



**HAL**  
open science

# Accurate Detection of Convergent Mutations in Large Protein Alignments With ConDor

Marie Morel, Anna Zhukova, Frédéric Lemoine, Olivier Gascuel

► **To cite this version:**

Marie Morel, Anna Zhukova, Frédéric Lemoine, Olivier Gascuel. Accurate Detection of Convergent Mutations in Large Protein Alignments With ConDor. *Genome Biology and Evolution*, 2024, 16 (4), pp.evae040. 10.1093/gbe/evae040 . pasteur-04601784v2

**HAL Id: pasteur-04601784**

**<https://pasteur.hal.science/pasteur-04601784v2>**

Submitted on 5 Jun 2024


**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NonCommercial 4.0 International License

# Accurate Detection of Convergent Mutations in Large Protein Alignments With ConDor

Marie Morel<sup>1,2</sup>, Anna Zhukova<sup>1,3</sup>, Frédéric Lemoine <sup>1,3,4,\*</sup>, and Olivier Gascuel <sup>1,5,\*</sup>

<sup>1</sup>Institut Pasteur, Université Paris Cité, Unité Bioinformatique Evolutive, Paris, France

<sup>2</sup>Université Claude Bernard Lyon 1, LBBE, UMR 5558, CNRS, VAS, Villeurbanne, 69100, France

<sup>3</sup>Institut Pasteur, Université Paris Cité, Bioinformatics and Biostatistics Hub, Paris, France

<sup>4</sup>Institut Pasteur, Université Paris Cité, CNR Virus Des Infections Respiratoires, Paris, France

<sup>5</sup>Institut de Systématique, Evolution, Biodiversité (UMR 7205—CNRS, Muséum National d'Histoire Naturelle, SU, EPHE, UA), Paris, France

\*Corresponding authors: E-mails: frederic.lemoine@pasteur.fr; olivier.gascuel@mnhn.fr.

Accepted: February 22, 2024

## Abstract

Evolutionary convergences are observed at all levels, from phenotype to DNA and protein sequences, and changes at these different levels tend to be correlated. Notably, convergent mutations can lead to convergent changes in phenotype, such as changes in metabolism, drug resistance, and other adaptations to changing environments. We propose a two-component approach to detect mutations subject to convergent evolution in protein alignments. The “Emergence” component selects mutations that emerge more often than expected, while the “Correlation” component selects mutations that correlate with the convergent phenotype under study. With regard to Emergence, a phylogeny deduced from the alignment is provided by the user and is used to simulate the evolution of each alignment position. These simulations allow us to estimate the expected number of mutations in a neutral model, which is compared to the observed number of mutations in the data studied. In Correlation, a comparative phylogenetic approach, is used to measure whether the presence of each of the observed mutations is correlated with the convergent phenotype. Each component can be used on its own, for example Emergence when no phenotype is available. Our method is implemented in a standalone workflow and a webserver, called ConDor. We evaluate the properties of ConDor using simulated data, and we apply it to three real datasets: sedge PEPC proteins, HIV reverse transcriptase, and fish rhodopsin. The results show that the two components of ConDor complement each other, with an overall accuracy that compares favorably to other available tools, especially on large datasets.

**Key words:** molecular evolution, phylogenetics, selection, adaptation, convergence, C4 metabolism, HIV, resistance to drugs, rhodopsin.

## Significance

Many examples of evolutionary convergence are known, such as the appearance of wings in insects, birds, and bats. The objective here is to detect mutations at the molecular level that could explain these convergent phenotypes. The proposed method allows the analysis of large sets of homologous proteins, it gives very good results on the tested datasets, and the software is freely available, notably via a website.

© The Author(s) 2024. Published by Oxford University Press on behalf of Society for Molecular Biology and Evolution.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (<https://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact reprints@oup.com for reprints and translation rights for reprints. All other permissions can be obtained through our RightsLink service via the Permissions link on the article page on our site—for further information please contact journals.permissions@oup.com.

## Introduction

Convergent evolution is often defined as the independent acquisition of similar traits in distinct lineages over the course of evolution (Arendt and Reznick 2008; Losos 2011; Stern 2013). The studied traits can be behavioral, morphological, molecular, etc. In each category, traits can be quantitative (size, length, weight, etc.), binary (presence or absence of a given phenotype), or categorical (a trait is subdivided into several categories). The presence of convergence, especially at the phenotypic level, is often seen as evidence of adaptation in the sense that similar evolutionary paths were found in response to the same evolutionary constraints (Castoe et al. 2009; Losos 2011). Many studies focus on the molecular level, assuming that convergent phenotypes may result from the same genetic changes (Stern 2013; Rosenblum et al. 2014; Storz 2016). At the protein level, it is common to distinguish (Zhang and Kumar 1997) between parallel mutations (a change toward the same amino acid is observed from the same ancestral amino acid), convergent mutations (change toward the same amino acid, from different ancestral amino acids), and reversions (mutations that restore an amino acid previously lost during evolution). For the sake of simplicity, in what follows we will refer to these three types of mutations as “convergent mutations”, unless explicitly stated.

Examples of evolutionary convergence at the molecular level have been demonstrated in eukaryotes, related to adaptation to certain environments (Muschick et al. 2012; Foll et al. 2014; Foote et al. 2015; Hill et al. 2019; Lu et al. 2020; Xu et al. 2020), diet (Zhang 2006; Zhen et al. 2012; Ujvari et al. 2015; Hu et al. 2017), changes in metabolism (Besnard et al. 2009; Parto and Lartillot 2018), morphological transformations (Larter et al. 2018), and acquisition of new abilities (Davies et al. 2012; Parker et al. 2013; Lee et al. 2018; Marcovitz et al. 2019; Chai et al. 2020). Similarly, when submitted to constraints such as harsh experimental conditions and drug treatments, viruses and microorganisms adapt and are likely to exhibit similar escapes. This has been demonstrated in HIV after exposure to antiviral drug treatments in several patients (Crandall et al. 1999) and within a single treated patient (Holmes et al. 1992). Similarly, Cuevas et al. (2002) found adaptive convergence in experimental populations of RNA viruses, and van Ditmarsch et al. (2013) in pathogenic bacteria. In natural conditions, evolutionary convergence was found in viruses having experienced host shifts (Longdon et al. 2018; Escalera-Zamudio et al. 2020; Martin et al. 2021a) and changes in vector specificity (Tsetsarkin et al. 2007).

Several methods have been developed to detect convergent evolution at the molecular level (Zhang and Kumar 1997; Zhang 2006; Tamuri et al. 2009; Parker et al. 2013; Thomas and Hahn 2015; Zou and Zhang 2015a; Parto

and Lartillot 2017; Chabrol et al. 2018; Rey et al. 2018). Most of them are based on prior knowledge of a convergent phenotype and aim to identify the protein mutations underlying the phenotypic trait studied. However, they differ in the scale at which molecular convergence is sought and the definition of what a convergent mutation is.

Some approaches aim to identify which coding genes harbor mutations supporting a convergent phenotype, while others study which amino acid changes can explain convergent changes at the scale of a single protein. Methods of the first category are commonly applied to eukaryotic and prokaryotic genomes and perform genome-wide analyses to detect convergent genes by considering simultaneously all positions of the corresponding protein sequences; for example, the methods developed by Parker et al. (2013), Zou and Zhang (2015b), Thomas and Hahn (2015) and Chabrol et al. (2018) were applied to the search of genes responsible for echolocation in mammals. In the second configuration, the coding genes responsible for the convergent phenotype have already been identified and the methods focus on the detection of convergent evolution at the position level; for example, Zhang and Kumar (1997) identified convergent and parallel mutations in stomach lysozyme sequences of foregut fermenters. Similarly, Zhang (2006) found parallel substitutions in colobine pancreatic ribonucleases, and Rey et al. (2018) found positions with convergent substitutions in the PEPC protein occurring jointly with the transition toward C4 metabolism in sedges. In fact, testing the significance of convergent changes at individual protein positions has many potential applications. In the case of complex eukaryotic and bacterial organisms, there are few examples of a single amino acid change that could explain a convergent phenotype (Storz 2016). However, in the case of viruses with rapid evolution, and whose (small) genomes are strongly constrained, only a few amino acid changes are generally possible at a given position (Pond et al. 2012) and position-wise convergent evolution is expected to be relatively frequent (Gutierrez et al. 2019). Determining molecular changes that deviate from what is expected by chance can thus be indicative of adaptive phenomena. This was the case for SARS-CoV-2, where one first identified mutations in the Spike protein, which were spreading within the viral population and appeared multiple times independently, before being demonstrated to be evolutionarily advantageous for the virus (Korber et al. 2020; van Dorp et al. 2020; Martin et al. 2021b). Note, however, that mutations that were initially thought to be adaptive were eventually shown to be simply the result of founder events (Hodcroft et al. 2021), demonstrating the difficulty of detecting convergent mutations without access to the phenotype.

Most importantly, different methods have different ways of selecting which mutations underlie the studied

convergent phenotype. In the most intuitive definition, one aims to detect mutations toward the same amino acid, which occurred in all clades with the convergent phenotype. This is the definition used first in Zhang and Kumar (1997) and then in Zhang (2006), Foote et al. (2015), Thomas and Hahn (2015), and Zou and Zhang (2015b). An extension was proposed by Chabrol et al. (2018), where the convergent amino acid may only be found in a subset of the convergent species, as well as in some nonconvergent species. Considering that a change toward the same amino acid may be too strict since several amino acids have similar physicochemical properties, Rey et al. (2018) relaxed this constraint in the PCOC program, by considering changes in amino acid profiles (Le et al. 2008a). Their work on amino acid profiles follows previous works aimed at detecting positions under condition-dependent selection, but which did not focus solely on convergent evolution (Tamuri et al. 2009; Parto and Lartillot 2017, 2018). A radically different approach, proposed by Parker et al. (2013) and inspired from Castoe et al. (2009), relies on the fact that convergence can lead to errors in phylogenetic reconstruction by artificially bringing convergent species together. These authors proposed selecting positions that best support the phylogeny that groups species with the convergent phenotype together, rather than the species tree (but see the critiques of this method by Thomas and Hahn 2015 and Zou and Zhang 2015b).

One of the main challenges in detecting molecular convergence is to identify only the convergent mutations that are linked to the studied convergent phenotype. In their review of methods for detecting molecular convergence, Rey et al. (2019) referred to this type of mutation as foreground convergence (or foreground convergent mutations) in opposition to background convergence which is unrelated to the convergent phenotype. Indeed, at the molecular level, one can find patterns of convergent mutations linked to another convergent phenotype, or occurring because of mutational biases, protein conformation limitations, constraints at the molecular level and epistatic forces (Zhang and Kumar 1997; Rokas and Carroll 2008; Storz 2016; Stoltzfus and McCandlish 2017). It has been shown that most (if not all) substitution models may fail at distinguishing between foreground convergent mutations and background ones (also called nonadaptive convergent mutations), especially in close taxa between highly exchangeable amino acids, and on fast-evolving sites (Goldstein et al. 2015; Zou and Zhang 2015a). In other words, finding multiple independent mutations resulting in the same (or a similar) amino acid should be tested carefully, even when the number of such mutations appears to be high. We shall see that our findings tend to confirm this.

Another difficulty is the definition of the convergent phenotype and the annotation of taxa that do or do not have this phenotype. For example, in the case of viruses,

we usually do not know the exact phenotype, but use a proxy instead. In the case of drug resistance mutations (DRMs) that occur repeatedly in different patients treated with antiviral drugs, we use the treatment status as a proxy for the resistance status. Although we expect that most (but not all, e.g. due to poor adherence) sequences from patients who fail drug treatment will contain resistance mutations, we also expect that some DRMs will be found in untreated (naive) patients in the case of resistance transmission (Blassel et al. 2021b). Similarly, environmental constraints are not strictly speaking phenotypes, but act as selective forces that can lead to phenotypic and molecular convergence. However, we do not expect all organisms living under the same environmental conditions to show the same phenotype and recurrent mutations.

In some respects, the identification of convergent mutations has similarities with the detection of positions under positive selection (Goldman and Yang 1994). The idea is indeed to identify mutations that might be advantageous, as they are found more often than expected in a neutral (or purifying) model of evolution. In the positive selection framework, these mutations can be directed to a specific amino acid (directional), or correspond to any change that differs from the original amino acid (diversifying). This is the case, for example, with immune avoidance where mutations toward any new amino acid at antigenic sites are generally favorable and positively selected. Conversely, in the case of convergent evolution, we are interested in substitutions toward one or a few similar amino acids, in the branches leading to the convergent taxa (Starr et al. 2020; Bloom and Neher 2023). Thus, a large number of nonsynonymous substitutions on convergent positions are expected, but the criterion of positive (or relaxed purifying) selection alone is not sufficient to assert convergence, as our results with drug resistance in HIV show. In fact, several authors have already noted this limitation of positive selection approaches in convergence detection in HIV (Crandall et al. 1999; Lemey et al. 2005), which motivated the development of directional approaches, including EDEPS and MEDS (Murrell, Oliveira, et al. 2012) now replaced by FADE in the HyPhy suite (Pond et al. 2005). These methods test whether positions in a protein alignment are subject to directional selection (or mutational bias) within a specified set of “foreground” branches that typically correspond to convergent taxa. These tools thus have their roots in positive selection approaches, but are closely related to convergence detection.

Here, we propose a new method for detecting convergent evolution at the position (or site) scale in large amino acid alignments, while relaxing the constraint that convergent mutations must be found only in organisms with the convergent phenotype and in all of them. Our method does not require specifying the branches where molecular convergence occurred (as with PCOC and FADE, for

example), which is a complex step, especially with large datasets and when using a proxy for the phenotypic convergence. The taxa are simply annotated as convergent or nonconvergent, and the mutations associated with this status are then detected. We are interested in mutations leading to a target amino acid, regardless of the ancestral amino acids at this position. In other words, parallel, convergent mutations and reversions are considered indifferently, and we consider mutations resulting in different target amino acids as different events. With this definition our method is in line with methods aiming at detecting changes toward the same amino acid, as opposed to detecting changes in profiles (Rey et al. 2018). Indeed, there are many examples of known convergent mutations, where the changes involve highly exchangeable amino acids that have very similar biochemical profiles. For example with drug resistance in HIV, there are convergent mutations from Isoleucine to Valine and from Tyrosine to Phenylalanine (the two most exchangeable amino acid pairs, cf. BLOSUM62) that confer resistance to certain drugs (Wensing et al. 2019).

In the following sections, we describe this approach, which is implemented in a workflow called ConDor (for Convergence Detector), available as a web service ([condor.pasteur.cloud](https://condor.pasteur.cloud)) and as a standalone workflow. We assess its properties under different conditions using simulated data and evaluate its performance on three real datasets involving sedge PEPC protein, HIV reverse transcriptase, and fish rhodopsin. The results are compared to those of PCOC and FADE, which are based on different assumptions.

## New Approaches

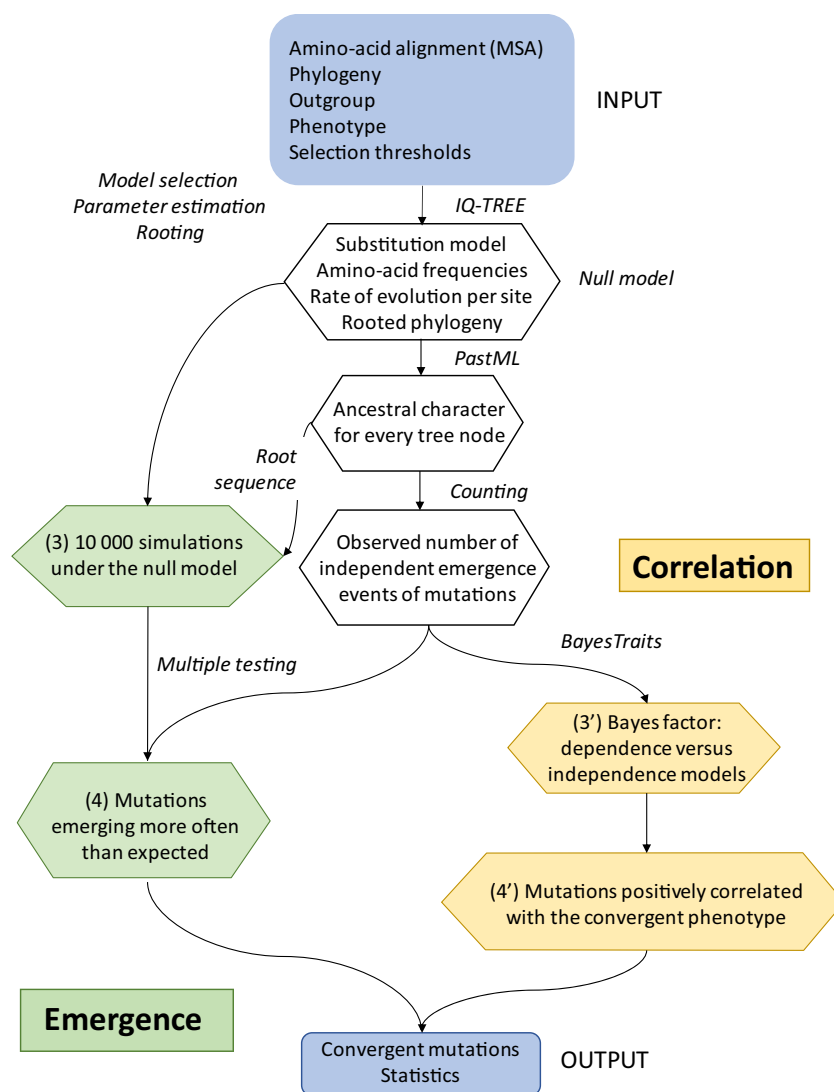
### Method Overview

Any method for detecting molecular convergence relies on a definition of what a convergent mutation is. Our definition is a form of common sense: convergent mutations have emerged several times in independent lineages, have occurred more frequently than expected by chance, and their presence is correlated with the presence of the phenotype of interest in the taxa studied. Our method is subdivided into two independent and complementary components: (1) the “Emergence” component that detects mutations emerging more often than expected in a neutral (or null) substitution model and (2) the “Correlation” component that identifies mutations that are positively correlated with the convergent phenotype. The combination of the two components accurately identifies amino acid mutations resulting from convergent evolution associated with the convergent phenotype (foreground mutations), although it is also possible to execute and interpret the results of the two components independently.

