1C). The model shows, that even if

nonsynonymous substitution matrix is extremely asymmetric, and y=x, *P* test does not give values higher than one under purely negative selection.

3.2.5. Normalization by different mutation types, sensitive to negative selection constraints

In this work I also compare the numbers of parallel and divergent substitutions, calculating divergent substitutions in another way (section 3.2.3.). It gives estimation of nonsynonymous evolution rate higher, than synonymous due to either negative selection constraints or adaptation. Inability to distinguish between these two causes makes this method similar to the method of amino acid convergence estimation, applied in the study of Castoe et al. (2009).

Here I describe behavior of this more traditional approach under strong negative selection constraints, for the comparison with *P* test. For the purposes of comparison, I handle divergent and convergent substitution patterns as it is done in *P* test, separately calculating the number of synonymous and nonsynonymous parallel substitutions, normalized by divergent substitutions (called $dNp_{\text{different}}$ and $dSp_{\text{different}}$). I call this test $P_{\text{different}}$ (for different divergent substitutions, which are used for normalization).

I apply that test on two substitution matrices (Figure 3.2A,B) exactly as it was done for *P* test in section 3.2.4. First of these matrices models evolution in neutral nucleotide (four-fold degenerate) sites (Figure 3.2A). Second models evolution under negative selection constraints in non-degenerate coding sites: strong matrix asymmetry

imitates amino acid site, in which only switches between two amino acids are permitted (Figure 3.2B).

I suppose that frequencies of all nucleotides are equal in last common ancestor of four species. I also assume that terminal branch lengths in path I and path II are same, and substitution probabilities do not change along the tree. I also do not consider mutations in inner branches, as if both of them are very short.

Similarly to the test of *P* score (in previous section), I take as x probability of neutral substitution between any 2 of 4 nucleotides (A,C,T,G) in time t. In nonsynonymous sites probability of substitution between any nucleotide pair I take as y. Is is typically 10 times lower, than x, yet sometimes it could be equal to x (y≤x).

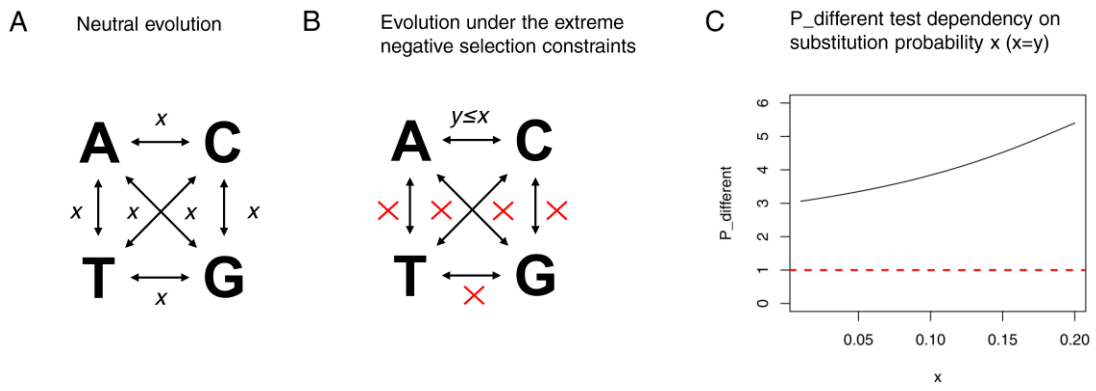


Fig. 3.2: $P_{different}$ test under purely negative selection.

In normalization by different substitution types, evolution rates are counted independently for 6 mutation types (AC, AT, AG, CT, CG, TG). Let us consider the normalization approach, applied to the AC substitution type. In that case, number of (A,C)(A,C) patterns are normalized by number of parallel and divergent substitutions, which occurred in path II in condition of AC substitution in path I. For normalization are

used following substitution patterns: (A,C)(A,C), (A,C)(A,T), (A,C)(A,G), (A,C)(C,T), (A,C)(C,G), (A,C)(T,G).

First, I estimate probabilities of different patterns in neutral model (Figure 3.2A), when each substitution is equally probable. In each path it may be required either one substitution - than its probability is $2x(1-3x)$, or two substitutions – than its probability is $2x^2$. All six analyzed patterns, depending on ancestral state, have probabilities either $(2x(1-3x))^2$, or $4x^4$, or $4x^3(1-3x)$. Considering all combinations of mutation patterns and all ancestral states, we obtained metric, partially close to *dSp*, but normalized by different substitution types: let us call it $dSp_{different}$.

$$dSp_{different} = (2(2x(1-3x))^2 + 8x^4)/(6(2x(1-3x))^2 + 24x^4 + 48x^3(1-3x)).$$

Probabilities of different substitution patterns in model with negative selection constraints (Figure 3.2B) are calculated similarly, yet ancestral states T and G are not considered, as substitution probabilities from T and G to A and C are taken as equal to zero. Mutations to T and G nucleotides are also absent. That leads to normalization of parallel substitutions by same parallel substitutions, as divergent substitutions are forbidden.

$$dNp_{different} = (2y(1-3y))^2 / (2y(1-3y))^2 = 1$$

Finally, if we take y=x to obtain highest possible values of $P_{different}$, the equation looks as follows:

$$P_{different} = dNp_{different} / dSp_{different} = (6(2x(1-3x))^2 + 24x^4 + 48x^3(1-3x)) / (2(2x(1-3x))^2 + 8x^4).$$

The function is defined at the interval from x=0 to x= 0.33, yet most precise values the model gives for low values of substitution rate x (Figure 3.2C). $P_{different}$ values are higher than one as a consequence of negative selection constraints, without any contribution of adaptive evolution to the model.

3.2.6. Traditional dn/ds calculation

Between the two species of path I in each quartet, I also calculated the numbers of nonsynonymous differences at non-degenerate sites, and of synonymous differences in four-fold degenerate sites, per such site across all codon sites of the alignment. I calculated $\frac{dN}{dS}$, as the ratio of these values.

3.2.7. Choice of quartets and filtering

For each of the three datasets, amphipods, cichlids and vertebrates, I randomly assembled 300 quartets of species (300 for amphipods, 300 for cichlids and 300 for vertebrates) with the tree topology shown in Figure 3.4A. For cichlids, I was concerned that the incongruence between gene trees constructed from different genes is rather high (Malinsky *et al*. 2018), which could make the inference of tree topology erroneous. Nevertheless, the nodes separating the six ecological groups are stable (Malinsky *et al*. 2018). Therefore, to ensure that unstable tree topology does not affect our results, I additionally constructed 300 quartets of cichlid species by picking two species from one ecological group and two from another, ensuring that the last common ancestor (LCA) of the two pairs was older than the LCA of each pair (Figure 3.4).

The orthologous nucleotide sites of the four species within a quartet were then filtered for data quality. I required the site to be surrounded by 20 nucleotide sites (10 to the left and 10 to the right) with no gaps in any of the four species of the quartet, and by 2 nucleotide sites (1 to the left and 1 to the right) with no substitutions in any of the four species of the quartet. I also required the upstream site to be non-C, and the downstream, to be non-G, to avoid the potential biases associated with the hypermutable CpG context. Finally, to ensure local alignment quality, I required the codon carrying the analyzed site to carry no nucleotide substitutions, and to be surrounded by 20 codons (10 to the left and 10 to the right) carrying in total no more than 3 nonsynonymous substitutions, between any of the species of the quartet.

Some quartets yielded very few parallel substitutions for some mutation types, making the estimates of *P* unreliable. To account for this, I only retained a quartet if the number of both synonymous and nonsynonymous parallel substitutions exceeded five for each of the six mutation types. This filtering retained 195 out of 300 quartets of amphipods, 147 of 300 quartets of vertebrates, and all 300 quartets of cichlids. Filtering and exclusion of quartets did not radically affect the results of the *P* test (Figure 3.7).

3.2.8. Validation by Sanger sequencing

The primers used are listed in Table A9. Purified PCR products were bidirectionally sequenced on an ABI 3500 Genetic Analyzer (Applied Biosystems) using the BigDye Terminator v 3.1 Cycle Sequencing Kit (Applied Biosystems) and the same primers as for PCR.

3.2.9. Polymorphism data

I used two sources of data on within-species polymorphism in amphipods at sites that had experienced a parallel pair of substitutions.

First, using TopHat (Trapnell 2009), I remapped the Illumina sequencing reads (Naumenko *et al*. 2017) corresponding to each sample back onto the assembly of that sample, and filtered out positions with quality<10 or with coverage<20. To detect polymorphic sites, i.e., sites heterozygous within the analyzed individual, I performed SNP calling by samtools-1.3.1 (Li, 2011). This analysis is further referred to as "individual-based polymorphism test".

Second, I used additional pooled transcriptomics data obtained from multiple individuals for one of the studied species, *Eulimnogammarus verrucosus* (Gerstfeldt, 1858). The dataset included 19 samples, each pooled across 4 individuals (Drozdova *et al*. 2019). The transcriptomics Illumina reads for each sample were mapped onto the reference assemblies of these species, filtering out positions with quality<10 or with coverage<20. To detect polymorphic sites, I performed SNP calling by samtools-1.3.1 (Li, 2011) for each sample individually. I considered a site polymorphic if 8 or more samples were preserved by filtering, and one or more sample was either heterozygous or homozygous with respect to a non-reference allele. This analysis is further referred to as "population-based polymorphism test".

3.2.10. Polymorphism at sites of a parallel substitution

Using the obtained amphipod SNP data, I estimated, among the sites that had underwent the same parallel A↔B substitution according to the reference genomes, the fraction of those that also carry both alleles A and B within a single species.

The individual-based polymorphism test was performed on the same 300 quartets of amphipod species that were previously used for the *P* test. For each quartet, I counted the sites corresponding to the ((A,B)(A,B)) or ((A,B)(B,A)) pattern, and, among those sites, the polymorphic sites were variants A and B were also both present in at least one of the four species. These values were summed over all 300 quartets, and the ratio of these sums is shown in Figure 3.8B. I compare the thus assessed SNP fractions for the nonsynonymous and synonymous sites (parallel pN/pS statistic).

For the population-based polymorphism test, I repeated the procedure for generating 300 species quartets, but this time required each quartet to include *E. verrucosus.* The fraction of polymorphic sites among the parallel sites was estimated in the same way as for the individual-based polymorphism test (Figure 3.8D).

To obtain the baseline polymorphism level at sites of a (non-parallel) substitution, I additionally calculated the number of polymorphic and monomorphic sites at positions that underwent a substitution between species of Path I (independently of whether a substitution has occurred in Path II), pooled these numbers over the 300 quartets, and showed the ratio of these sums in Figure 3.8AC. The comparison of these values for the nonsynonymous and synonymous sites yielded the nonparallel pN/pS statistic.

3.2.11. Search for possible phenotypic parallelism

To test for possible dependencies between parallel phenotypic and genotypic changes in amphipods, I randomly selected 40 quartets such that each path included one deepwater and one shallow water species (Table A10). Additionally, I considered detailed phenotypes of species that formed the quartets with the highest and the lowest values of the *P* statistic (Table A11). All phenotype descriptions are based on (Bazikalova *et al*. 1945).

3.2.12. Alignments of Eulimnogammarus clade

For GO analysis and for estimation of phylogeny incongruence I used orthologous groups and functional annotation from the study of Naumenko *et al*. (2017). That alignment contains 425 orthologous groups, which presents in each species of the clade and has no paralogs. Presence of all paralogs in all analyzed sequences is necessary for aforementioned analysis and thus data from Naumenko *et al*. study fits perfectly. GO analysis was performed with blast2GO software (Götz *et al*. 2008). Tree incongruence was visualized by DensiTree R package (Bouckaert 2010). Individual gene phylogenies were reconstructed with RAxML 8.1.20 (Stamatakis 2014) with GTR+Gamma model for 74 genes of 1000 or more nucleotides length.

3.3. Results

3.3.1. Phylogeny of Lake Baikal amphipods

The phylogenetic tree of Lake Baikal amphipods that was obtained based on 4366 orthologs (many of which were found only in a fraction of species) is similar to that obtained previously on the basis of ~175 groups of universal orthologs (Naumenko *et al*. 2017), indicating that phylogenetic reconstruction is robust. In particular, it confirms that amphipods populated Lake Baikal at least twice (Figure 3.3A), corresponding to the two invasion events. Most nodes have bootstrap support above 80.

3.3.2. High rate of parallel nonsynonymous evolution

To illustrate the amount of parallel and divergent evolution in amphipods, I calculate the number of substitutions of these two types in each of the considered quartets; this analysis is similar to that of Castoe *et al*. (2009). It gives a linear dependence between the numbers of divergent and parallel substitutions (Figure 3.3B,C) reflective of the differences in evolutionary distances between species quartets. The rate of parallel substitutions, relative to that of divergent substitutions, is higher for the nonsynonymous substitutions than for synonymous ones, as indicated by a steeper slope of the regression line for the former.

Although this finding implies an increased rate of parallel evolution among functional positions, this can arise both from parallel adaptation and differences in mode of constraint between synonymous and nonsynonymous sites (Bazykin *et al*. 2007, Rokas and Carroll 2008, Povolotskaya and Kondrashov 2010). To distinguish between these

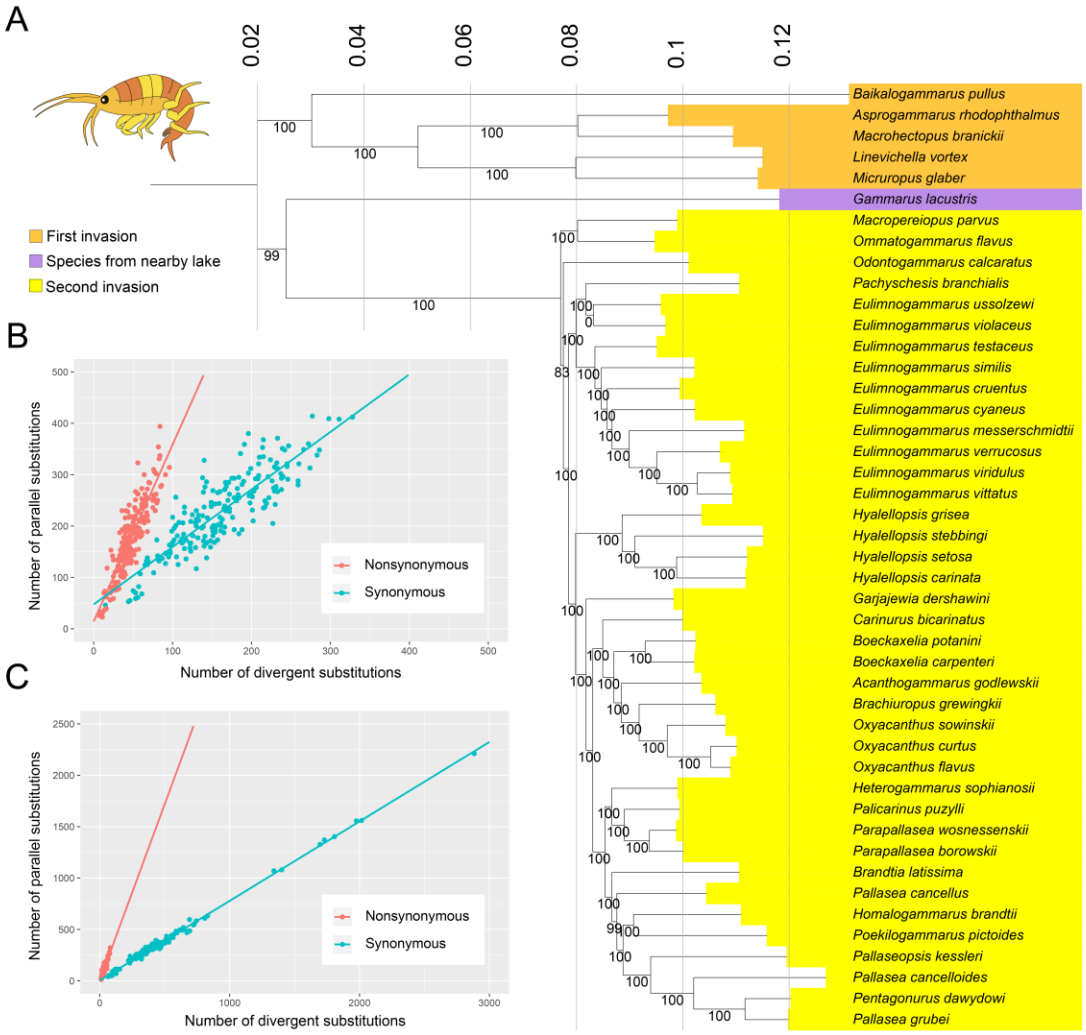


Fig. 3.3: Phylogenetics of Lake Baikal amphipods and analysis of parallelism. A: Phylogenetic tree of Lake Baikal amphipods. Values at inner nodes indicate bootstrap support. Two independent invasions are colored by orange (first invasion) and yellow (second invasion). Gammarus lacustris (sampled in a small lake adjacent, but not connected to Baikal) colored violet. B,C: Parallel vs divergent substitutions in amphipods. Each dot corresponds to one species quartet. B: quartets were selected only from the closely related group originating from the second invasion. C: quartets consist of two species from the first invasion and two species from the second invasion; these pairs are genetically more distant. Slopes of linear regression models

differ significantly (ANOVA test, p <2.2e-16) for synonymous and nonsynonymous substitutions at both pictures.

alternatives, we next compare the rate of nonsynonymous parallel evolution to that expected neutrally.

3.3.3. Parallel amino acid differences between amphipod species are more frequent than expected neutrally

In genes of compared amphipod species, the *dN/dS* ratio averaged over all sites and all quartets of compared species (n=195) equaled 0.11, consistent with previous findings (Naumenko *et al*. 2017) and in line with the universal prevalence of negative selection (Figure 3.4C). In stark contrast, the ratio of nonsynonymous and synonymous parallel substitutions *P* equaled 2.17, which implies that the rate of parallel nonsynonymous substitutions exceeded that of parallel synonymous substitutions by 117%. *P* exceeded 1 for 193 out of the 195 considered quartets of amphipod species (Figure 3.4B, Table A1). The excess of nonsynonymous parallel, compared to synonymous parallel, substitutions was observed for all types of mutations (AC, AT, AG, CT, CG and TG), and therefore is not due to heterogeneity of mutation rates (Figure 3.5, Table A1). The value of *P* was nearly independent of phylogenetic distance between the two species pairs (Table A4).

To confirm that our results are not an artifact of the NGS sequencing technology used, a subset of observed parallel sites were confirmed using Sanger technology. I randomly picked 3 sites of parallel nonsynonymous substitutions for each of the 2 species of amphipod: *Ommatogammarus albinus* (Dybowsky, 1874) and *Eulimnogammarus*

marituji (Bazikalova, 1945) (Table A8). The individuals used for resequencing were different from the ones used for the transcriptomic analysis. Sequences for all 6 samples were obtained; in 5 of these 6 cases, the only allele present was the one called for this species, and there was no evidence for the presence of an alternative allele. (In the remaining case, the alternative allele observed in *Eulimnogammarus marituji* coincided with a SNP observed in this species; see below).

