Published online 25 May 2020 *Nucleic Acids Research, 2020, Vol. 48, Web Server issue* ***W321–W331***
doi: 10.1093/nar/gkaa408

Pseudo*Checker*: an integrated online platform for gene inactivation inference

Luís Q. Alves[1,*], Raquel Ruivo[1,*], Miguel M. Fonseca[1], Mónica Lopes-Marques[1], Pedro Ribeiro[2] and L. Filipe C. Castro[1,3,*]

[1]CIIMAR-Interdisciplinary Centre of Marine and Environmental Research, U. Porto-University of Porto, Matosinhos, 4450-208, Portugal, [2]CRACS & INESC-TEC Department of Computer Science, FCUP, Porto, 4169-007, Portugal and [3]Department of Biology, FCUP, Porto, 4169-007, Portugal

Received March 10, 2020; Revised April 22, 2020; Editorial Decision May 04, 2020; Accepted May 06, 2020

ABSTRACT

The rapid expansion of high-quality genome assemblies, exemplified by ongoing initiatives such as the Genome-10K and i5k, demands novel automated methods to approach comparative genomics. Of these, the study of inactivating mutations in the coding region of genes, or pseudogenization, as a source of evolutionary novelty is mostly overlooked. Thus, to address such evolutionary/genomic events, a systematic, accurate and computationally automated approach is required. Here, we present Pseudo*Checker*, the first integrated online platform for gene inactivation inference. Unlike the few existing methods, our comparative genomics-based approach displays full automation, a built-in graphical user interface and a novel index, PseudoIndex, for an empirical evaluation of the gene coding status. As a multi-platform online service, Pseudo*Checker* simplifies access and usability, allowing a fast identification of disruptive mutations. An analysis of 30 genes previously reported to be eroded in mammals, and 30 viable genes from the same lineages, demonstrated that Pseudo*Checker* was able to correctly infer 97% of loss events and 95% of functional genes, confirming its reliability. Pseudo*Checker* is freely available, without login required, at http://pseudochecker.ciimar.up.pt.

INTRODUCTION

Understanding the molecular signatures underlying the evolution of phenotypic traits is a key challenge in both contemporary evolutionary biology and genomics (1). While events of gene duplication and amino acid divergence have frequently been associated with the evolution of novel traits, gene loss, on the other hand, has been less regarded as an evolutionary force *per se* (2). In fact, events of redundant loss have been thoroughly associated with the non-functionalization of genes arising from the accumulation of deleterious mutations, a process termed pseudogenization, following gene duplication (duplicated pseudogenes) or events of transposition of processed transcripts (processed pseudogenes) (3–5). Yet, non-redundant gene loss mechanisms, including complete gene elimination or pseudogenization (unitary pseudogenes), have been increasingly linked to phenotypic modifications, from adaptive and regressive perspectives (1,2,6–27).

Despite the current wealth of genome availability and rapid increase of high-quality genome assemblies, exemplified by the recent sequencing of 48 bird genomes (28), and ongoing projects such as Genome-10K (29) and i5k (30), the assessment of gene loss events still suffers from technical inertia. Additionally, some studies suggest that real pseudogenes can be mistakenly annotated as functional protein-coding genes. ORF-disrupting mutations, including frameshifts or in-frame premature stop codons, are often weighed as sequencing or assembly artefacts, being automatically corrected by whole-genome annotators (19–22,31). This is particularly relevant for the mammalian lineage. Mammals represent a diverse group of species, occupying a wide range of ecological niches, and displaying iconic phenotypic adaptations to their surrounding environment: including specialized dentition, placentation, enlarged brains, lactation, increased sensitivity of sense organs and hair to preserve heat and skin (1). Importantly, some of such mammalian-specific phenotypic modifications have been assigned to gene loss events in response to specific environmental cues (7–17,19–22). Nonetheless, the repertoire of gene loss in mammals, including affected genes and lineages, is still vastly incomplete (2), placing this group as a

*To whom correspondence should be addressed. Tel: +350 223 401 800; Fax: +350 223 401 800; Email: filipe.castro@ciimar.up.pt
Correspondence may also be addressed to Luís Q. Alves. Email: luis.alves@ciimar.up.pt
Correspondence may also be addressed to Raquel Ruivo. Email: ruivoraquel@gmail.com
Present address: Mónica Lopes-Marques, Population Genetics and Evolution Group, i3S- Instituto de Investigação e Inovação em Saúde, Universidade do Porto, Rua Alfredo Allen 208, 4200-135 Porto, Portugal.

reference test case to address the role and magnitude of gene loss as a major driver of morphological diversification and adaptation.

Although automatic and semi-automatic pipelines are currently available for the identification of duplicated and processed pseudogenes (32–34), the few systematic approaches capable of inferring episodes of non-redundant gene inactivation events display some restrictions including: (i) a reduced degree of automation, the requirement of whole genomes and absence of multiple sequence alignment-based methods (1); (ii) the lack of an objective metric capable of measuring gene erosion (6,17,19–22); or (iii) the necessity of exhaustive manual curation at every stage (6,17,19–22), which is less practical when dealing with the hundreds of genomes currently available.

To circumvent these bottlenecks, we developed Pseudo*Checker,* to the best of our knowledge, the first integrated online platform for gene inactivation inference. Pseudo*Checker* aims to facilitate and promote the study of gene inactivation as a driver of evolutionary change, providing an easy to use, systematic, highly accurate and computationally automatic approach. Our comparative genomics-based method consists of an online three-step-based computational pipeline able to infer the coding status of a given eukaryotic nuclear protein-coding gene in single or multiple species of interest by taking advantage of existing genomic data.

While making use of minimal user input and a set of established parameters, Pseudo*Checker* is capable of: (i) identifying gene inactivation events, automatically, remotely and in a relative short amount of time, highlighting the mutational evidence for a set of unlimited target species with available genomic data; (ii) unveiling ancestral gene inactivation events by accurately displaying conserved gene inactivating mutations across closely related *taxa* within a given analysis; (iii) measuring the erosion level of a candidate gene in any target species by assigning an index of pseudogenization, the PseudoIndex; (iv) including external functional gene datasets into the analyses; (v) and exporting the produced data throughout the analysis, useful for performing downstream complementary tasks including phylogenetic reconstructions and selection analyses.

Our software was built to accompany the emerging need of a convenient, comprehensive and complementary analysis tool for the fast-developing gene loss research field, being freely accessible, without login required, including supporting documentation and example data, at http://pseudochecker.ciimar.up.pt/.

ANALYSIS WORKFLOW

Pseudo*Checker*'s overview

A gene is considered inactivated in a given lineage if it complies with two conditions: first, it must derive from an ancestral sequence yielding an intact protein-coding gene; second, it should display evidence of erosion such as the complete absence of the corresponding orthologous genomic *locus*, or accumulation of open reading frame (ORF) disrupting mutations that likely results in non-functionalization (establishing a unitary pseudogene) (1,35). Pseudo*Checker* infers the coding status of a given candidate gene in a target species using, as a reference, an orthologous coding sequence. Pseudo*Checker* takes into account the coding sequence conservation across related species (36), requiring previous phylogenetic contextualization. Gene annotation is followed by the screening of gene sequence eroding features.