A representation of the ConDor workflow is shown in Fig. 1. Inputs are constituted of: (i) a multiple protein

sequence alignment (MSA); (ii) a phylogeny; (iii) an outgroup; (iv) the phenotype of each of the taxa; and (v) user supplied thresholds to select convergent mutations. The quality of the MSA and phylogeny is critical and should be carefully checked by users of the method, as, for example, running ConDor on poorly aligned sites of an MSA can lead to poor results and incorrect conclusions. The input phylogeny is intended to be the tree deduced from the input MSA (not a species tree), possibly using nucleotide-level sequences (rather than amino acids) for topology inference, when it is suspected that convergent mutations could affect the topology and group together the convergent taxa and clades. The two first steps of the workflow (Fig. 1) are common to the Emergence and Correlation components. In step (1), we estimate the parameters of the null model from the MSA and input phylogeny (parameters of the amino acid substitution model, ML-based branch lengths of the phylogeny, evolutionary rate per position, etc.). In step (2), we reconstruct the substitution history for each position in the MSA and count the number of emergence events of mutation (EEMs) observed for every mutation present at that position in at least  $m$  sequences. Mutations with less than  $n$  EEMs are not processed further. Both  $m$  and  $n$  are user-defined ( $m \geq n$ ), depending on the dataset, to avoid testing too many mutations and losing statistical power (and wasting computing time). In the Emergence component, for each position and mutation of interest, the two main steps are: (3) simulation of new datasets in the null model and counting of simulated EEMs; (4) comparison of observed and simulated numbers of EEM to identify the mutations that occurred significantly more often than expected assuming the null model. The Correlation component is applied mutation by mutation within each position. In step (3'), we compute the log Bayes factor of the model assuming a dependence between the presence/absence of the phenotype and the presence/absence of the given mutation, versus the model assuming their independence, using BayesTraits (Pagel 1994; Pagel and Meade 2006). In step (4'), we select among the significantly correlated mutations (user-defined threshold) those that are positively correlated with the phenotype. The final step (5) combines the results of steps (4) and (4') and provides a list of potential convergent mutations. The results of steps (4) and (4') are also provided to the user and can be interpreted independently. Users can thus keep a mutation selected by Correlation and with several EEMs even if the statistical significance of Emergence is not particularly high. Reciprocally, when a mutation selected by Correlation emerged very few times (e.g. only once, see examples below), it can be rejected as convergent, even if it is significantly correlated with the phenotype. Note that step (4) does not require knowledge of the phenotype (or a proxy for it, as with DRMs in HIV). For all selected mutations, ConDor provides the evolutionary rate of the corresponding position, the nature of the mutation (convergent, parallel, revertant), the number of EEMs,





**FIG. 1.**—Flowchart of the method. The method takes as input an amino acid alignment as well as the corresponding phylogeny and phenotype metadata. The MSA and phylogeny are used for inference of the null model (branch lengths, substitution model and its parameters, evolutionary rate per site, etc.) and ACR. In the Emergence component, the tree and root sequence are used to simulate 10,000 alignments under the null model; the output is the list of mutations that emerged more often in the input alignment than in the simulations. In the Correlation component, we select mutations that are positively correlated with the phenotype. The combination of the two components gives the list of mutations proposed as convergent by ConDor, but the results of both components are provided to users and can be analyzed independently.

the genetic barrier (minimum number of mutations at the DNA level), the relative rate of substitution between the two amino acids, etc. All these statistics are described in the user guide (<https://condor.pasteur.cloud/help>) and can be used to further analyze the results and select the most relevant mutations, depending on the dataset studied and the nature of the mutations (e.g. revertant, which emerged less often than convergent and parallel mutations, see below).

The null model and all its parameters are inferred from the input alignment and the phylogeny using ModelFinder (Kalyaanamoorthy et al. 2017) and IQ-TREE (Nguyen et al. 2015). The selected substitution model, along with amino

acid frequencies, rates-across-sites distribution parameters, branch lengths, and evolutionary rate per site are assumed to represent the data without convergence. We make this assumption because using large alignments (>1,000 sequences), we consider that mutations resulting from convergent evolution are rare enough to have a negligible influence on parameter inference. The phylogeny with optimized branch lengths is then rooted using the user supplied outgroup. This is necessary to infer the ancestral sequence at the root of the tree, run simulations starting from this sequence, and count simulated EEMs. Ancestral character reconstruction (ACR) for positions with mutations of interest



uncertainty (e.g. two amino acids with posteriors of 0.55 and 0.45). After the simulation of sequence evolution along the tree, we count the number of EEMs (10,000 values per position and per mutation of interest) using the algorithm detailed above. Consider, for example, the M41L mutation from our HIV dataset, in which a methionine (M) is substituted by a leucine (L) at position 41 in 211 sequences. The observed number of EEMs toward L is 47, which is smaller than 211 as in some subtrees all tips have L, corresponding to only 1 EEM. Note that in this example, the ancestral amino acid is always M, but we would have considered any ancestral amino acid in the counting of the EEMs toward L (see algorithm above). Then, 47 is compared to the distribution of the number of EEMs toward L (always starting from an M at the tree root since there is no ambiguity in ACR with this example), among 10,000 simulations in the null model; this distribution ranges from 0 to 31 with an average of 12. From the observed number of EEMs and the distribution of simulated EEMs, we estimate a *P*-value. To avoid zero *P*-values when all simulations result in fewer EEMs than the observed EEM, we use a pseudo-count of 0.5, which means that the (uncorrected) *P*-value is equal to  $(0.5 + \text{number of simulated EEMs} \geq \text{observed EEM})/10,001$  ( $\approx 5 \times 10^{-5}$  in our M41L example). Since we test many positions and mutations, we use the Holm–Bonferroni method (Holm 1979) to correct for multiple testing, with a default rejection threshold of 10% (user adjustable). We consider that mutations passing threshold after *P*-value correction did not occur by chance. These mutations can be studied on their own in the absence of an identified phenotype. However, we know from previous studies that background convergent mutations in real data tend to be more frequent than expected under any available substitution model, due to model approximations, epistatic constraints, etc. (Rokas and Carroll 2008; Castoe et al. 2009; Goldstein et al. 2015; Zou and Zhang 2015a). Moreover, some of these highly frequent mutations may be truly adaptive and convergent, but for other phenotypic traits than the one studied. Thus, we expect that a significant fraction of these frequent mutations are false positives (FP) for the studied phenotype. The Correlation component complements the Emergence component to focus on mutations that correlate positively with the phenotype, that is, foreground convergent mutations.

### Correlation With the Convergent Phenotype

The Correlation component of ConDor is based on the “Discrete” method from BayesTraits (Pagel 1994; Pagel and Meade 2006), which combines Markovian modeling of trait evolution and Bayesian model comparison, to distinguish between the two hypotheses of independent ( $H_0$ ) versus dependent ( $H_1$ ) evolution of two traits along a phylogeny. Here, we apply BayesTraits to the analysis of two

binary traits: presence/absence (1/0) of the mutation and convergent/nonconvergent phenotype (1/0). If, for a given position, there are several mutations toward different amino acids that meet the conditions of number of EEMs (*n*) and number of sequences (*m*), the correlation of the phenotype with each of these amino acids will be explored. For each of the hypotheses (corresponding to different evolutionary models), the marginal log-likelihood (approximated by the harmonic mean of the likelihoods after several millions of iterations) is calculated using a stepping stone sampler. BayesTraits then calculates the log Bayes factor (logBF) to decide if the ( $H_1$ ) dependence hypothesis is supported:

$$\log\text{BF} = 2 \log \left[ \frac{\text{MarginalLikelihood}(H_1)}{\text{MarginalLikelihood}(H_0)} \right]$$

As described in the BayesTraits manual ([www.evolution.reading.ac.uk/Files/BayesTraits-V1.0-Manual.pdf](http://www.evolution.reading.ac.uk/Files/BayesTraits-V1.0-Manual.pdf)), a “logBF greater than 2 is considered as ‘positive’ evidence, greater than 5 is ‘strong’ and greater than 10 is ‘very strong’ evidence”. To take into account the different sizes of the datasets and according to the results on the variation of the thresholds (supplementary tables S1, S2, S5, and S8, Supplementary Material online), we chose different thresholds for logBF (2 for the sedge PEPC dataset, with 78 sequences, and 20 for the other datasets, with >1,000 sequences). It should be noted, however, that the logBF value is the result of a stochastic procedure and may vary slightly from trial to trial (e.g. between ~60 and ~64 with a median of ~63, among 10 trials for the M41L mutation in HIV), which means that visual inspection is desirable for mutations that are close to the threshold and for making an appropriate decision. For the mutations that pass the threshold, we determine the direction of the correlation: is the presence of the mutation favored (i.e. more frequent) in the convergent taxa (positive correlation) or in the nonconvergent taxa (negative correlation)? For all the analyses presented here, we retained only positive correlations, corresponding to convergent molecular adaptations to the phenotype of interest (e.g. drug resistance in HIV). However, with the rhodopsin dataset, we were interested in adaptations to both environmental conditions (marine versus brackish/fresh water), and thus launched ConDor (and the other programs tested) twice, with each condition in turn considered “convergent”.

BayesTraits has been widely used in evolutionary biology and ecology to test correlations among behavioral, morphological, genetic and cultural characters, and for predicting functional gene linkages (Barker and Pagel 2005). To our knowledge, it has not been used to detect evolutionary convergence. One of the main advantages of this method is that it takes into account the phylogenetic correlation between taxa (as opposed to simple association tests, such



as Fisher's exact test that is commonly used for the detection of DRMs in HIV; Blassel et al. 2021b). Furthermore, it does not force the emergence of molecular convergence in all species with the convergent phenotype, as does the "One Change" (OC) model of PCOC, for example (Rey et al. 2018). This characteristic is especially important as in most analyses we do not know the exact phenotype, but use a proxy. However, it should be kept in mind that the Correlation component in isolation can identify mutation events that fall outside the scope of convergent evolution. For example, a perfect correlation between a mutation and phenotype can arise from a single mutation event which is then propagated to all the taxa of the corresponding subtree (a so-called "founder" event; Bhattacharya et al. 2007; Gutierrez et al. 2019). We will detail such mutations with the rhodopsin dataset, where a unique EEM is associated to a highly significant logBF. Although such a mutation may be of interest, it cannot be qualified as a convergent mutation since it has only occurred once, hence the need to combine Correlation with EEMs counting as implemented in Emergence.

### Assessment of ConDor Using Simulated Data

In this section, we study ConDor's properties and performance on simulated datasets where the convergence mutations are known. These datasets allow us to characterize ConDor's behavior under different conditions and facilitate the interpretation of results on real datasets. We also evaluate the accuracy of ConDor and PastML in estimating the number of EEMs, as these are key to prefilter the most relevant mutations (step (2) in the ConDor pipeline) and to select the significant ones in Emergence [steps (3) and (4)]. Several authors (Goldstein et al. 2015; Zou and Zhang 2015a; Rey et al. 2019) have pointed out the differences between simulated and real data due to the approximate nature of substitution models. Our simulated datasets do not completely escape this (as we will see), but are still realistic, following the pattern of sedge PEPC dataset (Rey et al. 2018) analyzed below.

The principle is as follows (see Materials and Methods for details). To generate the sequences and perform the analyses, we use the same rooted phylogenetic tree and annotations as for the sedge PEPC dataset, and we base the procedure on the original MSA and the 12 convergence mutations that we have retained for this dataset, following (Besnard et al. 2009; Rey et al. 2018; see below for details). This MSA contains 78 protein sequences and 458 positions; 23 sequences have convergent annotation, while the remaining 55 correspond to the ancestral phenotype. ModelFinder (Kalyaanamoorthy et al. 2017) selects the JTT+R3 model, which we use throughout the simulation. The first step is to remove the 12 convergence mutations in the real MSA by replacing all corresponding residues

with gaps (unknowns). Next, with this modified MSA, the model parameters, tree branch lengths and site rates are re-estimated using IQ-TREE (Nguyen et al. 2015), and the ancestral amino acids at the tree root are reconstructed for each position using PastML (Ishikawa et al. 2019). The simulation evolves the thus reconstructed ancestral sequence along the tree using JTT+R3, but with the parameters and site rates estimated in the absence of convergence. The 12 convergence mutations are then added to the sequences and positions where they occur in the real MSA. The resulting simulated MSAs thus have no convergence events except for the realistically added convergence mutations, which should be similarly difficult to detect as the real ones. ConDor analyses are performed using the same thresholds as for the real data (i.e. minimum number of sequences  $m = 3$ , minimum number of EEMs  $n = 3$ , Emergence corrected  $P$ -value  $< 0.1$ , Correlation logBF  $> 2$ ). This process is repeated 10 times to obtain representative average results.

To study the behavior of ConDor, we carried out three complementary experiments; each was repeated 10 times with the same tools and options as in the original experiment described above.

### Low/High Divergence

The tree used in the original condition above represents moderate divergence among sequences (maximum root-to-tip patristic distance = 0.19). Sequence divergence is expected to play an important role in the results. If the divergence is high, ancestral reconstruction will perform less well and Emergence should be penalized. Conversely, when divergence is low, ancestral amino acids are essentially conserved on the leaves of the tree for nonconvergent sites, while convergence mutations stand out clearly and are virtually the only recurrent mutations observed in the tree, meaning that Emergence should perform well. In contrast, Correlation should be little affected by these changes in scale, unless very long branches make it difficult to compare the Markov models on which this approach is based. To obtain low/high divergences throughout the tree, we multiplied all branches of the tree estimated from the sedge PEPC data after removing convergences by a factor of 1/3 and 3 (i.e. maximum root-to-tip distance = 0.064 and 0.57, respectively). The simulations for these new trees were rerun, along with the insertion of the convergence mutations. ConDor analyses were performed with the same parameters as in the original condition.

### Model Violation

Real data inevitably show deviations from the model used to analyze them. We therefore took the data simulated in the original condition with the JTT+R3 model estimated by IQ-TREE, but introduced a model violation by analyzing

them with ConDor under the usual LG+G4 model, which is often the default option in tree inference. It is a weak violation, but we expect it to affect Emergence, whose simulations will necessarily change, while Correlation, which does not use a substitution model, should not be affected. The other ConDor parameters are unchanged from the original condition, and the data are the same.

### Uncertain Phenotype

In many real datasets, the phenotype is uncertain and approximate. To study the impact of this effect, we introduced noise to the phenotype annotation from Rey et al. (2018). For each dataset simulated in the original condition, we randomly selected 8 converging sequences (out of 23) and 8 nonconverging sequences (out of 55) and swapped the annotations, creating 16 annotation errors compared with the original data, i.e. around 20%. We will see on real data (notably HIV) that this level of uncertainty is quite realistic. The analyses were then carried out as with the original simulations. In this condition, Correlation is expected to be perturbed, while Emergence (which does not use annotations) should retain the same accuracy.

To measure the accuracy of ConDor in counting EEMs, we used these simulated datasets before adding the convergent mutations. For each position in the MSA and each amino acid present at that position, we computed (i) the number of EEMs that occurred during the simulation and (ii) the number of EEMs predicted by ACR using PastML. For each of the above experimental conditions, we computed the correlation between these two values, for all (10) replicates and (458) positions. The result shows a high correlation of  $\sim 0.95$  in all five experimental conditions. This result is very reassuring, especially with respect to our EEMs-based prefiltering. It sounds somewhat counterintuitive, but we have shown that the ACR is surprisingly robust to model violation and accurate for the shallow nodes close to the tips that account for most EEMs (Gascuel and Steel 2014). In addition, counting EEMs is probably easier than ACR, since the actual number of EEMs (mutations) can be obtained even if there are a few errors in the reconstruction of ancestral amino acids on the nodes of the tree.

To assess the properties and performance of ConDor and its Emergence and Correlation components, we used common statistics, namely: the number of true positives (TP: number of detected convergent mutations), true negatives (TN: number of nondetected nonconvergent mutations), false positives (FP: number of detected nonconvergent mutations), and false negatives (FN: number of nondetected convergent mutations). We calculated the Type 1 error rate  $[FP/(FP+TN)]$  to check the risk level of the methods when they predict convergent mutations (i.e. reject the null hypothesis of nonconvergence). To determine the ability

of each method to discriminate between convergent and nonconvergent mutations, we calculated the recall  $[TP/(TP+FN)]$ ; called power in testing], precision  $[TP/(TP+FP)]$  and  $F1$  score, as is standard practice in supervised classification. The  $F1$  score is the harmonic mean of recall and precision. The  $F1$  score provides a balanced view between recall and precision, which are generally in tension (improving precision typically reduces recall and vice versa; in testing, a similar tension exists between Type 1 error rate and power). The  $F1$  score is robust to class imbalance, as is usually the case with convergent mutations that are much less frequent than nonconvergent mutations. The results are displayed in Table 1.

