

Supplemental Document 2: Sequence Recovery Analysis Seqrutinator.

Introduction

Seqrutinator identifies sequences based on its definitions (main document) and applied parameters and removes these as non-functional homologues. It shows consistent behaviour and various indirect analyses all indicate it has a good performance. However, since there is no reliable gold standard for benchmarking we have no quantifiable data on performance. Moreover, high performance does not exclude some of the removed sequences may have been incorrectly removed. We performed recovery analysis using CD-Hit to identify eventual false positives and hmmscan to confirm true positives. The recovery analyses were performed with parametric sigma cut-offs rather than the default IQR based cut-off to get information on the sensitivity of the outlier removal modules

Results and Discussion

The outlier modules NHHR and PR remove a few sequences from the twilight zone

No sequences were removed by NHHR or PR from the SwissProt BAHDDome using 3σ as cut-off. The more strict 2σ identified Q39048, Q9LIS1 and Q9SVM9, entries belonging to the ECERIFERUM or CER subfamily, a distant BAHDD subfamily implicated in the mechanism of cell wall waxes elongation [1] characterized by the absence of the DFGWG motif and unknown molecular function. Hence, Seqrutinator is highly specific on gold standard SwissProt using the 3σ setting of the outlier modules. On the other hand, the default pipeline removed five and 30 sequences from the 16 species BAHDDomes, for NHHR and PR respectively. CD-Hit clustering identified 17 sequences that are part of that same CER subfamily, as shown by phylogeny (See Figure SD2-1). Another 40 CER subfamily sequences were not identified as NFH. Hence, the 2σ setting of the outlier modules appeared as too strict.

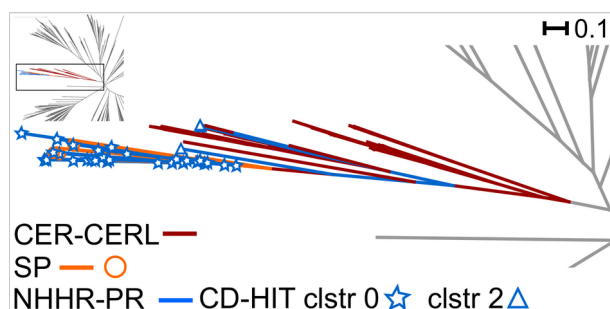


Figure SD2-1: NHHR-PR false positives cluster with the distant CER subfamily. Phylogeny of BAHDD SwissProt-accepted sequences combined with CER sequences removed by NHHR or PR. Reintroduced sequences fall in a 100% P&R cluster, which includes a subclade with all SwissProt CER entries. No other annotated sequences were found in the clade.

Similar results were obtained with the CYP superfamily. No sequences were removed by NHHR or PR from the SwissProt CYPome, whereas 17 sequences, part of the allene oxide synthase (AOS) subfamily were removed using the strict 2σ cut-off. One of these is a partial sequence that is normally removed by CGSR. The default pipeline removed 15 and 61 sequences from the 16 species' CYPomes, for NHHR and PR respectively. These were first analysed using hmmscan that identified 59 of the outliers as CYP sequences and 17 as not CYP (either no hit or other hits). Since Pfam does not have descriptions for CYP subfamilies, we again applied CD-Hit, which identified 50 similar sequences, of which 48 were shown by BLAST analysis to encode AOS, one is a partial AOS and one is an AOS pseudogene. From the remaining nine sequences, we identified one as an incorrectly removed, functional C-22 sterol desaturase whereas eight sequences were qualified as correctly removed. Sequence B3LF83.1 is a partial sequence removed by NHHR from the species proteome but by CGSR from the SwissProt dataset. Five sequences showed low identity and/or low cover to other SwissProt sequences, suggesting they are pseudogenes. Another sequence qualified as pseudogene since it contains a partial repeat. Two other

sequences, likely resulting from incorrect gene modelling, yielded truncated alignments and the last NFH shows the most similarity to a Monocopper oxidase-like protein SKS2 which is not a CYP, even though Pfam accepts it as CYP with an E value of 6E-22. Together we identified 49 false positives and 27 true positives using the default settings on the 16 species CYPomes.

