



HAL
open science

Alienness: rapid detection of candidate horizontal gene transfers across the tree of life

Corinne Rancurel, Ludovic Legrand, Etienne Danchin

► To cite this version:

Corinne Rancurel, Ludovic Legrand, Etienne Danchin. Alienness: rapid detection of candidate horizontal gene transfers across the tree of life. *Genes*, 2017, 8 (10), pp.248. 10.3390/genes8100248 . hal-01609604

HAL Id: hal-01609604

<https://hal.science/hal-01609604v1>

Submitted on 3 Oct 2017

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.



Distributed under a Creative Commons Attribution 4.0 International License

Technical Note

Alieness: Rapid Detection of Candidate Horizontal Gene Transfers across the Tree of Life

Corinne Rancurel ¹, Ludovic Legrand ² and Etienne G. J. Danchin ^{1,*} 

¹ INRA, CNRS, ISA, Université Côte d'Azur, 06903 Sophia Antipolis Cedex, France; corinne.rancurel@inra.fr

² LIPM, INRA, CNRS, Université de Toulouse, 31326 Castanet-Tolosan Cedex, France; ludovic.legrand@inra.fr

* Correspondence: etienne.danchin@inra.fr; Tel.: +33-492-386-402

Received: 31 August 2017; Accepted: 22 September 2017; Published: 29 September 2017

Abstract: Horizontal gene transfer (HGT) is the transmission of genes between organisms by other means than parental to offspring inheritance. While it is prevalent in prokaryotes, HGT is less frequent in eukaryotes and particularly in Metazoa. Here, we propose Alieness, a taxonomy-aware web application available at <http://alieness.sophia.inra.fr>. Alieness parses BLAST results against public libraries to rapidly identify candidate HGT in any genome of interest. Alieness takes as input the result of a BLAST of a whole proteome of interest against any National Center for Biotechnology Information (NCBI) protein library. The user defines recipient (e.g., Metazoa) and donor (e.g., bacteria, fungi) branches of interest in the NCBI taxonomy. Based on the best BLAST E-values of candidate donor and recipient taxa, Alieness calculates an Alien Index (AI) for each query protein. An AI > 0 indicates a better hit to candidate donor than recipient taxa and a possible HGT. Higher AI represent higher gap of E-values between candidate donor and recipient and a more likely HGT. We confirmed the accuracy of Alieness on phylogenetically confirmed HGT of non-metazoan origin in plant-parasitic nematodes. Alieness scans whole proteomes to rapidly identify possible HGT in any species of interest and thus fosters exploration of HGT more easily and largely across the tree of life.

Keywords: horizontal gene transfer; alien index; lateral gene transfer

1. Introduction

Horizontal gene transfer (HGT), is the transmission of genes between organisms by other way than direct (vertical) inheritance from parental lineages to their offspring. HGT is prevalent in prokaryotes [1], with substantial proportions of bacterial genes jumping horizontally, rather than being vertically inherited [2]. These horizontally-acquired genes play important functions in bacteria, including spreading of antibiotic resistance and emergence of pathogenicity [3–6]. Although HGT is much less prevalent in eukaryotes, and particularly multicellular eukaryotes, there are reported cases in the literature, including in viridiplantae and Metazoa [7–12]. Some of the reported examples also evoke important associated roles in the recipient organism. This suggests that the genomic and biological impact of HGT could be more widespread in the tree of life than initially thought [13].

These HGT challenge the view of a purely tree-like backbone underlying inheritance of genetic information across species [14]. With the technological progress and cost reduction for genome sequencing, more and more genomes from a wider diversity of organisms now become available and this trend will continue. To provide a more comprehensive view of the contribution of HGT to the genomes of the different organisms across the tree of life, methods to rapidly detect candidate HGT are needed.

Two classical ways to identify candidate HGT are (i) to study the intrinsic (e.g., GC content, codon usage distribution) and/or (ii) the extrinsic (e.g., percent identity, BLAST [15] E-value against

other species) characteristics of the genes of a species of interest. For instance, in the ‘intrinsic’ category, genes that have GC content and codon usage deviating from the rest of the genes of a species of interest (receiver) might be considered as a sign for horizontal acquisition. Similarly, in the ‘extrinsic’ category, if a gene from a species of interest (receiver) shows higher similarity (lower E-value, higher bit-score in BLAST) to sequences from distantly related (donor) species than to genes of close relatives; this gene is a candidate HGT. Another frequent ‘extrinsic’ approach is to assess the phylogenomic distribution of genes across a panel of diverse organism by clustering genes in groups of orthologs and looking for those with a patchy distribution (phylogenomic profiles). For a more comprehensive overview of the different categories of methods to detect HGT, please refer to this recent review by Ravenhall et al. [16].

Methods in the intrinsic category can be efficient to identify HGT, as long as the GC content and codon usage distributions of the receiver and donor genomes are sufficiently different. However, if these intrinsic metrics are closely related or if the HGT event is ancient and the acquired gene has adapted to the GC content and codon usage of the receiver organism, such methods will not be able to identify HGT [17].

Methods in the extrinsic category do not suffer from undistinguishable GC content and codon usage as they rely only in a difference of magnitude in the BLAST or other similarity metrics between closely related and distant taxa. These methods are better suited to identify HGT between distant species (e.g., trans-kingdom) and require an as comprehensive as possible reference sequence library covering a diversity of species. One such method in the extrinsic category is the Alien Index metrics. It was first introduced by Gladyshev et al. [9] to identify HGT of non-metazoan origin in a metazoan species, the bdelloid rotifer *Adineta vaga*. Briefly, an Alien Index (AI) is calculated to measure a difference of magnitude between the best non-metazoan and best metazoan E-value. If the best non-metazoan E-value is closer to 0 than the best metazoan E-value, the AI will be positive, if the AI is ≥ 45 , it has been assumed that an HGT event is very likely. This AI method was used again to assess the total contribution of genes of non-metazoan origin in the whole genome of the bdelloid rotifer once it was sequenced [18]. Using custom Perl scripts, our laboratory and collaborators also used AI to identify candidate HGT of non-metazoan origin in the genome of the plant-parasitic nematode *Globodera rostochiensis* [19], in the genomes of several panagrolaimid nematodes [20] as well as in the transcriptomes of the nematodes *Nacobbus aberrans* [21], *Xiphinema index* [22] and other as yet unpublished genomes. Calculation of AI scores for whole proteomes had previously been implemented in a software called AlienG [23]. It was used to highlight the importance of HGT in the colonization of land by plants [8] and in several other studies of HGT across different lineages [24–26]. In the original AlienG, publication, an AI > 30 was deemed good indicator of acquisition via HGT. However, as far as we know, this software is neither publicly available for download nor deployed on a web server. Hence, no user-friendly web tool or downloadable software is available, so far to compute AI scores directly from BLAST results. Furthermore, extracting the taxonomic information from BLAST results against the National Center for Biotechnology Information’s (NCBI) (Bethesda, MD, USA) protein libraries can be a long and difficult task, which prevents popularization of such methods.

To circumvent this difficulty of retrieving taxonomic information, it is tempting to divide the NCBI’s non-redundant protein library (NR) or other sequence libraries into taxonomic subsets (i.e., a non-alien subset consisting of sequences from species closely related to the receiver species of interest, and one or several subsets consisting of sequences from distantly related (alien) candidate donors). This approach has been implemented in a Perl software named `alien_index` [27]. However, E-values obtained from BLAST against different sequence libraries are not comparable which makes calculation of an AI questionable, especially if the libraries are of different sizes. This is one of the reasons why, another score, named HGT index (or *h*), has been proposed [28]. The principle is very similar to the AI score, except that it is calculated based on a difference between the best donor and recipient species bit scores. In contrast to E-values, bit scores are comparable between sequence libraries of different sizes. Thus, this allows tackling the problem of taxonomic identification from a single big BLAST results by running several different BLASTs against several taxonomic subdivisions

of a sequence library (e.g., NCBI's NR, Swissprot, Uniprot [29]). However, this imposes to predefine a priori, candidate donor taxa and to run multiple BLASTs against different libraries that have either to be downloaded and formatted by the user or constructed by the user. The HGT index method was first used to determine genes of non-metazoan origin in the transcriptome of the bdelloid rotifer *Adineta ricciae* [28] and later used to infer the contribution of HGT to the transcriptomes of several different metazoan species, including vertebrates [11].

A refined scoring method has been recently proposed that not only takes into account sequence similarity between receiver and candidate donors, measured in BLAST bit score, but also taxonomic distance measured as the number of step in the NCBI's taxonomic lineage to reach the common ancestor of the query and subject species [30]. The software is publicly available and has been initially developed to identify HGT in fungal genomes. However, this requires local installation of voluminous data, manual configuration, and expert skills are required to tune and adapt the method to identify HGT in other phyla.

Similarly, a method called DarkHorse and initially developed to identify HGT in bacteria is available as a software that can be downloaded and installed locally [31]. This method proposes a lineage probability index score (LPI) based on the taxonomic ranks of top hits to identify candidate HGT from BLAST against protein databases. Here again, the method can be generalized to other taxa of interest. However, the installation is restricted to users with computational skills on a Unix-based platform and requires installing, configuring and managing a MySQL database as well as downloading the whole NCBI's NR database and the NCBI's taxonomy.

Phylogenetic methods that compare a gene or protein tree to a reference species tree and identify inconsistencies are the gold standard to predict HGT. While such methods exist [32–35] they need to be implemented by expert users. These phylogeny-based methods require a reference species tree for comparison. However, such reference trees are not always available for the group of species of interest but some methods propose to generate a reference species trees as well as computing individual gene trees in parallel [36]. Furthermore, producing phylogenetic trees for a whole proteome or large gene set can be extremely time consuming, especially if the initial homology search retrieves numerous homologs for alignments and phylogenetic reconstruction. Because of their computationally demanding nature, phylogeny-based methods are currently hardly adapted to analysis of large genomes or to a high number of genes. Hence, AI or HGT-index based methods constitute an interesting method to rapidly identify candidate HGT and narrow down the number of genes to be phylogenetically analyzed afterwards.

As far as we know, there is currently no publicly available web tool that allows to easily and rapidly identify HGT from large datasets directly from a BLAST result.

Here, we propose Alieness, a user-friendly web application that requires no installation of any software and that is publicly accessible at <http://alieness.sophia.inra.fr>. Alieness requires nothing else than BLAST results against any sequence library at the NCBI and a few parameters to calculate AI for a set of query sequences (e.g., a whole proteome). Alieness can be applied to any genome of interest and to identify candidate HGT from any donor to any recipient taxonomic group.