Correlation performs better than Emergence overall, as expected, since Emergence does not use convergence annotation, and the difference is more pronounced in the case of high divergence and model violation. There are two exceptions, however, which are also expected: when the divergence is low, in which case observing mutations is sufficient to detect convergence, and when the annotation is uncertain. ConDor, which combines these two components, does not have a high overall  $F1$  score, but it has the best precision ( $=1.0$ ) and Type 1 error ( $=0.0$ ), since it only retains mutations already retained by Emergence and Correlation.

Type 1 error is generally well controlled and precision is high, except for Emergence in the presence of model violation (as expected) and with low divergence (but the numbers are so small in this setting that nothing can be inferred). We will see that real data pose similar difficulty of a high number of FP for all methods, due to hypothesis and model violations, and other factors that distinguish them from simulated data.

Recall is moderate overall, as might be expected with so few sequences, with certain mutations difficult to detect (they are even more so with real data).

The level of divergence does play a role in the performance of the methods: Correlation is little affected, but Emergence becomes the best method with low divergence and vice versa with high divergence. In the latter condition, we have  $\sim 0.6$  mutations per site (in average; much more for fast sites) between the tree tips and the root. Then, the number of mutations tested is very high, which affects first of all Emergence, due to the multiple tests and the difficulty of the ACR, but also Correlation, whose recall and precision decrease slightly compared to the original condition.

With model violations, as expected, Emergence is significantly penalized, as is ConDor, although the violation is small. Correlation remains unaffected.

With an uncertain phenotype, Correlation is significantly affected, but neither Emergence nor ConDor are impacted. It should be noted that with a more significant annotation perturbation (10 exchanges instead of 8), Correlation no longer finds any convergence mutations (results not shown).

**Table 1**

Assessment of ConDor using simulated data

	TP	FP	FN	TN	Type1	Recall	Precision	F1 score
PEPC-like simulations								
35.5 mutations tested								
Emergence	6.50	1.00	5.50	22.50	0.04	0.54	0.87	0.67
Correlation	8.00	0.00	4.00	23.50	<b>0.00</b>	<b>0.67</b>	<b>1.00</b>	<b>0.80</b>
ConDor	3.90	0.00	8.10	23.50	<b>0.00</b>	0.32	<b>1.00</b>	0.49
Low Divergence								
14.3 mutations tested								
Emergence	11.50	1.10	0.50	1.20	0.48	<b>0.96</b>	0.91	<b>0.93</b>
Correlation	8.40	0.00	3.60	2.30	<b>0.00</b>	0.70	<b>1.00</b>	0.82
ConDor	8.00	0.00	4.00	2.30	<b>0.00</b>	0.67	<b>1.00</b>	0.80
High Divergence								
124.7 mutations tested								
Emergence	1.70	0.80	10.30	121.9	0.01	0.14	0.68	0.22
Correlation	7.10	0.70	4.90	122.0	0.01	<b>0.59</b>	0.91	<b>0.72</b>
ConDor	1.60	0.00	10.40	122.7	<b>0.00</b>	0.13	<b>1.00</b>	0.23
Model Violation								
35.5 mutations tested								
Emergence	6.10	2.80	5.90	20.70	0.12	0.51	0.68	0.58
Correlation	7.90	0.00	4.10	23.50	<b>0.00</b>	<b>0.66</b>	<b>1.00</b>	<b>0.79</b>
ConDor	3.10	0.00	8.90	23.50	<b>0.00</b>	0.26	<b>1.00</b>	0.41
Uncertain Phenotype								
35.5 mutations tested								
Emergence	6.60	0.90	5.40	22.60	0.04	<b>0.55</b>	0.88	<b>0.68</b>
Correlation	6.40	1.00	5.60	22.50	0.04	0.53	0.86	0.66
ConDor	4.00	0.00	8.00	23.50	<b>0.00</b>	0.33	<b>1.00</b>	0.50

TP: true positives. FN: false negatives. FP: false positives. TN: true negatives. Type 1 error rate [FP/(FP+TN)]: proportion of FP among nonconvergent mutations. Recall or power in testing [TP/(TP+FN)]: proportion of TP among convergent mutations. Precision [TP/(TP+FP)]: proportion of TP among all mutations retained by the given method. F1 score: harmonic mean between recall and precision. Emergence: mutations showing a number of EEMs statistically higher than expected with  $P$ -value  $< 0.1$  (after Holm-Bonferroni correction for multiple tests). Correlation: mutations positively correlated with convergence annotation of sequences (Rey et al. 2018), with log Bayes factor  $> 2$ . ConDor: combination of Emergence and Correlation. In bold: best result for each indicator. The five experimental conditions are described in text, in all five there are 12 convergence mutations. Results are averaged over 10 replicates.

These last two conditions are particularly interesting, as real data combine model violation and uncertain phenotype, hence the interest in combining the two methods Emergence and Correlation, each imperfect but complementary, to focus on a reasonable number of candidate mutations. This is particularly evident with the rhodopsin dataset, where the phenotypic annotation is a proxy for a complex annotation that is not available.

## Results

### Overview: Data, Methods and Comparison Criteria

We applied ConDor to three datasets with widely studied convergent mutations: (i) a sedge phosphoenolpyruvate carboxylase (PEPC) protein dataset with mutations associated with the acquisition of C4 metabolism; (ii) an HIV dataset of reverse transcriptase with  $\sim 33\%$  sequences with DRMs; and (iii) a dataset of fish rhodopsin, a light-sensitive receptor protein that is highly conserved but known to vary at certain positions among species depending on their environment.

For HIV and rhodopsin datasets, we reconstructed the phylogeny from the sequences (nucleotide data and protein data, respectively), using ModelFinder (Kalyaanamoorthy et al. 2017) and IQ-TREE (Nguyen et al. 2015) with standard options (see Materials and Methods). For sedge PEPC data, we used the provided phylogeny. Each phylogeny (with branch lengths reoptimized with amino acid sequences for HIV and sedge PEPC) was used as input of ConDor, PCOC (Rey et al. 2018) and FADE (Murrell, Oliveira, et al. 2012). For all tested methods, we evaluated the same mutations and positions, corresponding to the mutations present in at least 0.5% of the sequences and with at least 3 EEMs. The latter value of 3 is above the minimum of two independent emergences (2 EEMs) implicitly required to speak of convergence. It was chosen to benchmark the various methods studied here because we observed that they all suffer from a large number of FP and a loss of statistical power when too many mutations are tested (see, e.g. fish rhodopsin results, Table 4). These thresholds correspond to ConDor's default options, but can be modified by the user. It should be noted, however, that prefiltering is essential for most datasets, given that there are, for example, 767

different amino acids (i.e. 767 possible tests) on all of the positions in our sedge PEPC dataset (HIV: 870, rhodopsin: 1198). Most of these amino acids are rare and have emerged infrequently, and there is no need for computationally expensive tests to check that they are probably nonconvergent.

Given a rooted phylogeny, an alignment of amino acid sequences, and a list of convergent clades, PCOC performs a detection analysis for its three models (Profile Change, PC; OC; and both) for which we can set independent significance thresholds. Instead of detecting a change toward the same amino acid, the PC component aims at detecting positions for which the general use in amino acid preference has changed in the convergent clades. This preference is modeled by a vector of amino acid frequencies or “profile” and, at a convergent position, the profile used in all convergent clades must be different from the ancestral profile used in the rest of the tree. Conversely, for a nonconvergent position, the same profile is used all along the tree. In addition, the OC model forces that the switch of profile occurs along with at least one substitution in the branches rooting the convergent clades. Positions that verify the two submodels are retained as convergent by PCOC, using a specific approach to combine the posterior probabilities from both submodels. For the profiles, we used the C10 model that combines 10 profiles to represent the diversity of biochemical and mutational properties among amino acids (Le et al. 2008a; default option in PCOC). Before running PCOC, users have to annotate the clades for their convergent status, using the list of species having the convergent phenotype. According to Rey et al. (2018), a clade is said to be convergent if all its tips possess the convergent phenotype, and the branches yielding convergence (where OC is expected) are those rooting the maximal convergent clades. PCOC aims to detect positions with molecular convergence, and does not return a list of mutations, but a list of positions. We considered in our experiments that a mutation was detected by PCOC (or one of its components), when (i) it was present at a position with a confidence value larger than 0.8 (the same was applied to PC and OC when used separately) and (ii) the mutation in question was more frequent in species with the convergent phenotype than in the nonconvergent ones. We also conducted experiments to vary the selection threshold of 0.8 and check the impact on accuracy (supplementary tables S1, S2, S5, and S8, Supplementary Material online).

FADE is one of the methods to detect selection available in the HyPhy package (Pond et al. 2005; <https://www.hyphy.org/>). FADE replaces previous approaches to test for episodic directional selection in protein alignments, which showed high detection power with DRMs in HIV (Murrell, Oliveira, et al. 2012). To run FADE, the users first have to specify the branches that are expected to have undergone directional selection, called “foreground”

branches. These typically correspond to all branches in convergent clades (and not solely to the clade rooting branches, as with the OC component of PCOC; see Materials and Methods for details). FADE tests for each position in the alignment if there is a “substitution bias toward a particular amino acid in the foreground branches, compared to the background branches”. The method relies on a Bayesian framework and a Bayes factor >100 (default) provides strong evidence that the site is evolving under directional selection. As with PCOC, we conducted experiments varying this selection threshold (supplementary tables S1, S2, S5, and S8, Supplementary Material online), and used higher thresholds with the larger datasets.

ConDor aims at detecting mutations emerging more often than expected under a null model and which are correlated with the convergent phenotype (or its proxy). In our experiments, the null model corresponded to the best substitution model according to BIC (Bayesian Information Criterion), as inferred by ModelFinder (Kalyaanamoorthy et al. 2017). However, we also tested alternative models to check the robustness of the method to model violation (inevitable with real data). Both components of ConDor use selection thresholds that were set, after examination (supplementary tables S1, S2, S5, and S8, Supplementary Material online), at a corrected *P*-value <10% and a log Bayes factor >20 for Emergence and Correlation, respectively, except for the small sedge PEPC dataset, for which we used a less stringent log Bayes factor threshold of 2. A mutation verifying both conditions was retained as bearing a foreground convergence signature.

We compared the three methods (PCOC, FADE, and ConDor) using the same statistics as with simulated data (i.e. TP, TN, FP, FN, Type 1 error rate, recall, precision and *F1* score). The Type 1 error rate is used to control that we rarely falsely reject the null hypothesis (i.e. no convergence), while the *F1* score provides a balanced view between recall (fraction of convergent mutations that are selected) and precision (fraction of selected mutations that are truly convergent).

### Sedge PEPC Protein Dataset

We selected this dataset on C4 metabolism because it was the one used as a reference to evaluate PCOC in Rey et al. (2018). This dataset comprises 78 sequences and allows comparison of convergence detection methods with a small dataset. C4 metabolism is a recognized case of convergence in plants and arose multiple times independently from the ancestral C3 metabolism. It is thought to be an adaptation to arid and warm environments (Ehleringer et al. 1997). Among the many proteins involved in the C4 photosynthetic pathway, phosphoenolpyruvate carboxylase (PEPC) has been studied to find a molecular basis for phenotypic convergence. PEPC is shared by both C3 and C4 plants and is encoded by a



multigene family. The standard, well-supported hypothesis is that the *pepc* gene responsible for C4 metabolism has derived from an ancestral *pepc* gene responsible for C3.

In our analyses, we focused on sedges, a plant family with multiple independent emergence events of C4 metabolism. We based our analysis on the dataset used in Besnard et al. (2009) and later in Rey et al. (2018). This dataset consists of an alignment of 78 sequences and 458 positions. We annotated the phenotype of the sequences according to two methods. First, based on the global phenotype of the plant (C3 or C4 metabolism) according to Bruhl and Wilson (2007). With this annotation, multiple copies of PEPC in the same plant have the same annotation. However, some sedges are intermediate between C3 and C4, following (Bruhl and Wilson 2007), and the 7 corresponding sequences were removed with the phenotype-based annotation. This resulted in a dataset composed of 71 protein sequences, 22 being annotated as C4. Second, we kept the genotype-based annotation used by Besnard et al. (2009). This annotation is grounded in the fact that, in the PEPC amino acid sequence, the A780S mutation (i.e. a change from A to S on reference position 780) has been experimentally demonstrated to be a major determinant of C4-specific characteristics (Bläsing et al. 2000). They then predicted the metabolism associated with the sedge PEPC sequences according to the presence or absence of the A780S mutation. Compared to the first annotation, this resulted in a change of annotation for 4 sequences, from C4 to C3 metabolism. Moreover, 5 of the proteins from C3-C4 intermediate sedges were annotated as C4 and 2 as C3. This dataset, using the genotype-based annotation by Besnard et al. (2009), thus contains 78 sequences, 23 annotated as C4.

Mutations at positions 780 and 665 were confirmed experimentally to have an impact on the catalytic activity and folding of PEPC (Svensson et al. 2003; Christin et al. 2007). Including these two positions, Besnard et al. (2009) found 16 positions under positive selection that carry parallel amino acid mutations in genes associated with C4 metabolism. Although most of these positions have not been experimentally confirmed (unlike HIV DRMs and convergent mutations in rhodopsin, see below), Rey et al. (2018) used these potentially convergent positions to evaluate the application of PCOC (and other approaches) to this dataset. Our approach was similar. Mutations belonging to positions predicted to be convergent by Besnard et al. (2009) were annotated as convergent. We tested all mutations with at least 3 EEMs, for both phenotype- and genotype-based sequence annotations. One effect of this constraint is that for each tested position predicted to be convergent by Besnard et al. only one (convergent) mutation is tested. Due to the difference in size of the two datasets, we tested a different number of mutations, as some mutations did not pass the EEM threshold. With the “phenotypic” annotation, we tested 59 mutations spread over 51 positions, with 11 mutations (positions)

presumed to be convergent. With the “genotypic” annotation, we tested 66 mutations spread over 56 positions, with 12 mutations (positions) presumed to be convergent. All the mutations tested with the “phenotypic” annotation were also tested with the “genotypic” annotation. The 4 and 5 positions predicted to be convergent by Besnard et al. (2009) and not tested here, since they showed less than 3 EEMs, have not been confirmed experimentally (i.e. positions 733, 770, 810, 906, and 839 for the “phenotypic” annotation). In fact, these positions are difficult to distinguish using convergence criteria and are not predicted as convergent by PCOC and its OC and PC components “because those sites do not fit PCOC’s definition of a convergent site” (Rey et al. 2018). Note that these 4 and 5 positions are not tested here, meaning that they are not included in any of the performance indicators for any of the methods being compared.

The results of the method comparisons for the two analyses are presented in Table 2 and Fig. 3 (see also supplementary tables S1–S4, Supplementary Material online). We first describe the results of the genotypic annotation as it was the one used in previous analyses (Besnard et al. 2009; Rey et al. 2018).