Seqrutinator's output includes a number of plots that can be useful to support a recovery analysis. Figure SD2-2 shows the example of the CYP case and the SwissProt dataset. On top (Figure SD2-2A to C) output from the pipeline with NHHR at 3σ whereas the lower panels (Figure SD2-2-D to F) correspond with the more strict 2σ setting. A and D show that no and 17 sequences are removed. The final GIR plots B and E show that at least one of the sequences has a large combined gap, which is not actively used but does indicate that one of the sequences removed by NHHR at high stringency has poor alignment explaining its lower score. The final plots (D and F) show the scores of the finally accepted sequences. The plot from the 3σ pipeline shows two outgroups. The largest consists of exactly the 17 sequences that NHHR removes at 2σ . The second smaller group with a higher score corresponds to five C-22 sterol desaturases, among which the sequence removed by the 3σ cut-off in one of the 16 species scrutinies. Hence, the 3σ cut-off results in both false and true positives.

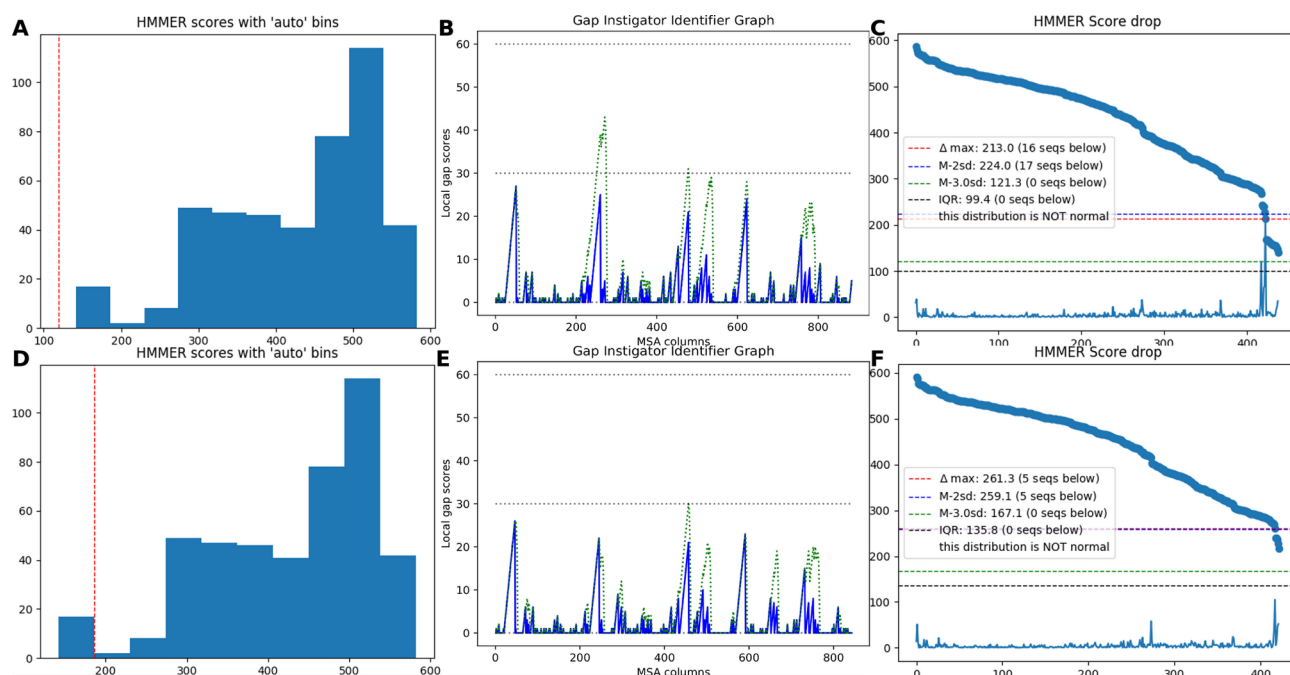


Figure SD2-2: Recovery SwissProt CYP Case. **A and D:** Score distribution histograms of NHHR modules on CYP, curated SwissProt Sequence set, with cut-off (red dashed line) at 3σ and 2σ , respectively. **B and E:** Gap score plots of GIR modules applied to sequence sets following NHHR with 3σ and 2σ , respectively. The blue line shows the applied continuous gap score and the dashed green line the more strict combined gap score. **C and F:** Score plots of PR modules applied to sequence sets following CGSR, after NHHR with 3σ and 2σ , respectively. Indicated are putative cut-off thresholds. The lineplot shows the score drop between respective hit scores.