To put these results into perspective, I considered other groups of species. Among the 100 species of vertebrates, the value of the *P* statistic is below 1 for all quartets (mean = 0.40, n = 147, Table A3), consistent with previous results (Bazykin *et al*. 2007; Zou and Zhang 2015b). In this group of species, *P* declines with phylogenetic distance between species (Figure 3.4B, Figure 3.5, Table A6). Among the 100 species of cichlids, *P* is close to 1 (mean = 1.13, n = 300; Figure 3.4B, Table A2). Similarly to the amphipod sample, it shows almost no dependence on phylogenetic distance (Table A5). The overall *dN/dS* ratio in vertebrates was below that in amphipods (0.02), and the *dN/dS* ratio in cichlids, above that in amphipods (0.34) (Figure 3.4C); the differences in these values may reflect data filtering, differences in effective population sizes leading to differences in efficiency of negative selection against slightly deleterious mutations (Nikolaev *et al*. 2007), and, perhaps more importantly, an excess of slightly deleterious unfixed variants in comparisons of closely related amphipods.

3.3.4. High amino acid parallelism shows no clear link with phenotypic parallelism

Species of Lake Baikal amphipods inhabit depths between 0 and >1500m. Reasoning that excess parallel evolution is more likely in those species which undergo similar adaptations to the same environment, I hypothesised that those quartets which include both deepwater and shallow water species in each path should have elevated values of the *P* statistic. However, I found that the data did not support this hypothesis: the values of the *P* statistic were similar to those of randomly selected 300 quartets (Figure 3.4D).

In addition, I analyzed the quartets with the highest and the lowest values of the *P* statistic, among the 300 quartets used in the main test. I hypothesised that the quartets with the high values of the *P* statistic will possess some phenotypic changes parallel between the two species pairs. However, these quartets demonstrated no clear excess of similarity between species of different pairs in any of the considered phenotypic traits (Table A11). Since the excess of the *P* statistic is rather uniform among amphipod quartets (Figure 3.4B) and is driven by a large number of genes, it is perhaps unsurprising that most of these parallel events have no obvious phenotypic manifestation.

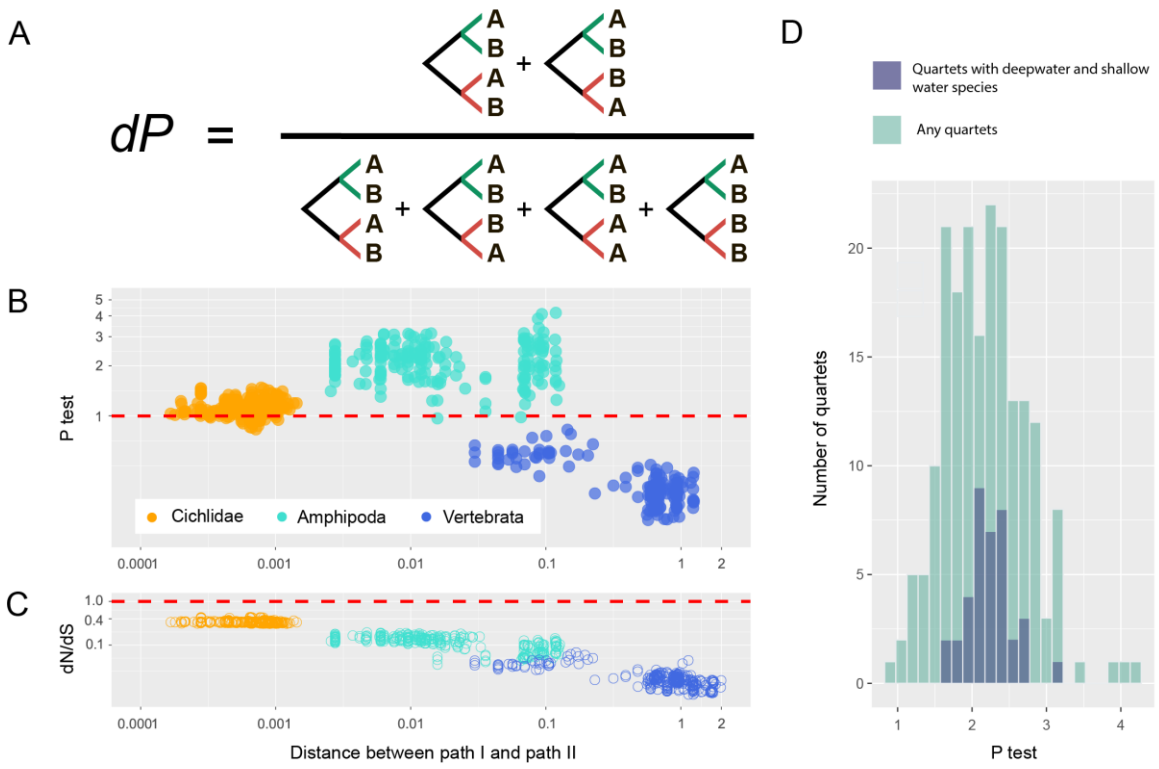


Fig. 3.4: The P test for excess parallelism. A: A schematic representation of the calculation of the level of parallelism d_P. In a quartet of species with the depicted topology, green and red colors identify evolutionary paths I and II respectively. d_P measures the probability that a substitution that has occurred along path I has also occurred along path II. P is the ratio of d_P values at nonsynonymous and synonymous sites. B: P for quartets of cichlids, amphipods and vertebrates. Each point represents the mean value of P for a species quartet across the 6 mutation types; only those 642 out of the 900 quartets for which sufficient data are available for each mutation type are shown (see Methods). The horizontal axis shows the phylogenetic distance between the last common ancestors of paths I and II, measured in number of substitutions per nucleotide site (note the logarithmic axis). C: dN/dS values for two species in path I, calculated for three groups of species. D: Distribution of values of the P statistic in all the 300 analyzed quartets (cyan) and in

the 40 quartets with evidence of parallel phenotypic adaptations to abyssal environment (blue). The mean values of the two distributions are similar (Wilcoxon test, p=0.2709).

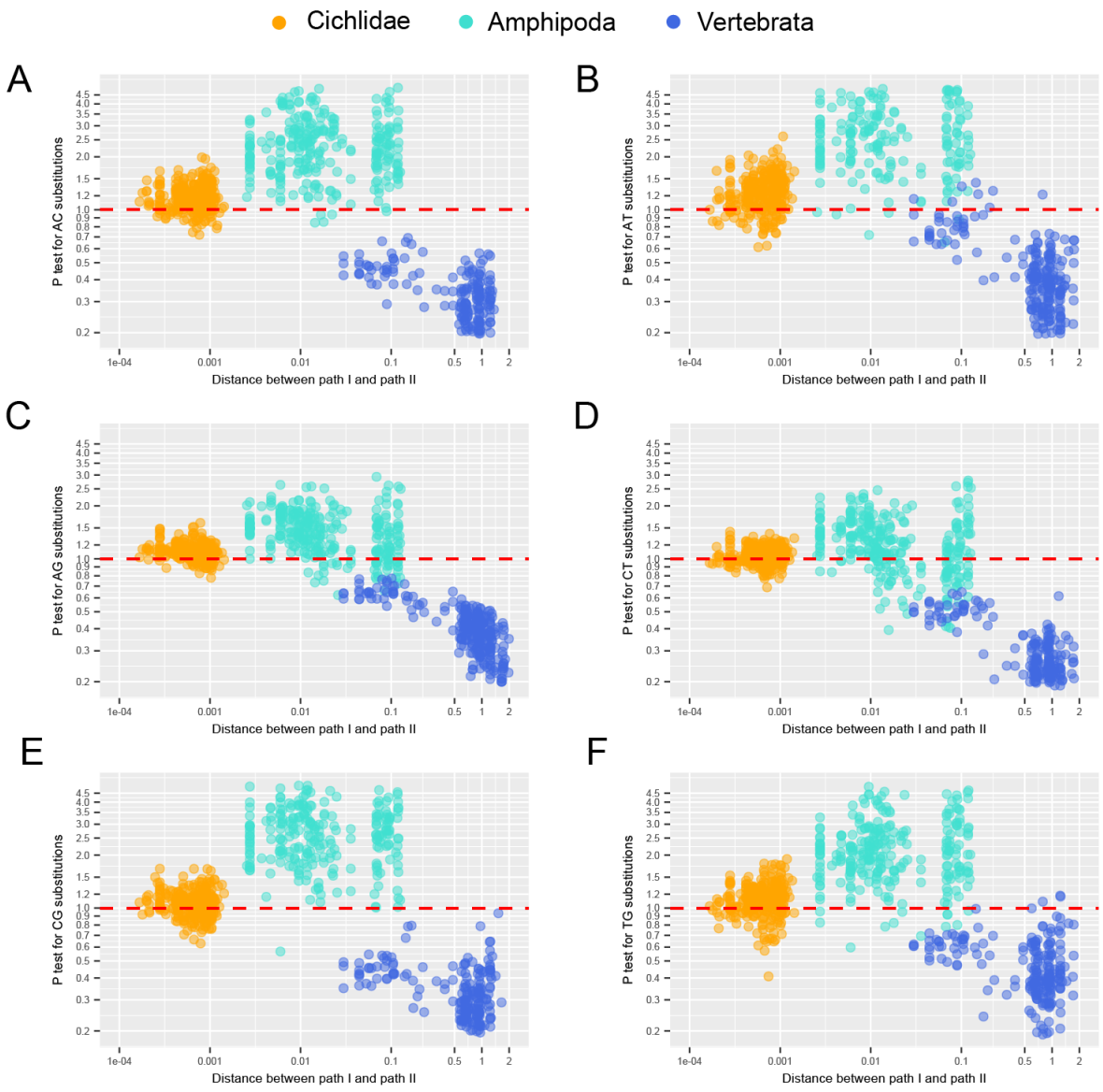


Fig. 3.5: P for quartets of cichlids, amphipods and vertebrates. Each panel represents a particular mutation type: AC (A), AT (B), AG (C), CT (D), CG (E) and TG (F). Each point represents the value of P for one species quartet; only those 642 out of the 900 quartets for which sufficient data

are available for each mutation type are shown (see Methods). The x axis corresponds to the distance between the last common ancestors of two species pairs (note the logarithmic axis).

3.3.5. Methodological factors that could affect P test values

To ensure that the *P* test calculation was not affected by experiment design, I calculated *P* values with different parameters. First, I were interested in how quartet selection in Cichlids may affect *P* test values. According to Malinsky *et al*. (2018), the phylogeny in different ecological groups of Malawi cichlids is quite unstable, as their speciation took place at last 10.000 years – a really short time for a species flock of hundreds of species. Meanwhile, the branching order of ecological groups was estimated as stable. In the main analysis, I selected species quartets randomly in order to use similar approaches in Amphipoda, Cichlidae and Vertebrata datasets. These quartets vary in the confidence of their topology. To ensure proper tree topology of a quartet, I separately analyzed those quartets for which the species of Path I were selected from one ecological group, while species of Path II were selected from another ecological group. Figure 3.6 compares the *P* tests for 300 randomly sampled species quartets (these data are included in the above analysis) and *P* tests for 300 species quartets sampled from different ecological groups. For both samples, the values of the *P* statistic are above one, indicating that the method of quartet sampling does not affect *P* test estimation seriously.

Another factor that could influence the results of the *P* test is the alignment filtering parameters. Although all filtering is designed to improve alignment quality, I checked Amphipoda results for possible side effects of filtering. In the main test, I only

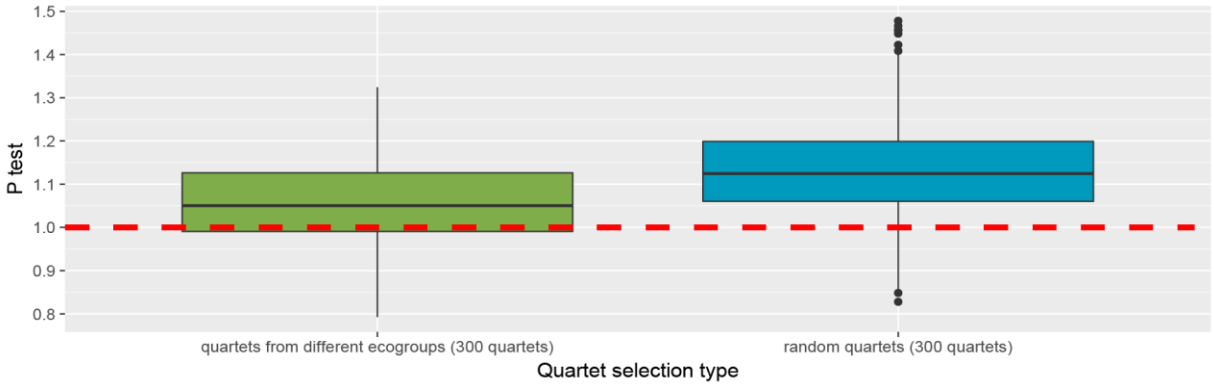


Fig. 3.6: P test for cichlids: both distributions have mean value higher than one according to two-tailed sign test (p = 1.998E-15 in quartets, sampled randomly, p < 2.2e-16 in quartets, sampled from different ecological groups).

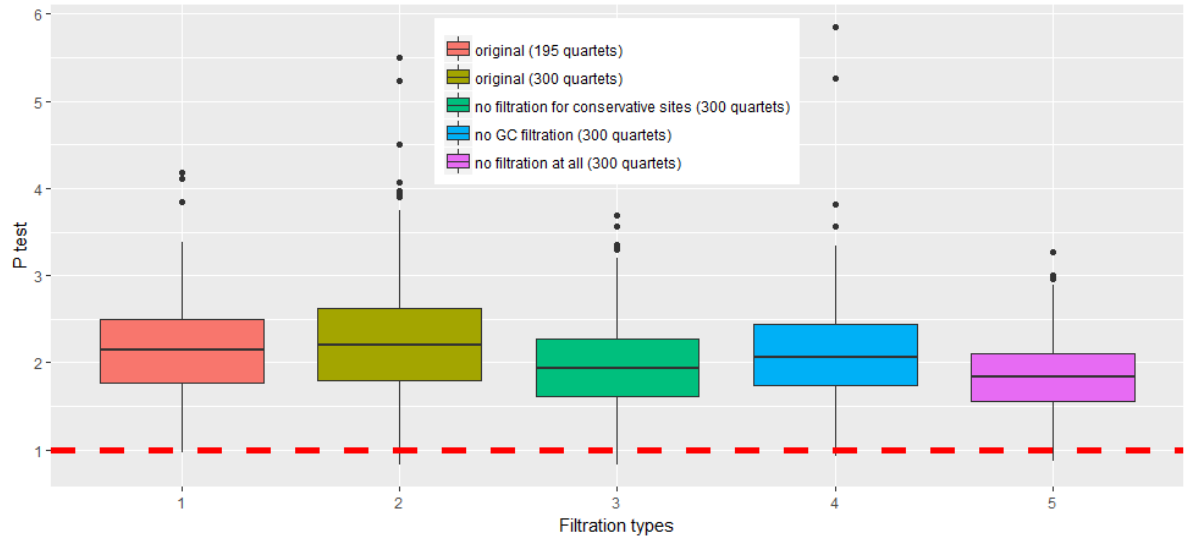


Fig. 3.7: Different filtering types do not affect the values of the P statistic in amphipods (all mean P-test values are around 2, 1st quartile is always over 1). The left box shows original data with all filters applied, while others combine different types of filtering relaxation. Pseudocounts were added to numerators and denominators used in calculations of all fractions for all boxes except the first one.

retained a quartet if the number of both synonymous and nonsynonymous parallel substitutions exceeded five for each of the 6 mutation types. This filtering kept just 195 out of 300 quartets for further analysis (Figure 3.7, 1st boxplot); the remaining quartets were excluded. To test if this filtering affects out conclusion, I designed an alternative dataset, in which I retained all 300 quartets, adding pseudocounts to the numerators and denominators of all ratios involved in calculation of the *P* statistic. In separate analyses, I also excluded other filters that were applied for our main dataset. Neither of these changes affected the result substantially (Figure 3.7, boxplots 2-5).

3.3.6. Among the sites with parallel changes between species, many are polymorphic within species

The observed differences between genomes or transcriptomes of different species could correspond to fixed differences between these species or to SNPs segregating within them. I asked whether the SNPs at sites of nonsynonymous parallel changes between species occur less or more often, compared to the analogous synonymous sites. A deficit of nonsynonymous SNPs is indicative of negative selection against one of the variants, while an excess of nonsynonymous SNPs can indicate balancing selection maintaining polymorphism at these sites.

To address this, I used within-species polymorphism data from two sources: heterozygous sites within the reference transcriptomes (individual-based polymorphism test) or SNPs detected in population samples of two amphipod species (population-based

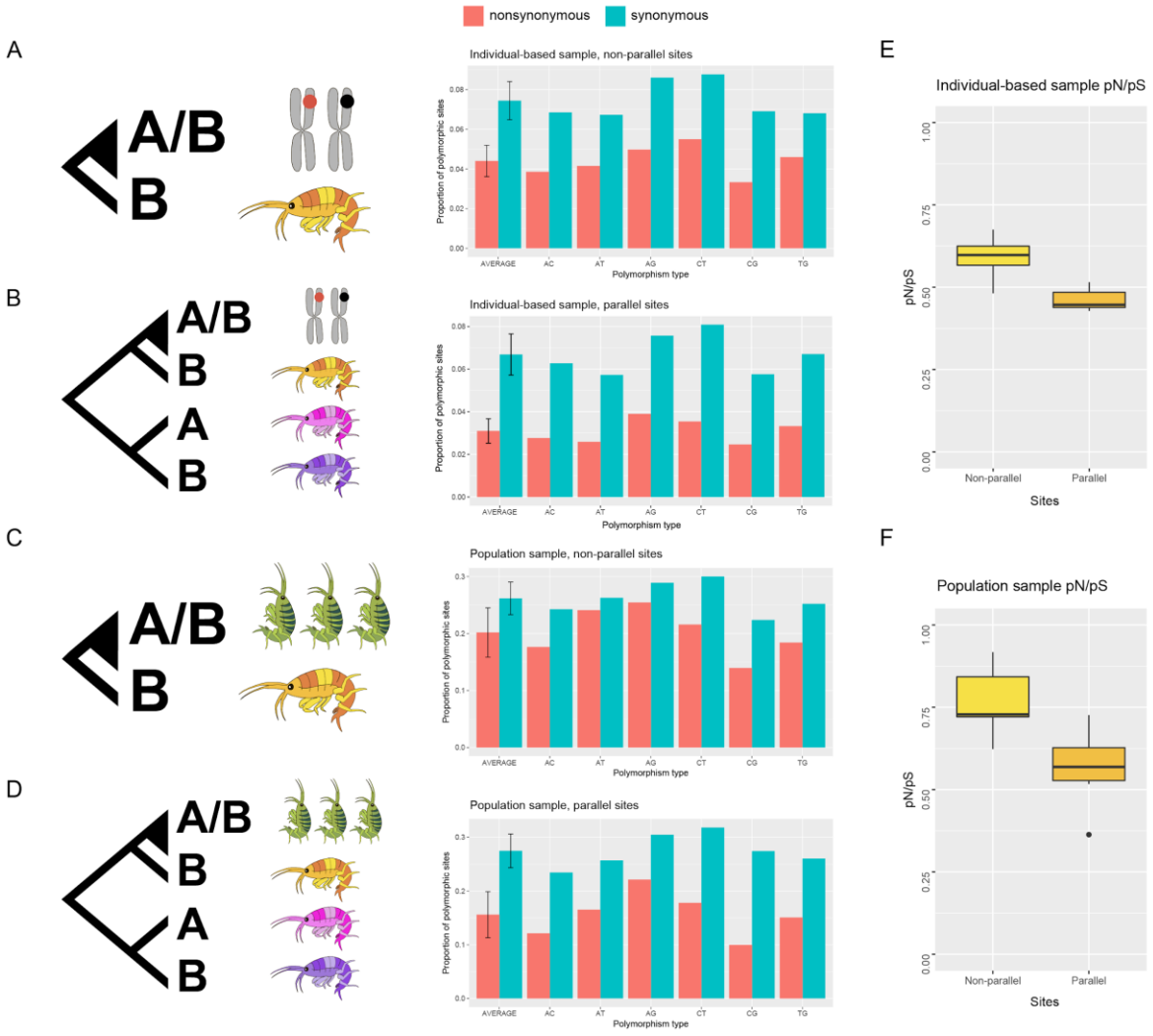


Fig. 3.8: Polymorphism at sites of parallel substitutions. A-D: Proportion of sites carrying a SNP among those with a nucleotide substitution between species in path I (A, C) or among sites with parallel substitutions (B, D). A,B: Individual-based polymorphism test. C,D: Population-based polymorphism test. E, F: Ratio of proportions of sites carrying a SNP between nonsynonymous and synonymous sites (pN/pS). E: Individual-based polymorphism test (paired t-test, p=0.0031; unpaired t-test, p=0.0033). F: Population-based polymorphism test (paired t-test, p=0.0091; unpaired t-test, p=0.0100).

polymorphism test). For all six possible mutation types, both tests showed that among nonsynonymous parallel sites, a lower fraction carried SNPs, compared to synonymous parallel sites (Figure 3.8B,D and Table A7).

