

Sequence analysis

PseudoChecker2 and PseudoViz: automation and visualization of gene loss in the Genome Era

Rui Resende-Pinto^{1,2,*} , Raquel Ruivo¹, Josefin Stiller³, Rute Fonseca⁴ ,
Luís Filipe C. Castro^{1,2,*} 

¹CIMAR/CIIMAR – Interdisciplinary Centre of Marine and Environmental Research, University of Porto, 4450-208 Matosinhos, Portugal

²FCUP – Department of Biology, Faculty of Sciences, University of Porto (U. Porto), 4169-007 Porto, Portugal

³Department of Biology, University of Copenhagen, 2100 Copenhagen, Denmark

⁴Center for Global Mountain Biodiversity, GLOBE Institute, University of Copenhagen, 2100 Copenhagen, Denmark

*Corresponding authors. Rui Resende-Pinto, CIMAR/CIIMAR – Interdisciplinary Centre of Marine and Environmental Research, University of Porto, Avenida General Norton de Matos, S/N, 4450-208 Matosinhos, Portugal. E-mail: rui.pinto@ciimar.up.pt; Luís Filipe C. Castro, CIMAR/CIIMAR – Interdisciplinary Centre of Marine and Environmental Research, University of Porto, Avenida General Norton de Matos, S/N, 4450-208 Matosinhos, Portugal. E-mail: filipe.castro@ciimar.up.pt

Associate Editor: Michael DeGiorgio

Abstract

Summary: High-fidelity genome assemblies provide unprecedented opportunities to decipher mechanisms of molecular evolution and phenotype landscapes. Here, we present PseudoChecker2, a command-line version of the web-tool PseudoChecker with expanded functions. It identifies gene loss via drastic mutational events such as premature stop codons, deletions and insertions. It enables the investigation of cross-species genomic datasets through: (i) integration into automated workflows, (ii) multiprocessing capability, and (iii) creation of a functional reference from annotation files. In addition, we introduce PseudoViz, a novel graphical interface designed to help interpret the results of PseudoChecker2 with intuitive visualizations. These tools combine the versatility and automation of a command-line tool with the user-friendliness of a graphical interface to tackle the challenges of the Genome Era.

Availability and implementation: PseudoChecker2 and PseudoViz are fully available at <https://github.com/rresendepinto/PseudoChecker2> and <https://github.com/rresendepinto/PseudoViz>.

1 Introduction

Gene inactivation is a widespread process with a dramatic influence on evolutionary change (Albalat and Cañestro 2016). The phenotypic consequences of lineage-specific events of loss have been reported across diverse animal lineages, leading to noticeable selective advantages or adaptations in several cases (Jeffery 2009, Castro *et al.* 2014, Emerling and Springer 2014). Genome initiatives aimed at creating large scale datasets have gained significant momentum in recent years, such as the Bird 10K (Feng *et al.* 2020), the Genome 10K (Rhie *et al.* 2021), and the Earth BioGenome (Lewin *et al.* 2018) projects. Concurrently, large repositories such as National Center for Biotechnology Information's (NCBI) RefSeq (Pruitt *et al.* 2005) and Ensembl (Birney *et al.* 2004) have made the data generated by these projects accessible. This wealth of genomic data, coupled with more efficient bioinformatic tools, has improved the inference of gene loss (e.g. Kirilenko *et al.* 2023).

The PseudoChecker (<http://pseudochecker.ciimar.up.pt/pseudochecker/>) web-tool was first designed to infer the coding status of a given candidate gene in a target species using an orthologous coding sequence as the reference (Alves *et al.* 2020). This tool was designed to be accessible to all users, particularly those without programming knowledge or familiarity with

command-line interfaces. However, while PseudoChecker offered a straightforward method for pseudogene detection with minimal user input, its use on large datasets (i.e. >20 sequences) proved intractable. In effect, the PseudoChecker web-server is not indicated for carrying out large Multiple Alignment of Coding Sequences (MACSE) (Ranwez *et al.* 2011) alignments (Alves *et al.* 2020).