Using PCOC on the genotypic dataset with a posterior probability threshold of 0.8 (as used in Rey et al. 2018; similar results are obtained with a threshold of 0.9, supplementary tables S1 and S2, Supplementary Material online), 11 mutations are detected among which 7 are TP. We thus find a large intersection between Besnard et al. and PCOC results, as previously described in Rey et al. (2018). PCOC results are mostly driven by the OC component, which detects 15 mutations including 8 TP. The PC component, on the other hand, finds only 1 TP and 6 FP. With this dataset, PCOC results are derived primarily from the OC component that “assumes that convergent positions must have undergone a substitution on the branches where the adaptation took place” (Rey et al. 2019). With small datasets like this one, one can reasonably use PCOC recommended method, which infers “the branches where the adaptation took place” as the ones rooting the convergent clades, where all tips have the convergent phenotype. This works very well here (see Fig. 4 in Rey et al. 2018), hence the performance of PCOC. However, it is generally difficult (if not impossible) to define the position of these branches in larger and more complex phylogenies, due to phylogenetic uncertainty, reconstruction errors, and the use of a proxy for the phenotype. In this regard, with the phenotypic annotation PCOC and its subcomponents are no longer able to recover any TP. This shows that by annotating the sequences based on the genotype (presence or absence of the A780S mutation) there is a perfect match between the convergent clades and the mutations, which is advantageous for PCOC. However, on this dataset, PCOC fails with the phenotypic



**Table 2**

Method comparison on sedge PEPC dataset

	TP	FP	FN	TN	Type1	Recall	Precision	F1 score
Genotypic annotation, 66 mutations tested, 12 convergent mutations								
PC	1	6	11	48	0.11	0.08	0.14	0.11
OC	8	7	4	47	0.13	0.67	0.53	0.59
PCOC	7	4	5	50	0.07	0.58	0.64	0.61
FADE	11	4	1	50	0.07	<b>0.92</b>	<b>0.73</b>	<b>0.81</b>
Emergence	7	16	5	38	0.30	0.58	0.30	0.40
Correlation	8	6	4	48	0.11	0.67	0.57	0.62
ConDor	5	3	7	51	<b>0.06</b>	0.42	0.62	0.50
Phenotypic annotation, 59 mutations tested, 11 convergent mutations								
PC	0	4	11	44	0.08	0.0	0.0	0.0
OC	0	1	11	47	0.02	0.0	0.0	0.0
PCOC	0	0	11	48	<b>0.00</b>	0.0	0.0	0.0
FADE	8	9	3	39	0.19	<b>0.73</b>	0.47	<b>0.57</b>
Emergence	6	18	5	30	0.38	0.55	0.25	0.34
Correlation	6	7	5	41	0.15	0.55	0.46	0.50
ConDor	4	3	7	45	0.06	0.36	<b>0.57</b>	<b>0.44</b>

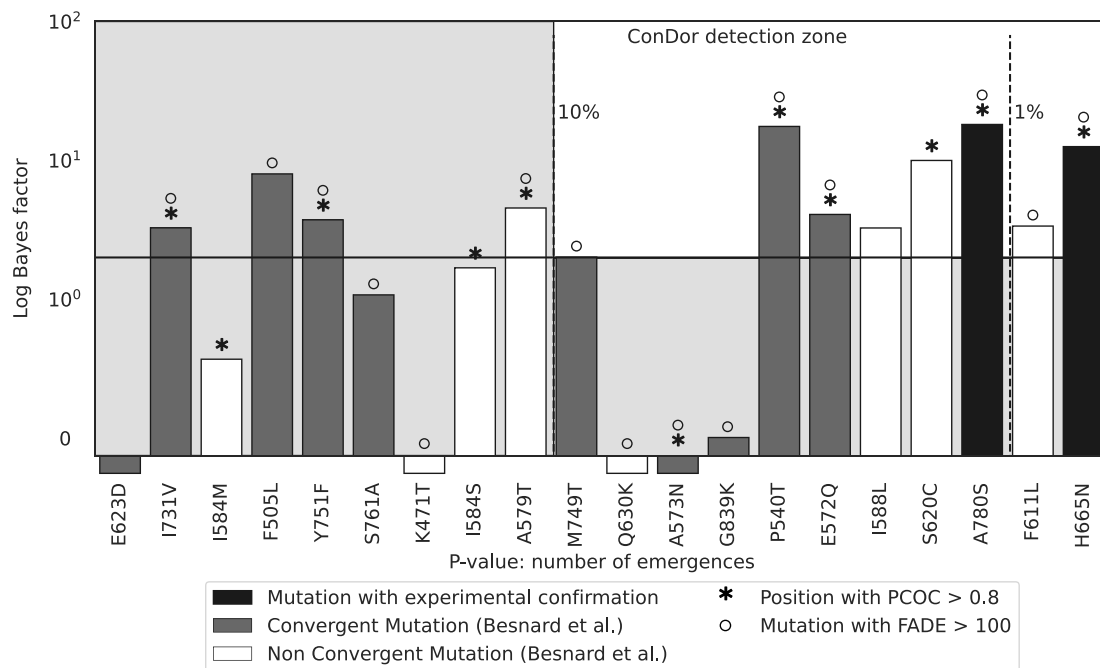
We display for PCOC, FADE, ConDor, and their subcomponents several performance indicators on the detection of convergent mutations either with genotypic annotation according to Besnard et al. (2009) or phenotypic annotation. TP: true positives. FN: false negatives. FP: false positives. TN: true negatives. Type 1 error rate [FP/(FP+TN)]: proportion of FP among nonconvergent mutations. Recall or power in testing [TP/(TP+FN)]: proportion of TP among convergent mutations (see text for details on selection of convergent mutations derived from Besnard et al. 2009). Precision [TP/(TP+FP)]: proportion of TP among all mutations retained by the given method. F1 score: harmonic mean between recall and precision. PC: mutations detected on positions with a PC in convergent clades with a posterior probability >0.8 (as used in Rey et al. 2018). OC: mutations detected on positions where changes occur on the branches leading to the convergent clades, with posterior probability >0.8. PCOC: combination of PC and OC components with posterior probability >0.8. FADE: mutations showing evolution under directional selection in convergent clades, with Bayes factor >100. Emergence: mutations showing a number of EEMs statistically higher than expected with *P*-value <0.1 (after Holm–Bonferroni correction for multiple tests). Correlation: mutations positively correlated with C4 annotation, with log Bayes factor >2. ConDor: combination of Emergence and Correlation. In bold: best result for each indicator. See supplementary tables S1 and S2, Supplementary Material online for results with other selection thresholds for all tested methods; those retained here give the best F1 score.

annotation, which is mutation agnostic and probably more realistic for many convergent evolution studies.

Using the genotypic annotation and a Bayes factor larger than 100 FADE detects 15 mutations, including 11 TP (the accuracy is lower with BF >1,000, supplementary tables S1 and S2, Supplementary Material online). Even though this tool and model were designed in a different context (typically the detection of DRMs in viruses; Murrell, Oliveira, et al. 2012), it performs very well and outperforms PCOC with a F1 score of 0.81 (against 0.61 for PCOC). With the phenotypic annotation, FADE accuracy decreases but still leads to the best results (F1 = 0.57), demonstrating the robustness of this method. This decrease is explained by the fact that there are fewer TP recovered at the same time as more FP. This behavior is expected, as with this annotation, convergent mutations are no longer exclusively found in convergent clades. Like PCOC, FADE requires the user to define the foreground branches where the adaptation occurred, which leads to similar difficulties. However, the hypotheses behind directional selection are less strict than with OC, as one simply assumes a mutational bias toward a certain amino acid in all branches of the convergent clades.

On this dataset, ConDor selected as null model the JTT substitution matrix associated with “freerate” rates-across-sites model (Susko et al. 2003; Soubrier et al. 2012) with 3

categories (R3). The Emergence component of ConDor detects 23 mutations with higher than expected number of EEMs (Holm–Bonferroni adjusted *P*-value <10%), 7 of which are TP. Emergence does not use any phenotype information and likely detects convergent mutations linked to factors other than C4 metabolism, hence the high number of detected mutations that do not belong to Besnard et al. (2009) list. Another factor is likely oversimplifications in the substitution model (e.g. epistasis, see above discussion and references). Overall, this results in a high Type 1 error rate (0.30) and a low precision (0.30). Based on the genotypic annotation, the Correlation component refines these results, as expected since it accounts for the genotype and focuses on foreground convergent mutations: 8 of these 23 mutations carry mutations that are positively correlated with C4 metabolism, among which 5 are TP (M749T, P540T, E572Q, A780S, and H665N). These correspond to the mutations present in the “ConDor detection zone” in Fig. 3. We notice that Correlation alone works fairly well (Table 2), without using any information on amino acid exchangeability and biochemistry, as constitutive of the Emergence component. The combination of the two components in ConDor, using the genotypic annotation, increases the precision of the two components individually without however reaching the one of FADE and PCOC, resulting in mild F1 (0.50). In fact, the F1 score of Correlation alone is higher (0.62) and similar to



**Fig. 3.**—ConDor, PCOC, and FADE convergent-mutation detections on sedge PEPC protein dataset with the genotypic annotation. We display the mutations (with  $\geq 3$ EEMs) that are predicted to be associated with a change in metabolism from C3 to C4 by (Besnard et al. 2009), as well as the mutations (with  $\geq 3$ EEMs) that are predicted by ConDor, PCOC, and FADE, using the same selection thresholds as in Table 2. The two experimentally demonstrated mutations are in black, the other convergent mutations retained by Besnard et al. are in gray, and all other (possibly nonconvergent) mutations detected by PCOC, FADE, and ConDor are in white. Mutations proposed by PCOC and FADE are indicated with an asterisk and a circle on the top, respectively. Mutations proposed by ConDor are present in the “ConDor detection zone”, corresponding to the upper-right white rectangle. Mutations are sorted on the x-axis by the  $P$ -value associated with the number of emergences (EEMs). The dashed lines represent various thresholds of Holm–Bonferroni adjusted  $P$ -values. We report on the y-axis the log Bayes factor as obtained with BayesTraits. The plain horizontal line represents the threshold for “positive evidence” of dependence between mutations and genotypic annotation ( $\log BF > 2$ ). See supplementary tables S3 and S4, Supplementary Material online for additional details on these mutations and ConDor results.

PCOC’s (0.61). Among the 5 TP found by ConDor, 4 are also found by the two other methods and especially the 2 mutations that were demonstrated experimentally (A780S and H665N), while the last one (M749T) is found by both FADE and ConDor, but not by PCOC. Note that this mutation (M749T) emerged only 3 times and is present in 4 sequences, showing the ability of ConDor to detect rare mutations between amino acids with different biochemical properties (supplementary table S3, Supplementary Material online). All methods detect other convergent candidates, most of which being different between methods (Fig. 3). This confirms that the experimental evidence on this dataset is still partial. Other convergent mutations could probably be found, and some of the positions proposed by Besnard et al. might not actually be involved in C4 metabolism. With the phenotypic annotation, the  $F1$  scores of the Emergence and Correlation components alone decrease as for the other methods: Emergence does not consider annotation, but fewer mutations are tested than with genotypic annotation; Correlation still performs well and has the second best  $F1$  score (0.50). ConDor as a whole is less affected by

this annotation change (and the reduction of tests) and has the best precision among all methods (0.57). These results indicate that ConDor is robust to the detection of convergent mutations, even when a convergent mutation is not present in all the convergent clades or when a convergent clade also contains nonconvergent mutations. Further analyses will confirm this finding.

#### HIV Reverse Transcriptase Dataset

DRMs occur independently in patients undergoing drug therapy and are therefore a prime example of molecular convergence. In the case of HIV, they are well characterized and extensively studied, as their occurrence can lead to treatment failure and transmission of resistant viral strains. In particular, mutations must meet certain criteria to be identified as DRMs, including experimental validation (Wensing et al. 2019). DRMs are primarily found in proteins targeted by antiretroviral therapies: protease, reverse transcriptase, and integrase. The list of known DRMs affecting these proteins is publicly available at <https://hivdb.stanford>.

edu/ and is updated regularly. As in the previous sedge PEPC dataset, DRMs are written in the form “XposY”, with X the ancestral (or wild-type) amino acid, “pos” the position of the substitution according to the HXB2 reference sequence, and Y the mutated amino acid, that is, the amino acid conferring resistance. We will use this notation for all our analyses.

In our case, we are interested in mutations on the reverse transcriptase, where DRMs are numerous, diverse, and have been experimentally confirmed. Furthermore, not all mutations occurring at a resistance-associated position make the virus resistant, but only a small subset, and frequently only one. This case is therefore well suited to our method, which aims to detect convergence at the level of mutations and not only at the level of positions. Here, we analyze an HIV-1 group M reverse transcriptase dataset sampled from 10 countries in West and Central Africa, and associated with metadata such as patient treatment status. We use treatment status as a proxy for phenotype, assuming that most patients with a detectable viral load are either treated patients whose treatment has failed due to the development of DRMs, or untreated (naive) patients without DRMs. However, some treated patients may have unsuppressed viral loads for other reasons (e.g. poor adherence to treatment), and some naive patients may have been infected with resistant strains harboring DRMs. This dataset was first studied in Villabona-Arenas et al. (2016) and then in Blassel et al. (2021a), from which we retrieved the data. After removal of recombinant sequences (those for which the recombination occurs within the reverse transcriptase), it contains 1,858 sequences of 747 nucleotide positions that have been translated into 249 amino acid positions. Ten subtypes and circulating recombinant forms (CRFs) are represented in this data, the major one being subtype C (37%). This dataset has several advantages for benchmarking

convergence detection methods, compared to the UK dataset studied in Blassel et al. (2021a). First, a large percentage of the sequences are from treated patients (31%). Second, the DRMs are relatively frequent: ~26% of the sequences harbor at least one DRM that is present in at least 10 sequences. Finally, there is relatively little transmitted resistance (12% of naive sequences have one or more DRMs, Villabona-Arenas et al. 2016). For example, the mutation M184V, which is the most frequent, is observed in 378 and 5 sequences with treated and naive status, respectively, corresponding to relative frequencies of 66% and 0.4%. It is expected that such a DRM will be found by any reasonable convergence detection method. In contrast, rare DRMs are much more difficult to detect. For example, DRM Y188L is found in only 21 sequences (all from treated patients), which corresponds to 3.7% of treated sequences and 1% of all sequences. Note that in these two examples, we are far from observing a perfect correlation between the presence of the DRM and (the proxy used for) the phenotype (i.e. treatment status).

We tested 240 mutations present in at least 10 sequences and showing at least 3 EEMS, corresponding to 95 positions. Among these 240 mutations, 29 are DRMs distributed on 24 positions. We focused on these 29 DRMs to assess and compare the performance of our approach. Results are displayed in Table 3 and Fig. 4 (see also supplementary tables S5–S7, Supplementary Material online).

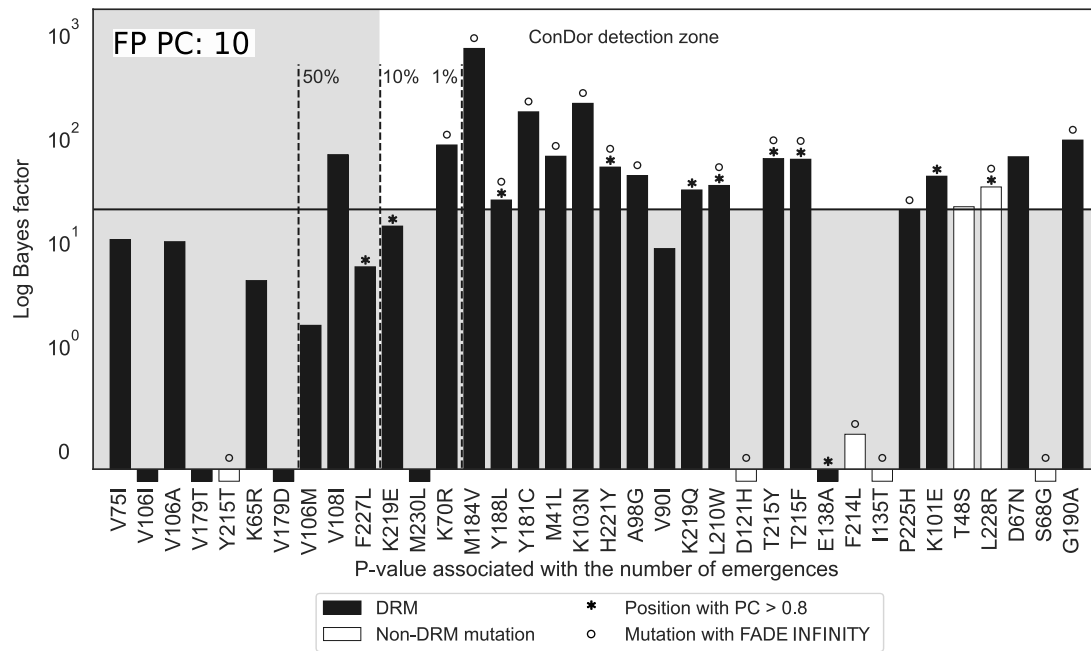
The PC component of PCOC works with a mild accuracy ( $F1 = 0.41$ ; Table 3). The Type 1 error rate is well controlled, but the recall and precision are low. PC finds 12 positions associated with a shift in profile, 8 of which harbor DRMs. These 12 positions correspond to 20 mutations positively correlated with the treatment status, among which

**Table 3**

Method comparison on HIV reverse transcriptase dataset

HIV dataset—29 DRMs, 240 mutations tested	TP	FP	FN	TN	Type 1	Recall	Precision	F1 score
PC	10	10	19	201	0.05	0.35	0.50	0.41
FADE HIVb	13	6	16	205	0.03	0.45	0.68	0.54
FADE JTT	14	8	15	203	0.04	0.48	0.64	0.55
Emergence HIVb	20	67	9	144	0.32	0.69	0.23	0.34
Emergence JTT	21	78	8	133	0.37	<b>0.72</b>	0.21	0.33
Correlation	16	3	13	208	0.01	0.55	0.84	0.67
ConDor HIVb	15	2	14	209	0.01	0.52	0.88	0.65
ConDor JTT	16	1	13	210	<b>0.00</b>	0.55	<b>0.94</b>	<b>0.7</b>

Several performance indicators of the detection of convergent mutations are displayed for PC, FADE, ConDor, and ConDor subcomponents. We display the results using the best substitution matrix (HIVb) and when possible with JTT, which is a form of model misspecification (surprisingly, the results with JTT tend to be better than with HIVb). TP: DRMs found by the given method. FN: DRMs not found by the given method. FP: mutations found by the given method, which are not DRM. TN: non-DRM mutation, not found by the given method. Type 1 error rate [FP/(FP+TN)]: proportion of FP among nonconvergent mutations. Recall or power in testing [TP/(TP+FN)]: proportion of TP among convergent mutations (29 DRMs tested). Precision: proportion of TP among all mutations found by the given method. F1 score: harmonic mean of recall and precision. PC: mutations detected on positions with a PC in convergent clades with a posterior probability >0.8. FADE: mutations showing evolution under directional selection, with a Bayes factor >10<sup>16</sup> (noted “INFINITY” in FADE’s outputs). Emergence: mutations showing a number of EEMs statistically higher than expected at a *P*-value <0.1 after Holm–Bonferroni correction for multiple testing. Correlation: mutations positively correlated with the treatment status, with a log Bayes factor >20. ConDor: combination of Emergence and Correlation using the same selection thresholds. In bold: best result for each indicator. See supplementary table S5, Supplementary Material online for results with other selection thresholds for all tested methods; those retained here give the best F1 score.



