Genome Biology

METHOD Open Access

Seqrutinator: scrutiny of large protein superfamily sequence datasets for the identification and elimination of non-functional homologues

Agustín Amalfitano[1†], Nicolás Stocchi[2†^], Hugo Marcelo Atencio[3], Fernando Villarreal[2*] and Arjen ten Have[2]

†Agustín Amalfitano and Nicolás Stocchi contributed equally to this work.

^Nicolás Stocchi is deceased.

*Correspondence: fernandovillarreal@mdp.edu.ar

[1] Laboratorio de Procesamiento de Imágenes, ICyTE-CONICET-UNMdP, Mar del Plata, Argentina
[2] Computational Biology and Comparative Genomics, IIB-CONICET-UNMdP, Mar del Plata, Argentina
[3] Banco Activo de Germoplasma de Papa Andina, EEA-Balcarce INTA, Balcarce, Argentina

Abstract

Seqrutinator is an objective, flexible pipeline that removes sequences with sequencing and/or gene model errors and sequences from pseudogenes from complex, eukaryotic protein superfamilies. Testing Seqrutinator on major superfamilies BAHD, CYP, and UGT removes only 1.94% of SwissProt entries, 14% of entries from the model plant *Arabidopsis thaliana*, but 80% of entries from *Pinus taeda*'s recent complete proteome. Application of Seqrutinator on crude BAHDomes, CYPomes, and UGTomes obtained from 16 plant proteomes shows convergence of the numbers of paralogues. MSAs, phylogenies, and particularly functional clustering improve drastically upon Seqrutinator application, indicating good performance.

Keywords: Phylogenomics, Multiple sequence alignment, Phylogeny, Pseudogene, Gene model, Classification, BAHD, Cytochrome P450, UDP-glycosyltransferase

Background

General introduction

Protein superfamilies, here defined as protein families with subfamilies that have different functional characteristics, are the subject of many computational studies [1–10] and form the target of many computational platforms [11–13]. Structure-function analysis aims not only to identify which residues and/or subsequences are involved in functional diversification; it also tries to explain and predict the functional differences and can identify hitherto nondescript subfamilies [2, 3]. A large set of methods [14–16] is available and novel methods are published regularly (for review, see [17]) in a research field often referred to as phylogenomics.

The basis for many protein bioinformatics tools is formed by phylogenies and their underlying multiple sequence alignments (MSAs). Reliable methods for superfamily phylogeny reconstruction use maximum likelihood (PhyML [18], RAxML [19]; FastTree

[20]) or Bayesian inference (MrBayes [21]). MSA construction has improved significantly in recent years [22–26] but, since more complex protein families are analyzed in the post-genome era, still forms a major research area [27–29].

Only recently, attention has been paid to the automated identification and removal of sequences from Non-Functional Homologues (NFHs) [30–34]. Information that NFH sequences provide to an MSA is often considered noise that has no significant effect on the results. MSAs are often trimmed to remove not so reliable columns [35, 36], which at least in part have resulted from NFHs. This trimming has been shown to lead to improved trees [37]. We argue that besides that trimming an MSA is a loss of information, NFH sequences provide erroneous signals that hinder the correct processing of the MSA, thereby deteriorating the output.

The two major sources of NFH sequences are pseudogenes and erroneous sequences. Pseudogenes are no longer under functional constraint and not only accumulate point mutations, leading to low similarity, but also obtain inserts and or deletions of subsequences, especially when the original gene contained introns. Erroneous sequences can result from both sequencing and assembly errors as well as from incorrect gene models. Notably, recently published complete proteomes that have not yet been subjected to community corrections, often contain many incorrect gene models.

MSAs with superfamily sequences from many complete proteomes are often prohibitively long due to an accumulation of various alignment errors (e.g., sequence-specific inserts provide information that can derail proper MSA). Identification and removal of NFH sequences is therefore required but demands a huge effort on large datasets. Existing methods are either not fully objective [30], directed at improving existing MSAs by removing subsequences [32], or only remove outliers [31, 33, 34]. None of these methods is fully automated and directed at removing NFH sequences from large sequence sets in order to obtain clean datasets. Most of the existing algorithms are only tested on simulated datasets and none directs the problem of large gap regions. A likely reason is that defining inclusion thresholds is troublesome and will by definition result in both false positives and false negatives. More importantly, no real benchmark datasets of functional homologues (FHs) and NFHs exist and any attempt to construct a benchmark set will result in a set that is too restrictive.

We have developed a method for objective sequence scrutiny directed at NFH detection and removal, named Seqrutinator. The method was developed and tested by performing the sequence mining of three single-domain superfamilies in plants: cytochrome P450 (CYP), UDP-glycosyltransferase (UGT), and BAHD acyltransferases (BAHD is an acronym derived from the first characterized enzymes: benzyl alcohol O-acetyl transferase; anthocyanin O-hydroxycinnamoyl transferase; N-hydroxycinnamoyl anthranilate benzoyl transferase; and deacetylvindoline 4-O-acetyltransferase). These families form part of our major biological research interest which is to model flavonoid metabolism in potato (*Solanum tuberosum*) (see Fig. 1), for which it is crucial to functionally assign sequences to their functional subfamilies. In addition, they were selected as being challenging cases.

The three superfamilies are rather large with in between 50 and 500 functional homologues in seed plants. CYP and UGT constitute ~ 4% and ~ 8%, respectively, of the known flavonoid metabolism in potato. The BAHD acyltransferase superfamily constitutes a

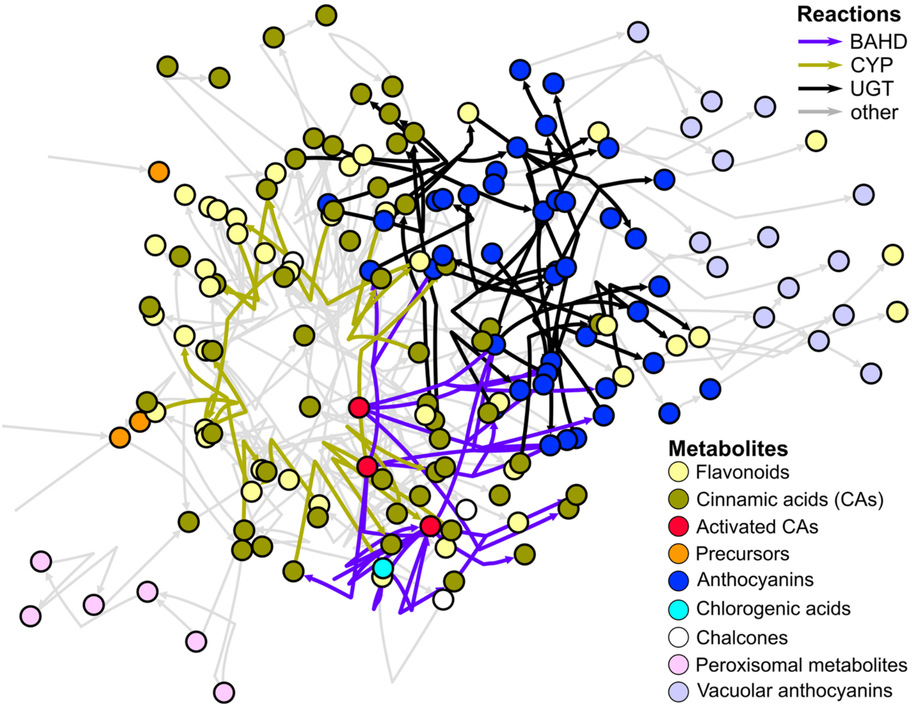


Fig. 1 Cartoon depicting the phenylpropanoid/flavonoid biosynthesis network in potato. Three major superfamilies cover approximately 25% of the enzymatic reactions. The superfamily subnetworks show a partial overlap. The BAHD and UGT superfamilies are shifted towards end-products that are conjugated and transported to the vacuole, thereby changing the effective sink

further ~ 12%, of a total of 384 known potato proteins involved in flavonoid metabolism, of which all are classified in 59 Pfam domains. Since enzymes from specialized metabolism concern secondary rather than primary metabolites, they are under less functional constraint [38]. Moreover, specialized metabolism generates a plethora of chemically related compounds via parallel, diverging and converging pathways [38], reflected in Fig. 1. These complex superfamilies result from a process in which duplications result in functional redundancy, which allows for sequence diversification that, in its turn, is required for functional diversification. Hence, these enzymes often also show high substrate permissiveness. Together, the functional constraint often acts on part of the superfamily rather than on specific paralogues. This contributes even further to sequence and functional diversification. The resulting low functional constraint explains why high evolutionary rate, high sequence diversity and complexity.

The complexity of the selected superfamilies is so high that no reliable function annotation is at hand for many homologues. CYP and UGT sequences are classified as CYP or UGT by Pfam or using a system based on the percentage of identity maintained by classification committees [39, 40]. Panther [41] assigns at the subfamily level whereas we recently developed HMMERCTTER [42], a software for the clustering and classification of protein superfamily sequences, which outperformed Panther in classifying the alpha-crystallin domain, glycosyl hydrolase 28 and phospholipase C superfamilies. The high HMMERCTTER performance depends on its 100% precision and recall (100% P&R) rule. This means that for each cluster, when screened with its cluster-specific HMMER profile, all sequences of the cluster have a HMMER score that is higher than the score of any non-cluster sequence from the superfamily sequence set. As such, it depends on sets that lack NFH sequences, such as partial or pseudogene sequences. Note that, on the one hand, Seqrutinator is a tool to assist HMMERCTTER and that HMMERCTTER, on the other hand, can be used to determine Seqrutinator performance.

Here, we present the pipeline and script, with some complementary scripts and methods. We show numerical data generated by the three case studies alongside, given the lack of a valid benchmark dataset, experiments directed at the validation of the method. We show that the MSAs of filtered datasets are significantly more reliable and that the method is flexible and robust. Most importantly, Seqrutinator shows high precision since it classified very few FHs as NFH, whereas no false negatives (NFHs classified as FH) were detected. Furthermore, we present a detailed recovery analysis that, besides that it confirms the high performance of Seqrutinator, shows the relative ease of sequence analysis once a high-quality MSA has been obtained by Seqrutinator.