The NHHR module did remove two SwissProt sequences that encode proteins from Glycosyl Transferase Family 28 (GT28) in the UGT case. Also, all 61 GT28 sequences were identified from the separate proteome sets, albeit few by NHHR (6) or PR (10) and more by GIR (12) or CGSR (33). The sequences were aligned to the final MSA of accepted sequences, introducing large gaps (Figure SD2-3A). The phylogeny we made (Figures SD2-3B and 3C) shows the removed GT28 sequences cluster in a distant clade. Hence, although GT28 and UGT sequences are similar they also differ substantially.

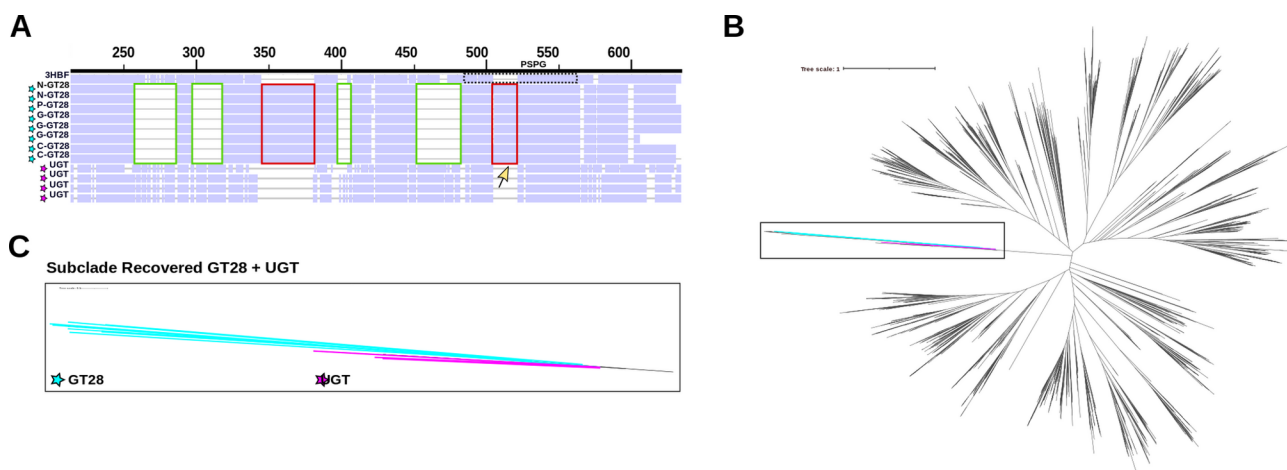


Figure SD2-3. Removed Glycosyl Transferase 28 Sequences are True Positives. **A:** Excerpt of MSA with GT28 sequences removed by NHHR (N), PR (P), GIR (G) or CGSR (C) module aligned to final UGT MSA. The GT28 sequences have inserts (red boxes) among which a 17 residue insert at position 505 (indicated with yellow arrow) that disturbs the PSPG motif (black dotted box). GT28 sequences also have large gap regions (green boxes). Four sequences demonstrated below the GT28 sequences fall in the same clade (see panels B and C). 3HBF (R) is the reference sequence for UGT. **B:** Phylogeny of accepted UGT and reintroduced GT28 sequences. **C:** Detail of GT28-UGT clade with GT28 sequences in light blue and four UGTs in magenta.

Interestingly, the clade appears as a subclade from a clade with four UGT sequences (one from *Amborella trichopoda* and three from *Physcomitrella patens*). None of these four sequences has a match against either UGT or GT28 in a Pfam scan, two hit GT28 glycosyltransferases in a BLAST against SwissProt. Figure SD2-3A shows how GT28 sequences but not the other four sequences from the mentioned clade disturb, among others, the PSPG motif. Indeed and although arbitrary, the Glycosyl Transferase Family 28 is homologous to but does not form part of the UGT superfamily [2].

Interestingly, some sequences were removed by CGSR or GIR. We analysed the corresponding MSAs and concluded that the large distance these homologues have with UGTs can result in gaps, either in the GT28 sequences or in the UGT sequences, gaps that are sometimes detected by the GIR or CGSR modules. If not detected the sequences are removed by the final PR outlier module. All together, this shows Seqrutinator correctly identifies and removes sequences from related protein families. As such its reach ends in the twilight zone of similarity detection.

