Methods to identify and study the evolution of pseudogenes using a phylogenetic approach

Jacques Dainat, Pierre Pontarotti

Double space

Methods to identify and study the evolution of pseudogenes using a phylogenetic approach

J. Dainat[1*] and P. Pontarotti[2,3]

[1]National Bioinformatics Infrastructure Sweden (NBIS), SciLifeLab, Uppsala Biomedicinska Centrum (BMC), Husargatan 3, S-751 23 Uppsala, Sweden.
[2]Evolutionary Biology team. Aix Marseille Univ IRD, APHM, MEPHI, IHU Méditerranée Infection, Marseille France.
[3]SNC5039 CNRS, 19-21 Boulevard Jean Moulin 13005 Marseille, France.

[*]Corresponding author:
Jacques Dainat, PhD

Uppsala Biomedicinska Centrum (BMC)
Department of Medical Biochemistry Microbiology, Genomics
Husargatan 3, S-751 23 Uppsala, Sweden.
+46 18 471 46 25

E-mail: jacques.dainat@gmail.com

Running head: Phylogenetic analysis of pseudogenes

Abstract

The discovery that pseudogenes are involved in important biological processes has excited enthusiasm and increased the research interest on them. An accurate detection and analysis of pseudogenes can be achieved using comparative methods, but only the use of phylogenetic tools can provide accurate information about their birth, their evolution and their death, hence about the impact that they have on genes and genomes. Here, phylogenetic methods that allow studying pseudogene history are described.

Key words: Pseudogenes, phylogeny, speciation, orthology, paralogy

1. Introduction

Genome analyses have shown the presence of numerous pseudogenes in all organisms. For example, *Mus musculus* and *Homo sapiens* both contain around 20 000 pseudogenes [1–4]. Pseudogenes are categorized into three main groups according to the underlying mechanisms involved in their emergence: processed pseudogenes, unprocessed pseudogenes and unitary pseudogenes (**Fig. 1**).

- The first are retrotransposed pseudogenes. They are considered as dead-on-arrival [5], having suffered lethal damage during the copying process or missing important flanking regions [6–8].
- The second are duplicated pseudogenes. They originate from a duplication event of a functional gene. Gene duplication can represent an important source of diversity for organisms [9, 10]. There are four potential outcomes for a duplicated gene:
    - It maintains the original function; this process is known as gene conservation.
    - It acquires a new function by neo-functionalization.
    - It shares the original function with the gene it originated from by sub-functionalization.
    - Not subject to any selection pressure, it accumulates mutations and becomes a pseudogene.
- The third group of pseudogenes is called unitary pseudogenes because they have no functional orthologous counterpart [11]. They originate from the pseudogenization of a gene that has been functional and subjected to selection for a long time period. Indeed, orthologs of such unitary pseudogenes still exists in other species. They are of special interest because their losses may be linked to functional losses.

Non-functional pseudogenes are not subjected to any selection pressure and drift until becoming unrecognizable [12]. They eventually become part of the genomic noise, formerly called “junk” DNA. The divergence rate of pseudogenes in avian and human lineages is estimated to be around 2% by million years [13, 14], therefore the timeframe of a complete disappearance of a pseudogene can be estimated to 50 million years. If negative selection is in place, as it might be assumed in the case of the “less-is-more” hypothesis [15], the pseudogenization process can be even faster. Consequently, after a certain divergence time, it might become impossible to identify the pseudogene on the basis of similarity to its orthologs.

On the contrary, pseudogenes that maintain a transcriptional or even translational activity and are involved in biological processes [16–19] can be preserved by positive selection and evolve like genes.

Currently, the majority of methods to study the life of pseudogenes are based on comparative genomics [20, 21], which analyse sequence similarity and synteny (physical

colocalization of multiple genetic loci on the same chromosome) to deduce the relationship occurring among different sequences. However, when the pseudogenization process has to be studied on a large evolution scale (superior to some million years), it should be considered that the structure of pseudogenic sequences can change compared to that of their parental genes, and the same may apply to synteny. These changes can in turn induce analysis difficulties and miss-interpretations and render comparative genomics a suboptimal approach.