Specifically, our general-purpose bioinformatics tool was designed to be easily applied in two different situations: (i) *de novo* candidate gene annotation, for instance for unannotated genomes; (ii) re-annotation of candidate genes in previously automatic annotated genomes, to verify previous annotations and identify unitary pseudogenes erroneously annotated as functional protein-coding genes.

The pipeline, input data and parameters

To run a Pseudo*Checker* analysis the user will require two main inputs: (i) a single reference nucleotide coding sequence (CDS) and the respective exon nucleotide sequence(s), both annotated and retrieved from a given reference species (FASTA format) (if a distinct gene isoform exists, i.e. splice variants, the user must select a single reference sequence); (ii) and, for each target species, a corresponding genomic sequence, against which the reference coding exons will be mapped to predict the gene CDS in the target species (FASTA format). The user is responsible for ensuring that each inserted sequence is orthologous to the in-study gene. As target sequences, our tool supports either partial/full genomic contigs, scaffolds or genomic sequencing reads. Optionally, and also in FASTA format, the user may include complete functional nucleotide coding sequences into a given analysis - referred to as the predetermined coding sequences, further incorporated into the second component of Pseudo*Checker* 's pipeline and, consequently, in the final output (see below). Detailed information on how the data should be formatted and submitted into the software is available at Pseudo*Checker*'s instructions page (http://pseudochecker.ciimar.up.pt/pseudochecker/instructions.html).

Once the input data are correctly assigned, and the available parameters, underlying the different components of the three-step integrated pipeline, are selected, the latter is executed as follows (Figure 1):

- Coding sequence prediction: for each target species, Pseudo*Checker* annotates the orthologous exons and, consequently, predicts the sequence of the in-study gene; this is done by performing a progressive deterministic nucleotide pairwise alignment of each reference species' coding exon, from the 5′ to the 3′ end of the gene, against the corresponding inputted genomic sequence of each target species. Our method makes use of the semi-global variation of the classical global alignment Needleman-Wunsch algorithm (1970) (37) for computing each alignment.
- Alignment by MACSE: once the first step is concluded, Pseudo*Checker* runs MACSE v2 (38), a standalone sequence aligner software, already used in gene pseudogenization studies (10,39). Here, a pairwise or multiple alignment between predicted sequence(s) (coding or

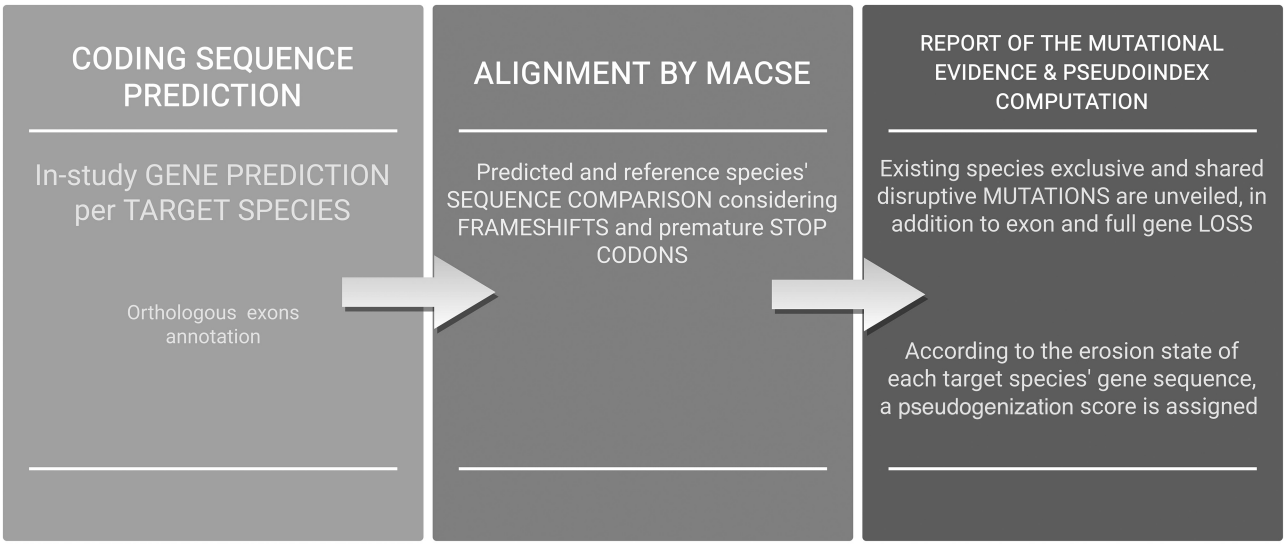

Figure 1. Conceptualization of the three-step-based Pseudo*Checker*'s computational pipeline.

pseudogenized), predetermined coding sequences (optional) and the reference coding sequence (functional) is produced. The alignment is computed considering the underlying amino acid translation of each sequence and the eventual presence of frameshifts and premature stop codons, while preserving the underlying codon structure. Algorithmically speaking, the MACSE solution is an improved version of the Needleman-Wunsch algorithm (37) that, for computing each optimal pairwise alignment, adds alignment costs associated with frameshift mutations and stop codons. If >2 sequences are inputted to the MACSE dataset, it extends the pairwise to a multiple sequence alignment by using a progressive alignment strategy in order to obtain an initial draft that is subsequently refined.

- Report of the mutational evidence and PseudoIndex computation: for each target species, and in agreement with the previous alignment produced by MACSE, existing cross-species conserved and non-conserved gene deleterious mutations, including frameshift mutations and in-frame premature stop codons, relative to the reference CDS, are identified. Following the first component of the pipeline, splice site disrupting mutations (any deviation to the consensus GT/GC-AG splice site pairs), full gene loss or exon loss are also revealed. Finally, considering the presence/absence of full or partial target gene sequences and degree of mutational evidence, a pseudogenization score, the PseudoIndex, is assigned to each target species corresponding sequence.

Importantly, parameter selection at the Pseudo*Checker*'s homepage, preceding each analysis, will have an impact on the outcome. Pseudo*Checker* divides the total set of required parameters into two different groups: (i) parameters related with the first component of the pipeline, the coding sequence prediction and (ii) the parameters related with the second component, the MACSE alignment.

Concerning the former, it includes:

- The similarity scoring scheme used for computing each exon alignment at the coding sequence prediction step of Pseudo*Checker*'s pipeline, with three available options: (a) one to be used when the reference and the target species are closely related - the closely related species optimized similarity scoring scheme; (b) other to be used if the reference and the target species are slightly more evolutionary divergent; and finally, (c) the best-fit similarity scoring scheme (the default and recommended), to be used when there is no clear idea about the evolutionary divergence between both the reference and test species and/or the candidate gene's conservation state. This corresponds to a (slightly more time consuming) dynamic similarity scoring scheme intending to make Pseudo*Checker*'s produced exon alignments more resistant to the evolutionary divergence of the in-study species and/or gene itself. Following the best-fit similarity scoring scheme, Pseudo*Checker* tests, in the worst-case scenario, 40 different combinations of match/mismatch alignment punctuating schemes per reference exon, with the ultimate goal of finding an alignment yielding a predicted exon presenting conserved adjacent splice sites (GT/GC-AG splice site pairs) and without underlying reading frame-disrupting indels.
- A parameter allowing the automatic trimming of any untranslated regions (UTR's) lying within the reference species' 5′ and/or 3′ coding exon of the candidate gene, as their absence is mandatory for an accurate prediction of the in-study gene's coding sequence for each test species.
- An additional parameter related with the optional extension of the reference species' 3′ coding exon alignment to allow the identification of a missing (in the original alignment) downstream final stop codon. This constitutes a particularly useful feature to select for cases where the C-terminus of the protein encoded by the in-study gene is slightly divergent in size across the tested lineages: the original alignment being putatively not capable of detecting an eventual more distant final stop codon in the test