We tested the accuracy of Alieness on the genomes of two plant-parasitic nematodes, for which phylogenetically supported HGT of a whole series of genes involved in plant parasitism had been previously identified [37,38]. We found that all phylogenetically supported cases could be retrieved by Alieness with an AI > 9 and that this AI threshold corresponded to a low rate of putative false positives.

We believe Alieness will promote a more rapid exploration of candidate HGT across the tree of life and will contribute to assessing more globally the evolutionary significance of HGT.

2. Materials and Methods

2.1. Input Data

Typically, Alieness takes as input a compressed BLAST result (in .zip or .gz format) of a proteome performed against any NCBI protein database (Figure 1). Currently, the uploaded compressed BLAST result must not exceed a size of 500 Mb (ca. 2 to 2.5 Gb uncompressed). The BLAST result must be obtained by the BLASTp program available at the NCBI FTP website in BLAST+ package. The required format is the tabular format with comment lines obtained by using the `-outfmt 7` BLAST option. The user must also define the NCBI taxonomy identifier (TaxID) of the taxonomic group of interest (recipient group, e.g., Metazoa, Viridiplantae) and one or several TaxIDs separated by a comma for the taxonomic groups to be ignored for the calculation of the AI (excluded group(s)). The latter parameter must at least include the TaxID of the query species used to produce the BLAST result to avoid self-BLAST results preventing the identification of putative HGT. Any hit to any taxon included in the entered taxonomic node(s) will be ignored and excluded from the calculation of AI. How taxonomic information is retrieved from BLAST tabular result is explained in the next section. By default, the different BLAST hits are categorized in the root nodes of the NCBI taxonomy: Archaea, Bacteria, Eukaryota, Viroids and Viruses. The groups 'other' and 'unclassified' are ignored, as they cannot be assigned to a known taxon. The user can define any additional category to classify the BLAST results in groups of interest (e.g., Fungi, Gammaproteobacteria, Deltavirus . . .). For this, the user just need to input a list of one or several corresponding TaxIDs.

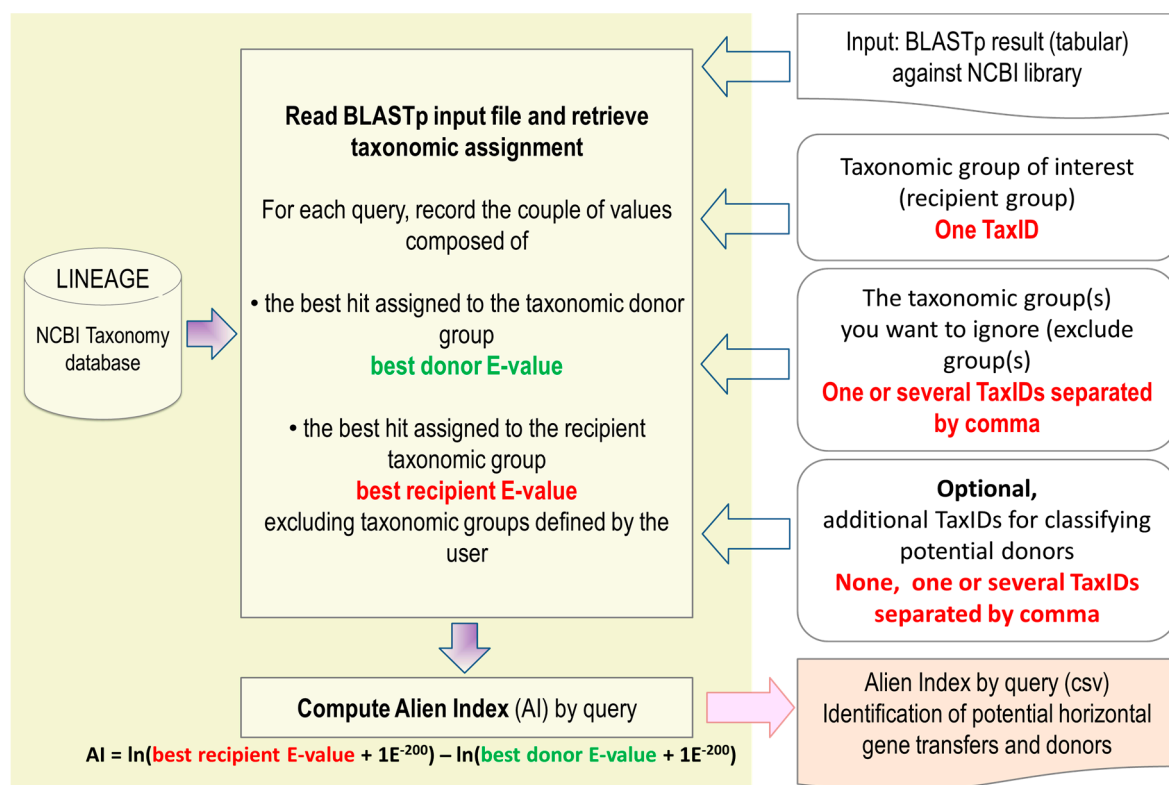


Figure 1. Schematic flow chart of Alieness processing BLAST results and providing Alien Indexes (AIs). NCBI: National Center for Biotechnology Information; TaxID: taxonomy identifier.

2.2. Taxonomic Assignment

The taxonomic assignment of BLAST hit sequences is performed using the taxonomic databases available on the NCBI FTP site (<ftp://ftp.ncbi.nlm.nih.gov/pub/taxonomy/>). This repository contains

the full taxonomy database along with flat files associating nucleotide and protein sequence records with their TaxIDs. The challenge lies in reconstructing the taxonomic lineage of species just from GenBank identifier (GI) extracted from BLAST results. To perform this taxonomic identification, we use the files 'gi_taxid_prot.dmp.gz' and 'taxdump.tar.gz', both available for download at <ftp://ftp.ncbi.nlm.nih.gov/pub/taxonomy/>. The first file provides, for each GI of protein record, its corresponding TaxID. The second file contains NCBI Taxonomy database dump files, and more precisely nodes.dmp and names.dmp files. Alieness retrieves the parent node ID of each node ID in GenBank taxonomy database from nodes.dmp file and associates a name for each of them using names.dmp file. Thus, Alieness re-builds the lineage of species based on the GI accessions.

2.3. Processing

The BLAST result of each query protein is read from the best BLAST hit to the last significant hit (Figure 1). Using the extracted taxonomic information (see above), Alieness records a couple of values composed of the best hit assigned to the taxonomic donor group called best donor E-value and the best hit assigned to the recipient taxonomic group called best recipient E-value. When a taxonomic assignment matches a group to be excluded, as defined by the user or belongs to the 'Unclassified' and 'Other' categories, the hit is ignored and discarded from the rest of the analyses.

2.4. Calculation of an Alien Index

To illustrate the calculation of an AI we will take as an example identification of candidate HGT of non-metazoan origin in a metazoan genome. The principle remains the same for identification of inter-phylum HGT to the genome of any species of interest from any donor (alien) taxonomic groups.

Typically, Alieness takes as input the result of a BLASTp search of a whole set of predicted proteins of interest (e.g., from a whole genome or a transcriptome) against the NCBI's NR library or any protein library available at NCBI. The BLAST result must be in tabular format and when performed against NR, we recommend using an E-value threshold of $1E^{-3}$ and no filtering for low complexity regions. Based on the BLAST hits GI accession numbers and on the NCBI taxonomy, Alieness extracts the best metazoan and best non-metazoan E-values to calculate an AI, first defined in [9], with the following formula.

$$AI = \ln(\text{best metazoan Evalue} + 1E^{-200}) - \ln(\text{best non_metazoan Evalue} + 1E^{-200})$$

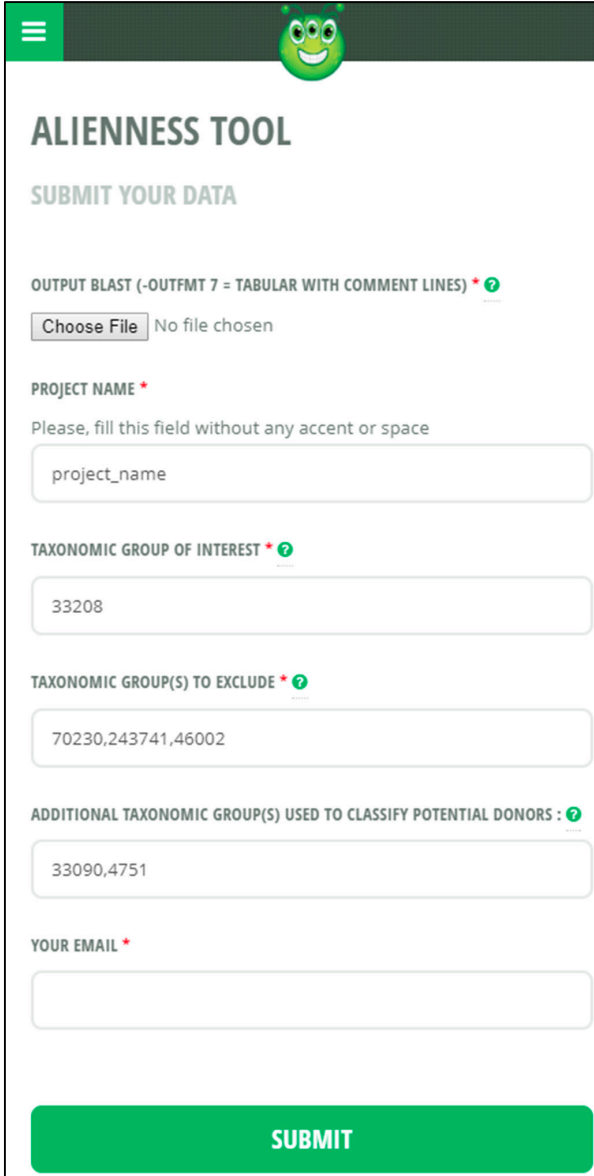
When either no metazoan or non-metazoan significant BLAST hit is found, a penalty E-value of 1 is automatically assigned as the best metazoan or non-metazoan E-value, respectively. Hence, E-values of the best metazoan and non-metazoan hits vary between 0 and 1 and, consequently AI scores vary between -460.5 and 460.5 . An $AI > 0$ indicates a better hit to a non-metazoan species than to a metazoan species and possible acquisition via HGT of non-metazoan origin. However, if the difference of magnitude between the best metazoan and non-metazoan E-values is low, this difference might not be significant and could just reflect a slight difference in the ranking of the best hits. For instance, a highly conserved protein between metazoan and non-metazoan species could by chance return a better hit to non-metazoan than metazoan. In the original publication of AI calculation, Gladyshev et al. [9] estimated, based on a few phylogenetic analyses that an $AI \geq 45$ corresponds to HGT candidates highly supported by phylogenetic trees. They thus suggested that an AI threshold of 45 is a good indicator for likely HGT. In the results section of this paper, we re-evaluated this threshold on the genomes of plant-parasitic nematodes and showed that even lower AI scores provide a good balance in terms of sensitivity (recall of phylogenetically supported HGT) and specificity (low number of false positives).