During the last years, the increase in calculation capacity has enabled the use of phylogenetic methods to study robustly and rapidly variable biological entities such as pseudogenes. The phylogenetic methods are more advantageous than the classical methods of comparative genomics and particularly suitable to elucidate pseudogene birth, age and loss. This is because they allow inferring more robust relationships between the studied sequences. For example, a missing gene in a phylogeny is likely to be caused by a pseudogenization event [22]. By deciphering the phylogeny, it is possible to deduce in which ancestor the gene was present and starting from which ancestor the gene was lost. The analysis of these features (presence/absence) by the parsimony method [23–26], allows determining the appearance period of the pseudogene, which corresponds to the period of gene deactivation. Another case that exemplifies the use of phylogeny in the study of pseudogenes is represented by duplication events occurring before a speciation. Such events give rise to co-orthologs. The pseudogenization of one of the two duplicated genes should not lead to the loss of the function of the gene, due to the presence of a still functional copy. Consequently, this kind of clue allows excluding the hypothesis that the pseudogene is of the unitary kind.

In this chapter, we describe how phylogenetic methods can be applied to the identification and characterization of pseudogenes.

2. Methods

2.1. Pseudogene identification using phylogenetic tools

To identify pseudogenes the strategy is to look for sequences similar to functional genes and containing disrupting mutations. It can be performed “manually” using the sequence of the gene of interest as query and searching for similar sequences with a Blast tool (BlastP, TBlastN, BlastZ, etc)[27, 28]. Among the retrieved sequences, those not annotated as gene or/and having deleterious mutations that disrupt the open reading frame (ORF) can be considered as pseudogenes. Some pipelines automate that principle [2, 29, 30]. This approach does not allow to determine precisely the relationship between the sequence used as a query and the putative pseudogenic sequences detected. Excluded rare cases of convergence, those sequences are either orthologs or paralogs. This information is crucial

for meaningful interpretation of the pseudogenization process. Moreover, there is no way to infer with precision temporality of the event.

A phylogenetic approach has the benefit to determine precisely the relationship of the studied sequences and the relative period when a pseudogenisation event has occurred. This relative period is determined by the speciation event of the species used for the phylogeny (*see* **Notes 1**).

The first step for a phylogenetic approach would naturally consist in building a phylogenetic tree of the gene or gene family of interest (*see* **Notes 1-4**).

Then, reading the phylogeny we can detect the missing genes and consequently the putative gene loss events (*see* **Note 5**). The putative losses are found by comparing the species tree of the chosen species with the phylogenetic tree of the gene of interest in those species (**Fig. 2**). The phylogenetic tree analysis can be done using software such as Phylopattern, Bio::Phylo and DendroPy [31–33].

When a gene is detected as missing in the genome of a species, the third step consists in searching for putative homologous sequences of that gene by a genome-wide BLAST. To optimize the chance to detect the sequence, the Blast should be performed using an orthologous gene retrieved from the phylogeny and such gene should be as close as possible to the investigated species. In the example reported in **Figure 2**, the loss of the gene of interest in *H. sapiens* should be investigated using the orthologous gene in *M. mulatta*. When no Blast hit is found, considering the genome assembly of the species as complete we can conclude that the gene has likely become genomic noise. But the available genome assemblies have variable numbers of missing regions, so any interpretation should be cautious. On the contrary, when a hit is found, its homology needs to be verified.

In the fourth step, we try to confirm the homology of the sequence retrieved in the previous step. The verification can be performed either by a synteny study or by a phylogeny study. To be orthologous by synteny, the retrieved sequence has to be located at the same position in the genome as the orthologous sequence use as query. Furthermore, the neighbouring genes should be the same. However, if distant species are compared in the synteny study (old speciation event), it is possible that the synteny has changed during evolution. In this respect, the phylogeny study ensures a higher grade of accuracy in inferring orthology, while excluding paralogy, or no homology (just similarity). This can be of special importance when looking at unitary pseudogenes (**Fig. 3**). To verify the relationship of the retrieved sequence with the sequence used as query, a new phylogeny must be built (*see* **Note 6**). Then, the orthologs of the retrieved sequence must be identified in this new phylogeny and compared with the orthologs defined in the old one (*see* **Note 7**). This can be done using software allowing reading phylogenies [31–33]. If both analyses give common orthologs, then the orthology of the Blast hit sequence is confirmed (*see* **Notes 8 and 9**).

