Sensitive Detection of Site-wise Convergent Evolution in Large Protein Alignments with ConDor

Marie Morel, Frédéric Lemoine, Olivier Gascuel

Sensitive Detection of Site-wise Convergent Evolution in Large Protein Alignments with ConDor

Marie MOREL[1,2], Frédéric LEMOINE[1,3] and Olivier GASCUEL[1,4]

1- Unité de Bioinformatique Évolutive - Département Biologie computationnelle, Institut Pasteur, 25-28 rue du Dr Roux, 75015 - Paris, France

2- Université de Paris, 5 rue Thomas Mann, 75013 - Paris, France

3- Hub de Bioinformatique et Biostatistique - Département Biologie computationnelle, Institut Pasteur, 25-28 rue du Dr Roux, 75015 - Paris, France

4- Institut de Systématique, Evolution, Biodiversité (ISYEB - UMR 7205, CNRS, Muséum National d'Histoire Naturelle, EPHE, SU, UA), 57 rue Cuvier, 75005 - Paris, France [Current address]

Corresponding Authors: marie.morel@pasteur.fr, olivier.gascuel@mnhn.fr.

Abstract

Evolutionary convergences are observed at all levels, from phenotype to DNA and protein sequences, and the changes observed at these different levels tend to be strongly correlated. Here we propose a simulation-based method to detect positions under convergent evolution in large protein alignments, without prior knowledge on the phenotype and environmental constraints. A phylogeny is inferred from the data and used in simulations to estimate the expected number of amino-acid changes in stable evolutionary constraints (null model) for each position. Similarly, we count the number of mutations towards the same amino acid in the data and test if they are occurring more often than expected.

We applied our method to two real datasets: HIV reverse transcriptase and fish rhodopsin, and to HIV-like simulated data. On the latter, with known convergent events and substitution model, we detected on average two third of these events, with a low fraction of false positives. With HIV data, one knows that drug resistance mutations (DRMs) are convergent. Even without any knowledge of patient treatment status, we retrieved more than 70% of positions corresponding to known DRMs. On the rhodopsin dataset, four substitutions are supposed to be convergent, as they change the maximum wavelength absorption of the photoreceptor and occurred several times independently during evolution. We detected three of them.

These results demonstrate the potential of the method to target specific mutations to be further studied experimentally or, for example, using a nonsynonymous/synonymous rate ratio approach. Our software named ConDor is available at http://condor.pasteur.cloud.

Key Words: molecular evolution, phylogenetics, selection, adaptation, convergence, HIV, resistance to drugs, rhodopsin.

Introduction

Convergent evolution can be defined as the independent acquisition of similar traits in distinct lineages over the course of evolution. The studied traits can be behavioral, morphological, molecular, etc. In each category, traits can be quantitative (size, length, etc.), binary (presence or absence of a given phenotype) or categorical (a trait is subdivided into several classes). Recently, several studies focused on the molecular level, following the hypothesis that convergent phenotypes generally result from the same genetic changes (Stern 2013; Rosenblum et al. 2014; Storz 2016). At the protein level, we commonly distinguish parallel mutations (a change towards the same amino acid is observed from the same ancestral amino acids), convergent mutations (change towards the same amino acid, from different ancestral amino acids) and reversions (mutations that restore an amino acid previously lost during evolution).

Examples of evolutionary convergence at the molecular level have been highlighted in higher eukaryotes in relation 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), metabolic or morphological changes and the acquisition of new abilities (Davies et al. 2012; Parker et al. 2013; Thomas and Hahn 2015; Parto and Lartillot 2018; Marcovitz et al. 2019; Chai et al. 2020; Lee et al. 2018 Jan 8). Similarly, when submitted to constraints such as experimental conditions or drug treatments, microorganisms and viruses 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, several authors found adaptive convergence in experimental populations of RNA viruses (Cuevas et al. 2002) and in pathogenic bacteria (van Ditmarsch et al. 2013). In natural conditions, evolutionary convergence was found in viruses having experienced host shifts (Longdon et al. 2018) and changes of 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; Thomas and Hahn 2015; Zou and Zhang 2015; Chabrol et al. 2018; Rey et al. 2018; etc). They are all based on the prior knowledge or observation of a convergent phenotype and aim to identify protein mutations that correlate with the presence of the converging trait. Two major types of approaches can be distinguished, depending on the scale at which evolutionary convergence is studied. Some approaches aim to identify which coding genes show 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 Zou *et al*, Thomas *et al* and Chabrol *et al* were applied to the search of genes responsible for echolocation in mammals (Thomas and Hahn 2015; Zou and Zhang 2015; Chabrol et al. 2018). In the second configuration, the coding genes responsible for the convergent phenotype have already been identified and the methods focus on the detection of converging mutations at the site level; for example, Zhang *et al* identified convergent and parallel mutations in stomach lysozyme sequences of foregut fermenters (Zhang and Kumar 1997). Similarly, Zhang found parallel substitutions in colobine pancreatic ribonucleases (Zhang 2006), and Rey *et al* found positions with convergent substitutions in the PEPC protein occurring jointly with the transition toward C4 metabolism in sedges (Rey et al. 2018).

As mentioned above, both types of approaches rely on the identification of mutations correlated with a convergent phenotype. More specifically, they rely on ancestral sequence reconstruction to detect independent changes that occurred at the same position towards the same amino acid within different lineages having the convergent phenotype of interest. Considering that a change towards the exact same amino acid could be too strict, since several amino acids can have similar physico-chemical properties, Rey *et al* looked for shifts in amino-acid profiles (Le, Gascuel, et al. 2008; Rey et al. 2018). Identifying those amino-acid changes or shifts is not sufficient as they could occur by chance. To statistically test the strength of convergent evolution at the site level, Chabrol *et al* defined a new "convergence index" and used simulations to estimate de distribution of this index in a null, non-convergent model (Chabrol et al. 2018). Rey *et al* selected convergent positions based on the log ratio of the posterior of the studied position assuming the convergent model (shift of the amino-acid profile) versus the one obtained with the null model (homogeneous amino-acid profile along all branches) (Rey et al. 2018).

Testing the significance of convergent (or parallel or revertant) changes at the site level in proteins has many potential applications. In the case of complex eukaryotic or 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 site-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 is in fact what was observed with 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 evolutionary advantageous for the virus (Korber et al. 2020; Martin et al. 2021; van Dorp et al. 2020). Indeed, in viruses it is often

easier to identify a mutation of interest than to observe the effects of that mutation given how difficult the phenotype of a virus and the environmental conditions in which it evolves are to access.

In some ways, identifying those mutations of interest presents similarities with the detection of positions under positive selection (Goldman and Yang 1994). The idea is indeed to identify mutations that could be advantageous as they are found independently more often than expected under a neutral (or purifying) model of evolution. Positive selection can be inferred at a position if the rate of non-synonymous substitutions exceeds the rate of synonymous substitutions. These substitutions can be towards a specific amino acid or any change from the original amino acid. This is, for example, the case in immune avoidance where the trend to mutate towards any new amino acid at the antigenic positions is generally favorable and positively selected. Conversely, in the case of convergent evolution, we are interested in substitutions towards one or a few similar amino acids. We thus expect that positive selection is a necessary, but not sufficient condition for a position to be convergent. We shall see that our results confirm this intuition.

Here we propose a method designed to detect site-wise convergent evolution in large amino-acid alignments without prior knowledge of phenotype. This method performs detailed analysis at the gene/protein level, with typical application to viruses, but also to specific genes known to be involved in phenotypic convergence (Hill et al. 2019). We are interested in changes towards a target amino acid regardless of the ancestral amino acids that lead to the difference in the extant amino acid sequences. In other words, parallel, convergent and revertant mutations are considered indifferently and we consider different target amino acids as different events. The observed number of amino-acid changes is estimated using ancestral character reconstruction, and their expected number in a null model using computer simulations. In the following sections, we describe this approach that is implemented in a software and web service named ConDor (Convergence Detector; condor.pasteur.cloud). Its performance is assessed on HIV-like simulated datasets, on a real HIV reverse transcriptase dataset and on a fish rhodopsin dataset. We notably compare its performance in retrieving drug resistance mutations in HIV, with the results of a standard positive selection-based approach.

New Approaches

Simulation-based approach

Our method identifies amino-acid mutations that emerged several times in independent lineages and occurred significantly more frequently than expected under a null model of evolution. The workflow of the method is presented in Figure 1. It is made up of four main steps: (1) estimate the parameters of the null model from the real data (phylogenetic tree, parameters of the substitution model, site-wise evolutionary rates, etc.); (2) infer ancestral amino acids and count the number of

observed emergence events of mutations (EEMs) for every position and amino acid of interest in the real data; (3) simulate new datasets under the inferred null model and count simulated EEMs; (4) compare the observed and the simulated numbers of EEMs, and then determine which mutations occurred significantly more often than expected by chance, assuming the null model. Such mutations are considered as convergent events.

The null model is inferred from the input alignment. The selected substitution model, along with amino-acid frequencies, rates-across-sites distribution parameters, tree topology, branch lengths and site-wise evolutionary rates are assumed to represent the data without convergence. We make this assumption because using large alignments (>1000 sequences), we consider that mutations resulting from convergent evolution are rare enough to have a negligible influence on tree and parameter inference. The reconstructed phylogeny is then rooted using the provided outgroup. This is essential 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) is achieved using a maximum likelihood approach, implemented in PastML (Ishikawa et al. 2019). We use the "maximum a posteriori" (MAP) method in which the state with the highest marginal posterior is selected at each tree node. Once all ancestral sequences are reconstructed and associated to the nodes in the phylogeny, we identify where independent amino-acid changes occurred in the tree and count them as explained in the subsection "Counting emergence events". This counting gives the observed number of EEMs for each alignment position and amino acid under study, that is, those that are observed at the given position enough time (≥12 sequences in our HIV experiments) and in more than 2 independent clades.

We then simulate the expected evolution without convergence of each position of the alignment many times (10,000 in our experiments). We do not use the root sequence reconstructed by ACR as a start, but draw amino acids based on their marginal posterior probabilities. Taking only the amino acid with the highest posterior tends to bias the simulations and yields poorer results (not shown), especially if the reconstruction is uncertain (e.g., two amino acids with posteriors of 0.55 and 0.45). Simulations are carried out along the inferred tree, and then we count the simulated numbers of EEMs (10,000 values per position and per studied AA) using the algorithm presented below. For example, let us consider the mutation M41L from our real HIV dataset, where at position 41, a Methionine (M) is substituted by a Leucine (L) in 211 sequences. The observed number of EEMs towards L is 47, which is smaller than 211 as in some subtrees all tips have L, corresponding to only 1 EEM. This number is compared to the distribution of the number of EEMs towards L, starting from an M at the tree root every time (no ambiguity in ACR), among 10,000 simulations in the null model; this

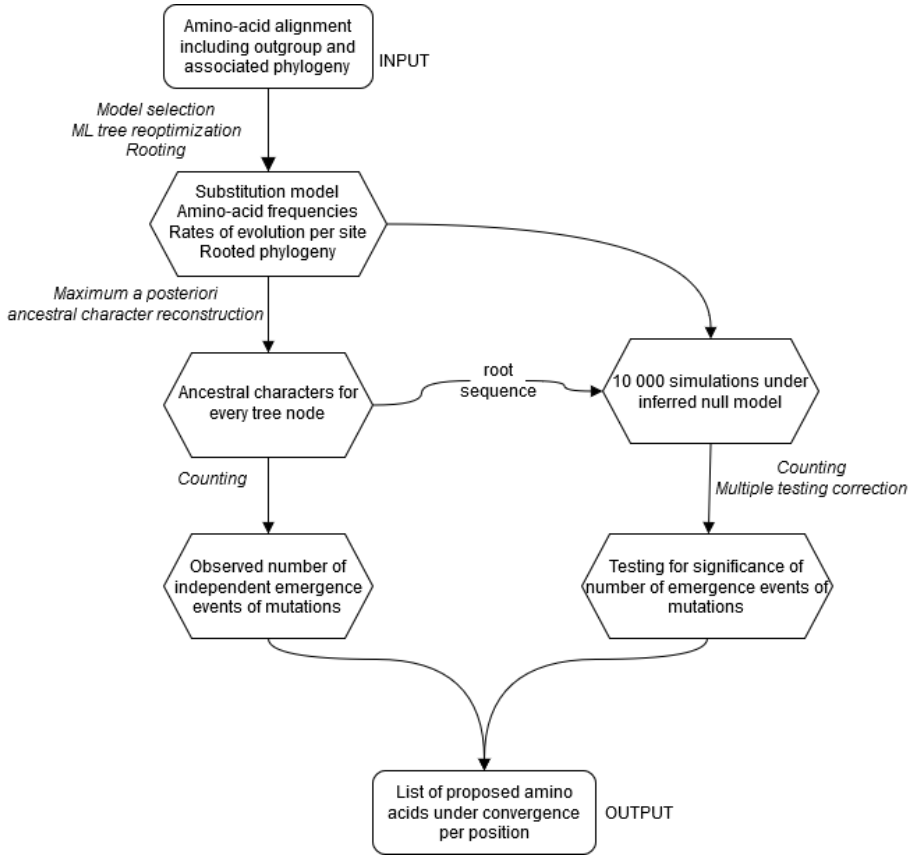


Figure 1: Flowchart of the method. *The method takes as input an amino-acid alignment, which is used for inference of the null model (phylogenetic tree, substitution model and its parameters, site-wise evolutionary rates, etc.) and ancestral reconstruction. The reconstructed tree and (probabilistic) root sequence are used to simulate 10,000 alignments under the null model. The output is a list of amino-acid changes per position that seem to be convergent, as they emerge more often in the input alignment than in simulations.*

number of simulated EEMs 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 for each observed mutation, which is equal to 0 in our M41L example. After correction for multiple testing, mutations with p-values lower than the rejection criterion are considered as resulting from convergent evolution.