3. Results

3.1. Alienness Web Portal and Interface

The Alienness web portal has been developed in HTML5 and CSS3 and launches a series of Perl programs via a CGI Perl module. Alienness is publicly and freely available at this URL: <http://alienness.sophia.inra.fr> and does not require creation of an account or downloading any software. Alienness just requires a pre-computed BLAST result in tabular format (-outfmt 7) against a protein library. The BLAST result file must be compressed in .gz or .zip format. The protein library must have GI or accession numbers that exist in the NCBI taxonomy database to retrieve the taxonomic information necessary to calculate AI values (i.e., any downloadable BLAST protein library at the NCBI). For a better coverage of the biodiversity, we recommend using the NCBI's NR library.

To run an Alienness analysis, click the 'Alienness Tool' tab and the following form will appear (Figure 2).



The screenshot shows the 'ALIENNESS TOOL' web interface. At the top, there is a green header with a menu icon and a green alien mascot. Below the header, the title 'ALIENNESS TOOL' is displayed in large, bold, dark green letters. Underneath, the section 'SUBMIT YOUR DATA' is shown in a smaller, grey font. The form consists of several fields, each with a circled number (1-6) to its left:

- 1. 'OUTPUT BLAST (-OUTFMT 7 = TABULAR WITH COMMENT LINES) * ?' with a 'Choose File' button and the text 'No file chosen'.
- 2. 'PROJECT NAME *' with a text input field containing 'project_name' and a note: 'Please, fill this field without any accent or space'.
- 3. 'TAXONOMIC GROUP OF INTEREST * ?' with a text input field containing '33208'.
- 4. 'TAXONOMIC GROUP(S) TO EXCLUDE * ?' with a text input field containing '70230,243741,46002'.
- 5. 'ADDITIONAL TAXONOMIC GROUP(S) USED TO CLASSIFY POTENTIAL DONORS : ?' with a text input field containing '33090,4751'.
- 6. 'YOUR EMAIL *' with an empty text input field.

At the bottom of the form is a large green button labeled 'SUBMIT' in white capital letters.

Figure 2. Graphical interface of Alienness to upload BLAST results and input the different parameters. Detailed information on 1–6 is given in the main text.

Fill the following information (all fields in Figure 2 except field 5 are mandatory).

1. Upload your BLAST result file here in .gz or .zip format.
2. Give a name to your project.
3. Enter the NCBI TaxID of the recipient group of interest (only 1 TaxID should be put here). For instance, if you are interested in HGT of non-metazoan origin to a metazoan species, please input 33208 (NCBI TaxID for Metazoa) as recipient. If you are interested in HGT of non-green plant origin to a green plant species, please input 33090 (NCBI TaxID for Viridiplantae). This is valid for any TaxID and this information is necessary to retrieve the best 'TaxID of recipient' E-value and best 'TaxID of candidate donor' E-value for calculation of an Alien index.
4. Enter the NCBI TaxIDs (one or several) for the taxonomic groups you want to ignore in the calculation of the AI. You must at least input the TaxID of the query species you used to produce the BLAST result. For instance, imagine that your query proteome is from *Mus musculus* and you are looking for HGT of non-metazoan origin. Most of the best BLAST hits (if not all) will be self-hits of *M. musculus* against itself. Hence, there is no chance to identify a *M. musculus* protein that would return a better hit to a non-metazoan than to a metazoan species. It is thus necessary to ignore all self-hits to *M. musculus* for the calculation of AI values. In this case, you have to input 10090 (NCBI TaxID for *M. musculus*). If you suspect that some HGT may have arisen earlier, in an ancestor of the rodents, for instance, then you have to input 9989 (TaxID for Rodentia). This will ignore all BLAST hits to Rodentia. Note that you can input several TaxIDs separated by comma and no space in this field if you want to ignore several non-overlapping taxonomic groups. This is useful if there is no monophyletic group in the NCBI taxonomy corresponding to the ensemble of species you want to ignore. Because Alienness is fast, you can also run it several times with different ignored TaxIDs and check the effect of this parameter on the number of candidate HGT identified (e.g., Rodentia, Mammalia, Vertebrata, etc.).
5. This field is optional and does not affect calculation of AI scores. This field controls the classification of the best BLAST hits in taxonomic categories in the output files. If left blank, the best hits are classified in the NCBI taxonomy basal categories Archaea, Bacteria, Eukaryota, Viroids and Viruses (the NCBI categories 'Other' and 'Unclassified' are ignored as they cannot be assigned to a species). If a user wants to further classify the best hits in other categories any additional TaxID can be entered. For instance, if you want to further categorize the Eukaryota into Fungi, Stramenopiles and Viridiplantae, you have to input 4751, 33634 and 33090 (separated by comma and no space).
6. Input your e-mail address. Please double check that your e-mail is correct because the link to download the results will be sent to this address.

Then click the submit button and wait for your compressed BLAST results to be uploaded to the server (do not refresh the page). Once the compressed BLAST result file is uploaded, a page will open and notify that the upload was successful and that you will receive two automatic e-mails. A first e-mail confirming the submission with a summary of the parameters selected by the user. A second one, a few minutes later, indicating that the job is finished and providing a summary of your job and a link to download a .zip archive with the results.

So far, we have run a dozen of eukaryotic whole proteomes with Alienness, and on average, the whole process from retrieval of taxonomic information to calculation of AI scores and production of result files takes 10–15 min (excluding upload time for the compressed BLAST results).

3.2. Results produced by Alienness

Once you download and uncompress the .zip result archive via the link in the automatic e-mail (Alienness_jobnumber.zip), the following result files will be available:

- Project_name_HGT_ai_full.csv: table presenting AI values and taxonomic information for all the proteins that returned an AI value.

- Project_name_HGT_ai_junk.csv: all the proteins that have an AI ≤ 0 and are thus probably not originated by HGT.
- Project_name_HGT_ai_likely_contamination.csv: proteins with an AI > 0 and $\geq 70\%$ identity to a protein from candidate donors (these cases are considered as possible contaminations and should be carefully manually examined).
- Project_name_HGT_ai_likely_contamination_stat.csv: statistics on the taxonomic distribution (species and kingdoms) of candidate donors for the possible contamination category.
- Project_name_HGT_ai_possible.csv: proteins with an AI > 0 and $< 70\%$ identity to candidate donors constitute the pool of possible HGT.
- Project_name_HGT_ai_possible_stat.csv: statistics on the taxonomic distribution (species and kingdoms) of candidate donors for the possible HGT category.
- Project_name_HGT_ai_very_likely.csv: proteins with an AI > 30 and $< 70\%$ identity to candidate donors constitute the pool of very likely HGT.
- Project_name_HGT_ai_very_likely_stat.csv: statistics on the taxonomic distribution (species and kingdoms) of candidate donors for the very likely HGT category.
- Project_name_HGT_blastp_no_hits.csv: proteins for which no AI could be calculated because they returned no significant BLAST hits.
- Project_name_index.html: an html file that allows visually exploring the BLAST results with a color code.
- Project_name_summary.txt: log file providing information on execution time, parameters selected by the user, etc.

Note that the files Project_name_HGT_ai_full.csv, Project_name_HGT_ai_junk.csv, Project_name_HGT_ai_likely_contamination.csv, Project_name_HGT_ai_possible_stat.csv and Project_name_HGT_ai_very_likely.csv all have the same template ("Project_name" is substituted by the user-defined name set in the Alieness field number 2 in Figure 2). To exemplify this template, we use the .csv result file (very likely AI) of Alieness run on the 2008 version of the protein set of the root-knot nematode *Meloidogyne incognita* [39] (Supplementary Table S1).

- Column 1 (AI): the Alien Index
- Column 2 (query hits number): the number of hits returned by the protein in consideration.
- Column 3 (query name): the query accession number.
- Columns 4–8: the best E-values for the basal NCBI taxonomy categories Archaea, Bacteria, Eukaryota, Viroids and Viruses. When the best E-value does not belong to the taxonomic category in consideration, a "-" character is present.
- Columns 9–10 (optional): the best E-values for the user-defined additional taxonomic categories (field 5 of the form in Figure 2). In this example, there are two user-defined additional categories: Viridiplantae and Fungi.
- Column 11: the best E-value for the user-defined taxonomic group of interest (recipient group). In this example, the group of interest is Metazoa as we are interested in candidate HGT of non-metazoan origin in a metazoan species (the root-knot nematode *M. incognita*).
- Column 12 (best hit GI)*: the NCBI GI accession of the best BLAST hit.
- Column 13 (best hit prct ident)*: the percent identity between the query sequence and the best hit.
- Column 14 (best hit org nickname)*: an abbreviated species name for the best hit.
- Column 15 (best hit org full name)*: full species name of the best hit
- Column 16 (best hit taxo group)*: abbreviated taxonomic classification of the best hit.
- Column 17 (best hit taxid)*: NCBI TaxID of the best hit.
- Column 18 (best hit lineage): full taxonomic lineage for the best hit.

* Note that hits to the user-defined excluded taxonomic groups (field 4 of form in Figure 2) are ignored as well as hits to the NCBI categories 'Other' and 'Unclassified'.

In addition, all the files named with a “_stat” suffix are built on a same two-column template and provide basic statistics on the candidate donors (or contaminant). The first column contains names of kingdom or taxa and the second the occurrence of these taxa in the set of candidate donors or possible contaminants.

The file “Project_name_index.html” allows exploring the BLAST results of interest regarding candidate HGT and candidate contaminants. By opening this file in a web browser, the following page will open (Figure 3).

