Detecting the molecular basis of phenotypic convergence

Olivier Chabrol, Manuela Royer-Carenzi, Pierre Pontarotti, Gilles Didier

Detecting the molecular basis of phenotypic convergence

Olivier Chabrol[1], Manuela Royer-Carenzi[1], Pierre Pontarotti[1,2] and Gilles Didier[1]

[1] Aix Marseille Univ, CNRS, Centrale Marseille, I2M, Marseille, France

[2] Aix Marseille Univ, IRD, APHM, MEPHI, IHU Méditerranée Infection, Marseille, France

November 19, 2020

Abstract

1. Convergence is the process by which several species independently develop similar traits. This evolutionary process is not only strongly related to fundamental questions such as the predictability of evolution and the role of adaptation, its study also may provide new insights about genes involved in the convergent phenotype. We focus on this latter question and aim to detect the molecular basis of a given phenotypic convergence.
2. After pointing out a number of concerns with current detection methods based on ancestral reconstruction, we propose a novel approach combining an original measure, called *convergence index*, which associates to any proteic site a quantity reflecting the extent to which it supports a phenotypic convergence, with a statistical framework for selecting genes from the convergence indices of all their sites.
3. First, our measure of the "convergence level" outperforms two previous ones in distinguishing simulated convergent sites from non-convergent ones. Second, by applying our detection approach to the well-studied case of convergent echolocation between dolphins and bats, we identified a set of genes which is very significantly annotated with audition-related GO-terms.
4. This result constitutes an indirect evidence that genes involved in a phenotypic convergence can be identified with a genome-wide approach, a point which was highly debated, notably in the echolocation case. Our approach paves the way to systematic studies of numerous examples of convergent evolution in order to link convergent phenotypes to genotypes.

1 Introduction

Evolutionary convergence, which is part of what is called homoplasy in cladistics, is a key concept in evolutionary biology (Stayton, 2015b; Pontarotti and

Hue, 2016). It is indeed an important and rather common evolutionary phenomenom, which may involve all kinds of traits, including behavioral, morphological, developmental and molecular ones. (Losos *et al.*, 1998; Losos, 2009; Mahler *et al.*, 2013; Gallant *et al.*, 2014; Pfenning *et al.*, 2014; Vidal-García and Keogh, 2015; Ujvari *et al.*, 2015; Friedman *et al.*, 2016; Davis *et al.*, 2016).

As a key concept, convergence has been considered from various points of view. Intuitively, the main underlying idea is that convergence arises as soon as two or more species independently evolve similar phenotypes that are not derived from a common ancestral phenotype. In order to avoid confusion, let us start by giving a formal "working" definition of convergence for a binary character (i.e., the presence/absence of a given phenotype), which will be discussed and refined in Section 2.2. We say that a phenotype is *convergent* over two taxa $\mathbf{s}_1$ and $\mathbf{s}_2$ if the two following assertions are true:

1. both taxa $\mathbf{s}_1$ and $\mathbf{s}_2$ have the phenotype;
2. the the Most Recent Common Ancestor (MRCA) of both taxa does not have the phenotype.

We emphasize the fact that our study deals only with binary characters. Considering more complex (e.g., quantitative) characters raises many questions which shall not be addressed in the present work.

Deciding whether a given phenotype is convergent in two or more taxa is generally not obvious. Several approaches have been developed for studying this question (Revell *et al.*, 2007; Ingram and Mahler, 2013; Arbuckle *et al.*, 2014; Arbuckle and Speed, 2016; Stayton, 2015a; Speed and Arbuckle, 2017).

The question that we shall address here is slightly different from identifying convergence events. Given the presence of a binary character (typically the presence or absence of some phenotype), which is assumed to be convergent for at least two extant taxa, we aimed at detecting genes showing convergences at the amino acid level which support the convergence of the phenotype. In particular, this is not the same as detecting molecular convergences *per se*, i.e., not related with a phenotype as considered in Zhang and Kumar (1997) and Storz (2016). More specifically, the inputs required to address this question are:

- the phylogenetic tree of a set of extant taxa,
- the information about the taxonomic distribution the phenotype of interest,
- the alignments of the clusters of orthologous genes of the extant taxa,

from which we aim to output a selection of genes which significantly support the convergence of the considered phenotype. Note that we implicitly expect the observed phenotypic convergence to result, at least partially, from molecular convergences that we are aiming at detecting.

An expected outcome of identifying such genes is, at first, a better understanding of the evolutionary mechanisms leading to the acquisition of new phenotypes. Second, genes supporting the convergence are suspected of playing a role not only in the emergence of the phenotype but also in its functions, possibly yielding new insights into the biological processes involved. This is thus

an important question, which has been addressed by several previous works (Yokoyama *et al.*, 2011; Parker *et al.*, 2013; Foote *et al.*, 2015; Thomas and Hahn, 2015; Zou and Zhang, 2016). All these previous approaches first measure the strength of convergence, with regard to the phenotype considered, for all sites of a given dataset and then select genes according to the convergence level of their sites. They differ mainly in the way of measuring the convergence level of sites.

A first class of measures of the convergence level of a site is conceptually very close to the aforementioned definition (Foote *et al.*, 2015; Thomas and Hahn, 2015; Zou and Zhang, 2016). Its main idea is to check if the taxa with the convergent phenotype show the same amino acid at the studied site and if this amino acid has been derived independently. To this end, the "convergent" amino acids are compared with the ancestral reconstructed ones, not necessarily at the MRCA level. For instance, Foote *et al.* (2015) compare the amino acid of each marine mammal with the reconstructed amino acid of its most recent ancestor having a terrestrial descendant. Assuming that the amino acids are accurately reconstructed allows counting the number of times that a given amino acid has been derived independently toward a taxon with the phenotype of interest (we shall see in Section 2.2 that this is not completely true). Since our approach was mainly designed to address this concern, we will further discuss ancestral reconstruction in the next section.

Another way of measuring the convergence level, introduced by Castoe *et al.* (2009) and used by Parker *et al.* (2013), consists in testing, for each site, the "real" phylogeny against an alternative phylogeny that separates the extant taxa having the convergent phenotype from the other one s. The convergence strength of a site is then measured in terms of ΔSSLS (site-wise log-likelihood support) that is the difference between the log-likelihoods obtained from these two phylogenies. The approach is conceptually far from the definition of convergence, in the sense that ΔSSLS tests an evolutionary hypothesis corresponding to the alternative phylogeny separating taxa with from those without the convergent phenotype, which is not obvious to interpret (Zou and Zhang, 2015b), rather than several independent substitutions toward a same amino acid. We refer to Zou and Zhang (2015b) and Thomas and Hahn (2015) for a thorough discussion and a critical evaluation of this method. Castoe *et al.* (2009) considered molecular convergence between all pairs of independent lineages and sought for excess of convergence in branches related to the phenotypic convergence.

The stage in which these approaches select genes from the convergence level of their sites is generally straightforward. For instance, Thomas and Hahn (2015) and Foote *et al.* (2015) considered genes that contain at least one convergent site, according to the measure used. Parker *et al.* (2013) ranked genes according to the mean ΔSSLS of their sites and considered the top-ranked ones.