Remaining available pipelines for pseudogene inference (e.g. van Baren and Brent 2006, Zhang *et al.* 2006, Baertsch *et al.* 2008, Syberg-Olsen *et al.* 2021, Kirilenko *et al.* 2023) have other constraints, such as being hard to integrate with other tools or unsuitable for assessing the coding status of genes across distant eukaryotic lineages. To circumvent these restrictions, we developed a pipeline that infers mutations based on exon alignment, allowing independent search for mutations on each species.

Here, we present PseudoChecker2, an improved version of the previous tool. By creating a command-line version of the tool, we aim to expand PseudoChecker's capabilities for large datasets, while enabling the integration with other tools and adding additional features. To maintain the intuitiveness and ease of visualization of the first version, we developed PseudoViz, a tool to display the results of PseudoChecker2 in a genome browser.

Received: May 5, 2025; Revised: July 9, 2025; Editorial Decision: July 27, 2025

© The Author(s) 2025. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

2 Application description

PseudoChecker2 is written in Python3, a programming language widely used by the scientific community due to its accessibility, versatility and vast ecosystem of frameworks and libraries. The input for PseudoChecker2 is the same of the original web-tool and consists of the following in a fasta format: (i) the target genomic region, (ii) individual sequences of the targeted exons, and (iii) the coding sequence for the corresponding gene from a reference (Fig. 1). This reference can now be created automatically from genome annotations (in GTF/GFF format) by leveraging the AGAT v1.0.0 tool (Dainat 2023). Due to the alignment-based nature of the software, it is advisable to avoid regions containing duplicates of the target ortholog or references with similar exons as these may lead to incorrect sequence alignments.

To annotate the orthologous exons in the target genomic sequences, PseudoChecker2 uses the Needleman-Wunsch algorithm (Needleman and Wunsch 1970) implemented in the Emboss module of the Biopython library (Cock *et al.* 2009). By using the best-fit similarity scoring scheme (Alves *et al.* 2020), PseudoChecker2 tests a combination of match/mismatch alignment punctuating schemes (Needleman and Wunsch 1970) per reference exon. Under this approach, tested alignments will then be sorted by splice site conservation in the target sequence, where alignments that present viable splice sites will be given preference, followed by percent similarity to the reference sequence, and the best scoring alignment will be selected. Frameshift mutations and premature termination codons are then inferred from exon alignments. The PseudoIndex metric (Alves *et al.* 2020) is

calculated based on the percentage of the ortholog's coding frame that is disrupted due to these mutations or due to perceived absence of one or more orthologous exons. This metric can be used to infer the overall state of sequence erosion of an ortholog.

The output from PseudoChecker2 can be visualized in PseudoViz, a Flask (Grinberg 2014) application that is designed to mimic and improve upon the user experience from the first PseudoChecker version. The Flask package was chosen due to its simple nature, providing an accessible, flexible and lightweight framework. PseudoViz is easily installed and used on a desktop computer, and its results can be displayed in a browser. PseudoViz highlights the main results from PseudoChecker2, enabling a better understanding of potentially inactivating mutations. PseudoViz requires minimal user input (i.e. specific output files from the PseudoChecker analysis; Fig. 1; Fig. 1, available as [supplementary data](#) at *Bioinformatics Advances* online). The user will then be presented with exon alignments from multiple target sequences against a single reference exon (Fig. 2, available as [supplementary data](#) at *Bioinformatics Advances* online). Under each target display, the exon alignment percent identity of the target exon relative to the reference exon is shown, as well as the start and end of the aligned target exon in the respective genomic sequence. By clicking on the “Region” link of the target displays in this tab, the user is redirected to a tab depicting the position of exons in the genomic region, with exon sequences highlighted in blue within the target sequences (Fig. 3, available as [supplementary data](#) at *Bioinformatics Advances* online). In another page, the user can view a multiple sequence alignment with MACSE