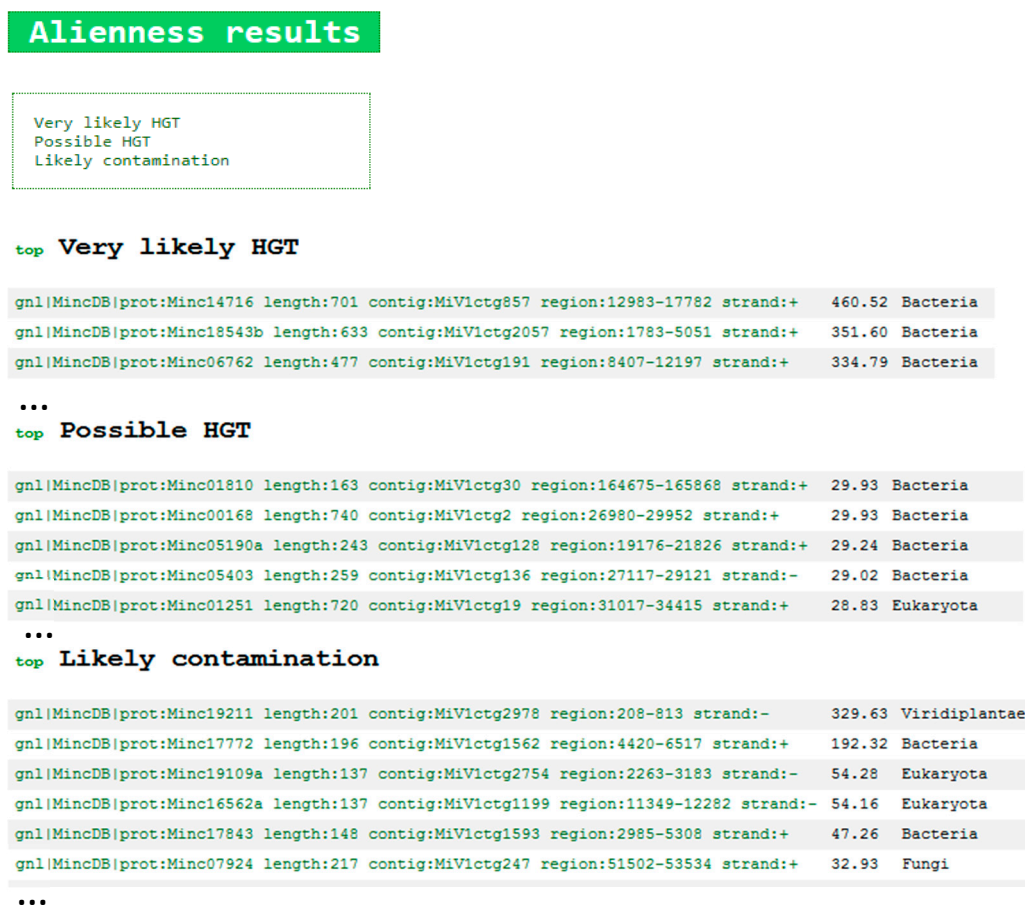


Figure 3. The .html index file opened in a web browser to explore the BLAST results by category (Very likely HGT, Possible HGT and Likely contamination).

BLAST results are classified in 3 categories, (i) very likely HGT (AI > 30 and <70% identity to candidate donor); (ii) possible HGT (AI > 0 and <70% identity to candidate donor) and (iii) likely contamination (AI > 0 and ≥70% identity to candidate donor). For each category, a list of accession numbers from the query proteome as well as the AI value and the taxonomic category of the candidate donor are indicated. By clicking on any of these accession numbers, a color-coded BLAST result in .html format opens and allows exploring the results in more details. An example on a *M. incognita* protein (Minc14047b, a GH5 cellulase acquired by HGT [37]) is given below (Figure 4).

In the BLAST result, the color code is as follows: hits that belong to the taxonomic group(s) to exclude are highlighted in blue. In this example, we chose to exclude the Tylenchida (TaxID: 6300) that include the root-knot nematodes and many other related plant-parasitic nematodes. The best ‘alien’ hit is indicated in green. In this example, we were interested in HGT of non-metazoan origin in a metazoan, so the alien taxon is anything but Metazoa. Here, the best alien taxon is from the bacterium *Cellulophaga algicola* with an E-value of $8E^{-81}$. The best hit that belongs to the taxonomic

group of interest (=recipient taxon, as defined in the field 3 of the form in Figure 2) is highlighted in red. In this example, the taxonomic group of interest is Metazoa and the best metazoan hit is from the phytophagous insect *Apriona japonica* with an E-value of $9E^{-68}$. These are thus the green (here best non-metazoan) and red (here best metazoan) E-values that are used to calculate the AI. Hits for which a taxonomic group could not be assigned are highlighted in light violet. This corresponds to protein that have either been removed from the NCBI or replaced by another protein with a different accession number. Hits that belong to the categories ‘other’ and ‘unclassified’ are highlighted in light grey. Note that no color code is given after the best green and best red taxonomic groups have been identified because these are necessary and sufficient to calculate an AI.

```
# BLASTP 2.2.26+
# Query: gn1|MincDB|prot:Minc14047b length:472 contig:MiV1ctg757 region:2001-5852 strand:-
# AI : 30.05
# very_likely_HGT
# Database: nr
# Fields: query id, subject id, % identity, alignment length, mismatches, gap opens, q. start, q. end, s. start, s. end, evaluate, bit score
# 250 hits found
gn1|MincDB|prot:Minc14047b gi|354549469|gb|AER27791.1| 52.96 253 108 4 22 272 1 244 2e-81 263
gn1|MincDB|prot:Minc14047b gi|503316845|ref|WP_013551506.1| 45.19 312 158 6 11 320 65 363 8e-81 266
gn1|MincDB|prot:Minc14047b gi|354549445|gb|AER27764.1| 51.78 253 113 4 22 272 1 246 8e-79 257
gn1|MincDB|prot:Minc14047b gi|646938477|ref|WP_025614577.1| 45.45 308 155 5 11 316 64 360 8e-79 261
gn1|MincDB|prot:Minc14047b gi|385655149|gb|AFI63769.1| 57.14 224 89 3 9 230 13 231 1e-78 256
gn1|MincDB|prot:Minc14047b gi|647297669|ref|WP_025741749.1| 48.06 283 134 6 12 292 85 356 2e-78 260
gn1|MincDB|prot:Minc14047b gi|354549447|gb|AER27780.1| 50.20 255 119 4 22 273 1 250 6e-78 254
gn1|MincDB|prot:Minc14047b gi|354549485|gb|AER27799.1| 51.19 252 110 5 24 272 1 242 3e-77 253
gn1|MincDB|prot:Minc14047b gi|402314775|gb|AFQ55680.1| 42.43 337 173 9 1 325 6 333 4e-77 255
gn1|MincDB|prot:Minc14047b gi|354549481|gb|AER27797.1| 51.59 252 109 5 24 272 1 242 8e-77 251
gn1|MincDB|prot:Minc14047b gi|639134539|ref|WP_024480035.1| 44.81 308 157 6 11 316 64 360 1e-76 255
gn1|MincDB|prot:Minc14047b gi|227278211|gb|ACP20205.1| 44.31 325 161 10 1 320 4 313 2e-76 254
gn1|MincDB|prot:Minc14047b gi|503317944|ref|WP_013552605.1| 38.21 458 243 14 1 430 15 460 4e-76 260
gn1|MincDB|prot:Minc14047b gi|586954898|gb|AHJ96632.1| 45.10 306 155 6 19 320 19 315 6e-75 256
gn1|MincDB|prot:Minc14047b gi|303760955|ref|WP_013995031.1| 43.67 300 156 6 19 316 65 351 7e-75 250
gn1|MincDB|prot:Minc14047b gi|639135527|ref|WP_024481016.1| 44.44 324 165 7 1 318 52 366 1e-74 257
gn1|MincDB|prot:Minc14047b gi|390979540|dbj|BAH21527.1| 45.40 326 160 8 1 322 24 335 2e-74 256
gn1|MincDB|prot:Minc14047b gi|612135178|dbj|BAO65801.1| 46.69 317 151 7 6 318 44 346 3e-74 256
gn1|MincDB|prot:Minc14047b gi|499788978|ref|WP_011469712.1| 45.48 321 157 8 6 322 34 340 3e-74 256
gn1|MincDB|prot:Minc14047b gi|269965254|dbj|BAI50016.1| 45.40 326 160 8 1 322 24 335 3e-74 256
gn1|MincDB|prot:Minc14047b gi|499788976|ref|WP_011469710.1| 45.40 326 160 8 1 322 24 335 8e-74 255
gn1|MincDB|prot:Minc14047b gi|260894277|dbj|BAI44493.1| 40.95 337 170 11 5 325 4 327 8e-74 246
gn1|MincDB|prot:Minc14047b gi|599567310|gb|EYF05358.1| 42.44 311 165 5 14 322 83 381 4e-73 247
gn1|MincDB|prot:Minc14047b gi|503385191|ref|WP_013619852.1| 43.93 305 157 5 19 320 62 355 4e-73 246
gn1|MincDB|prot:Minc14047b gi|504668320|ref|WP_014855431.1| 40.24 338 172 7 1 320 9 334 4e-73 245
gn1|MincDB|prot:Minc14047b gi|584429593|gb|EWHI4260.1| 43.93 305 157 5 19 320 62 355 1e-72 245
gn1|MincDB|prot:Minc14047b gi|646952097|ref|WP_025015336.1| 43.83 324 167 7 1 318 52 360 1e-72 252
gn1|MincDB|prot:Minc14047b gi|260894281|dbj|BAI44495.1| 40.48 336 173 10 4 325 5 327 2e-72 243
gn1|MincDB|prot:Minc14047b gi|527821125|ref|WP_023450454.1| 48.73 326 177 7 7 328 16 333 5e-72 242
gn1|MincDB|prot:Minc14047b gi|612135164|dbj|BAO65790.1| 47.47 297 142 6 26 318 9 295 7e-72 249
gn1|MincDB|prot:Minc14047b gi|499904749|ref|WP_011585483.1| 42.28 324 167 9 4 322 16 324 5e-71 240
gn1|MincDB|prot:Minc14047b gi|501465119|ref|WP_012488564.1| 46.80 297 143 8 22 314 44 329 8e-71 246
gn1|MincDB|prot:Minc14047b gi|806574|emb|CAA60493.1| 46.80 297 143 8 22 314 44 329 9e-71 246
gn1|MincDB|prot:Minc14047b gi|522824834|ref|WP_020737533.1| 41.80 311 167 5 8 316 45 343 1e-70 243
gn1|MincDB|prot:Minc14047b gi|517155055|ref|WP_018343873.1| 39.43 317 167 7 8 316 18 317 7e-70 236
gn1|MincDB|prot:Minc14047b gi|648518011|ref|WP_026209762.1| 39.43 317 167 7 8 316 21 320 7e-70 236
gn1|MincDB|prot:Minc14047b gi|402314779|gb|AFQ55682.1| 42.04 333 171 9 1 325 6 324 1e-69 236
gn1|MincDB|prot:Minc14047b gi|499903764|ref|WP_011584498.1| 44.44 306 149 6 19 317 34 325 1e-69 251
gn1|MincDB|prot:Minc14047b gi|284159269|gb|ADB00152.1| 42.30 305 162 5 19 320 63 356 2e-69 236
gn1|MincDB|prot:Minc14047b gi|354549448|gb|AER27781.1| 57.84 204 74 5 71 272 1 194 2e-69 230
gn1|MincDB|prot:Minc14047b gi|620947060|gb|KCZ93343.1| 37.50 352 191 9 1 338 5 341 7e-69 234
gn1|MincDB|prot:Minc14047b gi|303483556|gb|AAM50039.1| 40.80 326 178 5 1 321 15 330 1e-68 233
gn1|MincDB|prot:Minc14047b gi|4699637|pdb|1EGZ|A 46.21 277 134 7 22 294 5 270 3e-68 231
gn1|MincDB|prot:Minc14047b gi|37725601|gb|AAO25506.1| 40.49 326 179 5 1 321 15 330 3e-68 232
gn1|MincDB|prot:Minc14047b gi|647298145|ref|WP_025742218.1| 39.81 319 185 3 1 318 18 330 4e-68 243
gn1|MincDB|prot:Minc14047b gi|502329494|ref|WP_012765495.1| 43.12 320 159 8 2 318 30 329 5e-68 235
gn1|MincDB|prot:Minc14047b gi|517154372|ref|WP_018343190.1| 41.84 337 172 7 1 328 13 334 7e-68 247
gn1|MincDB|prot:Minc14047b gi|578359008|gb|AHI15747.1| 41.30 322 168 8 1 311 4 315 9e-68 231
gn1|MincDB|prot:Minc14047b gi|6578950|gb|AAF18152.1|AF208495_1 46.40 278 134 7 22 295 48 314 1e-67 233
```