Counting independent emergence events of mutations (EEMs)

The observed number of EEMs is inferred by ACR based on the input sequences, while the expected number of EEMs and its distribution are estimated from many simulations evolving the probabilistic root sequence along the inferred tree. In simulations, changes may appear which cannot be inferred by ACR, in particular when they are not transmitted to any tree leaf. In this case, the expected number of changes artificially deviates from the ACR-based “observed” number of EEMs. This effect is even more pronounced on positions with rapid evolution since more changes are expected. Thus, only the changes transmitted to at least one leaf are counted in our method, since they are the only ones that could be found by ACR. Note, moreover, that generally we are only interested by the amino acids present in the available, actual sequences, and rarely by those that are never observed.

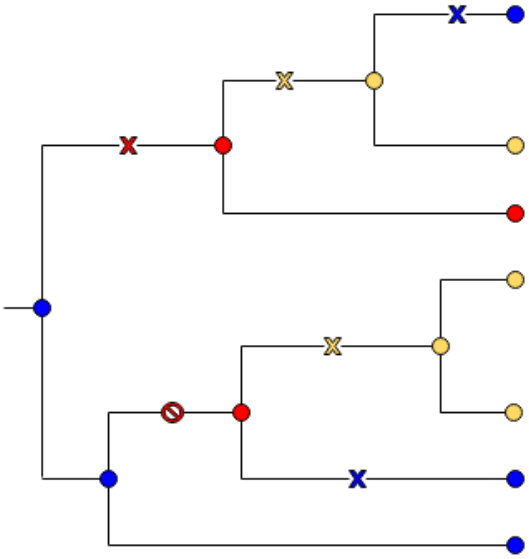


Figure 2: Counting independent emergence-events. *In this tree we count two (parallel) EEMs towards the yellow state, two reversions towards the blue state and only one EEM towards the red state since the blue-to-red mutation at the bottom of the tree is not transmitted to any leaf and thus not counted.*

In the tree illustrated in Figure 2, we have 6 changes along the branches, which are represented either by a cross or by a NO symbol. The NO symbol stands for a change of the blue state towards the red one, but the red state is then lost in the subtree. With ACR, this node would have been either blue or yellow, but never red, while this might occur in simulations. Thus, in our counting, we do not consider this EEM towards the red state, and only the one in the upper subtree is counted. The two yellow crosses mark changes transmitted to one leaf in the upper subtree and two tips in the bottom subtree; thus, both are counted. The two blue crosses mark a return to the ancestral state present at the tree root and are reversions. Even though we count EEMs without making the difference between convergent, parallel and revertant events, we retain the information during the counting process for interpretation afterwards.

Creation of a synthetic HIV-like dataset

To our knowledge, there is no convergent evolution model allowing simulating thousands of sequences without prior knowledge of the phenotype or environmental constraints. We therefore created our own convergent dataset inspired from a real case of convergence. Drug resistance mutations (DRMs) occur independently in patients receiving a drug treatment and thus are a perfect example of evolutionary convergence. In HIV, they are well characterized and studied since their emergence can lead to treatment failure and transmission of resistant viruses. They are mostly found in proteins targeted by the antiretroviral treatment: the protease, the reverse transcriptase and the integrase. The list of known DRMs affecting these proteins is publicly available at https://hivdb.stanford.edu/. DRMs are written as “XposY”, with X the ancestral amino acid, “pos” the

position of the substitution in the protein alignment, with numbering based on the reference sequence HXB2, and Y the mutated amino acid.

We extracted positions and sequences with DRMs from a real HIV polymerase amino-acid dataset (described in Material and Methods) and replaced the corresponding residues with gaps in the multiple sequence alignment (MSA). DRMs were retrieved from the Stanford University Drug resistance database "Essential DRM Data" lists (https://hivdb.stanford.edu/pages/poc.html). We then reconstructed a tree, estimated the substitution model parameters, and inferred the rates of evolution per position (see Material and Methods for details). This tree represents the evolutionary relationships between the sequences in the real data and its structure is not affected by the DRMs. Such procedure is standard when inferring trees from HIV sequences, to avoid convergence perturbations of the inferred tree. We then simulated the evolution of HXB2 (GenBank™ accession number K03455) reference sequence (reverse transcriptase only) along this tree. We performed the simulations 5 times for robustness purposes, resulting in 5 MSAs without convergence. DRMs were then manually added in the sequences and positions where they were found in the real MSA. For example, mutation M41L was found in the real data in 211 sequences, so we replaced in the corresponding sequences of our simulated alignment, any amino acid found at position 41 (which turned out to all be methionine (M), the same as HXB2 at that position) by a leucine (L). Since we used the real data tree to create the synthetic dataset, we did not randomly place DRMs, which would make the detection task easy. Methionine and leucine are closely related, with a genetic barrier of 1 (M and L can exchange through a single nucleotide substitution), and the detection is much more difficult than between highly different amino acids (e.g., D and W with a genetic barrier of 3). As M41L is at the same position and in the same tip sequences in our synthetic dataset as in real data, detecting this convergence in the synthetic and real datasets should be of similar difficulty. This guarantees a certain realism of our simulated data. We implemented this insertion procedure for the 37 most common DRMs of our real dataset (i.e., present in ≥ 12 sequences and > 2 independent clades). The five-resulting synthetic MSAs thus have no convergent events, except the "realistic" added DRMs.

Results

We applied our approach on three datasets for which we knew a priori mutations due to convergent evolution: (1) an HIV-like synthetic dataset with "realistic" added DRMs; (2) a real HIV dataset of reverse transcriptase with 20% sequences with DRMs; and (3) a real 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. On simulated data, we know exactly which mutations are true convergent events or not. With real data, even if we know certain convergent

events, some other mutations are likely convergent, but are unknown. In other words, we will be able to assess the rate of false positives with simulated data, but not with real data, where the method sensitivity in detecting the known convergence events will be the main criterion.

Synthetic HIV-like alignment

The dataset consists in 5 MSAs of 3,551 sequences and 250 amino acids each, mimicking HIV reverse transcriptase and simulated with HIVb model of evolution (Nickle et al. 2007). In total, 37 DRMs were placed in each of the MSAs at 27 distinct positions (see above). These DRMs are found in at least 12 sequences. The most common one, M184V, is found in 273 (8%) sequences. However, 19 DRMs are found in less than 1% of the sequences (i.e., in 12 to 33 sequences) so they are expected to be difficult to detect.

Performance of the method on the detection of true convergence

The model inferred from the datasets by ModelFinder (Kalyaanamoorthy et al. 2017) was HIVb, which is the model we used for generating them. Thus, the whole analysis (tree reconstruction, ACR and simulations) was first achieved with HIVb, which is the true model of evolution. In a second stage, we also used JTT (Jones et al. 1992), to study the impact of model misspecification. We tested on average 441 mutations per dataset, the ones present in at least 12 sequences and with more than 2 EEMs.

Using HIVb, on average 27.4 mutations are found to be convergent by our method, 26 of which are true DRMs (out of 37 added DRMs). The numbers of EEMs for the DRMs range from 9 (K101P) to 225 (M184V). We detect DRMs with a higher number of EEMs better, and especially those with more than 30 EEMs that we detect 90% as convergent, as illustrated in Figure 3a. If there are several DRMs at one position, we often only detect the most frequent one(s). For example, at position 219 (Fig3a, bold characters) we detect mutations towards Q and E but not N. Similarly, at position 215, we do not detect mutations towards S and D. However, we detect mutation T215C although there are fewer EEMs towards C. This is explained by the low substitution rate between T and C (BLOSUM62 score = -1), and thus few changes are expected from T to C; 2 EEMs between T and C at position 215 are observed on average over the 10,000 null-model simulations, while 17 are found in the synthetic, convergent MSAs. If we focus on detecting positions with convergence (e.g., position 219) rather than DRMs (e.g., K219Q, K219E, K219N, etc.) accuracy increases and we detect on average 22 of the 27 convergent positions for all datasets, while the number of false positives remains equal to 1.4 on average (Tab. 1).

		DRMs		Non-DRMs		Total	
	Model	**Mutation**	**Position**	**Mutation**	**Position**	**Mutation**	**Position**
Detected	**HIVb**	26 (±1.2)	22 (±1.2)	1.4 (±1.14)	1.4 (±1.1)	27.4 (±1.5)	23.4 (±1.3)
	JTT	25.2 (±0.8)	22.4 (±0.5)	17.2 (±3)	15.8 (±2.6)	42.4 (±2.5)	38.2 (±2.3)
Not detected	**HIVb**	11 (±1.2)	5 (±1.2)	402.8 (±6.8)	91.2 (±1.6)	413.8 (±7.15)	96.2 (±2.5)
	JTT	11.8 (±0.8)	4.6 (±0.5)	392.8 (±13.7)	70.2 (±4.3)	404.6 (±14)	74.8(±4.3)
Total	**HIVb**	37	27	404.2 (±6)	92.6 (±1.7)	441.2 (±5.6)	113.2 (±1.5)
	JTT	37	27	410 (±14.3)	86 (±3.4)	447 (±14.3)	113 (±3.4)

Table 1: Method accuracy with synthetic data. *We display the number of mutations and positions (1) detected and (2) tested but not detected on the 5 synthetic HIV-like MSAs, analyzed with HIVb and JTT substitution models. We report the average for the 5 datasets and the standard deviation between parentheses. True positives are at the intersection between detected and DRMs, and false positives at the intersection between detected and non-DRMs (i.e., mutations resulting from the evolution of the root sequence under the null model). In these experiments, we tested all mutations (regardless of their DRM status) exhibiting more than 2 EEMs and found in at least 12 sequences.*

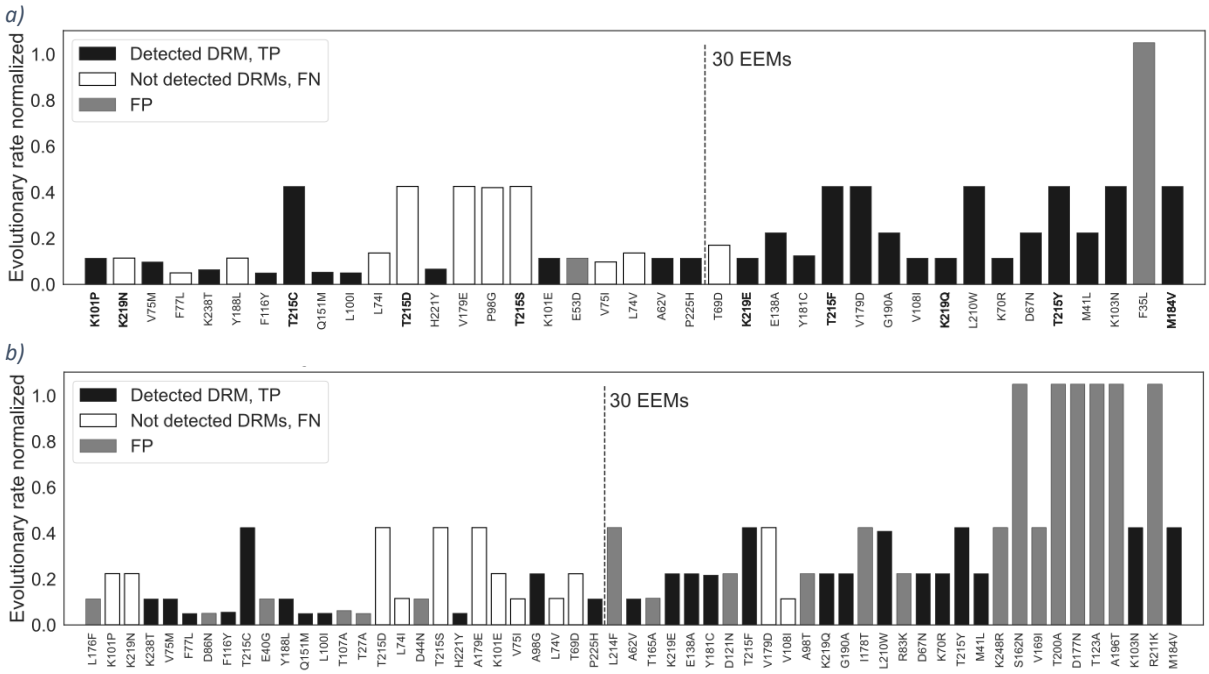


Figure 3: DRMs detection and convergent candidates with synthetic data. *We display the detected DRMs (true positives, TP, in black), not detected DRMs (false negatives, FN, in white) and false positives (FP, in grey) on one of the synthetic HIV-like MSA, analyzed with (a) HIVb substitution model or (b) JTT. Mutations are sorted by their number of EEMs on the x-axis. Mutations in bold characters are discussed in text. The evolutionary rate (normalized between 0 and 1) of each mutation position is reported on the y-axis. The dashed vertical line represents the limit of 30 EEMs, all events on the left have less than 30 EEMS.*

In the MSA represented in Figure 3a, we find two false positives, one of which exhibits a very high evolutionary rate. The same results are observed in the other MSAs: as the evolutionary rate increases, more changes are observed at the given position and thus more variability and uncertainty in the simulations. Thus, very fast positions can bias convergence detection and lead to the detection of false positives. As expected, we observe very few false positives when analyzing our synthetic datasets with the true model of evolution. Since we never have the true model of evolution with real data, we tested the effect of model violation on the synthetic dataset. We run the whole analysis with JTT, for tree reconstruction, ACR and simulations, instead of letting the workflow infer and use the best model of evolution (here HIVb).