Design of the pipeline and its modules

Objective

The objective of this work is to design, provide and test an objective, automated but flexible pipeline for the scrutiny of sequence sets for NFHs. It is directed at large, complex protein superfamilies from eukaryotes. Sequences, obtained by sensitive sequence mining are classified as either functional or non-functional. The NFH sequence sets can be subjected to a recovery analysis in order to prevent inadvertent false positives, i.e., FHs classified as NFH. Seqrutinator should be a method to obtain sequence sets that are representative of a functional protein superfamily. Since there are no reliable benchmark sets for the quantification of Seqrutinator's results, we cannot quantify performance by which we need to test performance in qualitative ways.

Definition of non-functional homologue

To scrutinize protein superfamily sequence sets for the presence of NFHs, we need to define NFH in terms of sequence characters. We discriminate two major classes of NFH sequences that are further subdivided. First, NFH sequences can result from either incorrect gene modeling or sequencing and/or assembling errors. Second, an NFH sequence can correspond to a pseudogene, which we define as a gene that no longer encodes its either supposed or original function.

Incorrect gene modeling and sequencing errors come with several issues. First, some sequences will lack or have additional N- or C-terminal subsequences as a result of a missed start or stop codon. Then, not all introns are identified from eukaryotic sequences whereas, on the other hand, coding subsequences that are incorrectly identified as intron form a fourth problem. Both intron issues can lead to intron-sized gaps in an MSA or to a switch in the reading frame and the untimely stop of the coding sequence.

Pseudogenes are no longer under constraint and will as such rapidly accumulate mutations. This can have two consequences. It will result in increased evolutionary distances, as can often be observed in phylogenetic trees. It may also result in the loss or gain of start and stop codons as well as splice donor and acceptor sites. As such, many pseudogenes will have issues similar to those mentioned for erroneous sequences. A last issue is that of the definition of the superfamily. Superfamilies can have related superfamilies of which some sequences may be identified in an initial, sensitive data mining but which should be identified by Seqrutinator as NFH.

Based on the above problem description we hypothesize that:

1) Relatively short sequences are unlikely functional.
2) NFH sequences with intron-derived subsequences incorrectly called as exons during gene modeling or presented by pseudogenes can instigate large continuous regions of gap-rich columns in MSAs. Incorrect identification of start or stop codons can instigate large N- or C-terminal extensions.
3) NFH sequences that lack exon-derived subsequences incorrectly called as intron during gene modeling or presented by pseudogenes can present large continuous gaps at otherwise amino acid-rich columns of MSAs. Incorrect identification of start or stop codons can instigate large N- or C-terminal gap regions.
4) Distant pseudogenes and otherwise similar sequences such as from related protein families will have low similarity and show low scores to a superfamily's HMMER profile.

Modules and algorithms were designed to find NFH sequences based on these four hypotheses. The resulting method is based on the concept of homology and it depends on MSAs, which makes Seqrutinator in principle not a valid method for the scrutiny and cleaning of protein families with different domain architectures.

The default pipeline and its modules

The fully automated Seqrutinator pipeline, described in detail in Additional file 1: Supplemental Document 1, is implemented in a larger procedure (see Fig. 2) that starts with a user-guided sequence collection and ends with user-guided recovery analyses, described in detail in Additional file 2: Supplemental Document 2. Sequence collection from multiple sequence sets is fully automated using the complementary script Multiple Fasta Aligner (MuFasA, see Additional file 1: Supplemental Document 1). This requires a HMMER [43] profile as well as initial sequence sets (e.g., complete proteomes) as input. All sequences from a complete proteome identified by the HMMER profile with a score higher than the inclusion threshold are collected in a single fasta file and aligned automatically. We recommend to align a user-selected reference sequence to each MSA, in order to remove non-homologous N- and C-terminal subsequences that may negatively interfere with the automated procedure. This is can be automatically achieved with an optional MuFasA parameter.

Seqrutinator is a flexible pipeline made of five different modules. The user can select the modules, in which order they will be implemented, and change settings that will affect the stringency of the automated scrutiny and filtration. Here, we summarize the default procedure of Fig. 2 and its reasoning. Details and optional settings are in the pipeline and module description in Additional file 1: Supplemental Document 1.

The first step is the Short Sequence Remover (SSR) directed at relatively short sequences. By default, sequences that have a length of 65% or less of the reference sequence are removed. Sequences of proteins with a resolved structure are preferred

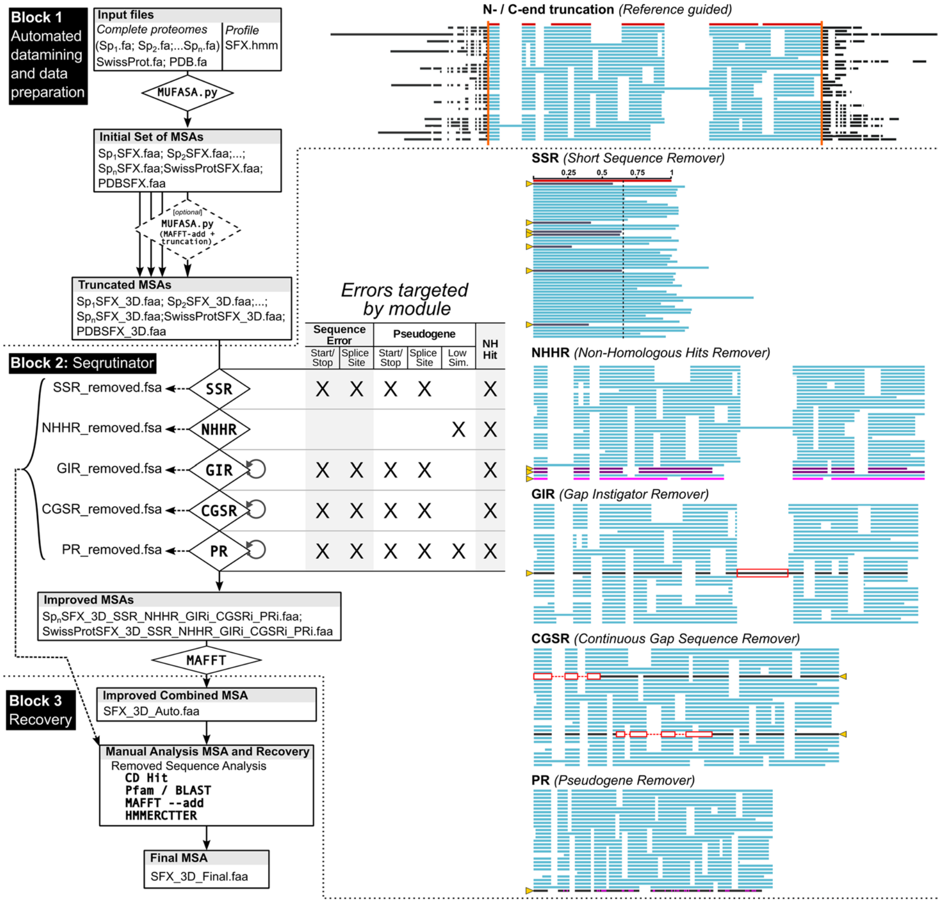

Block 1 Automated datamining and data preparation
Input files
Complete proteomes (Sp1.fa; Sp2.fa;...Spn.fa) SwissProt.fa; PDB.fa
Profile SFX.hmm
MUFASA.py
Initial Set of MSAs
Sp1SFX.faa; Sp2SFX.faa;...; SpnSFX.faa;SwissProtSFX.faa; PDBSFX.faa
[optional] MUFASA.py (MAFFT-add + truncation)
Truncated MSAs
Sp1SFX_3D.faa; Sp2SFX_3D.faa;...; SpnSFX_3D.faa;SwissProtSFX_3D.faa; PDBSFX_3D.faa
Block 2: Seqrutinator
SSR_removed.fsa
NHHR_removed.fsa
GIR_removed.fsa
CGSR_removed.fsa
PR_removed.fsa
SSR
NHHR
GIR
CGSR
PR
Errors targeted by module
Sequence Error
Pseudogene
NH Hit
Start/ Stop
Splice Site
Start/ Stop
Splice Site
Low Sim.
X X X X X
X X
X X X X X
X X X X X
X X X X X X
Improved MSAs
SpnSFX_3D_SSR_NHHR_GIRi_CGSRi_PRi.faa; SwissProtSFX_3D_SSR_NHHR_GIRi_CGSRi_PRi.faa
MAFFT
Block 3 Recovery
Improved Combined MSA
SFX_3D_Auto.faa
Manual Analysis MSA and Recovery
Removed Sequence Analysis
CD Hit
Pfam / BLAST
MAFFT --add
HMMERCTTER
Final MSA
SFX_3D_Final.faa
N- / C-end truncation (Reference guided)
SSR (Short Sequence Remover)
0.25
0.5
0.75
1
NHHR (Non-Homologous Hits Remover)
GIR (Gap Instigator Remover)
CGSR (Continuous Gap Sequence Remover)
PR (Pseudogene Remover)

Fig. 2 Schematic of the procedure with default seqrutinator pipeline. The workflow for protein superfamily sequence mining consists of three blocks (left). Block 1 concerns the preparation of the input for the automated Seqrutinator pipeline in the second block. Block 2 illustrates Seqrutinator's modules in default order, including eventual iterations indicated by circular arrows and described in the main text and Additional file 1: Supplemental Document 1. NH Hit means non-homologous hit. The various "_removed.fsa" are archives with the removed sequences for each of the modules that can be analyzed in block 3, directed at the identification and recovery of inadvertently removed FH sequences. The schematic MSAs on the right show the truncation of the MSA in block 1 and, for each of the five modules of the automated pipeline, which sequences (indicated by triangles) are removed and why

since these typically concern final, active proteins, resulting from possible post-translational modifications such as prepeptide cleavage.

The second step is the Non-Homologous Hit Remover (NHHR). It constructs a HMMER profile and screens all sequences by hmmsearch [43]. Outlier sequences with a HMMER score below Q1-1.5 IQR are removed.

The third module in the default pipeline is the Gap Instigator Remover (GIR) which removes sequences that instigate large gaps in MSAs. By default, the GIR removes sequences that induce regions of 30 or more continuous gap columns, where a gap column is defined as a column occupied by at least 90% of gaps. A 10% of residues is allowed since certain residues of FH sequences may have become aligned to the insert. The default region size is set according to the minimal intron size observed among most eukaryotes but may depend on the organism [44]. Note that sequences with long

N- and/or C-terminal subsequences are also removed, by which truncation of the input set is highly recommended. GIR removes sequence-by-sequence using realignment and iteration.