species relative to the reference species' stop codon. Importantly, the present version of Pseudo*Checker* does not feature an analogous optional parameter for searching a missing (in the original alignment) upstream start codon within a given test species' sequence. In effect, during development of this software, we did not face difficulties in detecting ATG codons as the first codons of predicted coding sequences.
- At last, the minimum percentage of alignment identity for an exon alignment to be considered as a valid alignment. This parameter will define if a given exon alignment corresponds to a real, biologically meaningful alignment, rather than a non-specific alignment. Each predicted exon for which the corresponding alignment identity is inferior to this value is considered as lost at the corresponding target species (or so eroded that any similarity is destroyed), thus, being excluded from the final annotated/predicted sequence. Predicted coding sequences under this condition are considered as partial coding sequences. Importantly, the value for this parameter should be adjusted according to the evolutionary relationship between the reference and target species. Particularly, if the reference species is highly divergent from the test species, the user might prefer choosing a lower value for this parameter, whereas, for the opposite case, a higher value might constitute a better option.

With respect to the parameters related with the MACSE alignment, these comprise the cost of each possible alignment event. The introduced values for these costs will be followed during the MACSE alignment computation, and these are clustered into three sub-groups, according to the type of sequence each parameter will interfere with:

- One related to the MACSE alignment costs associated with reliable sequences: including the reference species' candidate gene CDS, the coding sequences predicted as functional during the first step of Pseudo*Checker*'s pipeline, that is, sequences that do not exhibit frameshift gaps and/or in-frame premature stop codons within each exon alignment underlying its prediction, and predetermined coding sequences.
- A second concerning the MACSE alignment costs associated with less reliable sequences, which include predicted sequences during the first step of Pseudo*Checker*'s pipeline displaying frameshift gaps and/or in-frame premature stop codons within at least one exon alignment.
- Finally, a third associated with the MACSE alignment costs targeted for both types of sequences (reliable and less reliable sequences).

The default MACSE alignment costs provided by Pseudo*Checker* constitute the default similarity scoring scheme provided by the MACSE authors that from their experience has proven to be effective for most cases (38). As amino acid substitution matrix, Pseudo*Checker* uses the BLOSUM62 matrix, for which the default alignment costs are optimized, and a description of each MACSE alignment cost can be found within the advanced options section of the Pseudo*Checker*'s home page. Detailed information regarding these parameters can be found at the following MACSE documentation webpage https://bioweb.supagro.inra.fr/macse/index.php?menu=doc/delegations/costs.

Notably, not every MACSE produced alignment is viable for running a Pseudo*Checker* analysis. Even though from our experience such event is unlikely to occur, since the produced alignment depends on the defined MACSE similarity scoring scheme at Pseudo*Checker*'s home page, inadequate choices for a given in-study sequence dataset might lead to the appearance of frameshift mutations and/or premature stop codons at the reference species' CDS and/or at input predetermined coding sequences (optional). As these sequences are supposed to be functional, therefore, not presenting any frameshifts or premature stop codons arising within their aligned sequences, Pseudo*Checker* automatically interrupts any analysis concealing these erroneous situations.

Output

When a Pseudo*Checker* job is completed, the software automatically redirects the user to the corresponding results page. This interactive and intuitive web interface is divided into different sections, each providing different levels of information regarding an executed analysis (Figure 2).

- At the top of the results page, the Alignment by MACSE is provided. MACSE produces an alignment containing information at the nucleotide and amino acid levels for each aligned sequence, represented by Pseudo*Checker* as the top and bottom sequence, respectively. For a convenient visualization, the reference species' CDS is always shown at the top, and the alignment is colour graded according to the resulting codon structure, with each set of adjacent blocks of three nucleotides represented with different background colours. In detail, at the nucleotide level, frame-preserving gaps are represented by a codon '- - -' on a white background. At the amino acid level, in contrast, no special representation is applied. Regarding frameshift mutations, at the nucleotide level, these are represented by a partial codon with one or two exclamation marks (!), each highlighted in orange and, at the amino acid level, no representation is used. For in-frame stop codons, at the nucleotide level, these are represented in red font and white background colour and, at the amino acid level, by an asterisk (*), also in red font. Finally, if a given amino acid differs from the reference sequence at the same alignment position, the target sequence amino acid is represented with a grey background colour (Figure 2).
MACSE exclamation marks arise within partial codons that derive from frameshift mutations in order to preserve the structure of the reading frame. These partial codons may appear in different forms,'!!N', '!N!' or 'N!!', with 'N' corresponding to any of the four DNA nucleotide bases. Partial codon annotations may pinpoint different interpretations. For instance, a partial codon represented by a '!!N', '!N!' or 'N!!' can either represent the deletion of two nucleotides or a single nucleotide insertion. However, since frameshift mutations are inferred with respect to a reference functional sequence, the mutational interpretation underlying each partial codon must be made accord-

