KO-Annotation transfer based on feature-aware orthology inference

# Project aim

Proteins that are orthologous to each other are likely to have similar functions. The quality of orthology-based annotation transfer methods depends strongly on the accuracy of the ortholog prediction. Here we are introducing HamFAS, a robust annotation transfer pipeline based on feature-aware orthology inference. HamFAS has been shown to have higher sensitivity and comparable specificity in comparison to two state-of-the-art annotation tools KAAS and BlastKOALA from KEGG. A feature that makes HamFAS different than BlastKOALA and KAAS is the controllable ability of the annotation process. Users can choose different methods and threshold for increase or reduce the stringency of the annotation pipeline. Besides, HamFAS can be run locally through command lines. It provides a better solution for large-scale analysis than using online tools like KAAS and BlastKOALA.

# Data and method

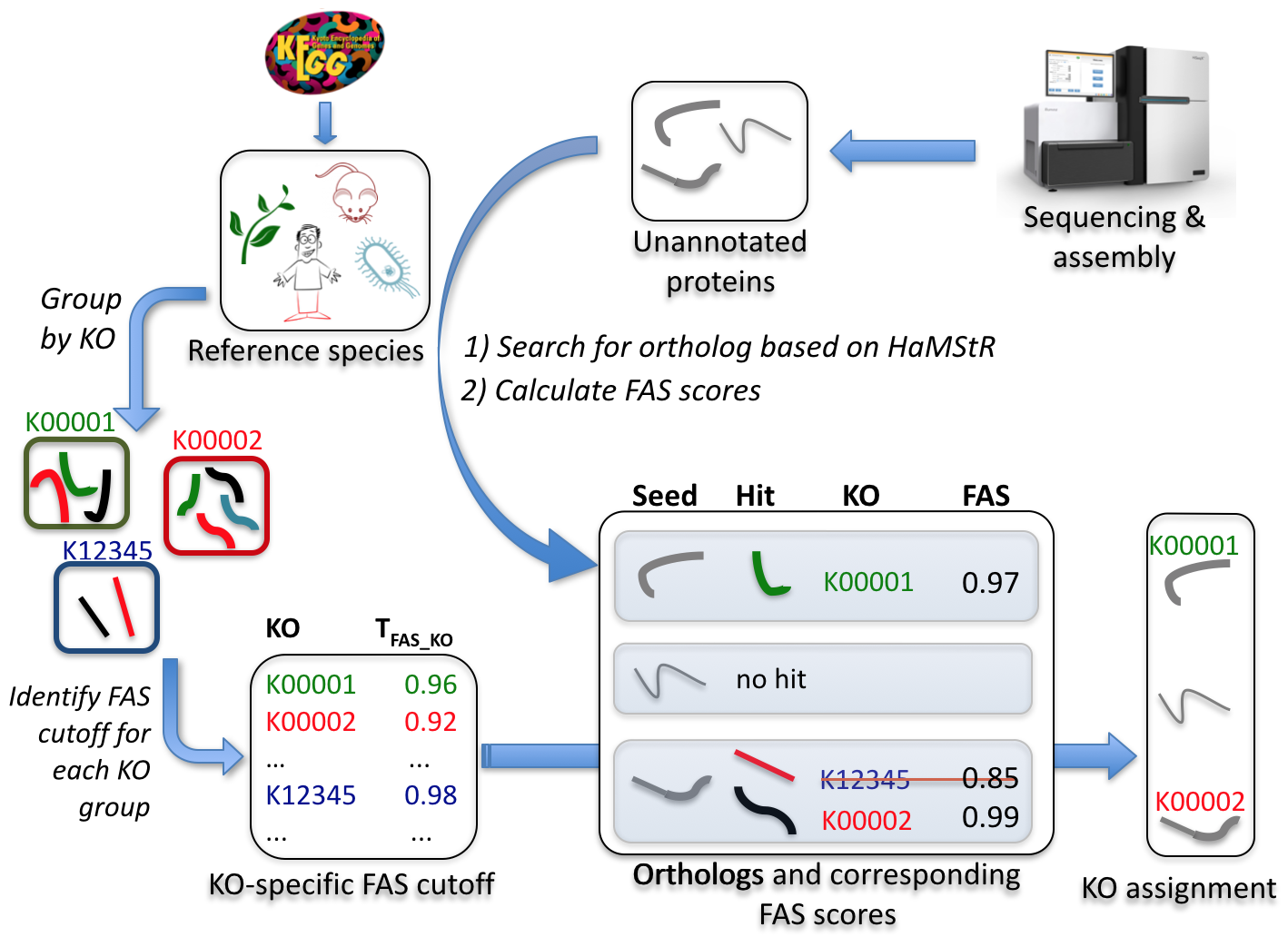


Figure 2.1. KO annotation transfer using HamFAS approach.

Figure 2.1 describes the annotation transfer pipeline using HamFAS approach. Protein sets of 30 manually KO-annotated reference species (Table 2.1) have been downloaded from KEGG database. Pairwise FAS scores of all reference proteins within a KO group have been calculated. A group's mean FAS score serves then as a cutoff (TFAS\_KO) that must be exceeded to warrant transfer of its KO identifier to the seed proteins.

Table 2.1. List of 30 manually KO-annotated reference taxa

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **No.** | **Name** | **No.** | **Name** | **No.** | **Name** |
| 1 | *A.gossypii* | 11 | *D.melanogaster* | 21 | *A.pernix* |
| 2 | *S.pombe* | 12 | *C.elegans* | 22 | *E.coli* |
| 3 | *C.albicans* | 13 | *M.brevicollis* | 23 | *N.meningtidis* |
| 4 | *S.cerevisiae* | 14 | *N.vectensis* | 24 | *H.pylori* |
| 5 | *N.crassa* | 15 | *E.histolytica* | 25 | *B.subtilis* |
| 6 | *A.nidulans* | 16 | *T.brucei* | 26 | *L.lactis* |
| 7 | *H.sapiens* | 17 | *A.thaliana* | 27 | *M.genitalium* |
| 8 | *M.musculus* | 18 | *P.falciparum 3D7* | 28 | *M.tuberculosis* |
| 9 | *R.norvegicus* | 19 | *C.hominis* | 29 | *Synechocystis sp.* |
| 10 | *D.rerio* | 20 | *M.jannaschii* | 30 | *A.aeolicus* |

Given a list of uncharacterized proteins (seed), we search for their orthologs in the reference species using HaMStR (with *-checkCorothologsRef*, *-rbh* options and *-hit\_limit=5*). FAS scores between seed proteins and their orthologs will be identified. If the calculated FAS scores is not smaller then the corresponding TFAS\_KO, the available KEGG identifiers of those paired orthologs will be transferred to the seed proteins.

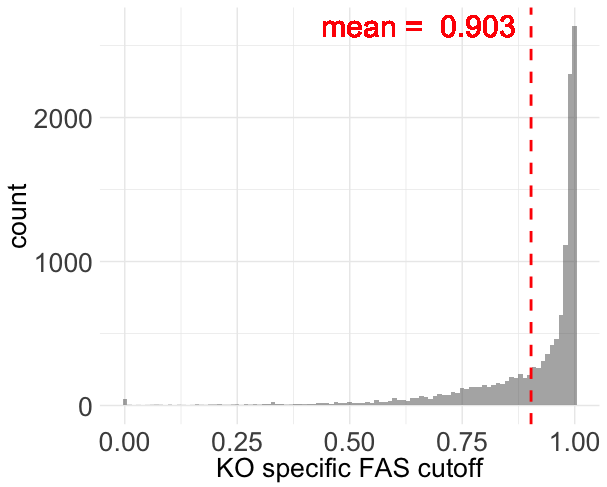


Figure 2.2. Distribution of TFAS\_KO for 12,748 KO groups

Figure 2.2 represents the distribution of all 12,748 TFAS\_KO values. Only about 3% of KOs have TFAS\_KO smaller than 0.5, 27% lie between 0.5 and 0.9, while 70% has TFAS\_KO greater than 0.9. The low TFAS\_KO values are caused mostly by the uninformative protein members. Figure 2.3 shows 2 examples, the FAS scores distribution of K00542, which represents low TFAS\_KO group, and K07888, which represents high TFAS\_KO group.