Figure 4. The color-coded BLAST results that can be explored from a web browser, color code and further explanations are detailed below.

3.3. Validation of Alienness on Nematode Genomes

Alienness identified 649 proteins with an AI > 0 (i.e., that have better hits to non-metazoan species than to metazoan) on the 2008 version of the *M. incognita* proteome [39]. Some of these proteins could originate from HGT events. To assess which AI threshold maximizes the number of very likely HGT and minimizes the number of false positives, we compared the results of Alienness to a previous analysis we published on the same proteome in 2012 [40]. Briefly, in this 2012 analysis we first identified *Meloidogyne* proteins having no clear ortholog in Metazoa via an OrthoMCL analysis [41]. We used these proteins to reconstruct automatic phylogenies with FIGENIX [32] and then searched for topologies supporting an HGT event using PhyloPattern [33]. We identified 205 *M. incognita* proteins for which an AI value could be calculated by Alienness and an automatic phylogenetic tree was previously reconstructed (Supplementary Table S2). To estimate the recall rate of Alienness, we plotted the percentage of candidate HGT proteins supported by phylogenies (Figure 5) for different AI threshold values (from -10 to >66 with incremental steps of 2). We differentiate phylogenies only

supported by a node 'A' linking the plant-parasitic nematodes to non-metazoan putative donors and those supported by both a node 'A' and an additional node 'B', linking the node 'A' to any species but plant-parasitic nematodes [40]. To estimate the risk of false positives produced by Alieness, we plotted the percentage of proteins returning phylogenetic trees not supporting HGT at the same range of AI values.

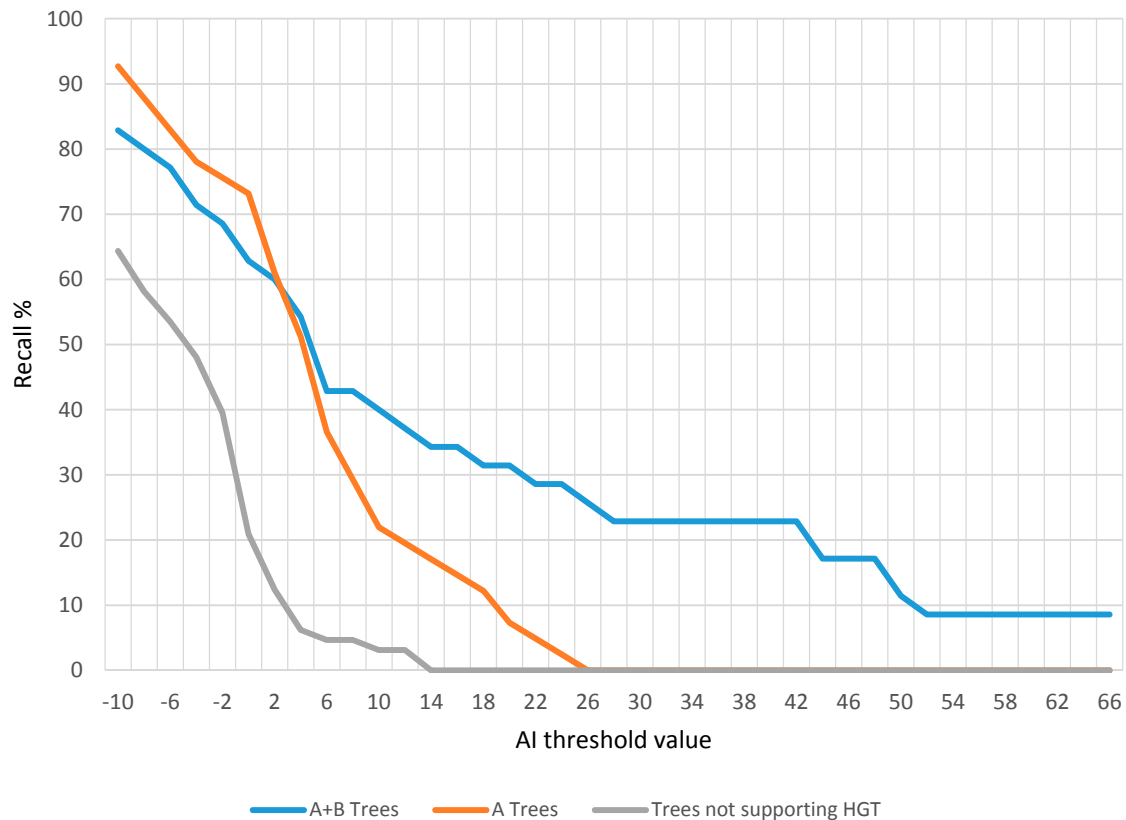


Figure 5. Recall rate (sensitivity) and risks of false positives of Alieness at different AI thresholds. The percentage of retrieved phylogenetically supported (with A or A + B nodes) as well as of phylogenetically unsupported cases (y-axis) as a function of the AI threshold (x-axis) is represented on this graph.

At an AI value > -10 we recall 83% and 93% of HGT cases supported by automatic phylogenies with A + B and A support, respectively. However, at this AI value, 64% of the proteins that returned phylogenies without support for HGT are also retrieved. Note that the AI cut-off as to be set to > -37 to retrieve all phylogenetically supported proteins, but at this threshold, 93% of the proteins returning trees without support for HGT are also retrieved. At an E-value > 0 ; 63% of the HGT proteins with A + B phylogenetic support are retrieved and 68% of those with A support only. However, at this threshold, 21% of the proteins returning trees without support for HGT are also retrieved. From an AI threshold of > 14 , all proteins returning trees not supporting HGT are eliminated. An AI > 14 corresponds to a difference of magnitude $> 1E^6$ between the best non-metazoan and best metazoan hits. At this threshold, 34% and 17% of the A + B and A trees, respectively are still retrieved. At an AI threshold > 26 , only A + B phylogenetically supported proteins remain, the A-only supported cases are eliminated.

To further measure the ability of Alieness to recall highly supported candidate HGT, we tested whether previously reported cases of HGT, supported by meticulous manual phylogenetic analysis were found by Alieness. From a survey of the literature, we identified 23 cases of HGT in plant-parasitic nematodes supported by manual phylogenetic analysis (Table 1, Supplementary Table S3).

Table 1. Horizontal gene transfer (HGT) cases supported by manual phylogenetic analysis in root-knot and cyst nematodes.

Protein / Protein Family [refs]	Highest AI in <i>M. incognita</i>	Highest AI in <i>G. rostochiensis</i>	Process
GH5_2 Cellulases [37,42–47]	39.14	198.94	PCW Degradation
GH30 Xylanase [37,48]	259.49	-	PCW Degradation
GH28 Polygalacturonase [37,49]	351.60	-	PCW Degradation
Expansin-like proteins [37,39,50,51]	86.11	29.93	PCW Degradation
GH43 candidate Arabinanase [37] *	69.07	-	PCW Degradation
GH53 candidate Arabinogalactanase [52] *	-	349.30	PCW Degradation
PL3 Pecate Lyase [37,53–55]	137.46	137.06	PCW Degradation
Chorismate Mutase [56–58]	15.02	42.36	Def. manipulation
Candidate Isochorismatase [59] *	91.41	66.08	Def. manipulation
Candidate Cyanate Lyases [60,61] *	9.90	11.51	Detoxification
GH32 invertase [62]	154.42	241.26	Nutrient processing
VB1 thiD [63] *	-	154.50	Nutrient processing
VB1 thiE [63] *	-	163.99	Nutrient processing
VB1 thi4 [63] *	-	108.07	Nutrient processing
VB1 thiM [63] *	-	46.05	Nutrient processing
VB1 tenA [63] *	-	108.33	Nutrient processing
VB5 panC [63] *	16.52	183.11	Nutrient processing
VB6 SNO [64]	-	-	Nutrient processing
VB6 SOR-SNZ [64]	-	12.72	Nutrient processing
Candidate GSI Glutamine Synthase [40,65] *	35.59	29.24	Nutrient processing
NodL—like [65,66] *	-	13.12	Feed. site induction
Candidate L-threonine aldolase [40,65] *	-	164.69	??
Candidate Phosphorybosyl transferase [40,65] *	202.63	198.13	??

* function not biochemically confirmed so far; PCW: Plant Cell Wall; AI: Alien Index; ??: no inferred process

We retrieved the PFAM domains [67] present in the corresponding proteins and scanned these domains against the proteomes of the root-knot nematode *M. incognita* [39] and of the cyst nematode *G. rostochiensis* [19]. We found corresponding protein-coding genes for 13 and 19 of these HGT cases in the proteomes of *M. incognita* and *G. rostochiensis*, respectively. We reported the AI scores for the proteins corresponding to these HGT cases (Table 1, Supplementary Table S3). In some of the cases, the genes acquired via HGT form multigene families in the recipient species genomes [37]. In these cases, we reported the highest AI value and indicated how many more proteins were from this gene family in Supplementary Table S3.