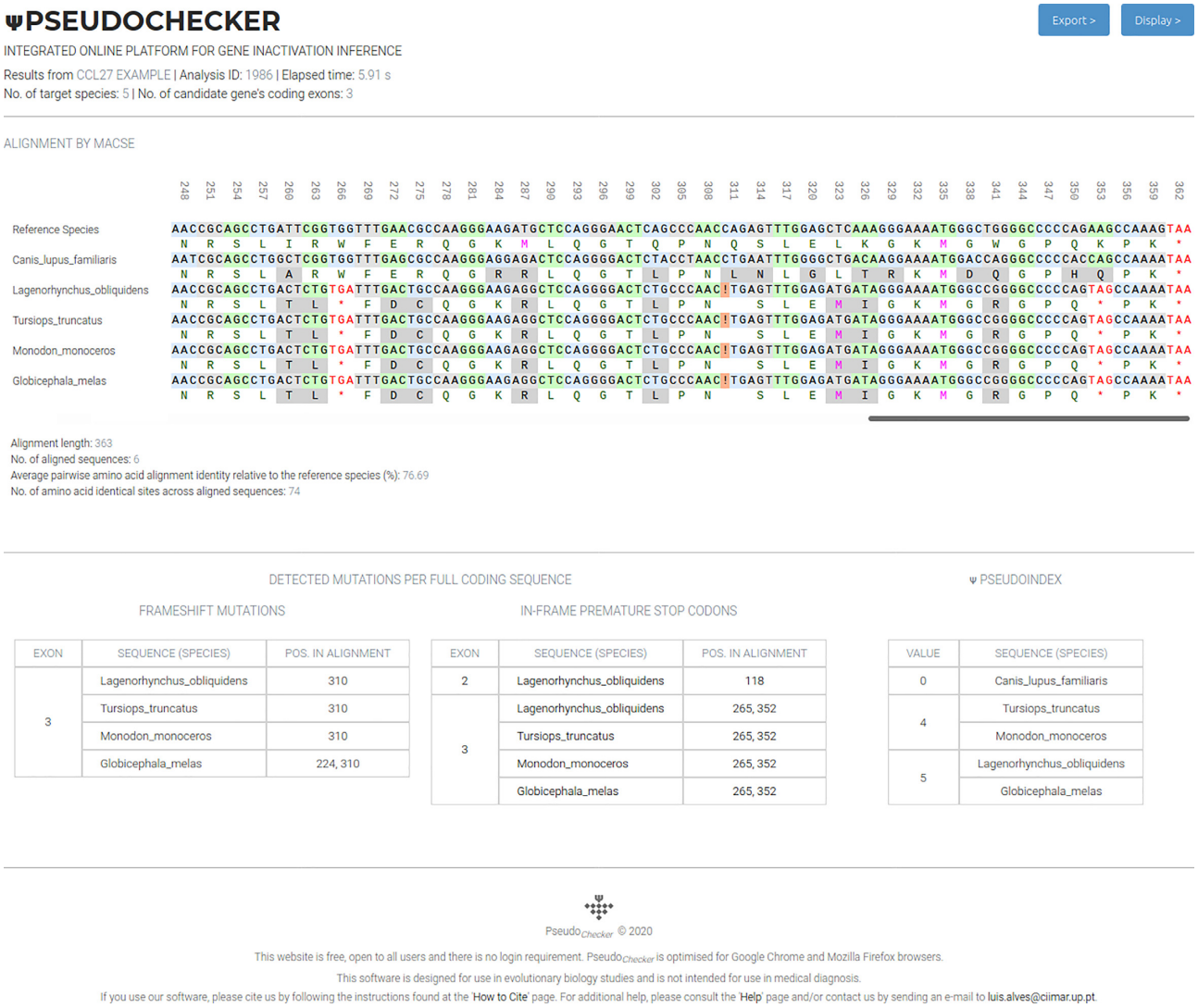

Figure 2. Typical output of a Pseudo*Checker* analysis. General information related to the executed analysis is shown at the top. Below, the alignment executed by MACSE, detected gene inactivating mutations per full coding sequence and PseudoIndex value assigned for each target species corresponding sequence are presented. At the upper right corner, both 'Export' and 'Display' buttons are available for additional options.

ingly, taking into account the reference species' codon observed at the same alignment site.

First, if a given reference codon is represented by 'NNN', where 'N' represents any of the four DNA nucleotide bases, and for the corresponding target species' sequence, the observed codon at the same alignment site is represented by a partial codon containing at least one '!', this should be understood as resulting from frameshift deletions, wherein each exclamation mark represents a single nucleotide deletion. In contrast, if a given reference codon is represented by a set of three gaps '- - -', and the corresponding target species' codon aligning at the same site is, represented by a '!!N', '!NN', 'N!!', 'NN!' or '!N!', this should be interpreted as an insertion of the 'N' DNA nucleotide base(s).

Four levels of information related with the alignment are also presented: alignment length, number of aligned sequences, average pairwise amino acid alignment identity relative to the reference species' sequence and the number of amino acid identical sites across aligned sequences. When supplied, the number of predetermined coding sequences included in the alignment is also shown. Also, partial coding sequences are enumerated, and absent sequences (sequences that are not included in the alignment since their respective species do not present any exon orthologous to the in-study gene) are equally mentioned. Additionally, by clicking in the button 'Export' at the top of the page (Figure 2), it is possible to export the produced alignment by MACSE, as well as the predicted coding sequences at both nucleotide and amino acid levels, the first, particularly useful for directly performing downstream phylogenetic and selection analyses with methods based on codon models of sequence evolution.

- Under the MACSE alignment, a summary of the detected frameshift mutations and stop codons per target species,

corresponding exon, as well as their respective position within the alignment is presented (Figure 2).
Importantly, partial coding sequences are excluded from this feature, only declaring detected mutations when a full sequence is predicted. As aforementioned, MACSE automatically imposes exclamation marks (!) in the most appropriate alignment location to maintain the original structure of the reading frame. However, when single or multiple exons are missing, which is the case for partial coding sequences, if their absence results into the disruption of the reading frame, exclamation marks will arise adjacently to the exons neighbouring the missing ones. This constitutes an issue since it might be difficult for the user to distinguish between real biological mutations from alignment adjustments produced by MACSE aiming to preserve the integrity of the reading frame. Nonetheless, an additional tool, the PseudoIndex, is supplied for target species exhibiting either partial or full coding sequences.

- Optionally, and viewable by clicking in the 'Display' button at the top of the page (Figure 2), a section is displayed presenting additional MACSE alignment metrics for each aligned sequence relative to the reference sequence.
- By clicking in the same button, a section containing several levels of information regarding the prediction of the coding sequence of the in-study gene in each target species can also be displayed. Here, it is possible to, for each target species, export the individualized predicted exons, as well as to visualize the alignments between each reference coding exon and the inputted corresponding genomic sequence.
- Finally, resorting of the same action, the input parameters used for the computation of the in-analysis Pseudo*Checker*'s job might be also accessed.
To each analysis, an ID is assigned. This is presented while waiting for a Pseudo*Checker*'s job to conclude, but also within the results page of a concluded analysis (Figure 2). By inserting a given analysis ID at the Pseudo*Checker*'s 'Submitted Jobs' page (http://pseudochecker.ciimar.up.pt/pseudochecker/submitted_jobs.html), the software will automatically redirect the user to the corresponding results page. This allows the user not only to avoid waiting for an analysis to be completed, but also to consult the results of a previously finished analysis in a later moment.

PseudoIndex

Accurately measuring the level of pseudogenization of a given gene poses several challenges. For instance, evolutionary changes in the exon–intron structures of conserved genes, including splice site shifts over evolution, lineage-specific exons and precise intron deletions, all mimic inactivating mutations in genes that, in fact, might be functional. Additionally, even real mutations might not indicate gene loss: for example, when a given frameshift indel arises but is downstream compensated by an additional frameshift restoring the original reading frame, or when such frameshifts and/or premature stop codons arise close to the sequence region encoding the C-terminus of the resulting protein, which is under less evolutionary constraints (1).

Considering all these factors, the manual screening of a given predicted DNA sequence might be a labouring-intensive and puzzling task. To overcome these challenges, we have built into Pseudo*Checker* the PseudoIndex, a user assistant metric that intends to, at a glance, measure the erosion state of a given gene at a given species by inspecting the presence and magnitude of the mutational evidence.

Explicitly, for each target species, the PseudoIndex takes into account three different components: (i) the absent-exons component that takes into account the percentage of exonic content present in the reference sequence that does not align with the corresponding target genomic sequence; (ii) the shifted codons component that takes into account the percentage of codons that are read out of the reference reading frame; (iii) and the truncated sequence component that measures the percentage of the target sequence that is not translated into protein, due to the presence of a premature stop codon.

Splice site abolishing mutations are not considered for the PseudoIndex calculation since splice site shifts may arise during evolution. In fact, a given splice site may be silenced due to the emergence of a novel, and phylogenetically alternative, splice site. Furthermore, non-canonical splice sites may also occur (6). As such, and to maintain misclassification rate at low values, we decided not to penalize these mutational events within PseudoIndex. Yet, splice site mutations are reported within the coding sequence prediction section, in the Pseudo*Checker*'s results page, and can thus be further scrutinized by users.