The fourth module is the Continuous Gap Sequence Remover (CGSR). This removes sequences that show one or more instances of large continuous gaps in the MSA. Again, the default setting is at 30 columns, provided that columns are occupied with less than 50% gaps. This is in order to allow subfamily-specific subsequences. CGSR removes sequences in a threshold-controlled batch, as described in detail in Additional file 1: Supplemental Document 1, and is iterated. Note that not all sequences that lack a subsequence due to incorrect gene modeling will be detected. Residues that enclose the gap caused by the absent subsequence may align to any of the columns of the gap, thereby splitting this into two or more continuous gap regions. These instances are not detected.

The last module in the default pipeline is the Pseudogene Remover (PR) which is identical to the NHHR module except that it is iterated. The names of the NHHR and PR modules are as such based on the intent of the modules. Although there is no clear-cut threshold that can discriminate between a non-homologous hit and a pseudogene, the first is expected to be less similar but more disturbing. As such, NHHR is by default the first module and not iterated. PR is iterated since pseudogene identification is more delicate and works with a much improved MSA that corresponds with a different HMMER score distribution. As a result, PR is more sensitive than NHHR and by default the last module of the pipeline.

We recommend performing a recovery analysis as block 3 since certain FH sequences may be inadvertently removed. The Seqrutinator output provides data and graphs that show the removal of sequences in detail and can assist in this analysis.

Design of the performance analysis

The performance of a binary classifier such as Seqrutinator is usually determined in terms of P&R. Since, as stated before, there are no reliable benchmark sets for the quantification of Seqrutinator's results, we need to incorporate different measures and experiments to show Seqrutinator performance. We designed experiments to show behavior and determine consistency and performance using mostly indirect measures.

UniProtKB/Swiss-Prot (SwissProt) is, to the best of our knowledge, the most appropriate dataset for semiquantitative benchmarking of the pipeline. It contains a large number of sequences that come with biochemical and/or transcript evidence alongside sequences that lack wetlab evidence. More importantly, even the most stringently curated sequences, those with protein evidence do include partial sequences and pseudogenes. Hence, we used SwissProt expecting that few entries will be NFH sequences which should be reflected in Seqrutinator results of three SwissProt homologue sets.

We also tested the pipeline by performing comparative sequence scrutiny and cleaning of the same three superfamilies in 16 complete plant proteomes with different degrees of quality. The three superfamilies have many homologues to allow for the statistical approach used in the outlier modules. Automated NFH sequence identification of three superfamilies in 15 land plants and the algae *Chlamydomonas reinhardtii*, used as an outgroup, allows for a comparative analysis that will shed light on performance. This is based on the hypothesis that performance on complete proteomes depends on the quality of the complete proteome rather than the superfamily that is analyzed. Each of the applied algorithms should identify no or only a few NFH sequences in a high-quality sequence set such as the complete proteome from the model organism *Arabidopsis thaliana* (TAIR10). On the other hand, less curated complete proteomes (that have either been published recently or do not count on a large research community) are more likely to contain many NFH sequences. Indeed, there are large differences between the number of sequences of the complete proteomes, and although different plants will have different numbers of functional paralogues, we foresee that Seqrutinator not only merely removes sequences but will also result in a convergence of the number of retained sequences throughout the process. For *A. thaliana*, we included TAIR v6, besides the latest and supposedly superb set of TAIR v10. The superfamily analyses will be published elsewhere; here, we report the numerical data from the NFH sequence identification to show the performance of the methods.

Finally, we considered how to detect false positives, i.e., inadvertently removed FH sequences. To do so, we must understand the method and the biases of the initial datasets. For instance, a complete proteome from a single organism has a different bias than the SwissProt sequence set or all sequences from a single Pfam seed alignment.

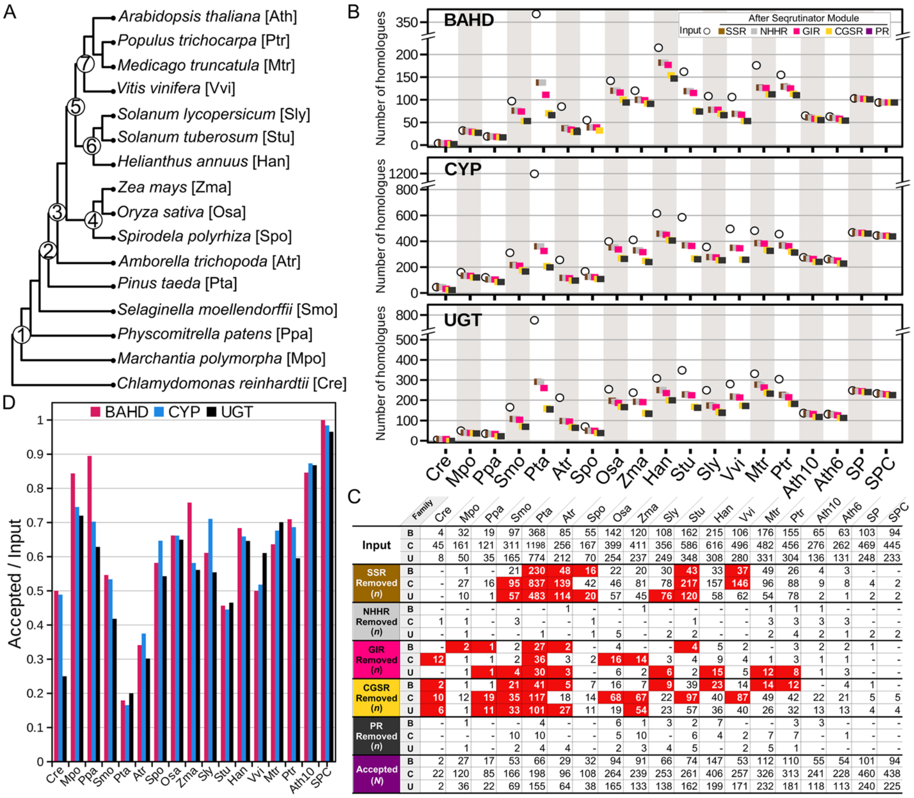
A first minor concern is that of sequences that have been removed by the PR module because of a biased HMMER profile. Although HMMER profiles are weighted, they cannot account for large differences in clade distance. In a superfamily with various equidistant subfamilies and a single, more distant subfamily, sequences of the distant subfamily will show low scores in a hmmsearch and may be inadvertently removed by the PR. This problem is exacerbated when working with sequence sets from complete proteomes, which come with divergent MSAs and HMMER profiles with low P&R. These inadvertently removed sequences can be identified by recovery analyses of the combined sequences that were removed from the various species sequence sets.

Another concern is that of taxon-specific sequences. Complex superfamilies show a high rate of evolution by which taxon-specific sequences are expected. Any type of mutation can result in a novel functional subfamily and as such functional sequences can have been removed by GIR, CGSR, or PR. Cluster analysis is likely to fail when the sequence is taxon-specific; hence, additional sequence mining in a specific part of sequence space (e.g., genus, family or order) that is to be analyzed may be required.

Results

Seqrutinator is consistent and performs well on 19 complete proteomes

We subjected 19 sequence sets to sensitive HMMER searches with HMMER profiles for BAHD, CYP, and UGT using MuFasA. This resulted in a total of three times 19 sequence sets representing the crude BAHDomes, CYPomes, and UGTomes of the 16 plant species (Fig. 3A), of which *A. thaliana* is represented by two versions (v6 and v10), as well as two SwissProt plant sequence sets (standard and the more strict subset with protein/transcript evidence, from here on referred to as curated). All crude sequence sets were then prepared for and subjected to the Seqrutinator pipeline using default settings and default order of SSR-NHHR-GIR-CGSR-PR, also indicated as 12345. The numbers of input and retained homologues after each module were recorded (Additional file 3: Supplemental Table S1) and are shown in Fig. 3B. This shows similar patterns for the



	Family	Cre	Mpo	Ppa	Smo	Pta	Atr	Spo	Osa	Zma	Sly	Stu	Han	Vvi	Mtr	Ptr	Ath10	Ath6	SP	SPC
Input	B	4	32	19	97	368	85	55	142	120	108	162	215	106	176	155	65	63	103	94
	C	45	161	121	311	1198	256	167	399	411	356	586	616	496	482	456	276	262	469	445
	U	8	50	35	165	774	212	70	254	237	249	348	308	280	331	304	136	131	248	233
SSR Removed (*n*)	B	-	1	-	21	230	48	16	22	20	30	43	33	37	49	26	4	3	-	-
	C	-	27	16	95	837	139	42	46	81	78	217	157	146	96	88	9	8	4	2
	U	-	10	1	57	483	114	20	57	45	76	120	58	62	54	78	2	1	2	2
NHHR Removed (*n*)	B	-	-	-	-	-	1	-	-	1	-	-	-	-	1	1	1	-	-	-
	C	1	1	-	3	-	-	1	-	-	-	1	-	-	3	3	3	3	-	-
	U	-	1	-	-	1	-	1	5	-	2	2	-	-	2	4	2	1	2	2
GIR Removed (*n*)	B	-	2	1	2	27	2	-	4	-	-	4	5	2	-	3	2	2	1	-
	C	12	1	1	2	36	3	2	16	14	3	4	9	4	1	3	1	1	-	-
	U	-	1	1	4	30	3	-	6	2	6	2	15	5	12	8	1	3	-	-
CGSR Removed (*n*)	B	2	1	1	21	41	5	7	16	7	9	39	23	14	14	12	-	4	1	-
	C	10	12	19	35	117	18	14	68	67	22	97	40	87	49	42	22	21	5	5
	U	6	1	11	33	101	27	11	19	54	23	57	36	40	26	32	13	13	4	4
PR Removed (*n*)	B	-	1	-	-	4	-	-	6	1	3	2	7	-	-	3	3	-	-	-
	C	-	-	-	10	10	-	-	5	10	-	6	4	2	7	7	-	1	-	-
	U	-	1	-	2	4	4	-	2	3	4	5	-	2	5	1	-	-	-	-
Accepted (*N*)	B	2	27	17	53	66	29	32	94	91	66	74	147	53	112	110	55	54	101	94
	C	22	120	85	166	198	96	108	264	239	253	261	406	257	326	313	241	228	460	438
	U	2	36	22	69	155	64	38	165	133	138	162	199	171	232	181	118	113	240	225

Fig. 3 Seqrutinator performance on 19 BAHDomes, CYPomes, and UGTomes. **A** Taxonomy of selected species. 1, embryophytes (land plants); 2, spermatopsida (seed plants); 3, angiosperms (flowering plants); 4, monocots; 5, eudicots; 6, asterids; 7, rosids. **B** Numbers of BAHD, CYP, and UGT homologues per species found (input) and retained after each step of the default Seqrutinator pipeline. **C** Number of removed sequences. Shown are the numbers of the initial and finally accepted sequences as well as the number of removed sequences, per module and superfamily (B: BAHD, C: CYP and U: UGT). Red shading indicates a proportionally high number of NFH was removed (see also main text and Additional file 3: Supplemental Table S1, SSR, GIR, and CGSR only). **D** Seqrutinator performance for BAHDomes, CYPomes, and UGTomes. Bars show the proportions of the number of finally accepted sequences over the number of initial sequences. Species in **B**, **C**, and **D** are presented by three letter codes according to A. Ath10 and Ath6 indicate proteome versions 10 and 6 of Ath. SP, SwissProt; SPC, SwissProt Curated

selection of FH sequences from the three superfamilies. This suggests Seqrutinator's performance is consistent.