Genome-wide detection of molecular signature of convergence is an emergent area of research which is still controversial. The article of Parker *et al.* (2013), about echolocation, was followed by two responses: from Thomas and Hahn (2015) and Zou and Zhang (2015b), who conclude that there is "no genome-wide protein sequence convergence for echolocation".

We propose here a new measure of the convergence level of a site, called *convergence index*, altogether with a statistical framework for ranking and selecting significant genes with regard to the convergence level of their sites. By applying our detection approach to the dataset of Thomas and Hahn (2015),

still about echolocation, we draw the opposite conclusion to that of Zou and Zhang (2015b). The set of genes significantly convergent between dolphins and microbats (the two echolocating taxa) show a very significant enrichment in GO-terms associated with audition (defined from a set of keywords given in Section SI-2.6) in contrast to those detected from the other pairs of terminal taxa in the dataset. These results provide an indirect evidence that molecular signatures of a phenotypic convergence may be detected with a suitable approach.

2 Material and methods

With a given set of extant taxa including some with a phenotype assumed convergent as well as the phylogenetic tree and alignments of orthologous genes of the extant taxa, the question is to identify the genes that show molecular convergences consistent with that of the phenotype. To this end, we follow the same general outline as the previous approaches, i.e., by first considering a convergence measure on alignment sites, then by selecting genes from the convergence level of their sites.

Our convergence measure, called *convergence index*, is essentially an attempt to address some concerns raised by ancestral reconstruction which are discussed in Section 2.1. The convergence index itself is presented in Section 2.2 (see also Section SI-1).

The convergence index of a site is not directly used for measuring the strength of its convergence. We rather consider its significance under a null "neutral" evolutionary model in order to normalize effects due to the number of convergent extant taxa, to the phylogenetic tree and to the evolution ary rate of its gene (Section 2.3).

The last stage consists in selecting the genes which contain a significant number of sites detected as convergent with regard to their index (Section 2.4).

2.1 Ancestral reconstruction approaches

Methods for identifying molecular signatures of convergence from ancestral sequence reconstruction (Foote *et al.*, 2015; Thomas and Hahn, 2015) raise several concerns, among which:

1. in order to decide whether there is convergence for a given site, one has to choose the ancest ral nodes whose reconstructed amino acids will be compared with those of convergent extant taxa;

2. ancestral reconstruction always comes with a certain amount of uncertainty, which is not taken into account by standard ancestral reconstructions;

3. approaches based on ancestral reconstruction implicitly assume that if one observes a same amino acid both at an ancestral taxa and in its direct descendant, then it was continuously present all along the branch (i.e., no substitution occurred during the corresponding timeline).

The first concern is not a big issue in the case where only two extant taxa have the convergent phenotype since, in this case, the MRCA is quite a natural

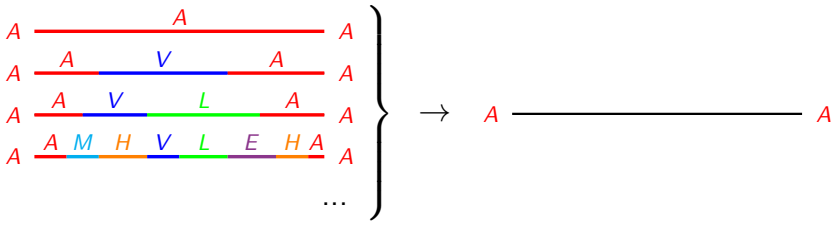

Figure 1: Several evolutionary histories (left) leading to the presence of amino acid A at the beginning and the end of a branch of a phylogenetic tree (right). The parts in red (resp. in blue, in green, ...) correspond to the times when amino acid A (resp. V, L, ...) was continuously present at the corresponding site.

choice. Things get more complicated with datasets containing a greater number of convergent extant taxa. The ancestral nodes to be compared have then to be chosen with regard to inferences about whether they ha d the phenotype of interest or not (e.g., Foote *et al.* (2015)).

The second concern may be easier to address. Ancestral reconstruction approaches based on stochastic evolutionary models are able to provide the probabilities of reconstructing any given amino acid at a specified ancestral node. This makes it possible to compute the expected number of convergent events in the sense of ancestral reconstruction approaches (Zou and Zhang, 2015a). In a similar way, Castoe *et al.* (2009) were interested in the expected number of convergences between all pairs of branches of a phylogenetic tree by considering the posterior probabilities of all ancest ral amino acids.

Let us remark that the definition of convergence given in the introduction is not completely consistent with the intuitive idea that there is convergence as soon as a phenotype (or here an amino acid) appeared independently. There may be an independent substitution toward an amino acid X inside a branch, even in the case where X is present at both nodes beginning and ending the branch. Figure 1 illustrates this point by displaying four different evolutionary histories of a site along a branch, all leading to observe *Alanine* (A) both at its beginning and at its end. All the histories but the one at the top-left of the figure show an independent substitution toward *Alanine*, which will not be considered as such in an ancestral reconstruction framework, even if it deals with the reconstruction uncertainty like Zhang and Kumar (1997); Castoe *et al.* (2009); Zou and Zhang (2015a). In short, assuming that no substitution occurred on a branch which starts and ends with a same amino acid is an oversimplification which may lead to underestimate the actual number of molecular convergences.

2.2 Convergence index of an alignment site

In order to introduce our convergence measure, let us start by assuming that the whole evolutionary history of a site is known. By whole evolutionary history, we mean that the amino acid present in all lineages and at all times encompassed by the phylogenetic tree is known (i.e., we know which amino acid is present not only at the nodes but anywhere in the tree, including inside branches). In this situation, and for all amino acids X present in a convergent extant

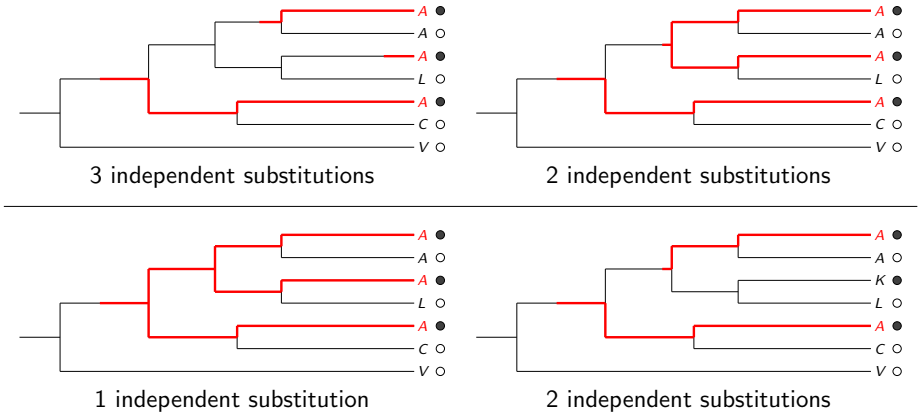

Figure 2: Independent substitutions toward amino acid A with regard to a given phenotype assumed convergent. Extant taxa are represented with $\bullet$ or $\circ$ depending on whether they have the convergent phenotype or not. Parts of the tree where amino acid A is continuously present until an extant convergent taxon are red-colored.

taxon, it is straightforward to count the number of substitutions toward X which are conserved from the substitution to an extant taxon with the convergent phenotype. This number reflects intuitively the extent to which substitutions toward X support the convergence of the phenotype for the site considered. It will be referred to as the *number of independent substitutions toward* X. Figure 2 displays several evolutionary histories leading to different numbers of independent substitutions toward A, which correspond to the number of starting points of the red parts of branches in the figure. Note that the amino acid considered (A in Figure 2) may not be present at all in the extant taxon having the convergent phenotype (e.g., evolutionary history at the bottom-right of Figure 2).