To put this observation into perspective, I compared this deficit of SNPs at nonsynonymous parallel sites to that observed at nonparallel ones, i.e., those where a single substitution had occurred between the path I species, independent of the similarity between the path II species (Figure 3.8A,C). The values of the pN/pS statistic were lower at parallel than at nonparallel sites, suggesting stronger negative selection in the former. This was true both for individual-based (Figure 3.8E) and population-based (Figure 3.8F) tests.

To validate SNP calling, I included in our sample for Sanger resequencing (see above) one position that was polymorphic in the resequenced species (*Eulimnogammarus marituji*) according to the NGS data. Resequencing of that position also gave us a polymorphic site, with the same pair of nucleotides at the SNP (A/C; Table A8).

3.3.7. Parallel and non-parallel amino acid substitutions are similar in their properties

Among the 4366 analyzed genes of amphipods, 2514 (57.6%) carried a parallel nonsynonymous substitution in at least one of the species quartets. I was unable to detect any specific properties of these genes. According to the GO analysis, the set of genes with parallel nonsynonymous substitutions was indistinguishable from the remaining genes.

I also asked whether the parallel amino acid substitutions in amphipods differ in their properties from the substitutions that occurred in just one of the species. I suppose that if parallel substitutions are adaptive, some rare amino acid transitions, which are characterized by higher Miyata distance, could be more frequent, than those governed mainly by neutral evolution. For this, I compared the distribution of Miyata distances between those substitutions that occurred in paths I and II (parallel substitutions, which are presumably adaptive) and those that only occurred in path I (non-parallel substitutions, which are mainly neutral). The two distributions were indistinguishable (X-squared goodness of fit test, X-squared = 2019.8, df = 55, p < 2.2e-16; Figure 3.9).

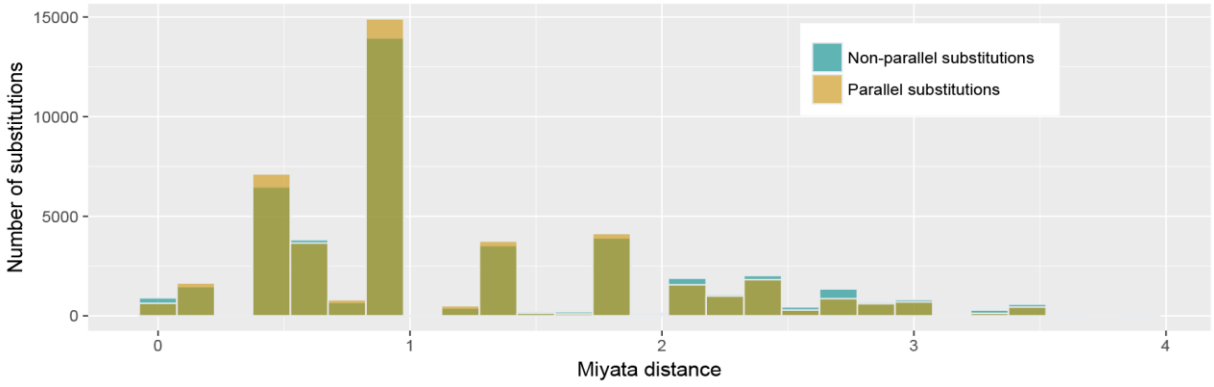


Fig. 3.9: Distribution of Miyata distances in parallel and non-parallel substitutions.

3.3.8. Contribution of homoplasy and hemiplasy in parallel evolution

The observed excess of parallel nonsynonymous evolution in amphipods could be a result of homoplasy or hemiplasy. Importantly, this distinction does not affect our main conclusion that adaptation via positive or balancing selection needs to be invoked to

explain $P>1$. Having said that, I tested whether hemiplasy or homoplasy is the main contributor to P in our analysis.

Hemiplasy usually affects contiguous regions of DNA and thus changes the phylogeny of specific genes or any longer closely located regions. Even good bootstrap values of phylogenetic reconstruction cannot give a guarantee that gene phylogeny does not differ from that based on all available sequences. The most straightforward way to check if observed parallelism could be result of hemiplasy is construction of phylogenies for particular genes. To estimate the possible contribution of hemiplasy to our inference, I plot phylogenies of 74 genes (each of more than 1000 nucleotides length) together at Figure 3.10. There is no consensus in deep branchig order of Eulimnogammarus clade. However, the tree at Figure 3.10 is ultrametric and thus does not show how short the internal branches are. In fact, deep internal branches in Eulimnogammarus clade have lengths of 0.1% or similar (Figure 3.3A). Obviously, orthologous groups of length 1000 are not sufficient for resolving branching that is so fast. Terminal branches in Eulimnogammarus group have lengths of 2%, which can be considered as sufficient for incongruence estimation. Topology of terminal branches seems to be also quite unstable, and thus hemiplasy may be a source of adaptive parallel substitutions, which were discovered by the P test in close species quartets. To estimate the amount of hemiplasy more accurately, long continuous genomic sequences are needed.

However, in addition to close amphipod species quartets I also analyzed species pairs, which belong to two independent invasions into Baikal lake. In these groups the maximum distance between path I and path II in amphipod species quartets exceeds 10%,

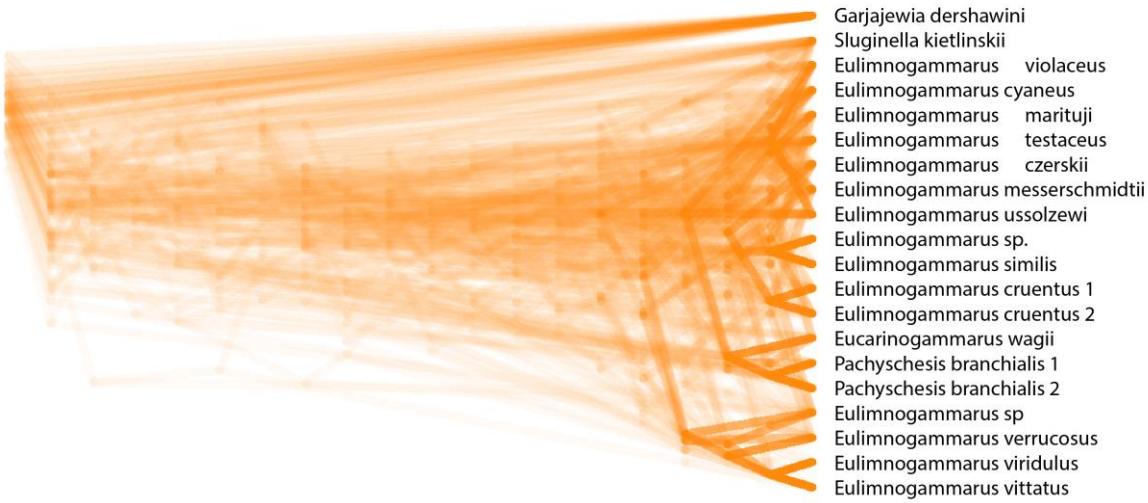


Fig. 3.10: Incongruence in phylogenetic reconstructions of Eulimnogammarus clade. Cladograms are based on 74 orthologous groups, each of more than 1000 nucleotide length.

and we observe $P>1$ even for that remote species pairs. At these distances, hemiplasy seems to be an impossible scenario, as neutral polymorphism cannot persist that long and hybridization is also unlikely (Jančúchová-Lásková *et al*. 2015, Fitzpatrick 2004). It implies that homoplasy also contributed to adaptive parallel evolution of baikalian amphipods at least at higher phylogenetic distances.

The origin of parallel substitutions in cichlid quartets should be discussed separately. In cichlids, the P test stays slightly higher than one no matter how species quartets are selected (Figure 3.6), and although comparing to amphipods the pattern is much weaker, it merits attention. Malinsky *et al*. (2018) demonstrates that the phylogeny of Malawi cichlids reconstructed on different genomic regions is substantially different and thus hemiplasy is widespread in the species flock. I also discovered one more evidence that all discovered adaptive parallelisms are result of hemiplasy. I compared

amount of parallel and divergent evolution in cichlids by exactly the same method as was used on amphipods (Figures 3.3B and 3.3C). Surprisingly, I found that in a quartet of cichlid species (300 quartets from main *P* test sample was analyzed) there is on average 762 synonymous parallel substitutions, 976 nonsynonymous parallel substitutions and 0 divergent substitutions (both synonymous and nonsynonymous). The absence of divergent mutations indicates that independent parallel substitutions also likely had no time to emerge, and thus are result of hemiplasy.

3.4. Discussion

If the substitution rate is uniform between sites and invariable in time, the probability that a substitution occurs is independent of whether the same substitution has occurred at the same position in a related genome. However, multiple processes can make the substitution rates heterogenic between genomic sites, and such heterogeneity should inflate the observed rate of parallel substitutions at orthologous sites. Specifically, the rate of a parallel substitution, compared to a non-parallel one, can be elevated even in the absence of selection due to differences between genomic sites in point mutation rates (Hodskinson and Eyre-Walker 2011, Seplyarskiy *et al*. 2012). Besides, differences in strength of negative selection between genomic regions should result in an excess of substitutions at those regions were this selection is relaxed, and if this difference in selection pressures is conserved between species, this should lead to an excess of cases whereby the same (beneficial, neutral or even deleterious) substitution has occurred in

parallel, compared to the expectation formulated without regard to selection heterogeneity.

Our approach of comparing the rates of nonsynonymous and synonymous parallel substitutions, while accounting for the mutation type, aims to single out the effect of parallel adaptation, mostly controlling for these confounding effects. Indeed, mutational heterogeneity should equally affect synonymous and nonsynonymous sites, which is the logic underlying the conventional dN/dS test as well as its derivatives such as the McDonald-Kreitman test (McDonald and Kreitman 1991, Smith and Eyre-Walker 2002). Both in the conventional dN/dS test and in the *P* test, a deficit of nonsynonymous substitutions compared to the synonymous control implies negative selection preventing their fixation. By contrast, in the conventional dN/dS test, an excess of nonsynonymous substitutions implies positive selection in their favor; similarly, in the P test, $P>1$ suggests selection in favor of the substitutions parallel to those that also occurred at another lineage (Bazykin *et al*. 2007).

Consistent with previous results using few species, I find $P\sim0.8$ in vertebrates. By contrast, and strikingly, I observe a genome-wide $P>1$ in amphipods, over all substitution types. The cichlids demonstrate an intermediate pattern ($P\sim1$).

What is the cause of high nonsynonymous parallelism in amphipods? While high parallelism could potentially arise artifactually, most sources of artifacts would equally affect synonymous and nonsynonymous sites, and it is difficult to think of a mechanism that could lead to $P>1$. Conceivably, some of the observed parallel differences between species could arise from artifactual misassembly of paralogous segments of DNA if

different copies of the paralogs are included in analyses for different species. Since I use transcriptomics data, similar problems can also result from alternative splicing involving the segment covering the parallel site. Such artifacts, however, should affect all categories of sites equally, and are not expected to lead to the observed excess of nonsynonymous parallel substitutions. The presence of just a single variant in the species DNA, coinciding with the variant determined by transcriptomics sequencing, was also confirmed by Sanger resequencing. Additionally, cross-sample contamination is made unlikely by the fact that different individuals were used for DNA and transcriptome sequencing. While some of the observed parallelisms could still be caused by artifactual assembly, it is not clear how it could lead to $P>1$ observed in our data.

Phylogenetic patterns consistent with parallel evolution can arise artifactually due to survival of ancestral polymorphism throughout the time between subsequent divergence events (incomplete lineage sorting), hybridisation or errors in phylogenetic reconstruction. This leads to hemiplasy, i.e., the situation when two alleles that seem to have originated in parallel are in fact identical by descent (Hahn and Nakhleh 2016). Yet these patterns are generally not expected to lead to $P>1$, unless they are accompanied by selection in favor of new variants. It may easily be the case in quartets of closely related amphipods and cichlids. One of conceivable scenario involving a hemiplasy that could result in $P>1$ is adaptive hybridization, i.e., hybridization followed by preferential fixation of introgressed loci. Adaptive hybridization has been proposed as a mechanism for species flock emergence, as it increases within-species diversity facilitating further adaptation (Seehausen 2004); and reticular speciation has been observed in other

crustaceans (daphnia, Giessler *et al*. 1999). Another likely scenario is independent fixation of ancestral polymorphisms in different linages, driven by positive selection. However, the maximum distance between path I and path II in some amphipod species quartets exceeds 10%, while we observe $P>1$ even for species that are that remote. At these distances hybridization is very unlikely (Jančúchová-Lásková *et al*. 2015, Fitzpatrick 2004), and ancestral polymorphism also cannot survive that long.

By exclusion, our observation of $P>1$ requires parallel selection favoring the novel variant. There are two potential mechanisms for it: long-term balancing selection and recurrent episodes of positive selection. Balancing selection may increase the lifetime of two alleles cosegregating at a site, and their allele frequencies, over that at neutral (synonymous) sites; at a fraction of sites, the alleles observed in the reference genomes of the four compared species will differ between species according to the pattern in Figure 3.4A, and this will increase P compared to that at synonymous sites. Alternatively, $P>1$ can arise from positive selection favoring the novel adaptive variant at least at two of the four compared species.

There are aspects of data that argue against both these options. On the one hand, while many of the observed parallel nonsynonymous sites are polymorphic within a species, the level of polymorphism is lower than that at parallel synonymous sites or non-parallel nonsynonymous sites, while balancing selection is expected to increase the within-site diversity, compared to neutral. The deficit of polymorphism suggests ongoing negative selection in one species pair in favor of the variant also acquired in the other; together with the $P>1$, this implies that selection that had favored this acquisition was

positive. On the other hand, the presence of some polymorphism is inconsistent with the action of positive selection, whereby positive selection should rid sites of polymorphism. The available data is not sufficient to distinguish between all alternative sources of adaptive parallel evolution. Moreover, we see no systematic directionality in the amino acid substitutions (Figure 3.9) or preference for particular gene categories, arguing against a genome-wide trend in adaptation, e.g. towards a change in protein thermostability; although this doesn't exclude the possibility that the parallel adaptation in this system is promiscuous with respect to the substitutions it favors or genes that it affects.

Detection of adaptation from molecular data is complicated by the fact that it is hard to distinguish from the background of neutral and deleterious mutations. Analyzing substitutions between closely related species alleviates this problem, because at low phylogenetic distances, neutral (or slightly deleterious) substitutions have not yet had sufficient time to accumulate (Wolf *et al*. 2009, Stolyarova *et al*. 2019). The observation of $P>1$ in the flock of closely related amphipod species, but not in more distantly related vertebrates, as well as the decrease in P with phylogenetic distance in vertebrates (Figure 3.4B), is consistent with this explanation. Still, in a group of even more closely related cichlid fish, P is lower than that in amphipods; and we see no dependence of P on phylogenetic distance in amphipods or cichlids (Figure 3.4B). Moreover, vertebrate and amphipod species at close phylogenetic distances demonstrate very different values of P (which are much higher in amphipods). In total, while dependent on phylogenetic distance in one of the groups (vertebrates), P is more strongly determined by the identity

of the analyzed group (Figure 3.4B). Therefore, the enormous observed amount of nonsynonymous parallel evolution could be a specific feature of the baikalian amphipod species flock.

Speaking about causes of excessive parallelism, I conclude that there could be no reasons for it, except different types of adaptive evolution. Even if limitations of the fitness landscape change permitted substitution directions in parallel, *P* test will give values lower than one unless new states are preferable for selection (and thus adaptive).

Except adaptation to the environment, discovered adaptive parallel evolution could be caused by epistatic interactions. For example, if some slightly deleterious allele, inherited from the common ancestor, needs a particular compensatory mutation – it could emerge in parallel and be fixed, as a new state is preferable. Of particular interest could be compensation of Dobzhansky-Muller incompatibilities, if these incompatibilities emerged in the last common ancestor of all species. Similarly, if some substitution can enhance fitness of species due to optimization of epistatic interactions – it could be gained in parallel if an alternative state is preferrable. All these scenarios include positive selection and would cause $P>1$.

More extensive, preferably whole-genome, data on within- and between-species variation is needed to clarify cause of excessive nonsynonymous parallelism. Such data, for example, could provide evidence of selective sweeps from reduced polymorphism in the vicinity of parallel sites, which would constitute an independent signature of positive selection. More generally, parallel adaptation may be not as rare as it seems, and worth a systematic survey in many groups, particularly those of closely related species.

Chapter 4. Convergent adaptation in mitochondria of phylogenetically distant birds: does it exist?

4.1. Introduction

At the last decade, many attempts have been made to find single-position adaptive convergence in phylogenetically distant species. Same parallel adaptations were discovered and refuted many times. The most of the discussions revolved around similar echolocating adaptations in bats and dolphins and about convergent adaptations in marine mammals. Many works were refuted (Parker *et al*. 2013, Foote *et al*. 2015), many methods changed (Chikina *et al*. 2016), and finally the discussion leaves an impression that only neutral single-position convergence do exist in phylogenetically distant species. At the same time, it seems strange that convergent echolocation and marine adaptations attracted so much attention, while many other perspective phenotypic convergences are not studied. Recent work of Natarajan *et al*. (2016) describes broad study of hemoglobin adaptation to high altitudes in close and distant species of birds. They demonstrated, that close species develop similar substitutions for high-altitude adaptations, while in distant species mutations are different. I decided that as bird mitochondrial genes are broadly sequenced, adaptations of wildlife species are diverse, and science methodology teaches us to make multiple hypothesis testing, I should make one more try to find single-position parallelism in diverse species. In this work, by analyzing phylogenetically recurrent

substitutions in distant species, I quantify the occurrence of single-position parallel adaptations in the natural environment.

The history of mitochondrial adaptive changes goes back decades (Toews *et al*. 2014). Mitochondrial genes were repeatedly claimed to adapt in response to lifestyles that require oxygen starvation or elevation of metabolism rate in a wide range of eukaryotic taxa (Das 2006). The most obvious candidate species for evolution of hypoxia tolerance are those inhabiting high altitudes. Indeed, high-altitude adaptations have been described for mitochondrial genes of many species, including the COX3 gene of the bar-headed goose (Scott *et al*. 2011), ATP6 and ATP8 in shrimps from genus *Artemia* (Zhang *et al*. 2013), ND5 in caterpillars of genus *Gynaephora* (Yuana *et al*. 2018), COX1 in Tibetan antelope *Pantholops hodgsonii* (Xu *et al*. 2005), and ND2, ND4, and ATP6 in Tibetan galliform birds (Zhou *et al*. 2014). Adaptations to high altitude in human populations typically involve mutations in ND1 as well as probably other genes (Kang *et al*. 2013, Ji *et al*. 2012). Besides high-altitude hypoxia, some rodents likely developed adaptations to subterranean hypoxia (Tomasco and Lessa 2011).