The PseudoIndex attributed value for each in-study target gene varies on a discrete scale from 0 to 5, with a PseudoIndex of 0 suggesting the full functionality of the candidate gene and a PseudoIndex of 5 indicating its full inactivation. In detail, a PseudoIndex of 0 indicates that the corresponding species presents an intact, or almost intact sequence version of the in-study gene, and a PseudoIndex of 1 and 2 indicates that, although the predicted gene has shown some mutational evidence, this likely does not affect the functionality of the resulting protein. A PseudoIndex of 3, on the other hand, indicates a doubtful case, for which the coding status of the corresponding gene should be manually inspected, and finally, a PseudoIndex equal to 4 or 5 suggests ORF disabling mutations.

Each of the three mentioned components of PseudoIndex will yield a sub-PseudoIndex: with respect to the exonic content, shifted codons and sequence truncation, also varying on a discrete scale from 0 to 5.

- Within the absent-exons component of the PseudoIndex, Pseudo*Checker* measures the harmful impact of the absence of single or multiple coding exons on the in-study gene in each target species. For this, Pseudo*Checker* starts by computing the weight that each coding exon displays at the reference gene by dividing its nucleotide length over the entire reference coding sequence length. Then, the percentage of absent gene content computed for a target species is the result of the sum of this computed ratio for each absent exon, multiplied by 100. Different obtained

Table 1. Percentage value of absent gene content and attributed sub-PseudoIndex value for the absent-exons component of PseudoIndex

Absent gene content (%)	Sub-PseudoIndex
≤ 10	0
> 10 and ≤ 15	1
> 15 and ≤ 20	2
> 20 and ≤ 25	3
> 25 and ≤ 30	4
> 30	5

Table 2. Percentage value of shifted codons and attributed preliminary sub-PseudoIndex value for the shifted codons component of PseudoIndex

Shifted codons (%)	Preliminary sub-PseudoIndex
≤ 10	0
> 10 and ≤ 15	1
> 15 and ≤ 20	2
> 20 and ≤ 25	3
> 25 and ≤ 30	4
> 30	5

values will yield different sub-PseudoIndex values (Table 1).

- For the shifted codons component of the PseudoIndex, Pseudo*Checker* measures the impact that frameshift mutations have on the in-study gene predicted sequence for a given target species. Here, our approach considers isolated frameshifts (a single frameshift that occurs within a given sequence), multiple frameshifts that do not compensate each other, compensatory frameshifts, and reading frame disruptive effects caused by the absence of single or multiple exons. To this aim, the shifted codons component considers two factors. First, it calculates the percentage of shifted codons that a given sequence displays. In detail, Pseudo*Checker* starts by counting the total number of codons retrieved in a shifted reading frame (gapped codons, read as '- - -' are not considered) from the 5′ end to the 3′ end of the sequence, then it divides the obtained number by the number of total codons within the sequence, and further multiplies it by a factor of 100 (gapped codons are not, once again, considered). This value is only calculated within the predicted coding sequence initiating with the first observed in-frame start codon and ending in the last available codon. Different values obtained for this ratio will result in different computed preliminary sub-PseudoIndex values (Table 2).
  This rationale considers that frameshift mutations, arising before the first observable start codon, do not correspond to real mutational events. Most commonly, the first codon of a predicted coding sequence will correspond to a start codon; however, if such does not occur (due to, for instance, start codon shifts during evolution or alignment related problems), alternative start codons, downstream from the first codon should not be dismissed. In such scenarios, frameshifts that occur upstream from the first observable start codon should be less penalized than frameshifts arising downstream of it.
  Thus, if at least one frameshift mutation arises upstream of the first observable start codon in a given sequence, a minimum value of 3 will be attributed to the sub-PseudoIndex for this component. Consequently, the final sub-PseudoIndex value obtained for this component results in the highest value between 3 and the preliminary sub-PseudoIndex that resulted from the previously computed percentage value of shifted codons (Table 2).
  In contrast, in the absence of frameshift mutations arising upstream of the first in-frame start codon, the computed sub-PseudoIndex value will solely be dependent on the value computed for the preliminary sub-PseudoIndex (Table 2). If no start codons are detected within a sequence of a given species, hindering the assessment of frameshifts, a value of 3 is attributed to the sub-PseudoIndex for this component.

Table 3. Truncated sequence percentage and attributed preliminary sub-PseudoIndex value for the truncated sequence component of PseudoIndex

Truncated sequence (%)	Preliminary sub-PseudoIndex
≤ 10	0
> 10 and ≤ 15	1
> 15 and ≤ 20	2
> 20 and ≤ 25	3
> 25 and ≤ 30	4
> 30	5

- Lastly, the truncated sequence component of PseudoIndex considers the percentage of truncated sequence that each target gene sequence displays. This is defined as the number of non-gapped codons that are not translated into protein (following either an in-frame premature stop codon or an out-of-frame premature stop codon, translated as a real stop codon, as a consequence of an upstream disruption of the reading frame) further divided by the number of codons within the sequence, multiplied by 100.
  Similarly to the previous component of PseudoIndex, this ratio is only calculated between the first observable in-frame start codon and the last available codon. Different values obtained for this metric will also yield different preliminary sub-PseudoIndex values (Table 3).
  Nevertheless, if an in-frame or out-of-frame premature start codon (again, translated as an effective stop codon due to the upstream disruption of the original reading frame) arises prior to the first observed start codon of the in-analysis sequence, the minimum corresponding species assigned sub-PseudoIndex value for this component of PseudoIndex will be equal to 3. Consequently, the final attributed sub-PseudoIndex will be defined by the maximum value between 3 and the preliminary sub-PseudoIndex value, which relates to the percentage of truncated sequence as explained above (Table 3).
  In contrast, if a sequence does not display any premature stop codons arising upstream of the first observed start codon, the attributed sub-PseudoIndex value will only depend on the percentage of truncated sequence (Table 3). Finally, if no start codons are found, the assigned sub-PseudoIndex concerning this PseudoIndex's component will be equal to 3.

The final PseudoIndex value attributed to each target species will correspond to the highest value amongst the computed sub-PseudoIndex values and should be understood as a user-friendly metric that considers multivariate

factors to assist the user in the interpretation of the coding status of a given in-study gene in a given species.

VALIDATION

Experimental design

To test the performance of our approach, we applied it to: (i) genes previously reported as inactivated in mammals; (ii) and to a subset of presumably functional genes in the same group of organisms. In the first case, we scrutinized recently published studies of mammalian gene loss occurrences, affected lineages and confronted Pseudo*Checker* with a total of 30 lost genes. In the latter, Pseudo*Checker* was applied to a set of 30 presumably functional genes across the mammalian lineage, determined according to the following pre-established criteria.

Assuming that highly expressed protein-coding genes are less prone to suffer deleterious mutations (2), we first inspected The Human Protein Atlas (40) database to recover 30 highly expressed genes: 15 retrieved from 15 randomly selected tissues with available expression data, the tissue-specific atlas, and an additionally 15 collected from the cell-specific atlas, a sub-database that contains expression information for different human cellular compartments. In detail, to establish the functional protein-coding gene repertoire to be used in Pseudo*Checker*, we imposed two initial filters to ensure the orthology and viability of each selected gene across mammals. For each of the selected tissue or cellular compartment samples, we (i) verified if the most expressed gene was annotated in at least one species from 19 out of the 20 orders of mammals (41), with available and annotated genomes at the National Center for Biotechnology Information (NCBI), and (ii) if the given gene was not reported as lost in any mammalian lineage. If both conditions were met, the corresponding gene was included in the repertoire of functional genes. If not, the next most expressed genes from the same inspected samples were screened until finding a suitable gene, which obeyed to both conditions.