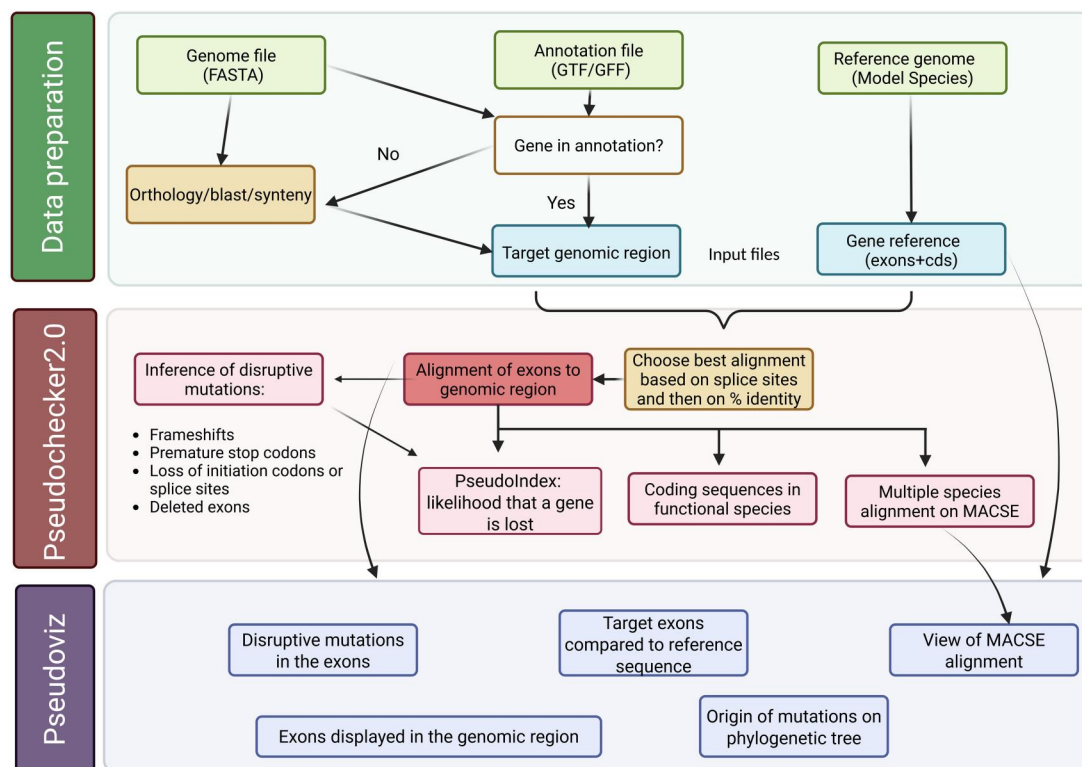


Figure 1. Overview of the input and analysis steps of PseudoChecker 2.0 and PseudoViz. The process begins with data preparation, involving genome, annotation, and reference files to define the target genomic region and extract gene reference sequences. Exons are aligned to the genomic region and the best alignments are selected based on splice sites and identity. Disruptive mutations (e.g. frameshifts, stop codons, splice site losses) are inferred, contributing to a PseudoIndex estimating the likelihood of gene loss. The final step, Pseudoviz, provides visualization of results, including exon mapping, mutation annotation, MACSE alignments, and phylogenetic context. Created with BioRender.com.

(Ranwez *et al.* 2011) (Fig. 4, available as [supplementary data](#) at *Bioinformatics Advances* online). This alignment is optional, since it becomes exponentially slower with larger amounts of target genomic sequences. Alternatively, after the exon alignments are completed, multiple sequence alignments for groups of species/targets of interest can be created (e.g. one monophyletic clade in which several species have inactivating mutations).

PseudoViz also offers a way of visualizing the mutational events in a phylogenetic context by displaying a dendrogram with information on the mutations and PseudoIndex metric (Fig. 5, available as [supplementary data](#) at *Bioinformatics Advances* online). Node colors correspond to the assigned PseudoIndex metric. The PseudoIndex values of the terminal nodes correspond to the PseudoIndex of the target sequence represented. The PseudoIndex of the middle nodes corresponds to the maximum PseudoIndex of the descendant nodes. This provides an easy overview of the erosion of the gene along a given phylogeny. When hovering over the nodes, a small text box will display the present mutations in the target or, in the case of the middle nodes, the mutations that are present in all its descendant nodes.

2.1 Installation and availability of both tools

To facilitate the accessibility and usability of the developed tools, we have provided multiple installation methods to accommodate various user preferences and system requirements. Both tools can be readily installed through GitHub, by cloning the repositories. Once cloned, the dependencies for each tool can be seamlessly installed using Conda. This method ensures an isolated environment with all the required dependencies, avoiding any conflicts with existing packages. Alternatively, for those who prefer containerized applications, we have provided Docker images for both tools. These images encapsulate the tools and their dependencies, ensuring consistent performance across various computing environments. The Docker images can be pulled with docker or singularity, which are popular container solutions in high-performance computing environments. Users can then run the tools in containers, which are isolated from the host system, thus preventing any dependency-related issues. These installation methods ensure flexibility and ease of use, enabling researchers to quickly deploy and utilize the tools for scientific research.