Life at extreme temperatures and extraordinary physical activity could also cause adaptive evolution in mitochondrial genes through their effect on energy metabolism. It has been shown that genes involved in oxidative phosphorylation could take part in adaptation to arctic environment: adaptations in *ND1*, *ND3* and *ND4* genes were described in the Atlantic salmon (Consuegra *et al*. 2015), and adaptations in *CYTB* gene were found in European anchovy (Silva *et al*. 2014). Furthermore, there is evidence for

adaptation to long-range migrations in the yellow-rumped warbler (Toews *et al*. 2014). Flight is another energy-consuming adaptation, and some studies confirm adaptation of mitochondrial genes to flight in bats (Shen *et al*. 2010).

Most evidence for selection in mitochondria is indirect. It often comes down to description of differences in frequencies of a few alleles, which could be a consequence of random drift in small populations. Positions discovered in different studies rarely overlap, suggesting that different organisms use different mechanisms of OXPHOS system adaptation to similar environments, or that some of the findings are erroneous. Still, some gene regions are more likely to be affected by adaptation (da Fonseca *et al*. 2008). Furthermore, the role of mitochondrial electron transport chain in physiological acclimation was demonstrated in many experimental studies. For example, gene expression, protein abundance, and the enzyme activity changes in plants and animals in process of cold acclimation (Armstrong *et al*. 2008, Lucassen *et al*. 2006).

To study this systematically, I decided to conduct a broad search for adaptive convergent evolution in mitochondria of birds in an attempt to find universal genotype-to-phenotype associations. I concentrated on mitochondrial adaptations in bird species that likely face hypoxia (high altitude and diving) or requirement for elevated (long-distance migration, wintering at high latitudes and unusual flight abilities) or reduced (loss of flight) rate of metabolism, hypothesizing that detected adaptations may confer resistance to these types of physiology.

To estimate potential convergence, I develop upon an existing phylogenetic test for detection of parallel adaptation (TreeWAS package, Collins and Didelot 2018). This test is based on reconstruction of ancestral states in the internal nodes of the phylogeny, and then counting the number of coincident changes of phenotype and genotype at each amino acid site at the same branch of the phylogenetic tree. Additionally, at each site, I measure the change of amino acid propensities associated with phenotype change (PCOC package, Rey *et al*. 2018). While I detect a number of candidate sites that could be associated with convergent adaptation to high altitude and long-distance migration in birds, I only detect one site that was associated according to both tests, and this association was borderline significant. Overall, I find little or no signal of recurrent adaptation, indicating that adaptation to extreme physiology in birds can proceed via different routes in different species, and/or that it can be largely driven by non-mitochondrially encoded genes.

4.2. Materials and Methods

4.2.1. Phenotypes

I analyzed 415 species of birds. All species were divided into 7 groups according to their phenotypic characteristics. Phenotypes were classified in accordance with the Bird of the World research database (https://birdsoftheworld.org, accessed August 20, 2020). As high altitude I classified those species for which the lower boundary of the range was over 2000 meters, and the upper boundary of the range, over 4000 meters above sea level. As divers I classified those species which can spend at least several

minutes underwater. As species with ability for long-distance migration I considered those species with non-overlapping or weakly overlapping breeding and wintering ranges. As wintering I classified those species which are typically exposed to sub-zero temperatures and snow cover for many months each year. I also formed two samples of species with specific flight-related phenotypes: flightless birds, a phenotype which has originated repeatedly in different groups of island birds; and birds with outstanding flight abilities. The latter group united swifts, hummingbirds, swallows, terns and gulls, scuas, gannets, tropicbirds, falcons and accipiters. Although the similarity of these adaptations may be controversial, I hypothesize that the lifestyles of all these groups involve high energy demand and thus could affect the mitochondrial genes similarly.

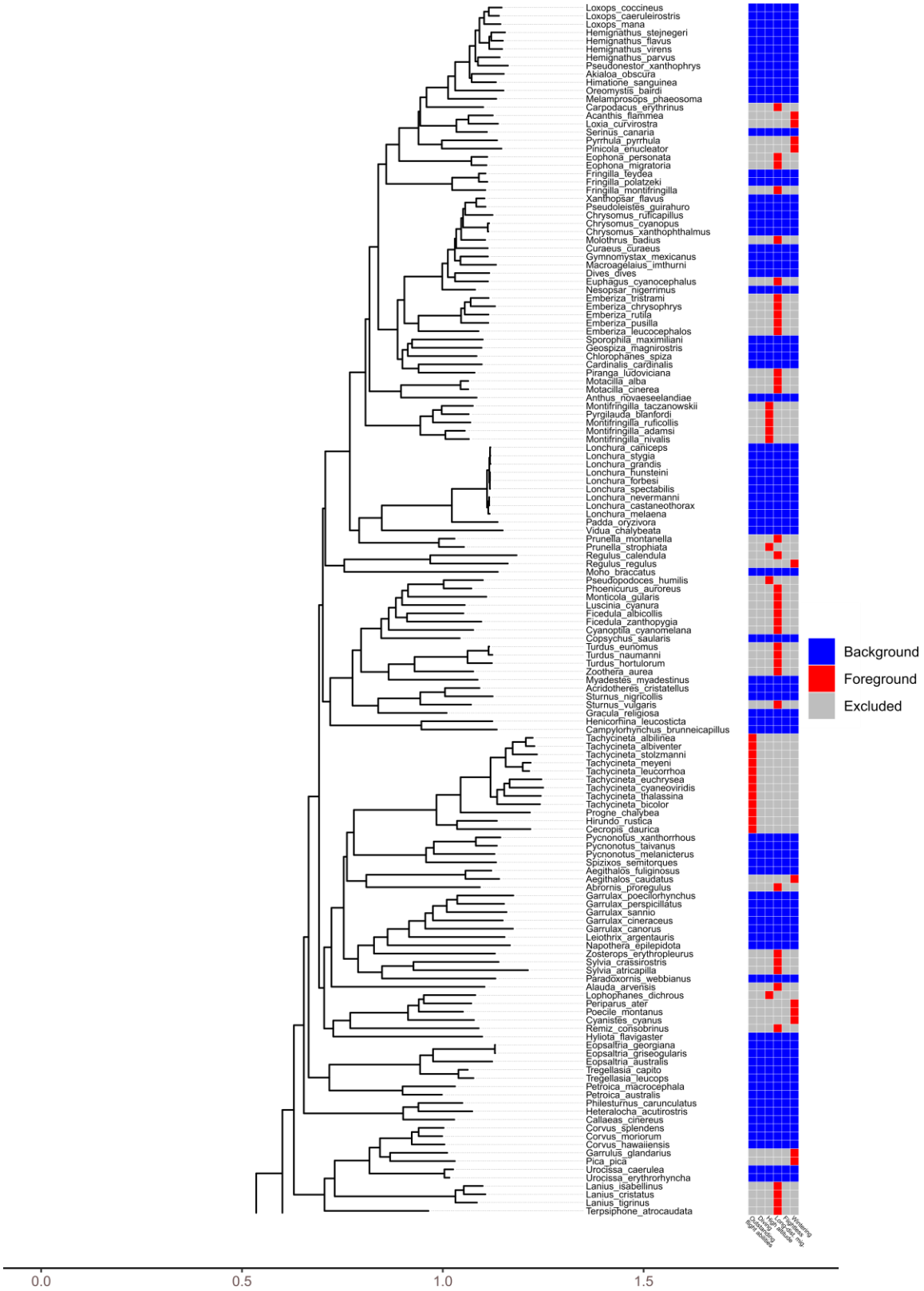
To study phenotypic associations, I also need a reference group of species which do not carry the specific adaptations considered in this work. The choice of such a reference is a complicated task, because of the complexity of natural ecological adaptations. As the reference group, I decided to use tropical and subtropical birds with ranges not extending above 2500 meters and which have none of the specific aforementioned adaptations. The number of species in each phenotypic group is provided in Table 4.1, and the list of species is provided in Figure 4.1.

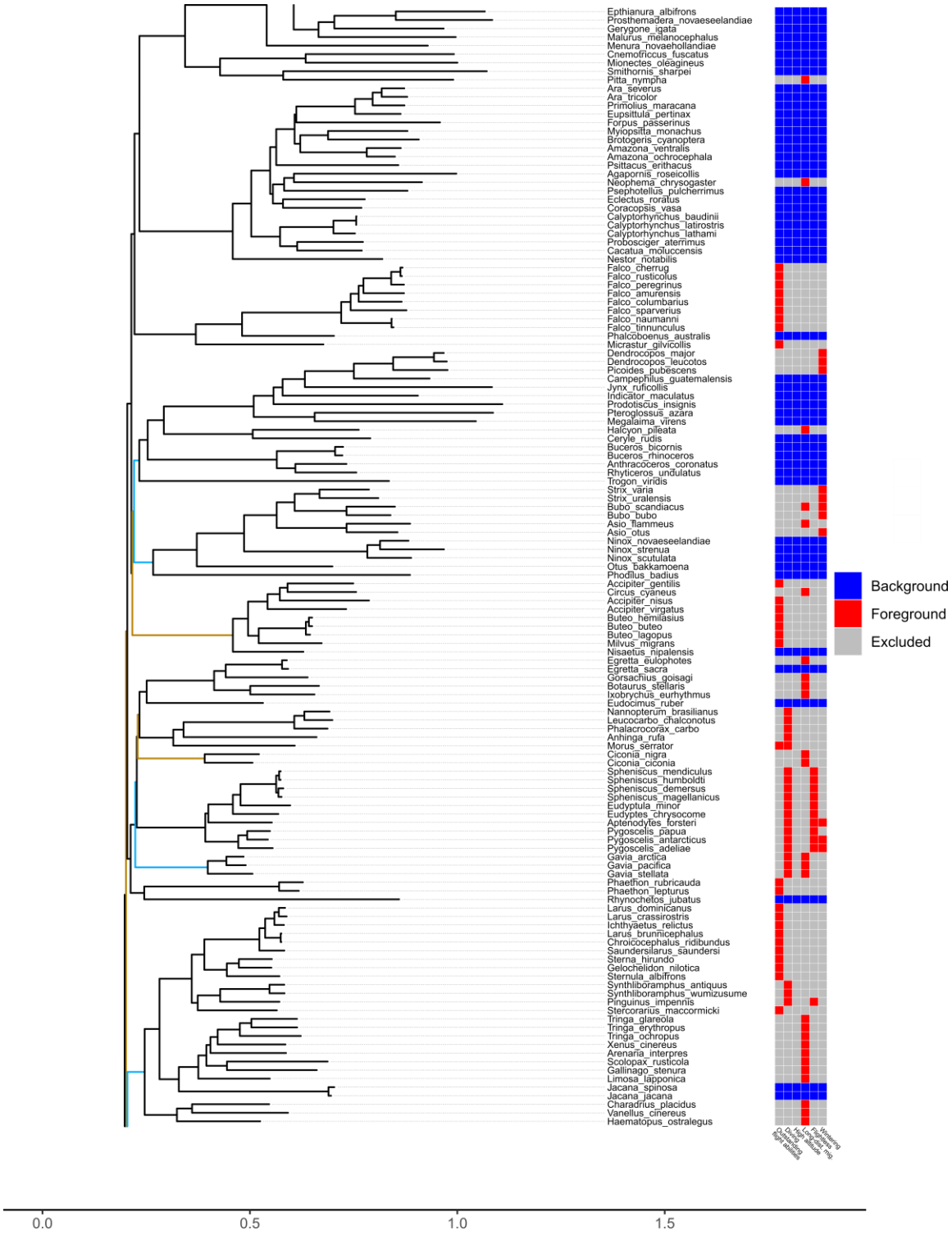
Table 4.1: Number of species in each phenotypic group.

	Number of species	Number of times the phenotype emerged independently
High altitude	23	7
Diving	25	3
Long distance migration	91	33
Wintering	28	11
Flightless	33	6
Outstanding flight abilities	58	7
Reference	174	-

4.2.2. Gene sequences and phylogeny

Complete mitochondrial sequences of birds were downloaded from Genbank (https://www.ncbi.nlm.nih.gov/genbank/, accessed August 20, 2020). The sequence of each of the 13 genes was obtained according to the GenBank annotation. Species with duplicated genes were excluded from analysis. Sequences of each gene were aligned independently with MACSE toolkit (Ranwez *et al*. 2011). Columns of alignment with gaps were excluded by trimAl software (Capella-Gutierrez *et al*. 2009). Phylogenetic analysis was performed in IQ-Tree 2 package (Minh *et al*. 2020) based on nucleotide alignment, split into 39 partitions by gene and codon position. As early divergences in the bird tree of life are discordant (Jarvis *et al*. 2014), I used constraints for bird orders branching in our reconstruction. I applied constraints from the most recent revision of bird phylogeny (Kimball *et al*. 2019), which combines nuclear and mitochondrial data to construct a consensus supertree for 707 bird species. Amino acid ancestor states


Loxops_coccineus
Loxops_caeruleirostris
Loxops_mana
Hemignathus_stejnegeri
Hemignathus_flavus
Hemignathus_virens
Hemignathus_parvus
Pseudonestor_xanthophrys
Akialoa_obscura
Himatione_sanguinea
Oreomystis_bairdi
Melamprosops_phaeosoma
Carpodacus_erythrinus
Acanthis_flammea
Loxia_curvirostra
Serinus_canaria
Pyrrhula_pyrrhula
Pinicola_enucleator
Eophona_personata
Eophona_migratoria
Fringilla_teydea
Fringilla_polatzeki
Fringilla_montifringilla
Xanthopsar_flavus
Pseudoleistes_guirahuro
Chrysomus_ruficapillus
Chrysomus_cyanopus
Chrysomus_xanthophthalmus
Molothrus_badius
Curaeus_curaeus
Gymnomystax_mexicanus
Macroagelaius_imthurni
Dives_dives
Euphagus_cyanocephalus
Nesopsar_nigerrimus
Emberiza_tristrami
Emberiza_chrysophrys
Emberiza_rutila
Emberiza_pusilla
Emberiza_leucocephalos
Sporophila_maximiliani
Geospiza_magnirostris
Chlorophanes_spiza
Cardinalis_cardinalis
Piranga_ludoviciana
Motacilla_alba
Motacilla_cinerea
Anthus_novaeseelandiae
Montifringilla_taczanowskii
Pyrgilauda_blanfordi
Montifringilla_ruficollis
Montifringilla_adamsi
Montifringilla_nivalis
Lonchura_cantans
Lonchura_stygia
Lonchura_grandis
Lonchura_hunsteini
Lonchura_forbesi
Lonchura_spectabilis
Lonchura_nevermanni
Lonchura_castaneothorax
Lonchura_melaena
Padda_oryzivora
Vidua_chalybeata
Prunella_montanella
Prunella_strophiata
Regulus_calendula
Regulus_regulus
Moho_braccatus
Pseudopodoces_humilis
Phoenicurus_auroreus
Monticola_gularis
Luscinia_cyanura
Ficedula_albicollis
Ficedula_zanthopygia
Cyanoptila_cyanomelana
Copsychus_saularis
Turdus_eunomus
Turdus_naumanni
Turdus_hortulorum
Zoothera_aurea
Myadestes_myadestinus
Acridotheres_cristatellus
Sturnus_nigricollis
Sturnus_vulgaris
Gracula_religiosa
Henicorhina_leucosticta
Campylorhynchus_brunneicapillus
Tachycineta_albiliinea
Tachycineta_albiventer
Tachycineta_stolzmanni
Tachycineta_meyeni
Tachycineta_leucorrhoa
Tachycineta_euchrysea
Tachycineta_cyaneoviridis
Tachycineta_thalassina
Tachycineta_bicolor
Progne_chalybea
Hirundo_rustica
Cecropis_daurica
Pycnonotus_xanthorrhous
Pycnonotus_taivanus
Pycnonotus_melanicterus
Spizixos_semitorques
Aegithalos_fuliginosus
Aegithalos_caudatus
Abrornis_proregulus
Garrulax_poecilorhynchus
Garrulax_perspicillatus
Garrulax_sannio
Garrulax_cineraceus
Garrulax_canorus
Leiothrix_argentauris
Napothera_epilepidota
Zosterops_erythropleurus
Sylvia_crassirostris
Sylvia_atricapilla
Paradoxornis_webbianus
Alauda_arvensis
Lophophanes_dichrous
Periparus_ater
Poecile_montanus
Cyanistes_cyanus
Remiz_consobrinus
Hyliota_flavigaster
Eopsaltria_georgiana
Eopsaltria_griseogularis
Eopsaltria_australis
Tregellasia_capito
Tregellasia_leucops
Petroica_macrocephala
Petroica_australis
Philesturnus_carunculatus
Heteralocha_acutirostris
Callaeas_cinereus
Corvus_splendens
Corvus_moriorum
Corvus_hawaiiensis
Garrulus_glandarius
Pica_pica
Urocissa_caerulea
Urocissa_erythrorhyncha
Lanius_isabellinus
Lanius_cristatus
Lanius_tigrinus
Terpsiphone_atrocaudata
Background
Foreground
Excluded
0.0
0.5
1.0
1.5


Epthianura_albifrons
Prosthemadera_novaeseelandiae
Gerygone_igata
Malurus_melanocephalus
Menura_novaehollandiae
Cnemotriccus_fuscatus
Mionectes_oleagineus
Smithornis_sharpei
Pitta_nympha
Ara_severus
Ara_tricolor
Primolius_maracana
Eupsittula_pertinax
Forpus_passerinus
Myiopsitta_monachus
Brotogeris_cyanoptera
Amazona_ventralis
Amazona_ochrocephala
Psittacus_erithacus
Agapornis_roseicollis
Neophema_chrysogaster
Psephotellus_pulcherrimus
Eclectus_roratus
Coracopsis_vasa
Calyptorhynchus_baudinii
Calyptorhynchus_latirostris
Calyptorhynchus_lathami
Probosciger_aterrimus
Cacatua_moluccensis
Nestor_notabilis
Falco_cherrug
Falco_rusticolus
Falco_peregrinus
Falco_amurensis
Falco_columbarius
Falco_sparverius
Falco_naumanni
Falco_tinnunculus
Phalcoboenus_australis
Micrastur_gilvicollis
Dendrocopos_major
Dendrocopos_leucotos
Picoides_pubescens
Campephilus_guatemalensis
Jynx_ruficollis
Indicator_maculatus
Prodotiscus_insignis
Pteroglossus_azara
Megalaima_virens
Halcyon_pileata
Ceryle_rudis
Buceros_bicornis
Buceros_rhinoceros
Anthracoceros_coronatus
Rhyticeros_undulatus
Trogon_viridis
Strix_varia
Strix_uralensis
Bubo_scandiacus
Bubo_bubo
Asio_flammeus
Asio_otus
Ninox_novaeseelandiae
Ninox_strenua
Ninox_scutulata
Otus_bakkamoena
Phodilus_badius
Accipiter_gentilis
Circus_cyaneus
Accipiter_nisus
Accipiter_virgatus
Buteo_hemilasius
Buteo_buteo
Buteo_lagopus
Milvus_migrans
Nisaetus_nipalensis
Egretta_eulophotes
Egretta_sacra
Gorsachius_goisagi
Botaurus_stellaris
Ixobrychus_eurhythmus
Eudocimus_ruber
Nannopterum_brasilianus
Leucocarbo_chalconotus
Phalacrocorax_carbo
Anhinga_rufa
Morus_serrator
Ciconia_nigra
Ciconia_ciconia
Spheniscus_mendiculus
Spheniscus_humboldti
Spheniscus_demersus
Spheniscus_magellanicus
Eudyptula_minor
Eudyptes_chrysocome
Aptenodytes_forsteri
Pygoscelis_papua
Pygoscelis_antarcticus
Pygoscelis_adeliae
Gavia_arctica
Gavia_pacifica
Gavia_stellata
Phaethon_rubricauda
Phaethon_lepturus
Rhynochetos_jubatus
Larus_dominicanus
Larus_crassirostris
Ichthyaetus_relictus
Larus_brunnicephalus
Chroicocephalus_ridibundus
Saundersilarus_saundersi
Sterna_hirundo
Gelochelidon_nilotica
Sternula_albifrons
Synthliboramphus_antiquus
Synthliboramphus_wumizusume
Pinguinus_impennis
Stercorarius_maccormicki
Tringa_glareola
Tringa_erythropus
Tringa_ochropus
Xenus_cinereus
Arenaria_interpres
Scolopax_rusticola
Gallinago_stenura
Limosa_lapponica
Jacana_spinosa
Jacana_jacana
Charadrius_placidus
Vanellus_cinereus
Haematopus_ostralegus
Background
Foreground
Excluded
0.0
0.5
1.0
1.5

