Supplementary Figures and Tables

The EFI Web Resource for Genomic Enzymology Web Tools: Leveraging Protein, Genome, and Metagenome Databases to Discover Novel Enzymes and Metabolic Pathways

Rémi Zallot[1], Nils O. Oberg[1], and John A. Gerlt[*,1, 2, 3]

[1]Institute for Genomic Biology, [2]Department of Biochemistry, and [3]Department of Chemistry, University of Illinois at Urbana-Champaign, 1206 West Gregory Drive, Urbana, Illinois 61801, United States.

Corresponding author: John A. Gerlt (j-gerlt@illinois.edu).

Phone: +1-217-244-7414

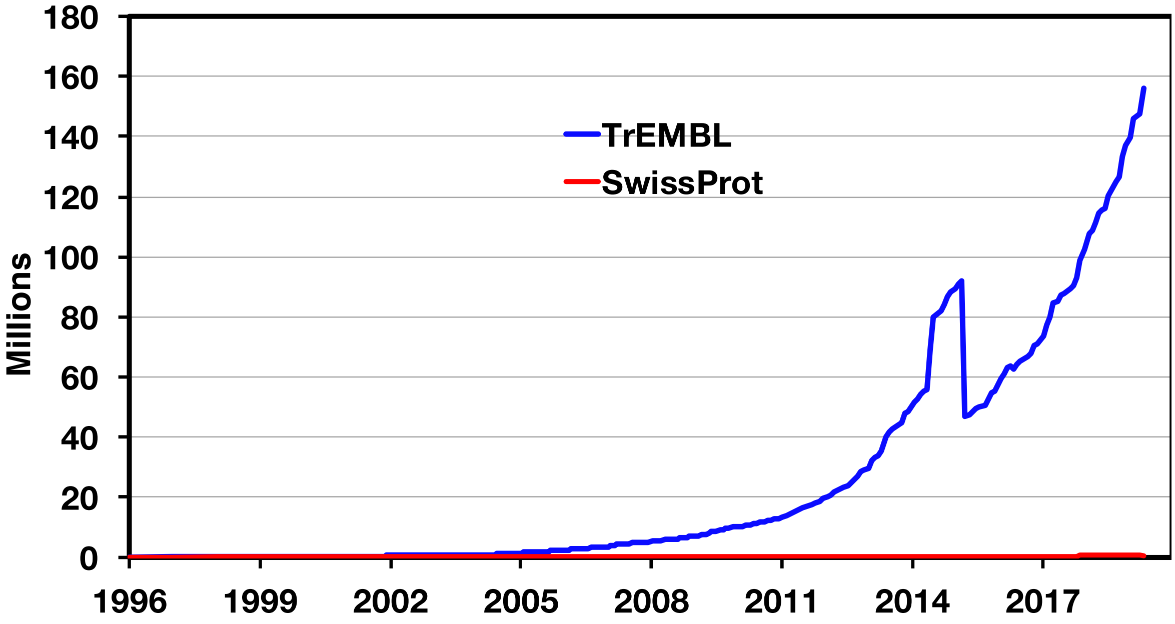
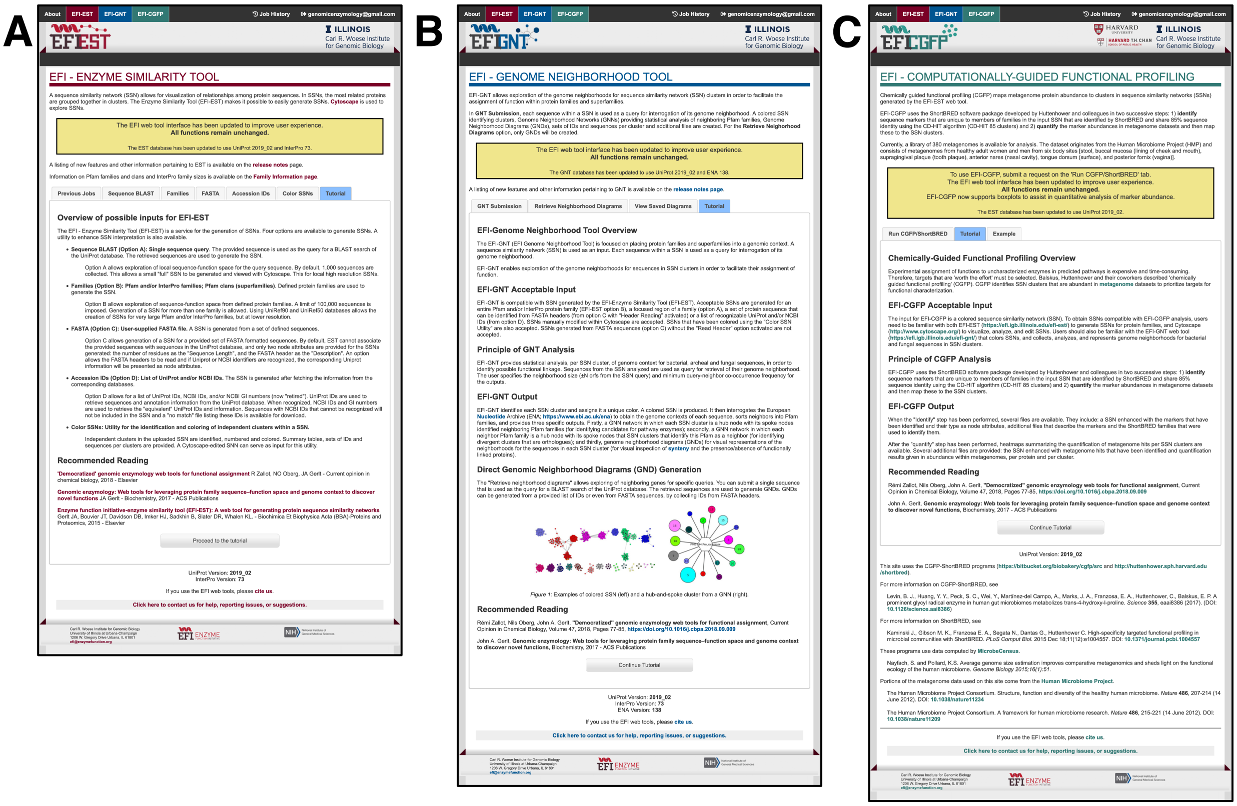


Figure S1. **The growth of the UniProtKB database**. The web tools use the UniProtKB database for sequences used to generate SSNs and bioinformatic data included as node attributes in the SSNs. UniProtKB is the aggregate of the UniProtKB/TrEMBL database that contains computationally annotated entries [156,077,686 in Release 2019_04 (08-May-2019)] and the UniProtKB/SwissProt database that contains manually curated entries (560,118 in Release 2019_04). The decrease in 2015 is the result of archiving sequences from redundant proteomes in UniParc to manage the growth of the database.



EFI - ENZYME SIMILARITY TOOL

A sequence similarity network (SSN) allows for visualization of relationships among protein sequences. In SSNs, the most related proteins are grouped together in clusters. The Enzyme Similarity Tool (EFI-EST) makes it possible to easily generate SSNs. **Cytoscape** is used to explore SSNs.

The EFI web tool interface has been updated to improve user experience.
All functions remain unchanged.

The EST database has been updated to use UniProt 2019_02 and InterPro 73.

A listing of new features and other information pertaining to EST is available on the **release notes** page.

Information on Pfam families and clans and InterPro family sizes is available on the **Family Information page**.

Overview of possible inputs for EFI-EST

The EFI - Enzyme Similarity Tool (EFI-EST) is a service for the generation of SSNs. Four options are available to generate SSNs. A utility to enhance SSN interpretation is also available.

- **Sequence BLAST (Option A): Single sequence query**. The provided sequence is used as the query for a BLAST search of the UniProt database. The retrieved sequences are used to generate the SSN.

  Option A allows exploration of local sequence-function space for the query sequence. By default, 1,000 sequences are collected. This allows a small "full" SSN to be generated and viewed with Cytoscape. This for local high resolution SSNs.
- **Families (Option B): Pfam and/or InterPro families; Pfam clans (superfamilies)**. Defined protein families are used to generate the SSN.

  Option B allows exploration of sequence-function space from defined protein families. A limit of 100,000 sequences is imposed. Generation of a SSN for more than one family is allowed. Using UniRef90 and UniRef50 databases allows the creation of SSNs for very large Pfam and/or InterPro families, but at lower resolution.
- **FASTA (Option C): User-supplied FASTA file.** A SSN is generated from a set of defined sequences.

  Option C allows generation of a SSN for a provided set of FASTA formatted sequences. By default, EST cannot associate the provided sequences with sequences in the UniProt database, and only two node attributes are provided for the SSNs generated: the number of residues as the "Sequence Length", and the FASTA header as the "Description". An option allows the FASTA headers to be read and if Uniprot or NCBI identifiers are recognized, the corresponding Uniprot information will be presented as node attributes.
- **Accession IDs (Option D): List of UniProt and/or NCBI IDs.** The SSN is generated after fetching the information from the corresponding databases.

  Option D allows for a list of UniProt IDs, NCBI IDs, and/or NCBI GI numbers (now "retired"). UniProt IDs are used to retrieve sequences and annotation information from the UniProt database. When recognized, NCBI IDs and GI numbers are used to retrieve the "equivalent" UniProt IDs and information. Sequences with NCBI IDs that cannot be recognized will not be included in the SSN and a "no match" file listing these IDs is available for download.
- **Color SSNs: Utility for the identification and coloring of independent clusters within a SSN.**

  Independent clusters in the uploaded SSN are identified, numbered and colored. Summary tables, sets of IDs and sequences per clusters are provided. A Cytoscape-edited SNN can serve as input for this utility.

Recommended Reading

'Democratized' genomic enzymology web tools for functional assignment R Zallot, NO Oberg, JA Gerlt - Current opinion in chemical biology, 2018 - Elsevier

Genomic enzymology: Web tools for leveraging protein family sequence–function space and genome context to discover novel functions JA Gerlt - Biochemistry, 2017 - ACS Publications

Enzyme function initiative-enzyme similarity tool (EFI-EST): A web tool for generating protein sequence similarity networks Gerlt JA, Bouvier JT, Davidson DB, Imker HJ, Sadkhin B, Slater DR, Whalen KL. - Biochimica Et Biophysica Acta (BBA)-Proteins and Proteomics, 2015 - Elsevier

UniProt Version: **2019_02**
InterPro Version: **73**

If you use the EFI web tools, please **cite us**.

Click here to contact us for help, reporting issues, or suggestions.