We identified 36 additional outliers in the UGT case. CD-Hit analysis identified seven epimerases (PF01370), which are not related to UGT. We also found two sequences from potato and one from tomato that are highly similar, suggesting a Solanaceae-specific subfamily. BLAST with these sequences against Reference Proteins identified highly similar sequences from other Solanum or Solanaceae. A sudden drop of identity from ~80 to ~58% occurred from a hit against *Nicotiana tabacum* (tobacco, another Solanaceae) to the next hit from *Ipomea nil* (morning glory, a Convolvulaceae from the order of Solanales). The other outliers did not show clustering and are indeed NFHs

Together these analyses show that Seqrutinator modules NHHR and PR remove real outliers but also that outliers can encode FHs belonging to distant subfamilies. The applied cut-off of 3σ , which is often just above default threshold IQR, resulted in a good balance in terms of precision and recall, and shows why a recovery analysis is recommended.

The GIR module Identifies False Positives in the BAHD but not the CYP or UGT cases.

A total of 54 candidate BAHD sequences were removed by GIR in the default pipeline. All sequences presented significant hits to the BAHD Transferase domain, as determined by hmmscan. Three sequences also presented partial but significant hits to other families (TIR PF01582.23, Flavodoxin PF00258.28 and Peptidase_M14 PF00246.27), which explains why these instigate a large gap in an otherwise BAHD alignment. Then, many of the other removed sequences presented more than one significant hit with the BAHD transferase domain. These sequences were analysed using LALIGN local alignment in order to identify putative internal duplications.

Three scenarios, two of which are depicted in Figure SD2-4, describe 48 sequences removed by GIR. Scenario 1 accounts for nine sequences that show broken alignments with the BAHD Transferase model. The sequences show at least one large single repeat, hitting part of the model twice leading to total scores above the outlier threshold of the earlier NHHR module. LALIGN finds internal duplications with an E-value < 0.0001, suggesting recent duplication events or errors made during sequencing, assembling or gene modelling. In scenario 2 that accounts for 25 sequences, part of the BAHD Transferase profile cannot detect its counterpart in the query sequence. This is combined with a repeat with a rather high E value. This repeat is in turn repeated, as is shown by the parallel secondary alignments. As such, the sequence is relatively long, which explains also these sequences instigate a gap. In scenario 3, with the remaining 12 sequences, the model is typically well covered and no internal duplications are found (not shown).

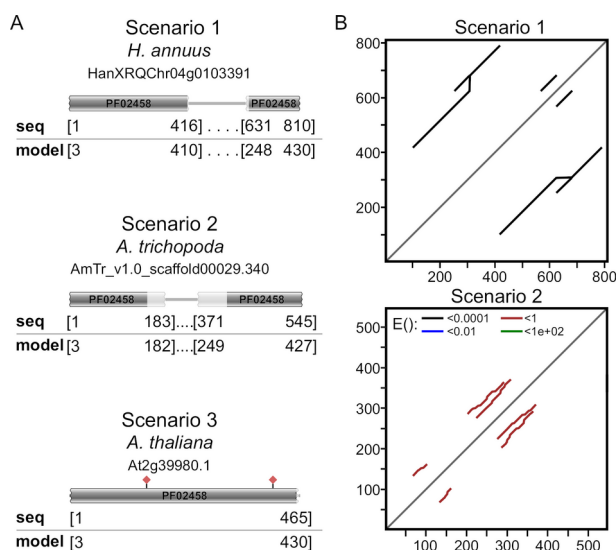


Figure SD2-4: Different Types of Repetitions Instigate Gaps. **A:** Sequences removed by GIR were analysed with hmmscan and mostly correspond to one of three scenarios. Shown are paradigmatic sequences (IDs indicated) for scenarios 1 and 2. “seq” and “model” indicate the start and end position of the hits of the query and the BAHD Transferase model as determined by Pfam. Red squares indicate conserved motifs. **B** Dotplot representation of LALIGN self-alignment. Secondary alignments indicate sequence repeats (E value according to colours).