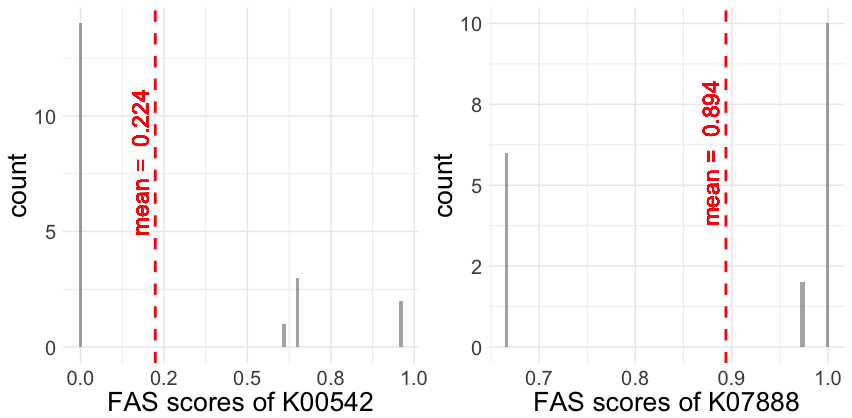


Figure 2.3. FAS score density of KO group K00542 (left) and K07888 (right)

In ortholog group K00542 (guanidinoacetate N-methyltransferase), only one protein member (rat rno:25257) has one Pfam domain (Orn\_DAP\_Arg\_deC). The lack of Pfam domain annotation of other proteins (human hsa:2593, mouse mmu:14431, zebrafish dre:796865 and *N.vectensis* nemve:1432) caused FAS scores of 0 for 14/20 pairwise comparisons and led to the low TFAS\_KO (0.224) for the whole group. On the contrary, the rich annotation of protein members of group K07888 (Ras-related protein Rab-5B) is the reason for its high TFAS\_KO.

# Benchmarking HamFAS

We used *S.cerevisiae* as a test species to benchmark our approach HamFAS. The protein set of yeast has been obtained from KEGG containing 3457 KO-annotated and 3158 un-annotated sequences. The annotated proteins have been used for evaluating the accuracy of the approach, while the un-annotated set has been used for estimating its sensitivity. The output of HamFAS is also compared with KAAS and BlastKOALA.

We removed *S.cerevisiae* out of the reference species list for avoiding redundant information while doing orthology search. The same reference species have been used for KAAS approach. With BlastKOALA, however, we couldn't remove yeast annotations out of the reference data.

The ortholog search has been also performed with different parameters to find the best settings for HaMStR (*-rbh*, *-checkCoorthologRef*).

## 3457 KO-annotated yeast proteins

With this data set, we tried to evaluate the accuracy of HamFAS in comparison to KAAS and BlastKOALA by calculating the recall, precision and F1 score.

recall = TP / (TP + FN)

precision = TP / (TP + FP)

F1 = (2\*precision\*recall)/(precision+recall)

Table 3.1 shows the evaluations of HamFAS, BlastKOALA and KAAS. HamFAS performed best in term of precision, while F1-score is lower then KAAS due to its lower recall. Interestingly, the latest annotation tool from KEGG, BlastKOALA, has the lowest scores in both recall and precision.

Table 3.1. Recall, precision and F1-score of HamFAS in comparison to BlastKOALA and KAAS. Second column shows values of HamFAS after filtering the orthology assignment with InParanoid's orthologs.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Approach** | **HamFAS** | **supported\_HamFAS** | **BlastKOALA** | **KAAS** |
| Recall | 0.915 | 0.861 | 0.905 | 0.931 |
| Precision | 0.985 | 0.985 | 0.979 | 0.984 |
| F1-score | 0.949 | 0.919 | 0.940 | 0.957 |

For checking the ortholog prediction result obtained by HaMStR, we evaluated the annotation transfer again using only orthologs that are supported by both HaMStR and InParanoid. Predicted KOs from HamFAS of 188 yeast proteins has been removed after filtering based on InParanoid's orthologs. It leads to the decrease of recall and F1-score. However, the precision is not affected (see Table 3.1). FAS scores of unsupported orthologs are slightly smaller than the ones of supported orthologs, with mean score of 0,918 and 0,988 respectively (see Figure 3.1).

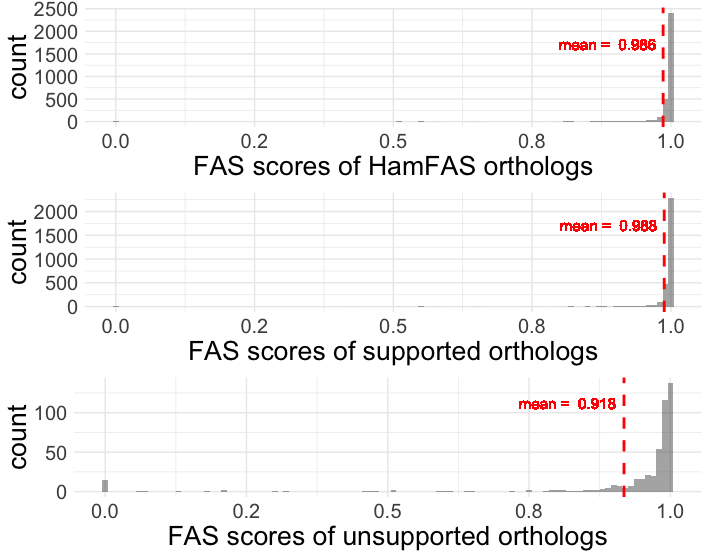


Figure 3.1. FAS score distribution of all HamFAS orthologs, only supported orthologs and unsupported orthologs

For a more detailed comparison between 3 approaches, we compare the fractions of proteins annotated by HamFAS, BlastKOALA and KAAS. 85,6% of the seed proteins has been annotated by all 3 approaches (Figure 3.2).

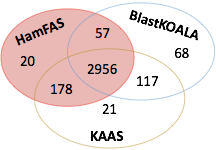


Figure 3.2. Fraction of proteins annotated by HamFAS, BlastKOALA and KAAS

There is a small difference between the KEGG identifiers annotated by each approach, which is shown in Table 3.2 below.

Table 3.2. Compare KEGG identifiers annotated by HamFAS, BlastKOALA and KAAS. Number in parentheses are the different KOs after filtered by synonymous KOs.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Approach** | **All 3 approaches** | **HamFAS +**  **BlastKOALA** | **HamFAS + KAAS** | **KAAS +**  **BlastKOALA** |
| Same KOs | 2951 | 54 | 168 | 108 |
| Diff. KOs | 5 (1) | 3 (1) | 10 (5) | 9 (6) |
| Total | 2956 | 57 | 178 | 117 |

Although those KEGG identifiers are different, most of them are "synonymous" KOs. They either have the same EC numbers, same EC classes, same GO numbers, or are the same components in KEGG pathways, responsible for the same reactions, etc.

Some examples of synonymous KOs:

* 1 KO is very general described (putative ABC transport system ATP-binding protein) while the other is specific (phospholipid/cholesterol/gamma-HCH transport system ATP-binding protein).
* Synonym/Alternative name: "septin" and "sporulation-regulated protein 3" (also septin); or "tristetraprolin" (ZFP36) and "butyrate response factor 1" (ZFP36L1).
* Involved in the same process: "cleavage stimulation factor subunit 2" and "polyadenylate-binding protein 2" are involved in 3-end formation of pre-mRNAs

## 3158 un-annotated yeast proteins

HamFAS could annotate 257 proteins, in which 164 proteins are HamFAS specific (HamFAS-only annotated proteins) (see Figure 3.3). In comparison to 150 and 116 annotated proteins from KAAS and BlastKOALA, HamFAS has annotated more proteins than BlastKOALA and KAAS (257 proteins versus 116 and 150 proteins, respectively)~~.~~

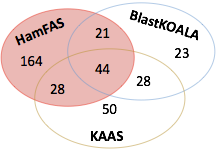


Figure 3.3. Fraction of proteins annotated by HamFAS, BlastKOALA and KAAS

Here the interesting part is the HamFAS-only proteins. So what are the differences of those proteins in comparison to others?

1. Was HaMStR so inclusive to include many false positive orthologs?

We also compared the orthology search of HamFAS with InParanoid. After removing ortholog pairs that are not predicted by InParanoid, 150 out of 257 proteins still can be annotated, 55 of them belong to HamFAS-only annotated proteins. It proved the reliability of the orthology assignment of HamFAS-only proteins.

1. Are HamFAS-only proteins short and uninformative?

We compared the sequence length and the informative content of protein domains between HamFAS-only proteins and proteins annotated by at least 2 approaches including HamFAS and KAAS and/or BlastKOALA.