With respect to the target lineages to be included in each analysis, and concerning the ones involving lost genes, the respective affected species with annotation available at NCBI for the in-study gene were inspected, and if existent and displaying no assembly gaps at the annotated genomic sequence (represented by contiguous N's, that could negatively influence the performance analysis' outcome), these were directly collected and inputted into the analysis. For the analyses actuating over presumable functional genes, per each of these, the same previously scrutinized 20 mammalian orders were inspected for the presence of at least one belonging species presenting annotation of it at NCBI, that, in addition to the absence of sequencing gaps, should not present the low-quality protein tag in at least one annotated gene isoform: an NCBI RefSeq tag indicating that the annotated sequence was modified to correct possible deleterious indels and stop codons, arising either from assembly artefacts or real biological mutations. If no suitable genomic sequence was found amongst the available species of a given order, the corresponding lineage was excluded for the analysis.

Next, a Pseudo*Checker* analysis was run for each candidate test gene and using the three similarity scoring schemes provided by the first component of Pseudo*Checker*'s pipeline. Thus, a total of 180 analyses were conducted, distributed between two categories, functional and lost, and three similarity scoring scheme variations, relevant to measure the impact of different schemes in the classification outcome.

For all analyses, we fixed human as the reference species, not only to standardize the reference lineage but due the quality, comprehensiveness and completeness of the genome, also including manual curation for the vast majority of the annotated genes (1). For each analyzed gene, the longest annotated and curated sequence was preferred, retrieved from the NCBI's Gene database and inputted into the analysis. No extension of the reference 3′ (or single-exon) alignment to search a missing (in the original alignment) final stop codon was requested, and the MACSE alignment costs were left with the default values, not only due to the impracticability to test all possible combinations of alignment costs but also, as previously mentioned, due to the fact that these are reliable for running the majority of analyses. Since we employed a great diversity of species within each analysis and used human as the reference species, for all analyses, the minimum exon alignment identity was fixed to 50%.

The PseudoIndex values for each tested species, and the time of execution, in seconds, were recorded for all the 180 analyses. The raw data obtained for the functional gene category is shown in Supplementary Table S1 that includes, for each tested gene and corresponding human isoform, the included target species and the PseudoIndex attributed to each according to the similarity scoring scheme tested. Similarly, Supplementary Table S2 summarizes the raw data associated with the three sets of 30 analysis performed with the lost gene set. For all the executed analyses, the elapsed time in seconds is also presented.

A set of P observations, in which P corresponds to the sum of the number of target species included in each composing analysis, was assigned to each of the six sets of analysis. To each observation, in other words, each species, corresponds a given PseudoIndex. Regarding functional genes, each of the three sets displays 479 observations (Supplementary Table S1) while each set from the lost gene category contains a total of 155 observations (Supplementary Table S2).

Evaluation metrics

For each tested similarity scoring scheme available in the Pseudo*Checker*'s coding sequence prediction step, we evaluated the percentage of correctly categorized gene functionality events, defined as the presumable presence of a viable gene in a given species, as well as the percentage of correctly categorized gene loss events, defined as the loss of a given gene in a given species.

Regarding the first metric, taking advantage of each computed PseudoIndex per species (or observation), for each set of 30 analysis involving functional genes from tested target species, we calculated the fraction of predicted gene functionality events by Pseudo*Checker* over the actual tested

Table 4. Percentage value of discarded observations per tested similarity scoring scheme and type of analysis (involving lost genes or functional ones). (CR): Optimized for closely related species; (SD): Optimized for slightly divergent species; (BF): Best-fit similarity scoring scheme

	Lost genes			Functional genes		
Similarity scoring scheme	CR	SD	BF	CR	SD	BF
Discarded observations	6	2	3	7	11	5
Total observations	155	155	155	479	479	479
Discarded observations (%)	3.87	1.29	1.93	1.46	2.29	1.04

gene functionality events. The percentage of well classified gene functionality events (WCGFE) is concretely given by the formula:

$$(\text{WCGFE}) = (\text{PGFE}/\text{AGFE}) * 100$$

where PGFE constitutes the number of **p**redicted **g**ene **f**unctionality **e**vents by Pseudo*Checker* and AGFE corresponds to the number of **a**ctual tested **g**ene **f**unctionality **e**vents.

Similarly, regarding each set of the 30 analyses involving lost genes at the inputted target species, we computed the fraction of the number of predicted gene loss events by Pseudo*Checker* over the number of actual tested gene loss events. This is given as the percentage of well classified gene loss events (WCGLE), expressed by the formula:

$$(\text{WCGLE}) = (\text{PGLE}/\text{AGLE}) * 100$$

where PGLE constitutes the number of **p**redicted **g**ene **l**oss **e**vents by Pseudo*Checker* and AGLE corresponds to the number of **a**ctual tested **g**ene **l**oss **e**vents.

To compute both these ratios, however, we first turned our approach into a binary classifier by converting the PseudoIndex scale into two different categories. Tested species with values of PseudoIndex between 0 and 2 yielded a predicted gene functionality event, whereas values of PseudoIndex equal to 4 or 5, in its turn, corresponded to a predicted gene inactivation event.

Finally, since these represented doubtful cases, species that presented PseudoIndex values equal to 3 were not be considered into Pseudo*Checker*'s classification performance evaluation, being, therefore, discarded.

Results

Prior to computing the final results concerning both mentioned classification evaluation metrics, we first assessed the percentage of discarded observations per tested similarity scoring scheme for both the 30 analyses involving lost genes and the 30 remaining ones involving functional ones (Table 4).

Our results show that percentage of cases discarded from quality evaluation is rather small: ranging from 1.04 to 3.87%. In fact, not only was Pseudo*Checker* capable of producing a low rate of doubtful predictions, but also the removal of such cases will not likely influence our classification quality evaluation outcome.

Once removed the doubtful observations from each set, for each tested similarity scoring scheme, the percentage of well classified gene functionality events (Table 5), as well as the percentage of well classified gene loss events (Table 6) were computed.

Table 5. Percentage value of well classified gene functionality events per tested similarity scoring scheme, including the obtained average value for all of these. (CR): Optimized for closely related species; (SD): Optimized for slightly divergent species; (BF): Best-fit similarity scoring scheme

Similarity scoring scheme	CR	SD	BF
PGFE	433	456	465
AGFE	472	468	474
WCGFE (%)	91.73	97.43	98.10
Average WCGFE (%)		95.75	

Table 6. Percentage value of well classified gene loss events per tested similarity scoring scheme, including the obtained average value for all of these. (CR): Optimized for closely related species; (SD): Optimized for slightly divergent species; (BF): Best-fit similarity scoring scheme

Similarity scoring scheme	CR	SD	BF
PGLE	147	147	149
AGLE	149	153	152
WCGLE (%)	98.65	96.07	98.02
Average WCGLE (%)		97.58	

Looking at the overall classification quality, the Pseudo*Checker* robustness for the identification of functional genes yielded satisfactory results. Particularly, this analysis highlights the importance of the accurate selection of the used similarity scoring scheme for the successful inference of gene functionality. In fact, and as shown in Table 5, the best-fit similarity scoring scheme was the best performer, displaying the highest percentage value of well classified gene functionality events amongst the remaining available similarity scoring schemes.