Carl R. Woese Institute for Genomic Biology
University of Illinois at Urbana-Champaign
1206 W. Gregory Drive Urbana, IL 61801
efi@enzymefunction.org

EFI - GENOME NEIGHBORHOOD TOOL

EFI-GNT allows exploration of the genome neighborhoods for sequence similarity network (SSN) clusters in order to facilitate the assignment of function within protein families and superfamilies.

In **GNT Submission**, each sequence within a SSN is used as a query for interrogation of its genome neighborhood. A colored SSN identifying clusters, Genome Neighborhood Networks (GNNs) providing statistical analysis of neighboring Pfam families, Genome Neighborhood Diagrams (GNDs), sets of IDs and sequences per cluster and additional files are created. For the **Retrieve Neighborhood Diagrams** option, only GNDs will be created.

The EFI web tool interface has been updated to improve user experience.
All functions remain unchanged.

The GNT database has been updated to use UniProt 2019_02 and ENA 138.

A listing of new features and other information pertaining to GNT is available on the **release notes page**.

EFI-Genome Neighborhood Tool Overview

The EFI-GNT (EFI Genome Neighborhood Tool) is focused on placing protein families and superfamilies into a genomic context. A sequence similarity network (SSN) is used as an input. Each sequence within a SSN is used as a query for interrogation of its genome neighborhood.

EFI-GNT enables exploration of the genome neighborhoods for sequences in SSN clusters in order to facilitate their assignment of function.

EFI-GNT Acceptable Input

EFI-GNT is compatible with SSN generated by the EFI-Enzyme Similarity Tool (EFI-EST). Acceptable SSNs are generated for an entire Pfam and/or InterPro protein family (EFI-EST option B), a focused region of a family (option A), a set of protein sequence that can be identified from FASTA headers (from option C with "Header Reading" activated) or a list of recognizable UniProt and/or NCBI IDs (from option D). SSNs manually modified within Cytoscape are accepted. SSNs that have been colored using the "Color SSN Utility" are also accepted. SSNs generated from FASTA sequences (option C) without the "Read Header" option activated are not accepted.

Principle of GNT Analysis

EFI-GNT provides statistical analysis, per SSN cluster, of genome context for bacterial, archeal and fungal sequences, in order to identify possible functional linkage. Sequences from the SSN analyzed are used as query for retrieval of their genome neighborhood. The user specifies the neighborhood size (±N orfs from the SSN query) and minimum query-neighbor co-occurrence frequency for the outputs.

EFI-GNT Output

EFI-GNT identifies each SSN cluster and assigns it a unique color. A colored SSN is produced. It then interrogates the European **Nucleotide** Archive (ENA; **https://www.ebi.ac.uk/ena**) to obtain the genome contexts of each sequence, sorts neighbors into Pfam families, and provides three specific outputs. Firstly, a GNN network in which each SSN cluster is a hub node with its spoke nodes identified neighboring Pfam families (for identifying candidates for pathway enzymes); secondly, a GNN network in which each neighbor Pfam family is a hub node with its spoke nodes that SSN clusters that identify this Pfam as a neighbor (for identifying divergent clusters that are orthologues); and thirdly, genome neighborhood diagrams (GNDs) for visual representations of the neighborhoods for the sequences in each SSN cluster (for visual inspection of **synteny** and the presence/absence of functionally linked proteins).

Direct Genomic Neighborhood Diagrams (GND) Generation

The "Retrieve neighborhood diagrams" allows exploring of neighboring genes for specific queries. You can submit a single sequence that is used as the query for a BLAST search of the UniProt database. The retrieved sequences are used to generate GNDs. GNDs can be generated from a provided list of IDs or even from FASTA sequences, by collecting IDs from FASTA headers.

Figure 1: Examples of colored SSN (left) and a hub-and-spoke cluster from a GNN (right).

Recommended Reading

Rémi Zallot, Nils Oberg, John A. Gerlt, **"Democratized" genomic enzymology web tools for functional assignment**, Current Opinion in Chemical Biology, Volume 47, 2018, Pages 77-85, **https://doi.org/10.1016/j.cbpa.2018.09.009**

John A. Gerlt, **Genomic enzymology: Web tools for leveraging protein family sequence–function space and genome context to discover novel functions**, Biochemistry, 2017 - ACS Publications

UniProt Version: **2019_02**
InterPro Version: **73**
ENA Version: **138**

If you use the EFI web tools, please **cite us**.

Click here to contact us for help, reporting issues, or suggestions.

Carl R. Woese Institute for Genomic Biology
University of Illinois at Urbana-Champaign
1206 W. Gregory Drive Urbana, IL 61801
efi@enzymefunction.org

EFI - COMPUTATIONALLY-GUIDED FUNCTIONAL PROFILING

Chemically guided functional profiling (CGFP) maps metagenome protein abundance to clusters in sequence similarity networks (SSNs) generated by the EFI-EST web tool.

EFI-CGFP uses the ShortBRED software package developed by Huttenhower and colleagues in two successive steps: 1) **identify** sequence markers that are unique to members of families in the input SSN that are identified by ShortBRED and share 85% sequence identity using the CD-HIT algorithm (CD-HIT 85 clusters) and 2) **quantify** the marker abundances in metagenome datasets and then map these to the SSN clusters.

Currently, a library of 380 metagenomes is available for analysis. The dataset originates from the Human Microbiome Project (HMP) and consists of metagenomes from healthy adult women and men from six body sites [stool, buccal mucosa (lining of cheek and mouth), supragingival plaque (tooth plaque), anterior nares (nasal cavity), tongue dorsum (surface), and posterior fornix (vagina)].

To use EFI-CGFP, submit a request on the 'Run CGFP/ShortBRED' tab.
The EFI web tool interface has been updated to improve user experience.
All functions remain unchanged.
EFI-CGFP now supports boxplots to assist in quantitative analysis of marker abundance.

The EST database has been updated to use UniProt 2019_02.

Chemically-Guided Functional Profiling Overview

Experimental assignment of functions to uncharacterized enzymes in predicted pathways is expensive and time-consuming. Therefore, targets that are 'worth the effort' must be selected. Balskus, Huttenhower and their coworkers described 'chemically guided functional profiling' (CGFP). CGFP identifies SSN clusters that are abundant in **metagenome** datasets to prioritize targets for functional characterization.

EFI-CGFP Acceptable Input

The input for EFI-CGFP is a colored sequence similarity network (SSN). To obtain SSNs compatible with EFI-CGFP analysis, users need to be familiar with both EFI-EST (**https://efi.igb.illinois.edu/efi-est/**) to generate SSNs for protein families, and Cytoscape (**http://www.cytoscape.org/**) to visualize, analyze, and edit SSNs. Users should also be familiar with the EFI-GNT web tool (**https://efi.igb.illinois.edu/efi-gnt/**) that colors SSNs, and collects, analyzes, and represents genome neighborhoods for bacterial and fungal sequences in SSN clusters.

Principle of CGFP Analysis

EFI-CGFP uses the ShortBRED software package developed by Huttenhower and colleagues in two successive steps: 1) **identify** sequence markers that are unique to members of families in the input SSN that are identified by ShortBRED and share 85% sequence identity using the CD-HIT algorithm (CD-HIT 85 clusters) and 2) **quantify** the marker abundances in metagenome datasets and then map these to the SSN clusters.

EFI-CGFP Output

When the "Identify" step has been performed, several files are available. They include: a SSN enhanced with the markers that have been identified and their type as node attributes, additional files that describe the markers and the ShortBRED families that were used to identify them.

After the "quantify" step has been performed, heatmaps summarizing the quantification of metagenome hits per SSN clusters are available. Several additional files are provided: the SSN enhanced with metagenome hits that have been identified and quantification results given in abundance within metagenomes, per protein and per cluster.

Recommended Reading

Rémi Zallot, Nils Oberg, John A. Gerlt, **"Democratized" genomic enzymology web tools for functional assignment**, Current Opinion in Chemical Biology, Volume 47, 2018, Pages 77-85, **https://doi.org/10.1016/j.cbpa.2018.09.009**

John A. Gerlt, **Genomic enzymology: Web tools for leveraging protein family sequence–function space and genome context to discover novel functions**, Biochemistry, 2017 - ACS Publications

UniProt Version: **2019_02**

This site uses the CGFP-ShortBRED programs (**https://bitbucket.org/biobakery/cgfp/src** and **http://huttenhower.sph.harvard.edu/shortbred**).

For more information on CGFP-ShortBRED, see

Levin, B. J., Huang, Y. Y., Peck, S. C., Wei, Y., Martínez-del Campo, A., Marks, J. A., Franzosa, E. A., Huttenhower, C., Balskus, E. P. A prominent glycyl radical enzyme in human gut microbiomes metabolizes trans-4-hydroxy-l-proline. *Science* **355**, eaai8386 (2017). (DOI: **10.1126/science.aai8386**)

For more information on ShortBRED, see

Kaminski J., Gibson M. K., Franzosa E. A., Segata N., Dantas G., Huttenhower C. High-specificity targeted functional profiling in microbial communities with ShortBRED. *PLoS Comput Biol.* 2015 Dec 18;11(12):e1004557. DOI: **10.1371/journal.pcbi.1004557**

These programs use data computed by **MicrobeCensus**.

Nayfach, S. and Pollard, K.S. Average genome size estimation improves comparative metagenomics and sheds light on the functional ecology of the human microbiome. *Genome Biology 2015;16(1):51.*

Portions of the metagenome data used on this site come from the **Human Microbiome Project**.

The Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207-214 (14 June 2012). DOI: **10.1038/nature11234**

The Human Microbiome Project Consortium. A framework for human microbiome research. *Nature* **486**, 215-221 (14 June 2012). DOI: **10.1038/nature11209**

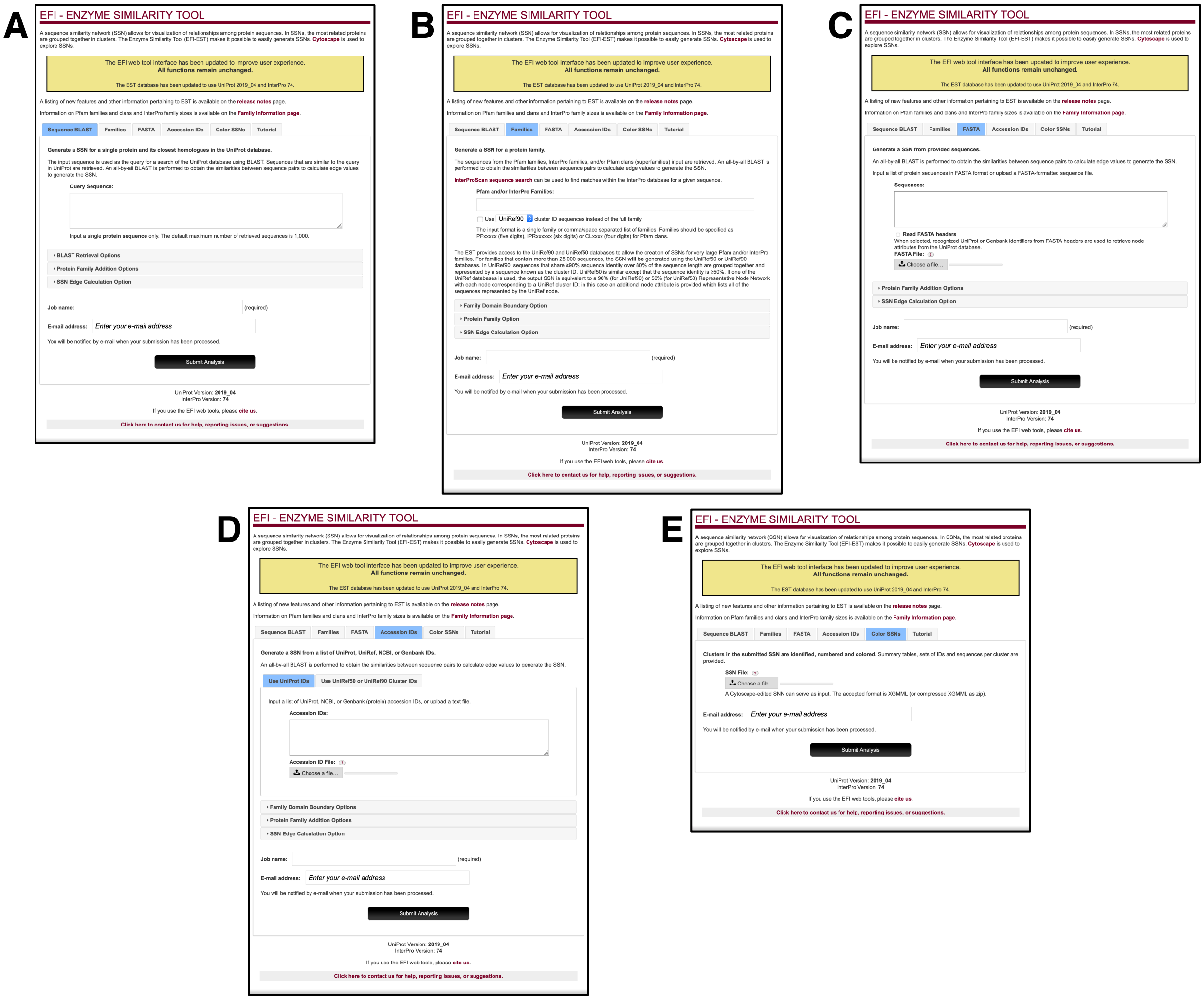
If you use the EFI web tools, please **cite us**.

Click here to contact us for help, reporting issues, or suggestions.

Carl R. Woese Institute for Genomic Biology
University of Illinois at Urbana-Champaign
1206 W. Gregory Drive Urbana, IL 61801
efi@enzymefunction.org

EFI ENZYME FUNCTION INITIATIVE
NIH National Institute of General Medical Sciences

Figure S2. **The home pages for the web tools**. **Panel A**, EFI-EST for generating sequence similarity networks (SSNs), accessible at https://efi.igb.illinois.edu/efi-est/. **Panel** B, EFI-GNT for collecting and analyzing genome context for bacterial, archaeal, and fungal proteins in SSNs, accessible at https://efi.igb.illinois.edu/efi-gnt/. **Panel C**, EFI-CGFP for mapping metagenome abundance to SSNs clusters, accessible at https://efi.igb.illinois.edu/efi-cgfp/.



EFI - ENZYME SIMILARITY TOOL

A sequence similarity network (SSN) allows for visualization of relationships among protein sequences. In SSNs, the most related proteins are grouped together in clusters. The Enzyme Similarity Tool (EFI-EST) makes it possible to easily generate SSNs. **Cytoscape** is used to explore SSNs.

The EFI web tool interface has been updated to improve user experience.
All functions remain unchanged.

The EST database has been updated to use UniProt 2019_04 and InterPro 74.

A listing of new features and other information pertaining to EST is available on the **release notes** page.

Information on Pfam families and clans and InterPro family sizes is available on the **Family Information page**.

Sequence BLAST | Families | FASTA | Accession IDs | Color SSNs | Tutorial

Generate a SSN for a single protein and its closest homologues in the UniProt database.

The input sequence is used as the query for a search of the UniProt database using BLAST. Sequences that are similar to the query in UniProt are retrieved. An all-by-all BLAST is performed to obtain the similarities between sequence pairs to calculate edge values to generate the SSN.

Query Sequence:

Input a single **protein sequence** only. The default maximum number of retrieved sequences is 1,000.

- **BLAST Retrieval Options**
- **Protein Family Addition Options**
- **SSN Edge Calculation Option**

Job name: (required)

E-mail address: *Enter your e-mail address*

You will be notified by e-mail when your submission has been processed.

UniProt Version: **2019_04**
InterPro Version: **74**

If you use the EFI web tools, please **cite us**.

Click here to contact us for help, reporting issues, or suggestions.

EFI - ENZYME SIMILARITY TOOL

A sequence similarity network (SSN) allows for visualization of relationships among protein sequences. In SSNs, the most related proteins are grouped together in clusters. The Enzyme Similarity Tool (EFI-EST) makes it possible to easily generate SSNs. **Cytoscape** is used to explore SSNs.

The EFI web tool interface has been updated to improve user experience.
All functions remain unchanged.

The EST database has been updated to use UniProt 2019_04 and InterPro 74.

A listing of new features and other information pertaining to EST is available on the **release notes** page.

Information on Pfam families and clans and InterPro family sizes is available on the **Family Information page**.

Sequence BLAST | Families | FASTA | Accession IDs | Color SSNs | Tutorial

Generate a SSN for a protein family.

The sequences from the Pfam families, InterPro families, and/or Pfam clans (superfamilies) input are retrieved. An all-by-all BLAST is performed to obtain the similarities between sequence pairs to calculate edge values to generate the SSN.

InterProScan sequence search can be used to find matches within the InterPro database for a given sequence.

Pfam and/or InterPro Families:

☐ Use UniRef90 cluster ID sequences instead of the full family

The input format is a single family or comma/space separated list of families. Families should be specified as PFxxxxx (five digits), IPRxxxxxx (six digits) or CLxxxx (four digits) for Pfam clans.

The EST provides access to the UniRef90 and UniRef50 databases to allow the creation of SSNs for very large Pfam and/or InterPro families. For families that contain more than 25,000 sequences, the SSN **will be** generated using the UniRef50 or UniRef90 databases. In UniRef90, sequences that share ≥90% sequence identity over 80% of the sequence length are grouped together and represented by a sequence known as the cluster ID. UniRef50 is similar except that the sequence identity is ≥50%. If one of the UniRef databases is used, the output SSN is equivalent to a 90% (for UniRef90) or 50% (for UniRef50) Representative Node Network with each node corresponding to a UniRef cluster ID; in this case an additional node attribute is provided which lists all of the sequences represented by the UniRef node.

- **Family Domain Boundary Option**
- **Protein Family Option**
- **SSN Edge Calculation Option**

Job name: (required)

E-mail address: *Enter your e-mail address*

You will be notified by e-mail when your submission has been processed.

UniProt Version: **2019_04**
InterPro Version: **74**

If you use the EFI web tools, please **cite us**.

Click here to contact us for help, reporting issues, or suggestions.

EFI - ENZYME SIMILARITY TOOL

A sequence similarity network (SSN) allows for visualization of relationships among protein sequences. In SSNs, the most related proteins are grouped together in clusters. The Enzyme Similarity Tool (EFI-EST) makes it possible to easily generate SSNs. **Cytoscape** is used to explore SSNs.

The EFI web tool interface has been updated to improve user experience.
All functions remain unchanged.

The EST database has been updated to use UniProt 2019_04 and InterPro 74.

A listing of new features and other information pertaining to EST is available on the **release notes** page.

Information on Pfam families and clans and InterPro family sizes is available on the **Family Information page**.

Sequence BLAST | Families | FASTA | Accession IDs | Color SSNs | Tutorial

Generate a SSN from provided sequences.

An all-by-all BLAST is performed to obtain the similarities between sequence pairs to calculate edge values to generate the SSN.

Input a list of protein sequences in FASTA format or upload a FASTA-formatted sequence file.

Sequences:

☐ **Read FASTA headers**
When selected, recognized UniProt or Genbank identifiers from FASTA headers are used to retrieve node attributes from the UniProt database.
FASTA File: ?

- **Protein Family Addition Options**
- **SSN Edge Calculation Option**

Job name: (required)

E-mail address: *Enter your e-mail address*

You will be notified by e-mail when your submission has been processed.

UniProt Version: **2019_04**
InterPro Version: **74**

If you use the EFI web tools, please **cite us**.

Click here to contact us for help, reporting issues, or suggestions.

EFI - ENZYME SIMILARITY TOOL

A sequence similarity network (SSN) allows for visualization of relationships among protein sequences. In SSNs, the most related proteins are grouped together in clusters. The Enzyme Similarity Tool (EFI-EST) makes it possible to easily generate SSNs. **Cytoscape** is used to explore SSNs.

The EFI web tool interface has been updated to improve user experience.
All functions remain unchanged.

The EST database has been updated to use UniProt 2019_04 and InterPro 74.

A listing of new features and other information pertaining to EST is available on the **release notes** page.

Information on Pfam families and clans and InterPro family sizes is available on the **Family Information page**.

Sequence BLAST | Families | FASTA | Accession IDs | Color SSNs | Tutorial

Generate a SSN from a list of UniProt, UniRef, NCBI, or Genbank IDs.

An all-by-all BLAST is performed to obtain the similarities between sequence pairs to calculate edge values to generate the SSN.

Input a list of UniProt, NCBI, or Genbank (protein) accession IDs, or upload a text file.

Accession IDs:

Accession ID File: ?

- **Family Domain Boundary Options**
- **Protein Family Addition Options**
- **SSN Edge Calculation Option**

Job name: (required)

E-mail address: *Enter your e-mail address*

You will be notified by e-mail when your submission has been processed.

UniProt Version: **2019_04**
InterPro Version: **74**

If you use the EFI web tools, please **cite us**.

Click here to contact us for help, reporting issues, or suggestions.