The sequences were subjected to CD-Hit clustering. The largest cluster has 14 sequences depicting broad taxonomic distribution. Since the dataset affects the performance, we decided to align these to the accepted SwissProt sequence set using MAFFT --add and applied the GIR module that now only removed three of the sequences, examples of either scenario 1 or 2. The 11 retained sequences correspond with scenario 3 and have an insert at a conserved position (~420-480, Figure SD2-5A). Phylogeny shows ten of the sequences fall in the same subgroup with some accepted sequence while the 11th sequence (from *P. taeda*) is a sibling orphan. All are part of a larger cluster that is 100% P&R-SD (Figure SD2-5C). Hence, these 11 sequences form a second example of sequences incorrectly removed that correspond to a distant subfamily.

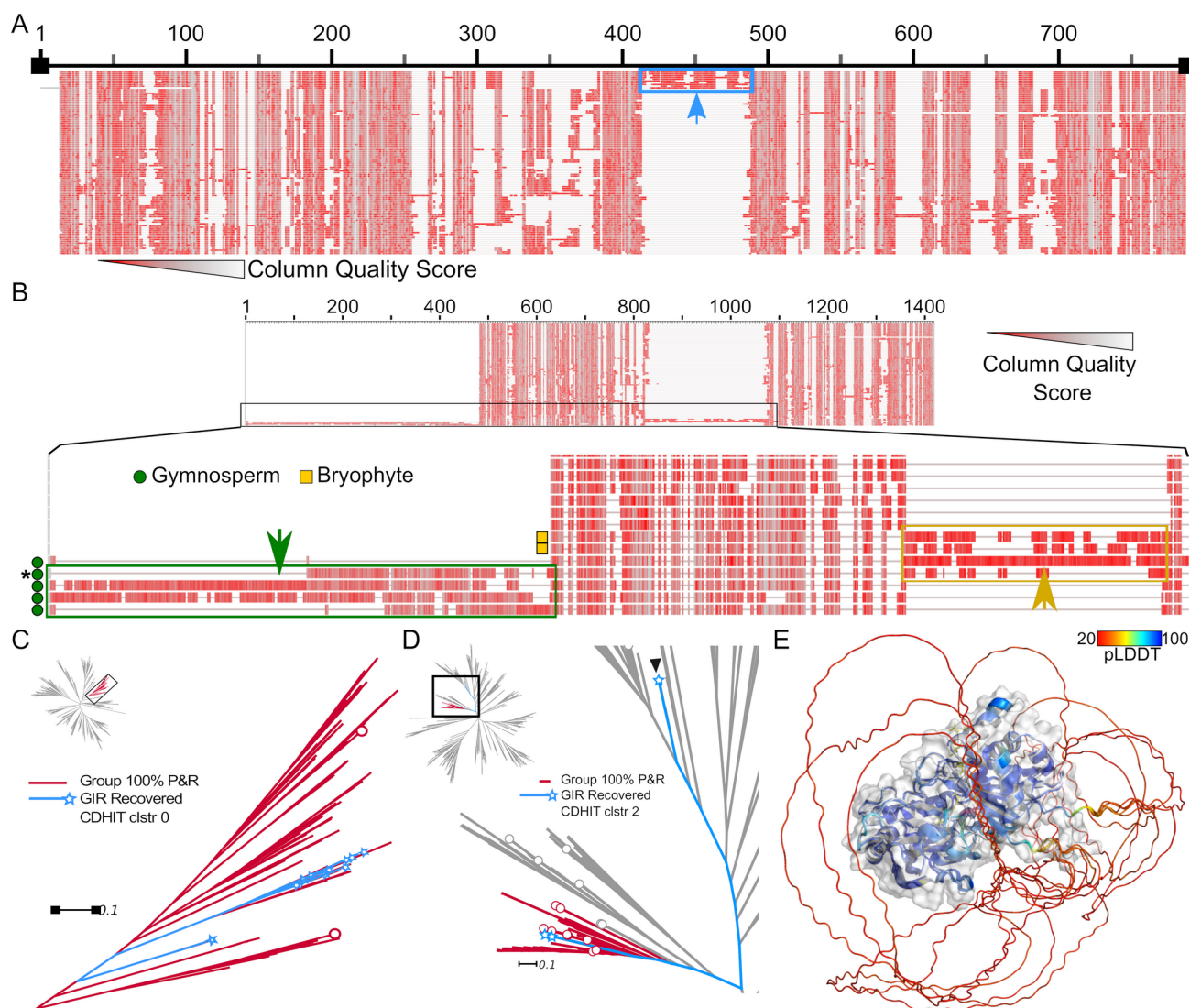


Figure SD2-5: Recovery Analysis of BAHD Sequences Removed by GIR. **A:** MSA excerpt showing 11 sequences removed by GIR (top) that instigate a large gap in the MSA of accepted SwissProt entries (~420-480, delimited by the box and indicated by an arrow). Column quality score is depicted according to the scale shown. **B:** MSA excerpt showing seven sequences removed by GIR aligned to the SwissProt entries MSA. Sequences from *P. taeda* and bryophytes are indicated with green circles and yellow boxes, respectively. Four sequences, two from bryophytes, two from *P. taeda* instigate a large gap (yellow box) whereas four of the five *P. taeda* sequences have a large N-terminal extension (green box). The star indicates a sequence with both an N-terminal extension and insert. **C:** Phylogeny of accepted SwissProt BAHD sequences and the 11 false GIR positives (shown in blue and with a star). **D:** Phylogeny of accepted SwissProt BAHD sequences and three false GIR positives (shown in blue and with a star). The two sequences from bryophytes *P. patens* and *M. polymorpha* fall in the 100% P&R clade with HCTs whereas the single *P. taeda* with only the insert (also indicated with an arrowhead) falls into a different clade. **E:** Five *PpaHCT* relaxed models generated with AlphaFold were aligned to *Coffea canephora* HCT (PDB identifier 4G0B) with average RMSD 0.6384 ± 0.04 . CcHCT is colored in white, with its surface (with transparency). Residues in *PpaHCT* models are coloured according to the per-residue confidence score (pLDDT) as shown in the scale. The ~150 residues insert of interest is surface oriented (opposed to the location of the substrates) and displays very low confidence scores, suggesting high disorder.

A cluster with seven sequences removed by GIR had sequences from *P. patens*, *M. polymorpha* and *P. taeda* and was rather complex. Both bryophyte and two of the *P. taeda* homologues have a large insert in the central region of the protein when aligned to the