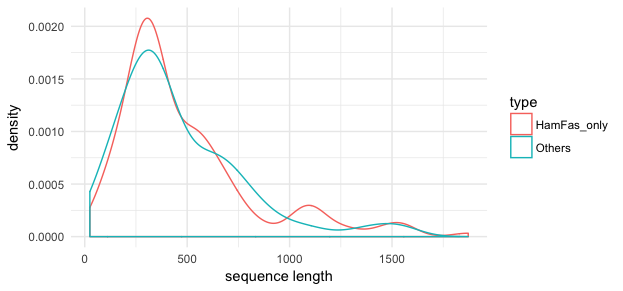


Figure 3.4. Length distribution of HamFAS-only proteins and others

Figure 3.4 and Figure 3.5 show no clear difference between those 2 protein sets. HamFAS-only proteins are not either extremely shorter or longer than other proteins. And the annotation transfer result was not driven by the uninformative domain annotation of those proteins (one Pfam domain that leads to the high FAS score).

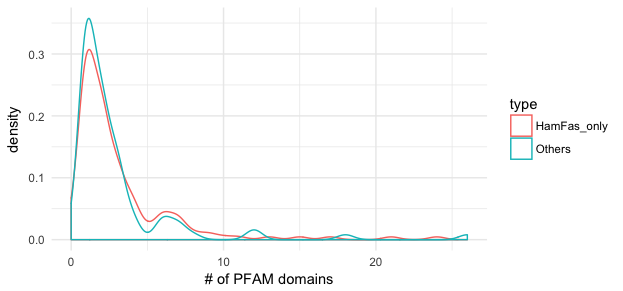


Figure 3.5. Number of Pfam domains distribution of HamFAS-only proteins and others

The distribution of FAS scores of all HamFAS orthologs in comparison to HamFAS-only orthologs shown in Figure 3.6 also confirms the rich domain annotations of HamFAS-only proteins.

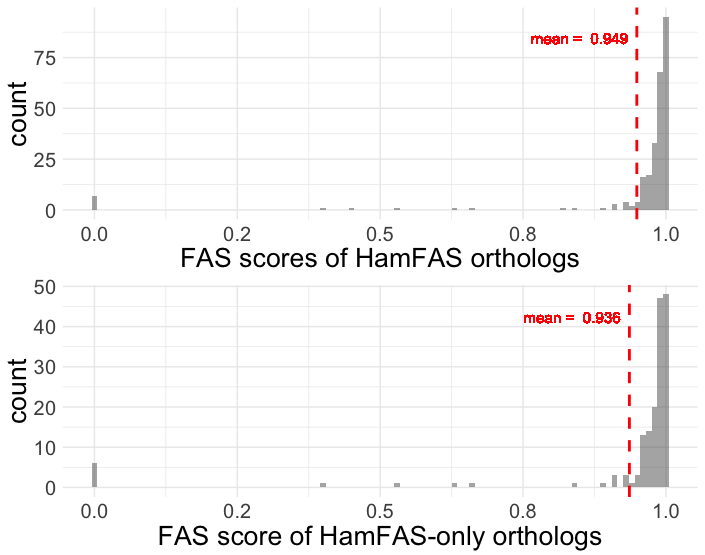


Figure 3.6. FAS score distribution of all HamFAS orthologs and HamFAS-only orthologs.