EFI - ENZYME SIMILARITY TOOL

A sequence similarity network (SSN) allows for visualization of relationships among protein sequences. In SSNs, the most related proteins are grouped together in clusters. The Enzyme Similarity Tool (EFI-EST) makes it possible to easily generate SSNs. **Cytoscape** is used to explore SSNs.

The EFI web tool interface has been updated to improve user experience.
All functions remain unchanged.

The EST database has been updated to use UniProt 2019_04 and InterPro 74.

A listing of new features and other information pertaining to EST is available on the **release notes** page.

Information on Pfam families and clans and InterPro family sizes is available on the **Family Information page**.

Sequence BLAST | Families | FASTA | Accession IDs | Color SSNs | Tutorial

Clusters in the submitted SSN are identified, numbered and colored. Summary tables, sets of IDs and sequences per cluster are provided.

SSN File: ?

A Cytoscape-edited SNN can serve as input. The accepted format is XGMML (or compressed XGMML as zip).

E-mail address: *Enter your e-mail address*

You will be notified by e-mail when your submission has been processed.

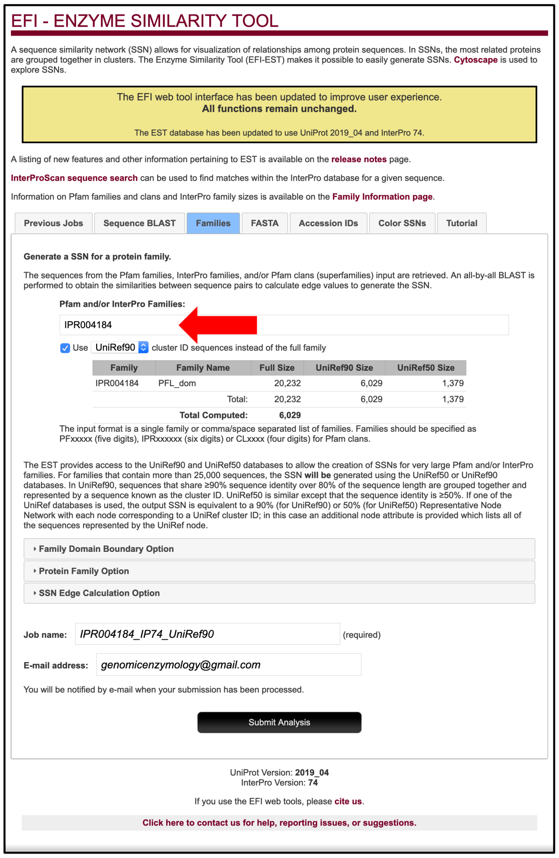
UniProt Version: **2019_04**
InterPro Version: **74**

If you use the EFI web tools, please **cite us**.

Click here to contact us for help, reporting issues, or suggestions.

Figure S3. **The EFI-EST pages for generating SSNs**. **Panel A**, Option A, a user-provided sequence is used as the query for a BLAST search of the UniProt database to collect homologues. **Panel B**, Option B, one or more user-specified protein families (Pfam, InterPro, and/or Pfam clans) is used to generate the SSN. **Panel C**, Option C, a user-provided FASTA file provides the sequences to generate the SSN. **Panel D**, Option C, a user-provided list of accession IDs (UniProt and/or NCBI) specifies the sequences used to generate the SSN. **Panel E**, Color SSN Utility, unique colors and numbers are assigned to the clusters in a user-provided SSN.

A



EFI - ENZYME SIMILARITY TOOL

A sequence similarity network (SSN) allows for visualization of relationships among protein sequences. In SSNs, the most related proteins are grouped together in clusters. The Enzyme Similarity Tool (EFI-EST) makes it possible to easily generate SSNs. **Cytoscape** is used to explore SSNs.

The EFI web tool interface has been updated to improve user experience.
All functions remain unchanged.

The EST database has been updated to use UniProt 2019_04 and InterPro 74.

A listing of new features and other information pertaining to EST is available on the **release notes** page.

InterProScan sequence search can be used to find matches within the InterPro database for a given sequence.

Information on Pfam families and clans and InterPro family sizes is available on the **Family Information page**.

Generate a SSN for a protein family.

The sequences from the Pfam families, InterPro families, and/or Pfam clans (superfamilies) input are retrieved. An all-by-all BLAST is performed to obtain the similarities between sequence pairs to calculate edge values to generate the SSN.

Pfam and/or InterPro Families:

IPR004184

☑ Use UniRef90 cluster ID sequences instead of the full family

Family	Family Name	Full Size	UniRef90 Size	UniRef50 Size
IPR004184	PFL_dom	20,232	6,029	1,379
	Total:	20,232	6,029	1,379
	Total Computed:	**6,029**		

The input format is a single family or comma/space separated list of families. Families should be specified as PFxxxxx (five digits), IPRxxxxxx (six digits) or CLxxxx (four digits) for Pfam clans.

The EST provides access to the UniRef90 and UniRef50 databases to allow the creation of SSNs for very large Pfam and/or InterPro families. For families that contain more than 25,000 sequences, the SSN **will be** generated using the UniRef50 or UniRef90 databases. In UniRef90, sequences that share ≥90% sequence identity over 80% of the sequence length are grouped together and represented by a sequence known as the cluster ID. UniRef50 is similar except that the sequence identity is ≥50%. If one of the UniRef databases is used, the output SSN is equivalent to a 90% (for UniRef90) or 50% (for UniRef50) Representative Node Network with each node corresponding to a UniRef cluster ID; in this case an additional node attribute is provided which lists all of the sequences represented by the UniRef node.

Job name: *IPR004184_IP74_UniRef90* (required)

E-mail address: *genomicenzymology@gmail.com*

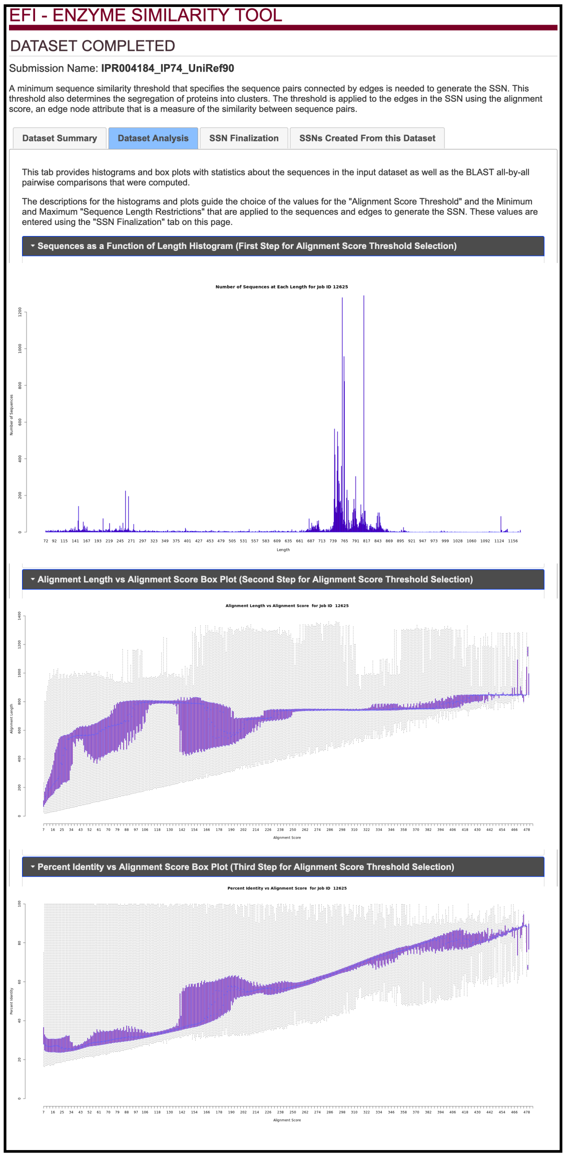
You will be notified by e-mail when your submission has been processed.

UniProt Version: **2019_04**
InterPro Version: **74**

If you use the EFI web tools, please **cite us**.

Click here to contact us for help, reporting issues, or suggestions.

B



EFI - ENZYME SIMILARITY TOOL

DATASET COMPLETED

Submission Name: **IPR004184_IP74_UniRef90**

A minimum sequence similarity threshold that specifies the sequence pairs connected by edges is needed to generate the SSN. This threshold also determines the segregation of proteins into clusters. The threshold is applied to the edges in the SSN using the alignment score, an edge node attribute that is a measure of the similarity between sequence pairs.

This tab provides histograms and box plots with statistics about the sequences in the input dataset as well as the BLAST all-by-all pairwise comparisons that were computed.

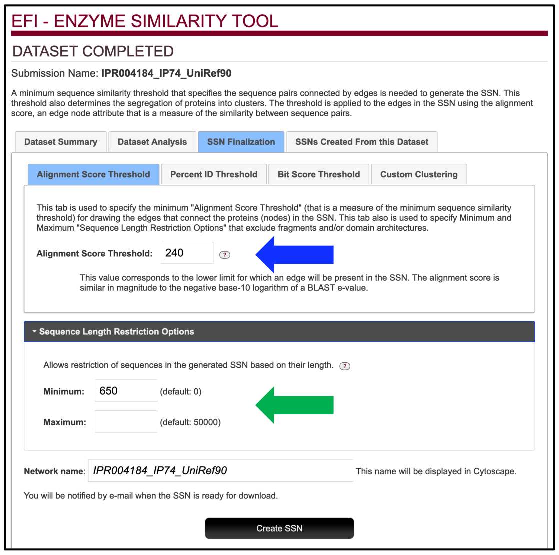
The descriptions for the histograms and plots guide the choice of the values for the "Alignment Score Threshold" and the Minimum and Maximum "Sequence Length Restrictions" that are applied to the sequences and edges to generate the SSN. These values are entered using the "SSN Finalization" tab on this page.

Sequences as a Function of Length Histogram (First Step for Alignment Score Threshold Selection)

Alignment Length vs Alignment Score Box Plot (Second Step for Alignment Score Threshold Selection)

Percent Identity vs Alignment Score Box Plot (Third Step for Alignment Score Threshold Selection)

C



EFI - ENZYME SIMILARITY TOOL

DATASET COMPLETED

Submission Name: **IPR004184_IP74_UniRef90**

A minimum sequence similarity threshold that specifies the sequence pairs connected by edges is needed to generate the SSN. This threshold also determines the segregation of proteins into clusters. The threshold is applied to the edges in the SSN using the alignment score, an edge node attribute that is a measure of the similarity between sequence pairs.

This tab is used to specify the minimum "Alignment Score Threshold" (that is a measure of the minimum sequence similarity threshold) for drawing the edges that connect the proteins (nodes) in the SSN. This tab also is used to specify Minimum and Maximum "Sequence Length Restriction Options" that exclude fragments and/or domain architectures.