Once the homology is defined, the Blast hit sequence should be subjected to in depth study (see **Note 10**). Two scenarios can occur. In the first case scenario, the sequence is already annotated within the genome: i) it is a gene present in the database used, but its sequence was not kept in the phylogeny building process; ii) it is a gene not present in the database used, but present in other databases; iii) it is indeed a pseudogene already described (see **Note 11**). In the second case scenario, the sequence is not annotated: i) it is a gene forgotten during the genome annotation process; ii) it is indeed an unannotated pseudogene. In reaching these conclusions, a more thorough investigation must be carried on (e.g., a gene prediction, ORF sanity, dN/dS ratio, use of RNAseq to check transcription and the sanity of the transcribed sequence). When a deleterious mutation is detected, it can be concluded that the sequence is a pseudogene. For ancient pseudogenes, the disruptions are easy to detect, but this can be more difficult when only few disruptive mutations are presents. This can be particularly tricky because genes exist with stop-codon read-through [34–36]. The mutations searched for can be of various types: micro-mutations, such as substitutions or indels that engender the appearance of premature stop codons in the ORF (**Fig. 4**), but also macro-mutations that engender the loss of entire exons of the original sequence. Macro-mutations can be the result of a series of micro-mutations or of a recombination event (*see* **Notes 12 and 13**).

2.2. Pseudogene dating and characterization using phylogenetic tools

Once a pseudogene has been identified, the phylogeny approach allows getting insights into its evolutionary history and in particular on the time frame of the pseudogenization events. If there are several pseudogenes in the gene family studied, putting them together in the same phylogeny can improve the accuracy of the study. Thanks to the use of parsimony methods [23–26], it is in fact possible to reconstruct the ancestral state starting from the contemporary states that are present on the leaves of the phylogeny. In the example reported in **Figure 5**, the parsimony reconstruction allows to define three pseudogenization events for four pseudogenes. The first pseudogenization event (P1) occurred in *R. norvegicus* after the split with *M. musculus* and produced a unitary pseudogene that has no functional co-ortholog counterpart. The second pseudogenization event (P2) is specific of *M. mulatta.* The third pseudogenization event (P3) is shared between *H. sapiens* and *M. mulatta* and concerns the Catarrhini phylum. Further analyses can be performed to infer the retrotransposed or duplicated nature of the pseudogenes produced by P2 and P3. Indeed, the loss of introns compared to the ancestral structure of the gene and the trace of a polyadenylation tail are evidence to conclude that a pseudogene arose by retrotransposition.

2.3. Pseudogene analysis at the nucleotide level using phylogenetic tools

The mutation analysis and its interpretation in an evolutionary perspective allow the analysis of the pseudogenization process.

As a first step, all the mutations occurred in the pseudogenic sequence must be identified. This can be done by performing a comparative analysis between the nucleotide sequence of the pseudogene under study and that of the corresponding functional gene(s) (**Fig. 6**). The functional reference sequence may be selected among the available orthologs. For a better accuracy, it is also possible to reconstruct the ancestral sequence and use it as reference [37, 38]. This is particularly suited when the divergence between the contemporary functional reference sequence and the pseudogene sequence is high.

Secondly, when all mutations have been documented, the use of a parsimony method over the defined species tree allows revealing their evolutionary history. We can now distinguish between the pseudogenization events that are shared and the independents ones. As exemplified in **Figure 6**, we can thus identify the first mutation that led to the pseudogenization and the history of mutations that accumulated ever since.