SwissProt dataset using MAFFT --add (Figure SD2-5B). However, given that some of the inserts are short, it is not clear if the inserts are homologous. Four of the five *P. taeda* sequences have a large N-terminal extension which strengthens the idea that the inserts of the two *P. taeda* homologues are not monophyletic with the insert of the gymnosperm sequences. We applied phylogeny in order to show how these sequences evolved. The sequences from *P. patens* and *M. polymorpha* fall in the cluster with enzymes with hydroxycinnamoyl transferase (HCT) activity where the entries from *P. taeda*, whether with or without the insert, belong to a another subfamily (Figure SD2-5D). The two independent inserts suggests that here it may concern a hot spot for acceptable inserts. Both bryophyte homologues have been demonstrated to have HCT activity [3]. Interestingly, *in vitro* *P. patens* HCT activity and even substrate preference are maintained in a mutant enzyme that has the region corresponding to the insert deleted. An AlphaFold [4] obtained structure model showed the insert as a highly disordered, surface-exposed region (Figure SD2-5E). This further corroborates the hypothesis of the hot spot for acceptable inserts providing an example where the assumption that large inserts are functionality fatal appears incorrect.

CD-Hit clustering of the 112 putative CYP sequences that were removed by GIR yielded few clusters, the largest containing a mere 10 sequences. Nine of these sequences instigated different gaps of 30 or more residues in the MSA made of that cluster (Figure SD2-6A). Thus, although the sequences cluster, they differ in the gap they instigate. The tenth sequence instigated a gap of only 18 residues. However, when aligned to the alignment of accepted SwissProt sequences, a gap of exactly 30 residues was instigated (Figure SD2-6B). Hence, this sequence was correctly removed. Other CD-Hit clusters contained only a single pair of sequences with a homologous insert that appeared to be an internal repeat. Hence, in the CYP case, GIR did not yield a single false positive.



Figure SD2-6: Gap instigation by GIR removed sequences. **A:** An MSA of the largest cluster of sequences obtained from the GIR module of the CYP case. Nine of the ten sequences have inserts that result in a large gap in most of the other sequences. **B:** Excerpt of MSA of accepted SwissProt sequences with the last putative GIR-FH being the second sequence showing a gap of exactly 30 residues. The reference sequence 5YLW is shown on top of both A and B.

In the case of UGTs, the three largest clusters contained seven, six and five sequences respectively. Alignment consistently showed that the gaps that were instigated are not conserved. The largest cluster showed some taxonomic distribution but the other two smaller CD-Hit clusters contain only sequences from *P. taeda*. Hence, we conclude that almost all sequences removed by GIR are most likely real non-functional homologues.