1. How different are the phylogenetic profile of KO-annotated proteins and un-annotated protein?

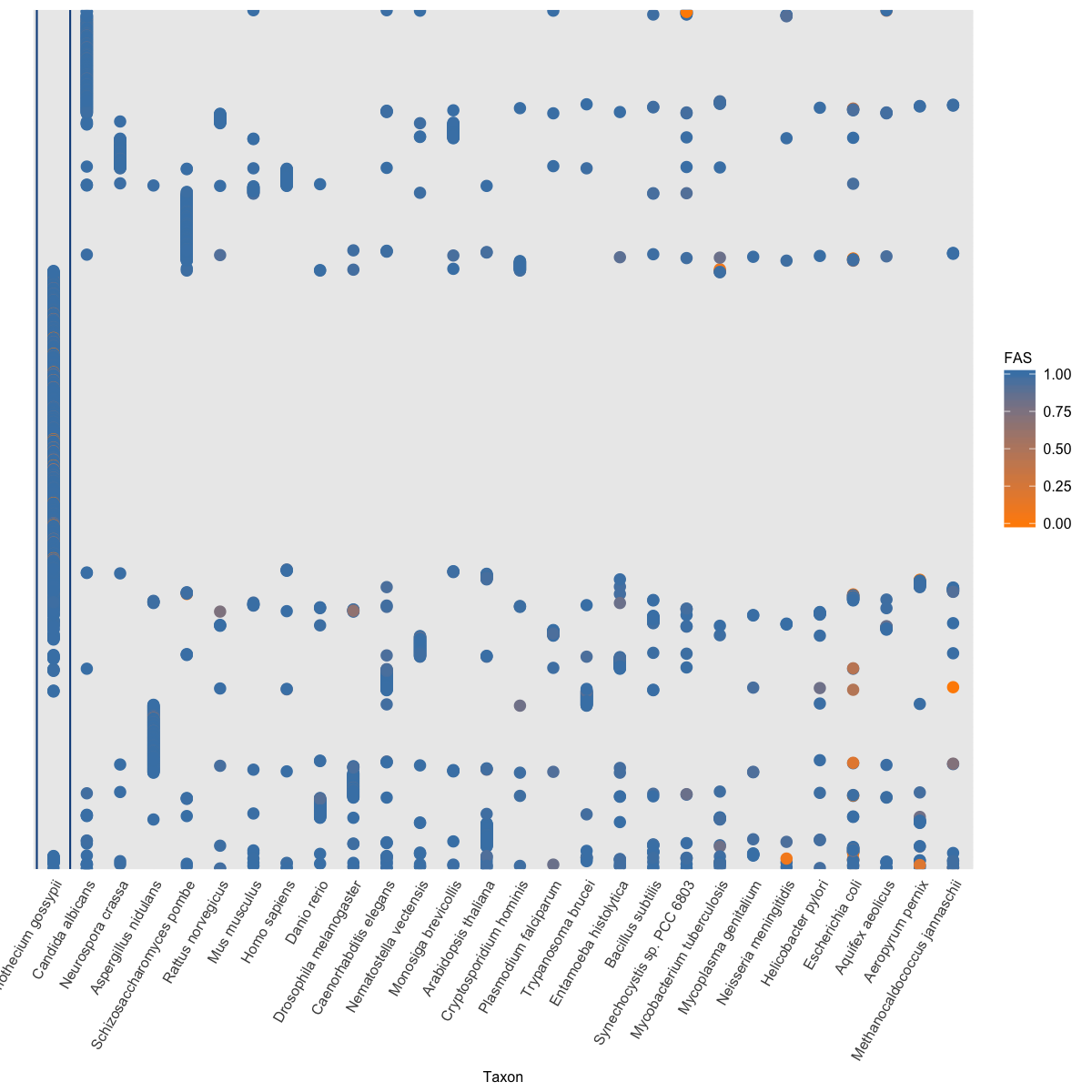


Figure 3.7. Phylogenetic profile of KO-annotated proteins

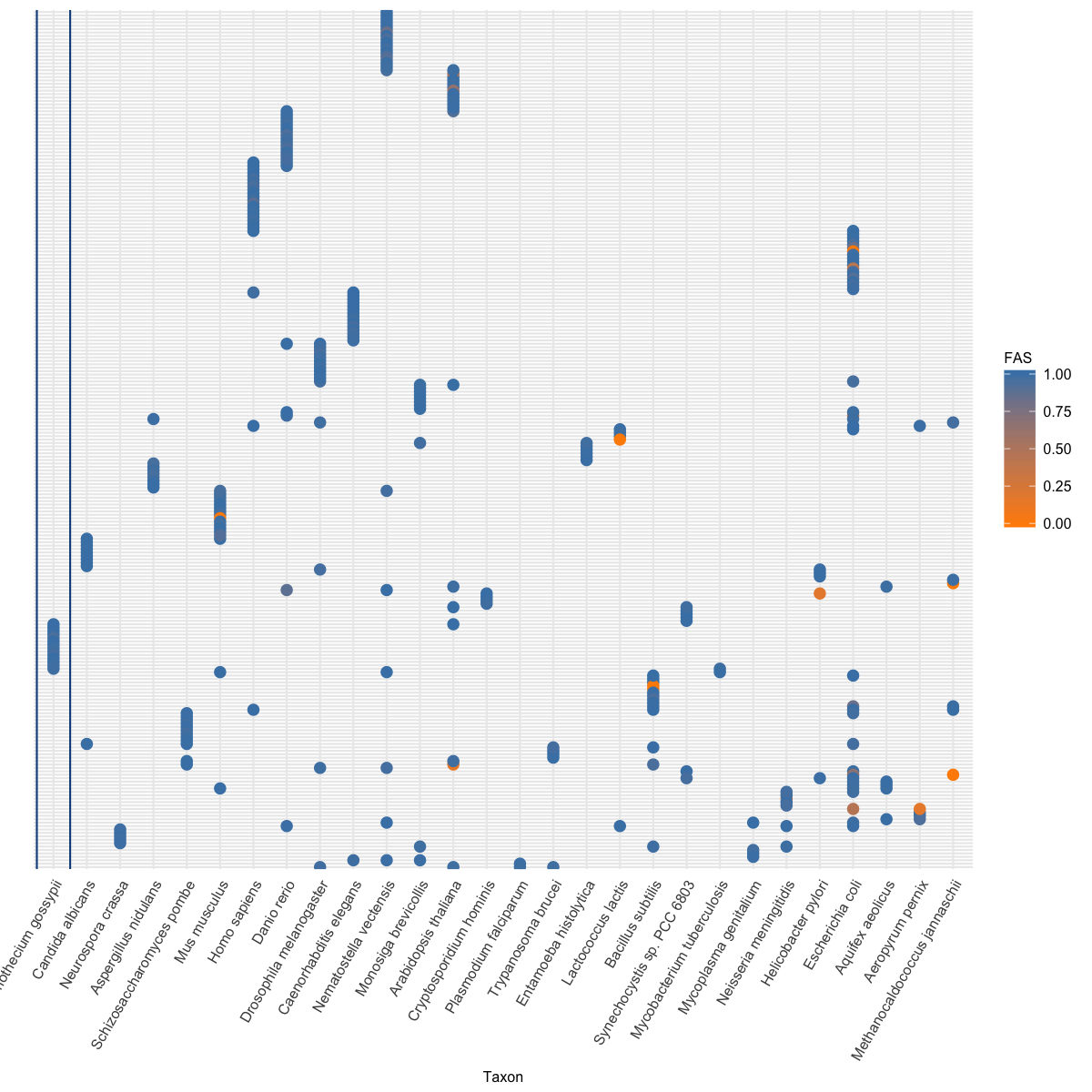


Figure 3.8. Phylogenetic profile of un-annotated proteins

Figure 3.7 and Figure 3.8 show that orthologs of un-annotated proteins are not broadly distributed like the one of annotated proteins. However, most of the proteins in both annotated and un-annotated set have only one ortholog (79% KO-annotated proteins, 80% un-annotated and 80% HamFAS-only proteins. See Figure 3.9). And more than 22% of un-annotated proteins have only orthologs in distantly related reference taxa (more detail in point 4).

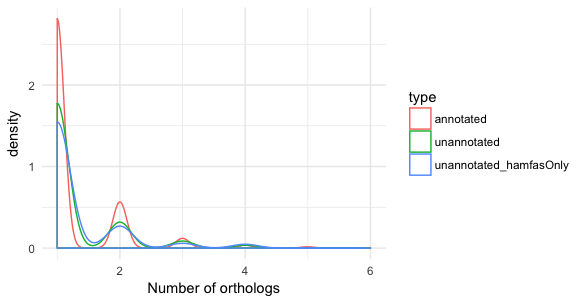


Figure 3.9. Distribution of number of orthologs for KO-annotated, un-annotated and HamFAS-only protein set

1. Do the annotations of HamFAS-only proteins come from distantly related species?

We checked for the origin of the annotations (i.e. the origin of reference orthologs) for all un-annotated proteins and compared with annotated set (Figure 3.10).

Figure 3.10. Origin of KO-annotations for annotated, un-annotated proteins and HamFAS-only proteins of un-annotated set

As expected, most annotations of annotated proteins come from their fungal orthologs (75%) while only few of them have obtained annotations from archaea or bacterial taxa (2,4%). In contrary, although large amount of annotations for un-annotated proteins originate from eukaryotes taxa (78%), there are still 22% (or 27% in case of HamFAS-only proteins) annotations are from distantly related taxa.

Analyzing the phylogenetic profile of proteins annotated by archaea and bacterial orthologs in Figure 3.11 and Figure 3.12, we can see that there is no difference between the HamFAS-only proteins and proteins that are annotated by both HamFAS and at least one other approach (HamFAS + BlastKOALA, HamFAS + KAAS or HamFAS + BlastKOALA + KAAS).

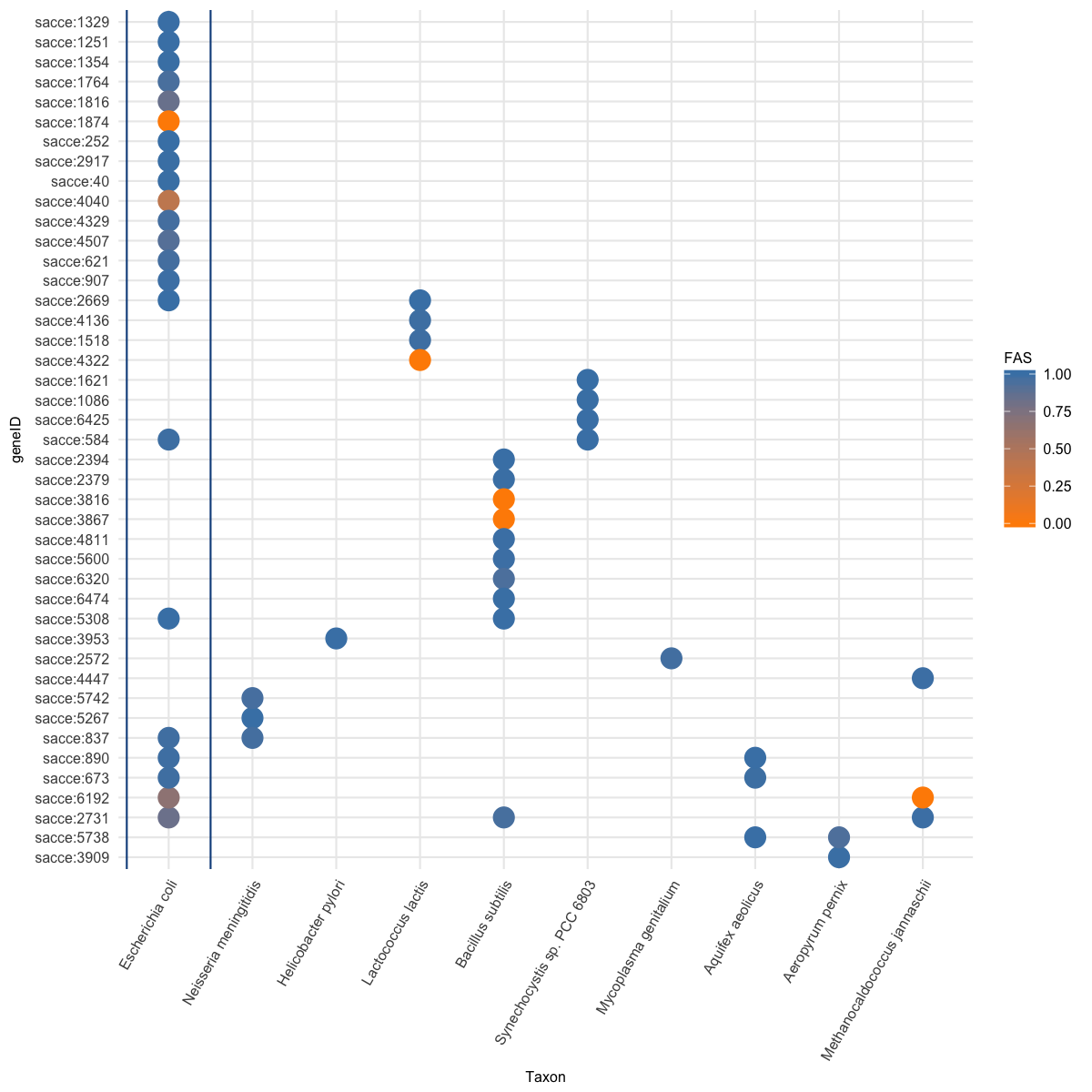


Figure 3.11. Phylogenetic profile of 44 HamFAS-only proteins that annotated based on archaea and bacterial orthologs.