Unfortunately, in a real situation, we do not have access to the whole evolutionary history of a site, but only to the amino acids of the extant taxa. All is not lost, however, since we are able to compute the expected number of independent substitutions under a standard continuous time Markov model of evolution (Section SI-1).

In order to obtain a synthetic measure, the convergent index of a site is defined as the maximum over all amino acids X of the expected number of substitutions toward X continuously conserved from the substitution to an extant convergent taxon, conditioned on amino acids of all the extant taxa (i.e., the corresponding alignment column, see Equation 5 of Section SI-1).

Let us note that the concerns stated at the beginning of Section 2.1 do no apply to the convergence index, since:

1. computing the expected number of substitutions toward an amino acid does not require to select any ancestor node;

2. it does take into account the uncertainty due to the stochastic nature of evolution, since it is an expectation under a probabilistic model;

3. our calculus distinguish es between the case where there is no substitution

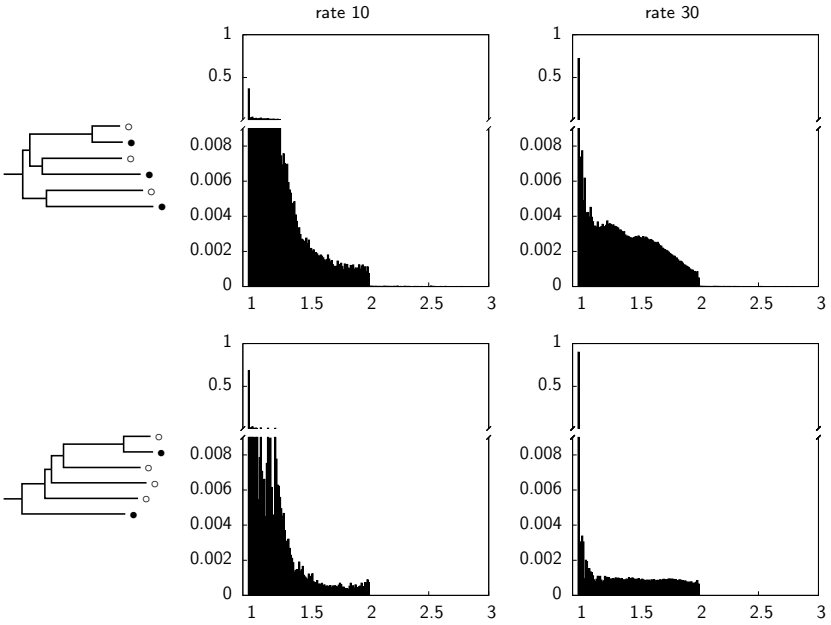

Figure 3: Simulated distributions of the convergence index under neutral evolution. Distributions of each line are simulated from the tree s displayed in Column 1 at evolution ary rates 10 (Column 2) and 30 (Column 3). The tree of the top row (resp. of the bottom row) has three (resp. two) convergent taxa (represented with •).

all along a branch of the phylogenetic tree and the case where an ancestor and its direct descendant share a same amino acid (Section SI-1).

The computation of the convergence index requires a continuous time Markov model of sequence evolution, which is generally given by its substitution rate matrix (Whelan and Goldman, 2001). In order to compute likelihoods over phylogenetic trees, this matrix has to be multiplied by a constant rate, standing for the evolution speed with regard to the time unit of the branch lengths. The choice of this rate has a great influence on the convergence index. In particular, a high rate leads systematically to convergence indices almost equal to the number of convergent taxa. After trying several alternatives, we devised a heuristic approach for calibrating the evolution ary rate used to compute the convergence index from the phylogenetic tree and the convergent taxa, which ensures that the convergence index takes values over a range as wide as possible (Section SI-2.2).

2.3 Significance of a site

For all amino acids X, the expected number of independent substitutions toward X which are conserved from the substitution to a convergent extant tax on is smaller or equal to the total number of convergent extant taxa. It follows that the convergence index heavily depends on the number of convergent extant taxa. Another factor which influences the expected number of substitutions toward

an amino acid is the evolutionary rate of the site. For instance, no molecular convergence can be identified from a completely conserved site. Figure 3 displays the simulated distributions of the convergence index over two different trees, having respectively three and two convergent taxa, and at evolutionary rates 10 and 30. We do observe that these distributions are quite different between each other. In particular, distributions simulated with rate 10 do not have the same general shape as those simulated with rate 30. Note that the rate used for computing the convergence index is here fixed, constant over all the plots, and different from that used for the simulations. Convergence indices of the bottom row distributions are bounded by 2 while those of top row distributions may actually reach 3, which is the number of convergent taxa, though the corresponding probabilities are very low and not visible at the scale of the plots (Figure 3).

In order to assess whether a site is convergent or not, we thus have to normalize its convergence index with regard to its evolution ary rate, the phylogenetic tree and the number and positions of extant taxa with the convergent phenotype. To this end, we consider the p-value of its convergence index from the empirical distribution of the convergence indices of proteic sites simulated under neutral evolution on the same phylogenetic tree and with the same convergent taxa. The parameters of the evolutionary model used for simulations are estimated from the whole gene to which the tested site belongs. We insist on the fact that the model used for simulating sites does not have to be the same as the one used for computing the convergence index. Convergence index is treated here as a statistics of the site of which we evaluate the distribution under an evolutionary model of its gene (this model may take into account rates heterogeneity etc.). In the current implementation of the method, the same substitution rate matrix is used both for convergence indices and for simulations but convergence indices are computed with a single evolution ary rate while simulations are performed from evolution ary rates drawn from a discretized Gamma distribution. Computing convergence indices from the exact same model as for simulations worked as well but was several times more time-consuming.