The strongest effect can be seen on the number of false positives, which increases from 1.4 to 17.2 (Tab. 1). Compared to the number of negatives, this remains very low, with 4% of falsely detected random mutations (i.e., mutations resulting from the evolution under the null model) among more than 400. Moreover, as we see Figure 3b, the 6 events with highest evolutionary rate are false positives. The same trend can be observed in the other MSAs, where the mutations at positions with highest rates are always false positives. Since the substitution model does not correspond exactly to the synthetic datasets being analyzed, we tend to detect more mutations as convergent, but this does not impact the detection of DRMs. DRM detection remains sensitive and is robust to model violation, as we still detect more than 25 (among 37) DRMs, and more than 22 DRM positions (among 27), that is, nearly the same true positive fractions as with the true HIVb substitution model. However, based on these results, we expect false positives with real data, representing a substantial fraction of detections (~40% on average in Tab 1 with JTT model). True positives tend to be mutations with the most EEMs, low substitution rate between amino acids and at positions with medium evolutionary rate. On the opposite, fast evolving positions tend to be detected as convergent, even if they are not.

Real HIV dataset

This dataset consists in a MSA of truncated polymerase from HIV-1 subtype B. It was retrieved from the paper by Lemoine *et al* (2018). After removal of recombinant sequences, it contains 3,546 sequences of 1,043 nucleotide positions that were translated into 347 amino acids. Among these 347 amino acids, 250 are on the reverse transcriptase and are analyzed here. Slightly more than 20% of the sequences have at least one known DRM (https://hivdb.stanford.edu/pages/poc.html) and, on average, the DRMs are found in ~11 sequences. The most common one, M184V, is found in 273 sequences. There are 37 DRMs present in at least 12 sequences, corresponding to those used to generate the synthetic datasets. They are distributed on 27 positions. We focused on these 37 DRMs to assess the performance of our approach, but, as already explained, we expect to detect other mutations, some being truly convergent but unknown, and some corresponding to false positives, likely located on fast positions, due to model misspecification.

The evolutionary model selected by ModelFinder (Kalyaanamoorthy et al. 2017) on this dataset is HIVb, with 'freerates' rates-across-site model and 9 rate categories. We tested 255 mutations in total: those present in at least 12 sequences and showing more than 2 EEMS. Among these, we detected 74 convergent events, after applying the Benjamini-Hochberg correction (Benjamini and Hochberg 1995) for multiple testing (non-corrected p-value threshold of 4e-4, corresponding to a corrected alpha level of 5%). Among these detections, 20 are DRMs, which represents 54% of true positives and is a higher proportion than what is expected by chance (Fisher's

exact test p-value = 4.8e-4). The non-DRM detected events correspond to 11 mutations on fast evolving positions (Fig. 4a), which are likely false positives, plus other events for which we cannot conclude using simple arguments (but see below). Regarding false negatives, 17 DRMs are not detected, 7 of which having a (non-corrected) p-value lower than 0.005, but higher than the significance threshold (4e-4). Interestingly, 4 of them are between amino acids with a low substitution rate, as shown in Figure 4b. Based on the substitution rate and small p-values, it could be possible to identify some false negatives, in particular Y188L with amino acids that are unlikely to substitute (BLOSUM62 score = -1). Positive selection analysis supports this approach (see below).

If we focus on positions instead of mutations, we detected 65 positions as convergent among which 19 (~70%) are positions with DRMs. To complement and comfort our findings, we analyzed the 109 positions presenting mutations in at least 12 sequences and more than 2 EEMs, to check for a signal of positive selection. In total, 32 positions were found to be under episodic positive selection using the mixed effects model of evolution (MEME ; Murrell et al. 2012), among which 26 intersected with our detections as presented Table 2 (exhaustive list of detections of MEME is given in Supplementary Table S3). There is thus a strong correlation between the two approaches as almost all the positions under positive selection harbor mutations found as convergent with ConDor. Positive selection identified 11 positions with DRMs, 9 of which were also found with ConDor.

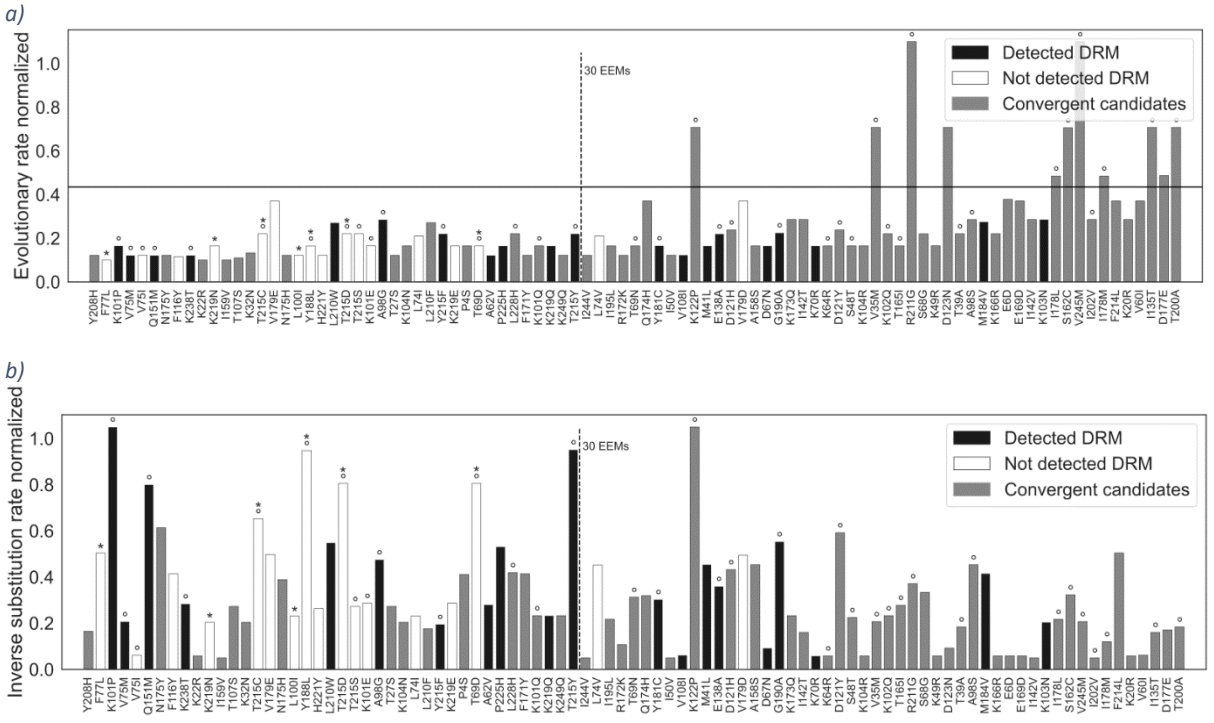


Figure 4: DRMs detection and convergent candidates with real HIV data. *We display DRMs, detected (black) or not (white) and convergent candidates (grey), as obtained using our method on the real HIV-1 subtype B MSA. Mutations are sorted by their number of EEMs on the x-axis. We report on the y-axis of Figure (a) the normalized evolutionary rate of each mutation position; the plain horizontal line represents the threshold of the 5% fastest positions on the whole dataset. We report on the y-axis of Figure (b) the inverse of the HIVb substitution rate (the higher the bar, the less likely a substitution between the two amino acids is). The stars on top of bars represent undetected DRMs with an uncorrected p-value lower than 0.005 and the circles stand for positions under positive selection.*

However, MEME did not detect positions with major DRMs (e.g., K103N, M184V, D67N and M41L, even after increasing the p-value threshold to 0.10) and seems to lack sensitivity for DRM detection. The other positions with positive selection that intersected with ConDor detections were 8 fast positions and 9 other non-DRM related positions. Among non-fast positions, only a small fraction (9/36) of the non-DRM events detected with ConDor were thus found at positions with positive selection. However, while reviewing the proposed convergent mutations found by ConDor, we could identify 20 mutations on 18 (among 36) positions previously described in the literature as potential accessory mutations, correlated with virological failure (see Supplementary Table S2). Moreover, as we previously noticed with synthetic datasets, our method tends to detect events at fast positions. The same holds with positive selection, thus leading us to consider the 5% of the fastest positions (calculated on the whole dataset) as false positives. This is comforted by several studies indicating that the main polymorphisms in HIV-1 are found on the fast positions (Bao et al. 2014; Mohanakrishnan et al. 2015; Cantão et al. 2018). In total, the number (46) of uncharacterized positions detected by ConDor is reduced by ~60% (10 fast + 18 found in the literature).

	All	DRMPos	Fast	Others
Convergence	65	19	10	36
Positive selection	32	11	10	11
Both	26	9	8	9
Total in the dataset	109	27	13	69

Table 2: Comparison of convergence detection and positive selection with real HIV data. *We display the positions harboring events detected with our method and MEME on the real HIV-1 subtype B MSA of reverse transcriptase. DRMPos: detected positions with DRM(s). Fast: detected positions that belong to the 5% positions with highest evolutionary rate on the whole dataset. Others: other detected positions.*

Fish Rhodopsin

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 selection among species that evolve in different environments (Spady et al. 2005). Depending on the habitat and the amount of light available, different amino acids are observed at the same positions, which result in variations in rhodopsin structure and different maximum wavelength absorption. Certain substitutions corresponding to these amino-acid changes have been described as resulting from convergent evolution. In particular D83N, E122Q, F261Y and A292S (using similar substitution encoding as with HIV) occurred several times independently (Yokoyama 2008).

The dataset we used comes from a study in which the authors characterized substitution F261Y as convergent in fish rhodopsin, as a possible result from a transition from marine to brackish or freshwater environments (Hill et al. 2019). It contains an alignment of 2,047 sequences with 308 amino-acid positions. The sequences have been classified by the authors in two groups: species found

only in marine water and species that can evolve (exclusively or not) in brackish or fresh water. Species annotated within the habitats brackish or fresh water can therefore also be found in marine water.

The neutral model inferred by ModelFinder on this dataset was 'MtZoa' and 'freerates' with 8 rate categories. The reconstructed tree is well supported with three quarters of the bootstrap supports above 70%. We tested 355 substitutions (present in at least 12 sequences and showing more than 2 EEMs) with our method and 55 were retained as possibly convergent (15%). They are distributed over 49 positions out of the 136 tested positions, which means that 36% of the tested positions are detected as convergent (16% considering all alignment positions). We were able to recover substitutions F261Y, D83N and A292S, and reversion N83D was also found as convergent. Substitution E122Q was not found as convergent since glutamine (Q) independently emerged only 3 times according to ACR, but emerged up to 7 times in simulations.

From the ancestral reconstruction of position 261 presented Figure 5, Tyrosine (Y) emerged 20 times independently from the phenylalanine (F), which confirms the observation made in (Hill *et al*, 2019). It is a reversion since Y is found as root amino acid at this position. After being acquired from amino acid F, amino acid Y shifts 3 times back to F without this change being detected as convergent. Hill *et al* observed a strong correlation between amino acid Y at position 261 and the brackish or

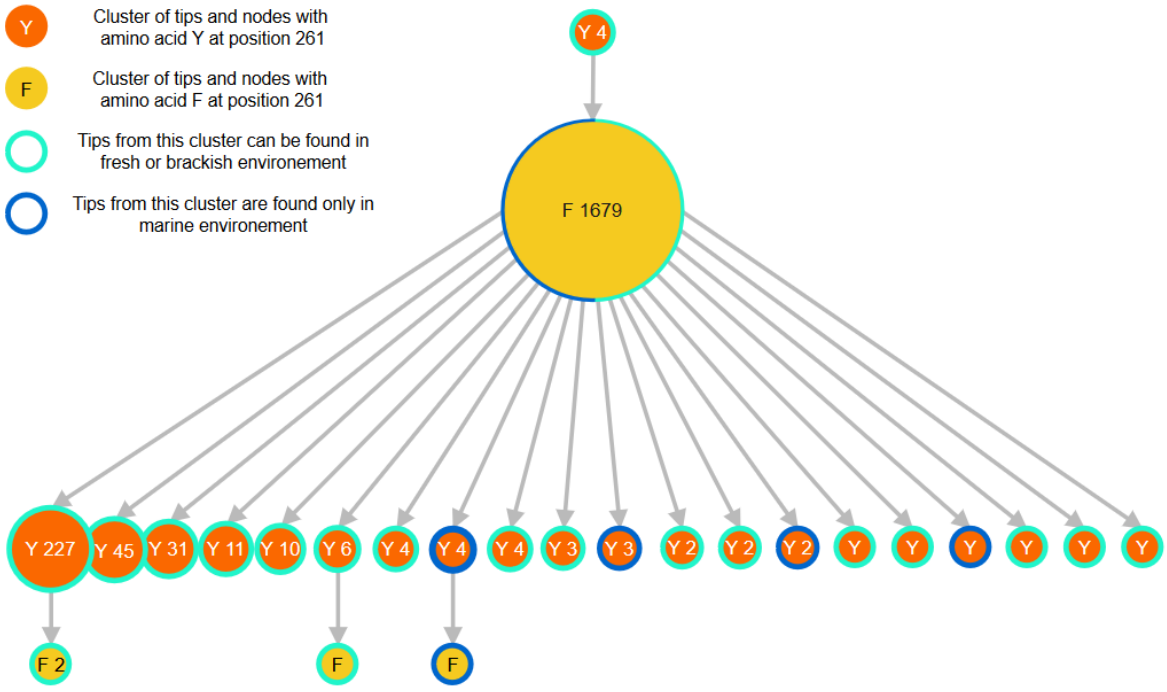


Figure 5: Fish rhodopsin data, ancestral reconstruction of position 261. *This visualization corresponds to a compressed representation of the ancestral scenario, after performing a vertical merge such as defined by PastML. Each disk corresponds to a cluster within which all nodes and tips have a common ancestor (a node, included in the cluster) and are predicted by ACR with the same state. Next to the inferred state (here Y or F), is written the number of tips in the cluster. If there is no number, the represented "cluster" is a tip. Arrows represent independent EEMs. The blue and green circles around each cluster represent the percentage of tips in the cluster annotated with marine and fresh/brackish water, respectively.*

freshwater habitat (Hill et al. 2019). To alleviate phylogenetic confounding factors, we reanalyzed this correlation using BayesTraits 'discrete dependent' model (Pagel and Meade 2006) and found a strong dependence between these two traits (Bayes Factor = 58). Indeed, this substitution between F and Y is known to shift wavelength absorption of the rhodopsin (Yokoyama 2000) and could be advantageous in brackish or fresh water. We also found a significant correlation between habitat and the emergence of amino acid S at position 292, as well as a correlation between habitat and mutations D83N/N83D (see Supplementary Figures S1 and S2).