2.4. Automation

With the increasing interest in the field of pseudogene function and evolution, the phylogeny methods become essential. The automation of these methods is an important issue as well, but it is still in its infancy. Many difficulties exist, such as the construction of reliable phylogenies, the phylogenetic analysis, the collection of evidence from heterogeneous data, and the ability to merge all the information available to deduce the best explanation among the diversity of possibilities. To our knowledge, the only available automated method using phylogeny for pseudogene analysis is GLADX (**G**ene **L**oss **A**nalyzer **D**AGOBAH e**X**tension) [37]. It follows the outline described in this chapter. It relies on many analytical tools (e.g. Blast, phylogeny, protein prediction, reconstruction of ancestral sequences), and integrates heterogeneous data. Based on an expert system [39], GLADX is in fact designed to be as close as possible to human expertise and, compatibly with the errors that can be present in the databases and with the limits of the tools used, it is able to provide reliable results and interpretations. However, GLADX has been designed to study specifically unitary pseudogenes and works exclusively with the Ensembl database. Furthermore, it has not been tested on recent databases and can be outdated. Finally, it is not maintained. The complexity of the framework in which it has been developed, as well as the uncommon programming language used (PROLOG), didn't help to find contributors to maintain it.

Using new pipeline frameworks such as Snakemake [40] or Nextflow [41] might simplify the automation of such analysis, probably going even beyond.

2.5. Closing remarks

The most accurate way to detect and analyse pseudogenes is by using phylogenetic approaches. The integration of evolutionary context provides unrivalled results for the understanding of pseudogenes and the underlying pseudogenization phenomenon. Unfortunately, the development of tools to automate the detection and analysis of pseudogenes using phylogenetic approaches is still required. Such tools should be able to analyse the three main types of pseudogenes: processed, unprocessed and unitary pseudogenes.

3. Notes

1. When a phylogenetic tree needs to be built, the choice of the species and the time occurring between speciation events need to be chosen wisely, in order to have a data set that ensures high accuracy. An example is reported in **Figure 7A**. If the aim of the study is to attribute a date to the loss of a given gene observed in *M. musculus*, the use of *H. sapiens* and *M. domestica* as other species allows estimating that the loss occurred during a period of 100 million years. However, adding one more species such as *R. norvegicus* enables to deduce that the gene was lost during a more restricted period of 40 Million years (**Fig. 7B**). If the gene is absent in *R. norvegicus* as well, it can be deduced that it was lost after the Eutheria speciation and before the Rodent speciation during a period of 60 Million years. The loss might have occurred two times independently after the Rodent speciation (once in the *R. norvegicus* lineage and once in the *M. musculus* lineage), but this scenario would be less parsimonious and is excluded.

2. The choice of species must also consider the annotation quality of their genomes, because the more the genome is badly sequenced, the higher is the probability that a gene or pseudogene is present in a non-sequenced DNA fragment. Thus, plenty of missing genes and pseudogenes might in fact be sequencing artefacts and be actually present in the genome. The chance of finding the gene sequence of interest can be increased using EST databases and the absence or the pseudogenization of a gene can be double-checked using a re-sequencing process. In any case, a compromise has to be found between the choice of species and the quality of the annotation of their genomes.

3. When a pseudogene is ancient, its study should be performed at the amino acidic level, where the sequence can be more conserved due to synonymous mutations (**Fig. 8**). The same applies to a recent pseudogene, when the only known functional gene exists in a very divergent species. In all the other cases, it is possible to study the pseudogene at the nucleotide level.

4. To build the phylogeny, firstly a blast should be performed using a query sequence against different genomes in order to find similar genes. Using these sequences in FASTA format, a multiple sequence alignment should be then performed using tools such as ClustalW, Muscle

or T-coffee (34–36). To increase the accuracy of the phylogeny, the alignment can be cleaned in order to remove sites, sequences and regions that are spurious or poorly aligned (37–39). Once the alignment is ready, the phylogeny can be built using software such as Phyml, Figenix, RaxML [42–44]. Note that Figenix integrates these steps and can automatically build a phylogeny from a query sequence.