2.4 Significance of a gene

In our context, assessing the significance of a gene requires to combine the (empirical) p-values of its sites. Since combining p-values is a question of broad interest, several methods have been developed to perform this task (Loughin, 2004). The widely used "quantile" approaches such as Fisher and truncated product are ill-suited to our particular question. Empirical p-values are prone to uncertainty, notably the smallest ones which have the greatest influence on these methods. Thus, we follow Wilkinson (1951) and start by choosing a significance level γ. A site is said *convergent at a significance level* γ, or γ-*convergent*, if the probability of observing a convergence index greater or equal to its own convergence index is smaller or equal to γ, in the empirical distribution associated to its gene as described in Section 2.3. We developed an adaptive sampling scheme which determines the number of simulations required for ensuring a given confidence level to the number of γ-convergent sites of a gene, with regard to its length and γ. All genes are then associated with the number of γ-convergent sites that they contain. By assuming independence between sites,

the number of convergent sites of a gene of length L follows a binomial distribution of parameters (γ, L). The p-value of the convergent status of a gene, which is the probability of observing a number superior or equal to the observed number of convergent sites in this binomial distribution, is thus straightforward to compute. This p-value has to be corrected for multiple testing with regard to all the genes/alignments in the dataset, in order to give the final significance of this gene.

2.5 Detection pipeline overview

The detection pipeline is schematically displayed in Figure 4. In order to detect molecular signatures of a given convergent phenotype inside a set of genes, the phylogenetic tree of a set of taxa among which some are convergent, the information about whether they carry the convergent phenotype and the alignments of orthologous genes of these taxa are required as inputs. Users have to provide four parameters: a substitution matrix $\mathcal{M}$, suited to the type of sequences considered, a significance threshold γ for deciding which sites are convergent and a significance threshold β for deciding if a gene is convergent with regard to its length, the convergent sites that it contains and the total number of genes. The execution of the pipeline follows three stages. Stage 1, "Method calibration", determines the evolutionary rate μ used for computing the convergence index with matrix $\mathcal{M}$ (Section SI-2.2). Stage 2 consists of treating all alignments/genes of the dataset by (i) estimating the parameter α of the discretized Gamma distribution for the evolutionary rates of the alignment protein from matrix $\mathcal{M}$ (Yang, 1994), (ii) simulating the empirical distribution of the convergence index from the estimated parameter α with $\mathcal{M}$, (iii) computing the convergence index of all sites with the method rate μ under $\mathcal{M}$ and (iv) determining the p-values associated to alignments/genes with regard to parameter γ, the number of sites of significance smaller than γ and the length of alignments/genes. In Stage 3, p-values are corrected for multiple testing. Finally, alignments/genes with corrected p-values smaller than parameter β are returned. The pipeline also returns the complete list of genes, sorted according to their p-values, and the positions of their γ-convergent sites (Supplementary information).

2.6 Simulating (non-)convergent sites

Simulated neutral "non-convergent" sites, used both in our simulation study (Section 3.1) and for computing the empirical distributions of convergent index, were obtained by simulating the evolution of amino acid s on the tree displayed in Figure 5, under the WAG model with an evolutionary rate of 10 (Whelan and Goldman, 2001) and by keeping only the amino acids of the extant taxa which give us our alignment columns.

Simulated convergent sites were obtained in two stages. First, we simulated the evolution of an amino acid in the very same way as for a non-convergent site. Second, for all simulated sites, we randomly picked an extant taxon with the convergent phenotype and "copied" its amino acid in all the other convergent extant taxa. Thus, we obtained an alignment column (i.e., a site) in which a same amino acid occurs at all the entries corresponding to the convergent taxa.

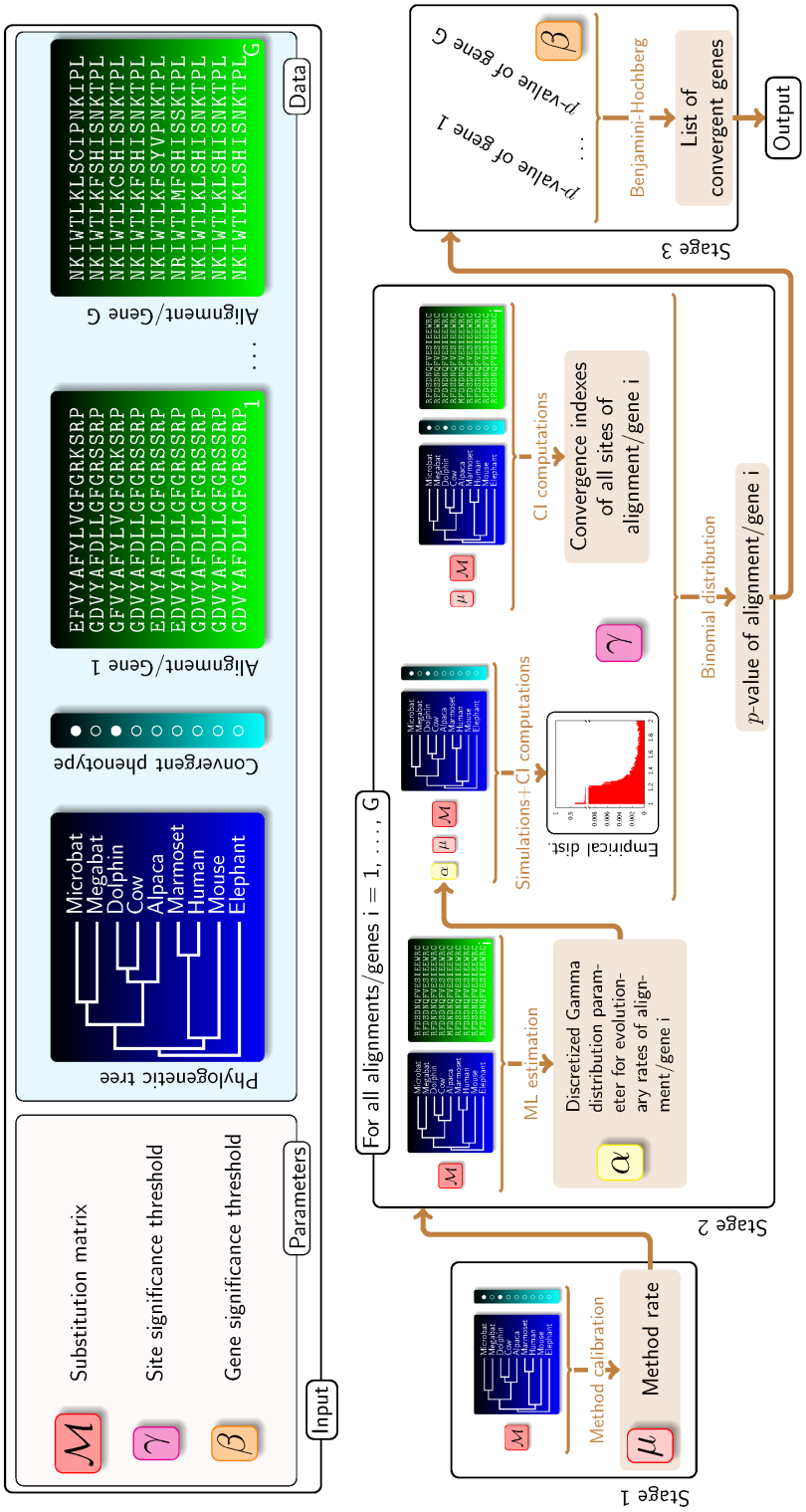

Figure 4: Schematic of the detection pipeline. *ML* stands for *Maximum Likelihood* and *CI* for *Convergence Index*.