Alignment Score Threshold: 240

This value corresponds to the lower limit for which an edge will be present in the SSN. The alignment score is similar in magnitude to the negative base-10 logarithm of a BLAST e-value.

Sequence Length Restriction Options

Allows restriction of sequences in the generated SSN based on their length.

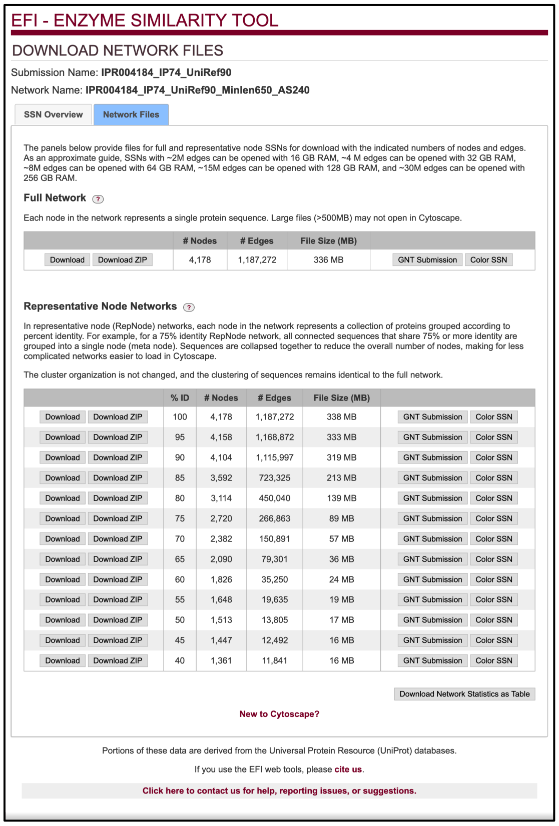
Minimum: 650 (default: 0)

Maximum: (default: 50000)

Network name: *IPR004184_IP74_UniRef90* This name will be displayed in Cytoscape.

You will be notified by e-mail when the SSN is ready for download.

D



EFI - ENZYME SIMILARITY TOOL

DOWNLOAD NETWORK FILES

Submission Name: **IPR004184_IP74_UniRef90**

Network Name: **IPR004184_IP74_UniRef90_Minlen650_AS240**

The panels below provide files for full and representative node SSNs for download with the indicated numbers of nodes and edges. As an approximate guide, SSNs with ~2M edges can be opened with 16 GB RAM, ~4 M edges can be opened with 32 GB RAM, ~8M edges can be opened with 64 GB RAM, ~15M edges can be opened with 128 GB RAM, and ~30M edges can be opened with 256 GB RAM.

Full Network

Each node in the network represents a single protein sequence. Large files (>500MB) may not open in Cytoscape.

	# Nodes	# Edges	File Size (MB)	
Download Download ZIP	4,178	1,187,272	336 MB	GNT Submission Color SSN

Representative Node Networks

In representative node (RepNode) networks, each node in the network represents a collection of proteins grouped according to percent identity. For example, for a 75% identity RepNode network, all connected sequences that share 75% or more identity are grouped into a single node (meta node). Sequences are collapsed together to reduce the overall number of nodes, making for less complicated networks easier to load in Cytoscape.

The cluster organization is not changed, and the clustering of sequences remains identical to the full network.

	% ID	# Nodes	# Edges	File Size (MB)	
Download Download ZIP	100	4,178	1,187,272	338 MB	GNT Submission Color SSN
Download Download ZIP	95	4,158	1,168,872	333 MB	GNT Submission Color SSN
Download Download ZIP	90	4,104	1,115,997	319 MB	GNT Submission Color SSN
Download Download ZIP	85	3,592	723,325	213 MB	GNT Submission Color SSN
Download Download ZIP	80	3,114	450,040	139 MB	GNT Submission Color SSN
Download Download ZIP	75	2,720	266,863	89 MB	GNT Submission Color SSN
Download Download ZIP	70	2,382	150,891	57 MB	GNT Submission Color SSN
Download Download ZIP	65	2,090	79,301	36 MB	GNT Submission Color SSN
Download Download ZIP	60	1,826	35,250	24 MB	GNT Submission Color SSN
Download Download ZIP	55	1,648	19,635	19 MB	GNT Submission Color SSN
Download Download ZIP	50	1,513	13,805	17 MB	GNT Submission Color SSN
Download Download ZIP	45	1,447	12,492	16 MB	GNT Submission Color SSN
Download Download ZIP	40	1,361	11,841	16 MB	GNT Submission Color SSN

New to Cytoscape?

Portions of these data are derived from the Universal Protein Resource (UniProt) databases.

If you use the EFI web tools, please **cite us**.

Click here to contact us for help, reporting issues, or suggestions.

Figure S4. **The sequence of steps in generating an SSN with EFI-EST, using the glycyl radical enzyme superfamily (IPR004184) as an example**. **Panel A**, the family identifier is specified, and the user chooses the database (UniProt, UniRef90, or UniRef50) for generating the SSN (red arrow). **Panel B**, the "Dataset Completed" page that provides histograms and boxplots that are used for selecting in minimum alignment score threshold for generating the initial SSN; see text for how the histograms and plots are used to select the alignment score. **Panel C**, the SSN "Finalization" tab for entering the "Alignment Score Threshold" (blue arrow) and "Minimum"/"Maximum" length filters (green arrow). **Panel D**, the "Download Network Files" page that provides access to the full and representative node SSNs. The details are described in the text.

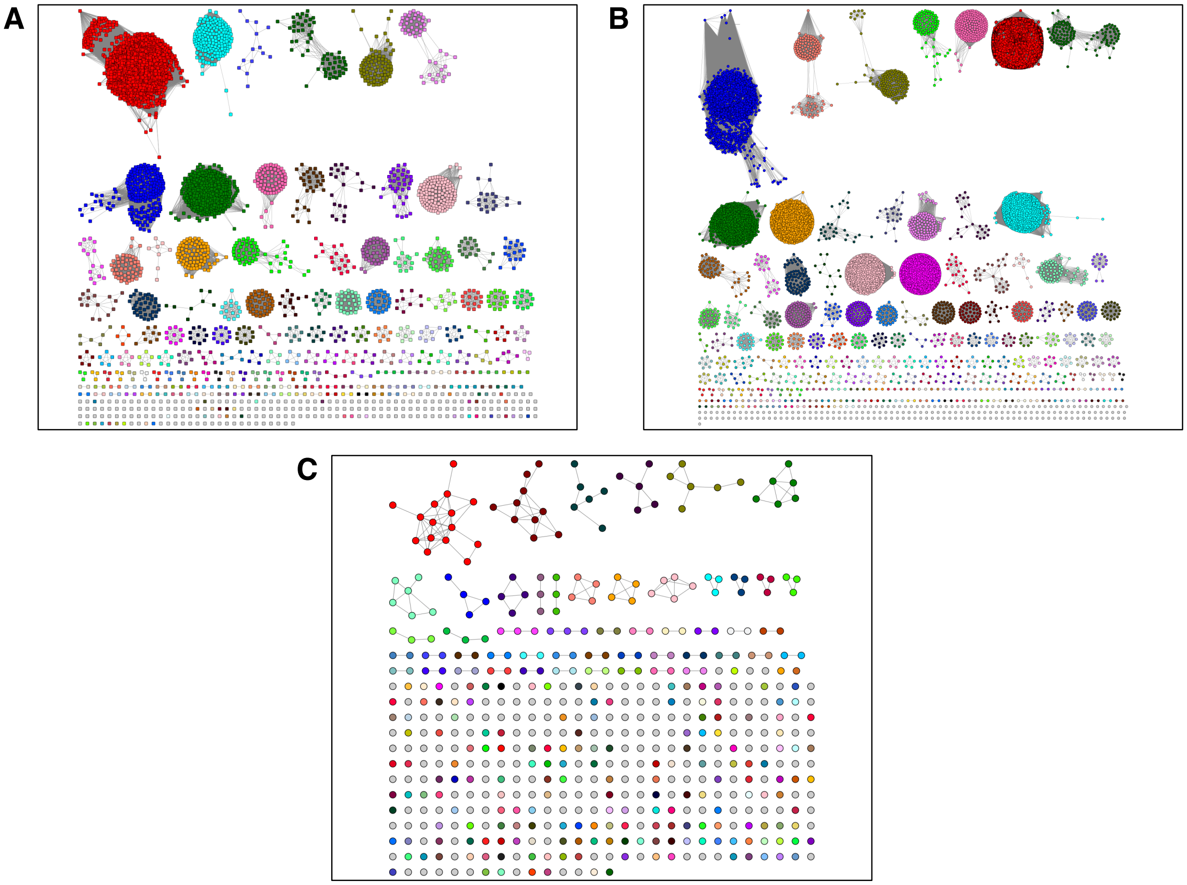
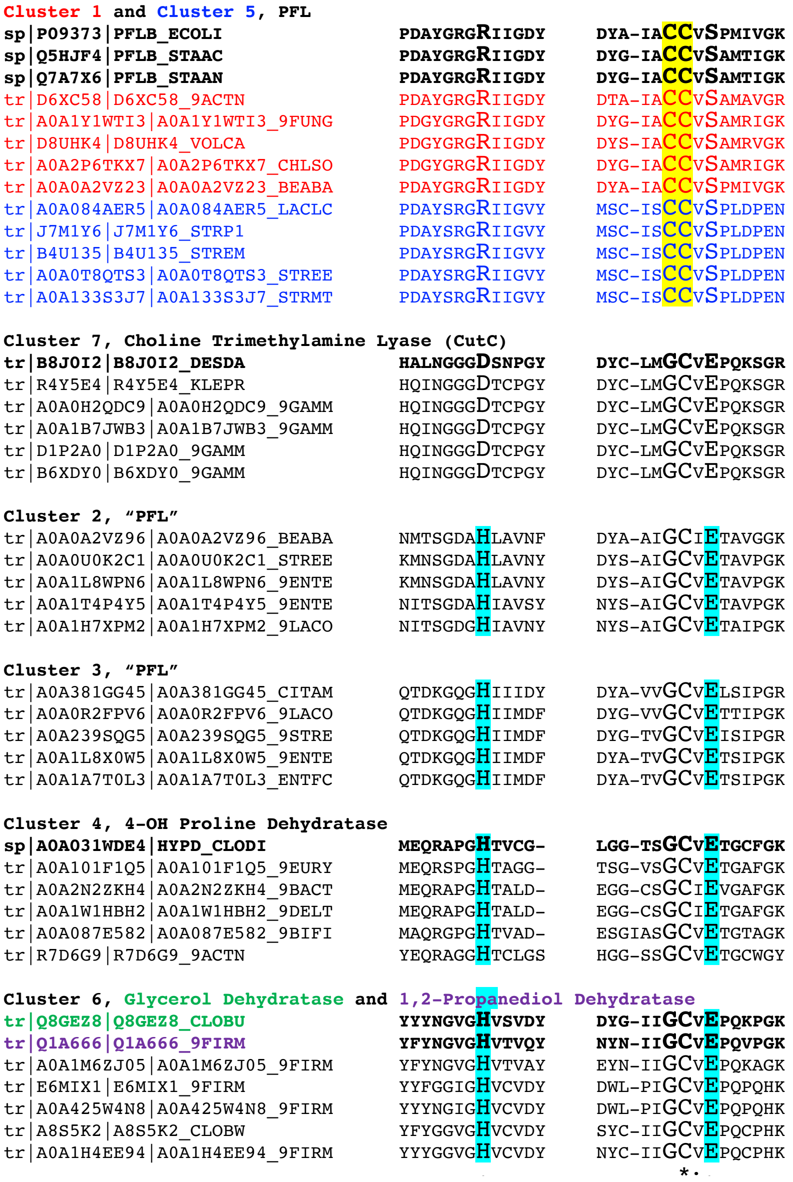
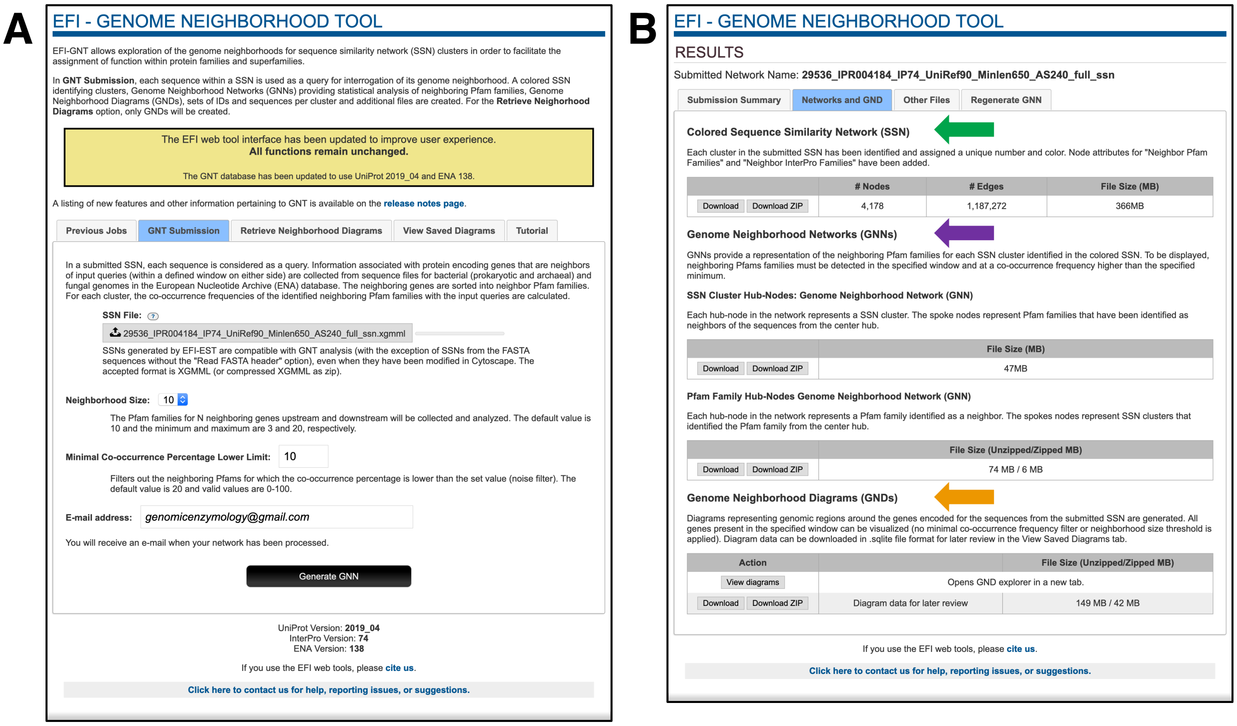