**Fig. 4.**—DRMs detection and convergent candidates on HIV data. We display all DRMs (black) and the non-DRM mutations (white) obtained using ConDor and FADE on the HIV-1 group M reverse transcriptase dataset (same selection thresholds as in Table 3). If these mutations were found on positions associated with a shift in profile using PC, the bar is surmounted with an asterisk. If they were found associated with an “INFINITY” ( $>10^{16}$ ) BF using FADE, the bar is surmounted with a circle. Mutations found by ConDor are present in the “ConDor detection zone”, corresponding to the upper-right white rectangle. Mutations are sorted by their *P*-value (Emergence component) on the x-axis. The dashed lines represent various thresholds of adjusted *P*-values using a Holm–Bonferroni correction. We report on the y-axis the log Bayes factor as obtained with BayesTraits. The plain horizontal line represents a strong evidence (log Bayes factor  $>20$ ) of dependence between a mutation and the treatment status. Mutations that display a bar below the x-axis were found to be independent or negatively correlated with treatment status. FP PC in the upper-left indicates the number of FP (=10) found with PC. See [supplementary table S7, Supplementary Material](#) online for additional details on these mutations and ConDor results.

10 are DRMs (TP). However, no position is significant for the OC component (nor PCOC). This result is somewhat expected as DRMs are only found in a subset of all sequences showing the convergent phenotype. Moreover, several DRMs are not associated with a shift in profile and occur between closely related amino acids such as V and I, or K and R (Fig. 4).

FADE with a default Bayes Factor threshold of 100 and with the HIVb substitution matrix has excellent recall but also detects many non-DRM mutations (66, [supplementary table S5, Supplementary Material](#) online), leading to a mild *F1* score (0.38) and poor Type 1 error rate and precision. Focusing on the detections with the highest BF ( $>10^{16}$ , noted “INFINITY” in FADE’s outputs; Table 3), FADE has much better Type 1 error rate and precision, and a significantly higher *F1* score (0.54). Overall, FADE’s performance on this dataset is good, which is consistent since the EDEPS and MEDS models (now replaced by FADE) were designed for drug resistance detection in HIV (Murrell, Oliveira, et al. 2012). In fact, FADE performs much better than FUBAR (Murrell et al. 2013) and MEME (Murrell, Wertheim, et al. 2012), two methods to

detect positive selection from the HyPhy suite (*F1* score: FUBAR = 0, MEME = 0.13; [supplementary table S6, Supplementary Material](#) online). This result was somewhat expected, since these two approaches do not account for the phenotype (see also the result of Crandall et al. (1999) with similar models, and of Lemey et al. (2005) with a branch-site model). Moreover, running FADE with the JTT matrix (instead of HIVb; Table 3), leads to similar Type 1 error rate, recall, precision and *F1* score, showing that FADE is robust to model misspecification.

The null substitution model inferred for this dataset using ModelFinder (Kalyaanamoorthy et al. 2017) is HIVb (Nickle et al. 2007), with “freerates” (Soubrier et al. 2012) rates-across-site model and 4 rate categories. Using the Emergence component of ConDor, we detect 87 mutations with more EEMs than expected, after applying a Holm–Bonferroni correction for multiple testing (adjusted *P*-value  $<10\%$ ). Of these detections, 20 are DRMs, which represents a recall of 0.69 and is higher than expected by chance (Fisher’s exact test *P*-value =  $2e-4$ ). However, 67 mutations are non-DRM events, resulting in poor Type 1 error rate

and precision. In fact, we do not know whether these 67 mutations are simply false positives (FP) or convergent mutations associated with phenotypes different from drug resistance. It is likely that both factors apply, a significant fraction of these FP being explained by oversimplifications in the modeling of substitutions (e.g. epistasis, see discussion above). The Correlation component, at a log Bayes factor  $> 20$ , detects 19 mutations including 16 DRMs (TP). With this dataset, Correlation is therefore sufficient, with an  $F1$  score (0.67) similar to that obtained by ConDor (0.65 with HIVb, 0.70 with JTT; Table 3). As expected, the correlation between DRM and treatment status is strong, and treatment status is a good proxy for the resistance phenotype. We shall see in the following section that this configuration does not occur on the rhodopsin dataset, where both components are needed. ConDor (with its two components) has a high  $F1$  score and the best precision and Type 1 error rate of all the methods tested (Table 3). Moreover, the 2 mutations counted as FP (T48S and L228R) could be true convergent mutations. In particular, L228R (also detected by PC and FADE, Fig. 4) has previously been described as an accessory mutation occurring in response to certain HIV treatments (Rhee 2003; Blassel et al. 2021a). In the case of model misspecification (using JTT, Table 3), the number of FP with Emergence increases slightly. However, the Correlation component smooths out this effect, and ConDor's results are even better than with HIVb ( $F1$  score = 0.7, compared to 0.55 with FADE and 0.41 with PC), showing that ConDor as a whole is robust to model misspecification and achieves high accuracy.

Regarding the ConDor FN (undetected DRMs), we see (Fig. 4) that 5 of them are not detected because they do not pass the threshold limit of the log Bayes factor even though they have a significant  $P$ -value in terms of EEMs (K219E, M230L, V90I, E138A, and P225H). However, 3 of them have a significant log Bayes factor ( $> 10$ , K219E, V90I, and P225H). Moreover, Sluis-Cremer et al. (2014) showed that FN E138A (with negative log-BF) is a polymorphic mutation found naturally in naive patients and particularly in subtype C. This subtype happens to be the main subtype sampled in this dataset and mainly from naive patients (Villabona-Arenas et al. 2016). The false negative E138A is therefore prevalent in our dataset with no significant difference between treated and naive patients, which explains our findings. Lastly, the false negative M230L is present in a small number of sequences ( $n = 14$ ) and the correlation with the phenotype is difficult to establish with such a small number. However, this DRM is significant for the Emergence component. There are 9 additional FN that were not detected by the Emergence component, 8 of which were also not detected by the Correlation component. Most of these mutations have a small number of EEMs and, as expected, both ConDor components here lack detection power.

The Emergence component of ConDor is mostly driven by the number of EEMs and the exchangeability between amino acids. A mutation with a high number of observed EEMs, and corresponding to amino acids with low exchangeability, will rarely emerge in the simulations and will be detected by the Emergence component. Conversely, mutations between highly exchangeable amino acids, such as V and I, will often occur in simulations, which explains why DRM V108I is only detected by the Correlation component. This dataset contains several examples of DRMs involving highly exchangeable amino acids (some of which are TP detected by ConDor: K70R, M41L and, almost, V90I) demonstrating that convergent mutations with effects on phenotype can occur even between amino acids sharing highly similar biochemical properties. In this case, the PC component of PCOC may not be appropriate because it is designed to detect changes in amino acid profiles. Moreover, detailed results show that DRMs occur in both fast and slowly evolving positions (e.g. M184V and V90I with evolutionary rates of 3.27 and 0.17, respectively, meaning that M184V evolves  $\sim 20$  times faster than V90I; supplementary table S7, Supplementary Material online).

In this dataset, there is a strong correlation between most of the DRMs and the phenotype (treated/naive), hence the success of Correlation that has very high  $F1$  score. In fact, with such HIV data, this genotype/phenotype correlation makes it possible to identify DRMs using simple association tests, with additional controls to account for the phylogenetic correlation between the sequences (Villabona-Arenas et al. 2016). Moreover, all DRMs emerged frequently (between 9 and 225 times, supplementary table S7, Supplementary Material online), which reduces the interest of the Emergence component. We shall see that this is not the case for the Rhodopsin dataset where the proxy for the phenotype correlates less well with convergent mutations. In this case, both ConDor components are needed.

### Rhodopsin Data

Rhodopsin is a photosensitive protein pigment responsible for the eye's sensitivity to light. It is found in many vertebrates and has been shown to be under positive (or relaxed purifying) selection among species that evolve in different environments (Spady et al. 2005; Larmuseau et al. 2009). Depending on the habitat and the amount of available light, different amino acids are observed at the same position, resulting in variations in structure of the rhodopsin and different maximum absorption wavelengths ( $\lambda_{max}$ ). Mutagenesis experiments of engineered pigments revealed that the difference of  $\lambda_{max}$  between most rhodopsins could be explained by 9 amino acid mutations (Yokoyama 2008). In particular, D83N, E122Q, F261Y, and A292S (using similar substitution encoding as with HIV and



**Table 4**

Method comparison on fish rhodopsin dataset

Marine and brackish/fresh, 358 mutations tested, 10 convergent mutations	TP	FP	FN	TN	Type 1	Recall	Precision	F1 score
PC	3	24	7	324	0.07	0.3	0.11	0.16
FADE	6	72	4	276	0.21	<b>0.6</b>	0.08	0.14
FADE JTT	4	81	6	254	0.24	0.4	0.05	0.08
Emergence	4	56	6	292	0.16	0.4	0.07	0.11
Emergence JTT	4	72	6	276	0.21	0.4	0.05	0.09
Correlation	6	67	4	281	0.19	<b>0.6</b>	0.09	0.14
ConDor	4	14	6	334	<b>0.04</b>	0.4	<b>0.22</b>	<b>0.29</b>
ConDor JTT	4	14	6	334	<b>0.04</b>	0.4	<b>0.22</b>	<b>0.29</b>

Several performance indicators on the detection of convergent mutations are displayed for PC, FADE, ConDor, and ConDor subcomponents. We display the results using the second-best model (LG) for comparison purpose and, when possible, with JTT, which is a form of model misspecification. TP: mutations affecting maximum absorption wavelength found by the given method. FN: mutations affecting maximum absorption wavelength not found by the given method. FP: mutations found by the given method, which are not experimentally demonstrated to affect absorption wavelength. TN: nondetected mutations that do not affect absorption wavelength. Type 1 error rate [FP/(FP+TN)]: proportion of FP among nonconvergent mutations. Recall or power in testing [TP/(TP+FN)]: proportion of TP among convergent mutations (10 mutations tested which affect maximum absorption wavelength). Precision: proportion of TP among all mutations found by the given method. F1 score: harmonic mean between recall and precision. PC: mutations detected on positions with a PC in convergent clades with a posterior probability >0.8. FADE: mutations showing evolution under directional selection, with a Bayes factor >1,000. Emergence: mutations showing a number of EEMs statistically higher than expected at a *P*-value <0.1 after Holm–Bonferroni correction for multiple testing. Correlation: mutations positively correlated with the proxy of the phenotype, with a log Bayes factor >20. ConDor: combination of Emergence and Correlation using the same thresholds. In bold: best result for each indicator. See [supplementary table S8, Supplementary Material online](#) for results with other selection thresholds for all tested methods; those retained here give the best F1 score.

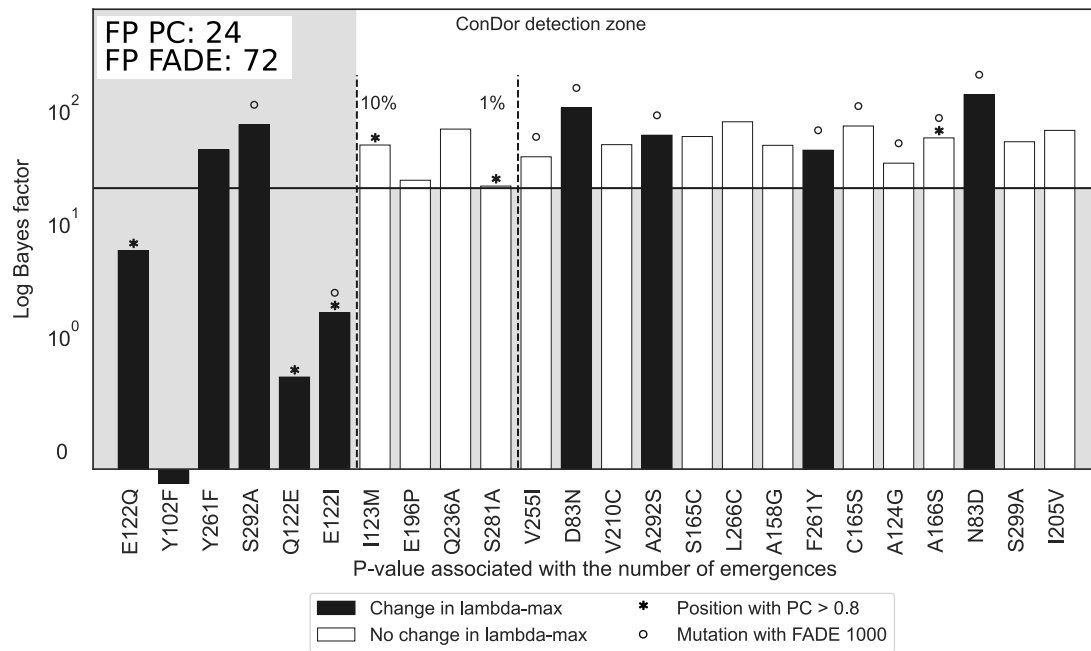
PEPC) occurred several times independently and resulted into functional changes.

The dataset we used comes from a study in which the authors characterized substitution F261Y as convergent in fish rhodopsin, as a possible result of a transition from marine to brackish or fresh water environments (Hill et al. 2019). This dataset contains an alignment of 2,047 sequences with 308 amino acid positions. The sequences have been classified by the authors into two groups: species found only in marine water and species that can live (exclusively or not) in brackish or fresh water. Some of the species associated with the habitat brackish/fresh water can therefore also be found in marine water. The proxy for the  $\lambda_{max}$  is thus given by the environmental condition, depending on whether the fish species are found exclusively in marine water (43%) or not (57%). This approximation of the phenotype is rather imprecise, and we expect the correlation component to work less well on this dataset than for the sedge PEPC and HIV datasets. The reconstructed tree is well supported with 75% of the ultrafast bootstrap supports (Hoang et al. 2018) above 70%. We tested 358 substitutions: the ones present in at least 11 sequences and with at least 3 EEMs. In addition to D83N, E122Q, F261Y, and A292S, the E122I and Y102F mutations have emerged several times in this dataset (Hill et al. 2019), and have been shown experimentally to affect absorption wavelength (Yokoyama 2008). We considered these 6 “direct” mutations as well as their reversions as our TP and explored their emergence for both habitats (marine and brackish/fresh water). Indeed, in the case of successive changes in the environment, we can consider the reversion to the ancestral amino acid as convergent if it has emerged multiple times independently, that is, N83D, Y261F, S292A, and Q122E and I122E, which were counted together as they both revert to the same ancestral amino

acid E. The F102Y mutation shows zero emergence in the dataset and was not tested. All this allowed us to study 10 (6 direct, 4 revertant) mutations as truly convergent. Because we were interested in adaptations to both the marine environment and brackish/fresh water, all programs were launched twice with each of the two conditions as the target. The results are presented in [Table 4](#) and [Fig. 5](#) (see also [supplementary tables S8 and S9, Supplementary Material online](#)), where all the counts are summed over the two annotations/runs of the program.

We applied PC and OC components of PCOC on this dataset for both environmental annotations. A total of 12 positions (corresponding to 27 mutations) were selected using PC, 1 of which has mutations involving a change in the absorption wavelength of rhodopsin (E122I, E122Q, and their reversion Q122E). This resulted in a low F1 score of 0.16 ([Table 4](#)). Moreover, as for the HIV dataset, no position was significant for OC and by extension for PCOC.

FADE with BF >100 ([supplementary table S8, Supplementary Material online](#)) showed a very high number of detections, with a total of 109 mutations for the two environmental annotations. This is hardly surprising given that the environment used as a proxy for the phenotype is very vague. A large proportion of the branches are labeled as foreground, which reduces the specificity of the method. Given this low specificity, FADE had a high recall, but a poor Type 1 error rate and precision, resulting in a F1 score of 0.10, which is the lowest of all methods tested. With a more conservative threshold (BF >1,000), FADE detects 78 mutations, including 6 truly convergent mutations (D83N, E122I, F261Y, A292S, and reversions S292A and N83D), resulting again in poor Type 1 error rate and precision, and a slightly better F1 score of 0.14 (similar to PC’s; [Table 4](#)). The F1 score decreases at higher thresholds. Results with JTT (model misspecification,



**Fig. 5.**—Detection of convergent mutations affecting maximum absorption wavelength on rhodopsin data. We display all mutations affecting maximum absorption wavelength (black) and the other detections (white) obtained using ConDor on the rhodopsin dataset. This figure combines both the detections correlated with brackish/fresh water and those correlated with marine water. If these mutations were found on positions associated with a shift in profile using PC with a posterior probability >0.8, the bar is surmounted with an asterisk. If they were found associated with BF >1,000 using FADE, the bar is surmounted with a circle. Mutations found by ConDor are present in the “ConDor detection zone”, corresponding to the upper-right white rectangle. Mutations are sorted by their *P*-value (Emergence component) on the x-axis. The dashed lines represent various thresholds of adjusted *P*-values using a Holm–Bonferroni correction. We report on the y-axis the log Bayes factor as obtained with BayesTraits. The plain horizontal line represents the threshold for strong evidence of dependence between a mutation and the environmental conditions (log Bayes factor >20). Mutations that display a bar below the x-axis were found to be independent or negatively correlated with treatment status. FP PC and FP FADE in the upper-left indicate the number of FP found with PC and FADE. See [supplementary table S9, Supplementary Material](#) online for additional details on these mutations and ConDor results.