5. An alternative method to detect gene losses is the use of clustering algorithms. Indeed, after the clustering of sequences of different genomes, the losses can be determined by a parsimony analysis of the genes contained in the clusters compared to the species tree of the species used. However, the use of the clustering method by itself can lead to artefactual groups. For large-scale studies, this approach has the advantage compared to a phylogenetic approach to be relatively fast. Thus, it can be used as first quick filter to be followed by a deepened analysis using phylogeny. The IODA database contains losses that have been detected in Chordate using this double approach [45].

6. Two approaches are available to build this new phylogeny: i) to use the sequence alignment employed to build the previous phylogeny (*see* **Note 4**) after having added the putative pseudogenic sequence; ii) to build a new alignment with the sequences obtained by Blast using the putative pseudogene as a query. Using the first approach, if the putative pseudogene comes from a distant family, it might be placed as an external sequence. Thus, no information about its evolution or its function can be deduced. On the contrary, using the second approach, the new sequences collected for this phylogeny can provide new information.

7. Depending on the method used to build the new phylogeny, the gene initially used as query in the Blast step may be missing. Therefore, using only this gene to test orthology is risky.

8. If the sequence investigated is too degenerated to be kept in the process (too short or too divergent from the other present sequences), the orthology cannot be determined by phylogeny.

9. Often, the first Blast hits found are not pseudogenes, but paralogous or co-orthologous functional genes (since pseudogenes evolve more quickly than genes, their similarity is often lower than those of paralogous genes), while the hits that correspond to pseudogenes do not appear among the best. Moreover, the Blast search can detect several pseudogenes. Consequently, depending on what is sought for (only one pseudogene or as many as possible), it is necessary to study a bigger or smaller fraction of the Blast hits found.

10. As annotation errors or mis-annotations may exist in databases, the orthologous sequence retrieved may be a pseudogene annotated as functional gene or, on the contrary, a functional gene annotated as pseudogene. These cases are rare and difficult to assess therefore they require deeper investigation. The suspicious to face up such case can be

raised when the pattern of the loss/pseudogenization events detected in the results are not parsimonious [37].

11. The checking may be done verifying the annotations present at the genome location where the sequence has been found.

12. Even if disruptive mutations are found, this does not mean that the identified pseudogene is non-functional. Pseudogenes have for a long time been considered without functional relevance, but numerous studies have indeed shown the transcriptional activity of some of them, and their involvement in biological processes [16–19]. However, the transcriptional or functional activity may be tough to detect and requires further analyses. Furthermore, when a functional activity is revealed, it remains to be determined whether this activity is a residual activity of the parental gene or if the pseudogene can be assimilated to a new gene having a new function.

13. Depending on the chosen approach and on the subsidiary information desired, it may be preferable to perform the verification that the retrieved sequence is indeed a pseudogene before performing the orthology verification by phylogeny. It is a considerable time optimization to avoid building a phylogeny to study a sequence that finally would turn out to be an already annotated gene.

Acknowledgements

Thanks to Philippe Gouret and Julien Paganini for their help on the development of the strategy to detect gene losses and unitary pseudogenes.