2.7 Availability

The source code of the software implementing the detection of the molecular signatures of phenotypic convergence is available at `https://github.com/gilles-didier/Convergence`.

3 Results

3.1 Simulation study: Comparison of 3 measures of convergence of a site

In order to assess the accuracy of measures of the convergence level, we simulated the evolution of non-convergent and convergent sites on the tree of Thomas and Hahn (2015) (Figure 5-left-top). We compared 3 measures, namely the convergence index, the number of convergences observed from the ancestral reconstruction (Foote *et al.*, 2015) and the ΔSSLS (Parker *et al.*, 2013). We used the tree of Thomas and Hahn (2015) with the same convergent extant taxa, also displayed in Figure 6. The ΔSSLS measure was computed with an alternative tree built following the ideas of Hypothesis H2 in Parker *et al.* (2013) (Figure 5-left-bottom). Following Foote *et al.* (2015), we compared the amino acids of convergent extant taxa with those reconstructed at their most recent ancestors with at least one descendant without the convergent phenotype, in order to count the number of convergent events for the ancestral reconstruction method. Since all the sites were simulated on the same tree with the same convergent taxa and under the same evolution ary rate, it is not required to normalize the convergence indices with regard to their empirical distribution. They are thus used directly.

The relevance of the convergence measures was next assessed with regard to their ability to distinguish between the simulated non-convergent and convergent sites (the status of all simulated sites is known). Figure 5-right displays the results obtained for the ancestral reconstruction, ΔSSLS and the convergence index. The ROC curves (Zhou *et al.*, 2009) reporting the results of each measure show that the convergence index better discriminates between convergent and non-convergent sites than the two other measures (Figure 5-Right). In this particular example, the convergence index identifies 99% of the convergent sites with an error rate of 14%, with a suitable threshold.

3.2 Application: Convergent genes related to echolocation

We applied the approach described in Section 2, to the dataset of Thomas and Hahn (2015). This dataset was designed for studying the emergence(s) of echolocating abilities in mammals, like that of Parker *et al.* (2013). It contains 6,332 alignments of orthologous genes from 9 mammal taxa, among which two have echolocating abilities (dolphins and microbats), and the phylogenetic tree of these 9 taxa (Figure 6 and Section SI-2.3).

Convergent sites are detected at a confidence level of 10^{-4} according to the empirical distribution simulated with regards to the genes to which they belong (Sections SI-2.4). We next computed the (binomial) p-values of all genes with regard to the number of convergent sites that they contain, ranked the genes

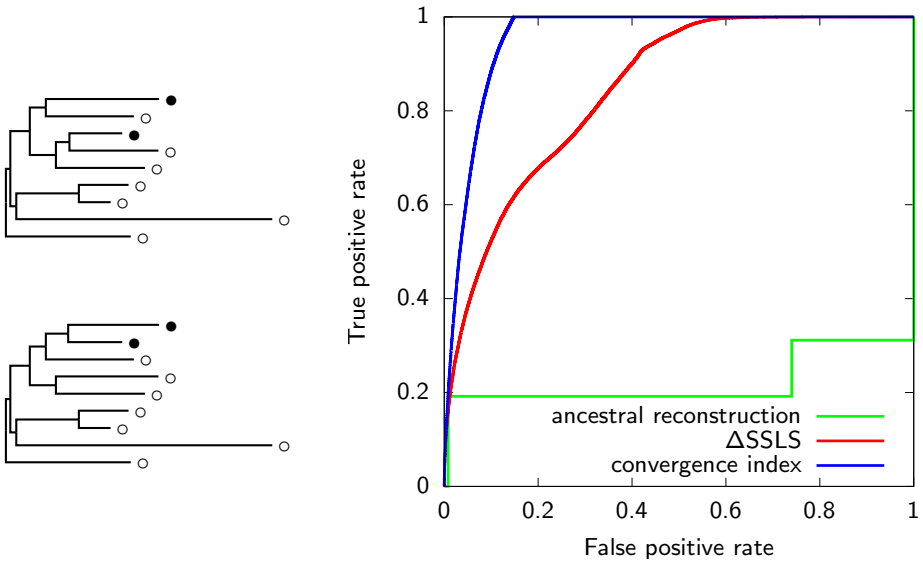

Figure 5: (left-top) Tree used for simulating non-convergent and convergent protein site. (left-bottom) Alternative tree used for ΔSSLS computation. Extant taxa are represented with • or ◦ depending on whether they have the convergent phenotype. (right) ROC curves obtained from 100,000 simulations. The closer a ROC curve is to the upper left corner, the higher the accuracy of the corresponding convergence measure.

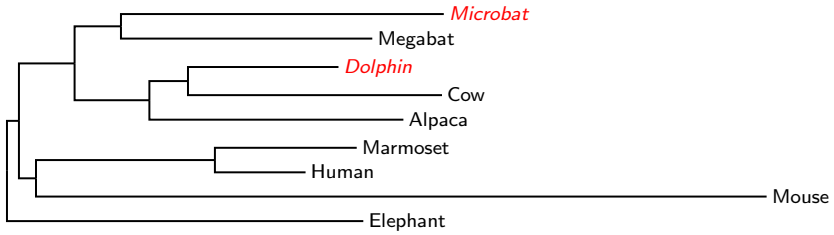

Figure 6: Phylogenetic tree from Thomas and Hahn (2015). Among the nine extant taxa, dolphins and microbats (in red italic) have echolocating abilities.

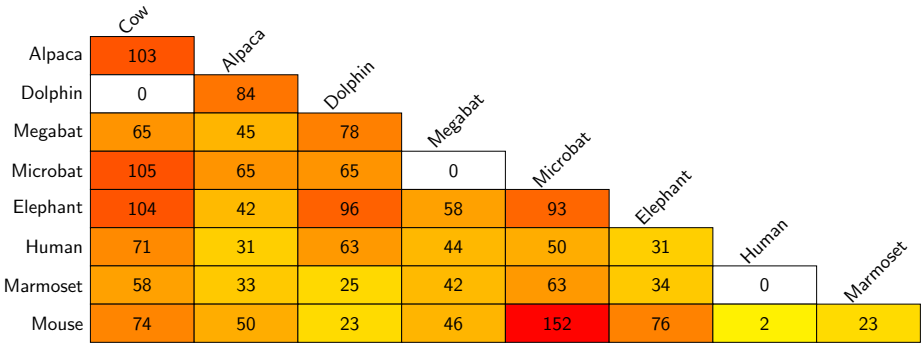

Figure 7: Number of genes detected convergent for all pairs of terminal taxa of the dataset.

according to their p-values and performed a Benjamini-Hochberg correction for multiple-testing (Section SI-2.5). We finally selected the genes with corrected p-values smaller than 5×10^{-2}, which give us the set of convergent genes. The detection of convergent genes was performed between all pairs of terminal taxa in the dataset for control purposes.