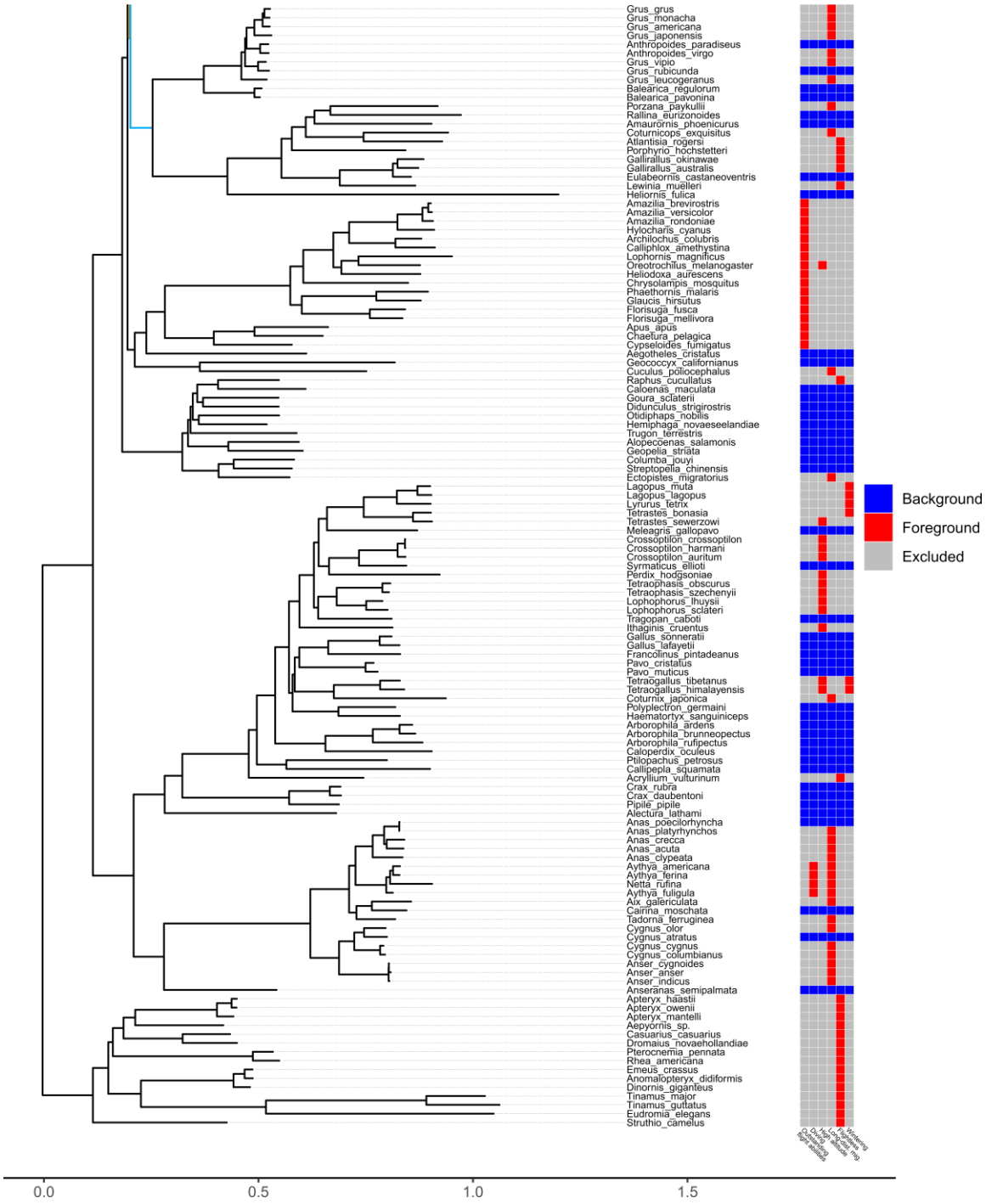


Fig. 4.1: Phylogentic tree and phenotype groups. Distances are measured in number of substitutions per nucleotide site.

reconstruction was also performed in IQ-Tree with usage of genewise partitioned mtVer evolutionary model.

4.2.3. Search for convergent evolution and phenotype to genotype associations

To count simultaneous changes of phenotype and genotype at the site, I use the simultaneous score of the TreeWAS package (Collins and Didelot, 2018). The simultaneous score is designed for the so-called phylogenetic GWAS analysis. It splits each alignment column into binary (two-state) SNPs and counts simultaneous changes of phenotype and genotype at tree branches. To estimate the probability that the observed association is non-random, TreeWAS simulates a "null" genetic dataset under the empirical phylogenetic tree and terminal phenotypes. It also takes from empirical data the distribution of numbers of substitutions per site. In each simulation, it counts the number of phylogenetic branches with simultaneous changes of phenotype and genotype, and combines these counts to obtain the null distribution. At the upper tail of the null distribution, a threshold of significance is drawn at the quantile corresponding to [1 - (alpha-level corrected for multiple testing)]. If a locus in the real dataset has more simultaneous changes than the threshold, it is considered to be significantly associated with the corresponding phenotype.

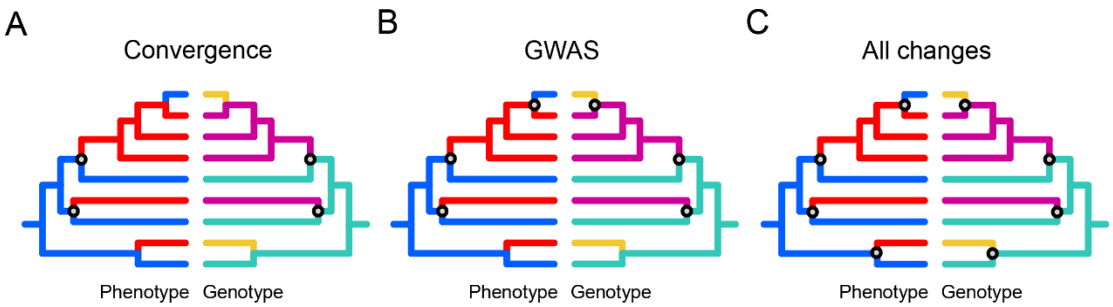


Fig 4.2: Three approaches to counting the number of simultaneous changes in the phenotype and the encoded amino acid. In each panel, the two trees facing each other show the same phylogeny, with the coloring corresponding to the trait states in the corresponding branches; phenotype in the left, and amino acid at a particular position in the right. Circles indicate changes in phenotype and genotype deemed coincident under the corresponding approach. A, under the Convergence approach, of the three gains of the foreground phenotypic trait state (blue to red), two coincide with the gains of the foreground amino acid (cyan to purple). B, under the GWAS approach, two gains (blue to red) and one loss (red to blue) of the foreground phenotypic trait state coincide respectively with two gains (cyan to purple) and one loss (purple to yellow) of the foreground amino acid. C, under the All changes approach, four changes in the phenotype coincide with four changes in the encoded amino acid. The corresponding numbers of simultaneous events is two (A), three (B) and four (C).

TreeWAS was originally developed for analysis of whole-genome datasets of closely related species, in which the assumption that no more than two variants may occur in any particular site generally holds. By contrast, in distantly related bird species considered here, the amino acid sequence of mitochondria has frequently undergone multiple substitutions per site. Therefore, I had to adapt TreeWAS for dealing with non-

binary SNPs. This is non-trivial, as there is have no *a priori* knowledge which of the amino acid changes may be associated with changes in the phenotype.

I used three approaches to identify the phenotype and genotype changes. Generally, each approach resulted in its own set of branches with phenotype and genotype changes, and therefore in different estimates of the simultaneous score.

In the first and in the second approach, I designated one amino acid as foreground, and did not distinguish between the remaining amino acids. I also designated one phenotype (of the two possible phenotypes) as foreground. In the first approach, as genotype changes, I only considered the gains of the foreground amino acid; and as phenotype changes, I only considered the acquisition of the foreground phenotype. As coincident changes, I considered the phylogenetic branches where these events coincided. This approach is referred to as "Convergence" (Figure 4.2A).

In the second approach, as genotype changes, I considered both the gains and the losses of the foreground amino acid, and as phenotype changes, both the gains and the losses of the foreground phenotype. As coincident changes, I considered the phylogenetic branches where both the foreground amino acid and the foreground phenotype were gained, or both were lost. This approach is referred to as "GWAS" (Figure 4.2B). The same approaches are used in studies of Pease *et al*. (2016) and Collins and Didelot (2018).

Finally, in the third approach, I assumed that any amino acid substitution constitutes a genotype change event, and any change in the phenotype counts

independently of its direction. This approach is referred to as “All changes” (Figure 4.2C).

These approaches correspond to different assumptions regarding the genotype-phenotype association. The “Convergence” approach assumes that the gains of the trait are associated with a gain of a specific amino acid variant, while its losses can proceed through multiple means. The “GWAS” approach assumes that both the gains and the losses of the trait are associated respectively with gains and losses of a specific amino acid variant. Finally, the “All changes” approach assumes that both the gains and the losses of the trait are associated with any changes in the encoded amino acid.

4.2.4. Change of site-specific amino acid propensities

To get an alternative view of amino acid changes associated with convergent phenotype characteristics, I asked, for each amino acid site, if changes in amino acid propensities correlated with phenotype change. Among many methods for assessment of position-specific amino acid profiles, I chose the Profile Change method (Rey *et al.* 2018) as it was developed for studies of parallel evolution. This method assigns two amino acid profiles to each site, one for foreground and one for background branches. It then estimates differences between these two profiles in a Bayesian framework and reports the posterior probability that amino acid preferences differ between the two classes of branches. As branch classes, I used the ones for which the corresponding phenotype state was reconstructed.

4.2.5. 3D Structure

To estimate the functional role of candidate mutations, were reconstructed the 3D structure of proteins coded by genes that carry candidate mutations. Sequences of *Gallus gallus* genes *ND1, ND2, ND4, ND5* and *ND6* were aligned with homologous genes of *Ovis aries*. The protein 3D structure based on homology with *Ovis aries* respiratory complex I (PDB:5LNK) was reconstructed by Modeller software (Webb and Sali 2016, Marti-Renom *et al*. 2000, Sali 1993, Fiser *et al.* 2000).

4.3. Results

4.3.1. Simultaneous change

All three types of simultaneous score metrics revealed that the number of significant associations was low. The highest number of sites (8) was detected in high altitude birds, all of them of marginal significance. The Convergence test detected two sites in *ND1* and *ND5* genes (Figure 4.3). The GWAS test detected another two sites in *ND2* gene (Figure 4.4). The All changes test detected 6 sites in *ND1, ND2, ND4, ND5* and *ND6* genes (Figure 4.5). Findings of different tests partially overlap (Table 4.2). Additionally, the All changes test detected the site associated with adaptation to long-distance migration in the *ND5* gene. When the stronger Bonferroni correction was applied, only two sites detected by the Convergence test in high altitude birds and one site detected by the All changes test in long-distance migrants remained significant (Table 4.2).

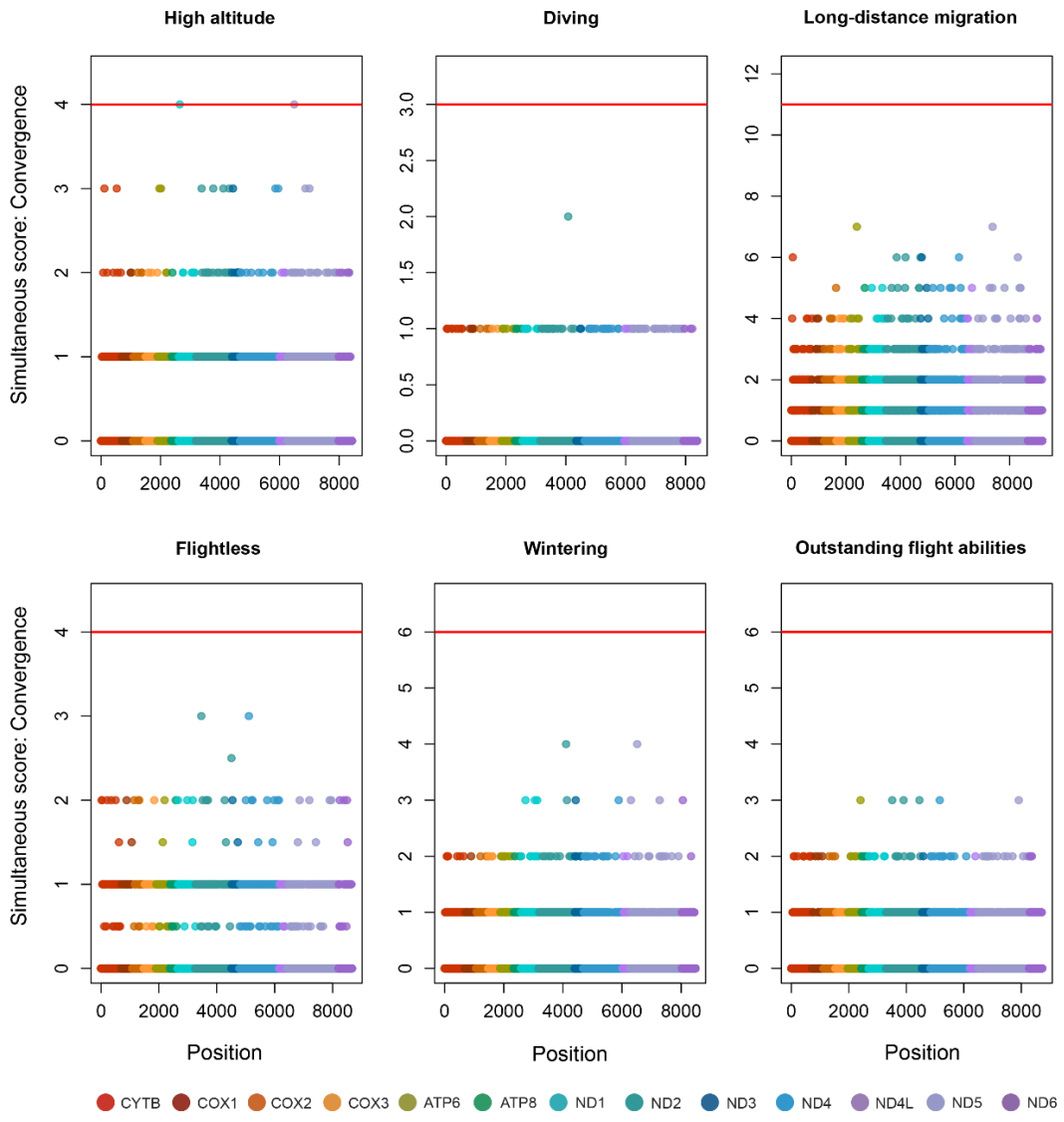


Fig 4.3: Simultaneous test based on the Convergence approach for each of the six considered phenotypic traits. Horizontal axis, position in the mitochondrial genes; vertical axis, number of simultaneous changes of phenotype and genotype. Red line corresponds to significance threshold 0.05 with Bonferroni correction accounting for the number of considered sites in particular test and phenotype (Correction 1).

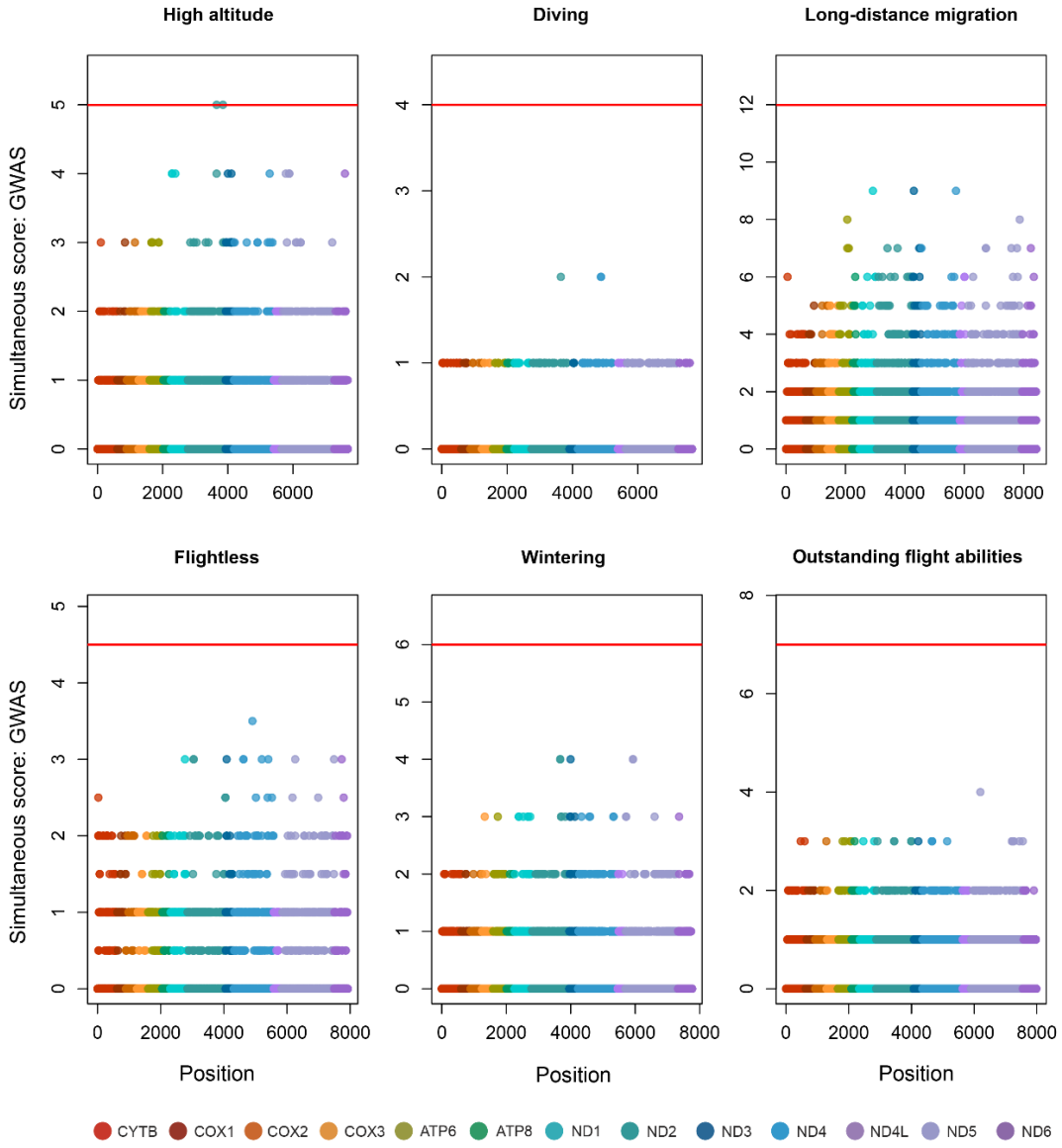


Fig. 4.4: Simultaneous score. GWAS. Horizontal axis, position in the mitochondrial genes; vertical axis, number of simultaneous changes of phenotype and genotype. Red line corresponds to significance threshold 0.05 with Bonferroni correction accounting for the number of considered sites in particular test and phenotype (Correction 1).