2.2 Application: inactivation of the *CYP2J19* gene in birds

To test PseudoChecker2 and PseudoViz, we assessed the coding status of *CYP2J19* (Figs 2–5, available as [supplementary data](#) at *Bioinformatics Advances* online), a gene that is responsible for the production of red retinal oil droplets in birds and turtles (Mundy *et al.* 2016, Twyman *et al.* 2016) and has been described as lost in penguins, owls and kiwis (Emerling 2018). These bird lineages display adaptations to dim light, such as having low or negligible amounts of red oil droplets (Gondo and Ando 1995). We were able to retrieve inactivating mutations in the same lineages as Emerling (2018), namely in penguins, owls and kiwis. Of note, we also identified previously unreported cases of gene inactivating mutations in the barred owl-nightjar *Aegotheles bennetti* and the oilbird *Steatornis caripensis* (Strisores), which are nocturnal birds and hence adapted to dim-light conditions (explored in more detail in File 1, available as [supplementary](#)

[data](#) at *Bioinformatics Advances* online). Although the mutation in the oilbird is potentially nullified by a change of the canonical start codon, the mutation in the nightjar is more likely to be disruptive as it consists of a premature stop codon in the middle of the coding sequence (Fig. 7, available as [supplementary data](#) at *Bioinformatics Advances* online). Nevertheless, these would represent two independent events because oilbirds and owl-nightjars are not closely related within Strisores (Chen *et al.* 2019, Stiller *et al.* 2024) and we did not find inactivating mutations in any of the other Strisores. This includes all other 10 Strisores genomes from the B10K dataset (Feng *et al.* 2020) and the European nightjar *Caprimulgus europaeus* genome from the DTOL project (GCA_907165065.1). The completeness of the *CYP2J19* gene in *C. europaeus* is in agreement with Twyman *et al.* (2018) and the fact that red-oil droplets are present in the retina of the related jungle nightjar *C. indicus* (Gondo and Ando 1995).

3 Concluding remarks

The goal of PseudoChecker2 is to standardize and facilitate pseudogene detection and to allow the visualization of the coding status of a single gene across a phylogenetic tree using PseudoViz. This approach aims to enable the discovery of patterns of gene loss and associate them with known phenotypes. This phylogenetic profiling is useful to not only to understand genomic adaptations of species to their respective ecological niches but can also help recognizing poorly understood functions of a single protein or protein interactions that underlie specific phenotypes (Date and Peregrín-Alvarez 2008). Lastly, its potential applications extend beyond comparative genomics, because it can be used to identify population-specific variants from the growing population genomics datasets. By incorporating PseudoChecker2 and PseudoViz into genomic studies, researchers can gain a more comprehensive understanding of gene dynamics across diverging lineages and populations.

Supplementary data

[Supplementary data](#) are available at *Bioinformatics Advances* online.

Conflict of interest

None declared.

Funding

This work was supported by Fundação para a Ciência e Tecnologia [PhD fellowship to RRP (2020.08608.BD)] and VILLUM FONDEN [research grant no 42153 to J.S.; Center for Global Mountain Biodiversity grant no 25925 to R.F.].

Data availability

PseudoChecker2 and PseudoViz, as well as example data pertaining to *CYP2J19*, are fully available at <https://github.com/rresendepinto/PseudoChecker2> and <https://github.com/rresendepinto/PseudoViz>.