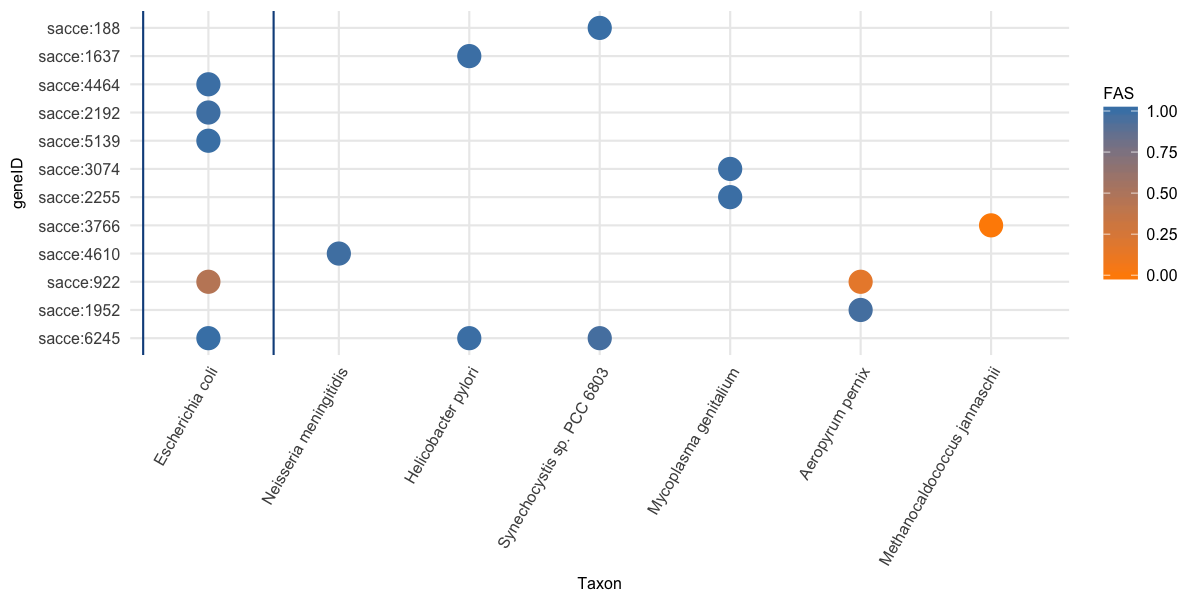


Figure 3.12. Phylogenetic profile of 12 un-annotated proteins that annotated by HamFAS and at least one other approach (BlastKOALA and/or KAAS), where their annotations originate from archaea or bacteria reference taxa.

1. How does the annotation result change by removing annotations from archaea and bacterial orthologs?

We filtered the annotations that originate from archaea and bacterial orthologs from both KO-annotated and un-annotated protein sets.

Table 3.3. Recall, precision and F1-score of filtered HamFAS in comparison to HamFAS, BlastKOALA and KAAS by applying on KO-annotated yeast proteins.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Approach** | **HamFAS after filtered** | **HamFAS** | **BlastKOALA** | **KAAS** |
| Recall | 0.9149 | 0.9152 | 0.905 | 0.931 |
| Precision | 0.9867 | 0.9854 | 0.979 | 0.984 |
| F1-score | 0.9496 | 0.9490 | 0.940 | 0.957 |

Table 3.3 shows a slightly increase in precision and F1-score of filtered HamFAS in comparison to original HamFAS due to a small number of annotations obtained from distantly related taxa.

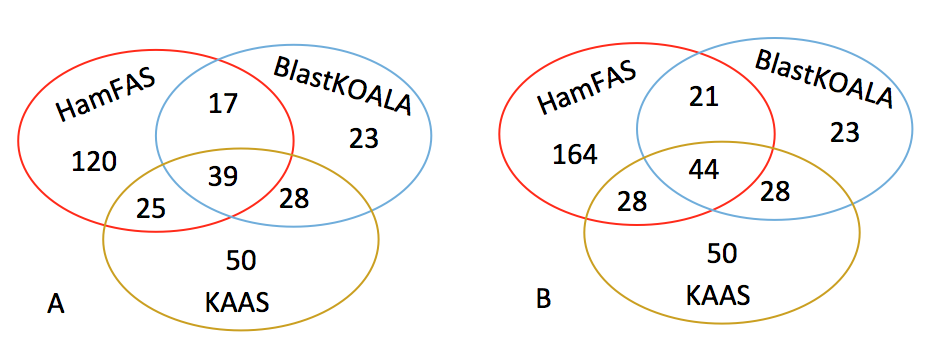


Figure 3.13. Fraction of proteins annotated by BlastKOALA, KAAS and filtered HamFAS (A) or original HamFAS (B)

In Figure 3.13 we observe a decrease of the number of proteins annotated by HamFAS. However there are still a large amount of proteins that are annotated only by HamFAS (120 proteins) in comparison to BlastKOALA (23 proteins) and KAAS (50 proteins).

1. Are annotated proteins involved in PPI networks or KEGG pathways?

We analyzed the connectivity of annotated proteins and the obtained KOs by calculating the node degree of those proteins in yeast protein-protein-interaction (PPI) networks and the occurrence of the annotated KOs in KEGG pathways. PPI data are retrieved from Yeast Interactome Project (http://interactome.dfci.harvard.edu/S\_cerevisiae/) and STRING database (https://string-db.org).

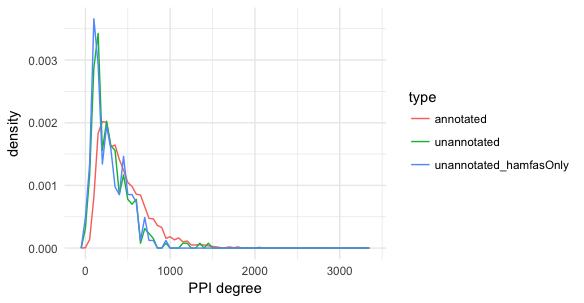


Figure 3.14. The PPI degree distribution of 3 protein sets

Figure 3.14 shows the distribution of PPI degree of KO-annotated, un-annotated and HamFAS-only proteins inside un-annotated set. KO-annotated proteins have in general more interacting partners (mean PPI degree 444) than un-annotated and HamFAS-only proteins (mean PPI degree 294 and 275 respectively). However, 99% of the proteins of un-annotated set have the PPI degree more than 10, while only 2 proteins don't have any interacting partner.

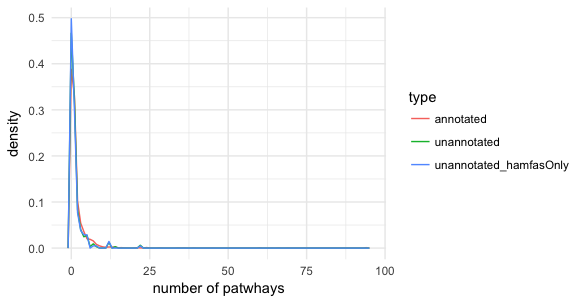


Figure 3.15. Distribution of the number of pathways in which annotated KOs are involved

For the annotated KOs, we calculate the number of pathways in which those KOs are involved. All 3 data sets show the same trend in Figure 3.15, that not less then 50% the KOs belong to at least one KEGG pathway (KO-annotated set 61%, un-annotated set 53% and HamFAS-only protein set 50%).

1. Are new annotations from HamFAS meaningless?

About 50% of KOs annotated only by HamFAS belongs to KEGG's pathways. Figure 3.16 shows the distribution of those KOs in different pathway categories.