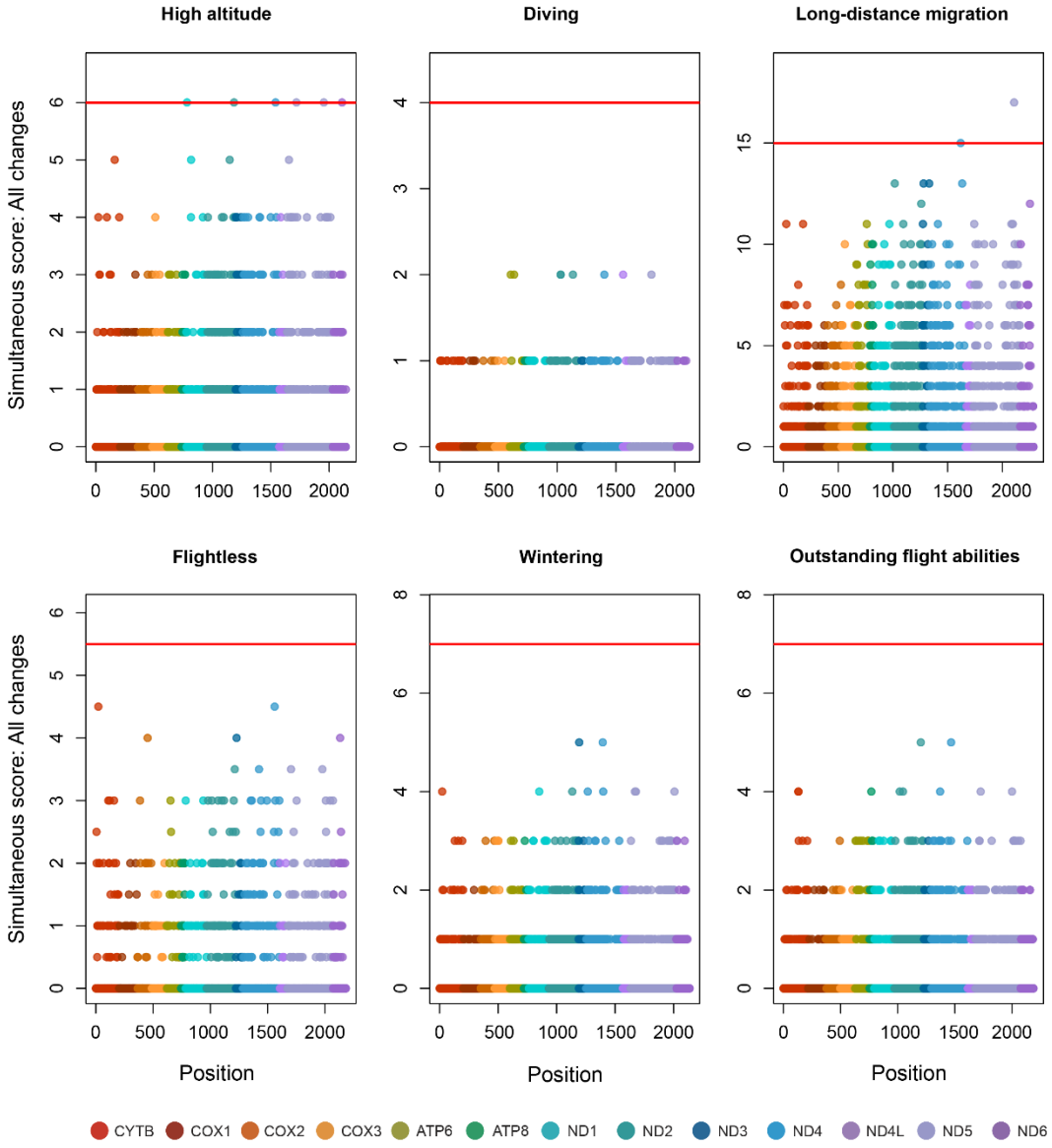


Fig. 4.5: Simultaneous score. All changes. Horizontal axis, position in the mitochondrial genes; vertical axis, number of simultaneous changes of phenotype and genotype. Red line corresponds to significance threshold 0.05 with Bonferroni correction accounting for the number of considered sites in particular test and phenotype (Correction 1).

Table 4.2: List of significant SNPs detected by the Simultaneous test. Correction 1 column indicates the sites that remain significant after Bonferroni correction accounting for the number of considered sites. Correction 2 column indicates the sites that remain significant after Bonferroni correction accounting for the number of sites and the number of tests (3 tests: Convergence, All changes and GWAS).

Test type	Gene	Position in *Gallus gallus*	Correction 1	Correction 2
High altitude				
Convergence	*ND1*	17aa(I)	+	+
	ND5	57aa(H)	+	+
All changes	*ND1*	17aa(I)	+	-
	ND2	329aa(T)	+	-
	ND4	418(T)	+	-
	ND5	114aa(F)	+	-
	ND5	495aa(T)	+	-
	ND6	135(V)	+	-
GWAS	*ND2*	278(M)	+	-
	ND2	329aa(T)	+	-
Long-distance migration				
All changes	*ND5*	533(T)	+	+

4.3.2. Profile change

As an additional test for convergence, I use the amino acid Profile change metric. I expect that recurrent mutations emerging simultaneously with convergent phenotype change could also lead to a change in the amino acid profile between the branches carrying the foreground and the background phenotypes. I compared the results of the Simultaneous tests under all three approaches with the Profile change test (Figure 4.6, Figure 4.7, 4.8).

First, I were interested to estimate profile change levels at sites that have a significant Simultaneous score. Among them, only one (57th position of *ND5*, detected by the Convergence test) has profile change score above 0.5 (0.82) and thus can be considered as potentially convergent. Others either result from convergent changes to the same amino acid without a profile change (17th position of *ND1*), or are a consequence of divergent evolution (all other positions).

Second, it could be expected that sites with higher simultaneous score could have higher profile change metric if a substantial fraction of these sites is involved in convergent adaptations. To test this assumption, I arbitrarily divided the plots (Figure 4.6, Figure 4.7, 4.8) into four parts, and tested if sites with higher simultaneous score have higher profile change score. I found no dependency in any of the tested phenotype groups (Fisher test, significance level 0.01).

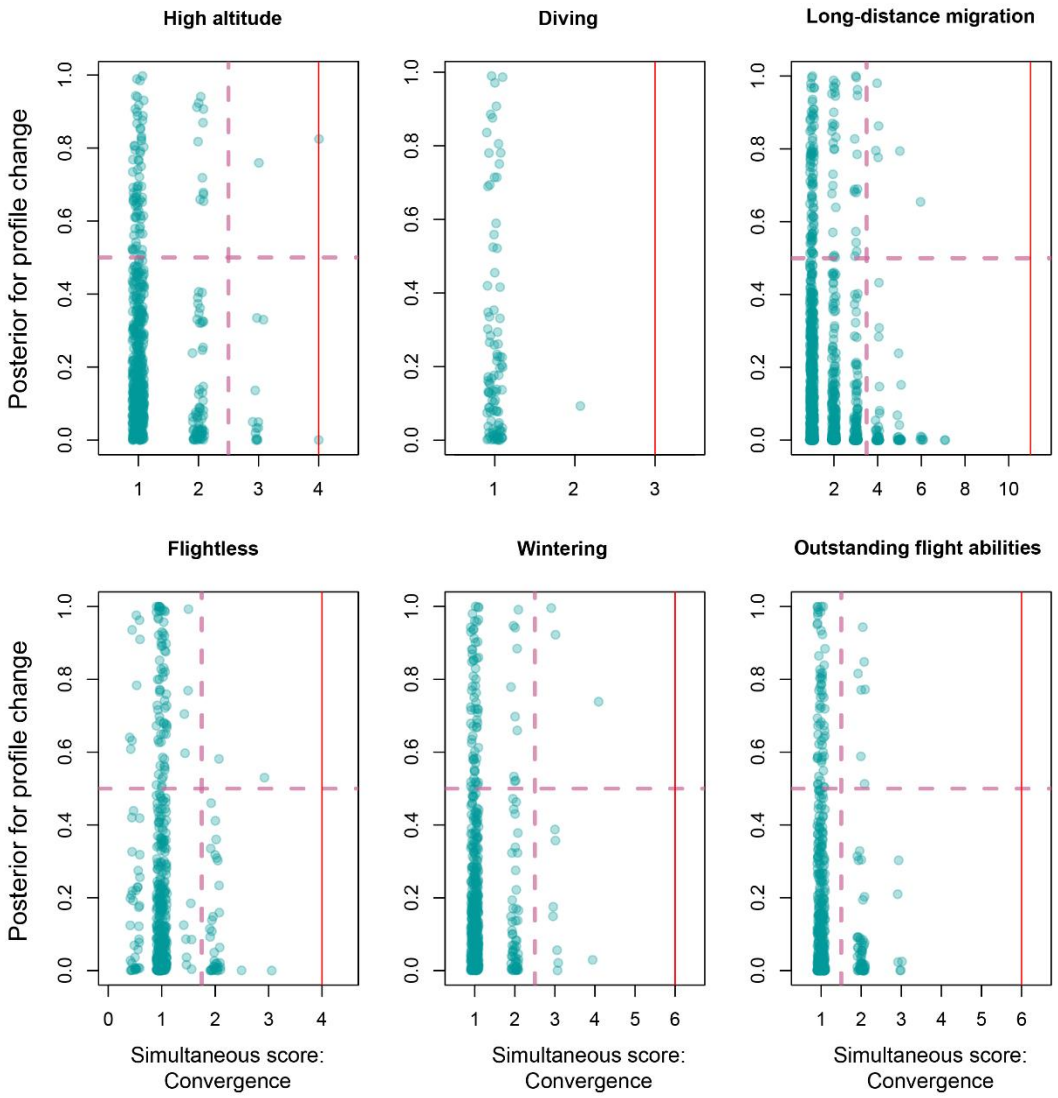


Fig. 4.6: Profile change vs Simultaneous score (Convergence approach). The red line shows the significance threshold for the Simultaneous test. The dashed lines show division of the plot for the Fisher test.

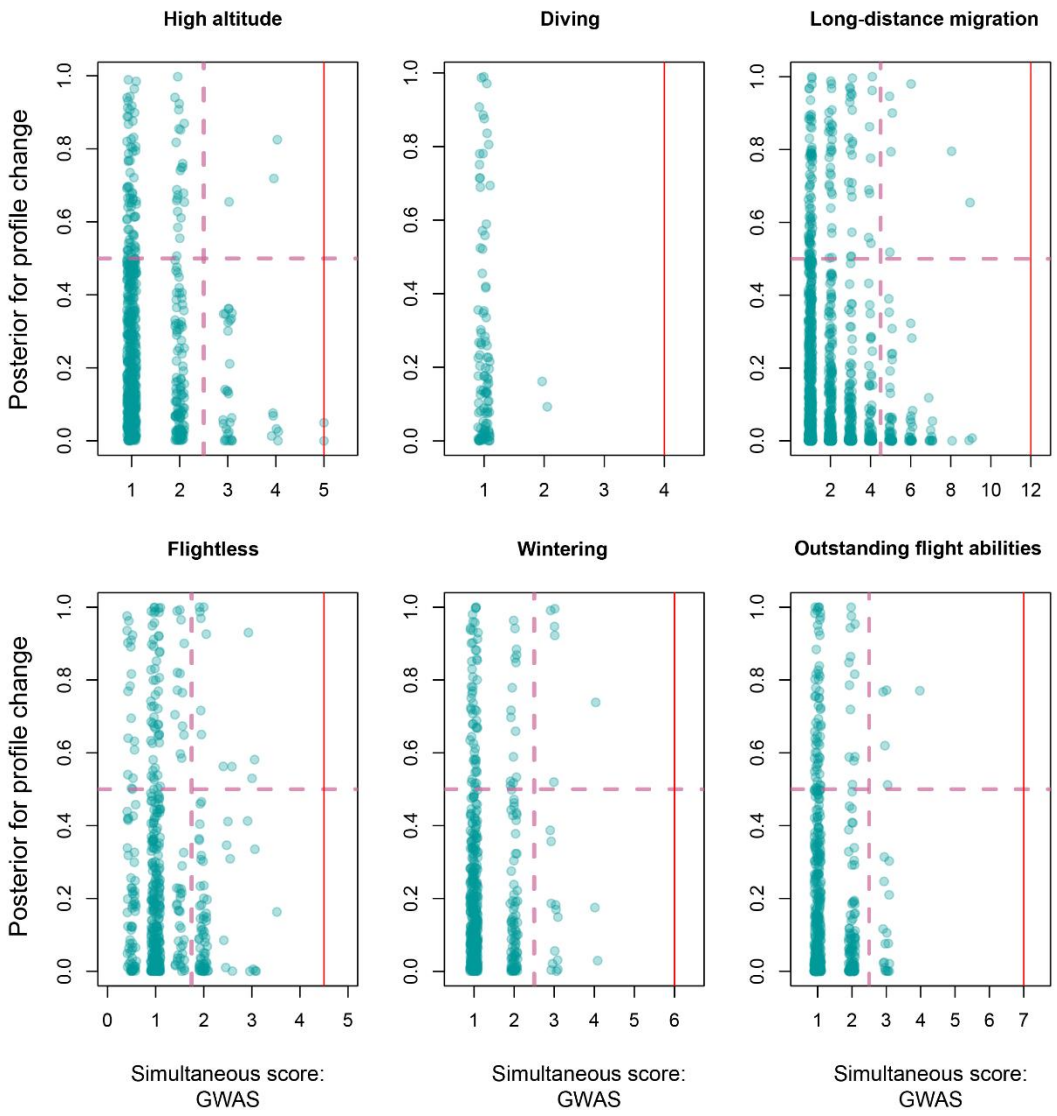


Fig. 4.7: Profile change vs simultaneous score (GWAS). Red line shows significance threshold for simultaneous test. Dashed lines show division of plot for Fisher test.

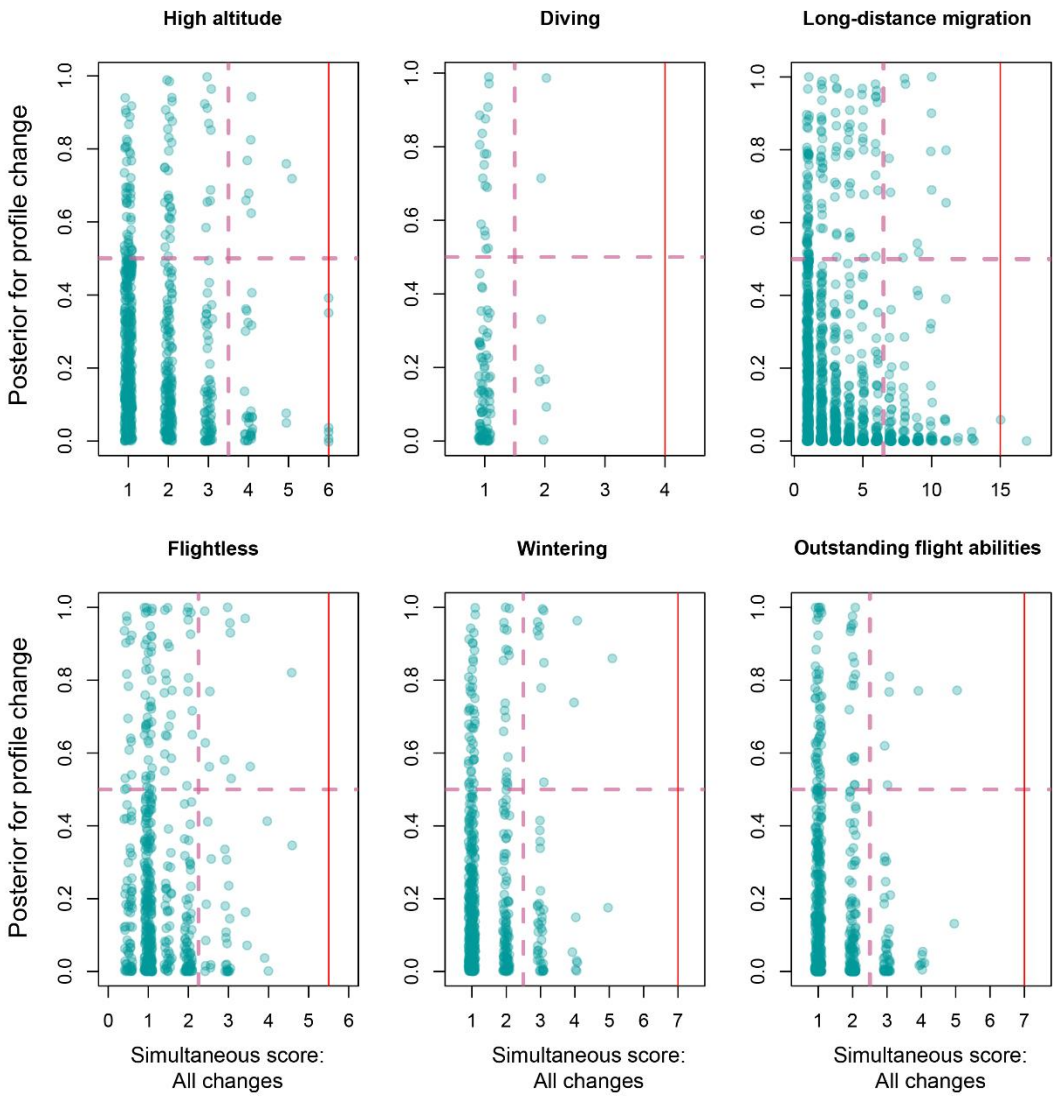


Fig. 4.8: Profile change vs simultaneous score (All changes). Red line shows significance threshold for simultaneous test. Dashed lines show division of plot for Fisher test.

4.3.3. Sites with evidence for phenotypic association

Position 57 in the *ND5* gene carries the highest signal of functional convergence, as both the Simultaneous and Profile change scores are rather high for it (4 and 0.82 respectively). However, except for the 4 substitutions from Histidine to Tyrosine that occurred synchronously with the phenotype change, there are tens of other substitutions between these two amino acids that occurred at various positions of the phylogeny (partially shown in Figure 4.9). This suggests that this convergence can be accidental.

Other positions with relatively high Simultaneous scores demonstrate low Profile change scores. Thus, at best some of these sites could be involved in divergent evolution associated with phenotype changes. I run the MEME tool (Kosakovsky Pond *et al*. 2005) on *ND* genes to find sites involved in recurrent positive selection, yet there was no overlap with previous findings.

As all 9 candidate positions were in *ND* genes, I suggested that adaptations could be associated with the respiratory complex I. Among all 3602 analyzed amino acid positions, *ND* genes account for 1998 positions (55%), so it is unlikely to be a coincidence. To ask if there is additional evidence for function, I mapped the candidate positions onto the 3D structure of the respiratory complex I (Figure 4.10). All the positions are far from the FeS electron-transport clusters and are buried into the membrane arm of the respiratory complex I. They are not grouped together, and they are not close to polar residues in proton channels that play a key role in proton transport (Fiedorczuk *et al*. 2016). In total, structural data provide no additional evidence to consider these sites as adaptive.