Figure S5. **A comparison of the SSNs for IPR004184 generated using the UniProt (Panel A), UniRef90 (Panel B), and UniRef50 (Panel C) databases**. The SSNs were colored with the Color SSNs Utility that assigns a unique color and number to the SSN clusters, with the numbers assigned in order of decreasing number of UniProt IDs in the clusters. Some singleton nodes are colored in the UniRef90 and UniRef50 SSNs—these contain multiple UniProt IDs so are considered clusters. The UniProt SSN (Panel A) is the highest resolution; the UniRef90 SSN (Panel B) provides similar resolution but the file size is significantly less than that for the UniProt SSN. The UniRef50 SSN (Panel B) should be used only for the largest families, with subsequent generation of daughter SSNs for sequences in individual clusters allowing higher resolution analyses. In this example, the UniProt and UniRef90 SSNs contain the same number of clusters.



```
Cluster 1 and Cluster 5, PFL
sp|P09373|PFLB_ECOLI                 PDAYGRGRIIGDY      DYA-IACCvSPMIVGK
sp|Q5HJF4|PFLB_STAAC                 PDAYGRGRIIGDY      DYG-IACCvSAMTIGK
sp|Q7A7X6|PFLB_STAAN                 PDAYGRGRIIGDY      DYG-IACCvSAMTIGK
tr|D6XC58|D6XC58_9ACTN               PDAYGRGRIIGDY      DTA-IACCvSAMAVGR
tr|A0A1Y1WTI3|A0A1Y1WTI3_9FUNG       PDGYGRGRIIGDY      DYG-IACCvSAMRIGK
tr|D8UHK4|D8UHK4_VOLCA               PDGYGRGRIIGDY      DYS-IACCvSAMRVGK
tr|A0A2P6TKX7|A0A2P6TKX7_CHLSO       PDGYGRGRIIGDY      DYG-IACCvSAMRIGK
tr|A0A0A2VZ23|A0A0A2VZ23_BEABA       PDAYGRGRIIGDY      DYA-IACCvSPMIVGK
tr|A0A084AER5|A0A084AER5_LACLC       PDAYSRGRIIGVY      MSC-ISCCvSPLDPEN
tr|J7M1Y6|J7M1Y6_STRP1               PDAYSRGRIIGVY      MSC-ISCCvSPLDPEN
tr|B4U135|B4U135_STREM               PDAYSRGRIIGVY      MSC-ISCCvSPLDPEN
tr|A0A0T8QTS3|A0A0T8QTS3_STREE       PDAYSRGRIIGVY      MSC-ISCCvSPLDPEN
tr|A0A133S3J7|A0A133S3J7_STRMT       PDAYSRGRIIGVY      MSC-ISCCvSPLDPEN

Cluster 7, Choline Trimethylamine Lyase (CutC)
tr|B8J0I2|B8J0I2_DESDA               HALNGGGDSNPGY      DYC-LMGCvEPQKSGR
tr|R4Y5E4|R4Y5E4_KLEPR               HQINGGGDTCPGY      DYC-LMGCvEPQKSGR
tr|A0A0H2QDC9|A0A0H2QDC9_9GAMM       HQINGGGDTCPGY      DYC-LMGCvEPQKSGR
tr|A0A1B7JWB3|A0A1B7JWB3_9GAMM       HQINGGGDTCPGY      DYC-LMGCvEPQKSGR
tr|D1P2A0|D1P2A0_9GAMM               HQINGGGDTCPGY      DYC-LMGCvEPQKSGR
tr|B6XDY0|B6XDY0_9GAMM               HQINGGGDTCPGY      DYC-LMGCvEPQKSGR

Cluster 2, “PFL”
tr|A0A0A2VZ96|A0A0A2VZ96_BEABA       NMTSGDAHLAVNF      DYA-AIGCIETAVGGK
tr|A0A0U0K2C1|A0A0U0K2C1_STREE       KMNSGDAHLAVNY      DYS-AIGCvETAVPGK
tr|A0A1L8WPN6|A0A1L8WPN6_9ENTE       KMNSGDAHLAVNY      DYS-AIGCvETAVPGK
tr|A0A1T4P4Y5|A0A1T4P4Y5_9ENTE       NITSGDAHIAVSY      NYS-AIGCvETAVPGK
tr|A0A1H7XPM2|A0A1H7XPM2_9LACO       NITSGDGHIAVNY      NYS-AIGCvETAIPGK

Cluster 3, “PFL”
tr|A0A381GG45|A0A381GG45_CITAM       QTDKGQGHIIIDY      DYA-VVGCvELSIPGR
tr|A0A0R2FPV6|A0A0R2FPV6_9LACO       QTDKGQGHIIMDF      DYG-VVGCvETTIPGK
tr|A0A239SQG5|A0A239SQG5_9STRE       QTDKGQGHIIMDF      DYG-TVGCvEISIPGR
tr|A0A1L8X0W5|A0A1L8X0W5_9ENTE       QTDKGQGHIIMDF      DYA-TVGCvETSIPGK
tr|A0A1A7T0L3|A0A1A7T0L3_ENTFC       QTDKGQGHIIMDF      DYA-TVGCvETSIPGK

Cluster 4, 4-OH Proline Dehydratase
sp|A0A031WDE4|HYPD_CLODI             MEQRAPGHTVCG-      LGG-TSGCvETGCFGK
tr|A0A101F1Q5|A0A101F1Q5_9EURY       MEQRSPGHTAGG-      TSG-VSGCvETGAFGK
tr|A0A2N2ZKH4|A0A2N2ZKH4_9BACT       MEQRAPGHTALD-      EGG-CSGCIEVGAFGK
tr|A0A1W1HBH2|A0A1W1HBH2_9DELT       MEQRAPGHTALD-      EGG-CSGCIETGAFGK
tr|A0A087E582|A0A087E582_9BIFI       MAQRGPGHTVAD-      ESGIASGCvETGTAGK
tr|R7D6G9|R7D6G9_9ACTN               YEQRAGGHTCLGS      HGG-SSGCvETGCWGY

Cluster 6, Glycerol Dehydratase and 1,2-Propanediol Dehydratase
tr|Q8GEZ8|Q8GEZ8_CLOBU               YYYNGVGHVSVDY      DYG-IIGCvEPQKPGK
tr|Q1A666|Q1A666_9FIRM               YFYNGVGHVTVQY      NYN-IIGCvEPQVPGK
tr|A0A1M6ZJ05|A0A1M6ZJ05_9FIRM       YFYNGVGHVTVAY      EYN-IIGCvEPQKAGK
tr|E6MIX1|E6MIX1_9FIRM               YYFGGIGHVCVDY      DWL-PIGCvEPQPQHK
tr|A0A425W4N8|A0A425W4N8_9FIRM       YYYNGIGHVCVDY      DWL-PIGCvEPQPQHK
tr|A8S5K2|A8S5K2_CLOBW               YFYGGVGHVCVDY      SYC-IIGCvEPQCPHK
tr|A0A1H4EE94|A0A1H4EE94_9FIRM       YYYGGVGHVCVDY      NYC-IIGCvEPQCPHK
                                                                *.
```

Figure S6. Partial multiple sequence alignments (MSAs) for the largest seven clusters in SSN_{240} for IPR004184. These first region includes a conserved His in the active sites of characterized dehydratases; the second region includes the conserved Cys-Cys motif that is characteristic of the PFL function. The bold sequence entries are either SwissProt-curated (“sp”; PFL and 4-OH Pro dehydratase) or from the literature (choline trimethylamine lyase, glycerol dehydratase, and 1,2-propanediol dehydratase). Based on their sequences, clusters 2 and 3, that have SwissProt-curated PFL functions (inferred from homology), are predicted to be dehydratases, consistent with their colocation with 4-OH Pro dehydratase, glycerol dehydratase, and 1,2-propanediol dehydratase in SSN$_{185}$. Choline trimethylamine lyase is neither PFL nor a dehydratase so it is lacking the conserved motifs for these functions.