We found that 100% of the phylogenetically supported cases returned an AI > 9 in *M. incognita* and an AI > 11 in *G. rostochiensis* (Figure 6). With an AI value > 14; more than 92% and 84% of the phylogenetically supported cases are retrieved in *M. incognita* and *G. rostochiensis*, respectively. As mentioned in the previous section, at this AI threshold, no case of automatic phylogeny not supporting HGT is retrieved.

Overall, depending on the aim of the user, we recommend using different AI thresholds. To have a more comprehensive overview of possible HGT, we recommend an AI > 0 but with a substantial proportion of possible false positives, to be manually curated. To focus on candidates that are likely to produce phylogenetic trees supporting HGT, and minimizing the rate of false positives, we recommend an AI > 14. To focus on the candidates that will likely produce phylogenies with higher support for HGT, we recommend an AI > 26.

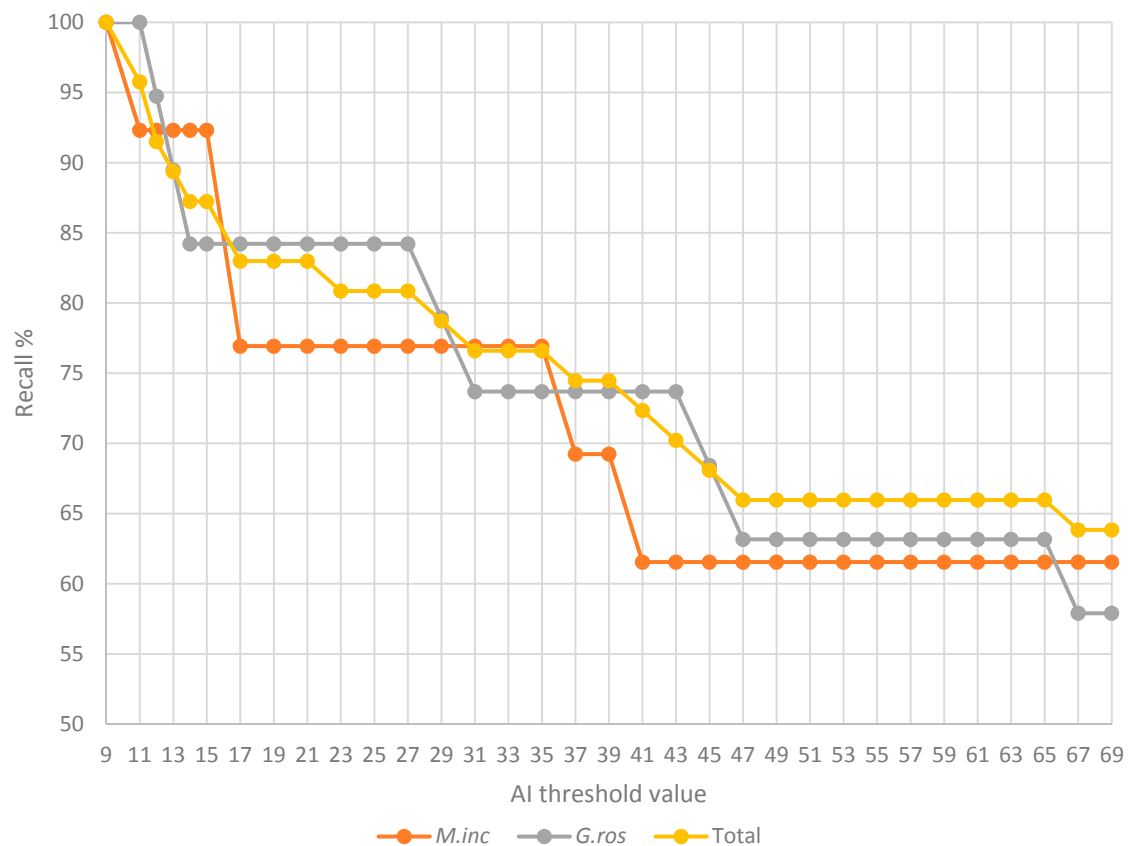


Figure 6. Phylogenetically supported HGT from the literature in the genomes of two plant-parasitic nematodes at different AI thresholds. Percentage of known and phylogenetically supported HGT cases (y -axis) at different AI thresholds (x -axis) in *M. incognita* (orange) and *G. rostochiensis* (grey).

4. Discussion

Alieness allows rapid identification of candidate HGT from whole proteomes and provides accompanying taxonomic information as well as basic statistics for the possible donors. An advantage of providing scores such as an AI is that the user can define its own threshold of significance and partition its dataset according to the AI scores. A high AI is an indication for putative acquisition by HGT. Obviously; a careful phylogenetic analysis remains the gold standard to support the hypothesis of HGT. However, careful phylogenetic analysis is time-consuming, can be complicated and requires expert skills, which is not (or hardly) applicable to large datasets of thousands of genes. In this perspective, Alieness provides an efficient metrics to quickly reduce the number of candidates that can be further analyzed via phylogenetic analysis. For instance, in the case of the plant-parasitic nematodes 20,359 and 14,309 proteins were present in the predicted proteome of *M. incognita* and *G. rostochiensis*, respectively. However, only 632 *M. incognita* and 519 *G. rostochiensis* proteins returned an AI > 0 and <70% identity to putative donors, which substantially narrows down the number of phylogenetic analyzes that must be undertaken. With an AI > 14, corresponding to low risk of false positives, the number of candidates is reduced to 165 in *M. incognita* and 136 in *G. rostochiensis*.

Alieness also provides a list of putative contaminants that can be useful for the user to prune the original dataset or even to identify possible endosymbiont. Any protein with an AI > 0 and $\geq 70\%$ identity is put apart in a category of possible contaminants. In a recent paper, Ku and Martin [68] defined what they called the 70% rule, above which candidate HGT of prokaryote origin to eukaryotes are more likely to represent contamination than actual HGT. This rule holds at the nucleotide level and for prokaryote to eukaryote transfers. In our own experience, none of the candidate HGT of non-metazoan origin to nematodes, supported by careful phylogenetic analysis, reported so far present

more than 70% identity at the protein level with a candidate donor. We thus, extended this 70% rule to the protein level. However, among these suspicious contaminants, rare recent or well-conserved HGT might be present, but we recommend being particularly cautious with these candidates to rule out the possibility of contamination. One evident first verification is to check that bona-fide genes from the recipient species surround candidate HGT genes in its genome. With Alieness, this can be easily achieved by looking for surrounding genes with negative AI values. A series of additional checks hold true for prokaryotic to metazoan HGTs, including presence of spliceosomal introns on the gene models, or presence of eukaryotic signal peptides for secretion in the protein. These and other features specific to the recipient species and absent or much less frequent in the candidate donors are good indicators to rule out the hypothesis of contamination. Support from expression data is also an independent element making HGT more likely than contamination.

Although Alieness provides rapid identification of candidate HGT, some limitations should be kept in mind while interpreting the results. First, as discussed in the previous section, Alieness just predicts candidate HGT that ideally need to be further supported by additional sourced of evidence. Second, the ability to detect candidate HGT depends on the quality and coverage of the queried sequence library. Both the accuracy of the taxonomic assignment and the diversity of taxa covered by the reference library influence the calculation of AI or any other method based on BLAST hits. Third, because BLAST E-values and bit scores also depend on the length of the query protein and matching hits, metrics such as AI and HGT Index will necessarily return lower scores for short sequences as the difference of E-value or bit score between the best donor and recipient organisms will have less difference in magnitude.

Overall, Alieness provides a fast and reliable service to rapidly detect candidate HGT in any genome of interest from any donor group. Although HGT is recognized as an evolutionary important and prevalent mechanism in prokaryotes, its importance and prevalence in eukaryotes remain actively debated and can be controversial [69]. Exploring more proteomes for a wider diversity of species will allow deciphering the contribution of HGT to the biology and genomes across the tree of life more comprehensively. Alieness will promote this effort by publicly offering a user-friendly method to rapidly scan existing and upcoming genomes.

Supplementary Materials: The following are available online at www.mdpi.com/2073-4425/8/10/248/s1. Table S1: AI-MiV1: Alien Indexes of the *M. incognita* whole proteome, Table S2: AI-Tree-MiV1: Alien Indexes of HGT supported by automatic phylogeny, Table S3: GeneHGT: Alien Indexes of previously reported HGT in plant-parasitic nematodes.

Acknowledgments: The authors are grateful to the GenoToul bioinformatics platform Toulouse Midi-Pyrenees, France as well as the INRA SPE BBRIC network and the LIPM bioinformatics platform at INRA Toulouse, France for providing computing resources.

Author Contributions: C.R. developed and tested Alieness and the web portal. E.G.J.D defined the specifications of Alieness and supervised the project. L.L. contributed to the code of Alieness. E.G.J.D. analyzed the data and wrote the paper with contribution from C.R.

Conflicts of Interest: The authors declare no conflict of interest. The founding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

References

1. Thomas, C.M.; Nielsen, K.M. Mechanisms of, and Barriers to, Horizontal Gene Transfer between Bacteria. *Nat. Rev. Microbiol.* **2005**, *3*, 711–721. [[CrossRef](#)] [[PubMed](#)]
2. Lukjancenko, O.; Wassenaar, T.M.; Ussery, D.W. Comparison of 61 Sequenced *Escherichia coli* Genomes. *Microb. Ecol.* **2010**, *60*, 708–720. [[CrossRef](#)] [[PubMed](#)]
3. Wiedenbeck, J.; Cohan, F.M. Origins of bacterial diversity through horizontal genetic transfer and adaptation to new ecological niches. *FEMS Microbiol. Rev.* **2011**, *35*, 957–976. [[CrossRef](#)] [[PubMed](#)]