The table in Fig. 3C shows how many sequences were removed by each module from each dataset. Shown are the numbers of removed sequences *n*, for each individual BAHDome, CYPome, and UGTome and module. We then calculated *n*/*N*, where *N* is the number of accepted sequences. The numbers *n* are highlighted in red if the corresponding *n*/*N* proportion is higher than the average for that module and superfamily, which as such indicates if a species has relatively many NFH of a certain superfamily that are removed by that particular module (only for SSR, GIR, and CGSR since the NHHR and the PR module remove too few sequences for meaningful comparison). As hypothesized, relatively few sequences were removed from the *A. thaliana* datasets (see also Fig. 3D). No real difference was found comparing v6 with the more recent v10. The *Marchantia polymorpha*, *Physcomitrella patens* (moss), *Spirodela polyrhiza* (duckweed), *Oryza sativa* (rice), *Zea mays* (corn), *Helianthus annuus* (sunflower), *Solanum lycopersicum* (tomato), *Vitis vinifera* (grape), *Medicago truncatula (*barrel clover), and *Populus trichocarpa* (black cottonwood poplar) datasets appear to have intermediate numbers of NFHs (< 50% for all three superfamily datasets). The remainder of the complete proteomes consistently show high numbers of NFHs (at least 40% but mostly > 50%). The most particular is *Pinus taeda* (loblolly pine), with over 80% NFHs. It is noteworthy that the majority of the NFHs are detected and removed by SSR and CGSR; intermediate and few numbers of NFH were detected by GIR and by NHHR and PR, respectively. Hence, the performance of the modules is largely explained by the provided dataset rather than by the superfamily. This is in correspondence with our hypothesis and indicates that Seqrutinator has a consistent performance.

The most effective step is SSR, which indicates that many sequences, in particular from the complete proteome of *P. taeda*, are partial. Both versions of the *A. thaliana* complete proteome appear with a few partials (4, 9, and 2 for BAHD, CYP, and UGT, respectively for v10). CGSR also removed many sequences, which results from the relaxed default setting of SSR (< 65% of length reference sequence). GIR was the third most effective module and removed only a few sequences per dataset, except for *P. taeda* and the algae *C. reinhardtii*. It appears to remove relatively many UGT superfamily sequences from dicotyledonous plants which suggests this case comes with false positives, which should be noted in the recovery analysis.

Figure 3D shows how many sequences are classified as FHs, relative to the initial number of sequences, for each species and each superfamily. The complete proteome of *A. thaliana* appears as the best, with 85, 87, and 86% of each of the originally identified BAHD, CYP, and UGT sequences classified as functional. On the other hand, the recently published complete proteome of *P. taeda* appears as very poor with a mere 18, 16, and 20% of the sequences classified as functional.

Consistency should also be found by comparing the proportions of classified sequences over initial sequences for each species. Since low initial numbers lead to large variations of this proportion, we only checked for species with at least 50 initial sequences for all three superfamilies. The largest differences in sequence removal among superfamilies are found for *Zea mays* (corn, 76, 58, and 56%) and *S. lycopersicum* (61, 71, and 55%). These are normal fluctuations that result from differences in the initial number of homologues per superfamily. For example, the *S. lycopersicum* crude CYPome has much fewer sequences than the crude CYPome of closely related *S. tuberosum*, 356 and 585 respectively, while the numbers of finally accepted CYP homologues are similar, 267 and 253. *Z. mays* has relatively few BAHD homologues, 120, of which 91 are functional. *O. sativa* has 142 homologues of which 94 are considered as functional. Hence, although the numbers of removed sequences can show large differences between species, the final numbers of accepted homologues per superfamily are similar by which Seqrutinator results in converging numbers of homologues, indicating good performance.

Most importantly, the SwissProt sequence set appears to have only a few sequences that were detected as NFH (4.6, 2.8 and 7.9% per superfamily (curated SwissProt dataset see Additional file 3: Supplemental Table S1)). This suggests the method is not overzealous in removing sequences but recovery analysis will have to show if all these removed sequences are indeed not functional.

Removal of NFH sequences results in improved MSA quality

One of the results of Seqrutinator should be a sequence set that can be aligned with improved fidelity. Several methods can be used to calculate the quality of an MSA. The sum-of-pairs [45] and TCS [46] are the most prominent measures but do not provide highly discriminative scores, which makes them poor benchmarking methods. Hence, we sought alternative methods. A simple method is to look at the length of the MSA, as compared to the length of the mature protein. Although this is a quantitative measure, MSA length does not accurately describe its quality since, for instance, a single large insert leads to a large MSA but does not necessarily result in either a good or a bad MSA.

A more sophisticated albeit indirect method is to determine the number of reliable columns using trimmed MSAs. Since MSAs of complex superfamilies by definition have regions that are either specific to certain subfamilies (i.e., not truly homologous) or not too reliable, MSAs are usually trimmed before phylogenetic reconstruction. Trimming tools such as BMGE [35] or trimAl [36] remove columns with either high amounts of gaps or high entropy. As such, the length of an MSA following trimming conceptually reflects the number of reliable columns and can as such be used as a quality measure to compare MSAs of the same or, as in this case, similar datasets. Figure 4A shows the number of columns of MSAs built throughout the application of Seqrutinator. The number of reliable columns increases in almost all steps for all species and the three superfamilies. Besides that the MSAs improve, typically with a factor of 2 to 4, there appears to be a convergence of trimmed MSA length. Note that scrutiny and filtration of the non-seed plants datasets end with only a few sequences (Fig. 3), by which the MSAs show low complexity and are typically larger following trimming (Fig. 4A). Another interesting detail is that, among a few others, the initial *P. taeda* MSAs of all three superfamilies appear completely unreliable. On the other end, the trimmed MSAs from the SwissProt and *A. thaliana* datasets increase little in length. In general, the largest increases in trimmed MSA length occur following SSR and CGSR.

We also checked if the presence of NFHs negatively affected the MSA processing. We removed all NFH sequences detected by Seqrutinator from the initial MSAs and removed the resulting 100% gap columns. The resulting pseudo-MSAs have the same sequences as the final MSAs produced by Seqrutinator but they were aligned in the presence of NFHs. Figure 4B and C show that the number of reliable columns, as determined by BMGE, is generally lower in the pseudo-MSAs and this effect was statistically significant ($p < 0.001$, Wilcoxon signed-rank test). Thus, Seqrutinator and removing NFHs results in sequence sets that show significant improvement in the quality of MSAs.

Comparison of Seqrutinator module performance

The performance of the five Seqrutinator modules was analyzed. First of all, we wondered how changing the cut-off threshold of the outlier modules affects the results. We used the customizable empirical rule of probability distribution threshold (mean – $\alpha^*\sigma$, where σ stands for standard deviation) We compared 3σ, which we applied as default in this study, with the more stringent 2.35σ, which corresponds with 95% inclusion according to a normal distribution. We also performed analyses with different pipelines in