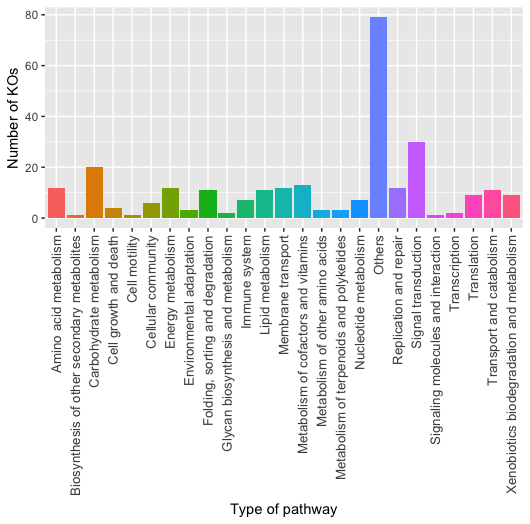
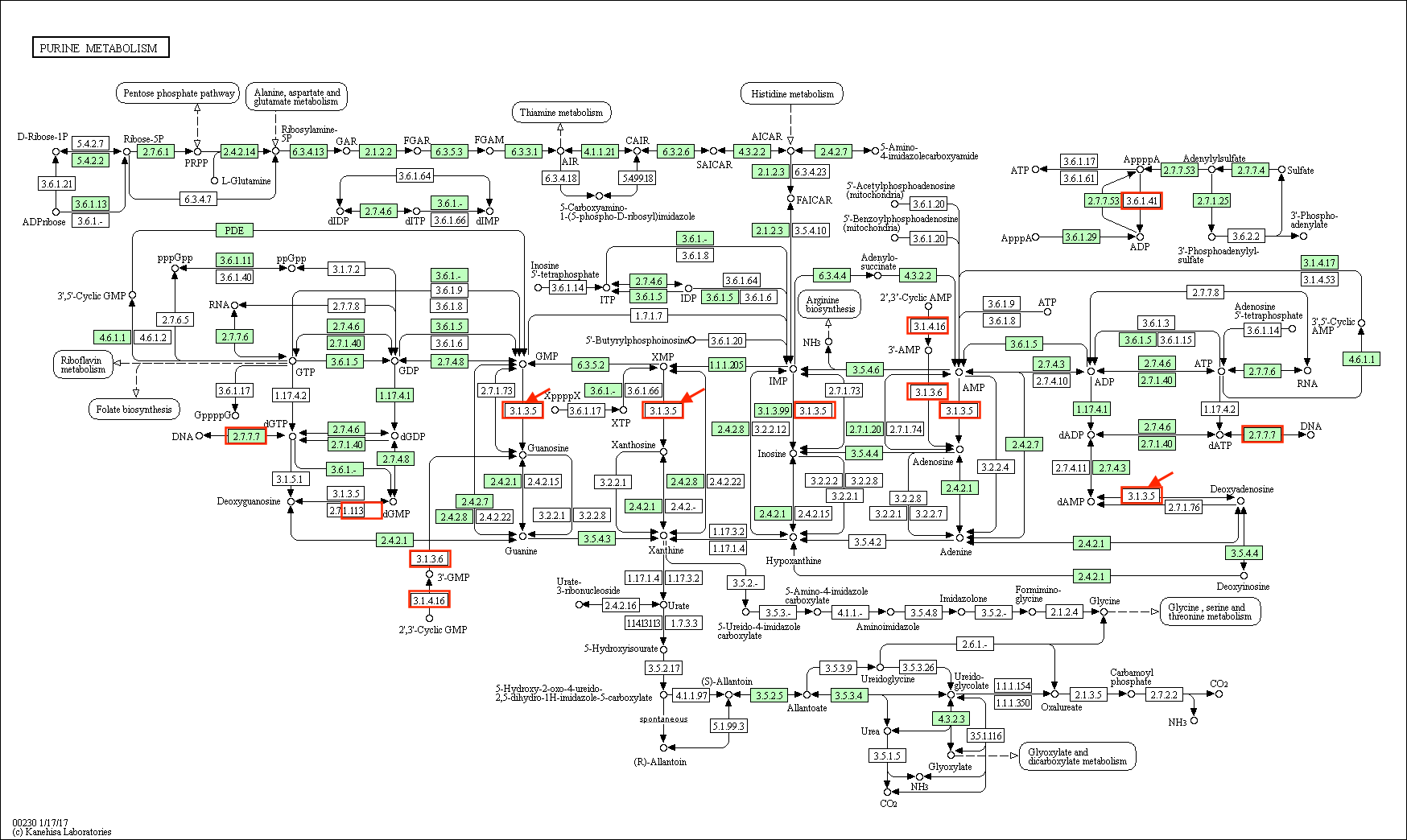
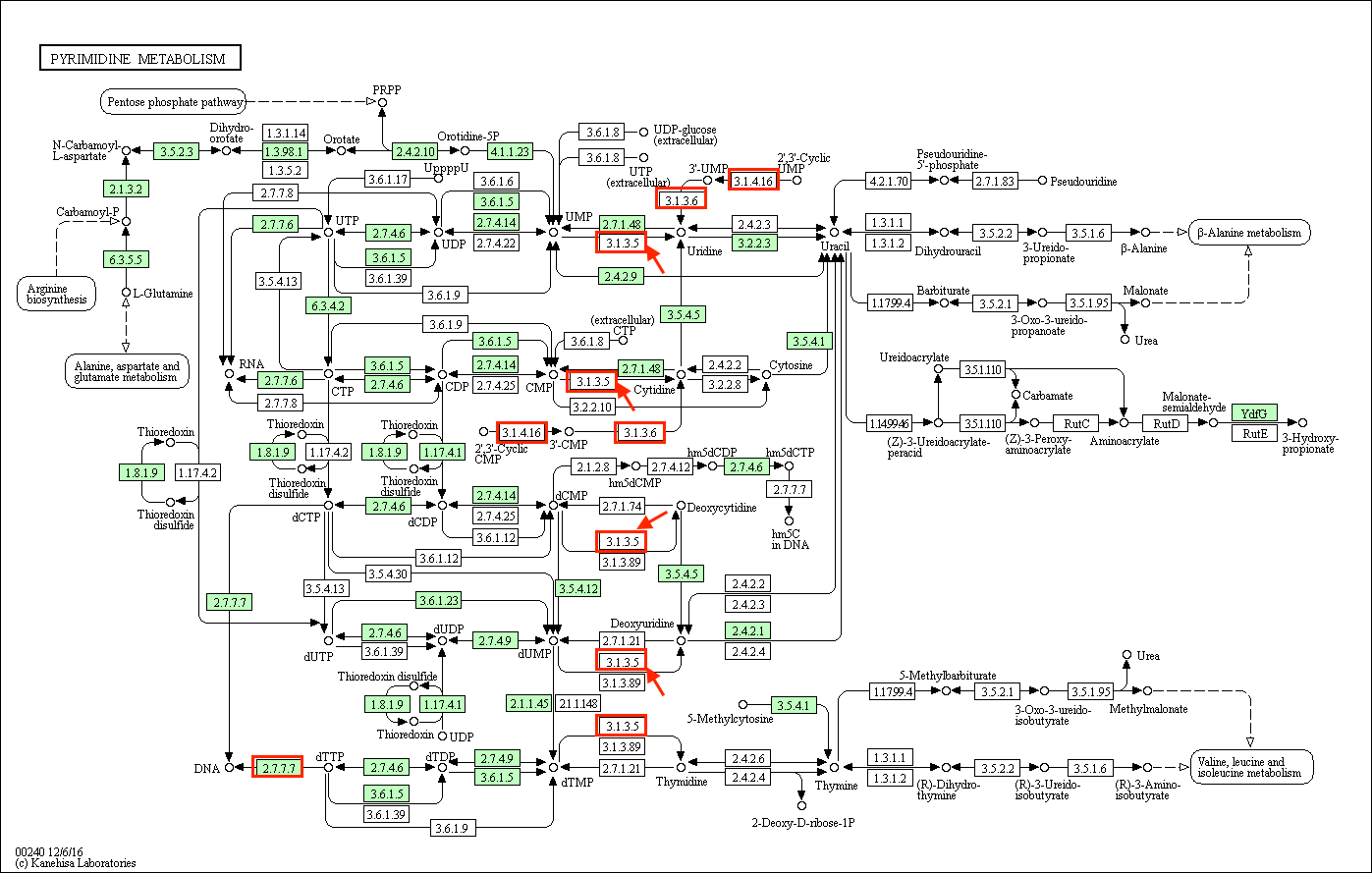
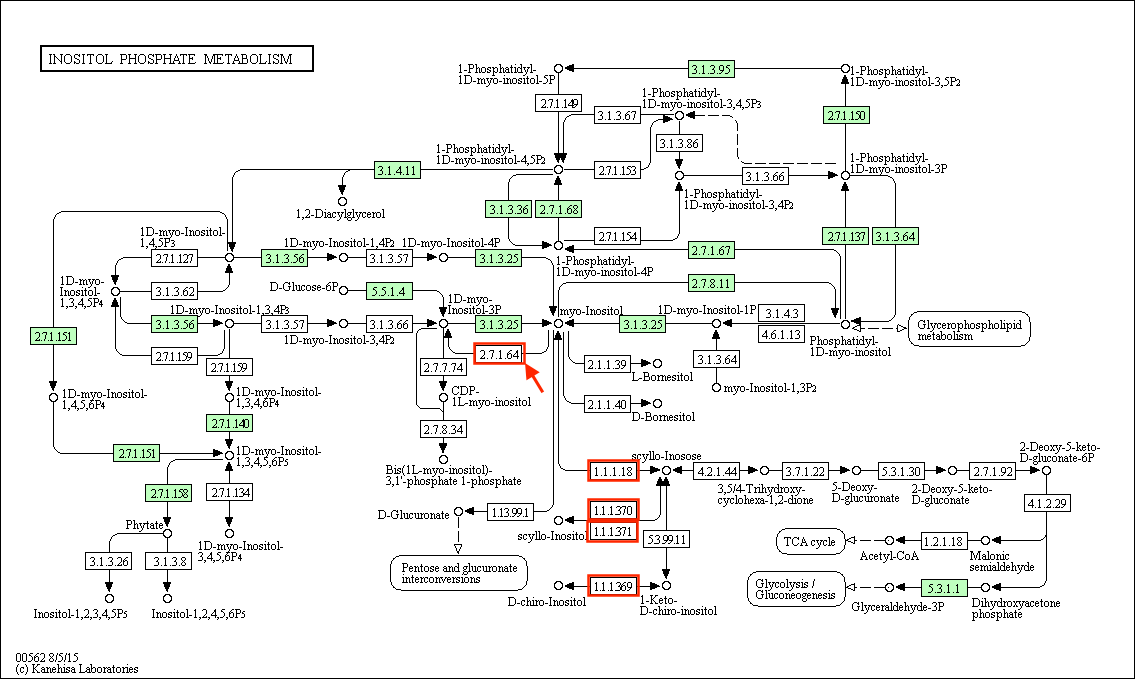