[Table 4](#), [supplementary table S8, Supplementary Material](#) online) are slightly worse, unlike the results with HIV.

The neutral model inferred by ModelFinder on this dataset was “MtZoa” and “freerates” with 8 rate categories. However, we analyzed the data using LG (second best substitution model) to ensure a fair comparison with FADE (MtZoa is not an available option). On this dataset, 60 mutations exhibit a number of EEMs significantly higher than expected as shown in [Table 4](#). Using the Correlation component alone with a log Bayes factor of 20, one detects 73 mutations (40 correlated with brackish/fresh water and 33 with marine water) among which 6 are TP, resulting in a *F1* score of 0.14. Combining both ConDor components, the Type 1 error rate is controlled and we find 18 convergent mutations that are correlated with the environment (9 with brackish/fresh water and 9 with marine water). Although predicting a few mutation candidates, ConDor still detects 4 out of the 10 convergent mutations experimentally confirmed by [Yokoyama \(2008\)](#), which corresponds to the best *F1* score (0.29), precision (0.22), and

Type 1 error rate (0.04) of all methods ([Table 4](#)). In case of model misspecification ([Table 4](#), ConDor JTT), we see that the Emergence component is slightly sensitive (like FADE) with more FP detected. However, ConDor is not affected by the change of model.

As illustrated in [Fig. 5](#), ConDor retrieves mutations F261Y, D83N, A292S, and reversion N83D. Mutation E122Q is not found as convergent (adjusted *P*-value of ~1 and log Bayes factor of 6 associated with the marine environment), because glutamine (Q) independently emerged only 3 times according to ACR, but emerged up to 11 times in simulations. Similarly, mutations E122I (3 EEMs), Y102F (3 EEMs) and reversions Q122E (3 EEMs), Y261F (3 EEMs), and S292A (11 EEMs) are not detected by ConDor as their number of EEMs was not significantly higher than expected ([supplementary table S9, Supplementary Material](#) online). However, the case of reversions is interesting, as these mutations inevitably emerged less frequently than “direct” mutations. With such mutations, users can modify the selection threshold of Emergence and give

more importance to Correlation. For example, S292A (11 EEMs) and Y261F (3EEMs) are selected by Correlation with high log Bayes factors of 72.8 and 43.6, respectively, and could be retained while considering their “Reversion” annotation that is available in ConDor outputs. Nevertheless, counting the number of emergences is mandatory with this dataset. All 358 mutations tested show at least 3 emergences. When relaxing this constraint, the number of FP is even higher, for all methods. For example, Correlation finds mutations with high log Bayes factor, which emerged only once, illustrating the presence of founder events that can mislead this method when used in isolation (e.g. “direct” mutations A16T, F52S, I54S, and N55T show 1 EEM and a significant logBF of 11.8, 14.0, 15.8, and 13.8, respectively; see detailed result on GitHub). All this indicates that both components of ConDor are needed to focus on a reasonable number of convergent candidates (PC and FADE exhibit, respectively, 24 and 72 FP, Table 4), even if the constraints imposed by Emergence can be relaxed in some cases (e.g. reversions).

Interestingly, convergent candidate mutation A166S is detected by the three methods (Fig. 5). This mutation is found by ACR to have 48 EEMs, whereas on simulations it emerged at most 38 times. Following previous results from Malinsky et al. (2015) and O’Reilly et al. (2016), it might be associated with a blue-shifting absorption wavelength.

## Discussion

In this work, we developed the ConDor approach, which detects evolutionary convergence at the amino acid mutation level using two components: Emergence and Correlation. Convergent (versus original) phenotypic annotations are given by users for extant taxa, without the need to define convergent clades and infer the phenotypic annotations of ancestral nodes. As we developed this method for the study of viruses and microorganisms for which the phenotype is difficult to access, ConDor allows the use of environmental conditions (and other factors inducing selection pressure) as a proxy for the phenotype. Thus, convergent mutations can be found even if they are present in only a subset of the convergent taxa and if they are found in some taxa that do not possess the convergent phenotype. This is particularly suitable to the analysis of large datasets with several thousand sequences, where inference of convergent clades and ancestral phenotypes are especially challenging. For example, we were able to find more than half of the DRMs on a large HIV dataset where the application of PCOC was not appropriate because the underlying assumptions (OC and PC) were poorly satisfied. We also detected more DRMs with ConDor than using FADE, while the assumptions of this software were made for DRM detection in HIV. Although it was primarily developed for the analysis

of large datasets of viruses and microorganisms, ConDor was able to detect several convergent mutations involved in the change in metabolism in a small dataset of sedge PEPC protein, and in the change in absorption wavelength in a large dataset of fish rhodopsin. For the latter, its accuracy was markedly better than that of PCOC and FADE. These results confirm that ConDor detects a realistic signal of convergent molecular evolution and that it can be applied to a wide range of organisms and datasets.

We tested the robustness of the Emergence component of ConDor to model violation by using the JTT substitution matrix (Jones et al. 1992) instead of HIVb (Nickle et al. 2007) as the neutral evolutionary model for the HIV dataset study. A similar experiment was performed with fish rhodopsin where JTT was used again, instead of LG. In doing so, the sensitivity of Emergence remained high, we still detected the most frequent convergent mutations, but the number of FP slightly increased. The same was observed with simulated data. However, with the addition of the Correlation component, ConDor proved to be robust to these model violations. Since we never know the true evolutionary model, we expect that a substantial number of FP may be observed with Emergence, even when using the best substitution matrix (as selected by IQ-TREE using BIC in our experiments). This behavior was observed in Goldstein et al. (2015) and Zou and Zhang (2015a), where the authors showed the difficulty to account for background convergent mutations using standard substitution models. More advanced substitution models based on CAT or CAT-JTT profile models (Lartillot and Philippe 2004; Le et al. 2008a), on mixture of matrix models accounting for structural features and evolutionary rates of the positions (Le et al. 2008b, 2012), or on Markov modulated Markov models as in Escalera-Zamudio et al. (2020), should likely improve the Emergence component, make simulations more realistic, and decrease the number of background convergent-mutation detections in this component. However, these approaches are resource-intensive, and the Correlation component already complements the Emergence component well.

The two components of ConDor can be used independently of each other. When using the Emergence component alone, there is no need to specify the taxa with the convergent phenotype, but one cannot distinguish between foreground and background mutations and the method tends to return a large number candidate mutations. On the other hand, using the correlation component alone, one loses the constraint of multiple emergence, and our results with fish rhodopsin show that founder events with a single emergence can be significantly correlated with phenotype. Furthermore, in this setting, we do not take into account the biochemical properties of the amino acids, the evolutionary rates of the positions, etc. The emergence component allows effective sorting of correlation detections by ranking mutations not only by their number of emergences, but also by their properties, allowing users

to focus on a few notable mutations. The detailed output of ConDor, which includes (among other things) the genetic barrier and exchangeability between the two amino acids, the rate of evolution of the position and the type of the convergent mutation, allows for the refinement of selection priorities, for example with reversions where a lower number of emergences is expected.

ConDor was developed to detect convergent amino acid mutations, not convergent positions, which makes it difficult to compare with existing approaches based on convergent position detection [e.g. PCOC (Rey et al. 2018)]. An adaptation of ConDor to work at the position level could be an interesting feature to add to the program. Our approach is made possible because we work at the scale of a single protein with thousands of sequences, which provides sufficient signal and detection power. Working on thousands or even millions of positions (e.g. with bacterial genomes), ConDor would probably not have the statistical power to work at the scale of a single mutation due to multiple testing. An extension of ConDor to work at the gene level (similarly to Chabrol et al. 2018; Fukushima and Pollock 2023; Duchemin et al. 2023), or to detect convergence in a sliding window, would certainly be a useful development, allowing for the discovery of adaptive mutational patterns involving multiple sites in the protein alignment, rather than isolated sites as with the current version of ConDor.

Other improvements could concern the Correlation component that currently uses discrete trait evolution models in a Bayesian framework, which requires a lot of computing resources (~30 min per mutation on the rhodopsin dataset). This computational burden could be greatly reduced using a similar maximum-likelihood approach [e.g. based on the “ace” routine from the Analyses of Phylogenetics and Evolution (APE) package (Paradis et al. 2004)]. In the same sense, simulations of the Emergence component are computationally expensive, and analytical approaches, inspired by those used in Chabrol et al. (2018), would also significantly reduce the computational burden of the approach.

## Materials and Methods

### Simulated Datasets

We used the sedge PEPC data (see next paragraph) consisting of a protein alignment of 78 sequences (458 positions), 1 outgroup, and a phylogenetic tree. ModelFinder (Kalyaanamoorthy et al. 2017; IQ-TREE v1.6.8; option `-m MFP`) was used for model selection (JTT+R3 model). We re-rooted the tree using the sequence *Chrysithr* as outgroup with Gtree (v0.4.4, options `reroot outgroup`). Convergent sites in convergent taxa (genotypic annotation) were replaced by gaps using Goalign (v.dev103ea5b, options `replace -posfile`). Outgroup was removed from the

alignment using goalign (v. dev103ea5b, option `subset`). Branch lengths were optimized and site-specific evolutionary rates were estimated by IQ-TREE (v1.6.8, option `-m JTT + R3 -te -wsr`). Ancestral root sequence was inferred using PastML with JTT matrix (v1.9.33, option `-prediction_method MAP` and parameter files giving the site rates and the JTT matrix). Before simulating the sequences, the branch lengths of the tree were scaled at different factors (0.33, 1.00, or 3.00), using gtree (v0.4.4, options `gotree brlen scale -f <scale>`). Simulated alignments were then generated as in the ConDor pipeline described below, with JTT matrix and estimated site rates. Finally, convergent mutations were reintroduced in the alignment using goalign (v.dev103ea5b, options `replace -posfile`).

Simulated datasets were analyzed using the Condor pipeline, using different options depending on the case (command `nextflow run condor.nf --outgroup outgroup.txt -model 'JTT+R3' (or 'LG+G4') -nb_simu 10000 -min_seq 3 -min_eem 3 -freqmode 'Fmodel' -branches condor -correction holm -alpha 0.1 -bays 2`).

### Sedge PEPC Protein Dataset

Protein data of sedge PEPC, associated phylogeny, and “genotypic” C3/C4 annotation were downloaded from <https://github.com/CarineRey/pcoc/tree/master/data/det>. The protein data consist of a multiple sequence alignment of 79 protein sequences and 458 positions. The sequences are highly conserved, except for a few long deletions, and well aligned with no problematic gappy regions. We used the sequence *Chrysithr* as outgroup to root the tree and then removed it from the analysis (as Rey et al. 2018), resulting in an alignment of 78 protein sequences. Following Besnard et al. (2009), 23 sequences have a convergent “genotypic” annotation (i.e. C4), based on the presence of the A780S mutation. For the “phenotypic” annotation, we annotated each gene using the annotation of the plant species in which it was sequenced from Bruhl and Wilson (2007), and we removed the 7 genes from *Eleocharis baldwinii* and *Eleocharis vivipara* that perform both C3 and C4 metabolisms. This resulted in a multiple sequence alignment of 71 proteins and 458 positions, with 22 sequences annotated as convergent. The 7 sequences from *E. baldwinii* and *E. vivipara* were pruned from the provided phylogeny.

### HIV Dataset

The HIV reverse transcriptase dataset we analyzed is based on the nucleotide alignment of Villabona-Arenas et al. (2016), which was also studied in Blassel et al. (2021a), and is available from [https://github.com/lucblassel/HIV-DRM-machine-learning/tree/main/data/African\\_dataset](https://github.com/lucblassel/HIV-DRM-machine-learning/tree/main/data/African_dataset). This is a high-quality alignment, thanks to the fact that the HIV proteins are highly conserved, with very few indels (see



Villabona-Arenas et al. 2016 for details). This alignment consists of 3,990 HIV-1 group M partial reverse transcriptase sequences, divided into treatment-naive and treated sequences, along with a metadata file indicating the treatment status, whether the sequence has one or more DRMs and the subtype or CRF. The subtype annotation indicates that 2,247 sequences are recombinant forms, which we removed if the recombinant breakpoints were found within the reverse transcriptase (if so, we cannot reconstruct a sound phylogeny,  $n = 2,008$ ). For example, 1,477 sequences were CRF02\_AG, which recombines within the reverse transcriptase gene (Kusagawa et al. 2001). We then ran jpHMM (version of March 2015) (Schultz et al. 2012) to identify other possible recombinant forms. We used the default settings for HIV -v HIV and the priors provided in the jpHMM folder: `-a priors/emissionPriors_HIV.txt -b priors/transition_priors.txt`. Based on jpHMM analysis, we removed 124 additional recombinant sequences with breakpoints in the reverse transcriptase gene.

To root the tree, we added to this nucleotide MSA 3 reference sequences from the N group, which we downloaded from <https://www.hiv.lanl.gov/content/sequence/NEWALIGN/align.html> (reverse transcriptase: user-defined range 2550 to 3297).

The tree was inferred from the nucleotide MSA with IQ-TREE 1.6.8 while selecting the model (GTR+R10 in this case) with Model Finder (IQ-TREE option `-m MFP`). After rooting the tree, the 3 reference sequences of group N were removed from the analysis. The resulting alignment contains 1,858 group M sequences of 747 nucleotides that we translated into 249 amino acids. DRMs were identified in the translated MSA using the 2,019 HIV-1 DRM list (Wensing et al. 2019). Five hundred and seventy-one sequences were annotated as treated and 1,287 sequences as naive. The tree branch lengths were reoptimized by ConDor using the protein MSA (see ConDor Pipeline Description below).

### Rhodopsin Dataset

Rhodopsin protein data and fish habitat were downloaded from [https://github.com/Clupeaharengus/rhodopsin/tree/master/phylogeny\\_habitat](https://github.com/Clupeaharengus/rhodopsin/tree/master/phylogeny_habitat). We extracted from the “`final_alignment.translated.fullrhodopsin.fasta`” alignment file 2,056 sequences corresponding to the identifiers indicated in the “`spp_to_keep.txt`” file. After a quick tree reconstruction with FastTree (Price et al. 2009), we removed 7 badly aligned sequences (based on their aberrant branch lengths). We checked the quality of the resulting alignment with Transitive Consistency Score (TCS; Chang et al. 2014) and obtained a score of 997/1,000 demonstrating high reliability. Rhodopsin phylogeny was reconstructed from this protein MSA, using Model Finder (IQ-TREE 1.6.8 option `-m MFP`) to select the evolutionary model (MtZOA+R8) and IQ-TREE with option `-bb 1,000` for ultrafast bootstrap

approximation (Hoang et al. 2018). We rooted the tree using the same sequences as in Hill et al. (2019) (*Huso huso* and *Polyodon spathula*) and removed them from the phylogeny for the analysis. This resulted in an alignment of 2,047 sequences, 883 annotated with marine water and 1,164 with brackish/fresh water. Habitat was provided in the “`ra-bo_allele_hab.tsv`” file from the repository provided in Hill et al. (2019).

### PCOC

We used PCOC v1.0.1 (Rey et al. 2018) to detect convergent positions based on knowledge of genotype/phenotype (C3 versus C4), treatment status (treated versus naive), and habitat (marine versus brackish/fresh water). We used the C10 profile (`-CATX_est 10`) with 4 gamma categories (`-gamma`) and set the posterior probability threshold  $>0.8$  for all models (PC, OC, PCOC) and datasets (`-f 0.8`). As described in the user guide (<https://github.com/CarineRey/pcoc>), the convergent scenario (`-m`) corresponds to the list of the maximal clades that exhibit the convergent phenotype. Each clade corresponds to an independent emergence event. Since it cannot be known exactly where the convergent transition occurred in the tree, the clades are first reconstructed by retrieving the tips with the convergent phenotype (C4 metabolism, treated, brackish/fresh water, marine water). Then, the internal nodes are recursively annotated with the convergent phenotype if the two child nodes also have the convergent phenotype.

### FADE

We used FADE 0.2 (unpublished to date; <https://hyphy.org/tutorials/CL-prompt-tutorial/>) from the HyPhy package (Pond et al. 2005) to detect mutations under directional selection. FADE requires as input a rooted tree with annotations for the set of foreground branches suspected to have undergone directional selection. We annotated the foreground branches using <http://phyloree.hyphy.org/>. This software allows us to select terminal branches leading to tips with a convergent phenotype as foreground branches. Then, we can label internal nodes as foreground based on several methods (maximum parsimony, conjunction and disjunction). We labeled the internal nodes using conjunction, which is based on logical “AND” (a node is labeled foreground if all its children are foreground). This follows, in fact, the same labeling process as for PCOC. FADE was then run using the same substitution matrices as ConDor (JTT, HIVb, LG) and providing the same amino acid alignments as for PCOC and ConDor.

### ConDor Pipeline Description

The ConDor Pipeline consists of several processes shown in Fig. 1. Here, we describe the implementation details of IQ-TREE 1.6.8 (Nguyen et al. 2015) for the reoptimization



step, PastML 1.9.33 (Ishikawa et al. 2019) for the ancestral reconstruction step, and BayesTraits 3.0.1 (Pagel 1994; Pagel and Meade 2006) for the correlation step. We also describe the implementation of ConDor.