EFI - GENOME NEIGHBORHOOD TOOL

EFI-GNT allows exploration of the genome neighborhoods for sequence similarity network (SSN) clusters in order to facilitate the assignment of function within protein families and superfamilies.

In **GNT Submission**, each sequence within a SSN is used as a query for interrogation of its genome neighborhood. A colored SSN identifying clusters, Genome Neighborhood Networks (GNNs) providing statistical analysis of neighboring Pfam families, Genome Neighborhood Diagrams (GNDs), sets of IDs and sequences per cluster and additional files are created. For the **Retrieve Neighborhood Diagrams** option, only GNDs will be created.

The EFI web tool interface has been updated to improve user experience.
All functions remain unchanged.

The GNT database has been updated to use UniProt 2019_04 and ENA 138.

A listing of new features and other information pertaining to GNT is available on the **release notes page**.

In a submitted SSN, each sequence is considered as a query. Information associated with protein encoding genes that are neighbors of input queries (within a defined window on either side) are collected from sequence files for bacterial (prokaryotic and archaeal) and fungal genomes in the European Nucleotide Archive (ENA) database. The neighboring genes are sorted into neighbor Pfam families. For each cluster, the co-occurrence frequencies of the identified neighboring Pfam families with the input queries are calculated.

SSN File: 29536_IPR004184_IP74_UniRef90_Minlen650_AS240_full_ssn.xgmml

SSNs generated by EFI-EST are compatible with GNT analysis (with the exception of SSNs from the FASTA sequences without the "Read FASTA header" option), even when they have been modified in Cytoscape. The accepted format is XGMML (or compressed XGMML as zip).

Neighborhood Size: 10

The Pfam families for N neighboring genes upstream and downstream will be collected and analyzed. The default value is 10 and the minimum and maximum are 3 and 20, respectively.

Minimal Co-occurrence Percentage Lower Limit: 10

Filters out the neighboring Pfams for which the co-occurrence percentage is lower than the set value (noise filter). The default value is 20 and valid values are 0-100.

E-mail address: *genomicenzymology@gmail.com*

You will receive an e-mail when your network has been processed.

UniProt Version: **2019_04**
InterPro Version: **74**
ENA Version: **138**

If you use the EFI web tools, please **cite us**.

Click here to contact us for help, reporting issues, or suggestions.

EFI - GENOME NEIGHBORHOOD TOOL

RESULTS

Submitted Network Name: **29536_IPR004184_IP74_UniRef90_Minlen650_AS240_full_ssn**

Colored Sequence Similarity Network (SSN)

Each cluster in the submitted SSN has been identified and assigned a unique number and color. Node attributes for "Neighbor Pfam Families" and "Neighbor InterPro Families" have been added.

	# Nodes	# Edges	File Size (MB)
Download Download ZIP	4,178	1,187,272	366MB

Genome Neighborhood Networks (GNNs)

GNNs provide a representation of the neighboring Pfam families for each SSN cluster identified in the colored SSN. To be displayed, neighboring Pfams families must be detected in the specified window and at a co-occurrence frequency higher than the specified minimum.

SSN Cluster Hub-Nodes: Genome Neighborhood Network (GNN)

Each hub-node in the network represents a SSN cluster. The spoke nodes represent Pfam families that have been identified as neighbors of the sequences from the center hub.

	File Size (MB)
Download Download ZIP	47MB

Pfam Family Hub-Nodes Genome Neighborhood Network (GNN)

Each hub-node in the network represents a Pfam family identified as a neighbor. The spokes nodes represent SSN clusters that identified the Pfam family from the center hub.

	File Size (Unzipped/Zipped MB)
Download Download ZIP	74 MB / 6 MB

Genome Neighborhood Diagrams (GNDs)

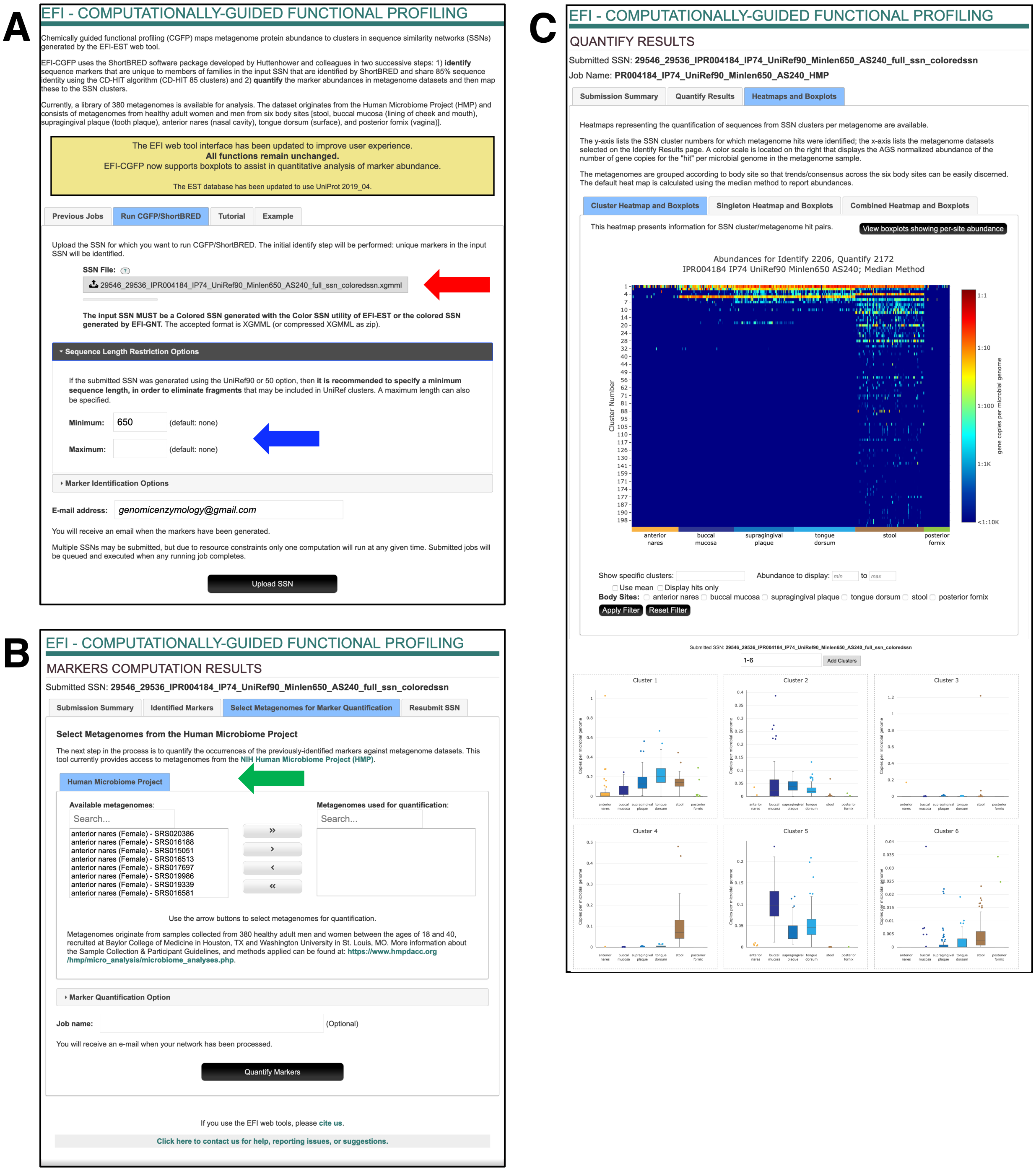
Diagrams representing genomic regions around the genes encoded for the sequences from the submitted SSN are generated. All genes present in the specified window can be visualized (no minimal co-occurrence frequency filter or neighborhood size threshold is applied). Diagram data can be downloaded in .sqlite file format for later review in the View Saved Diagrams tab.

Action		File Size (Unzipped/Zipped MB)
View diagrams	Opens GND explorer in a new tab.	
Download Download ZIP	Diagram data for later review	149 MB / 42 MB

If you use the EFI web tools, please **cite us**.

Click here to contact us for help, reporting issues, or suggestions.

Figure S7. **The sequence of steps in generating GNNs and GNDs with EFI-GNT, using the glycyl radical enzyme superfamily (IPR004184) as an example**. **Panel A**, the SSN is uploaded, with the user specifying the neighborhood size (±10 orfs is the default; red arrow) and query-neighborhood family co-occurrence frequency (20% is the default; the example in the text uses 10%; blue arrow) for generating GNNs. The scripts collect genome neighbors in a ±20 orf window, but the user-specified value is used to generate the GNNs; the GNNs can be recalculated using different neighborhood sizes and co-occurrence frequencies. **Panel B**, the "Results" page that provides the colored SSN (with unique cluster colors and numbers and "Neighbor Pfam Families" and "Neighbor InterPro Families" node attributes; green arrow), the GNNs (SSN cluster-hub nodes with Pfam family spoke nodes, with the SSN cluster-hub nodes colored/numbered; and Pfam family-hub nodes with SSN cluster spoke nodes, with the SSN cluster-spoke nodes colored/numbered; magenta arrow); and access to the GND viewer (orange arrow).