Among the 55 detected convergent events, 12 are found at fast positions and should be treated carefully (Supplementary Table S4). Interestingly, 9 of the detected events are reversions towards the amino acid reconstructed at the root by ACR, none of which is at fast positions. These 9 reversions most likely are true convergence events. We tested the correlation between the 55 predicted convergent events and the habitat and found a significant dependence for 32 of them (9 on fast positions), meaning that, again, a substantial fraction of these detections most likely corresponds to true convergence. This was confirmed by a literature review (see Supplementary Table S4 for details) showing that ConDor detected several interesting mutations such as T166S responsible for blue-shifting absorption (Luk et al. 2016), S299A found mainly in bathypelagic species (Dungan and Chang 2017), or A124G causing a spectral shift towards red wavelengths (Van Nynatten et al. 2021).

To confirm our results and compare our method, we also ran PCOC (Rey *et al* 2018) on this dataset, considering that the convergent clades were those annotated with the brackish or freshwater environment. PCOC identified 16 positions with a shift in profile compared to the ancestral one ("Profile Change" (PC) model was significant for 16 positions), but without any change within the convergent clades ("One Change" (OC) model not significant). This means that PCOC detected positions correlated with changes in constraints associated with the habitat but no strictly convergent positions, in the sense that the branches where the adaptation took place did not exhibit a substitution as required for the OC model. Several positions intersected with our detections: 123, 166, 198 and 281. None of these positions was part of the 5 percent fastest positions. However, PCOC did not recover positions 83, 122, 261 and 292. As PCOC detects changes in the amino-acid profiles and our method identifies strict changes in the amino acids, we expect the results to be different between both methods. Extensive description of PCOC results can be found in the Supplementary Table S5.

Implementation and availability

Our method is implemented in a workflow, named ConDor, which is accessible through a web service: http://condor.pasteur.cloud. ConDor inputs consist of: (1) a protein alignment in fasta format; (2) a file containing outgroup sequence identifiers; and (3) a newick file with a tree (rooted or not)

whose tips are the same sequences. The workflow is executed on the Institut Pasteur cluster and takes ~160 minutes for a dataset of ~2,050 sequences and ~300 positions. This time corresponds to ~140 minutes for the model inference and ~20 minutes for the convergence detection, using 10,000 simulations per position and target amino acid. The output consists of all the tested mutations with several statistics such as their p-value and if they passed or not the threshold to be considered as convergent mutations. We also provide the evolutionary rate of the corresponding position, the nature of the mutation (convergent, parallel, revertant), the number of EEMs, the genetic barrier, the BLOSUM62 score, etc. All these statistics can be used to analyze further and select the most relevant mutations.

Discussion

In this work, we have developed the ConDor approach, which makes it possible to detect evolutionary convergence at the resolution of a mutation, without prior knowledge of constraints and phenotype. We could retrieve more than 70% of the positions with DRMs on a real HIV dataset, and ~55% of all DRMs present in the sequences. On synthetic datasets mimicking the evolution of HIV, we showed that the detection power depended on the number of emergence-events of mutations and the exchangeability between amino acids. Even though ConDor was primarily developed for the analysis of viral datasets, it could detect several convergent mutations involved in the change in absorption wavelength in a fish rhodopsin dataset. These results confirm that our method detects realistic convergent evolution signal and could be applied to a broad range of organisms.

We tested the robustness of ConDor to model violation by using JTT (Jones et al. 1992) instead of HIVb (Nickle et al. 2007) as a neutral model of evolution for the study of synthetic HIV-like datasets. In doing so, the sensitivity remained high, and we still detected ~70% of positions with DRMs. However, the number of false positives increased, and thus we expect false positives with real data. The number of false positives can be partly explained as ConDor tends to be biased towards mutations found at positions with very high evolutionary rates. Indeed, our method relies on how realistic simulations are as a null model. Our results show that for most positions and most mutations, we are close to what is observed in real data, and our simulations represent a satisfactory null model. However, at certain fast positions we observed that simulations tend to differ from the real data, which resulted in an increased rate of false positives. Thus, we advise that ConDor detections should be cross-validated with approaches such as positive selection to decrease the number of potential false positives. More advanced substitution models, for example based on mixtures or some ideas derived from the CAT model (Le, Gascuel et al. 2008) also used in (Rey et al. 2018) but in a different setting, or mixture of matrix models which accounts for structural features of the positions (Le,

Lartillot et al. 2008) or their evolutionary rate (Le et al. 2012), should likely enhance our approach, make the simulations more realistic and lower the number of potentially erroneous detections.

ConDor was developed to detect convergent amino-acid changes and not convergent positions, which complicates the comparison with existing approaches based on convergent site detection (e.g. PCOC (Rey et al. 2018), and to some extent positive selection methods (Murrell et al. 2012)). 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 since we are working at the scale of a single protein with thousands of sequences, which provides sufficient signal and detection power. By working on thousands or even millions of positions (e.g., with bacterial genomes), ConDor would likely lack 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)), or to detect convergence within a sliding window, would certainly be a useful development.

We have designed this method for the study of specific genes, typically from viruses and microorganisms for which the phenotype is rarely available. Thus, we do not consider any prior knowledge of constraints and phenotype for our analyses. In the same way that methods consider that a position is convergent if it shows the same amino acid derived independently in species with a convergent phenotype (Foote et al. 2015; Zou and Zhang 2015), knowledge of the phenotype could be added to ConDor and only positions meeting the above criterion could be tested and selected.

Materials and Methods

Real HIV dataset

The HIV reverse transcriptase dataset we analyzed is based on the nucleotide alignment provided in (Lemoine et al. 2018) which we downloaded from https://github.com/evolbioinfo-/booster-workflows/tree/master/data/vih. This is an alignment of 9,147 HIV-1 group M polymerase sequences. The authors indicate that this alignment contains recombinant sequences, which we have removed based on the JPHMM output file provided in their supplementary data (https://github.com/evolbioinfo/booster-workflows). We then extracted the B subtype sequences from this alignment using their annotation file "pol_nonrecombinant.txt". Finally, we added the reference sequence HXB2 (GenBank™ accession number K03455) from which the position numbering and the ancestral amino acids in the DRMs notation are derived. The resulting alignment contains 3,557 B subtype sequences and 1,043 nucleotide positions.

As outgroup, we downloaded the group M subtype reference alignment (user-defined range 2258-3300) from the Los Alamos HIV database (https://www.hiv.lanl.gov/content/index). We

removed subtype B sequences from the outgroup and added the sequences to Lemoine *et al* MSA, resulting in an MSA of 3,592 nucleotide sequences which we translated into amino acids.

After tree reconstruction and rooting (as explained in subsection "Tree reconstruction by maximum likelihood") we visually checked the rooting of the tree with Itol (Letunic and Bork 2019). We removed from the MSA 11 sequences classified as B (but likely recombinant), which were placed among the outgroup and obtained a monophyletic group of 3,546 B subtype sequences and the corresponding rooted tree.

Synthetic HIV-like alignment

We used the sierrapy (https://github.com/hivdb/sierra-client/tree/master/python) fasta command to locate (positions and sequences) the DRMs present in the HIV-1 subtype B MSA of 3,557 sequences. The list of mutations used by sierrapy can be found at "https://hivdb.stanford.edu/hiv-db/by-mutations". Then, we replaced positions with DRMs with unknowns in the corresponding sequences. The phylogeny, thus unaffected by DRMs, was reconstructed as explained in subsection "Tree reconstruction by maximum likelihood". This phylogeny was therefore different from the one representing the real data because the aim here was not the same. In the first case, with the real data, we did not remove the DRMs because we wanted to be agnostic of any convergence before applying ConDor. In the second case, we wanted to control where the convergence was in order to quantify the percentage of true and false positives. To do this, we removed all traces of (known) convergence, thus removing the DRMs and reconstructing a phylogeny whose structure was affected as little as possible by convergence.

We rooted the tree using the same outgroup as before and we deleted 6 sequences of subtype B that were placed among the outgroup sequences. As result, we obtained a true HIV phylogeny of 3551 truncated polymerases of subtype B. We simulated the evolution of HXB2 (reverse transcriptase only) along this tree 5 times using a homemade simulator implemented in python.

We did not replace exactly the DRMs found by sierrapy, but used the same list as the one used for the real data, for comparison purposes. This list corresponds to the "Essential DRM Data" data from the Stanford University Drug resistance database (https://hivdb.stanford.edu/pages/poc.html), from which we selected only mutations found in at least 12 sequences and showing more than 2 EEMs, i.e., found in at least 3 distinct subtrees. The list of 37 DRMs can be found in Supplementary Table S1.

Rhodopsin dataset

Protein data of rhodopsin and fish habitat were retrieved from https://github.com/Clupeaharengus/rhodopsin/tree/master/phylogeny_habitat. We extracted the 2,056 sequences from

"spp_to_keep.txt" from the file "final_alignment.translated.fullrhodopsin.fasta" and removed 7 badly aligned sequences. We used the same sequences used for rooting than in (Hill et al. 2019) (*Huso huso* and *Polyodon spathula*). The habitat was provided in the file "rabo_allele_hab.tsv" from the repository provided in (Hill et al. 2019).

Tree reconstruction by maximum likelihood

All phylogenies were reconstructed from the corresponding protein MSAs, using the following procedure and options. We used model finder (Kalyaanamoorthy et al. 2017), corresponding to parameter `-m MFP` in IQtree version 1.6.8 (Nguyen et al. 2015), to select the model of sequence evolution (substitution matrix, gamma categories or freerates model, presence of invariant positions, etc.). Amino acid equilibrium frequencies were estimated by maximum likelihood using IQ-tree option `+FO`, and site-specific evolutionary rates were estimated using option `-wsr`.

Ancestral character construction by maximum likelihood

Ancestral character reconstruction was achieved using PastML version 1.9.29.9 (Ishikawa et al. 2019) with option `--prediction_metho MAP`. We provided one parameter file (option `--parameter`) per position, in which are written (1) the amino acid frequencies for the whole alignment and (2) the scaling factor for the studied position, corresponding to the rate of evolution of the site as estimated by IQtree. The selected substitution matrix (HIVb, JTT, resp. MetaZoa) was given as input (`--rate_matrix`) using PastML option -m `CUSTOM_RATE`.

Technical details

The whole method is implemented in a Nextflow pipeline (Tommaso et al. 2017) taking as input an amino-acid alignment, a rooted/unrooted tree and a file containing outgroup sequences identifiers. 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 statistic tools they provide. We used `biopython` (Cock et al. 2009) for sequences and alignments manipulations. Tree traversals and analyses were achieved with `ETE 3` (Huerta-Cepas et al. 2016). Graphics were obtained using `matplotlib` (Hunter 2007) and `seaborn` libraries. All MSA (translation to amino acids, subalignments, etc.) and trees manipulations (pruning, rooting, etc.) were achieved using `goalign` and `gotree` (Lemoine and Gascuel 2021). Simulations and counting of EEMs were computed using homemade python scripts. Convergent candidate events were selected based on their p-value after correcting for multiple testing with a "Benjamini – Hochberg" correction, with a risk alpha of 0.05.

MEME

We used MEME (Murrell et al. 2012) to search for positions under episodic positive selection on the nucleotidic HIV-1 MSA. The threshold for significant positions was set at a p-value of 0.05, without correction for multiple testing, as their test tends to be conservative and "traditional setting of multiple testing (multiple tests on the same data) does not directly apply here" (https://github.com/veg/hyphy/issues/851 and https://github.com/veg/hyphy/issues/188). Thus, the proportion of false positives we expect is approximately equal to 0.05 multiplied by the number of positions that are under positive selection (here 0.05x32≈2).