### Model Selection, Branch-length and Evolutionary Rate Estimation

Given an input protein MSA and the corresponding phylogeny, we estimate the best evolutionary model and reoptimize the branch lengths and evolutionary rates for the protein MSA, using the `-m MFP` option of IQ-TREE, while fixing the topology using the `-te` option. This phylogeny with optimized branch lengths is used by ConDor, but also for PCOC and FADE analyses. In ConDor, site-specific evolutionary rates are retrieved from IQ-TREE with the `-wsr` option. For all analyses, we used the equilibrium frequencies corresponding to the substitution matrix, except for the model misspecification experiment with JTT on the HIV dataset, where we reoptimized the equilibrium frequencies (option +FO), in order to obtain a reasonable fit with the data, using a standard procedure. The best substitution model for each dataset was JTT+R3 (Sedge), HIVb+R4 (HIV), and MtZOA+R8 (Rhodopsin). For the rhodopsin analysis with ConDor, we used LG+R8 to allow a fair comparison with FADE.

### ACR by Maximum Likelihood

ACR in the ConDor pipeline is performed using PastML with option `-prediction_method MAP` (i.e. maximum a posteriori). PastML takes as input a parameter file (option `-parameters`) per position, containing (i) the amino acid frequencies for the entire alignment and (ii) a scaling factor for the position under study, corresponding to the evolutionary rate of the site, as estimated by IQ-TREE. This scaling factor (evolutionary rate) is used by PastML to rescale the branch lengths while computing the tree likelihood. The selected substitution matrix (JTT, HIVb, LG) was given as input (`-rate_matrix`) using PastML option `-m CUSTOM_RATE`.

### Correlation Measurement Using BayesTraits

Correlations between the convergent phenotype and mutations occurring more often than expected were measured using the BayesTraits “discrete dependent” model. The convergent phenotype was annotated as 1 and the nonconvergent phenotype as 0. Similarly, for a given position, the mutated amino acid of interest had value 1, and the other amino acids at that position had value 0. To assess whether dependence between the two traits was more likely than their independence, we followed [www.evolution.reading.ac.uk/Files/BayesTraits-V1.0-Manual.pdf](http://www.evolution.reading.ac.uk/Files/BayesTraits-V1.0-Manual.pdf). The dependence hypothesis was retained if the logarithm of the Bayes factor was greater than 2 for the sedge PEPC dataset and greater

than 20 for the others. Thresholds of 2, 5, and 10 are given, respectively, as positive, strong, and very strong evidence against  $H_0$  in Kass and Raftery (1995). Priors for transition rates were defined as uniform with a range between 0 and 100, as described in the user guide. Mutations that were found to be dependent of the phenotype by BayesTraits, were retained as convergent if the correlation was positive. To do this, we checked that the mutation frequency was greater in sequences with the convergent phenotype than in sequences with the nonconvergent phenotype. More formally, let us denote:  $m_C$  the number of sequences that have the mutation M and are annotated with the convergent phenotype;  $m_{NC}$  the number of sequences that also have the mutation M but are annotated with the nonconvergent phenotype;  $C$  the total number of sequences annotated with the convergent phenotype; and  $NC$  the total number of sequences annotated with the nonconvergent phenotype. If  $m_C/C > m_{NC}/NC$ , then the correlation is positive, and M is considered as a convergent mutation by ConDor.

### Implementation

ConDor method is implemented in a Nextflow DSL1 v20.10.0 pipeline (Tommaso et al. 2017), taking as input an amino acid alignment, a rooted tree, a file containing outgroup sequence identifiers and a file containing the list of sequences with the convergent phenotype. The python libraries `numpy` (Harris et al. 2020), `pandas` (McKinney 2010), and `scipy` (Virtanen et al. 2020) were used for data frames and matrices manipulations and for the statistics tools they provide. We used `biopython` (Cock et al. 2009) for sequence and alignment manipulations. Tree traversals and analyses were achieved with `ETE 3` (Huerta-Cepas et al. 2016). Graphs were obtained using the `matplotlib` (Hunter 2007) and `seaborn` libraries (Waskom 2021). MSA (translation to amino acids, subalignments, etc.) and tree manipulations (pruning, rooting, etc.) were performed using `goalign` and `gotree` (Lemoine and Gascuel 2021). Simulations and counting of EEMs were computed using homemade python scripts. Mutations emerging more frequently than expected were selected based on their  $P$ -value (with pseudocount) after Holm–Bonferroni multiple testing correction, with an alpha risk of 0.1 This pipeline is available on Github at <https://github.com/evolbioinfo/condor>, via a webserver at [condor.pasteur.cloud](http://condor.pasteur.cloud) and as standalone docker container (evolbioinfo/condor). A user guide provides full details on the input and output formats, including explanations on the statistics provided for each mutation tested ( $P$ -value, logBF, genetic barrier, etc.).

### Supplementary Material

Supplementary material is available at *Genome Biology and Evolution* online.

## Acknowledgments

We sincerely thank Luc Blassel, Jakub Voznica, and Bastien Boussau for their help and suggestions. We would also like to thank the GBE editors and anonymous reviewers for their helpful suggestions that improved our manuscript.

## Funding

This work was supported by INCEPTION program (Convention ANR-16-CONV-0005; M.M. PhD grant) and by PRAIRIE program (Convention ANR-19-P3IA-0001; O.G.). M.M. was a student from the FIRE PhD program funded by the Bettencourt Schueller foundation and the EURIP graduate program (ANR-17-EURE-0012).

## Data Availability

Our MSAs, phylogenetic trees, scripts and results analysis are accessible from the Github repository <https://github.com/evolbioinfo/condor-analysis>.

## Literature Cited

- Arendt J, Reznick D. Convergence and parallelism reconsidered: what have we learned about the genetics of adaptation? *Trends Ecol. Evol.* 2008;23(1):26–32. <https://doi.org/10.1016/j.tree.2007.09.011>.
- Barker D, Pagel M. Predicting functional gene links from phylogenetic—statistical analyses of whole genomes. *PLoS Comput Biol.* 2005;1(1):e3. <https://doi.org/10.1371/journal.pcbi.0010003>.
- Besnard G, Muasya AM, Russier F, Roalson EH, Salamin N, Christin P-A. Phylogenomics of C4 photosynthesis in sedges (Cyperaceae): multiple appearances and genetic convergence. *Mol Biol Evol.* 2009;26(8):1909–1919. <https://doi.org/10.1093/molbev/msp103>.
- Bhattacharya T, Daniels M, Heckerman D, Foley B, Frahm N, Kadie C, Carlson J, Yusim K, McMahon B, Gaschen B, et al. Founder effects in the assessment of HIV polymorphisms and HLA allele associations. *Science.* 2007;315(5818):1583–1586. <https://doi.org/10.1126/science.1131528>.
- Bläsing OE, Westhoff P, Svensson P. Evolution of C4 phosphoenolpyruvate carboxylase in *Flaveria*, a conserved serine residue in the carboxyl-terminal part of the enzyme is a major determinant for C4-specific characteristics. *J Biol Chem.* 2000;275(36):27917–27923. <https://doi.org/10.1074/jbc.M909832199>.
- Blassel L, Tostevin A, Villabona-Arenas CJ, Peeters M, Hué S, Gascuel O, Database O behalf of the UHDR. Using machine learning and big data to explore the drug resistance landscape in HIV. *PLoS Comput Biol.* 2021a;17(8):e1008873. <https://doi.org/10.1371/journal.pcbi.1008873>.
- Blassel L, Zhukova A, Villabona-Arenas CJ, Atkins KE, Hué S, Gascuel O. Drug resistance mutations in HIV: new bioinformatics approaches and challenges. *Curr Opin Virol.* 2021b;51:56–64. <https://doi.org/10.1016/j.coviro.2021.09.009>.
- Bloom JD, Neher RA. Fitness effects of mutations to SARS-CoV-2 proteins. *bioRxiv* 526314. <https://doi.org/10.1101/2023.01.30.526314>, 2023, preprint: not peer reviewed.
- Bruhl J, Wilson K. Towards a comprehensive survey of C3 and C4 photosynthetic pathways in Cyperaceae. *Aliso.* 2007;23(1):99–148. <https://doi.org/10.5642/aliso.20072301.11>.
- Castoe TA, de Koning APJ, Kim H-M, Gu W, Noonan BP, Naylor G, Jiang ZJ, Parkinson CL, Pollock DD. Evidence for an ancient adaptive episode of convergent molecular evolution. *Proc Natl Acad Sci.* 2009;106(22):8986–8991. <https://doi.org/10.1073/pnas.0900233106>.
- Chabrol O, Royer-Carenzi M, Pontarotti P, Didier G. Detecting the molecular basis of phenotypic convergence. *Methods Ecol Evol.* 2018;9(11):2170–2180. <https://doi.org/10.1111/2041-210X.13071>.
- Chai S, Tian R, Rong X, Li G, Chen B, Ren W, Xu S, Yang G. Evidence of echolocation in the common shrew from molecular convergence with other echolocating mammals. *Zool Stud.* 2020;59:e4. <https://doi.org/10.6620/ZS.2020.59-4>.
- Chang J-M, Di Tommaso P, Notredame C. TCS: a new multiple sequence alignment reliability measure to estimate alignment accuracy and improve phylogenetic tree reconstruction. *Mol Biol Evol.* 2014;31(6):1625–1637. <https://doi.org/10.1093/molbev/msu117>.
- Christin P-A, Salamin N, Savolainen V, Duvall MR, Besnard G. C4 photosynthesis evolved in grasses via parallel adaptive genetic changes. *Curr Biol.* 2007;17(14):1241–1247. <https://doi.org/10.1016/j.cub.2007.06.036>.
- Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics.* 2009;25(11):1422–1423. <https://doi.org/10.1093/bioinformatics/btp163>.
- Crandall KA, Kelsey CR, Imamichi H, Lane HC, Salzman NP. Parallel evolution of drug resistance in HIV: failure of nonsynonymous/synonymous substitution rate ratio to detect selection. *Mol Biol Evol.* 1999;16(3):372–382. <https://doi.org/10.1093/oxfordjournals.molbev.a026118>.
- Cuevas JM, Elena SF, Moya A. Molecular basis of adaptive convergence in experimental populations of RNA viruses. *Genetics.* 2002;162(2):533–542. <https://doi.org/10.1093/genetics/162.2.533>.
- Davies KTJ, Cotton JA, Kirwan JD, Teeling EC, Rossiter SJ. Parallel signatures of sequence evolution among hearing genes in echolocating mammals: an emerging model of genetic convergence. *Heredity (Edinb).* 2012;108(5):480–489. <https://doi.org/10.1038/hdy.2011.119>.
- Duchemin L, Lanore V, Veber P, Boussau B. Evaluation of methods to detect shifts in directional selection at the genome scale. *Mol Biol Evol.* 2023;40(2):msac247. <https://doi.org/10.1093/molbev/msac247>.
- Ehleringer JR, Cerling TE, Helliker BR. C4 photosynthesis, atmospheric CO<sub>2</sub>, and climate. *Oecologia.* 1997;112(3):285–299. <https://doi.org/10.1007/s004420050311>.
- Escalera-Zamudio M, Golden M, Gutiérrez B, Théze J, Keown JR, Carrique L, Bowden TA, Pybus OG. Parallel evolution in the emergence of highly pathogenic avian influenza A viruses. *Nat Commun.* 2020;11(1):5511. <https://doi.org/10.1038/s41467-020-19364-x>.
- Fitch WM. Toward defining the course of evolution: minimum change for a specific tree topology. *Syst Biol.* 1971;20(4):406–416. <https://doi.org/10.1093/sysbio/20.4.406>.
- Foll M, Gaggiotti OE, Daub JT, Vatsiou A, Excoffier L. Widespread signals of convergent adaptation to high altitude in Asia and America. *Am J Hum Genet.* 2014;95(4):394–407. <https://doi.org/10.1016/j.ajhg.2014.09.002>.
- Foot AD, Liu Y, Thomas GWC, Vinař T, Alföldi J, Deng J, Dugan S, van Elk CE, Hunter ME, Joshi V, et al. Convergent evolution of the genomes of marine mammals. *Nat Genet.* 2015;47(3):272–275. <https://doi.org/10.1038/ng.3198>.
- Fukushima K, Pollock DD. Detecting macroevolutionary genotype–phenotype associations using error-corrected rates of protein convergence. *Nat Ecol Evol.* 2023;7:155–170. <https://doi.org/10.1038/s41559-022-01932-7>.

- Gascuel O, Steel M. Predicting the ancestral character changes in a tree is typically easier than predicting the root state. *Syst Biol*. 2014;63(3):421–435. <https://doi.org/10.1093/sysbio/syu010>.
- Goldman N, Yang Z. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol Biol Evol*. 1994;11(5):725–736. <https://doi.org/10.1093/oxfordjournals.molbev.a040153>.
- Goldstein RA, Pollard ST, Shah SD, Pollock DD. Nonadaptive amino acid convergence rates decrease over time. *Mol Biol Evol*. 2015;32(6):1373–1381. <https://doi.org/10.1093/molbev/msv041>.
- Gutierrez B, Escalera-Zamudio M, Pybus OG. Parallel molecular evolution and adaptation in viruses. *Curr Opin Virol*. 2019;34:90–96. <https://doi.org/10.1016/j.coviro.2018.12.006>.
- Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, Wieser E, Taylor J, Berg S, Smith NJ, et al. Array programming with NumPy. *Nature*. 2020;585(7825):357–362. <https://doi.org/10.1038/s41586-020-2649-2>.
- Hill J, Enbody ED, Pettersson ME, Sprehn CG, Bekkevold D, Folkvord A, Laikre L, Kleinau G, Scheerer P, Andersson L. Recurrent convergent evolution at amino acid residue 261 in fish rhodopsin. *Proc Natl Acad Sci*. 2019;116(37):18473–18478. <https://doi.org/10.1073/pnas.1908332116>.
- Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. UFBoot2: improving the ultrafast bootstrap approximation. *Mol Biol Evol*. 2018;35(2):518–522. <https://doi.org/10.1093/molbev/msx281>.
- Hodcroft EB, Zuber M, Nadeau S, Vaughan TG, Crawford KHD, Althaus CL, Reichmuth ML, Bowen JE, Walls AC, Corti D, et al. Spread of a SARS-CoV-2 variant through Europe in the summer of 2020. *Nature*. 2021;595(7869):707–712. <https://doi.org/10.1038/s41586-021-03677-y>.
- Holm S. A simple sequentially rejective multiple test procedure. *Scand J Stat*. 1979;6:65–70. <https://www.jstor.org/stable/4615733>.
- Holmes EC, Zhang LQ, Simmonds P, Ludlam CA, Brown AJ. Convergent and divergent sequence evolution in the surface envelope glycoprotein of human immunodeficiency virus type 1 within a single infected patient. *Proc Natl Acad Sci U S A*. 1992;89(11):4835–4839. <https://doi.org/10.1073/pnas.89.11.4835>.
- Hu Y, Wu Q, Ma S, Ma T, Shan L, Wang X, Nie Y, Ning Z, Yan L, Xiu Y, et al. Comparative genomics reveals convergent evolution between the bamboo-eating giant and red pandas. *Proc Natl Acad Sci U S A*. 2017;114(5):1081–1086. <https://doi.org/10.1073/pnas.1613870114>.
- Huerta-Cepas J, Serra F, Bork P. ETE 3: reconstruction, analysis, and visualization of phylogenomic data. *Mol Biol Evol*. 2016;33(6):1635–1638. <https://doi.org/10.1093/molbev/msw046>.
- Hunter JD. Matplotlib: a 2D graphics environment. *Comput Sci Eng*. 2007;9(3):90–95. <https://doi.org/10.1109/MCSE.2007.55>.
- Ishikawa SA, Zhukova A, Iwasaki W, Gascuel O. A fast likelihood method to reconstruct and visualize ancestral scenarios. *Mol Biol Evol*. 2019;36(9):2069–2085. <https://doi.org/10.1093/molbev/msz131>.
- Jones DT, Taylor WR, Thornton JM. The rapid generation of mutation data matrices from protein sequences. *Bioinformatics*. 1992;8(3):275–282. <https://doi.org/10.1093/bioinformatics/8.3.275>.
- Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods*. 2017;14(6):587–589. <https://doi.org/10.1038/nmeth.4285>.
- Kass RE, Raftery AE. Bayes factors. *J Am Stat Assoc*. 1995;90(430):773–795. <https://doi.org/10.1080/01621459.1995.10476572>.
- Korber B, Fischer WM, Gnanakaran S, Yoon H, Theiler J, Abfalterer W, Hengartner N, Giorgi EE, Bhattacharya T, Foley B, et al. Tracking changes in SARS-CoV-2 spike: evidence that D614G increases infectivity of the COVID-19 virus. *Cell*. 2020;182(4):812–827.e19. <https://doi.org/10.1016/j.cell.2020.06.043>.
- Kusagawa S, Takebe Y, Yang R, Motomura K, Ampofo W, Brandful J, Koyanagi Y, Yamamoto N, Sata T, Ishikawa K, et al. Isolation and characterization of a full-length molecular DNA clone of Ghanaian HIV type 1 intersubtype A/G recombinant CRF02\_AG, which is replication competent in a restricted host range. *AIDS Res Hum Retroviruses*. 2001;17(7):649–655. <https://doi.org/10.1089/08922201300119761>.
- Larmuseau MH, Raeymaekers JA, Ruddick KG, Van Houdt JK, Volckaert FA. To see in different seas: spatial variation in the rhodopsin gene of the sand goby (*Pomatoschistus minutus*). *Mol Ecol*. 2009;18(20):4227–4239. <https://doi.org/10.1111/j.1365-294X.2009.04331.x>.
- Larter M, Dunbar-Wallis A, Berardi AE, Smith SD. Convergent evolution at the pathway level: predictable regulatory changes during flower color transitions. *Mol Biol Evol*. 2018;35(9):2159–2169. <https://doi.org/10.1093/molbev/msy117>.
- Lartillot N, Philippe H. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol Biol Evol*. 2004;21(6):1095–1109. <https://doi.org/10.1093/molbev/msh112>.
- Le SQ, Dang CC, Gascuel O. Modeling protein evolution with several amino acid replacement matrices depending on site rates. *Mol Biol Evol*. 2012;29(10):2921–2936. <https://doi.org/10.1093/molbev/mss112>.
- Le SQ, Gascuel O, Lartillot N. Empirical profile mixture models for phylogenetic reconstruction. *Bioinformatics*. 2008a;24(20):2317–2323. <https://doi.org/10.1093/bioinformatics/btn445>.
- Le SQ, Lartillot N, Gascuel O. Phylogenetic mixture models for proteins. *Philos. Trans. R. Soc. B Biol. Sci*. 2008b;363(1512):3965–3976. <https://doi.org/10.1098/rstb.2008.0180>.
- Lee J-H, Lewis KM, Mural TW, Kirilenko B, Boronovo B, Prange G, Koessl M, Huggenberger S, Kang C, Hiller M. Building superfast muscles: insights from molecular parallelism in fast-twitch muscle proteins in echolocating mammals. *bioRxiv* 244566. <https://doi.org/10.1101/244566>, 2018, preprint: not peer reviewed.
- Lemey P, Derdelinckx I, Rambaut A, Van Laethem K, Dumont S, Vermeulen S, Van Wijngaerden E, Vandamme A-M. Molecular footprint of drug-selective pressure in a human immunodeficiency virus transmission chain. *J Virol*. 2005;79(18):11981–11989. <https://doi.org/10.1128/JVI.79.18.11981-11989.2005>.
- Lemoine F, Gascuel O. Gotree/Goalign: toolkit and Go API to facilitate the development of phylogenetic workflows. *NAR Genom Bioinform*. 2021;3(3):lqab07. <https://doi.org/10.1093/nargab/lqab075>.
- Longdon B, Day JP, Alves JM, Smith SCL, Houslay TM, McGonigle JE, Tagliaferri L, Jiggins FM. Host shifts result in parallel genetic changes when viruses evolve in closely related species. *PLoS Pathog*. 2018;14(4):e1006951. <https://doi.org/10.1371/journal.ppat.1006951>.
- Losos JB. Convergence, adaptation, and constraint. *Evolution*. 2011;65(7):1827–1840. <https://doi.org/10.1111/j.1558-5646.2011.01289.x>.
- Lu B, Jin H, Fu J. Molecular convergent and parallel evolution among four high-elevation anuran species from the Tibetan region. *BMC Genomics*. 2020;21(1):839. <https://doi.org/10.1186/s12864-020-07269-4>.
- Malinsky M, Challis RJ, Tyers AM, Schifffels S, Terai Y, Ngatunga BP, Miska EA, Durbin R, Genner MJ, Turner GF. Genomic islands of speciation separate cichlid ecomorphs in an east African crater lake. *Science*. 2015;350(6267):1493–1498. <https://doi.org/10.1126/science.aac9927>.
- Marcovitz A, Turakhia Y, Chen HI, Gloude-mans M, Braun BA, Wang H, Bejerano G. 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*. 2019;116(42):21094–21103. <https://doi.org/10.1073/pnas.1818532116>.
- Martin DP, Weaver S, Tegally H, San EJ, Shank SD, Wilkinson E, Giandhari J, Naidoo S, Pillay Y, Singh L, et al. The emergence