References

- Albalat R, Cañestro C. Evolution by gene loss. *Nat Rev Genet* 2016; **17**:379–91.
- Alves LQ, Ruivo R, Fonseca MM *et al.* PseudoChecker: an integrated online platform for gene inactivation inference. *Nucleic Acids Res* 2020; **48**:W321–31.
- Baertsch R, Diekhans M, Kent WJ *et al.* Retrocopy contributions to the evolution of the human genome. *BMC Genomics* 2008; **9**:466.
- Birney E, Andrews TD, Bevan P *et al.* An overview of ensembl. *Genome Res* 2004; **14**:925–8.
- Castro LFC, Gonçalves O, Mazan S *et al.* Recurrent gene loss correlates with the evolution of stomach phenotypes in gnathostome history. *Proc Biol Sci* 2014; **281**:20132669.
- Chen A, White ND, Benson RBJ *et al.* Total-evidence framework reveals complex morphological evolution in nightbirds (Strisores). *Diversity* 2019; **11**:143.
- Cock PJA, Antao T, Chang JT *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* 2009; **25**:1422–3.
- Dainat J. Agat: another Gff analysis toolkit to handle annotations in any GTF/GFF format. [Computer software]. Zenodo. [10.5281/zenodo.4205393](https://doi.org/10.5281/zenodo.4205393). 2023.
- Date SV, Peregrín-Alvarez JM. Phylogenetic profiling. *Methods Mol Biol Clifton NJ* 2008; **453**:201–16.
- Emerling CA. Independent pseudogenization of CYP2J19 in penguins, owls and kiwis implicates gene in red carotenoid synthesis. *Mol Phylogenet Evol* 2018; **118**:47–53.
- Emerling CA, Springer MS. Eyes underground: regression of visual protein networks in subterranean mammals. *Mol Phylogenet Evol* 2014; **78**:260–70.
- Feng S, Stiller J, Deng Y *et al.* Dense sampling of bird diversity increases power of comparative genomics. *Nature* 2020; **587**:252–7.
- Gondo M, Ando H. Comparative and histophysiological study of oil droplets in the avian retina. *Kobe J Med Sci* 1995; **41**:127–39.
- Grinberg M. *Flask Web Development: Developing Web Applications with Python*. Sebastopol, CA, USA: O'Reilly Media, Inc., 2014.
- Jeffery WR. Regressive evolution in *Astyanax* cavefish. *Annu Rev Genet* 2009; **43**:25–47.
- Kirilenko BM, Munegowda C, Osipova E *et al.*; Zoonomia Consortium. Integrating gene annotation with orthology inference at scale. *Science* 2023; **380**:eabn3107.
- Lewin HA, Robinson GE, Kress WJ *et al.* Earth BioGenome project: sequencing life for the future of life. *Proc Natl Acad Sci USA* 2018; **115**:4325–33.
- Mundy NI, Stapley J, Bennison C *et al.* Red carotenoid coloration in the zebra finch is controlled by a cytochrome P450 gene cluster. *Curr Biol* 2016; **26**:1435–40.
- Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* 1970; **48**:443–53.
- Pruitt KD, Tatusova T, Maglott DR *et al.* NCBI reference sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* 2005; **33**:D501–4.
- Ranwez V, Harispe S, Delsuc F *et al.* MACSE: Multiple Alignment of Coding SEquences accounting for frameshifts and stop codons. *PLoS One* 2011; **6**:e22594.
- Rhie A, McCarthy SA, Fedrigo O *et al.* Towards complete and error-free genome assemblies of all vertebrate species. *Nature* 2021; **592**:737–46.
- Stiller J, Feng S, Chowdhury A-A *et al.* Complexity of avian evolution revealed by family-level genomes. *Nature* 2024; **629**:851–60.
- Syberg-Olsen MJ, Garber AI, Keeling PJ *et al.* Pseudofinder: detection of pseudogenes in prokaryotic genomes. *Mol Biol Evol* 2021; **39**:msac153.
- Twyman H, Andersson S, Mundy NI *et al.* Evolution of CYP2J19, a gene involved in colour vision and red coloration in birds: positive selection in the face of conservation and pleiotropy. *BMC Evol Biol* 2018; **18**:22.
- Twyman H, Valenzuela N, Literman R *et al.* Seeing red to being red: conserved genetic mechanism for red cone oil droplets and co-option for red coloration in birds and turtles. *Proc R Soc B* 2016; **283**:20161208.
- van Baren MJ, Brent MR. Iterative gene prediction and pseudogene removal improves genome annotation. *Genome Res* 2006; **16**:678–85.
- Zhang Z, Carriero N, Zheng D *et al.* PseudoPipe: an automated pseudogene identification pipeline. *Bioinformatics* 2006; **22**:1437–9.