

Orphan genes: Their identification and evolution in plant-parasitic nematodes

Ercan Seçkin, Dominique Colinet, Marc Bailly-Bechet, Edoardo Sarti, Etienne Danchin

▶ To cite this version:

Ercan Seçkin, Dominique Colinet, Marc Bailly-Bechet, Edoardo Sarti, Etienne Danchin. Orphan genes: Their identification and evolution in plant-parasitic nematodes. JOBIM 2024 - Journées ouvertes en biologie, informatique, et mathématiques 2024, Jun 2024, Toulouse, France. hal-04615706v2

HAL Id: hal-04615706 https://hal.science/hal-04615706v2

Submitted on 19 Jun 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers. L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



ORPHAN GENES: THEIR IDENTIFICATION AND

EVOLUTION IN PLANT PARASITIC NEMATODES

Ercan SEÇKIN*^{1,2}, Dominique COLINET¹, Marc BAILLY-BECHET¹, Edoardo SARTI², Etienne GJ DANCHIN¹

1 ISA - INRAE PACA, Université Côte d'Azur, CNRS, 400 Route des Chappes, 06903, Sophia Antipolis, France. 2 Centre Inria d'Université Côte d'Azur, 2004 Route des Lucioles, 06902, Sophia Antipolis, France.

MANAGEN AND MANAGE

- We developed a bioinformatics pipeline to identify orphan genes and investigate their emergence dynamics and possible *de novo* origin.
- The biological purpose is to apply the built pipeline to the genus Meloidogyne of root-knot nematodes, which cause a huge global crop loss. The species of this genus secrete parasitism effector, proteins, many of which are encoded by genes having no identifiable homologues [4].

INTRODUCTION

- **Orphan genes** are genes that have no identifiable homologs. They are systematically found in eukaryotic genomes, and represent up to 30% of the genes [1].
- The emergence of such genes is an important mechanism of functional acquisition during evolution. They may originate from duplication or horizontal transfer events followed by rapid divergence. Another possibility is that they may emerge from non-genic regions, which is known as de novo gene birth [2]. Hence, there is a debate on the real origin of orphan genes [3].
- The most common method for identifying orphan genes involves the search for homologues (identification of their absence). Then, their origin can be studied by syntenic approaches. However, these approaches cannot always distinguish between de novo genes and the ones that diverged significantly from their ancestor.

METHOD & RESULTS

PIPELINE <u>Identification of homology relationships in nematoda</u> Proteome completeness analysis with BUSCO [5] and OMArk [6] for all the nematoda proteomes present on WormBase Parasite [7] database or predicted by the GAME team Selection of 85 nematode proteomes (including 8 Meloidogyne species) and 1 outgroup of nematodes (BUSCO completeness > 60%) Comparative proteome analysis and establishment of homology relationships Clustering of best reciprocal hits to Clustering of best reciprocal hits to orthogroups by kmer approach between all orthogroups by alignments between all proteomes with SonicParanoid [9] proteomes with OrthoFinder [8] Comparison of orthogroups from both tools Meloidogyne orphan gene candidates in nematoda Double check against nr database with Diamond [10] (e-value 0.001) Expression validation by using transcriptomic data FPKM > defined threshold for at least 1 gene of an orthogroup **Expressed orphan genes** (figure 2) species tree of the Study of origin Emergence dynamics with genus of interest **Bridge** [16] (figure 3) with the closest MSA for each sequence of an outgroup species Emergence position on the orthogroup with MAFFT [11] species tree for each orthogroup Gene tree reconstruction for each orthogroup by IQTREE [12] and Alignment of ASR sequences to reconciliation of gene trees with orphan from the genomes of corresponding species tree by *TreeRecs* [13] pre-existing gene outgroups (according to if aligned on a position at phylogeny in figure 3) with which there is a coding Ancestral sequence reconstruction Exonerate [15] (figure 4A) gene in the outgroup (ASR) with ARPIP [14] de novo gene if aligned with the presence of a mutated splicing Verification of orphan regions via syntenic site, stop codon or indels (not multiple of 3) on approaches (figure 4B) non-coding region in the outgroup