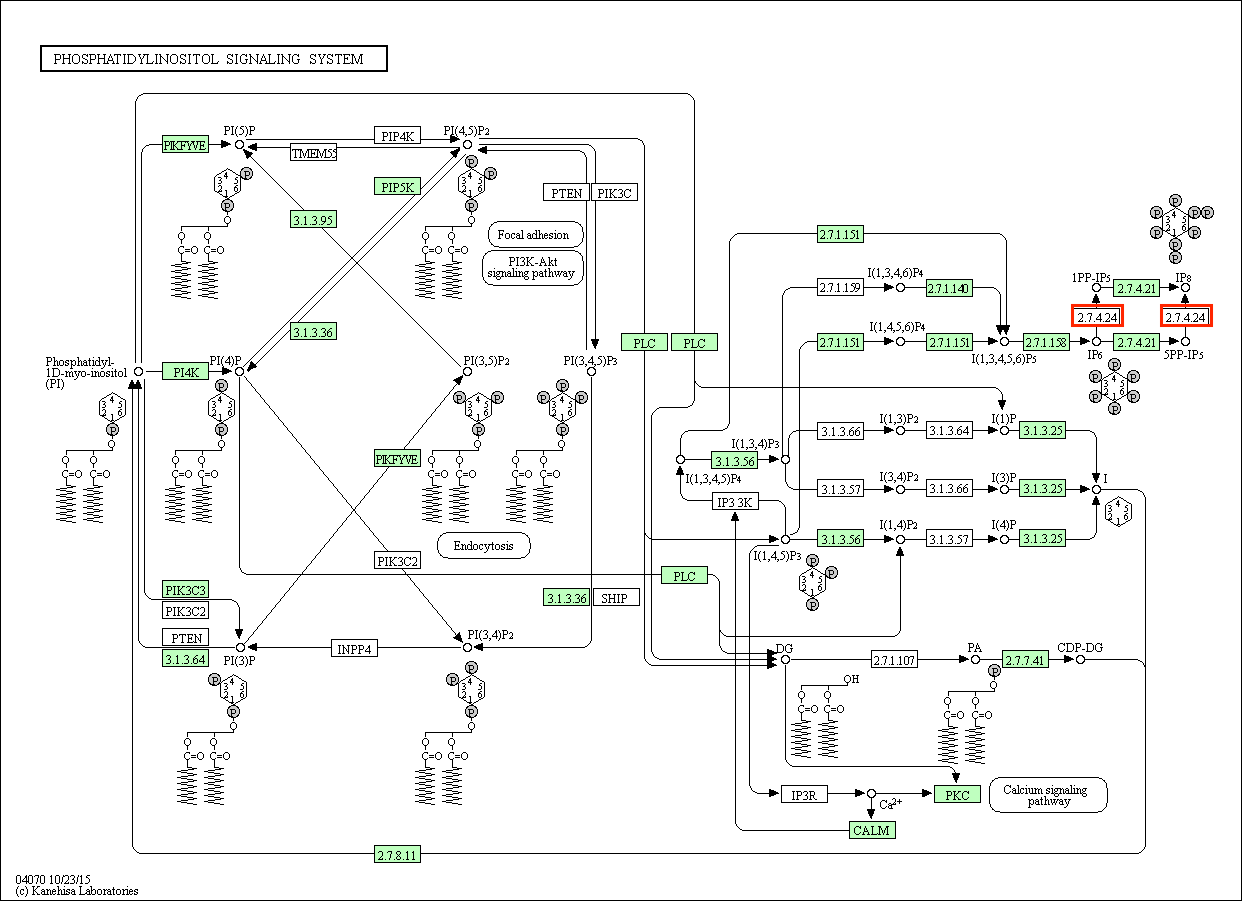
Figure 3.16. The numbers of HamFAS-only KOs distributed into different pathway categories

29 yeast pathways are further complemented by new KOs from HamFAS.









1. Why did BlastKOALA and KAAS fail to annotate HamFAS-only protein?

One reason could be, that the orthology prediction approaches used by KAAS and BlastKOALA are not as sensitive as HaMStR. The second reason is due the their "secret" filter criteria to select candidate for doing annotation transfer.

We have observed, that there are in total 86 predicted KOs that are common between annotated and unannotated data sets. In which 44 cases are also supported by InParanoid. Examples:

**K00077:** sce:4314, B.subtilis:BSU15110, B.subtilis:BSU14440 (KEGG's representative sequences)

HamFAS: sce:6474 (unannotated protein) is orthologous with B.subtilis:BSU14440

Inparanoid: 2 separate OGs: (sce:4314, B.subtilis:BSU15110) and (sce:6474, B.subtilis:BSU14440)

**K00799:** ath:AT1G02930,..., ath:AT2G30870,..., sce:5364, sce:1884

HamFAS: sce:2310 - ath:AT1G02930,...

(sce:1884 has no ortholog with *A.thaliana* according to InParanoid)

**K00877:** S.pombe:4570, S.pombe:875, S.pombe:1336, sce:1877, sce:997

HamFAS: sce:487 - S.pombe:1336

Those proteins have been probably either not predicted as orthologs or discarded after filtering through KEGG annotation pipeline.

# Conclusion

The ability of identifying distantly related orthologs of HaMStR leads to the result that more proteins have been annotated by HamFAS than BlastKOALA or KAAS. There is no strong evidence to distinct the difference between HamFAS-only proteins and proteins that are annotated by both HamFAS and other approaches. We can increase the stringency of HamFAS by allowing the annotations from only close related species. However, if doing so we will lose the benefit of HaMStR and therefore we have no reason to use HaMStR instead of other more stringent orthology search approaches like OMA or InParanoid. The principle factor that affects the annotation result is the accuracy of orthology assignment method. This HamFAS approach could be supported more by the analysis of QfO from Holger.

# KEGG ID

KEGG, Kyoto encyclopedia of genes and genomes, is a resource for analysis of gene functions. It contains 16 main databases, which divided into 4 groups: System information (PATHWAY, BRITE, MODULE), Genomic information (ORTHOLOGY, GENES, GENOME), Chemical information or KEGG ligand (COMPOUND, GLYCAN, REACTION, RPAIR, RCLASS, ENZYME), Health information (DISEASE, DRUG, DGROUP, ENVIRON)[[1]](#footnote-1),[[2]](#footnote-2).

KEGG Orthology (KO) database is a hub to link genomic information to system information and chemical information2. Each KO entry, defined by a K number, is comprised of context (pathway) - dependent similarity sequences[[3]](#footnote-3). Originally, KO database is developed together with KEGG pathway maps, BRITE functional hierarchies and KEGG modules based on experimental knowledge2,3. The assignment of K numbers to KEGG GENES using auto KOALA (KEGG Orthology And Links Annotation) algorithm is a highly computerized process. Manual KOALA checking is required if there are discrepancies between current annotations and new assignments (see Figure 5.1).