PCOC

We used PCOC (Rey et al. 2018) to detect convergent positions based on the knowledge of the habitat (marine vs fresh/brackish water). We used the profile C10 (`-CATX_est 10`) with 4 gamma categories (`--gamma`) and fixed the posterior probability threshold of a position to be above 0.8 (`-f 0.8`). For the convergent scenario (`-m`) corresponding to the different clades of nodes which exhibit the convergent transition, we considered that the convergent tips were those with the fresh/ brackish water environment. Then, we retrieved all internal nodes the tips of which were in the convergent environment and completed the scenario, as described in the user guide (https://github.com/CarineRey/pcoc).

BayesTraits

In the rhodopsin dataset, correlations between fish habitat and mutations detected as convergent by ConDor were measured with BayesTraits 'discrete dependent' model (Pagel 1994; Pagel & Meade 2006). Prior to running the software, we transformed our data into discrete binary traits. This way, marine habitat was annotated as 1 and fresh/brackish water as 0. Similarly, for a given position, the convergent change had the value 1 and the other amino acids at that position the value 0. We followed the procedure detailed in http://www.evolution.rdg.ac.uk/BayesTraitsV3/Files/-BayesTraitsV3.Manual.pdf to assess whether the dependence between both traits was more likely than their independence. The dependence hypothesis was retained if the Bayes factor was greater than 10. Priors for the transition rates were estimated using the output provided by the maximum likelihood models as described in the user guide.

Data availability

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

Acknowledgments

We sincerely thank Anna Zhukova, Luc Blassel and Jakub Voznica for their help and suggestions. This work was supported by INCEPTION program (Convention ANR-16-CONV-0005; MM PhD grant) and by PRAIRIE program (Convention ANR-19-P3IA-0001; OG).

References

Bao Y, Tian D, Zheng Y-Y, Xi H-L, Liu D, Yu M, Xu X-Y. 2014. Characteristics of HIV-1 Natural Drug Resistance-Associated Mutations in Former Paid Blood Donors in Henan Province, China. PLOS ONE. 9(2):e89291. doi:10.1371/journal.pone.0089291.

Benjamini Y, Hochberg Y. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J R Stat Soc Ser B Methodol. 57(1):289–300. doi:https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.

Cantão NM, Fogaça de Almeida L, Rodrigo Wolf I, Oliveira Almeida R, Alves de Almeida Cruz A, Nunes C, Barbosa AN, Valente GT, de Moura Campos Pardini MI, Grotto RMT. 2018. HIV Reverse Transcriptase and Protease Genes Variability Can Be a Biomarker Associated with HIV and Hepatitis B or C Coinfection. Sci Rep. 8(1):8280. doi:10.1038/s41598-018-26675-z.

Chabrol O, Royer-Carenzi M, Pontarotti P, Didier G. 2018. Detecting the molecular basis of phenotypic convergence. Methods Ecol Evol. 9(11):2170–2180. doi:https://doi.org/10.1111/2041-210X.13071.

Chai S, Tian R, Rong X, Li G, Chen B, Ren W, Xu S, Yang G. 2020. Evidence of Echolocation in the Common Shrew from Molecular Convergence with Other Echolocating Mammals. Zool Stud. 59:e4. doi:10.6620/ZS.2020.59-4.

Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, et al. 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics. 25(11):1422–1423. doi:10.1093/bioinformatics/btp163.

Crandall KA, Kelsey CR, Imamichi H, Lane HC, Salzman NP. 1999. Parallel evolution of drug resistance in HIV: failure of nonsynonymous/synonymous substitution rate ratio to detect selection. Mol Biol Evol. 16(3):372–382. doi:10.1093/oxfordjournals.molbev.a026118.

Cuevas JM, Elena SF, Moya A. 2002. Molecular Basis of Adaptive Convergence in Experimental Populations of RNA Viruses. Genetics. 162(2):533–542.

Davies KTJ, Cotton JA, Kirwan JD, Teeling EC, Rossiter SJ. 2012. Parallel signatures of sequence evolution among hearing genes in echolocating mammals: an emerging model of genetic convergence. Heredity. 108(5):480–489. doi:10.1038/hdy.2011.119.

van Ditmarsch D, Boyle KE, Sakhtah H, Oyler JE, Nadell CD, Déziel É, Dietrich LEP, Xavier JB. 2013. Convergent evolution of hyperswarming leads to impaired biofilm formation in pathogenic bacteria. Cell Rep. 4(4):697–708. doi: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. 2020. Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. Infect Genet Evol.:104351. doi:10.1016/j.meegid.2020.104351.

Dungan SZ, Chang BSW. 2017. Epistatic interactions influence terrestrial–marine functional shifts in cetacean rhodopsin. Proc R Soc B Biol Sci. 284(1850):20162743. doi:10.1098/rspb.2016.2743.

Foll M, Gaggiotti OE, Daub JT, Vatsiou A, Excoffier L. 2014. Widespread Signals of Convergent Adaptation to High Altitude in Asia and America. Am J Hum Genet. 95(4):394–407. doi:10.1016/j.ajhg.2014.09.002.

Foote AD, Liu Y, Thomas GWC, Vinař T, Alföldi J, Deng J, Dugan S, van Elk CE, Hunter ME, Joshi V, et al. 2015. Convergent evolution of the genomes of marine mammals. Nat Genet. 47(3):272–275. doi:10.1038/ng.3198.

Goldman N, Yang Z. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol. 11(5):725–736. doi:10.1093/oxfordjournals.molbev.a040153.

Gutierrez B, Escalera-Zamudio M, Pybus OG. 2019. Parallel molecular evolution and adaptation in viruses. Curr Opin Virol. 34:90–96. doi: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. 2020. Array programming with NumPy. Nature. 585(7825):357–362. doi: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. 2019. Recurrent convergent evolution at amino acid residue 261 in fish rhodopsin. Proc Natl Acad Sci. 116(37):18473–18478. doi:10.1073/pnas.1908332116.

Holmes EC, Zhang LQ, Simmonds P, Ludlam CA, Brown AJ. 1992. 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. 89(11):4835–4839.

Hu Y, Wu Q, Ma S, Ma T, Shan L, Wang X, Nie Y, Ning Z, Yan L, Xiu Y, et al. 2017. Comparative genomics reveals convergent evolution between the bamboo-eating giant and red pandas. Proc Natl Acad Sci. 114(5):1081–1086. doi:10.1073/pnas.1613870114.

Huerta-Cepas J, Serra F, Bork P. 2016. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. Mol Biol Evol. 33(6):1635–1638. doi:10.1093/molbev/msw046.

Hunter JD. 2007. Matplotlib: A 2D Graphics Environment. Comput Sci Eng. 9(3):90–95. doi:10.1109/MCSE.2007.55.

Ishikawa SA, Zhukova A, Iwasaki W, Gascuel O. 2019. A Fast Likelihood Method to Reconstruct and Visualize Ancestral Scenarios. Mol Biol Evol. doi:10.1093/molbev/msz131. [accessed 2019 Jun 7]. https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msz131/5498561.

Jones DT, Taylor WR, Thornton JM. 1992. The rapid generation of mutation data matrices from protein sequences. Bioinformatics. 8(3):275–282. doi:10.1093/bioinformatics/8.3.275.

Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat Methods. 14(6):587–589. doi:10.1038/nmeth.4285.

Korber B, Fischer WM, Gnanakaran S, Yoon H, Theiler J, Abfalterer W, Hengartner N, Giorgi EE, Bhattacharya T, Foley B, et al. 2020. Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus. Cell. 182(4):812-827.e19. doi:10.1016/j.cell.2020.06.043.

Le SQ, Dang CC, Gascuel O. 2012. Modeling Protein Evolution with Several Amino Acid Replacement Matrices Depending on Site Rates. Mol Biol Evol. 29(10):2921–2936. doi:10.1093/molbev/mss112.

Le SQ, Gascuel O, Lartillot N. 2008. Empirical profile mixture models for phylogenetic reconstruction. Bioinformatics. 24(20):2317–2323.

Le SQ, Lartillot N, Gascuel O. 2008. Phylogenetic mixture models for proteins. Philos Trans R Soc B Biol Sci. 363(1512):3965–3976. doi:10.1098/rstb.2008.0180.

Lee J-H, Lewis KM, Moural TW, Kirilenko B, Borgonovo B, Prange G, Koessl M, Huggenberger S, Kang C, Hiller M. 2018 Jan 8. Building superfast muscles: insights from molecular parallelism in fast-twitch muscle proteins in echolocating mammals. bioRxiv.:244566. doi:10.1101/244566.

Lemoine F, Entfellner J-BD, Wilkinson E, Correia D, Felipe MD, Oliveira TD, Gascuel O. 2018. Renewing Felsenstein's phylogenetic bootstrap in the era of big data. Nature. 556(7702):452–456. doi:10.1038/s41586-018-0043-0.

Lemoine F, Gascuel O. 2021 Jun 10. Gotree/Goalign : Toolkit and Go API to facilitate the development of phylogenetic workflows. bioRxiv.:2021.06.09.447704. doi:10.1101/2021.06.09.447704.

Letunic I, Bork P. 2019. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. Nucleic Acids Res. 47(W1):W256–W259. doi:10.1093/nar/gkz239.

Longdon B, Day JP, Alves JM, Smith SCL, Houslay TM, McGonigle JE, Tagliaferri L, Jiggins FM. 2018. Host shifts result in parallel genetic changes when viruses evolve in closely related species. PLOS Pathog. 14(4):e1006951. doi:10.1371/journal.ppat.1006951.

Lu B, Jin H, Fu J. 2020. Molecular convergent and parallel evolution among four high-elevation anuran species from the Tibetan region. BMC Genomics. 21(1):839. doi:10.1186/s12864-020-07269-4.

Luk HL, Bhattacharyya N, Montisci F, Morrow JM, Melaccio F, Wada A, Sheves M, Fanelli F, Chang BSW, Olivucci M. 2016. Modulation of thermal noise and spectral sensitivity in Lake Baikal cottoid fish rhodopsins. Sci Rep. 6(1):38425. doi:10.1038/srep38425.

Marcovitz A, Turakhia Y, Chen HI, Gloudemans M, Braun BA, Wang H, Bejerano G. 2019. A functional enrichment test for molecular convergent evolution finds a clear protein-coding signal in echolocating bats and whales. Proc Natl Acad Sci U S A. 116(42):21094–21103. doi: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. 2021. The emergence and ongoing convergent evolution of the N501Y lineages coincides with a major global shift in the SARS-CoV-2 selective landscape. Infectious Diseases (except HIV/AIDS). [accessed 2021 Apr 1]. http://medrxiv.org/lookup/doi/10.1101/2021.02.23.21252268.

McKinney W. 2010. Data Structures for Statistical Computing in Python. Austin, Texas. p. 56–61. [accessed 2021 Jan 12]. https://conference.scipy.org/proceedings/scipy2010/mckinney.html.

Mohanakrishnan K, Kasthuri A, Amsavathani SK, Sumathi G. 2015. HIV reverse transcriptase gene mutations in anti-retroviral treatment naïve rural people living with HIV/AIDS. Indian J Med Microbiol. 33(4):565–567. doi:10.4103/0255-0857.167326.

Murrell B, Wertheim JO, Moola S, Weighill T, Scheffler K, Pond SLK. 2012. Detecting Individual Sites Subject to Episodic Diversifying Selection. PLOS Genet. 8(7):e1002764. doi:10.1371/journal.pgen.1002764.

Muschick M, Indermaur A, Salzburger W. 2012. Convergent Evolution within an Adaptive Radiation of Cichlid Fishes. Curr Biol. 22(24):2362–2368. doi:10.1016/j.cub.2012.10.048.

Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. Mol Biol Evol. 32(1):268–274. doi:10.1093/molbev/msu300.

Nickle DC, Heath L, Jensen MA, Gilbert PB, Mullins JI, Pond SLK. 2007. HIV-Specific Probabilistic Models of Protein Evolution. PLOS ONE. 2(6):e503. doi:10.1371/journal.pone.0000503.

Pagel M. 1994. Detecting correlated evolution on phylogenies: a general method for the comparative analysis of discrete characters. Proc R Soc Lond B Biol Sci. 255(1342):37–45. doi:10.1098/rspb.1994.0006.

Pagel M, Meade A. 2006. Bayesian Analysis of Correlated Evolution of Discrete Characters by Reversible-Jump Markov Chain Monte Carlo. Am Nat. 167(6):808–825. doi:10.1086/503444.

Parker J, Tsagkogeorga G, Cotton JA, Liu Y, Provero P, Stupka E, Rossiter SJ. 2013. Genome-wide signatures of convergent evolution in echolocating mammals. Nature. 502(7470):228–231. doi:10.1038/nature12511.

Parto S, Lartillot N. 2018. Molecular adaptation in Rubisco: Discriminating between convergent evolution and positive selection using mechanistic and classical codon models. PLOS ONE. 13(2):e0192697. doi:10.1371/journal.pone.0192697.

Pond SLK, Murrell B, Poon AFY. 2012. Evolution of Viral Genomes: Interplay Between Selection, Recombination, and Other Forces. Evol Genomics.:239–272. doi:10.1007/978-1-61779-585-5_10.

Rey C, Guéguen L, Sémon M, Boussau B. 2018. Accurate Detection of Convergent Amino-Acid Evolution with PCOC. Mol Biol Evol. 35(9):2296–2306. doi:10.1093/molbev/msy114.