No clear false positives could be recovered from the CGSR datasets

Most of the NFH homologues are partial sequences, of which most were removed by SSR and the rather safe setting of 0.65 times the length of the user-selected reference sequence. In most cases, also larger sequences will likely not be functional. Most often a more stringent setting can be applied using criteria that are to be set by the user. In this study, we chose a safe setting such that a recovery screen was not required for SSR-

removed sequences. Hence in the CGSR recovery screen, we expect to identify mostly sequences that are incomplete at the N- and/or the C-terminal.

CD-Hit Clustering of the 740 putative CYP sequences that were removed by CGSR yielded seven clusters of more than 20 sequences. All these were aligned in the presence of the general reference sequence for CYP we used (PDB identifier 5YLW). Alignment by MAFFT tends to result in ordered MSAs in which often subfamilies can be visually identified. However, not a single subfamily could be detected. We then sorted the sequences such that:

- 1) Sequences that appear to lack N-terminal subsequence are at the top, ordered by the length of the missing subsequence. Note that these may also lack C-terminal subsequence.
- 2) Sequences that do not lack N-terminal but do lack C-terminal subsequence are at the bottom, ordered by the length of the missing subsequence.
- 3) The remainder of the sequences were ordered by the position of the first large (>30 aa) gap and occur as a block in the middle.

Not a single subfamily with a conserved gap was identified. Figure SD2-7 shows as an example the aligned sequences of the largest CD-Hit cluster. As expected many sequences lack part of either or both the N- or the C-terminus. Approximately 50 sequences have both the N- and the C term intact but have an internal gap. None of the internal gaps appear however as conserved. Interestingly some sequences also show intermediately long inserts. Similar results were obtained for the other large clusters.

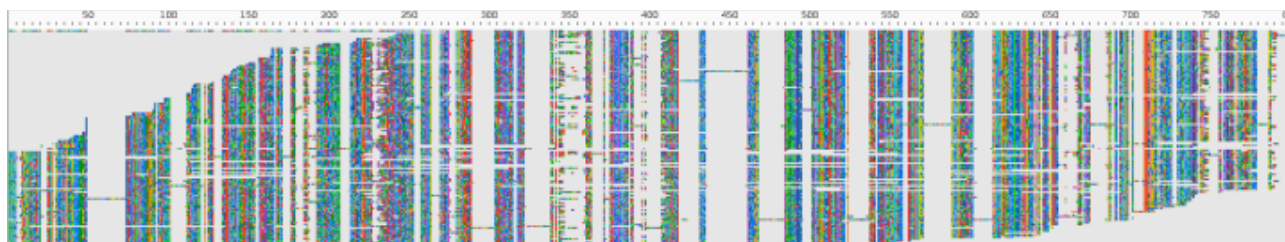


Figure SD2-7: MSA of the largest cluster of CGSR removed sequences. A total of 133 sequences that were identified by CGSR and clustered by CD-Hit were aligned with general reference 5YLW (top). Next, they were ordered based on lacking N-or C-terminal sequences. Approximately 50 sequences with both the N- and C-termini intact were ordered by the position of the first large gap.

A similar analysis was performed with the UGT datasets. We analysed the eight largest CD-Hit clusters. Most sequences lack an N-terminal sequence, many lack a C-terminal sequence and the remainder of the sequences all have large gaps whose position was not conserved.

Cited References, Information Resources and webtools

1. Haslam TM, Gerelle WK, Graham SW, Kunst L. The Unique Role of the ECERIFERUM2-LIKE Clade of the BAHD Acyltransferase Superfamily in Cuticular Wax Metabolism. *Plants* 2017, Vol 6, Page 23. 2017;6:23.
2. Breton C, Šnajdrová L, Jeanneau C, Koča J, Imberty A. Structures and mechanisms of glycosyltransferases. *Glycobiology*. 2006;16:29R-37R.
3. Kriegshauser L, Knosp S, Grienemberger E, Tatsumi K, Gütle DD, Sørensen I, et al. Function of the hydroxycinnamoyl-coa:shikimate hydroxycinnamoyl transferase is evolutionarily conserved in embryophytes. *The Plant Cell*. 2021;33:1472–91.
4. 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. 2021;596:583–9.