References

1. Bischof JM, Chiang AP, Scheetz TE, et al (2006) Genome-wide identification of pseudogenes capable of disease-causing gene conversion. Hum Mutat 27:545–52. https://doi.org/10.1002/humu.20335
2. Khelifi A, Adel K, Duret L, et al (2005) HOPPSIGEN: a database of human and mouse processed pseudogenes. Nucleic Acids Res 33:D59-66. https://doi.org/10.1093/nar/gki084
3. Torrents D, Suyama M, Zdobnov E, Bork P (2003) A genome-wide survey of human pseudogenes. Genome Res 13:2559–67. https://doi.org/10.1101/gr.1455503
4. Zhang Z, Carriero N, Gerstein M (2004) Comparative analysis of processed pseudogenes in the mouse and human genomes. Trends Genet 20:62–7. https://doi.org/10.1016/j.tig.2003.12.005
5. Vanin EF (1985) Processed pseudogenes: characteristics and evolution. Annu Rev Genet 19:253–72. https://doi.org/10.1146/annurev.ge.19.120185.001345
6. Gerstein M, Zheng D (2006) The real life of pseudogenes. Sci Am 295:48–55. https://doi.org/10.1038/scientificamerican0806-48
7. Satta Y (2011) Primate Evolution: Gene Loss and Inactivation. Life Sci 1–7. https://doi.org/10.1002/9780470015902.a0005121.pub2
8. Wang X, Grus WE, Zhang J (2006) Gene losses during human origins. PLoS Biol 4:e52. https://doi.org/10.1371/journal.pbio.0040052
9. Fischer I, Dainat J, Ranwez V, et al (2014) Impact of recurrent gene duplication on adaptation of plant genomes. BMC Plant Biol 14:1–15. https://doi.org/10.1186/1471-2229-14-151
10. Magadum S, Banerjee U, Murugan P, et al (2013) Gene duplication as a major force in evolution. J Genet 92:155–161. https://doi.org/10.1007/s12041-013-0212-8
11. Mitchell A, Graur D (2005) Inferring the pattern of spontaneous mutation from the pattern of substitution in unitary pseudogenes of Mycobacterium leprae and a comparison of mutation patterns among distantly related organisms. J Mol Evol 61:795–803. https://doi.org/10.1007/s00239-004-0235-0
12. Li W-H, Gojobori T, Nei M (1981) Pseudogenes as a paradigm of neutral evolution. Nature 292:237–239. https://doi.org/10.1038/292237a0
13. Nachman MW, Crowell SL (2000) Estimate of the mutation rate per nucleotide in humans. Genetics 156:297–304
14. Weir JT, Schluter D (2008) Calibrating the avian molecular clock. Mol Ecol 17:2321–8. https://doi.org/10.1111/j.1365-294X.2008.03742.x
15. Olson M V. (1999) When less is more: Gene loss as an engine of evolutionary change. Am J Hum Genet 64:18–23. https://doi.org/10.1086/302219
16. Chan W-L, Yuo C-Y, Yang W-K, et al (2013) Transcribed pseudogene ψPPM1K generates endogenous siRNA to suppress oncogenic cell growth in hepatocellular carcinoma. Nucleic Acids Res 41:3734–47. https://doi.org/10.1093/nar/gkt047
17. Hirotsune S, Yoshida N, Chen A, et al (2003) An expressed pseudogene regulates the messenger-RNA stability of its homologous coding gene. Nature 423:91–6. https://doi.org/10.1038/nature01535
18. Wen Y-Z, Zheng L-L, Qu L-H, et al (2012) Pseudogenes are not pseudo any more. RNA