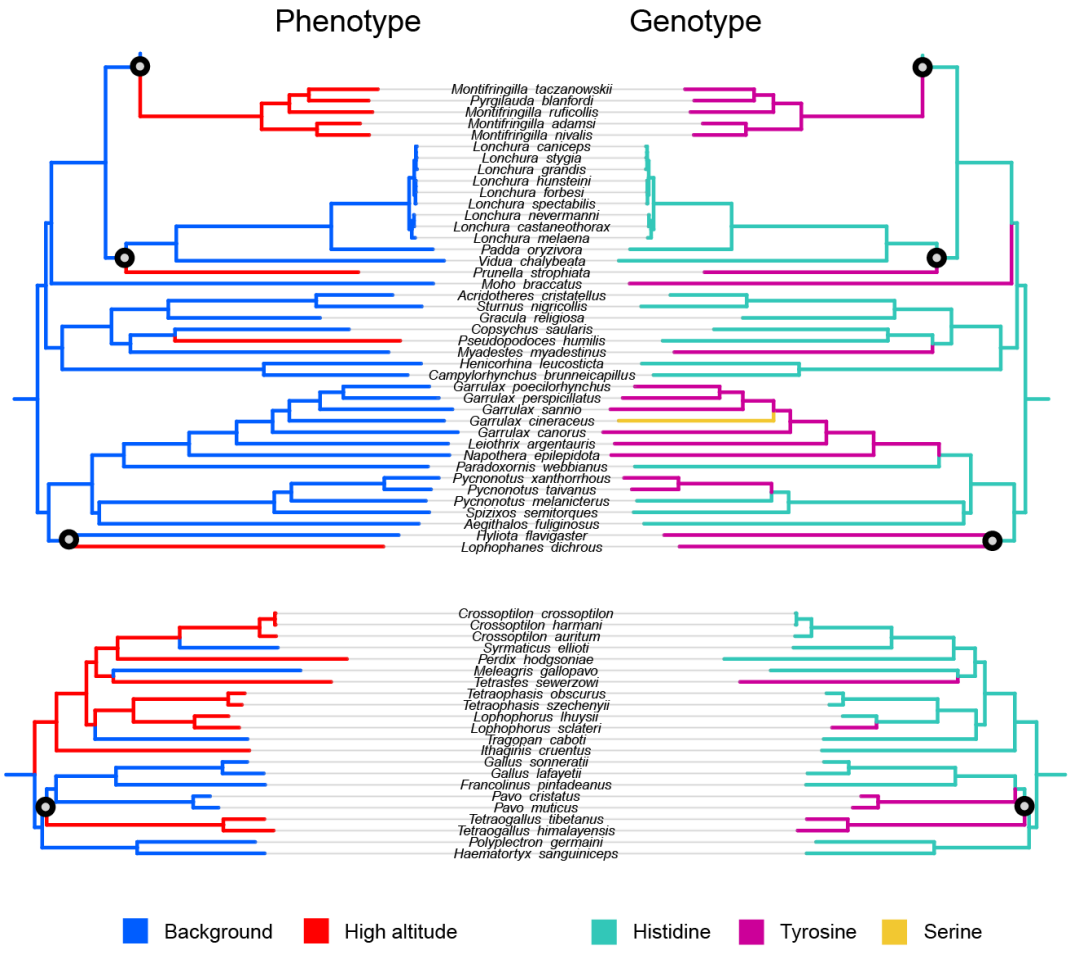


Fig 4.9: Position 57 in the ND5 gene associated with high altitude adaptation. This position carries 4 substitutions from Histidine to Tyrosine simultaneous with adaptation, and a Profile change score of 0.82.

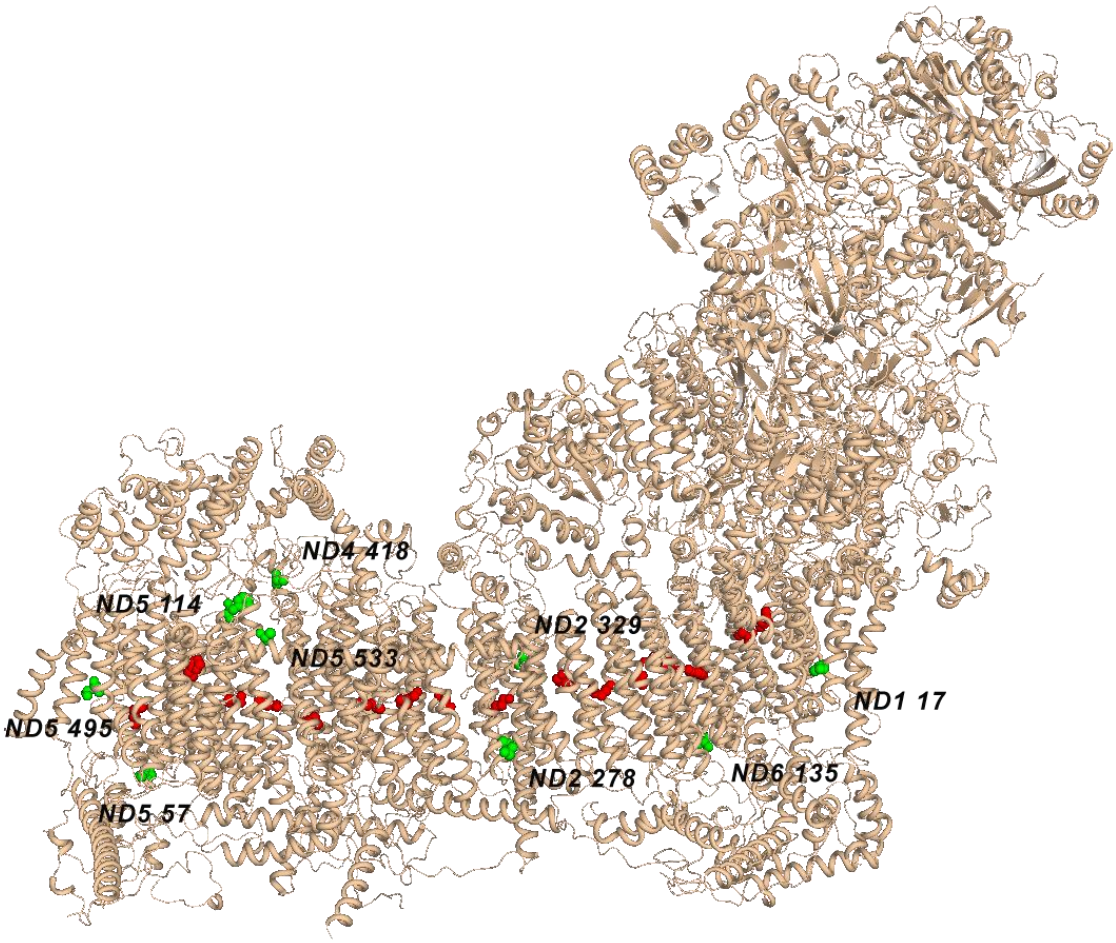


Fig 4.10: 3D structure of respiratory complex I. Candidate amino acid residues are colored green. Polar residues in proton channels, which play a key role in proton transport, are colored red.

4.4. Discussion

The site with the strongest signal of convergence detected in our analysis, which has high Convergence and Profile change scores (position 57 in *ND6*), could originate from functional convergence. Alternatively, the observed pattern of changes could still be coincident, as the mutation pattern rather looks like switches between the set of permitted amino acid under time-invariant amino acid constraints. Among other, basically divergent, positions, none provided additional evidence for selection. I suppose that the concentration of significant associated SNPs in the *ND* genes could be a consequence of higher mutation rates in *ND* genes.

Though it was detected no significant associations at the amino acid level, I hypothesized that sites with higher Simultaneous scores could have elevated Profile change scores. This could be the case if instead of a few strong associations, the data carried sites with convergent associations at a small group of phylogenetically close species, or if the associations were weak. However, I detect no excess Profile change score among the sites with higher Simultaneous scores.

Previous works have attempted to detect convergent single-nucleotide mutations in such distant groups as marine mammals (Foote *et al*. 2015) or echolocating bats and whales (Parker *et al*. 2013). Many of these attempts failed to find significant convergence or were disproved by later studies (Zou and Zhang 2015a, Thomas and Hahn 2015). Similar to those works, I here explore convergence between distantly related species. All the phenotypes analyzed here were acquired repeatedly (Table 4.1), supposedly making

the convergence-based analysis of adaptation more powerful. However, our results did not support this assumption. This suggests that adaptive convergent evolution is rare or hardly detectable in bird OXPHOS system at the considered phylogenetic distances.

There remains a possibility that convergent adaptations in the OXPHOS system could be found in groups of close relatives when the tree of life will be sequenced with higher density. As it was shown by Natarajan *et al.* (2016), similar mutations in hemoglobin subunits lead to high latitude adaptations only in a similar genetic context: there are typical "hummingbird high altitude mutations" and typical "duck high altitude mutations". If so, the near lack of signal in our study has to do with the fact that it was based on too distant organisms which have highly divergent evolutionary landscapes in the genes of interest. However, other work indicates that similar substitutions are rarely involved in independent adaptation to high altitudes even inside a group of closely related species like hummingbirds (Lim *et al*. 2019).

Further work may improve the search for simultaneous changes by using better statistical models. Specifically, these models could be improved by incorporating the heterogeneity of the substitution rates between sites or phylogenetic branches.

Chapter 5. Conclusions

In this work I analyze parallel evolution in organisms of different divergence level, under the assumption that the genetic distance is the main factor determining the reproducibility of evolution. The main conclusions of the work are as follows:

1. Adaptive single-position parallel evolution can prevail over neutral single-position parallel evolution at the whole-genome level in groups of closely related species, as it is detected in amphipods of lake Baikal.

2. Single-position adaptive parallel evolution is hardly detectable at large (between-order) phylogenetic distances, and presumably does not exist in bird mitochondrial OXPHOS genes.

The study is mainly remarkable by the finding of excessive adaptive parallelism in close species. While previous studies only found adaptive parallel evolution at the level of single sites, genes and pathways, I demonstrate prevalence of adaptive over neutral parallel evolution at the whole-genome level. The phenomenon is unique and so far, has only been observed in amphipods. Explosive speciation could have contributed to this observation, as well as high population size of these species.

The attempt to detect single-position adaptive convergence at between-order level was made under the assumption that contemporary methods, and especially data design with multiple phenotype acquisitions will make such findings possible. Our study, being conceptually close to the search for single-nucleotide convergences in echolocating bats and marine mammals, demonstrates the absence of convincing evidence for single-position convergent evolution in species of different taxonomic orders. Our findings are

similar to previous studies, yet datasets with multiple emergences of same phenotypes at the phylogenetic tree allows to judge it with more certainty, than in previous works. I also discuss the few candidate positions that could be associated with the change of phenotype.

Bibliography

1. Abascal F, Zardoya R, and Telford MJ. 2010. TranslatorX: Multiple Alignment of Nucleotide Sequences Guided by Amino Acid Translations. Nucleic Acids Res 38 (Web Server issue): W7–13.
2. Armstrong AF *et al*. 2008. Dynamic changes in the mitochondrial electron transport chain underpinning cold acclimation of leaf respiration. Plant, Cell & Environment, 31: 1156-1169.
3. Bailey SF., Rodrigue N, and Kassen R. 2015. The Effect of Selection Environment on the Probability of Parallel Evolution. Mol Biol Evol. 32(6): 1436-1448.
4. Baym M *et al*. 2016. Spatiotemporal Microbial Evolution on Antibiotic Landscapes. Science 353 (6304): 1147–51.
5. Bazikalova AY *et al*. 1945. Amphipods of Lake Baikal. Trudy Baikalskoj Limnologicheskoj Stantsii, 11, 1–439. (In Russian.)
6. Bazykin GA *et al*. 2007. Extensive Parallelism in Protein Evolution. Biol Direct 2:20.
7. Bazykin GA. 2015. Changing preferences: deformation of single position amino acid fitness landscapes and evolution of proteins. Biol Lett. 11(10): 20150315.
8. Bertels F, Leemann C, Metzner KJ, Regoes R. 2019. Parallel evolution of HIV-1 in a long-term experiment. Mol Biol Evol. 36(11): 2400–14.
9. Bouckaert RR. 2010. DensiTree: making sense of sets of phylogenetic trees, Bioinformatics, 26 (10): 1372–1373.

10. Capella-Gutierrez S, Silla-Martinez JM, and Gabaldon T. 2009. TrimAl: A Tool for Automated Alignment Trimming in Large-Scale Phylogenetic Analyses. Bioinformatics 25(15):1972-3.
11. Castoe TA *et al*. 2009. Evidence for an ancient adaptive episode of convergent molecular evolution. Proc Natl Acad Sci USA. 106(22): 8986-91
12. Chikina M, Robinson JD and Clark NL. 2016. Hundreds of Genes Experienced Convergent Shifts in Selective Pressure in Marine Mammals. Mol Biol Evol 33 (9): 2182–92.
13. Coll F *et al*. 2018. Author Correction: Genome-Wide Analysis of Multi- and Extensively Drug-Resistant Mycobacterium Tuberculosis. Nature Genetics 50 (5): 764.
14. Collins C, Didelot X. 2018. A Phylogenetic Method to Perform Genome-Wide Association Studies in Microbes That Accounts for Population Structure and Recombination. PLOS Comput Biol 14(2): e1005958.
15. Consuegra S *et al*. 2015. Patterns of natural selection acting on the mitochondrial genome of a locally adapted fish species. Genet Sel Evol 47:58.
16. Conte GL, Arnegard ME, Peichel CL, Schluter D. 2012. The Probability of Genetic Parallelism and Convergence in Natural Populations. Proc Biol Sci. 279 (1749): 5039–47.
17. Darwin CR. 1859. On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life. London: John Murray, 1st edition.

18. Darwin CR. 1868. The variation of animals and plants under domestication. London: John Murray, 1st edition, 1st issue. Volume 2.
19. Das J. 2006. The role of mitochondrial respiration in physiological and evolutionary adaptation. Bioessays. 28: 890-901.
20. Dasmahapatra K *et al*. 2012. Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. Nature 487, 94–98.
21. Drozdova P *et al*. 2019. Comparison between transcriptomic responses to short-term stress exposures of a common Holarctic and endemic Lake Baikal amphipods. BMC Genomics, 20(1): 712.
22. Edgar RC. 2004. MUSCLE: Multiple Sequence Alignment with High Accuracy and High Throughput. Nucleic Acids Res. 32(5): 1792-7.
23. Fiedorczuk K, Letts J, Degliesposti G *et al*. 2016. Atomic structure of the entire mammalian mitochondrial complex I. Nature 538, 406–410.
24. Fiser A, Do RK, Sali A. 2000. Modeling of loops in protein structures. Protein Science 9: 1753-1773.
25. Fitzpatrick BM. 2004. Rates of evolution of hybrid inviability in birds and mammals. Evolution, 58(8): 1865–1870.
26. da Fonseca RR *et al*. 2008. The adaptive evolution of the mammalian mitochondrial genome. BMC Genomics 9, 119.
27. Foote AD *et al*. 2015. Convergent Evolution of the Genomes of Marine Mammals. Nature Genetics. 47: 272–275.

28. Giessler S, Mader E, Schwenk K. 1999. Morphological evolution and genetic differentiation in Daphnia species complexes. J. Evol. Biol. 12(4), 710-723.
29. S. Götz *et al*. 2008. High-throughput functional annotation and data mining with the Blast2GO suite, Nucleic Acids Research, 36, 3420-3435.
30. Jančúchová-Lásková J, Landová E, Frynta D. 2015. Are genetically distinct lizard species able to hybridize? A review. Current Zoology 61(1): 155-180.
31. Jarvis ED *et al*. 2014. Whole-genome analyses resolve early branches in the tree of life of modern birds. Science 346(6215):1320-1331.
32. Ji F *et al*. 2012. Mitochondrial DNA variant associated with Leber hereditary optic neuropathy and high-altitude Tibetans. Proc. Natl. Acad. Sci. USA 109: 7391–7396.
33. Hahn MW, and Nakhleh L. 2016. Irrational Exuberance for Resolved Species Trees. Evolution 70 (1): 7–17.
34. Hiller M *et al*. A "Forward Genomics" Approach Links Genotype to Phenotype using Independent Phenotypic Losses among Related Species. Cell Reports, 2(4): 817-823.
35. Hodgkinson A, Eyre-Walker A. 2011. Variation in the mutation rate across mammalian genomes. Nat. Rev. Genet. 12(11): 756-66.
36. Kang L *et al*. 2013. MtDNA lineage expansions in Sherpa population suggest adaptive evolution in Tibetan highlands. Mol. Biol. Evol. 30: 2579–2587.
37. Karasov T, Messer PW, Petrov DA. 2010. Evidence that adaptation in Drosophila is not limited by mutation at single sites. PLoS Genet. 6(6): e1000924

38. Karolchik DR *et al*. 2007. The UCSC Genome Browser Database: 2008 Update. Nucleic Acids Res. 36(Database issue): D773-9.
39. Kimball RT *et al*. 2019. A Phylogenomic Supertree of Birds. Diversity 11(7): 109.
40. Kimura M. 1987. Evolutionary Rate at the Molecular Level. Nature 217: 624–626.
41. Klink GV, Bazykin GA. 2017. Parallel Evolution of Metazoan Mitochondrial Proteins. Genome Biol Evol 9(5): 1341-50.
42. Klink GV, Golovin AV, Bazykin GA. 2017. Substitutions into amino acids that are pathogenic in human mitochondrial proteins are more frequent in lineages closely related to human than in distant lineages. PeerJ 5:e4143.
43. Kosakovsky Pond SL, Frost SDW, Muse SV. 2005. HyPhy: hypothesis testing using phylogenies. Bioinformatics 21(5): 676-679.
44. Kreiner JM *et al*. 2019. Multiple modes of convergent adaptation in the spread of glyphosate-resistant Amaranthus tuberculatus. Proc Natl Acad Sci USA
45. Kryazhimskiy S, Rice DP, Jerison ER, Desai MM. 2014. Microbial Evolution. Global Epistasis Makes Adaptation Predictable despite Sequence-Level Stochasticity. Science 344 (6191): 1519–22.
46. Lefébure T *et al*. 2017. Less effective selection leads to larger genomes. Genome Res. 27(6): 1016-1028.
47. Li H. 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 27(21): 2987-93.

48. Li G *et al*. 2008. The hearing gene Prestin reunites echolocating bats. PNAS 105 (37): 13959-13964.

49. Li L, Stoeckert CJ Jr, Roos DS. 2003. OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. Genome Res. 13: 2178-2189.

50. Lim MCW, Witt CC, Graham CH, Dávalos LM. 2019. Parallel Molecular Evolution in Pathways, Genes, and Sites in High-Elevation Hummingbirds Revealed by Comparative Transcriptomics. Genome Biol Evol 11(6): 1573–1585.

51. Liu A *et al*. 2019. Convergent degeneration of olfactory receptor gene repertoires in marine mammals. BMC Genomics 20: 977.

52. Lucassen M, Koschnick N, Eckerle LG, Pörtner HO. 2006. Mitochondrial mechanisms of cold adaptation in cod (Gadus morhua L.) populations from different climatic zones. Journal of Experimental Biology 209: 2462-2471.