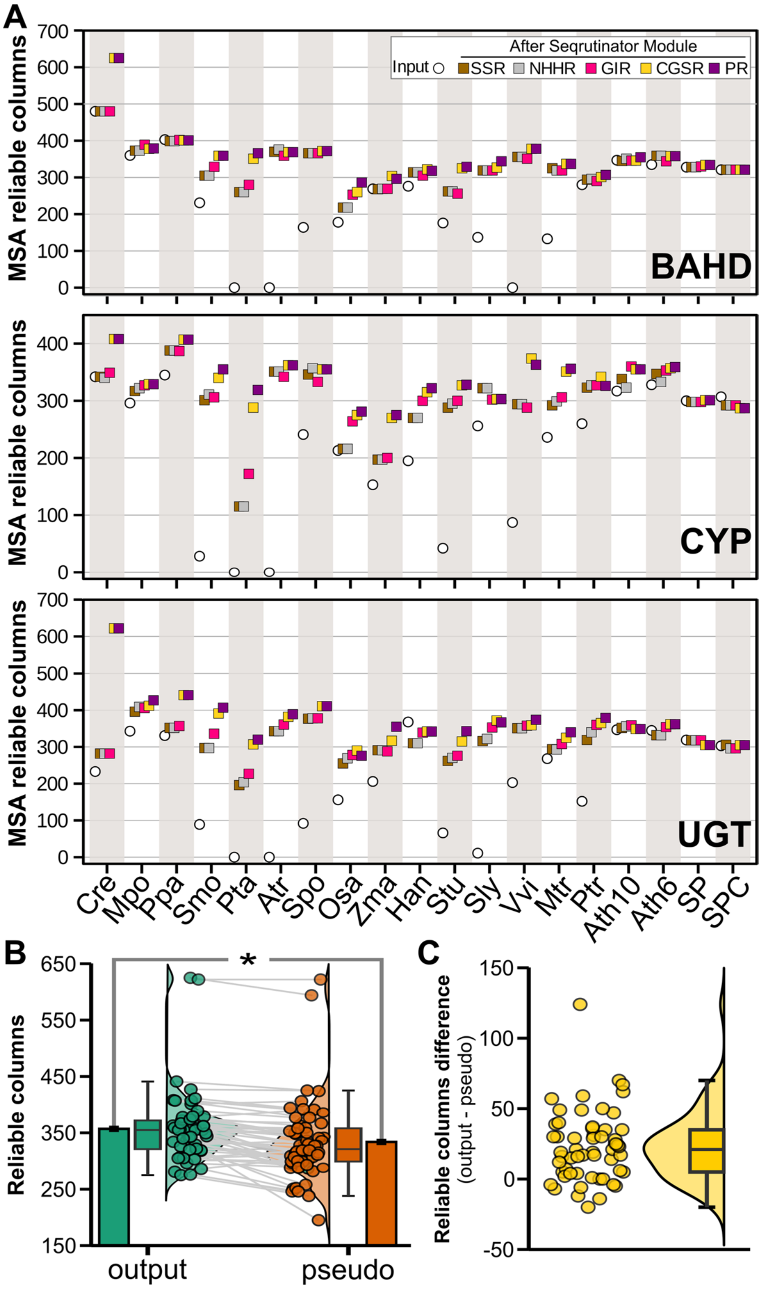
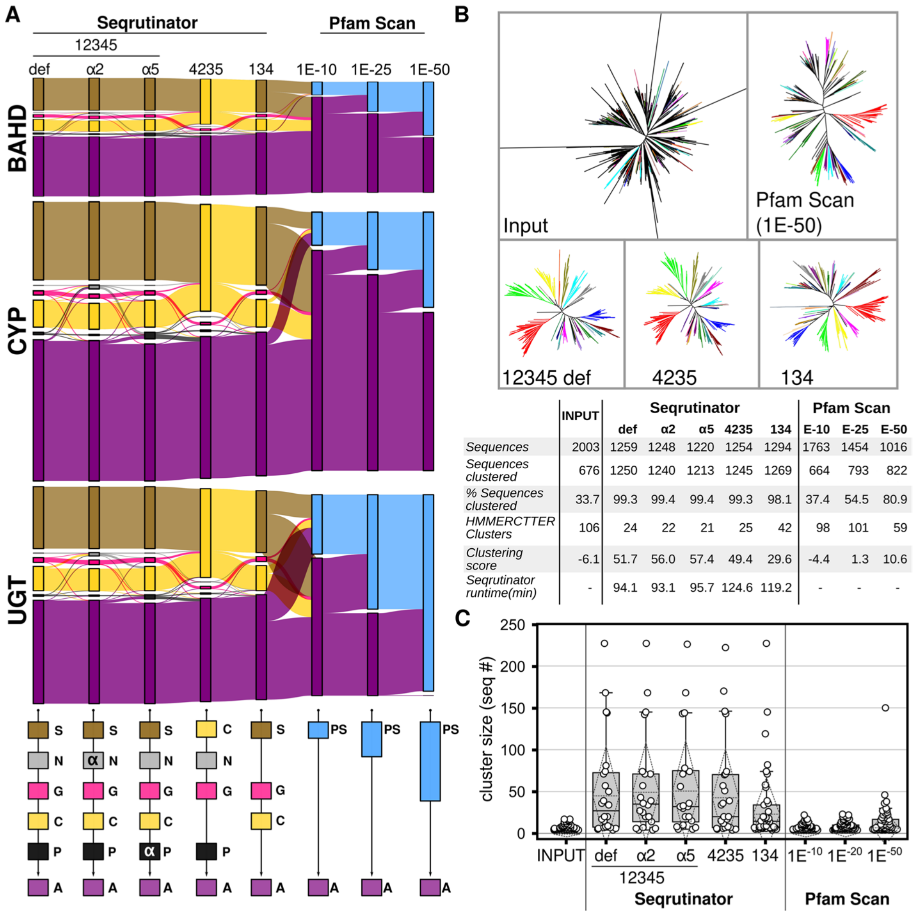


Fig. 4 Non-functional homologue sequences negatively affect MSA processing. **A** The number of reliable columns was computed with BMGE, entropy setting 0.8, for the MSAs made with the sequences before Seqrutinator (input) or the sequences retained after each module in the default pipeline. Species codes are as in Fig. 3. **B** Bar, boxplot, and raincloud density representation of number of reliable columns of the final output MSAs and the pseudo-input MSAs of the BAHD, CYP, and UGT cases. The pseudo-input MSA was obtained by removing all NFH sequences and subsequently all gap columns from the input MSAs. Gray lines connect output with corresponding pseudo-input set. * indicates significant difference with $p < 0.001$ (Wilcoxon signed-rank test). **C** Density and boxplot showing differences in reliable columns between pseudo-input and output MSAs of the BAHD, CYP, and UGT cases

order to determine if different modules can detect the same NFHs. We tested pipe 4235 to see whether CGSR can replace SSR and at what cost. We tested pipe 134 to see the effect of omitting outlier removal and to test if GIR and CGSR detect outliers. Moreover, Pfam scans with different cut-off thresholds were included as an external reference in order to shed light on performance in terms of P&R. Figure 5 summarizes the numbers as well as how the different pipelines affect phylogeny and clustering.

The Alluvial plot in Fig. 5A shows the correlation of the fate for every sequence among the Seqrutinator modules and accepted sequences, compared to the fate in three Pfam Scans. Although most removed sequences are always removed by the same module, we also observe most of all possible module swaps among the different pipelines. For instance, in the 4235 pipe, CGSR not only takes care of all sequences



	INPUT	Seqrutinator					Pfam Scan		
		def	α2	α5	4235	134	E-10	E-25	E-50
Sequences	2003	1259	1248	1220	1254	1294	1763	1454	1016
Sequences clustered	676	1250	1240	1213	1245	1269	664	793	822
% Sequences clustered	33.7	99.3	99.4	99.4	99.3	98.1	37.4	54.5	80.9
HMMERCTTER Clusters	106	24	22	21	25	42	98	101	59
Clustering score	-6.1	51.7	56.0	57.4	49.4	29.6	-4.4	1.3	10.6
Seqrutinator runtime(min)	-	94.1	93.1	95.7	124.6	119.2	-	-	-

Fig. 5 Seqrutinator is robust and flexible. **A** Sequence fate in different pipelines. Top: Alluvial plot showing the fate of initial BAHD, CYP and UGT representatives (2003, 6782, and 3994 sequences respectively from 16 species sets (Ath10 for *A. thaliana*) and the curated SwissProt set) in different pipelines of following Pfam scans with cut-off thresholds as indicated. Bottom: Schematic illustration of applied pipelines (S: SSR (1); N: NHHR (2); G: GIR (3) C: CGSR (4); P: PR (5); PS: Pfam Scan; and A: accepted). α2 and α5 indicate pipes with the more strict 2.35σ cut-off in NHHR and PR, respectively. **B** HMMERCTTER clustering of BAHD sequence sets. Top: Cluster-wise colored maximum likelihood trees and HMMERCTTER partitions of five BAHD sequence-sets as indicated: Input: partition of initial sequences; Pfam: partition of sequences obtained with most significant Pfam scan (expect value 1E-50); 12345 def: partition of sequences accepted by default Seqrutinator pipeline; 4235 and 134: partitions of sequences accepted by alternative Seqrutinator pipelines. Each cluster is automatically assigned a different color, black leaves are unclustered sequences or orphans. Bottom: Numerical abstract of clustering analysis of all nine tested datasets. Shown are the total number of sequences, the number and percentage of clustered sequences, the number of clusters and the cluster scores ((Clustered sequences-Orphans)/Total Sequences). **C** Boxplots of cluster sizes of obtained HMMERCTTER partitions. The dotted lines show the mean and the standard deviation

that SSR removes in the default pipeline; it also removes a number of sequences that are normally removed by GIR. This confirms the idea that many NFH sequences show more than one of the initially described issues. Applying a more stringent outlier cut-off for PR results in a significant increase of outliers.

The comparison with the Pfam scans also sheds light on performance. The original sequence sets were obtained with Pfam profiles but included all sequences with a score above HMMERs inclusion threshold. Pfam normally applies a more strict gathering threshold for each profile defined by a bitscore and corresponding *E* value that includes all sequences from the seed alignment. We applied cut-offs with different increasing levels of stringency. The most striking result is obtained with the most stringent Pfam scan at 1E-50. In the CYP case, it still included many sequences Seqrutinator tagged as NFH; for BAHD, it yielded a rather similar result as Seqrutinator, while only 12 sequences were accepted as FH for the UGT sequence sets. Intriguingly, one of the sequences accepted by Pfam is normally removed by the SSR and another one by the CGSR module. This last detail cannot be observed in this particular alluvial plot due to the order of the datasets we applied in the figure. Hence, Pfam Scans suffer from poor P&R, which is a recurrent issue in clustering and classification and related to the fact that Pfam has single, permissive profiles for complex superfamilies.

HMMERCTTER clustering is a method in which superfamily sequences are clustered based on phylogeny and a HMMER score cut-off that is determined to include all sequences of a monophyletic clade [42]. Only clusters with 100% P&R are accepted, and as such HMMERCTTER clusters are conserved. The presence of less conserved, low scoring NFH sequences in a monophyletic clade often prevents the clade to be accepted as 100% P&R cluster A sequence set that lacks NFH sequences should therefore result in larger and fewer clusters. As such, we performed unguided HMMERCTTER clustering on the BAHD datasets and compared the resulting partitions (Fig. 5B).

As expected, HMMERCTTER clustering of the crude dataset results in a very poor partition, with more orphans than clustered sequences as shown by the negative cluster score *((Clustered sequences-Orphans)/Final Sequences)*. Given the strict cut-off of HMMERCTTER, many partial sequences do end up as orphan sequence. The Pfam 1E-50 tree shows fewer orphans and has a better cluster score. The best cluster score (57.4, see table in Fig. 5B) is obtained with the dataset that results from the 1234α5 pipeline in which a substantial number of low-scoring outliers has been removed. The default pipeline has a cluster score of 51.7, which reflects a trade-off between a larger number of accepted sequences and a slightly lower number of clusters. As compared to the default pipeline, the 4235 pipeline has a slightly lower performance which, combined with its substantially longer runtime, (see table in Fig. 5B) indicates that the default pipeline is preferred. Not applying any outlier module (pipe 134) results in a substantially reduced cluster score of 29.6, even though only a few outliers are detected at the setting of 3σ.

Figure 5B shows that, based on the above-made assumption, the application of Seqrutinator in superfamily sequence mining results in largely improved datasets. This is also reflected by the distribution of cluster size (Fig. 5C). Although the strict Pfam scan removes more sequences than Seqrutinator, it has much smaller clusters than any of the Seqrutinator-derived datasets (Fig. 5C).