Rosenblum EB, Parent CE, Brandt EE. 2014. The Molecular Basis of Phenotypic Convergence. Annu Rev Ecol Evol Syst. 45(1):203–226. doi:10.1146/annurev-ecolsys-120213-091851.

Spady TC, Seehausen O, Loew ER, Jordan RC, Kocher TD, Carleton KL. 2005. Adaptive Molecular Evolution in the Opsin Genes of Rapidly Speciating Cichlid Species. Mol Biol Evol. 22(6):1412–1422. doi:10.1093/molbev/msi137.

Stern DL. 2013. The genetic causes of convergent evolution. Nat Rev Genet. 14(11):751–764. doi:10.1038/nrg3483.

Storz JF. 2016. Causes of molecular convergence and parallelism in protein evolution. Nat Rev Genet. 17(4):239–250. doi:10.1038/nrg.2016.11.

Tamuri AU, Reis M dos, Hay AJ, Goldstein RA. 2009. Identifying Changes in Selective Constraints: Host Shifts in Influenza. PLOS Comput Biol. 5(11):e1000564. doi:10.1371/journal.pcbi.1000564.

Thomas GWC, Hahn MW. 2015. Determining the Null Model for Detecting Adaptive Convergence from Genomic Data: A Case Study using Echolocating Mammals. Mol Biol Evol. 32(5):1232–1236. doi:10.1093/molbev/msv013.

Tommaso PD, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. 2017 Apr 11. Nextflow enables reproducible computational workflows. Nat Biotechnol. doi:10.1038/nbt.3820. [accessed 2018 Sep 6]. https://www.nature.com/articles/nbt.3820.

Tsetsarkin KA, Vanlandingham DL, McGee CE, Higgs S. 2007. A Single Mutation in Chikungunya Virus Affects Vector Specificity and Epidemic Potential. PLOS Pathog. 3(12):e201. doi: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. 2015. Widespread convergence in toxin resistance by predictable molecular evolution. Proc Natl Acad Sci. 112(38):11911–11916. doi:10.1073/pnas.1511706112.

Van Nynatten A, Castiglione GM, de A. Gutierrez E, Lovejoy NR, Chang BSW. 2021. Recreated Ancestral Opsin Associated with Marine to Freshwater Croaker Invasion Reveals Kinetic and Spectral Adaptation. Mol Biol Evol. 38(5):2076–2087. doi:10.1093/molbev/msab008.

Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J, et al. 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nat Methods. 17:261–272. doi:10.1038/s41592-019-0686-2.

Xu S, Wang J, Guo Z, He Z, Shi S. 2020. Genomic Convergence in the Adaptation to Extreme Environments. Plant Commun. 1(6):100117. doi:10.1016/j.xplc.2020.100117.

Yokoyama S. 2000. Molecular evolution of vertebrate visual pigments. Prog Retin Eye Res. 19(4):385–419. doi:10.1016/S1350-9462(00)00002-1.

Yokoyama S. 2008. Evolution of Dim-Light and Color Vision Pigments. Annu Rev Genomics Hum Genet. 9(1):259–282. doi:10.1146/annurev.genom.9.081307.164228.

Zhang J. 2006. Parallel adaptive origins of digestive RNases in Asian and African leaf monkeys. Nat Genet. 38(7):819–823. doi:10.1038/ng1812.

Zhang J, Kumar S. 1997. Detection of convergent and parallel evolution at the amino acid sequence level. Mol Biol Evol. 14(5):527–536.

Zhen Y, Aardema ML, Medina EM, Schumer M, Andolfatto P. 2012. Parallel Molecular Evolution in an Herbivore Community. Science. 337(6102):1634–1637. doi:10.1126/science.1226630.

Zou Zhengting, Zhang J. 2015. Are Convergent and Parallel Amino Acid Substitutions in Protein Evolution More Prevalent Than Neutral Expectations? Mol Biol Evol. 32(8):2085–2096. doi:10.1093/molbev/msv091.

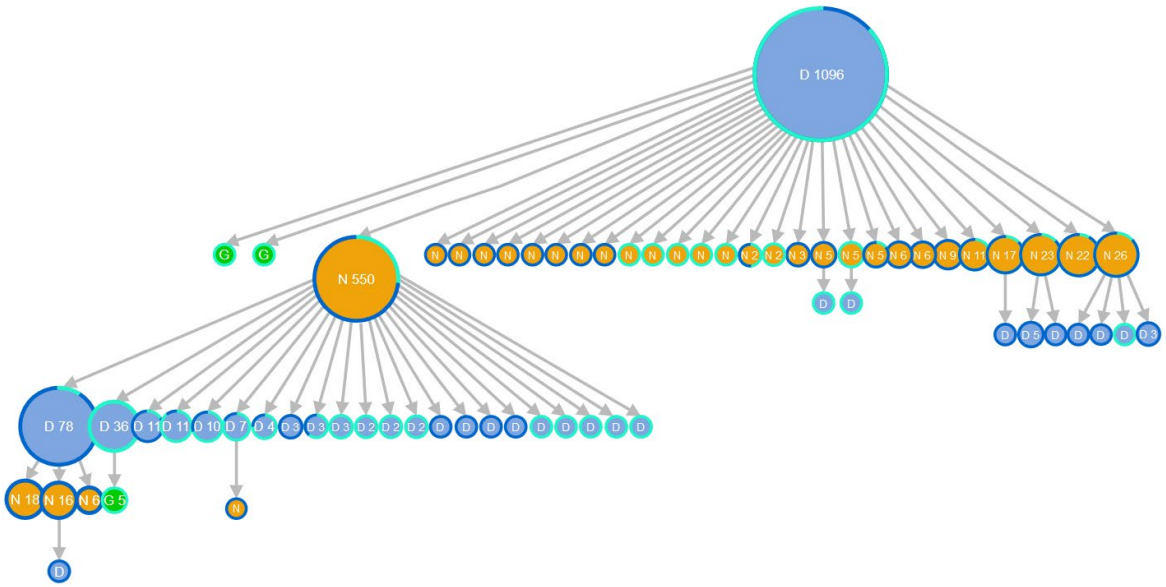
Zou Z., Zhang J. 2015. No Genome-Wide Protein Sequence Convergence for Echolocation. Mol Biol Evol. 32(5):1237–1241. doi:10.1093/molbev/msv014.

Sensitive Detection of Site-wise Convergent Evolution in Large Protein Alignments with ConDor

Marie MOREL, Frédéric LEMOINE and Olivier GASCUEL

Corresponding Authors: marie.morel@pasteur.fr, olivier.gascuel@mnhn.fr

Supplementary Material



	Brackish/fresh	Marine
D	**905**	258
N	99	**442**
G	7	**0**

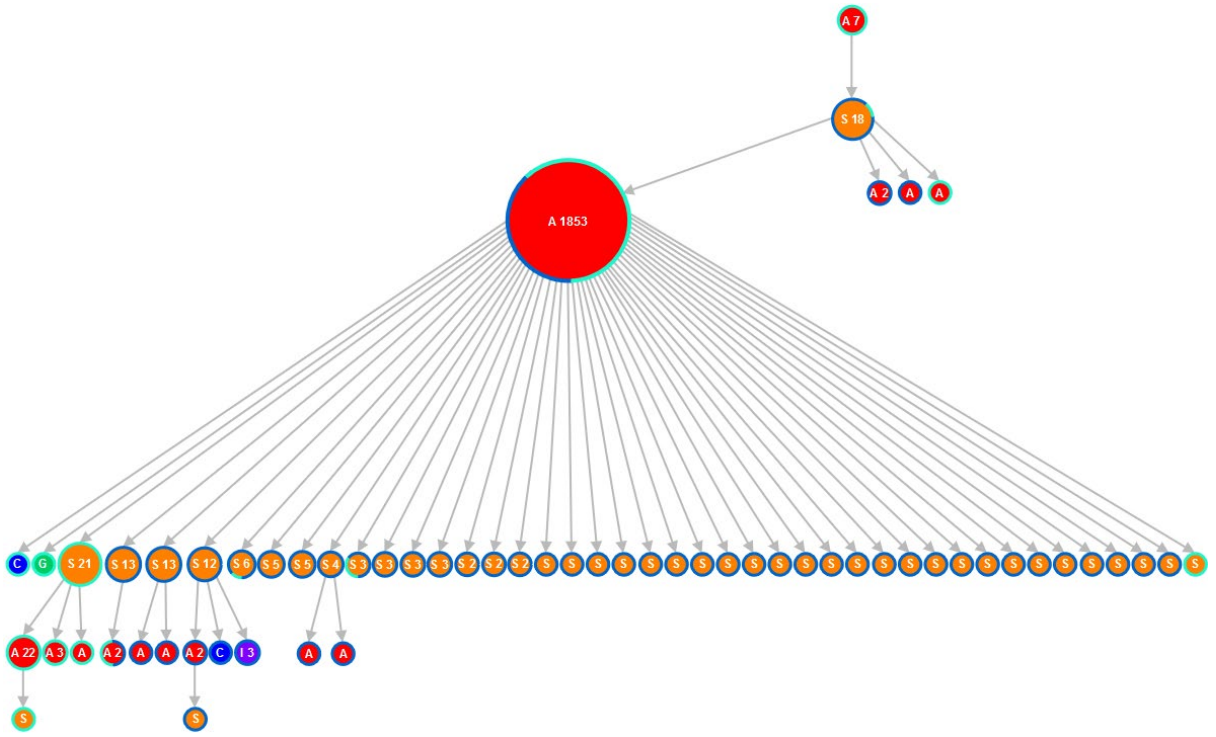
Figure S1: Fish rhodopsin data, ancestral reconstruction of position 83 and corresponding contingency table.

The visualisation corresponds to a compressed representation of the ancestral scenario, after performing a vertical merge such as defined by PastML (Ishikawa et al. 2019). Each disk corresponds to a cluster within which all nodes and tips have the same common ancestor (a node, included in the cluster) and are predicted by ACR with the same state. Next to the inferred state (here N, D or G), is written the number of tips in the cluster. If there is no number, the represented "cluster" is a tip. Arrows represent independent emergence-events of mutations (EEMs). The blue and green circles around each cluster represent the percentage of tips in the cluster annotated in marine and freshwater/brackish water, respectively.

The ancestral amino acid at position 83 is an aspartate (D) which was lost independently 28 times towards asparagine (N) and twice towards glycine (G). The reversion from N towards D then occurred independently 31 times. In some clusters this switch between D and N occurred again leading to 32 EEMs towards N and 32 EEMs towards D in total. Both mutations were found as convergent with ConDor.

D clusters seem to be more frequently associated with brackish/fresh water whereas species in N clusters seem to be found in marine water. This distribution can also be observed with the contingency table displaying the number of species having D or N at position 83 in function of their habitat. To alleviate phylogenetic confounding factors, we reanalysed this correlation using BayesTraits 'discrete dependent' model (Pagel and Meade 2006) and found a strong dependence between these two traits, with Bayes Factors equal to 104 (D83N) and 140 (N83D). The Bayes Factors are different because the data are not perfectly symmetrical between D83N and N83D, mainly due to the presence of other amino acids (G). Indeed, the input given to BayesTraits is binary with the

convergent change taking the value 1 and any other amino acids at the corresponding position the value 0. The correlation found between the habitat and position 83 is confirmed by previous work which found that position 83 harboured key blue-shifting substitutions, with amino acid N found in deep-diving species and D in non-deep-diving species (Sugawara et al. 2005; Yokoyama et al. 2008).



	Brackish/fresh	Marine
A	**1085**	717
S	27	**108**
C	1	1
G	1	0
I	0	3

Figure S2: Fish rhodopsin data, ancestral reconstruction of position 292 and corresponding contingency table.

This visualisation corresponds to a compressed representation of the ancestral scenario, after performing a vertical merge such as defined by PastML (Ishikawa et al. 2019). Each disk corresponds to a cluster within which all nodes and tips have the same common ancestor (a node, included in the cluster) and are predicted by ACR with the same state. Next to the inferred state (here A, S, C, G or I), is written the number of tips in the cluster. If there is no number, the represented cluster is a tip. Arrows represent independent EEMs. The blue and green circles around each cluster represent the percentage of tips in the cluster annotated in marine and freshwater/brackish water, respectively.

The two most common amino acids at position 292 are alanine (A) which is the root amino acid and serine (S). We count 44 EEMs towards S (all from A) and 13 EEMs towards A. Mutation A292S was found convergent with ConDor. This mutation could be associated with marine water as most of the species in the S clusters are also found in marine water. This distribution can also be observed with the contingency table displaying the number of species having A or S at position 292 in function of their habitat. To alleviate phylogenetic confounding factors, we reanalysed this correlation using BayesTraits 'discrete dependent' model (Pagel and Meade 2006) and found a strong dependence between these two traits with Bayes Factors equal to 70. This is consistent with previous works which found that amino acid S at position 292 could be an adaptation to the bathypelagic environment (Sugawara et al. 2005; Varela and Ritchie 2014).