Figure 7 displays the number of genes detected as showing convergence between all pairs of terminal taxa. No genes were detected as displaying convergence between pairs of sister taxa (cow-dolphin, microbat-megabat and human-marmoset). But there is a certain number of convergent genes for almost all the other pairs of terminal taxa, ranging from 23 to 152 genes, except for the pair human-mouse which has only 2 convergent genes. We did not detect more genes between the two echolocating taxa than between the other pairs, which is consistent with what was observed by Thomas and Hahn (2015) and with the fact that the gene sample of the dataset is unbiased, notably toward audition.

In order to assess if the convergent genes detected between dolphins and microbats were related to echolocation, we tested their enrichment with regards to GO-terms involved in audition (see Section SI-2.6). We performed the same test for all pairs of taxa, in order to ensure that there is no bias leading to observe more convergence on genes associated with these particular GO-terms. Results are displayed in Figure 8, which shows that the set of convergent genes between dolphins and microbats is by far the most significantly enriched in audition-related annotations. The corresponding Fisher's exact test p-value, i.e., 6.66×10^{-5}, is (at least) three orders of magnitude lower than those of other pairs of taxa.

Note that Prestin (a.k.a. SLC26A5), known to be involved in echolocation (Li *et al.*, 2010; Liu *et al.*, 2010, 2014), is the 3$^{\text{rd}}$ most significant convergent gene out of the $6,332$ ones of the dataset. Still for studying echolocation, Shen *et al.* (2012) screened three genes, namely CDH23, PCDH15 and OTOF, pointing out that "Convergent evolution and expression patterns of OTOF suggest the potential role of nerve and brain in echolocation". The two first genes were not in the dataset of Thomas and Hahn (2015) but Otoferlin (OTOF) was detected as displaying convergen ce with our approach (at the 54$^{\text{th}}$ rank). Davies *et al.* (2012) found signatures of sequence convergence in TMC1 and DFNB59 (aka

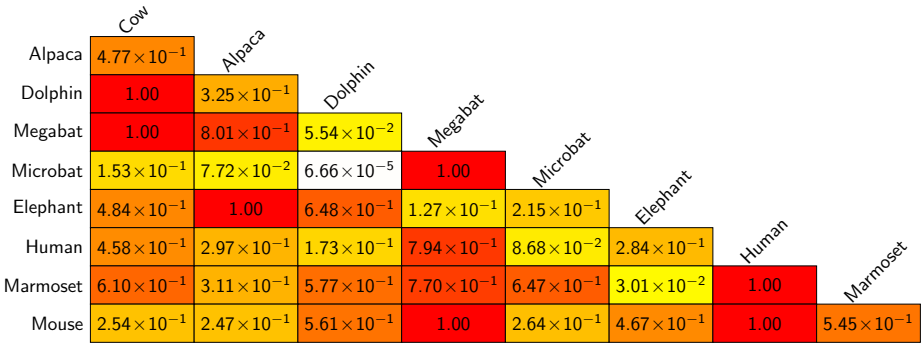

Figure 8: Enrichment p-values, with regard to the "audition-related" GO-terms, of the genes detected convergent for all pairs of taxa of the dataset.

PJVK). These genes are respectively the 5^{th} and 11^{th} most significantly convergent with our method (Supplementary Information). Among the seven genes pointed out by Parker *et al.* (2013) as previously reported for showing convergence and/or adaptation in echolocation, four were present in the dataset. We detected all of them (in bold in Table 1 of Supplementary Information). Table 1 displays the significant GO annotations of detected genes for the echolocating pair, which are related with audition (as defined in Section SI-2.6) and their Fisher's exact test p-values without multiple testing correction with regard to the total number of GO-terms.

Fisher's exact p-value	GO ID	Description	Genes
4.36×10^{-4}	GO:0007605	sensory perception of sound	SLC26A5, TMC1, DFNB59, COL11A1, OTOF
4.59×10^{-3}	GO:0050910	detection of mechanical stimulus involved in sensory perception of sound	TMC1, COL11A1
7.79×10^{-3}	GO:0090102	cochlea development	SLC26A5, OTOF
1.89×10^{-2}	GO:0007420	brain development	PTPRZ1, RELN, MED1, KCNAB1
3.10×10^{-2}	GO:0060117	auditory receptor cell development	TMC1

Table 1: Significant audition-related GO annotations of genes detected convergent between dolphins and microbats.

4 Discussion

There is mounting evidence that phenotypic convergence has a detectable molecular basis for many phenotypic characters, including echolocation (Li *et al.*, 2010; Liu *et al.*, 2010). Despite this fact and the importance of this matter, there is still no consensu al method for the genome-wide identification of the molecular signatures of a given phenotypic convergence. We presented here, first, a new measure for evaluating the extent to which an alignment site supports a phenotypic convergence and, second, a statistical framework for detecting genes displaying significant convergence.

A first result is that our convergence measure has better performance for detecting convergence than the two previous measures on simulated sites. In particular, the convergence measure based on ancestral reconstruction (the "historical" approach) showed poor results on discriminating between convergent and non-convergent simulated sites. This is not a surprise in view of the concerns we listed in Section 2.1 and of the lack of definition of this approach, which basically decides if the amino acids of the convergent taxa are convergent or not, without considering any nuance between the two situations. The ΔSSLS approach yields better results than the ancestral one but is outperformed by the convergence index whatever the alternative tree tested.

Our detection pipeline did not return a greater number of genes between the two echolocators than between the other pairs of taxa. This point was not completely unexpected since dolphins and microbats evolve in very different environments and, at first glance, do not share more phenotypic characters than the other pairs (echolocation excepted). Nevertheless, the fact that the number of convergent sites or genes detected between echolocators was not greater than between other pairs was put forward as an argument against the detectability of the molecular basis of echolocation (Thomas and Hahn, 2015; Zou and Zhang, 2015b). We argue that this point is not conclusive since it is based on the assumption that non-echolocating pairs have no phenotypic convergence (Thomas and Hahn (2015) used the non-echolocating pairs for determining a null distribution). The actual amount of phenotypic and molecular convergence between taxa remains difficult to predict since it may involve phenotypes not obvious to observe (e.g., metabolic pathways, proteins binding etc.). Evaluating the actual extent of convergence between taxa needs further investigations which are out of the scope of the present work. At this point, Figure 7 suggests either that convergence is quite a common mechanism whose molecular traces are detectable, or a possible issue in the approach.