4. Shoemaker, N.B.; Vlamakis, H.; Hayes, K.; Salyers, A.A. Evidence for extensive resistance gene transfer among *Bacteroides* spp. and among *Bacteroides* and other genera in the human colon. *Appl. Environ. Microbiol.* **2001**, *67*, 561–568. [[CrossRef](#)] [[PubMed](#)]
5. Treangen, T.J.; Rocha, E.P.C. Horizontal Transfer, Not Duplication, Drives the Expansion of Protein Families in Prokaryotes. *PLoS Genet.* **2011**, *7*, e1001284. [[CrossRef](#)] [[PubMed](#)]
6. Koonin, E.V.; Makarova, K.S.; Aravind, L. Horizontal gene transfer in prokaryotes: Quantification and classification. *Annu. Rev. Microbiol.* **2001**, *55*, 709–742. [[CrossRef](#)] [[PubMed](#)]
7. Dunning Hotopp, J.C.; Clark, M.E.; Oliveira, D.C.; Foster, J.M.; Fischer, P.; Torres, M.C.; Giebel, J.D.; Kumar, N.; Ishmael, N.; Wang, S.; et al. Widespread lateral gene transfer from intracellular bacteria to multicellular eukaryotes. *Science* **2007**, *317*, 1753–1756. [[CrossRef](#)] [[PubMed](#)]
8. Yue, J.; Hu, X.; Sun, H.; Yang, Y.; Huang, J. Widespread impact of horizontal gene transfer on plant colonization of land. *Nat. Commun.* **2012**, *3*, 1152. [[CrossRef](#)] [[PubMed](#)]
9. Gladyshev, E.A.; Meselson, M.; Arkhipova, I.R. Massive horizontal gene transfer in bdelloid rotifers. *Science* **2008**, *320*, 1210–1213. [[CrossRef](#)] [[PubMed](#)]
10. Eyres, I.; Boschetti, C.; Crisp, A.; Smith, T.P.; Fontaneto, D.; Tunnacliffe, A.; Barraclough, T.G. Horizontal gene transfer in bdelloid rotifers is ancient, ongoing and more frequent in species from desiccating habitats. *BMC Biol.* **2015**, *13*. [[CrossRef](#)] [[PubMed](#)]
11. Crisp, A.; Boschetti, C.; Perry, M.; Tunnacliffe, A.; Micklem, G. Expression of multiple horizontally acquired genes is a hallmark of both vertebrate and invertebrate genomes. *Genome Biol.* **2015**, *16*, 50. [[CrossRef](#)] [[PubMed](#)]
12. Dunning Hotopp, J.C. Horizontal gene transfer between bacteria and animals. *Trends Genet.* **2011**, *27*, 157–163. [[CrossRef](#)] [[PubMed](#)]
13. Huang, J. Horizontal gene transfer in eukaryotes: The weak-link model. *BioEssays* **2013**, *35*, 868–875. [[CrossRef](#)] [[PubMed](#)]
14. Raoult, D. The post-Darwinist rhizome of life. *Lancet* **2010**, *375*, 104–105. [[CrossRef](#)]
15. Altschul, S.F.; Madden, T.L.; Schaffer, A.A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D.J. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **1997**, *25*, 3389–3402. [[CrossRef](#)] [[PubMed](#)]
16. Ravenhall, M.; Škunca, N.; Lassalle, F.; Dessimoz, C. Inferring Horizontal Gene Transfer. *PLoS Comput. Biol.* **2015**, *11*, e1004095. [[CrossRef](#)] [[PubMed](#)]
17. Koski, L.B.; Morton, R.A.; Golding, G.B. Codon Bias and Base Composition Are Poor Indicators of Horizontally Transferred Genes. *Mol. Biol. Evol.* **2001**, *18*, 404–412. [[CrossRef](#)] [[PubMed](#)]
18. Flot, J.-F.; Hespels, B.; Li, X.; Noel, B.; Arkhipova, I.; Danchin, E.G.J.; Hejnl, A.; Henriessat, B.; Koszul, R.; Aury, J.-M.; et al. Genomic evidence for ameiotic evolution in the bdelloid rotifer *Adineta vaga*. *Nature* **2013**, *500*, 453–457. [[CrossRef](#)] [[PubMed](#)]
19. Eves-van den Akker, S.; Laetsch, D.R.; Thorpe, P.; Lilley, C.J.; Danchin, E.G.J.; Da Rocha, M.; Rancurel, C.; Holroyd, N.E.; Cotton, J.A.; Szitenberg, A.; et al. The genome of the yellow potato cyst nematode, *Globodera rostochiensis*, reveals insights into the basis of parasitism and virulence. *Genome Biol.* **2016**, *17*. [[CrossRef](#)] [[PubMed](#)]
20. Schiffer, P.H.; Danchin, E.; Burnell, A.M.; Schiffer, A.-M.; Creevey, C.; Wong, S.; Dix, I.; O'Mahony, G.; Culleton, B.A.; Rancurel, C.; et al. Signatures of the evolution of parthenogenesis and cryptobiosis in the genomes of panagrolaimid nematodes. *BioRxiv* **2017**, 159152. [[CrossRef](#)]
21. Eves-van den Akker, S.; Lilley, C.J.; Danchin, E.G.J.; Rancurel, C.; Cock, P.J.A.; Urwin, P.E.; Jones, J.T. The Transcriptome of *Nacobbus aberrans* Reveals Insights into the Evolution of Sedentary Endoparasitism in Plant-Parasitic Nematodes. *Genome Biol. Evol.* **2014**, *6*, 2181–2194. [[CrossRef](#)] [[PubMed](#)]
22. Danchin, E.G.J.; Perfus-Barbeoch, L.; Rancurel, C.; Thorpe, P.; Da Rocha, M.; Bajew, S.; Neilsen, R.; Sokolova, E.; Da Silva, C.; Guy, J.; et al. The transcriptomes of *Xiphinema index* and *Longidorus elongatus* reveal independent acquisition of parasitism genes by horizontal gene transfer in early-branching plant-parasitic nematodes. *Genes* **2017**, under review.
23. Tian, J.; Sun, G.; Ding, Q.; Huang, J.; Oruganti, S.; Xie, B. AlienG: An Effective Computational Tool for Phylogenomic Identification of Horizontally Transferred Genes. In Proceedings of the ISCA 3rd International Conference on Bioinformatics and Computational Biology, New Orleans, LA, USA, 23–25 March 2011.

24. Yue, J.; Sun, G.; Hu, X.; Huang, J. The scale and evolutionary significance of horizontal gene transfer in the choanoflagellate *Monosiga brevicollis*. *BMC Genom.* **2013**, *14*, 729. [[CrossRef](#)] [[PubMed](#)]
25. Sun, T.; Xu, Y.; Zhang, D.; Zhuang, H.; Wu, J.; Sun, G. An acyltransferase gene that putatively functions in anthocyanin modification was horizontally transferred from *Fabaceae* into the genus *Cuscuta*. *Plant Divers.* **2016**, *38*, 149–155. [[CrossRef](#)]
26. Ni, T.; Yue, J.; Sun, G.; Zou, Y.; Wen, J.; Huang, J. Ancient gene transfer from Algae to animals: Mechanisms and evolutionary significance. *BMC Evol. Biol.* **2012**, *12*, 83. [[CrossRef](#)] [[PubMed](#)]
27. Ryan, J.F. Alien_Index: Identify Potential Non-Animal Transcripts in Animal Transcriptomes. Available online: https://github.com/josephryan/alien_index. (accessed on 18 July 2017).
28. Boschetti, C.; Carr, A.; Crisp, A.; Eyres, I.; Wang-Koh, Y.; Lubzens, E.; Barraclough, T.G.; Micklem, G.; Tunnacliffe, A. Biochemical Diversification through Foreign Gene Expression in Bdelloid Rotifers. *PLoS Genet.* **2012**, *8*, e1003035. [[CrossRef](#)] [[PubMed](#)]
29. The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* **2017**, *45*, D158–D169.
30. Nguyen, M.; Ekstrom, A.; Li, X.; Yin, Y. HGT-Finder: A New Tool for Horizontal Gene Transfer Finding and Application to *Aspergillus* genomes. *Toxins* **2015**, *7*, 4035–4053. [[CrossRef](#)] [[PubMed](#)]
31. Podell, S.; Gaasterland, T. DarkHorse: A method for genome-wide prediction of horizontal gene transfer. *Genome Biol.* **2007**, *8*, R16. [[CrossRef](#)] [[PubMed](#)]
32. Gouret, P.; Vitiello, V.; Balandraud, N.; Gilles, A.; Pontarotti, P.; Danchin, E.G.J. FIGENIX: Intelligent automation of genomic annotation: Expertise integration in a new software platform. *BMC Bioinform.* **2005**, *6*, 198. [[CrossRef](#)] [[PubMed](#)]
33. Gouret, P.; Thompson, J.D.; Pontarotti, P. PhyloPattern: Regular expressions to identify complex patterns in phylogenetic trees. *BMC Bioinform.* **2009**, *10*, 298. [[CrossRef](#)] [[PubMed](#)]
34. Gouret, P.; Paganini, J.; Dainat, J.; Louati, D.; Darbo, E.; Pontarotti, P.; Levasseur, A. Integration of Evolutionary Biology Concepts for Functional Annotation and Automation of Complex Research in Evolution: The Multi-Agent Software System DAGOBAN. In *Evolutionary Biology—Concepts, Biodiversity, Macroevolution and Genome Evolution*; Pontarotti, P., Ed.; Springer: Berlin/Heidelberg, Germany, 2011; pp. 71–87. ISBN 978-3-642-20762-4.
35. Frickey, T.; Lupas, A.N. PhyloGenie: Automated phylome generation and analysis. *Nucleic Acids Res.* **2004**, *32*, 5231–5238. [[CrossRef](#)] [[PubMed](#)]
36. Jain, S.; Panda, A.; Colson, P.; Raoult, D.; Pontarotti, P. MimiLook: A Phylogenetic Workflow for Detection of Gene Acquisition in Major Orthologous Groups of Megavirales. *Viruses* **2017**, *9*, 72. [[CrossRef](#)] [[PubMed](#)]
37. Danchin, E.G.; Rosso, M.N.; Vieira, P.; de Almeida-Engler, J.; Coutinho, P.M.; Henrissat, B.; Abad, P. Multiple lateral gene transfers and duplications have promoted plant parasitism ability in nematodes. *Proc. Natl. Acad. Sci. USA* **2010**, *107*, 17651–17656. [[CrossRef](#)] [[PubMed](#)]
38. Haegeman, A.; Jones, J.T.; Danchin, E.G. Horizontal gene transfer in nematodes: A catalyst for plant parasitism? *Mol. Plant Microbe Interact.* **2011**, *24*, 879–887. [[CrossRef](#)] [[PubMed](#)]
39. Abad, P.; Gouzy, J.; Aury, J.-M.; Castagnone-Sereno, P.; Danchin, E.G.J.; Deleury, E.; Perfus-Barbeoch, L.; Anthouard, V.; Artiguenave, F.; Blok, V.C.; et al. Genome sequence of the metazoan plant-parasitic nematode *Meloidogyne incognita*. *Nat. Biotechnol.* **2008**, *26*, 909–915. [[CrossRef](#)] [[PubMed](#)]
40. Paganini, J.; Campan-Fournier, A.; Da Rocha, M.; Gouret, P.; Pontarotti, P.; Wajnberg, E.; Abad, P.; Danchin, E.G.J. Contribution of Lateral Gene Transfers to the Genome Composition and Parasitic Ability of Root-Knot Nematodes. *PLoS ONE* **2012**, *7*, e50875. [[CrossRef](#)] [[PubMed](#)]
41. Li, L.; Stoeckert, C.J.; Roos, D.S. OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Res.* **2003**, *13*, 2178–2189. [[CrossRef](#)] [[PubMed](#)]
42. Smant, G.; Stokkermans, J.P.; Yan, Y.; de Boer, J.M.; Baum, T.J.; Wang, X.; Hussey, R.S.; Gommers, F.J.; Henrissat, B.; Davis, E.L.; et al. Endogenous cellulases in animals: Isolation of β -1,4-endoglucanase genes from two species of plant-parasitic cyst nematodes. *Proc. Natl. Acad. Sci. USA* **1998**, *95*, 4906–4911. [[CrossRef](#)] [[PubMed](#)]
43. Rosso, M.N.; Favery, B.; Piotte, C.; Arthaud, L.; De Boer, J.M.; Hussey, R.S.; Bakker, J.; Baum, T.J.; Abad, P. Isolation of a cDNA encoding a β -1,4-endoglucanase in the root-knot nematode *Meloidogyne incognita* and expression analysis during plant parasitism. *Mol. Plant Microbe Interact.* **1999**, *12*, 585–591. [[CrossRef](#)] [[PubMed](#)]