Mutation	Number of sequences	Number of EEMs	Detected
M184V	273	124	Yes
K103N	221	143	Yes
M41L	211	45	Yes
T215Y	187	30	Yes
D67N	164	50	Yes
L210W	130	17	Yes
K70R	123	54	Yes
K219Q	78	29	Yes
G190A	64	50	Yes
T215F	60	22	Yes
V179D	58	47	No
Y181C	55	41	Yes
E138A	49	45	Yes
V108I	47	41	Yes
T69D	47	23	No
P225H	39	25	Yes
L74V	37	32	No
K219E	37	22	No
A62V	30	23	Yes
A98G	27	19	Yes
T215D	25	18	No
Q151M	22	9	Yes
L100I	22	17	No
K238T	22	11	Yes
T215C	21	16	No
T215S	21	18	No
F116Y	20	11	No
L74I	20	20	No
V75I	18	9	No
K101E	18	18	No
Y188L	18	17	No
H221Y	18	17	No
V179E	18	16	No
F77L	15	7	No
K219N	15	13	No
V75M	14	9	Yes
K101P	12	8	Yes

Table S1: List of the 37 DRMs and their detection status on the real HIV-1 subtype B MSA.

This table presents the 37 DRMs that we used as true convergent events for the real HIV-1 subtype B MSA. The chosen DRMs correspond to the "Essential DRM Data" data from the Stanford University Drug resistance database (https://hivdb.stanford.edu/pages/poc.html) from which we selected only mutations found in at least 12 sequences and more than 2 EEMs, i.e., found in two distinct clades in the real HIV-1 tree. Number of sequences: number of sequences in which the corresponding DRM was

found in the real data. Number of EEMs: number of independent emergence-events of mutation that was inferred for the DRM by ancestral reconstruction and counting. Detected: If they were detected or not as convergent by ConDor and thus found as true positives.

Mutation	P-value	Fast	Role of the mutation	Reference
P4S	0.0002	No	X	
E6D	0.0	No	Associated with AIDS related deaths	(Liu et al. 2017)
K20R	0.0	No	Reduced susceptibility	(Saracino et al. 2006)
K22R	0.0	No	X	
T27S	0.0	No	X	
K32N	0.0	No	X	
T35M	0.0	Yes	X	
T39A	0.0	No	Accessory mutation	(Gonzales et al. 2003; Svicher et al. 2006)
M41L	0.0	No	DRM	HIV database[1]
S48T	0.0	No	X	
K49R	0.0	No	X	
I50V	0.0	No	X	
V60I	0.0	No	Compensatory mutation	(Precious et al. 2000)
A62V	0.0	No	DRM	HIV database[1]
K64R	0.0	No	X	
D67N	0.0	No	DRM	HIV database[1]
S68G	0.0	No	Compensatory mutation	(Svarovskaia et al. 2008)
T69N	0.0	No	NRTI-selected mutation	(Winters and Merigan 2001)
K70R	0.0	No	DRM	HIV database[1]
V75M	0.0001	No	DRM	HIV database[1]
A98G	0.0	No	DRM	HIV database[1]
A98S	0.0	No	Increased virological success	(Alteri et al. 2016)
K101Q	0.0	No	Accessory mutation	(Melikian et al. 2014)
K101P	0.0004	No	DRM	HIV database[1]
K102Q	0.0	No	X	
K103N	0.0	No	DRM	HIV database[1]
K104R	0.0	No	X	
K104N	0.0	No	X	
T107S	0.0	No	X	
V108I	0.0	No	DRM	HIV database[1]
D121H	0.0	No	X	
D121Y	0.0	No	X	
K122P	0.0001	Yes	X	
D123N	0.0004	Yes	X	
I135T	0.0	Yes	Associated with NNRTI failure	(Ceccherini-Silberstein, Svicher, et al. 2007)
E138A	0.0001	No	DRM	HIV database[1]
I142T	0.0	No	X	
I142V	0.0001	No	Accessory mutation	(Kawamoto et al. 2008)
Q151M	0.0	No	DRM	HIV database[1]
A158S	0.0	No	X	

I159V	0.0004	No	X	
S162C	0.0	Yes	X	
T165I	0.0	No	Associated with Q151M	(Scherrer et al. 2012)
K166R	0.0	No	X	
E169D	0.0	No	Viral escape	(Zimbwa et al. 2007)
F171Y	0.0	No	X	
R172K	0.0002	No	Suppresses resistance (NRTIs & NNRTIs)	(Hachiya et al. 2012)
K173Q	0.0	No	X	
Q174H	0.0002	No	X	
N175H	0.0	No	X	
N175Y	0.0003	No	X	
D177E	0.0	Yes	X	
I178M	0.0	Yes	X	
I178L	0.0	Yes	Associated with M184V in subtype C	(Doualla-Bell et al. 2004)
Y181C	0.0	No	DRM	HIV database[1]
M184V	0.0001	No	DRM	HIV database[1]
G190A	0.0	No	DRM	HIV database[1]
I195L	0.0	No	X	
T200A	0.0	Yes	X	
I202V	0.0	No	Associated with Q151M	(Scherrer et al. 2012)
Y208H	0.0003	No	Reversion from accessory	(Nebbia et al. 2007)
L210F	0.0	No	X	
L210W	0.0	No	DRM	HIV database[1]
K211G	0.0	Yes	X	
F214L	0.0	No	Favourable virological response?	(Ceccherini-Silberstein, Cozzi-Lepri, et al. 2007)
T215F	0.0	No	DRM	HIV database[1]
T215Y	0.0001	No	DRM	HIV database[1]
K219Q	0.0	No	DRM	HIV database[1]
P225H	0.0	No	DRM	HIV database[1]
L228H	0.0	No	Reduced virological response	(Marcelin et al. 2006)
K238T	0.0001	No	DRM	HIV database[1]
I244V	0.0002	No	X	
K245M	0.0	Yes	X	
K249Q	0.0	No	X	

Note [1]: https://hivdb.stanford.edu/pages/poc.html

Table S2: Results of ConDor detection on real HIV-1 subtype B MSA, sorted by position.

Mutations: Mutations detected as convergent with ConDor on the real HIV-1 subtype B MSA of reverse transcriptase. The p-values are not corrected and below or equal to the acceptance threshold of 0.0004 after "Benjamini-Hochberg" correction. Fast: detected mutation is on a position that belong to the 5% positions with highest evolutionary rate on the whole dataset. Role of the mutation: possible role of the mutation found in the literature. Mutations that have "DRM" noted for the role of the mutation corresponds to the 37 DRMs that we used as true convergent mutations, and which can be found in Supplementary Table S1.

Positions	MEME p-values	ConDor detects	DRM	Fast position
35	0.0000	Yes	No	Yes
39	0.0002	Yes	No	No
40	0.0000	No	No	No
48	0.0000	Yes	No	No
64	0.0002	Yes	No	No
69	0.0000	Yes	Yes	No
75	0.0000	Yes	Yes	No
98	0.0005	Yes	Yes	No
101	0.0116	Yes	Yes	No
102	0.0000	Yes	No	No
111	0.0000	No	No	No
121	0.0037	Yes	No	No
122	0.0001	Yes	No	Yes
135	0.0000	Yes	No	Yes
138	0.0010	Yes	Yes	No
151	0.0000	Yes	Yes	No
162	0.0000	Yes	No	Yes
165	0.0003	Yes	No	No
176	0.0042	No	No	No
178	0.0000	Yes	No	Yes
181	0.0397	Yes	Yes	No
188	0.0000	No	Yes	No
190	0.0023	Yes	Yes	No
200	0.0000	Yes	No	Yes
202	0.0161	Yes	No	No
207	0.0000	No	No	Yes
211	0.0000	Yes	No	Yes
215	0.0000	Yes	Yes	No
228	0.0000	Yes	No	No
238	0.0267	Yes	Yes	No
245	0.0000	Yes	No	Yes
248	0.0000	No	No	Yes
total		26	11	10

Table S3: Results of MEME detection on the 109 positions tested with ConDor on real HIV-1 subtype B MSA, sorted by position.

Positions significantly detected with MEME to be under episodic positive selection (Murrell et al. 2012) on the real HIV-1 subtype B MSA of reverse transcriptase. The p-values are those provided by MEME and unmodified with an acceptance threshold at 0.05. ConDor detects: positions on which we also detected events with ConDor. DRM: position on which there is at least one DRM belonging to the list of 37 DRMs (see Supplementary Table S1). Fast: detected positions that belong to the 5% positions with highest evolutionary rate on the whole dataset.
Over the 32 positions found to be under positive selection, 26 intersect with events found by ConDor. Among all (32) MEME detections, 11 are at positions with DRMs and 10 are at fast positions.

Mutation	ConDor p-values	Fast	Bayes factor	Role of the mutation
V32A	0.0005	No	2,52	X
A33P	0.0	Yes	2,49	X
E64Q	0.0	No	2,57	X
D83N	0.0004	No	**104**	Shift lambda (Sugawara et al. 2005; Yokoyama 2008)
N83D	0.0007	No	**139,63**	Shift lambda (Yokoyama 2008)
L99M	0.0001	No	**54,59**	Parallel in bats, Role in vision? (Shen et al. 2010)
F115Y	0.0001	No	-3,31	X
L119H	0.0	No	9,43	X
I123M	0.0	No	**53,43**	X
A124G	0.0	No	**43,98**	Red-shifting mutation (Hunt et al. 2001; Van Nynatten et al. 2021)
S127C	0.0004	No	**15,47**	X
V137M	0.0	No	-2,45	Altered transducin activation (Athanasiou et al. 2018)
A158G	0.0	Yes	**43,71**	X
T160S	0.0007	No	-2,44	X
A163G	0.0	No	-3,73	X
S165C	0.0	Yes	**64**	X
C165S	0.0001	Yes	**75,31**	X
S166T	0.0007	No	**16**	Reversion from blue-shifting (Malinsky et al. 2015; O'Reilly et al. 2016)
A166S	0.0	No	**66,47**	Blue-shifting mutation (Malinsky et al. 2015; O'Reilly et al. 2016)
A168S	0.0	No	**17,83**	X
V169A	0.0	No	-2,45	X
E196P	0.0001	No	**35,58**	X
F198Y	0.0001	No	**10,52**	Shift lambda with K248R in RH2 (Matsumoto et al. 2020)
S202T	0.0	No	3,20	X
F203Y	0.0001	No	2,13	x
I205V	0.0	No	**66**	X
I209T	0.0	Yes	**14**	X
V210C	0.0005	No	**57,71**	X
L213M	0.0	Yes	-6	X
I214T	0.0	No	-2,76	Shift lambda with 83N (Yokoyama et al. 2008)
L216M	0.0	No	0,39	
I217T	0.0004	Yes	**28**	X
V218I	0.0	No	-5,2	X
V219I	0.0007	No	**22,49**	X
R248K	0.0	No	1,91	X
T251S	0.0	No	10	X
V254C	0.0006	No	-4,58	X
V255I	0.0	No	**46**	X
I256M	0.0	No	-4	X

I256L	0.0001	No	**18,53**	X
A260G	0.0	No	-8,02	X
F261Y	0.0	No	**58**	Shift lambda (Yokoyama 2008)
V263I	0.0003	No	**22**	X
L266C	0.0	Yes	**74,08**	X
S270G	0.0	Yes	**12**	Shift lambda in bovine and human (Shen et al. 2010; Morrow et al. 2017)
V271T	0.0	No	**12,75**	X
Y274W	0.0002	Yes	**25,49**	Sensitivity under low light? (Wu et al. 2021)
W274Y	0.0	Yes	**58**	Reversion
T277C	0.0	No	6,65	X
H278N	0.0	Yes	4	X
Q279K	0.0	No	**15,1**	X
K279Q	0.0004	No	**22**	X
S281T	0.0	No	8,99	X
A292S	0.0	No	**70**	Shift lambda (Yokoyama 2008)
S299A	0.0	No	**58**	Shift lambda (Yokoyama 2008; Dungan and Chang 2017)

Table S4: Results of ConDor detection on fish rhodopsin dataset, sorted by position.

Mutation: Mutations detected as convergent with ConDor on the fish rhodopsin dataset. The p-values are those given as output by ConDor and are not corrected. The list of mutations corresponds to those whose p-value is less than or equal to the acceptance threshold of 0.0007 after "Benjamini-Hochberg" correction with an alpha risk of 0.05. Fast: detected mutation on a position that belong to the 5% positions with highest evolutionary rate on the whole dataset. Bayes Factor: Bayes factor calculated by BayesTraits 'discrete dependent' model (Pagel 1994; Pagel and Meade 2006). If the Bayes factor is greater than 10 (highlighted in bold), the mutation is significantly correlated with the habitat (marine or fresh/brackish water). Role of the mutation: possible role of the mutation found in the literature, with the reference.

We detected 55 events with ConDor, 12 of which are found at fast positions, 32 are correlated with marine or fresh/brackish water environments, 15 were already discussed in the literature.

Positions	PCOC	PC	OC	ConDor detects
112	0.0	0.999241853902	0.0	No
122	0.0	0.999982215936	0.0	No
123	0.0	0.936513873132	0.0	Yes
130	0.0	0.999846947621	0.0	No
157	0.0	0.885208130113	0.0	No
166	0.0	0.950774252529	0.0	Yes
198	0.0	0.999964399102	0.0	Yes
212	0.0	0.838911340706	0.0	No
220	0.0	0.916154973344	0.0	No
250	0.0	0.872817517838	0.0	No
275	0.0	0.925024916922	0.0	No
281	0.0	0.99999982137	0.0	No
283	0.0	0.969703702425	0.0	No
288	0.0	0.918405596855	0.0	No
289	0.0	0.913253682925	0.0	Yes
293	0.0	0.939855892307	0.0	No

Table S5: Results of PCOC detection on fish rhodopsin dataset, sorted by position.