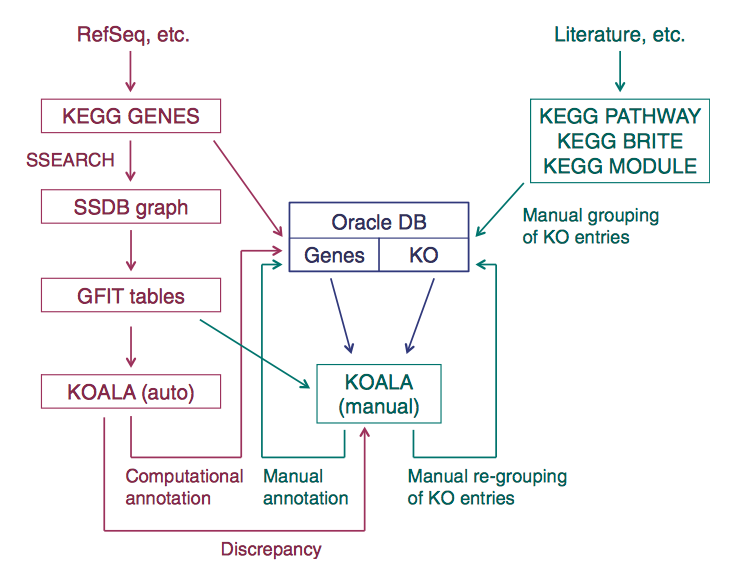


Figure 5.1. A schematic diagram of genome annotation in KEGG. It consists of two parts: manually defining KO entries represented by K numbers (right) and highly computerized assigning K numbers to genes in complete genomes (left)3.

# KAAS (KEGG Automatic Annotation Server)[[4]](#footnote-4)

KAAS is an online annotation server provided by KEGG for assigning K numbers to user-defined genes.

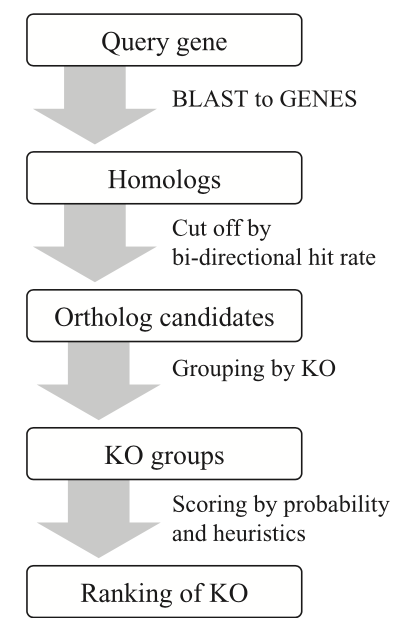


Figure 6.1. The overall procedure of KAAS.

Figure 6.1 shows the overall procedure of KAAS. Result from BLAST (bit score 60, % identity 35) is filtered by bi-directional hit rate (BHR) to obtain ortholog candidates. BHR is defined as:

with *R = S'/Sb*, where *S'* is the bit score of *a* against *b* and *Sb* is the score of *a* against the best hit in genome *B*, which can be different from *b*. *Rf* and *Rr* refer to forward and reverse BLAST, respectively. Default BHR cut-off is 0.95.

Ortholog groups (OGs) will than be ranked based on score SKO in order to find the best matching K number (best SKO) for the query gene.

where *Sh* is the highest score among all ortholog candidates in the OG (bit score),

m, n are lengths of query and target sequences of BLAST,

N is the number of OG members,

x is the number of organisms in the original OG (before removing hits < BHR),

p is the ratio of the size of the original OG versus the size of the entire GENES DB.

# BlastKOALA[[5]](#footnote-5)

BlastKOALA is a newly published automatic annotation server of KEGG. Set of query genes will be blasted against a non-redundant dataset created from GENES database. A modified KOALA algorithm is used to assign K numbers to query sequences based on the GFIT (Gene Function Identification Tool)-like table (see Figure 7.1) converted from BLAST result. KOALA computes the weighted sum of BLAST bit scores for each group and decides the best fitting K numbers. The weighting factors of the modified KOALA consist of overlap length of the alignment, the ratio of query and target sequence lengths, the degree of matches of taxonomic categories (if known) and the degree of matches Pfam domains. It does not take the bidirectional best-hit information into account like the original KOALA.

=> choose some examples where hamfas and inparanoid predicted orthologs but blastkoala didn't

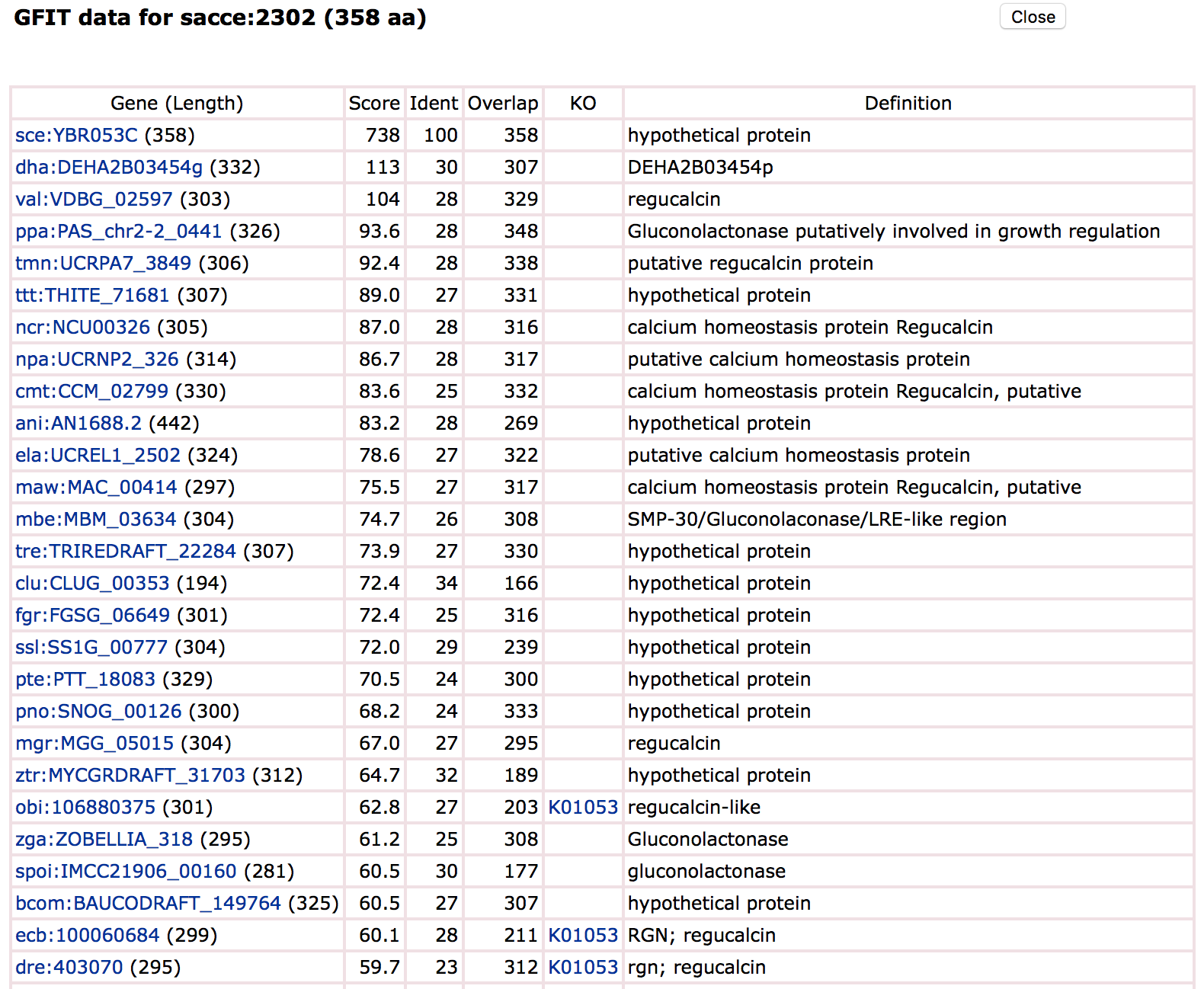


Figure 7.1. Part of the GFIT-like table for query sequence sacce:2302.

1. Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., & Kanehisa, M. (1999). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, *27*(1), 29-34. [↑](#footnote-ref-1)
2. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., & Tanabe, M. (2016). KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research*, *44*(D1), 457–462. [↑](#footnote-ref-2)
3. Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M., & Tanabe, M. (2014). Data, information, knowledge and principle: Back to metabolism in KEGG. *Nucleic Acids Research*, *42*(D1). [↑](#footnote-ref-3)
4. Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A. C., & Kanehisa, M. (2007). KAAS: An automatic genome annotation and pathway reconstruction server. *Nucleic Acids Research*, *35*(SUPPL.2). [↑](#footnote-ref-4)
5. Kanehisa, M., Sato, Y., & Morishima, K. (2016). BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences. *Journal of Molecular Biology, 428*(4), 726 - 731. [↑](#footnote-ref-5)