Biol 9:27–32. https://doi.org/10.4161/rna.9.1.18277
19. Pink RC, Wicks K, Caley DP, et al (2011) Pseudogenes: pseudo-functional or key regulators in health and disease? RNA 17:792–8. https://doi.org/10.1261/rna.2658311
20. Zhang ZD, Frankish A, Hunt T, et al (2010) Identification and analysis of unitary pseudogenes: historic and contemporary gene losses in humans and other primates. Genome Biol 11:R26. https://doi.org/10.1186/gb-2010-11-3-r26
21. Zhu J, Sanborn JZ, Diekhans M, et al (2007) Comparative genomics search for losses of long-established genes on the human lineage. PLoS Comput Biol 3:e247. https://doi.org/10.1371/journal.pcbi.0030247
22. Costello J, Han M, Hahn M (2008) Limitations of pseudogenes in identifying gene losses. In: Comparative Genomics. pp 14–25
23. Farris JS (1977) Phylogenetic Analysis Under Dollo's Law. Syst Biol 26:77–88. https://doi.org/10.1093/sysbio/26.1.77
24. Mirkin BG, Fenner TI, Galperin MY, Koonin E V (2003) Algorithms for computing parsimonious evolutionary scenarios for genome evolution , the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. BMC Evol Biol 3:2. https://doi.org/10.1186/1471-2148-3-2
25. Sankoff D, Rousseau P (1975) Locating the vertices of a Steiner tree in an arbitrary metric space. Math Program 9:240–246. https://doi.org/10.1007/BF01681346.
26. Sankoff D (1975) Minimal mutation trees in sequences. Soc Ind Appl Math 28:35–42
27. Altschul SF, Gish W, Miller W, et al (1990) Basic local alignment search tool. J Mol Biol 215:403–10. https://doi.org/10.1016/S0022-2836(05)80360-2
28. Schwartz S, Kent WJ, Smit A, et al (2003) Human-mouse alignments with BLASTZ. Genome Res 13:103–7. https://doi.org/10.1101/gr.809403
29. Ortutay C, Vihinen M (2008) PseudoGeneQuest - service for identification of different pseudogene types in the human genome. BMC Bioinformatics 9:299. https://doi.org/10.1186/1471-2105-9-299
30. Zhang Z, Carriero N, Zheng D, et al (2006) PseudoPipe: an automated pseudogene identification pipeline. Bioinformatics 22:1437–9. https://doi.org/10.1093/bioinformatics/btl116
31. Gouret P, Thompson JD, Pontarotti P (2009) PhyloPattern: regular expressions to identify complex patterns in phylogenetic trees. BMC Bioinformatics 10:298. https://doi.org/10.1186/1471-2105-10-298
32. Sukumaran J, Holder MT (2010) DendroPy: a Python library for phylogenetic computing. Bioinformatics 26:1569–71. https://doi.org/10.1093/bioinformatics/btq228
33. Vos RA, Caravas J, Hartmann K, et al (2011) BIO::Phylo-phyloinformatic analysis using perl. BMC Bioinformatics 12:63. https://doi.org/10.1186/1471-2105-12-63
34. Jungreis I, Chan CS, Waterhouse RM, et al (2016) Evolutionary dynamics of abundant stop codon readthrough. Mol Biol Evol 33:3108–3132. https://doi.org/10.1093/molbev/msw189
35. Stark A, Lin MF, Kheradpour P, et al (2007) Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures. Nature 450:219–232.

https://doi.org/10.1038/nature06340
36. Loughran G, Chou MY, Ivanov IP, et al (2014) Evidence of efficient stop codon readthrough in four mammalian genes. Nucleic Acids Res 42:8928–8938. https://doi.org/10.1093/nar/gku608
37. Dainat J, Paganini J, Pontarotti P, Gouret P (2012) GLADX: An automated approach to analyze the lineage-specific loss and pseudogenization of genes. PLoS One 7:. https://doi.org/10.1371/journal.pone.0038792
38. Paten B, Herrero J, Fitzgerald S, et al (2008) Genome-wide nucleotide-level mammalian ancestor reconstruction. Genome Res 18:1829–43. https://doi.org/10.1101/gr.076521.108
39. Gouret P, Paganini J, Dainat J, et al (2011) Integration of evolutionary biology concepts for functional annotation and automation of complex research in evolution: The multi-agent software system DAGOBAH. In: Pontarotti P (ed) Evolutionary Biology – Concepts, Biodiversity, Macroevolution and Genome Evolution, P. Pontaro. Springer Berlin Heidelberg, Berlin, Heidelberg, pp 71–87
40. Köster J, Rahmann S (2012) Snakemake-a scalable bioinformatics workflow engine. Bioinformatics 28:2520–2522. https://doi.org/10.1093/bioinformatics/bts480
41. DI Tommaso P, Chatzou M, Floden EW, et al (2017) Nextflow enables reproducible computational workflows. Nat Biotechnol 35:316–319. https://doi.org/10.1038/nbt.3820
42. Paganini J, Gouret P (2012) Reliable Phylogenetic Trees Building: A New Web Interface for FIGENIX. Evol Bioinform Online 8:417–21. https://doi.org/10.4137/EBO.S9179
43. Stamatakis A (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22:2688–90. https://doi.org/10.1093/bioinformatics/btl446
44. Guindon S, Dufayard J-F, Lefort V, et al (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol 59:307–21. https://doi.org/10.1093/sysbio/syq010
45. Levasseur A, Paganini J, Dainat J, et al (2012) The chordate proteome history database. Evol Bioinforma 2012:. https://doi.org/10.4137/EBO.S9186