Positions detected by PCOC considering that the convergent clades were those annotated with fresh/brackish water. PCOC: combination of posterior probabilities of PC (Profile change) and OC (One change) model. PC: posterior probabilities of the PC model. The threshold for significance was set to 0.8; OC: posterior probabilities of OC model. The threshold for significance was set to 0.8; ConDor detects: If we detect a convergent mutation with ConDor at the given position.

All the detections are due to the changes in amino acid profiles (PC model) which means that species with the convergent phenotype shifted to a different vector of amino acid probabilities compared to their ancestors. These shifts occurred at 16 different positions, 4 of which intersect with ConDor detections. However, the OC model is not significant as the model shift did not occur at the beginning of the branch supporting the clade of convergent species. Since the OC model is not verified, there are no strictly convergent positions detected for the given phenotype (fresh/brackish water).

Supplementary References

Alteri C, Surdo M, Di Maio VC, Santo F, Costa G, Parrotta L, Romeo I, Gori C, Santoro M, Fedele V, et al. 2016. The HIV-1 reverse transcriptase polymorphism A98S improves the response to tenofovir disoproxil fumarate+emtricitabine-containing HAART both in vivo and in vitro. Journal of Global Antimicrobial Resistance. 7. doi:10.1016/j.jgar.2016.06.005.

Athanasiou D, Aquila M, Bellingham J, Li W, McCulley C, Reeves PJ, Cheetham ME. 2018. The molecular and cellular basis of rhodopsin retinitis pigmentosa reveals potential strategies for therapy. Prog Retin Eye Res. 62:1–23. doi:10.1016/j.preteyeres.2017.10.002.

Ceccherini-Silberstein F, Cozzi-Lepri A, Ruiz L, Mocroft A, Phillips A, Olsen C, Gatell J, Günthard H, Reiss P, Perno C, et al. 2007. Impact of HIV-1 Reverse Transcriptase Polymorphism F214L on Virological Response to Thymidine Analogue—Based Regimens in Antiretroviral Therapy (ART)—Naive and ART-Experienced Patients. The Journal of infectious diseases. 196:1180–90. doi:10.1086/521678.

Ceccherini-Silberstein F, Svicher V, Sing T, Artese A, Santoro MM, Forbici F, Bertoli A, Alcaro S, Palamara G, d'Arminio Monforte A, et al. 2007. Characterization and Structural Analysis of Novel Mutations in Human Immunodeficiency Virus Type 1 Reverse Transcriptase Involved in the Regulation of Resistance to Nonnucleoside Inhibitors. J Virol. 81(20):11507–11519. doi:10.1128/JVI.00303-07.

Doualla-Bell F, Gaseitsiwe S, Ndung'u T, Modukanele M, Peter T, Novitsky V, Ndwapi N, Tendani G, Avalos A, Wester W, et al. 2004. Mutations and Polymorphisms Associated with Antiretroviral Drugs in HIV-1C-Infected African Patients. Antivir Chem Chemother. 15(4):189–200. doi:10.1177/095632020401500402.

Dungan SZ, Chang BSW. 2017. Epistatic interactions influence terrestrial–marine functional shifts in cetacean rhodopsin. Proceedings of the Royal Society B: Biological Sciences. 284(1850):20162743. doi:10.1098/rspb.2016.2743.

Gonzales MJ, Wu TD, Taylor J, Belitskaya I, Kantor R, Israelski D, Chou S, Zolopa AR, Fessel WJ, Shafer RW. 2003. Extended spectrum of HIV-1 reverse transcriptase mutations in patients receiving multiple nucleoside analog inhibitors: AIDS. 17(6):791–799. doi:10.1097/00002030-200304110-00003.

Hachiya A, Marchand B, Kirby KA, Michailidis E, Tu X, Palczewski K, Ong YT, Li Z, Griffin DT, Schuckmann MM, et al. 2012. HIV-1 Reverse Transcriptase (RT) Polymorphism 172K Suppresses the Effect of Clinically Relevant Drug Resistance Mutations to Both Nucleoside and Non-nucleoside RT Inhibitors. J Biol Chem. 287(35):29988–29999. doi:10.1074/jbc.M112.351551.

Hunt DM, Dulai KS, Partridge JC, Cottrill P, Bowmaker JK. 2001. The molecular basis for spectral tuning of rod visual pigments in deep-sea fish. Journal of Experimental Biology. 204(19):3333–3344. doi:10.1242/jeb.204.19.3333.

Ishikawa SA, Zhukova A, Iwasaki W, Gascuel O. A Fast Likelihood Method to Reconstruct and Visualize Ancestral Scenarios. Mol Biol Evol. doi:10.1093/molbev/msz131. [accessed 2019 Jun 7]. https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msz131/5498561.

Kawamoto A, Kodama E, Sarafianos SG, Sakagami Y, Kohgo S, Kitano K, Ashida N, Iwai Y, Hayakawa H, Nakata H, et al. 2008. 2'-deoxy-4'-C-ethynyl-2-halo-adenosines active against drug-resistant human immunodeficiency virus type 1 variants. Int J Biochem Cell Biol. 40(11):2410–2420. doi:10.1016/j.biocel.2008.04.007.

Liu P, Feng Y, Wu J, Tian S, Su B, Wang Z, Liao L, Xing H, You Y, Shao Y, et al. 2017. Polymorphisms and Mutational Covariation Associated with Death in a Prospective Cohort of HIV/AIDS Patients Receiving Long-Term ART in China. PLOS ONE. 12(1):e0170139. doi:10.1371/journal.pone.0170139.

Malinsky M, Challis RJ, Tyers AM, Schiffels S, Terai Y, Ngatunga BP, Miska EA, Durbin R, Genner MJ, Turner GF. 2015. Genomic islands of speciation separate cichlid ecomorphs in an East African crater lake. Science. 350(6267):1493–1498. doi:10.1126/science.aac9927.

Marcelin A-G, Flandre P, Furco A, Wirden M, Molina J-M, Calvez V, AI454-176 Jaguar Study Team. 2006. Impact of HIV-1 reverse transcriptase polymorphism at codons 211 and 228 on virological response to didanosine. Antivir Ther. 11(6):693–699.

Matsumoto Y, Oda S, Mitani H, Kawamura S. 2020. Orthologous Divergence and Paralogous Anticonvergence in Molecular Evolution of Triplicated Green Opsin Genes in Medaka Fish, Genus Oryzias. Genome Biol Evol. 12(6):911–923. doi:10.1093/gbe/evaa111.

Melikian GL, Rhee S-Y, Varghese V, Porter D, White K, Taylor J, Towner W, Troia P, Burack J, DeJesus E, et al. 2014. Non-nucleoside reverse transcriptase inhibitor (NNRTI) cross-resistance: implications for preclinical evaluation of novel NNRTIs and clinical genotypic resistance testing. J Antimicrob Chemother. 69(1):12–20. doi:10.1093/jac/dkt316.

Morrow JM, Castiglione GM, Dungan SZ, Tang PL, Bhattacharyya N, Hauser FE, Chang BSW. 2017. An experimental comparison of human and bovine rhodopsin provides insight into the molecular basis of retinal disease. FEBS Letters. 591(12):1720–1731. doi:10.1002/1873-3468.12637.

Murrell B, Wertheim JO, Moola S, Weighill T, Scheffler K, Pond SLK. 2012. Detecting Individual Sites Subject to Episodic Diversifying Selection. PLOS Genetics. 8(7):e1002764. doi:10.1371/journal.pgen.1002764.

Nebbia G, Sabin CA, Dunn DT, Geretti AM, UK Collaborative Group on HIV Drug Resistance, UK Collaborative HIV Cohort (CHIC) Study Group. 2007. Emergence of the H208Y mutation in the reverse transcriptase (RT) of HIV-1 in association with nucleoside RT inhibitor therapy. J Antimicrob Chemother. 59(5):1013–1016. doi:10.1093/jac/dkm067.

O'Reilly JE, Puttick MN, Parry L, Tanner AR, Tarver JE, Fleming J, Pisani D, Donoghue PCJ. 2016. Bayesian methods outperform parsimony but at the expense of precision in the estimation of phylogeny from discrete morphological data. Biology Letters. 12(4):20160081. doi:10.1098/rsbl.2016.0081.

Pagel M. 1994. Detecting correlated evolution on phylogenies: a general method for the comparative analysis of discrete characters. Proceedings of the Royal Society of London Series B: Biological Sciences. 255(1342):37–45. doi:10.1098/rspb.1994.0006.

Pagel M, Meade A. 2006. Bayesian Analysis of Correlated Evolution of Discrete Characters by Reversible-Jump Markov Chain Monte Carlo. The American Naturalist. 167(6):808–825. doi:10.1086/503444.

Precious HM, Günthard HF, Wong JK, D'Aquila RT, Johnson VA, Kuritzkes DR, Richman DD, Brown AJL. 2000. Multiple sites in HIV-1 reverse transcriptase associated with virological response to combination therapy. AIDS. 14(1):31–36.

Saracino A, Monno L, Scudeller L, Cibelli DC, Tartaglia A, Punzi G, Torti C, Caputo SL, Mazzotta F, Scotto G, et al. 2006. Impact of unreported HIV-1 reverse transcriptase mutations on phenotypic

resistance to nucleoside and non-nucleoside inhibitors. Journal of Medical Virology. 78(1). doi:10.1002/jmv.20500. [accessed 2021 May 12]. https://www.readcube.com/articles/10.1002%2Fjmv.20500.

Scherrer AU, von Wyl V, Götte M, Klimkait T, Cellerai C, Yerly S, Böni J, Held L, Ledergerber B, Günthard HF, et al. 2012. Polymorphic mutations associated with the emergence of the multinucleoside/tide resistance mutations 69 insertion and Q151M. J Acquir Immune Defic Syndr. 59(2):105–112. doi:10.1097/QAI.0b013e31823c8b69.

Shen Y-Y, Liu J, Irwin DM, Zhang Y-P. 2010. Parallel and convergent evolution of the dim-light vision gene RH1 in bats (Order: Chiroptera). PLoS One. 5(1):e8838. doi:10.1371/journal.pone.0008838.

Sugawara T, Terai Y, Imai H, Turner GF, Koblmüller S, Sturmbauer C, Shichida Y, Okada N. 2005. Parallelism of amino acid changes at the RH1 affecting spectral sensitivity among deep-water cichlids from Lakes Tanganyika and Malawi. PNAS. 102(15):5448–5453. doi:10.1073/pnas.0405302102.

Svarovskaia ES, Feng JY, Margot NA, Myrick F, Goodman D, Ly JK, White KL, Kutty N, Wang R, Borroto-Esoda K, et al. 2008. The A62V and S68G mutations in HIV-1 reverse transcriptase partially restore the replication defect associated with the K65R mutation. J Acquir Immune Defic Syndr. 48(4):428–436. doi:10.1097/QAI.0b013e31817bbe93.

Svicher V, Sing T, Santoro MM, Forbici F, Rodríguez-Barrios F, Bertoli A, Beerenwinkel N, Bellocchi MC, Gago F, d'Arminio Monforte A, et al. 2006. Involvement of Novel Human Immunodeficiency Virus Type 1 Reverse Transcriptase Mutations in the Regulation of Resistance to Nucleoside Inhibitors. J Virol. 80(14):7186–7198. doi:10.1128/JVI.02084-05.

Van Nynatten A, Castiglione GM, de A. Gutierrez E, Lovejoy NR, Chang BSW. 2021. Recreated Ancestral Opsin Associated with Marine to Freshwater Croaker Invasion Reveals Kinetic and Spectral Adaptation. Molecular Biology and Evolution. 38(5):2076–2087. doi:10.1093/molbev/msab008.

Varela A, Ritchie P. 2014. Critical amino acid replacements in the rhodopsin gene of 19 teleost species occupying different light environments from shallow-waters to the deep-sea. Environmental Biology of Fishes. 98:193–200. doi:10.1007/s10641-014-0249-4.

Winters MA, Merigan TC. 2001. Variants Other than Aspartic Acid at Codon 69 of the Human Immunodeficiency Virus Type 1 Reverse Transcriptase Gene Affect Susceptibility to Nucleoside Analogs. Antimicrob Agents Chemother. 45(8):2276–2279. doi:10.1128/AAC.45.8.2276-2279.2001.

Wu B, Feng C, Zhu C, Xu W, Yuan Y, Hu M, Yuan K, Li Y, Ren Y, Zhou Y, et al. 2021. The Genomes of Two Billfishes Provide Insights into the Evolution of Endothermy in Teleosts. Molecular Biology and Evolution.(msab035). doi:10.1093/molbev/msab035. [accessed 2021 May 14]. https://doi.org/10.1093/molbev/msab035.

Yokoyama S. 2008. Evolution of Dim-Light and Color Vision Pigments. Annu Rev Genom Hum Genet. 9(1):259–282. doi:10.1146/annurev.genom.9.081307.164228.

Yokoyama S, Tada T, Zhang H, Britt L. 2008. Elucidation of phenotypic adaptations: Molecular analyses of dim-light vision proteins in vertebrates. Proc Natl Acad Sci U S A. 105(36):13480–13485. doi:10.1073/pnas.0802426105.

Zimbwa P, Milicic A, Frater J, Scriba TJ, Willis A, Goulder PJR, Pillay T, Gunthard H, Weber JN, Zhang H-T, et al. 2007. Precise identification of a human immunodeficiency virus type 1 antigen processing mutant. J Virol. 81(4):2031–2038. doi:10.1128/JVI.00968-06.