Recovery screen

Benchmark analysis, in which one determines P&R, is not feasible since that would imply demonstrating that predicted NFHs are not functional, which is principally impossible. We can, however, analyze with other analyses if sequences identified as NFH are likely functional or not. We performed various analyses in order to identify eventually removed FH sequences. We used the SwissProt datasets as the gold standard, albeit that even the curated SwissProt dataset has both partial sequences and pseudogenes. The sequences removed from the 16 species proteomes were analyzed using tools such as hmmscan and the Pfam database, in order to verify correctly removed NFHs. On the other hand, we used CD-Hit clustering to identify incorrectly removed sequences, such as those from distant or small subfamilies. Additional file 2: Supplemental Document 2 describes the details and some more profound analyses that may be required to corroborate whether a certain removed sequence is an NFH or an FH. Here, we give a summary of the results of our recovery analyses in order to show Seqrutinator's performance only.

Out of 820 SwissProt sequences among three superfamilies, 21 sequences (~2.5%) were identified as NFH: 19 were removed from the complete and 18 from the curated set. Analyses suggest only two of the 21 removed SwissProt sequences may encode functional enzymes. GIR removed a single sequence from the BAHD set based on a 65 amino acid insert. This pseudogene sequence, included in SwissProt based on homology, lacks both the strictly conserved HxxxD and the highly conserved DFGWG motifs. NHHR removed two GT28 glycosyl transferase sequences from the UGT set. These are homologous to but not part of the UGT superfamily. SSR and CGSR removed 6 and 12 sequences, respectively. Alignment of these sequences to the MSA obtained for the accepted SwissProt sequences showed 16 of these were partial, while Q43078 and Q9LNE6 were considered as putative incorrectly removed sequences.

Q43078, removed from the CYP set by CGSR, has an internal gap in the MSA made against the accepted SwissProt sequences, hence its removal. However, a BLASTP against SwissProt did not show a large gap, suggesting the sequence represents an FH and is a false positive. The sequence was included in SwissProt based on transcript evidence.

Q9LNE6, removed from the UGT set by CGSR, appears to lack an N-terminal sequence according to the MSA made against the accepted SwissProt sequences. However, BLASTP against Reference Proteins identified a number of homologues that show good global alignment, which suggests this also concerns an FH and a second false positive. The inclusion of this sequence in SwissProt is supported by evidence at the protein level.

Since analysis suggests Seqrutinator removes only a few putative FHs, we wondered if the cut-off we applied, 3σ, was sufficiently strict and we performed outlier analysis with a 2σ threshold. This removed a number of FHs from the BAHD and the CYP family, rather than additional NFHs. Interestingly, this reflected problems with the default pipeline outputs from the species BAHDomes and CYPomes.

For BAHD, it concerned three sequences from the ECERIFERUM or CER subfamily, which is distant and lacks the generally conserved DFGWG motif. CD-Hit clustering of 35 sequences removed from 16 species BAHDomes by NHHR or PR identified 17 sequences as members of the CER subfamily. These we consider as false positives, while another 40 CER subfamily sequences were not identified as NFH. The other 18 removed sequences were confirmed as NFH.

In the CYP case, the strict 2σ cut-off of NHHR identified 17 allene oxide synthase sequences from SwissProt of which one is a partial that is normally removed by CGSR. Then, out of 76 outlier sequences removed from the species CYPomes, 48 were allene oxide synthase subfamily members, which concerns all AOS homologues. In addition, we identified a single C-22 sterol desaturase as FH. The remaining 27 outliers were true positives.

As stated above, the standard 3σ NHHR module did correctly remove two GT28 glycosyl transferases from the UGT SwissProt set. All 61 GT28 sequences were identified from the separate proteome sets albeit by NHHR, GIR, CGSR, or PR. In addition, the outliers contained seven epimerases. All were correctly removed as true positives. CD-Hit did identify a group of three sequences that appears to form a subfamily found in Solanaceae only. These are likely false positives.

In the BAHD case, we identified a subfamily that was inadvertently removed by GIR. It concerns 11 sequences that instigate a large gap in the MSA. Furthermore, we encountered four sequences that have non-homologous inserts at what appears as a hot spot of acceptable insertions. Finally, we analyzed sequences removed by CGSR. Here, it concerned between 200 and 500 sequences per superfamily. These were clustered by CD-Hit and we used MSAs of the large clusters to confirm these were partials. Most sequences lack N and/or C terminal subsequences whereas clustered sequences that had complete N- and C-termini showed consistently large gaps that were not conserved. As stated above, detailed analyses are described in Additional file 2: Supplemental Document 2.

The effect of the data on the performance

We also analyzed if the size of the sequence sets affects performance. Statistics on larger sequence sets are more reliable and these should therefore show improved performance. However, particularly using MAFFT global alignment, small increases in number of sequences come with large increases in computational cost. We tested the effect of sequence subset size using four different size conditions and randomized sequence sets. The analyses, presented in detail in Additional file 4: Supplemental Document 3 show that only the PR module is affected by size, removing more sequences when sequences are presented in larger sets. In general, a similar number of mostly the same sequences is removed.

Discussion

We present a flexible pipeline to clean superfamily protein sequence sets with the objective of obtaining sets encompassing most FH sequences and from which most NFH sequences have been removed. Since we cannot test for non-functionality, we cannot perform classic benchmark analysis showing P&R. Hence, we used other, indirect methods to gain insight into consistency and performance of Seqrutinator. This implies that Seqrutinator should not be seen as a method to determine whether a certain sequence corresponds to a FH. Seqrutinator is also likely to fail or make mistakes in the classification of spliceoforms.

We defined NFH sequences in order to provide an objective basis for sequence removal. Based on these definitions and the inherent characteristics of for instance pseudogenes, we expected that certain sequences could be removed by different modules. This we confirmed by comparing different pipelines (Fig. 5A) and by using randomized datasets (Additional file 4: Supplemental Document 3). Seqrutinator behavior sometimes depends on the dataset. The standard pipeline removed 19 and 18 sequences from the two SwissProt datasets, which correspond to 21 different sequences. The analyses in Additional file 4: Supplemental Document 3, where we compared different datasets obtained by randomization, also show a small number of sequences having different fates. This suggests Seqrutinator has only few false positives and few false negatives.

Another premise we made is that certain complete proteomes, such as from model plant *A. thaliana*, are superior in quality to, for instance, recently completed proteomes, and that this should be reflected by similar performance on the three different superfamilies. The data demonstrated in Fig. 3 show this premise to be true: Seqrutinator removed few sequences from *A. thaliana* and many from *P. taeda*, which sequence was obtained in 2017 [47]. Seqrutinator is also consistent since it removes similar percentages of sequences per species. Hence, although there is no quantitative measure for complete proteome quality, Seqrutinator's performance is in line with what was expected.

Application of Seqrutinator results in a convergence of the number of FH sequences per superfamily when comparing the different species, albeit that there are outliers. Non-seed plants have much smaller genomes and less specialized metabolism than seed plants. Outliers among seed plants are *Amborella trichopoda* that has relatively few final sequences and *H. annuus* that has relatively many final sequences. This can be explained by a relatively small and rather large genome size, respectively. The other seed plants do show variation but have rather similar amounts of accepted sequences for a specific superfamily. In all cases, performance on the three superfamilies is similar. Under the presumption that these species have similar amounts of FHs, this is as expected for a tool that removes relatively many NFHs and relatively few FHs from a crude sequence set.

Another assumption we made is that NFHs provide noise which negatively affects the quality of the MSA. This we tested by determining the amount of reliable columns using BMGE, a tool designed for trimming MSAs before phylogeny. The numbers of reliable columns converge towards approximately 300–400 (Fig. 4A), except for species with few homologues that yield more reliable columns. This is explained by the fact that alignment size increases with the number of sequences and additional lower sequence variation since these species have smaller, less complex specialized metabolism. A minor detail here is that the MSAs for the three superfamilies converge to similar sizes, which is as expected since the three protein superfamilies hold protein with also rather similar sizes. We also showed that the presence of NFHs negatively affects the MSA processing (Fig. 4B, C). Although the effect may seem marginal, it is significant and corroborates the need for sequence scrutiny and cleaning. We were surprised to see the largest effect on MSA quality following SSR and CGSR, rather than GIR. This is likely explained by complexity that results from large numbers of NFHs.

Particularly clustering of superfamily datasets shows Seqrutinator removes NFHs rather than FHs. Pseudogenes have no functional constraint and an elevated evolutionary rate by which they stand out in phylogenies. Four pseudogenes can be seen for the BAHD case in Fig. 5B. SSR and CGSR remove the majority of the NFH sequences. These partial sequence do often not show higher evolutionary rates and cannot be detected by phylogeny. They do, however, interfere in HMMERCTTER clustering. As mentioned earlier, HMMERCTTER clustering uses a classifier that identifies clades that as sequence cluster show 100% P&R using hmmsearch. Partial sequences will cluster perfectly in a phylogenetic reconstruction but will obtain lower scores in a cluster specific hmmsearch since hmmer score is cumulative. As such, removing partial sequences will result in larger clusters. Interestingly, the strict Pfam set (cut-off at 1E-50) has about 20% less sequences, which cluster in 59 clusters compared to 24 clusters for the default pipeline (Fig. 5B, C). Although the objective of Pfam is to provide sequence sets for all rather than only functional homologues, this shows Seqrutinator removes less sequences, more NFH sequences and as such few FH sequences.

The recovery analysis, presented in detail in Additional file 2: Supplemental Document 2, also gives valuable insight. It detected both false and true positives. Interestingly, it shows that, as expected, the modules targeting outliers sometimes detect complete subfamilies, irrespective of whether it concerns false positives, such as the AOS in the CYP case; true positives, such as the homologous glycosyl transferase family 28 and merely similar epimerases in the UGT case; or a taxon-specific subfamily also identified in the recovery analysis of the UGT case. In the BAHD case, the outlier removers identified only a part of the CER subfamily. Irrespective of the fact that it concerns a distant subfamily of sequences that lack the DFGWG motif, it is interesting to see that part of them were not removed. This is related to the applied threshold and a typical example of how cut-off threshold affects sensitivity and precision at the same time. When we applied a slightly more strict threshold (2.35σ instead of 3σ), the clustering score increases, particularly when applied to PR (see Fig. 5B). This reflects a trade-off between retaining FHs and removing NFHs. We recommend using the non-parametric IQR threshold since superfamilies tend to have score distributions that are not normal. We merely applied the parametric 3σ threshold since this facilitates comparisons (e.g., with 2 and 2.35σ). In order to facilitate the choice of the threshold, Seqrutinator output includes a number of graphs (see Fig. SD2-2 in Additional file 2: Supplemental Document 2,) of which particularly the score plot of the PR module is informative.