The relevance of the results obtained with our pipeline was assessed with regard to the particular phenotype studied in the dataset. Since echolocation requires special hearing capacities, one expects genes detected as showing convergence between the two echolocating taxa to be, at least for some of them, related to audition. This point is clearly observed in Figure 8 (see also Supplementary Information). On the contrary, Thomas and Hahn (2015) found no evidence of sensory enrichment in genes detected with the ancestral reconstruction approach on the same unbiased dataset. Though Parker *et al.* (2013) observed several hearing genes among the top 5% with the highest ΔSSLS, they did not provide any statistical support for this point (they obtained 117 genes among which only 4 were also detected convergent by our method). They showed that a selection of hearing (and sensory) genes have higher ΔSSLS than expected but

not at a level extremely significant, and without checking the non-echolocating pairs as pointed out by Thomas and Hahn (2015). The significance threshold γ, which is used for deciding if a site is convergent with regard to the empirical distribution associated to its gene, is a crucial parameter of our approach. Its choice is up to the user and relies on what is expected about molecular convergence, i.e., a signal spread out over the sites or concentrated on a few sites. We tested several values of γ from 10^{-3} to 10^{-5}. For all cases except $\gamma = 10^{-3}$, the "audition enrichment" p-value of the set of genes detected between the echolocating pair was at least one order of magnitude lower than that of the other pairs (Supplementary material). Though this is not an absolute rule, the number of detected genes tends to decrease with γ for all pairs of taxa. Significance with regard to audition-related GO-terms peaks at $\gamma = 5 \times 10^{-5}$ for the echolocators but only 36 convergent genes are detected at this level. There are only 7 genes left for $\gamma = 10^{-5}$ and there is no point in considering lower thresholds.

Though we aim at providing a rigorous framework for detecting convergent genes, the relevance of our results heavily depends on our assumptions with regard to protein evolution. Since deciding if a site is convergent relies on simulations from the evolutionary model chosen, the more realistic this model, the more accurate the results. By "evolutionary model", we mean here both the modeling of amino acid substitutions and that of the rate heterogeneity along a gene. The current version of the detection pipeline is based on the widely used model WAG+discretized Gamma distribution as implemented in PAML (Yang, 1994). Any evolutionary model, whatever its level of sophistication, may be easily plugged into the detection pipeline, since it is only used for simulating empirical null distributions. In order to test more realistic evolutionary models, we implemented CAT models (Lartillot and Philippe, 2004, 2006) for computing the null distribution of the convergence index. Since such models somehow constrain the simulated sites to evolve within a subset of amino acids, observing a high convergence index is more likely than under a WAG model (Figures SI-8 and SI-9). In other words, a given level of convergence index is less significant with regard to empirical distribution from CAT models than with regard to that of the WAG model that we used. We performed the same analyses as in Section 3.2 by using CAT models. Results are displayed in Section SI-5. As expected, we generally detected a smaller number of convergent genes between pairs of taxa. Genes detected as showing convergence between dolphins and microbats are still significantly enriched in audition-related GO-terms but to a lesser extent than in Figure 8. Nevertheless, though they are not all detected, the genes previously reported to be involved in echolocation (see Section 3.2) are still among the topmost-convergent genes, which may suggest that the empirical distribution from CAT models could underestimate the significance of some of the sites. This point deserves further investigations and illustrates the importance of the evolutionary model used for simulating the null distribution of the convergence index.

Though there is still room for improvement, the fact that genes we detected as showing convergence for the echolocating pair are annotated with audition-related GO-terms at a significance level far greater than for the other pairs of terminal taxa constitutes a proof of concept that a genome-wide approach may identify the molecular basis of a given phenotypic convergence. Whether such an identification was possible was debated. First, detecting genes displaying convergence is possible only if at least some of the substitutions leading to the

phenotype involve the same sites, thus the same genes, which corresponds to strong constraints on evolution. Second, the preceding condition is not sufficient to ensure that convergent sites are detectable. In a genome-wide context, this also requires a rigorous statistical framework for evaluating the significance of the molecular convergences observed at sites, in other words, a way of distinguishing convergence signal from evolutionary noise. This latter point is not a real concern when genes in which molecular signatures are expected are known *a priori* (Zhang, 2006; Ujvari *et al.*, 2015) but is essential for dealing with thousands of genes.

Since Conte *et al.* (2012) estimated that phenotypic convergence involves the same genes in around a third to half of the cases, the numerous occurrences of phenotypic convergence observed in natural settings constitutes a huge dataset that can be used for studying relations between genotypes and phenotypes.

Acknowledgement

All authors wish to thank Michel Laurin for many helpful suggestions, Gregg W.C. Thomas and Matthew W. Hahn for the quality of their dataset and for kindly providing it, Magdalen Lardière for correcting an earlier version of the manuscript and two anonymous referees as well as the editors of the journal for their careful reading and their useful comments and suggestions. GD thanks Bastien Boussau for many discussions about evolutionary convergence.

Authors contributions

OC contributed to the software development, fetched the data and ran the tests. MR-C provided statistical insights and a careful reading of the manuscript. PP produced the biological side of this work, provided evolutionary insights all along its development and most of the references. GD provided the original idea of the detection approach, supervised its development and its evaluation, derived the mathematical part, led the software development and wrote the manuscript. All authors read and approved the final manuscript.

References

Arbuckle, K. and Speed, M. P. (2016). Analysing Convergent Evolution: A Practical Guide to Methods. In P. Pontarotti, editor, *Evolutionary Biology: Convergent Evolution, Evolution of Complex Traits, Concepts and Methods*, pages 23–36, Cham. Springer International Publishing.

Arbuckle, K., Bennett, C. M., and Speed, M. P. (2014). A simple measure of the strength of convergent evolution. *Methods in Ecology and Evolution*, **5**(7), 685–693.

Castoe, T. A., de Koning, A. P. J., Kim, H.-M., Gu, W., Noonan, B. P., Naylor, G., Jiang, Z. J., Parkinson, C. L., and Pollock, D. D. (2009). Evidence for an ancient adaptive episode of convergent molecular evolution. *Proceedings of the National Academy of Sciences*, **106**(22), 8986–8991.

Conte, G. L., Arnegard, M. E., Peichel, C. L., and Schluter, D. (2012). The probability of genetic parallelism and convergence in natural populations. *Proceedings of the Royal Society of London B: Biological Sciences*.

Davies, K. T. J., Cotton, J. A., Kirwan, J. D., Teeling, E. C., and Rossiter, S. J. (2012). Parallel signatures of sequence evolution among hearing genes in echolocating mammals: an emerging model of genetic convergence. *Heredity*, **108**(5), 480–489.

Davis, M. P., Sparks, J. S., and Smith, W. L. (2016). Repeated and Widespread Evolution of Bioluminescence in Marine Fishes. *PloS one*, **11**(6), e0155154.

Foote, A. D., Liu, Y., Thomas, G. W., Vinař, T., Alföldi, J., Deng, J., Dugan, S., van Elk, C. E., Hunter, M. E., Joshi, V., and others (2015). Convergent evolution of the genomes of marine mammals. *Nature genetics*, **47**(3), 272–275.

Friedman, S. T., Price, S. A., Hoey, A. S., and Wainwright, P. C. (2016). Ecomorphological convergence in planktivorous surgeonfishes. *Journal of Evolutionary Biology*, **29**(5), 965–978.

Gallant, J. R., Traeger, L. L., Volkening, J. D., Moffett, H., Chen, P.-H., Novina, C. D., Phillips, G. N., Anand, R., Wells, G. B., Pinch, M., Güth, R., Unguez, G. A., Albert, J. S., Zakon, H. H., Samanta, M. P., and Sussman, M. R. (2014). Genomic basis for the convergent evolution of electric organs. *Science*, **344**(6191), 1522–1525.