In contrast, the similarity scoring scheme optimized for closely related species underperformed when compared to the former: with over 90% of correct classifications regarding gene functionality (Table 5). This can be explained by the chosen methodology regarding species selection, which integrated representatives from each mammalian order into the analyses, whenever possible, and used human as a reference species. Thus, and as anticipated, the similarity scoring scheme optimized for closely related species is less suitable when species divergence is increased (see Supplementary Table S1).

The similarity scoring scheme optimized for slightly divergent species, on the other hand, failed in fewer cases when compared to the similarity scoring scheme optimized for closely related species. This is likely due to the more relaxed mismatch penalty, when compared to the similarity scoring scheme optimized for closely related species, making it more tolerant to nucleotide substitutions within produced alignments. Yet, it is not a dynamic similarity scoring scheme, hence limiting a possible increased diversity of different alignments that is offered by the best-fit similarity scoring scheme.

As for Pseudo*Checker*'s robustness towards the classification of lost genes, a high percentage value of well classified lost genes was obtained, with the three similarity scoring schemes yielding similar results (Table 6).

Further inspection of the reduced number of misclassified cases (Table 6) revealed that these resulted either from a relaxed sequence prediction, possibly allowing for mismatch acceptance instead of frameshift gap openings (the case of similarity scoring scheme optimized for slightly divergent species), thus, erroneously predicting pseudogenes as coding genes, and/or due to different criteria applied by the authors that originally reported the gene as lost, notably regarding the penalization of splice site inactivating mutations, which, as aforementioned, are not considered in our PseudoIndex calculation. Since different approaches rely on distinct assumptions, display different biases and are likely to make use of different criteria to assign a given genomic fragment as a pseudogene, they occasionally yield divergent results, rendering difficult a systematic resolution of all the conflicts between such methods (42).

Additionally, the splice-site and reading-frame aware best-fit similarity scoring scheme underperformed the one optimized for closely related species, explained by the possibility of erroneous sequence prediction accomplished by this similarity scoring scheme, due to, again, the diversity of species included in the analyses. Such event could lead to the accumulation of false numerous and premature deleterious mutations, arising, for example, from bad exon boundaries predictions, culminating in the inference of a gene inactivation event, not due to the detection of real mutations, but by the presence of spurious frameshifts and/or in-frame stop codons, overestimating the level of gene erosion.

Yet, and considering the overall results, Pseudo*Checker* has shown to be suitable for accurately classifying genuine disabled genes by ORF-disrupting mutations, as well as predicted functional genes. Moreover, the embedded PseudoIndex metric displayed the required robustness and calibration to perform such binary classification. Finally, considering the temporal information recorded for each ran analysis, we achieved a very reasonable average value of 133.09 seconds per analysis (∼2.22 min), confirming its quickness towards gene inactivation inference.

CONCLUSION

To build a complete pseudogene database, it is ultimately required to develop computational approaches capable of reliably identifying and mapping gene inactivation over a phylogenetic blueprint. Overall, we suggest that an advance towards this aim was established with the development of Pseudo*Checker*. Our approach is designed to be repeatedly applied, by any, even non-experienced users in researches directed to identify and unveil the molecular signatures underlying gene inactivation occurrences in a straightforward, convenient, and highly accurate process, meeting the emerging need of an integrated analysis tool for this field of evolutionary biology.

IMPLEMENTATION

Pseudo*Checker* is hosted within a PHP/Apache environment under a Linux-based system equipped with 16 processing cores, 32 GB of RAM, with each job running as a single process. The front-end system is implemented in HTML, CSS and Javascript, while the back-end pipeline is built over Python 3. MACSE v2, the unique standalone software used in our pipeline, resorts of Java.

The source code of Pseudo*Checker* will be made available upon request.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

Author contributions: R.R. and L.F.C.C. conceived this study. L.Q.A., and P.R. conceived the webserver. L.Q.A., M.M.F., M.L-M., P.R., R.R. and L.F.C.C. designed the bioinformatics pipeline and constructed the web server. L.Q.A., R.R and L.F.C.C. wrote the manuscript with contributions from all co-authors.

FUNDING

This research was funded by COMPETE 2020, Portugal 2020 and the European Union through the ERDF, grant number 32030, and by FCT through national funds (PTDC/BIA-EUL/32030/2017).
Conflict of interest statement. None declared.

REFERENCES

1. Sharma,V., Hecker,N., Roscito,J.G., Foerster,L., Langer,B.E. and Hiller,M. (2018) A genomics approach reveals insights into the importance of gene losses for mammalian adaptations. *Nat. Commun.*, **9**, 1215.
2. Albalat,R. and Cañestro,C. (2016) Evolution by gene loss. *Nat. Rev. Genet.*, **17**, 379.
3. Mighell,A., Smith,N., Robinson,P. and Markham,A. (2000) Vertebrate pseudogenes. *FEBS Lett.*, **468**, 109–114.
4. Rouchka,E.C. and Cha,I.E. (2009) Current trends in pseudogene detection and characterization. *Curr. Bioinform.*, **4**, 112–119.
5. Cooke,S.L., Shlien,A., Marshall,J., Pipinikas,C.P., Martincorena,I., Tubio,J.M., Li,Y., Menzies,A., Mudie,L. and Ramakrishna,M.J.N.c. (2014) Processed pseudogenes acquired somatically during cancer development. *Nat. Commun.*, **5**, 3644.
6. Lopes-Marques,M., Machado,A.M., Barbosa,S., Fonseca,M.M., Ruivo,R. and Castro,L.F.C. (2018) Cetacea are natural knockouts for IL20. *Immunogenetics*, **70**, 681–687.
7. Hecker,N., Sharma,V. and Hiller,M. (2017) Transition to an aquatic habitat permitted the repeated loss of the pleiotropic KLK8 gene in mammals. *Genome Biol. Evol.*, **9**, 3179–3188.
8. Sharma,V., Lehmann,T., Stuckas,H., Funke,L. and Hiller,M. (2018) Loss of RXFP2 and INSL3 genes in Afrotheria shows that testicular descent is the ancestral condition in placental mammals. *PLoS Biol.*, **16**, https://doi.org/10.1371/journal.pbio.2005293.
9. Lee,J.-H., Lewis,K.M., Moural,T.W., Kirilenko,B., Borgonovo,B., Prange,G., Koessl,M., Huggenberger,S., Kang,C. and Hiller,M. (2018) Molecular parallelism in fast-twitch muscle proteins in echolocating mammals. *Sci. Adv.*, **4**, eaat9660.
10. Jebb,D. and Hiller,M. (2018) Recurrent loss of HMGCS2 shows that ketogenesis is not essential for the evolution of large mammalian brains. *Elife*, **7**, e38906.
11. Sharma,V. and Hiller,M. (2018) Loss of enzymes in the bile acid synthesis pathway explains differences in bile composition among mammals. *Genome Biol. Evol.*, **10**, 3211–3217.
12. Hecker,N., Sharma,V. and Hiller,M. (2019) Convergent gene losses illuminate metabolic and physiological changes in herbivores and carnivores. *Proc. Natl. Acad. Sci.*, **116**, 3036–3041.
13. Hecker,N., Lächele,U., Stuckas,H., Giere,P. and Hiller,M. (2019) Convergent vomeronasal system reduction in mammals coincides with convergent losses of calcium signalling and odorant-degrading genes. *Mol. Ecol.*, **28**, 3656–3668.