RESULTS 60000 Orphans Figure 2: Number of orphan and non Non orphans 40000 orphan genes for 8 Meloidogyne species <u>studied</u> Barplot representing the number of orphan genes 20000 identified for each of the 8 Meloidogyne species. Each bar corresponds to the total number of genes of each species. Light blue represents orphans and dark blue represents non orphan P.penetrans M.chitwoodi M.graminicola 326 M.enterolobii **507** M.arenaria M.javanica 4180 : Ancestral Meloidogyne M.luci 739 nodes with number of emerged orthogroups. M.incognita Figure 3: Species tree of Meloidogyne species with their closest outgroup species P.penetrans. Species tree with 8 Meloidogyne species corresponding to three different clades (represented at the right of the species) of Meloidogyne genus along with P. penetrans, which is the closest outgroup to this genus. Each ancestral node is represented by a dot. Number of orphan orthogroups emerged in each node is given by a value next to it. codon: *** Ancestral sequence positions: # Good splicing site Corresponding in white M. enterolobii Bad splicing site in red region Color code M. arenaria de novo M. javanica others M. incognita M. luci Figure 4: Example of a de novo gene birth that emerged inside the clade I A. Alignment of the reconstructed ancestral sequence of an example orthogroup common to all members of clade I except for M. enterolobii to the genome of M.enterolobii by Exonerate [15]. Brown represents exons and dark blue represents introns. Stop codons are represented by ***. Insertion-deletion positions (not a multiple of 3) are represented by #. Good splicing sites are in white,

mutated ones in red. B. Verification of the syntenic conservation of the identified de novo gene. For the position where ancestral sequence aligns to M.enterolobii, closest 3 genes in both directions are studied to determine the conservation of the region. Lines corresponds to introns and boxes corresponds to genes. Blue boxes corresponds to the de novo gene and blue cross is where it aligns on M.enterolobii. The color code indicates where the same genes align on different species.

CONCLUSION & DISCUSSION

This study introduces a pipeline that combines homology, transcriptomics, and ancestral sequence reconstruction to identify and investigate the origin of orphan genes. This makes it possible to correctly distinguish between orphan genes that have diverged from pre-existing genes and those that have emerged de novo. To date, results indicate that 24% of Meloidogyne genomes consist of orphan genes. At least 7% of these are the result of de novo gene birth. Yet, it should be kept in mind that the results depend highly on the quality of genomes in disposition.

Figure 1: Developed bioinformatics pipeline to identify orphan genes and to study their evolution in Meloidogyne.

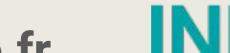
Despite the fact that some effectors in *Meloidogyne* are coded by non-homologous genes, there is still no connection between the orphan genes and their functions in Meloidogyne. Thus, we are exploring structural approaches to better characterize these gene-structure-functions associations.

References

- 1. Tautz D, Domazet-Lošo T. The evolutionary origin of orphan genes. Nat Rev Genet (2011).
- Weisman CM. The Origins and Functions of De Novo Genes: Against
- All Odds? J Mol Evol. (2022). Vakirlis, N. et al. Synteny-based analyses indicate that sequence
- divergence is not the main source of orphan genes. eLife (2020) Mitchum et al., Nematode effector proteins: an emerging paradigm
- of parasitism, New Phytologist (2013). Simão, F.A. et al. BUSCO: assessing genome assembly and
- annotation completeness with single-copy orthologs. Bioinformatics
- Nevers Y et al. Multifaceted quality assessment of gene repertoire
- https://parasite.wormbase.org/index.html Emms, D.M. et al. OrthoFinder: phylogenetic orthology inference for
- annotation with OMArk. Bioinformatics (2022).
- 9. Cosentino S, Iwasaki W. SonicParanoid2: fast, accurate, and comprehensive orthology inference with machine learning and language models, Bioinformatics (2023).
- 10. Buchfink et al., Fast and sensitive protein alignment using DIAMOND. Nat. Methods (2015).
- 11. Katoh, K., and Standley, D.M. MAFFT Multiple Sequence Alignment
- Software Version 7: Improvements in Performance and Usability. Mol. Biol. Evol. (2017).
- 12. Minh BQ et al. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Teeling E, editor. Mol Biol Evol.
- 13. Comte N et al. Treerecs: an integrated phylogenetic tool, from

biological sequence comparison. BMC Bioinformatics (2005)

sequences to reconciliations. Bioinformatics (2020). 14. Jowkar G. et al. ARPIP: Ancestral Sequence Reconstruction with Insertions and Deletions under the Poisson Indel Process. Syst Biol. (2023). 15. Slater, G., and Birney, E. Automated generation of heuristics for









comparative genomics. Genome Biol. (2019).