Ingram, T. and Mahler, D. (2013). Surface: detecting convergent evolution from comparative data by fitting Ornstein-Uhlenbeck models with stepwise Akaike information criterion. *Methods in Ecology and Evolution*, **4**(5), 416–425.

Lartillot, N. and Philippe, H. (2004). A bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Molecular Biology and Evolution*, **21**(6), 1095–1109.

Lartillot, N. and Philippe, H. (2006). Computing bayes factors using thermodynamic integration. *Systematic Biology*, **55**(2), 195–207.

Li, Y., Liu, Z., Shi, P., and Zhang, J. (2010). The hearing gene Prestin unites echolocating bats and whales. *Current Biology*, **20**(2), R55 – R56.

Liu, Y., Cotton, J. A., Shen, B., Han, X., Rossiter, S. J., and Zhang, S. (2010). Convergent sequence evolution between echolocating bats and dolphins. *Current Biology*, **20**(2), R53 – R54.

Liu, Z., Qi, F.-Y., Zhou, X., Ren, H.-Q., and Shi, P. (2014). Parallel Sites Implicate Functional Convergence of the Hearing Gene Prestin among Echolocating Mammals. *Molecular Biology and Evolution*, **31**(9), 2415.

Losos, J. (2009). *Lizards in an evolutionary tree: ecology and adaptive radiation of anoles*, volume 10. Univ of California Press.

Losos, J. B., Jackman, T. R., Larson, A., de Queiroz, K., and Rodriguez-Schettino, L. (1998). Contingency and determinism in replicated adaptive radiations of island lizards. *Science*, **279**(5359), 2115–2118.

Loughin, T. M. (2004). A systematic comparison of methods for combining p-values from independent tests. *Computational Statistics & Data Analysis*, **47**(3), 467 – 485.

Mahler, D. L., Ingram, T., Revell, L. J., and Losos, J. B. (2013). Exceptional Convergence on the Macroevolutionary Landscape in Island Lizard Radiations. *Science*, **341**(6143), 292–295.

Parker, J., Tsagkogeorga, G., Cotton, J. A., Liu, Y., Provero, P., Stupka, E., and Rossiter, S. J. (2013). Genome-wide signatures of convergent evolution in echolocating mammals. *Nature*, **502**(7470), 228–231.

Pfenning, A. R., Hara, E., Whitney, O., Rivas, M. V., Wang, R., Roulhac, P. L., Howard, J. T., Wirthlin, M., Lovell, P. V., Ganapathy, G., Mountcastle, J., Moseley, M. A., Thompson, J. W., Soderblom, E. J., Iriki, A., Kato, M., Gilbert, M. T. P., Zhang, G., Bakken, T., Bongaarts, A., Bernard, A., Lein, E., Mello, C. V., Hartemink, A. J., and Jarvis, E. D. (2014). Convergent transcriptional specializations in the brains of humans and song-learning birds. *Science*, **346**(6215).

Pontarotti, P. and Hue, I. (2016). Road map to study convergent evolution: A proposition for evolutionary systems biology approaches. In P. Pontarotti, editor, *Evolutionary Biology: Convergent Evolution, Evolution of Complex Traits, Concepts and Methods*, pages 3–21, Cham. Springer International Publishing.

Revell, L. J., Johnson, M. A., Schulte, J. A., Kolbe, J. J., and Losos, J. B. (2007). A phylogenetic test for adaptive convergence in rock-dwelling lizards. *Evolution*, **61**(12), 2898–2912.

Shen, Y.-Y., Liang, L., Li, G.-S., Murphy, R. W., and Zhang, Y.-P. (2012). Parallel evolution of auditory genes for echolocation in bats and toothed whales. *PLoS Genet*, **8**(6), e1002788.

Speed, M. P. and Arbuckle, K. (2017). Quantification provides a conceptual basis for convergent evolution. *Biological Reviews*, **92**(2), 815–829.

Stayton, C. T. (2015a). The definition, recognition, and interpretation of convergent evolution, and two new measures for quantifying and assessing the significance of convergence. *Evolution*, **69**(8), 2140–2153.

Stayton, C. T. (2015b). What does convergent evolution mean? the interpretation of convergence and its implications in the search for limits to evolution. *Interface Focus*, **5**(6).

Storz, J. F. (2016). Causes of molecular convergence and parallelism in protein evolution. *Nat Rev Genet*, **17**(4), 239–250.

Thomas, G. W. and Hahn, M. W. (2015). Determining the null model for detecting adaptive convergence from genomic data: a case study using echolocating mammals. *Molecular biology and evolution*, **32**(5), 1232–1236.

Ujvari, B., Casewell, N. R., Sunagar, K., Arbuckle, K., Wüster, W., Lo, N., O'Meally, D., Beckmann, C., King, G. F., Deplazes, E., and Madsen, T. (2015). Widespread convergence in toxin resistance by predictable molecular evolution. *Proceedings of the National Academy of Sciences*, **112**(38), 11911–11916.

Vidal-García, M. and Keogh, J. S. (2015). Convergent evolution across the australian continent: ecotype diversification drives morphological convergence in two distantly related clades of australian frogs. *Journal of Evolutionary Biology*, **28**(12), 2136–2151.

Whelan, S. and Goldman, N. (2001). A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Molecular biology and evolution*, **18**(5), 691–699.

Wilkinson, B. (1951). A statistical consideration in psychological research. *Psychological bulletin*, **48**(2), 156.

Yang, Z. (1994). Maximum likelihood phylogenetic estimation from dna sequences with variable rates over sites: Approximate methods. *Journal of Molecular Evolution*, **39**(3), 306–314.

Yokoyama, S., Altun, A., and DeNardo, D. F. (2011). Molecular convergence of infrared vision in snakes. *Molecular Biology and Evolution*, **28**(1), 45–48.

Zhang, J. (2006). Parallel adaptive origins of digestive RNases in Asian and African leaf monkeys. *Nat Genet*, **38**(7), 819–823.

Zhang, J. and Kumar, S. (1997). Detection of convergent and parallel evolution at the amino acid sequence level. *Molecular Biology and Evolution*, **14**(5), 527–536.

Zhou, X.-H., McClish, D. K., and Obuchowski, N. A. (2009). *Statistical methods in diagnostic medicine*, volume 569. John Wiley & Sons.

Zou, Z. and Zhang, J. (2015a). Are convergent and parallel amino acid substitutions in protein evolution more prevalent than neutral expectations? *Molecular biology and evolution*, **32**(8), 2085–2096.

Zou, Z. and Zhang, J. (2015b). No genome-wide protein sequence convergence for echolocation. *Molecular Biology and Evolution*, **32**(5), 1237–1241.

Zou, Z. and Zhang, J. (2016). Morphological and molecular convergences in mammalian phylogenetics. *Nature Communications*, **7**.