- and ongoing convergent evolution of the N501Y lineages coincides with a major global shift in the SARS-CoV-2 selective landscape. medRxiv. <https://doi.org/10.1101/2021.02.23.21252268>, 2021 Jul 25, preprint: not peer reviewed.
- Martin DP, Weaver S, Tegally H, San JE, Shank SD, Wilkinson E, Lucaci AG, Giandhari J, Naidoo S, Pillay Y, et al. The emergence and ongoing convergent evolution of the SARS-CoV-2 N501Y lineages. *Cell*. 2021;184(20):5189–5200.e7. <https://doi.org/10.1016/j.cell.2021.09.003>.
- McKinney W. Data structures for statistical computing in python. In: van der Walt S, Millman J, editors. *Proceedings of the 9th Python in Science Conference*; 2010. p. 51–56.
- Murrell B, de Oliveira T, Seebregts C, Pond SLK, Scheffler K, Consortium on behalf of the SAT and RN (SATuRN). Modeling HIV-1 drug resistance as episodic directional selection. *PLoS Comput Biol*. 2012;8(5):e1002507. <https://doi.org/10.1371/journal.pcbi.1002507>.
- Murrell B, Moola S, Mabona A, Weighill T, Sheward D, Pond SLK, Scheffler K. FUBAR: a fast, unconstrained Bayesian Approximation for inferring selection. *Mol Biol Evol*. 2013;30(5):1196–1205. <https://doi.org/10.1093/molbev/mst030>.
- Murrell B, Wertheim JO, Moola S, Weighill T, Scheffler K, Pond SLK. Detecting individual sites subject to episodic diversifying selection. *PLoS Genet*. 2012;8(7):e1002764. <https://doi.org/10.1371/journal.pgen.1002764>.
- Muschick M, Indermaur A, Salzburger W. Convergent evolution within an adaptive radiation of cichlid fishes. *Curr Biol*. 2012;22(24):2362–2368. <https://doi.org/10.1016/j.cub.2012.10.048>.
- Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol*. 2015;32(1):268–274. <https://doi.org/10.1093/molbev/msu300>.
- Nickle DC, Heath L, Jensen MA, Gilbert PB, Mullins JI, Pond SLK. HIV-Specific Probabilistic models of protein evolution. *PLoS One*. 2007;2(6):e503. <https://doi.org/10.1371/journal.pone.0000503>.
- O'Reilly JE, Puttick MN, Parry L, Tanner AR, Tarver JE, Fleming J, Pisani D, Donoghue PCJ. Bayesian methods outperform parsimony but at the expense of precision in the estimation of phylogeny from discrete morphological data. *Biol Lett*. 2016;12(4):20160081. <https://doi.org/10.1098/rsbl.2016.0081>.
- Pagel M. Detecting correlated evolution on phylogenies: a general method for the comparative analysis of discrete characters. *Proc R Soc Lond B Biol Sci*. 1994;255(1342):37–45. <https://doi.org/10.1098/rspb.1994.0006>.
- Pagel M, Meade A. Bayesian analysis of correlated evolution of discrete characters by reversible-jump Markov chain Monte Carlo. *Am Nat*. 2006;167(6):808–825. <https://doi.org/10.1086/503444>.
- Paradis E, Claude J, Strimmer K. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics*. 2004;20(2):289–290. <https://doi.org/10.1093/bioinformatics/btg412>.
- Parker J, Tsagkogeorga G, Cotton JA, Liu Y, Provero P, Stupka E, Rossiter SJ. Genome-wide signatures of convergent evolution in echolocating mammals. *Nature*. 2013;502(7470):228–231. <https://doi.org/10.1038/nature12511>.
- Parto S, Lartillot N. Detecting consistent patterns of directional adaptation using differential selection codon models. *BMC Evol Biol*. 2017; 17(1):147. <https://doi.org/10.1186/s12862-017-0979-y>.
- Parto S, Lartillot N. Molecular adaptation in Rubisco: discriminating between convergent evolution and positive selection using mechanistic and classical codon models. *PLoS One*. 2018;13(2):e0192697. <https://doi.org/10.1371/journal.pone.0192697>.
- Pond SLK, Frost SDW, Muse SV. Hyphy: hypothesis testing using phylogenies. *Bioinformatics*. 2005;21(5):676–679. <https://doi.org/10.1093/bioinformatics/bti079>.
- Pond SLK, Murrell B, Poon AFY. Evolution of viral genomes: interplay between selection, recombination, and other forces. *Methods Mol Biol*. 2012;856:239–272. [https://doi.org/10.1007/978-1-61779-585-5\\_10](https://doi.org/10.1007/978-1-61779-585-5_10).
- Price MN, Dehal PS, Arkin AP. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol Biol Evol*. 2009;26(7):1641–1650. <https://doi.org/10.1093/molbev/msp077>.
- Rey C, Guéguen L, Sémon M, Boussau B. Accurate detection of convergent amino-acid evolution with PCOC. *Mol Biol Evol*. 2018;35(9):2296–2306. <https://doi.org/10.1093/molbev/msy114>.
- Rey C, Veber P, Guéguen L, Lartillot N, Sémon M, Boussau B. Detecting adaptive convergent amino acid evolution. *Philos Trans R Soc B Biol Sci*. 2019;374(1777):20180234. <https://doi.org/10.1098/rstb.2018.0234>.
- Rhee S-Y. Human immunodeficiency virus reverse transcriptase and protease sequence database. *Nucleic Acids Res*. 2003;31(1):298–303. <https://doi.org/10.1093/nar/gkg100>.
- Rokas A, Carroll SB. Frequent and widespread parallel evolution of protein sequences. *Mol Biol Evol*. 2008;25(9):1943–1953. <https://doi.org/10.1093/molbev/msn143>.
- Rosenblum EB, Parent CE, Brandt EE. The molecular basis of phenotypic convergence. *Annu Rev Ecol Syst*. 2014;45(1):203–226. <https://doi.org/10.1146/annurev-ecolsys-120213-091851>.
- Schultz A-K, Bulla I, Abdou-Chekarou M, Gordien E, Morgenstern B, Zoulim F, Dény P, Stanke M. jpHMM: recombination analysis in viruses with circular genomes such as the hepatitis B virus. *Nucleic Acids Res*. 2012;40(W1):W193–W198. <https://doi.org/10.1093/nar/gks414>.
- Sluis-Cremer N, Jordan MR, Huber K, Wallis CL, Bertagnolio S, Mellors JW, Parkin NT, Harrigan PR. E138a in HIV-1 reverse transcriptase is more common in subtype C than B: implications for rilpivirine use in resource-limited settings. *Antiviral Res*. 2014;107:31–34. <https://doi.org/10.1016/j.antiviral.2014.04.001>.
- Soubrier J, Steel M, Lee MSY, Der Sarkissian C, Guindon S, Ho SYW, Cooper A. The influence of rate heterogeneity among sites on the time dependence of molecular rates. *Mol Biol Evol*. 2012;29(11):3345–3358. <https://doi.org/10.1093/molbev/mss140>.
- Spady TC, Seehausen O, Loew ER, Jordan RC, Kocher TD, Carleton KL. Adaptive molecular evolution in the opsin genes of rapidly speciating cichlid species. *Mol Biol Evol*. 2005;22(6):1412–1422. <https://doi.org/10.1093/molbev/msi137>.
- Starr TN, Greaney AJ, Hilton SK, Ellis D, Crawford KHD, Dingens AS, Navarro MJ, Bowen JE, Tortorici MA, Walls AC, et al. Deep mutational scanning of SARS-CoV-2 receptor binding domain reveals constraints on folding and ACE2 binding. *Cell*. 2020;182(5):1295–1310.e20. <https://doi.org/10.1016/j.cell.2020.08.012>.
- Stern DL. The genetic causes of convergent evolution. *Nat Rev Genet*. 2013;14(11):751–764. <https://doi.org/10.1038/nrg3483>.
- Stoltzfus A, McCandlish DM. Mutational biases influence parallel adaptation. *Mol Biol Evol*. 2017;34(9):2163–2172. <https://doi.org/10.1093/molbev/msx180>.
- Storz JF. Causes of molecular convergence and parallelism in protein evolution. *Nat Rev Genet*. 2016;17(4):239–250. <https://doi.org/10.1038/nrg.2016.11>.
- Susko E, Field C, Blouin C, Roger AJ. Estimation of rates-across-sites distributions in phylogenetic substitution models. *Syst Biol*. 2003;52(5):594–603. <https://doi.org/10.1080/10635150390235395>.
- Svensson P, Bläsing OE, Westhoff P. Evolution of C4 phosphoenolpyruvate carboxylase. *Arch Biochem Biophys*. 2003;414(2):180–188. [https://doi.org/10.1016/S0003-9861\(03\)00165-6](https://doi.org/10.1016/S0003-9861(03)00165-6).
- Tamuri AU, dos Reis M, Hay AJ, Goldstein RA. Identifying changes in selective constraints: host shifts in influenza. *PLoS Comput Biol*. 2009;5(11):e1000564. <https://doi.org/10.1371/journal.pcbi.1000564>.



- Thomas GWC, Hahn MW. Determining the null model for detecting adaptive convergence from genomic data: a case study using echolocating mammals. *Mol Biol Evol.* 2015;32(5):1232–1236. <https://doi.org/10.1093/molbev/msv013>.
- Tommaso PD, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. Nextflow enables reproducible computational workflows. *Nat Biotechnol.* 2017;35(4):316–319. <https://doi.org/10.1038/nbt.3820>.
- Tsetsarkin KA, Vanlandingham DL, McGee CE, Higgs S. A single mutation in chikungunya virus affects vector specificity and epidemic potential. *PLoS Pathog.* 2007;3(12):e201. <https://doi.org/10.1371/journal.ppat.0030201>.
- Ujvari B, Casewell NR, Sunagar K, Arbuckle K, Wüster W, Lo N, O’Meally D, Beckmann C, King GF, Deplazes E, et al. Widespread convergence in toxin resistance by predictable molecular evolution. *Proc Natl Acad Sci U S A.* 2015;112(38):11911–11916. <https://doi.org/10.1073/pnas.1511706112>.
- van Ditmarsch D, Boyle KE, Sakhtah H, Oyler JE, Nadell CD, Déziel É, Dietrich LEP, Xavier JB. Convergent evolution of hyperswarming leads to impaired biofilm formation in pathogenic bacteria. *Cell Rep.* 2013;4(4):697–708. <https://doi.org/10.1016/j.celrep.2013.07.026>.
- van Dorp L, Acman M, Richard D, Shaw LP, Ford CE, Ormond L, Owen CJ, Pang J, Tan CCS, Boshier FAT, et al. Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. *Infect Genet Evol.* 2020;83:104351. <https://doi.org/10.1016/j.meegid.2020.104351>.
- Villabona-Arenas CJ, Vidal N, Guichet E, Serrano L, Delaporte E, Gascuel O, Peeters M. In-depth analysis of HIV-1 drug resistance mutations in HIV-infected individuals failing first-line regimens in West and Central Africa. *AIDS.* 2016;30(17):2577–2589. <https://doi.org/10.1097/QAD.0000000000001233>.
- Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J, et al. Scipy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods.* 2020;17(3):261–272. <https://doi.org/10.1038/s41592-019-0686-2>.
- Waskom ML. Seaborn: statistical data visualization. *J. Open Source Softw.* 2021;6(60):3021. <https://doi.org/10.21105/joss.03021>.
- Wensing AM, Calvez V, Ceccherini-Silberstein F, Charpentier C, Günthard HF, Paredes R, Shafer RW, Richman DD. 2019 update of the drug resistance mutations in HIV-1. *Top Antivir Med.* 2019;27:111–121. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6892618/>.
- Xu S, Wang J, Guo Z, He Z, Shi S. Genomic convergence in the adaptation to extreme environments. *Plant Commun.* 2020;1(6):100117. <https://doi.org/10.1016/j.xplc.2020.100117>.
- Yokoyama S. Evolution of dim-light and color vision pigments. *Annu Rev Genom Hum Genet.* 2008;9(1):259–282. <https://doi.org/10.1146/annurev.genom.9.081307.164228>.
- Zhang J. Parallel adaptive origins of digestive RNases in Asian and African leaf monkeys. *Nat Genet.* 2006;38(7):819–823. <https://doi.org/10.1038/ng1812>.
- Zhang J, Kumar S. Detection of convergent and parallel evolution at the amino acid sequence level. *Mol Biol Evol.* 1997;14(5):527–536. <https://doi.org/10.1093/oxfordjournals.molbev.a025789>.
- Zhen Y, Aardema ML, Medina EM, Schumer M, Andolfatto P. Parallel molecular evolution in an herbivore community. *Science.* 2012;337(6102):1634–1637. <https://doi.org/10.1126/science.1226630>.
- Zou Z, Zhang J. Are convergent and parallel amino acid substitutions in protein evolution more prevalent than neutral expectations? *Mol Biol Evol.* 2015a;32(8):2085–2096. <https://doi.org/10.1093/molbev/msv091>.
- Zou Z, Zhang J. No genome-wide protein sequence convergence for echolocation. *Mol Biol Evol.* 2015b;32(5):1237–1241. <https://doi.org/10.1093/molbev/msv014>.

Associate editor: Carolin Kosiol