14. Huelsmann,M., Hecker,N., Springer,M.S., Gatesy,J., Sharma,V. and Hiller,M. (2019) Genes lost during the transition from land to water in cetaceans highlight genomic changes associated with aquatic adaptations. *Sci. Adv.*, **5**, eaaw6671.
15. Sharma,V. and Hiller,M. (2020) Losses of human disease-associated genes in placental mammals. *NAR Genomics Bioinform.*, **2**, lqz012.
16. Guijarro-Clarke,C., Holland,P.W. and Paps,J. (2020) Widespread patterns of gene loss in the evolution of the animal kingdom. *Nat. Ecol. Evol.*, **4**, 519–523.
17. Springer,M.S., Emerling,C.A., Gatesy,J., Randall,J., Collin,M.A., Hecker,N., Hiller,M. and Delsuc,F. (2019) Odontogenic ameloblast-associated (ODAM) is inactivated in toothless/enamelless placental mammals and toothed whales. *BMC Evol. Biol.*, **19**, 31.
18. Wang,X., Grus,W.E. and Zhang,J. (2006) Gene losses during human origins. *PLoS Biol.*, **4**, e52.
19. Lopes-Marques,M., Machado,A.M., Alves,L.Q., Fonseca,M.M., Barbosa,S., Sinding,M.-H.S., Rasmussen,M.H., Iversen,M.R., Frost Bertelsen,M. and Campos,P.F. (2019) Complete inactivation of sebum-producing genes parallels the loss of sebaceous glands in Cetacea. *Mol. Biol. Evol.*, **36**, 1270–1280.
20. Lopes-Marques,M., Ruivo,R., Alves,L.Q., Sousa,N., Machado,A.M. and Castro,L.F.C. (2019) The singularity of Cetacea behavior parallels the complete inactivation of melatonin gene modules. *Genes*, **10**, 121.
21. Lopes-Marques,M., Alves,L.Q., Fonseca,M.M., Secci-Petretto,G., Machado,A.M., Ruivo,R. and Castro,L.F.C. (2019) Convergent inactivation of the skin-specific CC motif chemokine ligand 27 in mammalian evolution. *Immunogenetics*, **71**, 363–372.
22. Alves,L.Q., Alves,J., Ribeiro,R., Ruivo,R. and Castro,F. (2019) The dopamine receptor D5 gene shows signs of independent erosion in toothed and baleen whales. *Peer J.*, **7**, e7758.
23. Olson,M.V. (1999) When less is more: gene loss as an engine of evolutionary change. *Am. J. Human Genet.*, **64**, 18–23.
24. Emerling,C.A., Delsuc,F. and Nachman,M.W.J.S.a. (2018) Chitinase genes (CHIAs) provide genomic footprints of a post-Cretaceous dietary radiation in placental mammals. *Sci. Adv.*, **4**, eaar6478.
25. Gaudry,M.J., Jastroch,M., Treberg,J.R., Hofreiter,M., Paijmans,J.L., Starrett,J., Wales,N., Signore,A.V., Springer,M.S. and Campbell,K.L.J.S.a. (2017) Inactivation of thermogenic UCP1 as a historical contingency in multiple placental mammal clades. *Sci. Adv.*, **3**, e1602878.
26. Meredith,R.W., Zhang,G., Gilbert,M.T.P., Jarvis,E.D. and Springer,M.S.J.S. (2014) Evidence for a single loss of mineralized teeth in the common avian ancestor. *Science*, **346**, 1254390.
27. Meyer,W.K., Jamison,J., Richter,R., Woods,S.E., Partha,R., Kowalczyk,A., Kronk,C., Chikina,M., Bonde,R.K. and Crocker,D.E.J.S. (2018) Ancient convergent losses of Paraoxonase 1 yield potential risks for modern marine mammals. *Science*, **361**, 591–594.
28. Jarvis,E.D., Mirarab,S., Aberer,A.J., Li,B., Houde,P., Li,C., Ho,S.Y., Faircloth,B.C., Nabholz,B. and Howard,J.T. (2014) Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science*, **346**, 1320–1331.
29. Scientists,G.K.C.o. (2009) Genome 10K: a proposal to obtain whole-genome sequence for 10 000 vertebrate species. *J. Hered.*, **100**, 659–674.
30. Robinson,G.E., Hackett,K.J., Purcell-Miramontes,M., Brown,S.J., Evans,J.D., Goldsmith,M.R., Lawson,D., Okamuro,J., Robertson,H.M. and Schneider,D.J. (2011) Creating a buzz about insect genomes. *Science*, **331**, 1386–1386.
31. Emerling,C.A., Widjaja,A.D., Nguyen,N.N. and Springer,M.S. (2017) Their loss is our gain: regressive evolution in vertebrates provides genomic models for uncovering human disease loci. *J. Med. Genet.*, **54**, 787–794.
32. Baertsch,R., Diekhans,M., Kent,W.J., Haussler,D. and Brosius,J. (2008) Retrocopy contributions to the evolution of the human genome. *BMC Genomics*, **9**, 466.
33. Zhang,Z., Carriero,N., Zheng,D., Karro,J., Harrison,P.M. and Gerstein,M. (2006) PseudoPipe: an automated pseudogene identification pipeline. *Bioinformatics*, **22**, 1437–1439.
34. van Baren,M.J. and Brent,M.R. (2006) Iterative gene prediction and pseudogene removal improves genome annotation. *Genome Res.*, **16**, 678–685.
35. Zhang,Z.D., Frankish,A., Hunt,T., Harrow,J. and Gerstein,M. (2010) Identification and analysis of unitary pseudogenes: historic and contemporary gene losses in humans and other primates. *Genome Biol.*, **11**, R26.
36. Sharma,V., Elghafari,A. and Hiller,M. (2016) Coding exon-structure aware realigner (CESAR) utilizes genome alignments for accurate comparative gene annotation. *Nucleic Acids Res.*, **44**, e103.
37. Needleman,S.B. and Wunsch,C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, **48**, 443–453.
38. Ranwez,V., Douzery,E.J., Cambon,C., Chantret,N. and Delsuc,F.J.M.B. (2018) MACSE v2: toolkit for the alignment of coding sequences accounting for frameshifts and stop codons. *Mol. Biol. Evol.*, **35**, 2582–2584.
39. Barrett,C.F., Sinn,B.T. and Kennedy,A.H. (2019) Unprecedented parallel photosynthetic losses in a heterotrophic orchid genus. *Mol. Biol. Evol.*, **36**, 1884–1901.
40. Uhlén,M., Fagerberg,L., Hallström,B.M., Lindskog,C., Oksvold,P., Mardinoglu,A., Sivertsson,Å., Kampf,C., Sjöstedt,E. and Asplund,A. (2015) Tissue-based map of the human proteome. *Science*, **347**, 1260419.
41. Burgin,C.J., Colella,J.P., Kahn,P.L. and Upham,N.S. (2018) How many species of mammals are there? *J. Mammal.*, **99**, 1–14.
42. Karro,J.E., Yan,Y., Zheng,D., Zhang,Z., Carriero,N., Cayting,P., Harrrison,P. and Gerstein,M. (2007) Pseudogene. org: a comprehensive database and comparison platform for pseudogene annotation. *Nucleic Acids Res.*, **35**, D55–D60.