53. Malinsky M *et al*. 2018. Whole Genome Sequences of Malawi Cichlids Reveal Multiple Radiations Interconnected by Gene Flow. Nat Ecol Evol 2(12): 1940-1950.

54. Marcovitz A *et al*. 2019. A functional enrichment test for molecular convergent evolution finds a clear protein-coding signal in echolocating bats and whales. Proc Natl Acad Sci U S A. 116(42): 21094-21103.

55. Martin CH, Feinstein LC. 2014. Novel trophic niches drive variable progress towards ecological speciation within an adaptive radiation of pupfishes - Molecular Ecology 23 (7): 1846-1862.

56. Marti-Renom MA *et al*. 2000. Comparative protein structure modeling of genes and genomes. Annu. Rev. Biophys. Biomol. Struct. 29: 291-325.

57. McDonald JH, Kreitman M. 1991. Adaptive protein evolution at the Adh locus in Drosophila. Nature 351(6328): 652-4.

58. Meier J *et al*. 2017. Ancient hybridization fuels rapid cichlid fish adaptive radiations. Nat Commun 8, 14363.

59. Meredith RW, Gatesy J, Cheng J, Springer MS. 2011. Pseudogenization of the tooth gene enamelysin (MMP20) in the common ancestor of extant baleen whales. Proc. R. Soc. B. 278: 993–1002

60. Meredith RW, Gatesy J, Springer MS. 2013. Molecular decay of enamel matrix protein genes in turtles and other edentulous amniotes. BMC Evol. Biol. 13:20.

61. Minh BQ *et al*. 2020. IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. Mol. Biol. Evol.37:1530-1534.

62. Natarajan C *et al*. 2016. Predictable Convergence in Hemoglobin Function Has Unpredictable Molecular Underpinnings. Science 354 (6310): 336–39.

63. Naumenko S *et al* 2017. Transcriptome-Based Phylogeny of Endemic Lake Baikal Amphipod Species Flock: Fast Speciation Accompanied by Frequent Episodes of Positive Selection. Mol Ecol. 26(2): 536-553.

64. Nikolaev SI, Montoya-Burgos JI, Popadin K, Parand L, Margulies EH. 2007. Life-history traits drive the evolutionary rates of mammalian coding and noncoding genomic elements. Proc Natl Acad Sci U S A. 104(51): 20443-8.

65. Parker J *et al*. 2013. Genome-wide signatures of convergent evolution in echolocating mammals. Nature 502, 228–231.

66. Partha R, Kowalczyk A, Clark NL, Chikina M. 2019. Robust Method for Detecting Convergent Shifts in Evolutionary Rates. Mol. Biol. Evol. 36(8): 1817–1830.

67. Pease JB, Haak DC, Hahn MW, Moyle LC. 2016. Phylogenomics Reveals Three Sources of Adaptive Variation during a Rapid Radiation. PLoS Biol. 14(2): e1002379.

68. Pessia E *et al*. 2012. Evidence for Widespread GC-biased Gene Conversion in Eukaryotes, Genome Biol. Evol. 4(7): 675–682.

69. Povolotskaya IS, Kondrashov FA. 2010. Sequence space and the ongoing expansion of the protein universe. Nature. 465: 922–926.

70. Ranwez V, Harispe S, Delsuc F, Douzery EJP. 2011. MACSE: Multiple Alignment of Coding SEquences accounting for frameshifts and stop codons. PLoS One 6(9): e22594.

71. Rey C, Guéguen L, Sémon M, Boussau B. 2018. Accurate Detection of Convergent Amino-Acid Evolution with PCOC. Mol Biol. Evol. 35(9): 2296–2306.

72. Rokas A, Carroll SB. 2008. Frequent and Widespread Parallel Evolution of Protein Sequences. Mol Biol Evol. 25(9): 1943–1953.

73. Sali A. 1993. Comparative protein modelling by satisfaction of spatial restraints. J. Mol. Biol. 234: 779-815.

74. Sawyer SA, Kulathinal RJ, Bustamante CD, Hartl DL. 2003. Bayesian analysis suggests that most amino acid replacements in Drosophila are driven by positive selection. J Mol Evol. 57.

75. Scott GR *et al*. 2011. Molecular Evolution of Cytochrome c Oxidase Underlies High-Altitude Adaptation in the Bar-Headed Goose. Mol. Biol. Evol. 28(1): 351–363.

76. Seehausen O. 2004. Hybridization and adaptive radiation. Trends Ecol Evol 19(4): 198-207.

77. Seplyarskiy VB, Kharchenko P, Kondrashov AS, Bazykin GA. 2012. Heterogeneity of the Transition/Transversion Ratio in Drosophila and Hominidae Genomes. Mol Biol Evol. 29(8): 1943-55.

78. Shen YY *et al*. 2010. Adaptive evolution of energy metabolism genes and the origin of flight in bats. PNAS 107(19): 8666-8671.

79. Silva G, Lima FP, Martel P, Castilho R. 2014. Thermal adaptation and clinal mitochondrial DNA variation of European anchovy. Proc. R. Soc. B. 281(1792):20141093.

80. Smith NG, Eyre-Walker A. 2002. Adaptive protein evolution in Drosophila. Nature, 415(6875): 1022-4.

81. Soria-Carrasco V *et al*. 2014. Stick Insect Genomes Reveal Natural Selection's Role in Parallel Speciation. Science. 344: 738–742.

82. Stamatakis A. 2014. RAxML Version 8: A Tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies. Bioinformatics 30(9):1312-3.

83. Stolyarova AV *et al*. 2019. Senescence and entrenchment in evolution of amino acid sites. Biorxiv.

84. Storz JF. 2016. Causes of Molecular Convergence and Parallelism in Protein Evolution. Nat Rev Genet. 17(4): 239–50.

85. Strauli NB, Hernandez RD. 2016. Statistical inference of a convergent antibody repertoire response to influenza vaccine. Genome Med 8, 60.

86. Terekhanova NV *et al*. 2014. Fast Evolution from Precast Bricks: Genomics of Young Freshwater Populations of Threespine Stickleback Gasterosteus Aculeatus. PLoS Genet. 10(10): e1004696.

87. Thomas GWC, Hahn MW. 2015. Determining the Null Model for Detecting Adaptive Convergence from Genomic Data: A Case Study Using Echolocating Mammals. Mol Biol Evol. 32(5): 1232-6.

88. Toews DPL, Mandic M, Richards JG, Irwin DE. 2014. Migration, mitochondria, and the yellow-rumped warbler. Evolution 68: 241-255.

89. Tomasco IH, Lessa EP. 2011. The evolution of mitochondrial genomes in subterranean caviomorph rodents: adaptation against a background of purifying selection. Molecular Phylogenetics and Evolution. 61(1):64-70.

90. Trapnell C, Pachter L, Salzberg SL. 2009. TopHat: Discovering Splice Junctions with RNA-Seq. Bioinformatics. 25(9): 1105-1111.

91. Usmanova DR, Ferretti L, Povolotskaya IS, Vlasov PK, Kondrashov FA. 2015. A Model of Substitution Trajectories in Sequence Space and Long-Term Protein Evolution. Mol Biol Evol. 32(2): 542–54.

92. Vavilov NI. 1922. The law of homologous series in variation. Journal of Genetics, 12(1), 47–89.

93. Webb B, Sali A. 2016. Comparative Protein Structure Modeling Using Modeller. Curr. Protoc. Bioinformatics 54: 5.6.1-5.6.37.

94. Woods R, Schneider D, Winkworth CL, Riley MA, and Lenski RE. 2006. Tests of Parallel Molecular Evolution in a Long-Term Experiment with Escherichia Coli. Proc Natl Acad Sci USA. 103(24): 9107-9112.

95. Xu SQ *et al*. 2005. A Mitochondrial Genome Sequence of the Tibetan Antelope (Pantholops hodgsonii), Genomics, Proteomics & Bioinformatics 3(1): 5-17.

96. Xue KS *et al*. 2017. Parallel evolution of influenza across multiple spatiotemporal scales. eLife 6: e26875.

97. Yuana ML *et al*. 2018. Mitochondrial phylogeny, divergence history and high-altitude adaptation of grassland caterpillars (Lepidoptera: Lymantriinae: Gynaephora) inhabiting the Tibetan Plateau. Mol.Phyl. Evol. 122: 116-124.

98. Zhang HX *et al*. 2013. Mitochondrial genome sequences of Artemia tibetiana and Artemia urmiana: assessing molecular changes for high plateau adaptation. Science China Life Sciences 56: 440–452.

99. Zhou T, Shen X, Irwin DM, Shen Y, Zhang Y. 2014. Mitogenomic analyses propose positive selection in mitochondrial genes for high-altitude adaptation in galliform birds. Mitochondrion 18: 70-75.

100. Zou Z, Zhang J. 2015a. No Genome-Wide Protein Sequence Convergence for Echolocation. Mol Biol Evol 32(5): 1237–41.

101. Zou Z, Zhang J. 2015b. Are Convergent and Parallel Amino Acid Substitutions in Protein Evolution More Prevalent Than Neutral Expectations? Mol Biol Evol. 32(8): 2085–96.

Appendix A

Table A1: Comparison of *P* test value with one in amphipods.

AMPHIPODS				
	n	# values less than one	# values higher than one	P-value (two-tailed sign test; alternative hypothesis: median is not equal to one)
Mean over 6 mutation types	195	2	193	< 2.2E-16
AC	195	2	193	< 2.2E-16
AT	195	5	190	< 2.2E-16
AG	195	22	173	< 2.2E-16
CT	195	55	140	9.807E-10
CG	195	1	194	< 2.2E-16
TG	195	9	186	< 2.2E-16
AC&AT&AG &CT&CG&TG	1170	94	1076	4.441E-16

Table A2: Comparison of *P* test value with one in cichlids.

CICHLIDS				
	n	# values less than one	# values higher than one	P-value (two-tailed sign test; alternative hypothesis: median is not equal to one)
Mean for 6 points	300	34	266	< 2.2E-16
AC	300	64	236	< 2.2E-16
AT	300	52	248	< 2.2E-16
AG	300	52	248	< 2.2E-16
CT	300	106	194	4.241E-07
CG	300	115	185	6.321E-05
TG	300	85	215	3.753E-14
AC&AT&AG &CT&CG&TG	1800	474	1326	2.22E-16

Table A3: Comparison of *P* test value with one in vertebrates.

VERTEBRATES				
	n	# values less than one	# values higher than one	P-value (two-tailed sign test; alternative hypothesis: median is not equal to one)
Mean for 6 points	147	147	0	< 2.2E-16
AC	147	147	0	< 2.2E-16
AT	147	140	7	< 2.2E-16
AG	147	147	0	< 2.2E-16
CT	147	147	0	< 2.2E-16
CG	147	147	0	< 2.2E-16
TG	147	145	2	< 2.2E-16
AC&AT&AG &CT&CG&TG	882	873	9	< 2.2E-16

Table A4: The *P* test value does not depend on distance between species pairs in amphipods. Median-based linear models were computed for double log-transformed data.

AMPHIPODS				
	n	slope	intercept	P-value (wilcoxon test)
Mean over 6 mutation types	195	-0.007007	0.755258	0.496
AC	195	0.007533	0.868408	0.492
AT	195	-0.005556	1.018691	0.958
AG	195	-0.08457	-0.02293	<2E-16
CT	195	-0.13979	-0.36952	<2E-16
CG	195	0.01055	1.01779	0.25
TG	195	0.04038	0.89232	0.000527
AC&AT&AG &CT&CG&TG	1170	-0.01557	0.61244	7.47E-05

Table A5: The *P* test value does not depend on distance between species pairs in cichlids. Median-based linear models were computed for double log-transformed data.

CICHLIDS				
	n	slope	intercept	P-value (wilcoxon test)
Mean for 6 points	300	0.022474	0.26996	2.52E-06
AC	300	0.0216	0.31956	0.016771
AT	300	0.082033	0.80653	5.24E-11
AG	300	-0.055139	-0.3141	1.02E-22
CT	300	0.022021	0.20386	0.0085445
CG	300	-0.036588	-0.22525	0.0031679
TG	300	0.050262	0.46828	1.69E-05
AC&AT&AG &CT&CG&TG	1800	0.0014662	0.096866	0.17639

Table A6: The *P* test value decreases with distance between species pairs in vertebrates. Median-based linear models were computed for double log-transformed data.

VERTEBRATES				
	n	slope	intercept	P-value (wilcoxon test)
Mean for 6 points	147	-0.2009	-1.11489	<2.00E-16
AC	147	-0.2178	-1.3289	2.91E-16
AT	147	-0.3178	-1.0214	<2E-16
AG	147	-0.1964	-0.90747	<2E-16
CT	147	-0.3073	-1.5038	<2E-16
CG	147	-0.1868	-1.3217	9.86E-10
TG	147	-0.1536	-0.9061	1.02E-11
AC&AT&AG &CT&CG&TG	882	-0.2106	-1.1370	<2.00E-16

Table A7: Polymorphism proportion in parallel sites.

	Non-parallel sample				Parallel samples			
	Non-Population sample		Population sample of Eulimnogammarus verrucosus (19 individuals)		Non-Population sample		Population sample of Eulimnogammarus verrucosus (19 individuals)	
	syn	nonsyn	syn	nonsyn	syn	nonsyn	syn	nonsyn
AC	0.0684 (13559/ 198018)	0.0385 (2410/6 2532)	0.2427 (4725/ 19466)	0.1764 (1215/ 6885)	0.0627 (345/55 01)	0.0276 (120/43 42)	0.2344 (683/29 13)	0.1211 (396/32 68)
AT	0.0672 (6907/1 02708)	0.0415 (2387/5 7505)	0.2627 (2785/ 10600)	0.2409 (1943/ 8063)	0.0572 (118/20 61)	0.0258 (101/39 01)	0.2567 (276/10 75)	0.1652 (618/37 40)
AG	0.0859 (27439/ 319062)	0.0497 (9099/1 83017)	0.2890 (8880/ 30717)	0.2545 (6542/ 25700)	0.0756 (1449/1 9149)	0.0389 (649/16 651)	0.3047 (3441/1 1293)	0.2212 (2993/1 3529)
CT	0.0876 (46606/ 531741)	0.0549 (4570/8 3114)	0.3001 (15930/ 53074)	0.2158 (2521/ 11678)	0.0808 (2716/3 3605)	0.0353 (231/65 28)	0.3182 (5915/1 8585)	0.1778 (923/51 90)
CG	0.0690 (11906/ 172380)	0.0332 (1883/5 6626)	0.2238 (3478/ 15539)	0.1394 (1050/ 7530)	0.0575 (238/41 33)	0.0246 (97/393 4)	0.2741 (706/25 75)	0.0996 (309/31 02)
TG	0.0680 (11852/ 174122)	0.0459 (2379/5 1784)	0.2522 (3837/ 15210)	0.1839 (1041/ 5658)	0.0670 (333/49 67)	0.0332 (113/34 03)	0.2603 (667/25 62)	0.1507 (382/25 34)

Table A8: Expected from transcriptomic data and observed after sanger sequencing nucleotides in sites with parallel substitutions. Dashes are positions for which sequences could not be obtained.

	Site 1		Site 2		Site 4	
	Transcrip tome	Sanger DNA	Transcrip tome	Sanger DNA	Transcrip tome	Sanger DNA
Ommatogamm arus albinus	C	C	T	T	C	C
Eulimnogamm arus marituji	A/C	A/C	C	C	C	C

Table A9: Primers for sanger resequencing, which were selected for conservative regions in homologous genes of *Hyalella azteca*.

	Site1 Hyalella azteca proteasome subunit alpha type-3-like (LOC108674219), mRNA	Site2 Hyalella azteca polyadenylate-binding protein 4-like (LOC108683085), mRNA	Site4 Hyalella azteca peptidyl-prolyl cis-trans isomerase FKBP4-like (LOC108674895), mRNA
Left primer	GTRAAGGAYAA GCAGTTT	CAGCAGTTCACT ATGCAG	TATTGAAAAACT GAAGAAGA
Right primer	YACATRTCTTCC TCTTCAT	GTRTACTTGTAV GCGTTG	AAATATTTGGTT CCTTTATC

Table A10: Deepwater and shallow species (by Bazikalova, 1945).

Deepwater species (usually below 200m):	
Brachiuropus grewingkii	140-1300m, usually below 300-400m
Eulimnogammarus ussolzewi	60-697m, usually below 100-200m
Garjajewia dershawini	80-1250m
Pachyschesis branchialis	42-1131m, usually below 200-300m
Shallow species (not more than 100m, usually above 30m):	
Pallaseopsis kessleri	1-61m, usually at 10-20m
Pallasea cancelloides	0.3-178m, usually at 1-10m
Hyalellopsis carinata	3-20m, rarely till 50m
Hyalellopsis grisea	15m
Hyalellopsis setosa	4-8m
Hyalellopsis stebbingi	0.5-52m, usually above 30m
Heterogammarus sophianosii	1.5m-100m
Eulimnogammarus viridulus	0.5-30m
Eulimnogammarus vittatus	0-30m, usually at 2-3m
Eulimnogammarus verrucosus	0.25m, rarely 5-10m
Eulimnogammarus similis	4-26m, rarely at 53-107m
Eulimnogammarus cruentus	0.5-35m, rarely till 100m
Eulimnogammarus cyaneus	0.25m
Brandtia latissima	3.5-32m, rarely found at 170m

Table A11: Phenotype characteristics (by Bazikalova, 1945) of species in two quartets, which show highest *P* values. No specific reasons for elevated parallel evolution have been found.

Quartet 1, *P* = 4.17			
Linevichella vortex	Micruropus glaber	Poekilogammarus pictoides	Eulimnogammarus similis
Body length: 5-7mm Depth: 0.5-209m, usually 5-10m Habitat: stones with seaweed	Body length: 6.5-7.2mm Depth: 1.5-70m, usually 10-15m Habitat: stones with seaweed	Body length: 10-12mm Depth: 3.5-100m Habitat: sandy and rocky soil	Body length: 10-15mm Depth: 4-6m, sometimes 53-107m Habitat: sandy and rocky soil
Quartet 2, *P* = 4.10			
Linevichella vortex	Macrohectopus branickii	Carinurus bicarinatus	Poekilogammarus pictoides
Body length: 5-7mm Depth: 0.5-209, usually 5-10 Habitat: stones with seaweed	Body length: 25-30mm Depth: 0-1410m Habitat: pelagic zone	Body length: 30mm Depth: 200m Habitat: muddy soil	Body length: 10-12mm Depth: 3.5-100m Habitat: sandy and rocky soil
Quartet 3, *P* = 0.96			
Baikalogammarus pullus	Macrohectopus branickii	Gammarus lacustris	Eulimnogammarus messerschmidtii
Body length: 3-6 mm Depth: 0.5-25, usually 5-10 Habitat: stones with seaweed	Body length: 25-30mm Depth: 0-1410m Habitat: pelagic zone	Body length: 7-15mm Depth: most common at 0-1 m Habitat: usually sand, silt or macrophytes	Body length: 7-18mm Depth: 0-0.3m, possibly deeper Habitat: pebbles and sand, macrophytes
Quartet 4, *P* = 0.97			
Linevichella vortex	Micruropus glaber	Gammarus lacustris	Eulimnogammarus messerschmidtii

Body length: 5-7mm Depth: 0.5-209m, usually 5-10m Habitat: stones with seaweed	Body length: 6.5-7.2mm Depth: 1.5-70m, usually 10-15m Habitat: stones with seaweed	Body length: 7-15mm Depth: most common at 0-1 m Habitat: usually sand, silt or macrophytes	Body length: 7-18mm Depth: 0-0.3m, possibly deeper Habitat: pebbles and sand, macrophytes

Skoltech

2020