44. Bera-Maillet, C.; Arthaud, L.; Abad, P.; Rosso, M.N. Biochemical characterization of MI-ENG1, a family 5 endoglucanase secreted by the root-knot nematode *Meloidogyne incognita*. *Eur. J. Biochem.* **2000**, *267*, 3255–3263. [[CrossRef](#)] [[PubMed](#)]
45. Yan, Y.; Smant, G.; Stokkermans, J.; Qin, L.; Helder, J.; Baum, T.; Schots, A.; Davis, E. Genomic organization of four β -1,4-endoglucanase genes in plant-parasitic cyst nematodes and its evolutionary implications. *Gene* **1998**, *220*, 61–70. [[CrossRef](#)]
46. Davis, E.L.; Hussey, R.S.; Baum, T.J.; Bakker, J.; Schots, A.; Rosso, M.-N.; Abad, P. Nematode Parasitism Genes. *Annu. Rev. Phytopathol.* **2000**, *38*, 365–396. [[CrossRef](#)] [[PubMed](#)]
47. Ledger, T.N.; Jaubert, S.; Bosselut, N.; Abad, P.; Rosso, M.N. Characterization of a new β -1,4-endoglucanase gene from the root-knot nematode *Meloidogyne incognita* and evolutionary scheme for phytonematode family 5 glycosyl hydrolases. *Gene* **2006**, *382*, 121–128. [[CrossRef](#)] [[PubMed](#)]
48. Mitreva-Dautova, M.; Roze, E.; Overmars, H.; de Graaff, L.; Schots, A.; Helder, J.; Goverse, A.; Bakker, J.; Smant, G. A symbiont-independent endo-1,4- β -xylanase from the plant-parasitic nematode *Meloidogyne incognita*. *Mol. Plant Microbe Interact.* **2006**, *19*, 521–529. [[CrossRef](#)] [[PubMed](#)]
49. Jaubert, S.; Laffaire, J.-B.; Abad, P.; Rosso, M.-N. A polygalacturonase of animal origin isolated from the root-knot nematode *Meloidogyne incognita*. *FEBS Lett.* **2002**, *522*, 109–112. [[CrossRef](#)]
50. Qin, L.; Kudla, U.; Roze, E.H.; Goverse, A.; Popeijus, H.; Nieuwland, J.; Overmars, H.; Jones, J.T.; Schots, A.; Smant, G.; et al. Plant degradation: A nematode expansin acting on plants. *Nature* **2004**, *427*, 30. [[CrossRef](#)] [[PubMed](#)]
51. Kudla, U.; Qin, L.; Milac, A.; Kielak, A.; Maissen, C.; Overmars, H.; Popeijus, H.; Roze, E.; Petrescu, A.; Smant, G.; et al. Origin, distribution and 3D-modeling of Gr-EXPB1, an expansin from the potato cyst nematode *Globodera rostochiensis*. *FEBS Lett.* **2005**, *579*, 2451–2457. [[CrossRef](#)] [[PubMed](#)]
52. Vanholme, B.; Haegeman, A.; Jacob, J.; Cannoot, B.; Gheysen, G. Arabinogalactan endo-1,4- β -galactosidase: A putative plant cell wall-degrading enzyme of plant-parasitic nematodes. *Nematology* **2009**, *11*, 739–747. [[CrossRef](#)]
53. Popeijus, H.; Overmars, H.; Jones, J.; Blok, V.; Goverse, A.; Helder, J.; Schots, A.; Bakker, J.; Smant, G. Degradation of plant cell walls by a nematode. *Nature* **2000**, *406*, 36–37. [[PubMed](#)]
54. Doyle, E.A.; Lambert, K.N. Cloning and Characterization of an Esophageal-Gland-Specific Pectate Lyase from the Root-Knot Nematode *Meloidogyne javanica*. *Mol. Plant Microbe Interact.* **2002**, *15*, 549–556. [[CrossRef](#)] [[PubMed](#)]
55. Kudla, U.; Milac, A.-L.; Qin, L.; Overmars, H.; Roze, E.; Holterman, M.; Petrescu, A.-J.; Goverse, A.; Bakker, J.; Helder, J.; et al. Structural and functional characterization of a novel, host penetration-related pectate lyase from the potato cyst nematode *Globodera rostochiensis*. *Mol. Plant Pathol.* **2007**, *8*, 293–305. [[CrossRef](#)] [[PubMed](#)]
56. Lambert, K.N.; Allen, K.D.; Sussex, I.M. Cloning and Characterization of an Esophageal-Gland-Specific Chorismate Mutase from the Phytoparasitic Nematode *Meloidogyne javanica*. *Mol. Plant Microbe Interact.* **1999**, *12*, 328–336. [[CrossRef](#)] [[PubMed](#)]
57. Jones, J.T.; Furlanetto, C.; Bakker, E.; Banks, B.; Blok, V.; Chen, Q.; Phillips, M.; Prior, A. Characterization of a chorismate mutase from the potato cyst nematode *Globodera pallida*. *Mol. Plant Pathol.* **2003**, *4*, 43–50. [[CrossRef](#)] [[PubMed](#)]
58. Vanholme, B.; Kast, P.; Haegeman, A.; Jacob, J.; Grunewald, W.; Gheysen, G. Structural and functional investigation of a secreted chorismate mutase from the plant-parasitic nematode *Heterodera schachtii* in the context of related enzymes from diverse origins. *Mol. Plant Pathol.* **2009**, *10*, 189–200. [[CrossRef](#)] [[PubMed](#)]
59. Bauters, L.; Haegeman, A.; Kyndt, T.; Gheysen, G. Analysis of the transcriptome of *Hirschmanniella oryzae* to explore potential survival strategies and host–nematode interactions. *Mol. Plant Pathol.* **2014**, *15*, 352–363. [[CrossRef](#)] [[PubMed](#)]
60. Opperman, C.H.; Bird, D.M.; Williamson, V.M.; Rokhsar, D.S.; Burke, M.; Cohn, J.; Cromer, J.; Diener, S.; Gajan, J.; Graham, S.; et al. Sequence and genetic map of *Meloidogyne hapla*: A compact nematode genome for plant parasitism. *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 14802–14807. [[CrossRef](#)] [[PubMed](#)]
61. Wybouw, N.; Balabanidou, V.; Ballhorn, D.J.; Dermauw, W.; Grbić, M.; Vontas, J.; Van Leeuwen, T. A horizontally transferred cyanase gene in the spider mite *Tetranychus urticae* is involved in cyanate metabolism and is differentially expressed upon host plant change. *Insect Biochem. Mol. Biol.* **2012**, *42*, 881–889. [[CrossRef](#)] [[PubMed](#)]

62. Danchin, E.G.J.; Guzeeva, E.A.; Mantelin, S.; Berepiki, A.; Jones, J.T. Horizontal Gene Transfer from Bacteria Has Enabled the Plant-Parasitic Nematode *Globodera pallida* to Feed on Host-Derived Sucrose. *Mol. Biol. Evol.* **2016**, *33*, 1571–1579. [[CrossRef](#)] [[PubMed](#)]
63. Craig, J.P.; Bekal, S.; Niblack, T.; Domier, L.; Lambert, K.N. Evidence for Horizontally Transferred Genes Involved in the Biosynthesis of Vitamin B-1, B-5, and B-7 in *Heterodera glycines*. *J. Nematol.* **2009**, *41*, 281–290. [[PubMed](#)]
64. Craig, J.P.; Bekal, S.; Hudson, M.; Domier, L.; Niblack, T.; Lambert, K.N. Analysis of a Horizontally Transferred Pathway Involved in Vitamin B6 Biosynthesis from the Soybean Cyst Nematode *Heterodera glycines*. *Mol. Biol. Evol.* **2008**, *25*, 2085–2098. [[CrossRef](#)] [[PubMed](#)]
65. Scholl, E.H.; Thorne, J.L.; McCarter, J.P.; Bird, D.M. Horizontally transferred genes in plant-parasitic nematodes: A high-throughput genomic approach. *Genome Biol.* **2003**, *4*, R39. [[CrossRef](#)] [[PubMed](#)]
66. McCarter, J.P.; Mitreva, M.D.; Martin, J.; Dante, M.; Wylie, T.; Rao, U.; Pape, D.; Bowers, Y.; Theising, B.; Murphy, C.V.; et al. Analysis and functional classification of transcripts from the nematode *Meloidogyne incognita*. *Genome Biol.* **2003**, *4*, R26. [[CrossRef](#)] [[PubMed](#)]
67. Finn, R.D.; Coghill, P.; Eberhardt, R.Y.; Eddy, S.R.; Mistry, J.; Mitchell, A.L.; Potter, S.C.; Punta, M.; Qureshi, M.; Sangrador-Vegas, A.; et al. The PFAM protein families database: Towards a more sustainable future. *Nucleic Acids Res.* **2016**, *44*, D279–D285. [[CrossRef](#)] [[PubMed](#)]
68. Ku, C.; Martin, W.F. A natural barrier to lateral gene transfer from prokaryotes to eukaryotes revealed from genomes: The 70% rule. *BMC Biol.* **2016**, *14*, 89. [[CrossRef](#)] [[PubMed](#)]
69. Danchin, E.G.J. Lateral gene transfer in eukaryotes: Tip of the iceberg or of the ice cube? *BMC Biol.* **2016**, *14*, 101. [[CrossRef](#)] [[PubMed](#)]



© 2017 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).