Another major issue is how to apply Seqrutinator. We present Seqrutinator as a pipeline embedded in between the already discussed third block of recovery analysis and the initial preparative block. This preparative block guides somewhat how Seqrutinator is best applied.

We recommend to use the complementary script MuFasA to automatically mine homologues from multiple datasets such as complete proteomes. However, an important parameter is the amount of sequences obtained. In principle, the higher the number of offered sequences, the better the results of particularly outlier removal should be. This is, however, also affected by computational cost of the alignments. In our hands using powerful PCs with sixteen threads, 50 up to 200 sequences run well, whereas 2003 sequences (complete BAHD case, Additional file 4: Supplemental

Document 3) was no longer feasible using the default alignment method (MAFFT Global). If fusion of sequence sets is feasible, we advise to do so and do this in a taxonomical meaningful manner since that prevents the loss of taxon-specific subfamilies.

A profound understanding of your biological system can be important. Settings for GIR and CGSR depend on the species that are involved. We can envisage Seqrutinator may be applied to bacterial and archaeal superfamilies but GIR and CGSR may be less functional. Knowledge on, for instance, how proteins derive from polycistronic operons is likely helpful in setting Seqrutinator parameters.

It is important to check the input MSA since Seqrutinator depends on it. Every combination of sequences that is hard to align should be approached with care. Differences in domain architectures and sequence repeats are the major no-gos or at least demand additional preparation. We envisage two types of protein classes with different domain architectures. The most problematic class has two or more families that share at least one domain and differ in at least one domain. This results in alignments with large gap regions or, depending on the exact architecture, nonsensical MSAs. An example is found for single chain ABC transporters that have two homologous modules, each with an ABC and a transmembrane region (TMR) domain. In the A and G families, the order of the ABC and TMR domain has been switched, resulting in two different possible ways of alignment. The second problematic protein class has one family that has an additional domain. Besides that these proteins will be aligned with large gaps, the shorter class will have low hmmer scores by which pseudogenes of the larger class may no longer be detected. This also occurs for a set with single chain and dimer ABC transporters. Such problematic domain architecture can be detected studying the MSA. A possible solution is to split the sets into the corresponding families and run more than a single Seqrutinator analysis. When the parameters for the sets are identical this can be done using the complementary script SeqYNeT, which automatically controls Seqrutinator to operate over various input datasets.

Sequence repeats, which often but not always correspond with low complexity regions, are more difficult to detect but can negatively affect Seqrutinator performance. An example from ABC transporters is that where B, C, and D families form a superfamily that has the same architecture as the A family, which is however not homologous for the TMR domain [48]. Each TMR consists of six helices with mostly, but not only, hydrophobic residues. Hence, they form non-homologous but nevertheless similar repeats with low complexity. These are infamous when it concerns alignment in general, do indeed result in MSAs with poorly aligned subsequences, and are fatal for Seqrutinator performance.

Care should also be taken when it concerns multidomain sequences with the same architecture. Domains are typically connected by loop regions which have the propensity to be intrinsically disordered. This may result in inadvertent removal of sequences because GIR or CGSR detects a correct gap region instigated by the intrinsically disordered region. An example of a superfamily with highly variable, intrinsically disordered linkers is that of plant phospholipase C [49].

In summary, Seqrutinator is an efficient tool that can assists in the automated sequence mining of protein superfamilies that should give good results as long as sequences are truly homologous. More complex cases should as such be split into truly homologous subsets for optimal results.

Methods

Initial data mining

The initial sequence mining by hmmsearch from HMMER [43] using Pfam profiles PF02458 for BAHD [50], PF00067 for CYP [51], and PF00201 for UGT [52] using HMMERs inclusion threshold as cut-off. Searches were performed in batch using the MuFasA script (See Additional file 1: Supplemental Document 1) and the 16 plant species sequence sets obtained from Phytozome v12.1.6 [53], TAIR v6 obtained from TAIR [54], and the SwissProt datasets that were obtained from UniProt [55, 56]. Sequences for structures, PDB identifiers 4G0B (BAHD [57]), 5YLW (CYP [58]), and 3HBF (UGT [59]), were identified as top scoring sequence with the respective superfamily Pfam profiles and used to guide the manual trim of the MSAs that formed the input for Seqrutinator.

Sequence alignment and other biocomputational analyses

All MSAs were performed using MAFFT-G-INS-i [26], except when indicated that FAMSA [25] was used. Trimming in Seqrutinator was performed using BMGE [35] using standard gap settings, BLOSUM62, and an entropy cut-off h of 0.8. Due to the poor quality of some datasets (e.g., dataset before Seqrutinator), which resulted in low or none reliable columns with BMGE, all datasets for phylogeny were first trimmed with trimAl with -gappyout settings followed by BMGE. CD-Hit clustering [60] was performed at the CD-Hit suite [61] at an identity cut-off of 0.3. BLAST [62, 63] analysis was performed at NCBI [64] against the database as indicated. Pfam scans were performed at EBI [65] using Pfam's gathering threshold for cut-off or locally if and with Expect values as indicated. Dotplots [66] were performed at the SIB [67]. Phylogenies were reconstructed by FastTree [20], using the WAG model and optimized Gamma20 likelihood, and drawn by Dendroscope [68]. Local alignments were performed with LALIGN/PLALIGN [69] at the UVA [70]. Alphafold [71] structure models were made the Colabfold [72] form. Alluvial diagrams were generated using RAWGraphs [73]. Boxplots and histograms were prepared with Plotly (Plotly Technologies Inc. Collaborative data science. Montréal, QC, 2015 [74]). Statistical analysis and raincloud plots were performed with JASP (v0.18.3, JASP Team 2024 [75]).

Seqrutinator

A full description of Seqrutinator and complementary scripts is in Additional file 1: Supplemental Document 1.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-024-03371-y.

Additional file 1: Supplemental Document 1. Technical Description of Seqrutinator and its auxiliary scripts.

Additional file 2: Supplemental Document 2. Details on sequence recovery analysis of sequences removed by Seqrutinator.

Additional file 3: Supplemental Table S1. BAHD, CYP and UGT Seqrutinator results with default pipeline.

Additional file 4: Supplemental Document 3. The Effect of the Data on Seqrutinator's Performance.

Additional file 5. Review history.

Acknowledgements
Not applicable.

Review history
The review history is available as Additional file 5.

Peer review information
Veronique van den Berghe and Kevin Pang were the primary editors of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Authors' contributions
AA and NS wrote Seqrutinator and MuFasA. AtH, FV, and HMA performed the analyses on the CYP, BAHD, and UGT cases, respectively. The design of the method was performed by all authors under supervision of AtH. The design of the performance analyses was performed by all authors under supervision of FV. The manuscript was written by AtH and FV with corrections by AA, NS, and HMA.

Authors' information
This manuscript is dedicated to NS, our beloved friend and co-worker that passed away much too soon and could not witness the publication of the results of his hard work.

Funding
AA and NS were CONICET fellows; AtH and FV are CONICET Career investigators. Funding was from ANPCyT, project PICT2017-1310.

Availability of data and materials
All datasets have resulted from public sequence sets. Seqrutinator software (including the auxiliary scripts MuFasA and SeqYNet) is free under GPL-3.0 license, and can be downloaded from the repositories either at GitHub [76] or Zenodo [77].

Declarations

Ethics approval and consent to participate
Not applicable. This study did not involve human participants or experiments with animals.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 30 August 2023 Accepted: 13 August 2024
Published online: 26 August 2024

References

1. Villarreal F, Stocchi N, ten Have A. Functional classification and characterization of the fungal glycoside hydrolase 28 protein family. J Fungi. 2022;8:217.
2. Bondino HG, Valle EM, ten Have A. Evolution and functional diversification of the small heat shock protein/α-crystallin family in higher plants. Planta. 2012;235:1299–313.
3. Bustamante JP, Radusky L, Boechi L, Estrin DA, ten Have A, Martí MA. Evolutionary and functional relationships in the truncated hemoglobin family. Keskin O, editor. PLoS Comput Biol. 2016;12:e1004701.
4. Valiñas MA, Have A ten, Andreu AB. Identification of the functions of 4-coumarate-CoA ligase/ acyl-CoA synthetase paralogs in potato. 2021. bioRxiv. https://doi.org/10.1101/2021.07.06.451337.
5. Revuelta MV, van Kan JAL, Kay J, ten Have A. Extensive expansion of A1 family aspartic proteinases in fungi revealed by evolutionary analyses of 107 complete eukaryotic proteomes. Genome Biol Evol. 2014;6:1480–94.
6. Kumar K, Mhetre A, Ratnaparkhi GS, Kamat SS. A superfamily-wide activity atlas of serine hydrolases in Drosophila melanogaster. Biochemistry. 2021;60:1312–24.
7. Spence MA, Mortimer MD, Buckle AM, Minh BQ, Jackson CJ. A Comprehensive phylogenetic analysis of the serpin superfamily. Mol Biol Evol. 2021;38:2915–29.
8. Lin LM, Guo HY, Song X, Zhang DD, Long YH, Xing ZB. Adaptive evolution of chalcone isomerase superfamily in Fagaceae. Biochem Genet. 2021;59:491–505.
9. Orts F, Ten Have A. Structure-function analysis of Sedolisins: evolution of tripeptidyl peptidase and endopeptidase subfamilies in fungi. BMC Bioinformatics. 2018;19:464.
10. Stocchi N, Revuelta MV, Castronuovo PAL, Vera DMA, Ten Have A. Molecular dynamics and structure function analysis show that substrate binding and specificity are major forces in the functional diversification of Eqolisins. BMC Bioinformatics. 2018;19:338.

11. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, et al. Pfam: the protein families database. Nucleic Acids Res. 2014;42:D222–30.
12. Haft DH, Selengut JD, White O. The TIGRFAMs database of protein families. Nucleic Acids Res. 2003;31:371–3.
13. Gough J, Karplus K, Hughey R, Chothia C. Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. J Mol Biol. 2001;313:903–19.
14. Simonetti FL, Teppa E, Chernomoretz A, Nielsen M, Marino BC. MISTIC: mutual information server to infer coevolution. Nucleic Acids Res. 2013;41:W8-14.
15. Mazin PV, Gelfand MS, Mironov AA, Rakhmaninova AB, Rubinov AR, Russell RB, et al. An automated stochastic approach to the identification of the protein specificity determinants and functional subfamilies. Algorithms for Molecular Biology. 2010;5:29.
16. Wilkins A, Erdin S, Lua R, Lichtarge O. Evolutionary trace for prediction and redesign of protein functional sites. Methods Mol Biol. 2012;819:29–42.
17. Chagoyen M, García-Martín JA, Pazos F. Practical analysis of specificity-determining residues in protein families. Brief Bioinform. 2016;17:255–61.
18. Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol. 2010;59:307–21.
19. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics (Oxford, England). 2014:1312–3. Oxford University Press. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24451623.
20. Price MN, Dehal PS, Arkin AP. FastTree 2 - approximately maximum-likelihood trees for large alignments. Poon AFY, editor. PLoS One. 2010;5:e9490.
21. Ronquist F, Huelsenbeck JP. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics. 2003;19:1572–4.
22. Notredame C, Higgins DG, Heringa J. T-coffee: a novel method for fast and accurate multiple sequence alignment 1 1Edited by J. Thornton. J Mol Biol. 2000;302:205–17.
23. Löytynoja A, Vilella AJ, Goldman N. Accurate extension of multiple sequence alignments using a phylogeny-aware graph algorithm. Bioinformatics. 2012;28:1684–91.
24. Szalkowski AM. Fast and robust multiple sequence alignment with phylogeny-aware gap placement. BMC Bioinformatics. 2012;13:1–11.
25. Deorowicz S, Debudaj-Grabysz A, Gudys A. FAMSA: fast and accurate multiple sequence alignment of huge protein families. Sci Rep. 2016;6:1–13.
26. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30:772–80.
27. Shen C, Zaharias P, Warnow T. MAGUS+eHMMs: improved multiple sequence alignment accuracy for fragmentary sequences. Bioinformatics. 2022;38:918–24.
28. Santus L, Garriga E, Deorowicz S, Gudyś A, Notredame C. Towards the accurate alignment of over a million protein sequences: current state of the art. Curr Opin Struct Biol. 2023;80:102577.
29. Baltzis A, Mansouri L, Jin S, Langer BE, Erb I, Notredame C. Highly significant improvement of protein sequence alignments with AlphaFold2. Bioinformatics. 2022;38:5007–11.
30. Tumescheit C, Firth AE, Brown K. CIAlign: a highly customisable command line tool to clean, interpret and visualise multiple sequence alignments. PeerJ. 2022;10:e12983.
31. Chiner-Oms A, González-Candelas F. EvalMSA: a program to evaluate multiple sequence alignments and detect outliers. Evol Bioinforma. 2016;12:277–84.
32. Mendoza MLZ, Nygaard S, Da Fonseca RR. DivA: detection of non-homologous and very divergent regions in protein sequence alignments. BMC Res Notes. 2014;7. Available from: https://pubmed.ncbi.nlm.nih.gov/25403086/. Cited 2022 Jun 24.
33. Jehl P, Sievers F, Higgins DG. OD-seq: outlier detection in multiple sequence alignments. BMC Bioinformatics. 2015;16. Available from: https://pubmed.ncbi.nlm.nih.gov/26303676/. Cited 2022 Jun 26.
34. Maldonado E, Antunes A. LMAP_S: Lightweight Multigene Alignment and Phylogeny eStimation. BMC Bioinformatics. 2019;20. Available from: https://pubmed.ncbi.nlm.nih.gov/31888452/. Cited 2022 Jun 26.
35. Criscuolo A, Gribaldo S. BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. BMC Evol Biol. 2010;10:210.
36. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics (Oxford, England). 2009;25:1972–3.
37. Talavera G, Castresana J. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst Biol. 2007;56:564–77.
38. Rieseberg TP, Dadras A, Fürst-Jansen JMR, Dhabalia Ashok A, Darienko T, de Vries S, et al. Crossroads in the evolution of plant specialized metabolism. Semin Cell Dev Biol. 2022. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1084952122000738. Cited 2022 Mar 14.
39. Cytochrome P450 Nomenclature Files. Cytochrome P450 Homepage. 2020. Available from: https://drnelson.uthsc.edu/nomenclature/. Cited 2024 Apr 10.
40. UGT Gene Names | Washington State University. Available from: https://labs.wsu.edu/ugt/. Cited 2024 Apr 10.
41. Mi H, Muruganujan A, Thomas PD. PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. Nucleic Acids Res. 2013;41:D377–86.
42. Pagnuco IA, Revuelta MV, Bondino HG, Brun M, Ten Have A. HMMER cut-off threshold tool (HMMERCTTER): supervised classification of superfamily protein sequences with a reliable cut-off threshold. PLoS One. 2018;13(3):e0193757.
43. Eddy SR. Accelerated profile HMM searches. PLoS Comput Biol. 2011;7:e1002195.
44. Hong X, Scofield DG, Lynch M. Intron size, abundance, and distribution within untranslated regions of genes. Mol Biol Evol. 2006;23:2392–404.
45. Carrillo H, Lipman D. The multiple sequence alignment problem in biology. SIAM J Appl Math. 1988;48:1073–82.

46. Chang J-M, Di Tommaso P, Lefort V, Gascuel O, Notredame C. TCS: a web server for multiple sequence alignment evaluation and phylogenetic reconstruction: Figure 1. Nucleic Acids Res. 2015;43:W3-6.
47. Zimin AV, Stevens KA, Crepeau MW, Puiu D, Wegrzyn JL, Yorke JA, et al. An improved assembly of the loblolly pine mega-genome using long-read single-molecule sequencing. Gigascience. 2017;6:1–4.
48. Thomas C, Aller SG, Beis K, Carpenter EP, Chang G, Chen L, et al. Structural and functional diversity calls for a new classification of ABC transporters. FEBS Lett. 2020;594:3767–75.
49. Robuschi L, Mariani O, Perk EA, Cerrudo I, Villarreal F, Laxalt AM. Arabidopsis thaliana phosphoinositide-specific phospholipase C 2 is required for Botrytis cinerea proliferation. Plant Sci. 2024;340:111971.
50. Pfam: Family: Transferase (PF02458). Available from: https://pfam.xfam.org/family/PF02458. Cited 2022 Mar 22.
51. Pfam: Family: p450 (PF00067). Available from: https://pfam.xfam.org/family/PF00067. Cited 2022 Mar 22.
52. Pfam: Family: UDPGT (PF00201). Available from: https://pfam.xfam.org/family/PF00201.21. Cited 2022 Mar 22.
53. Phytozome. Available from: https://phytozome-next.jgi.doe.gov/. Cited 2022 Mar 22.
54. TAIR - Home Page. Available from: https://www.arabidopsis.org/. Cited 2022 Mar 22.
55. Consortium U. UniProt: the universal protein knowledgebase. Nucleic Acids Res. 2017;45:D158–69.
56. UniProtKB. UniProtKB/Swissprot. Available from: https://www.uniprot.org/uniprot/?query=reviewed:yes. Cited 2022 Mar 22.
57. Lallemand LA, Zubieta C, Lee SG, Wang Y, Acajjaoui S, Timmins J, et al. A structural basis for the biosynthesis of the major chlorogenic acids found in coffee. Plant Physiol. 2012;160:249–60.
58. RCSB PDB - 5YLW: CYP76AH1 from Salvia miltiorrhiza. Available from: https://www.rcsb.org/structure/5ylw. Cited 2022 Jun 29.
59. Modolo LV, Li L, Pan H, Blount JW, Dixon RA, Wang X. Crystal structures of glycosyltransferase UGT78G1 reveal the molecular basis for glycosylation and deglycosylation of (iso)flavonoids. J Mol Biol. 2009;392:1292–302.
60. Huang Y, Niu B, Gao Y, Fu L, Li W. CD-HIT Suite: a web server for clustering and comparing biological sequences. Bioinformatics. 2010;26:680–2.
61. CD-HIT Suite. Available from: http://weizhong-lab.ucsd.edu/cdhit-web-server/cgi-bin/index.cgi?cmd=cd-hit. Cited 2022 Jun 29.
62. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215:403–10.
63. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25:3389–402.
64. Protein BLAST: search protein databases using a protein query. Available from: https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome. Cited 2022 Jun 29.
65. Pfam: Home page. Available from: https://pfam.xfam.org/. Cited 2022 Jun 29.
66. Junier T, Pagni M. Dotlet: diagonal plots in a web browser. Bioinformatics (Oxford, England). 2000;16:178–9.
67. Dotlet JS. Available from: https://dotlet.vital-it.ch/. Cited 2022 Jun 29.
68. Huson DH, Scornavacca C. Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks. Syst Biol. 2012;61:1061–7.
69. Pearson WR, Lipman DJ. Improved tools for biological sequence comparison. Proc Natl Acad Sci USA. 1988;85:2444–8.
70. LALIGN/PLALIGN local alignments. Available from: https://fasta.bioch.virginia.edu/fasta_www2/fasta_www.cgi?rm=lalign&pgm=pal. Cited 2022 Jun 29.
71. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596(7873):583–9.
72. Colabfold form. Available from: https://colab.research.google.com/github/sokrypton/ColabFold/blob/main/AlphaFold2.ipynb. Cited 2022 Dec 14.
73. Mauri M, Elli T, Caviglia G, Uboldi G, Azzi M. RAWGraphs: A visualisation platform to create open outputs. In: Proceedings of the 12th Biannual conference on Italian SIGCHI Chapter. New York: Association for Computing Machinery; 2017. p. 1–5.
74. Plotly: the front end for ML and data science models. Available from: https://plotly.com/. Cited 2022 Jun 29.
75. Download JASP. JASP - free and user-friendly statistical software. Available from: https://jasp-stats.org/download/. Cited 2024 Apr 17.
76. Amalfitano A, Stocchi N, Atencio HM, Villarreal F, ten Have A. Seqrutinator. Github; 2024. Available from: https://github.com/BBCMdP/Seqrutinator.
77. Amalfitano A, Stocchi N, Atencio HM, Villarreal F, ten Have A. Seqrutinator. Zenodo; 2024. Available from: https://zenodo.org/records/10980626.

Publisher's Note