EFI - COMPUTATIONALLY-GUIDED FUNCTIONAL PROFILING

Chemically guided functional profiling (CGFP) maps metagenome protein abundance to clusters in sequence similarity networks (SSNs) generated by the EFI-EST web tool.

EFI-CGFP uses the ShortBRED software package developed by Huttenhower and colleagues in two successive steps: 1) **identify** sequence markers that are unique to members of families in the input SSN that are identified by ShortBRED and share 85% sequence identity using the CD-HIT algorithm (CD-HIT 85 clusters) and 2) **quantify** the marker abundances in metagenome datasets and then map these to the SSN clusters.

Currently, a library of 380 metagenomes is available for analysis. The dataset originates from the Human Microbiome Project (HMP) and consists of metagenomes from healthy adult women and men from six body sites [stool, buccal mucosa (lining of cheek and mouth), supragingival plaque (tooth plaque), anterior nares (nasal cavity), tongue dorsum (surface), and posterior fornix (vagina)].

The EFI web tool interface has been updated to improve user experience.
All functions remain unchanged.
EFI-CGFP now supports boxplots to assist in quantitative analysis of marker abundance.

The EST database has been updated to use UniProt 2019_04.

Upload the SSN for which you want to run CGFP/ShortBRED. The initial identify step will be performed: unique markers in the input SSN will be identified.

SSN File: 29546_29536_IPR004184_IP74_UniRef90_Minlen650_AS240_full_ssn_coloredssn.xgmml

The input SSN MUST be a Colored SSN generated with the Color SSN utility of EFI-EST or the colored SSN generated by EFI-GNT. The accepted format is XGMML (or compressed XGMML as zip).

Sequence Length Restriction Options

If the submitted SSN was generated using the UniRef90 or 50 option, then **it is recommended to specify a minimum sequence length, in order to eliminate fragments** that may be included in UniRef clusters. A maximum length can also be specified.

Minimum: 650 (default: none)

Maximum: (default: none)

Marker Identification Options

E-mail address: *genomicenzymology@gmail.com*

You will receive an email when the markers have been generated.

Multiple SSNs may be submitted, but due to resource constraints only one computation will run at any given time. Submitted jobs will be queued and executed when any running job completes.

EFI - COMPUTATIONALLY-GUIDED FUNCTIONAL PROFILING

MARKERS COMPUTATION RESULTS

Submitted SSN: **29546_29536_IPR004184_IP74_UniRef90_Minlen650_AS240_full_ssn_coloredssn**

Select Metagenomes from the Human Microbiome Project

The next step in the process is to quantify the occurrences of the previously-identified markers against metagenome datasets. This tool currently provides access to metagenomes from the **NIH Human Microbiome Project (HMP)**.

Available metagenomes:

anterior nares (Female) - SRS020386
anterior nares (Female) - SRS016188
anterior nares (Female) - SRS015051
anterior nares (Female) - SRS016513
anterior nares (Female) - SRS017697
anterior nares (Female) - SRS019986
anterior nares (Female) - SRS019339
anterior nares (Female) - SRS016581

Metagenomes used for quantification:

Use the arrow buttons to select metagenomes for quantification.

Metagenomes originate from samples collected from 380 healthy adult men and women between the ages of 18 and 40, recruited at Baylor College of Medicine in Houston, TX and Washington University in St. Louis, MO. More information about the Sample Collection & Participant Guidelines, and methods applied can be found at: **https://www.hmpdacc.org/hmp/micro_analysis/microbiome_analyses.php**.

Marker Quantification Option

Job name: (Optional)

You will receive an e-mail when your network has been processed.

If you use the EFI web tools, please **cite us**.

Click here to contact us for help, reporting issues, or suggestions.

EFI - COMPUTATIONALLY-GUIDED FUNCTIONAL PROFILING

QUANTIFY RESULTS

Submitted SSN: **29546_29536_IPR004184_IP74_UniRef90_Minlen650_AS240_full_ssn_coloredssn**

Job Name: **PR004184_IP74_UniRef90_Minlen650_AS240_HMP**

Heatmaps representing the quantification of sequences from SSN clusters per metagenome are available.

The y-axis lists the SSN cluster numbers for which metagenome hits were identified; the x-axis lists the metagenome datasets selected on the Identify Results page. A color scale is located on the right that displays the AGS normalized abundance of the number of gene copies for the "hit" per microbial genome in the metagenome sample.

The metagenomes are grouped according to body site so that trends/consensus across the six body sites can be easily discerned. The default heat map is calculated using the median method to report abundances.

Cluster Heatmap and Boxplots
Singleton Heatmap and Boxplots
Combined Heatmap and Boxplots

This heatmap presents information for SSN cluster/metagenome hit pairs.

Show specific clusters: Abundance to display: min to max

□ Use mean □ Display hits only

Body Sites: □ anterior nares □ buccal mucosa □ supragingival plaque □ tongue dorsum □ stool □ posterior fornix

Submitted SSN: **29546_29536_IPR004184_IP74_UniRef90_Minlen650_AS240_full_ssn_coloredssn**

1-6

Figure S8. The sequence of steps in generating CGFP heatmaps and boxplots with EFI-CGFP, using the glycyl radical enzyme superfamily (IPR004184) as an example. Panel A, a colored SSN is uploaded (with unique cluster and singleton numbers to enable mapping of

metagenome abundance to clusters and singletons; red arrow); minimum and maximum length filters are recommended to ensure that the consensus sequences for ShortBRED families used for marker identification are not biased by the presence of fragments (blue arrow). **Panel B**, the “Markers Computation Results” page that allows the user to choose metagenomes for abundance mapping from a library of 380 metagenomes from six body sites from healthy individuals (green arrow). **Panel C**, the “Quantify Results” page that provides (top) heatmaps for metagenome abundance for clusters and singletons in the input SSN and (bottom) boxplots showing quantitative analyses of the metagenome abundances for selected clusters.

Table S1. Node Attributes for SSNs Generated by EFI-EST

Node Attribute	Options A, B, C with FASTA header reading, D
Name	Full network - UniProt or UniRef ID; Rep Node network - UniProt or UniRef ID for the longest sequence in the representative node (seed sequence for CD-Hit). For domain SSNs, ID:N-terminus:C-terminus
Shared name	Full network - UniProt or UniRef ID; Rep Node network - UniProt or UniRef ID for the longest sequence in the representative node (seed sequence for CD-Hit). For domain SSNs, ID:N-terminus:C-terminus
UniRef90 Cluster Size	Number of UniProt IDs in UniRef90 cluster
UniRef90 Cluster IDs	List of UniProt IDs in the UniRef90 cluster
UniRef50 Cluster Size	Number of UniProt IDs in UniRef50 cluster
UniRef50 Cluster IDs	List of UniProt IDs in the UniRef50 cluster
Number of IDs in Rep Node[1]	Number of UniProt IDs in the representative node
List of IDs in Rep Node[1]	List of UniProt IDs in the representative node
Sequence Source	Option A, "INPUT" if input sequence, "BLASTHIT" if identified in BLAST, "FAMILY" if from user-specified user-specified Pfam/InterPro family, "USER+BLASTHIT" if from BLAST and family
	Options B, C, and D, "USER" if from user-supplied file, "FAMILY" if from user-specified Pfam/InterPro family, "USER+FAMILY" if from both
Query IDs	Options C and D, Input Query ID(s) that identified a UniProt match in the idmapping file
Other IDs	Option C, headers for FASTA sequences that could not identify a UniProt match in the idmapping file
User IDs in Cluster	Options A, B, and C with UniRef family added and/or rep node SSNs, UniProt IDs for BLASTHITs or user-supplied sequences in metanode
Organism	organism genus/genera and species, from UniProt taxonomy.xml
Taxonomy ID	NCBI taxonomy identifier(s), from UniProt
UniProt Annotation Status	SwissProt - manually annotated; TrEMBL - automatically annotated; from UniProt
Description	protein name(s)/annotation(s), from UniProtKB
SwissProt Description	protein name(s)/annotation(s), from UniProtKB for SwissProt reviewed entries
Sequence Length	number(s) of amino acid residues, from UniProt
Cluster ID Sequence Length	Sequence length for Cluster ID in UniRef SSNs ("most informative" sequence in cluster, as designated by UniProt)
Gene name	gene name(s)
NCBI IDs	RefSeq/GenBank IDs and GI numbers, from UniProt idmapping
Superkingdom	domain of life of the organism, from UniProt taxonomy.xml
Kingdom	kingdom of the organism, from UniProt taxonomy.xml

Phylum	Phylogenetic phylum of the organism, from UniProt taxonomy.xml
Class	Phylogenetic class of the organism, from UniProt taxonomy.xml
Order	Phylogenetic order of the organism, from UniProt taxonomy.xml
Family	Phylogenetic family of the organism, from UniProt taxonomy.xml
Genus	Phylogenetic genus of the organism, from UniProt taxonomy.xml
Species	Phylogenetic species of the organism, from UniProt taxonomy.xml
EC	EC number, from UniProt
PFAM	Pfam family, from UniProt
PDB	Protein Data Bank entry, from UniProt
InterPro (Domain)	InterPro domain(s), from InterPro
InterPro (Family)	InterPro family(ies), from InterPro
InterPro (Homologous Superfamily)	InterPro homologous superfamily(ies), from InterPro
InterPro (Other)	Other InterPro classes (repeat, site), from InterPro
BRENDA ID	BRENDA Database ID, from UniProt
CAZY Name	Carbohydrate-Active enZYmes (CAZy) family name(s), from UniProt
GO Term	Gene Ontology classification(s), from UniProt
KEGG ID	KEGG Database ID, from UniProt
PATRIC ID	PATRIC Database ID, from UniProt
STRING ID	STRING Database ID, from UniProt
HMP Body Site	location(s) of organism(s) in/on the body, if human microbiome organism, spreadsheet from HMP
HMP Oxygen	oxygen requirement(s), if human microbiome organism, from HMP
P01 gDNA	availability of gDNA(s) at EFI Protein Core, custom
Sequence	Option C, Sequence from UniProt database if ID can be located

Node Attribute	**Option C without FASTA header reading**
Name	zzznnn, where nnn = number of the sequence in FASTA file
Shared Name	zzznnn, where nnn = number of the sequence in FASTA file
Description	FASTA Header
Sequence Length	Length of sequence in FASTA entry
Sequence	Sequence from FASTA entry

Additional Node Attributes	**Colored SSN (from Colored SSNs utility)**
Cluster Number	Number assigned to cluster, in order of decreasing number of sequences in the clusters
Cluster Sequence Count	Number of sequences in the cluster
Node.fillColor	Unique color assigned to cluster, in hexadecimal
Singleton Number	Number assigned to singleton

Table S2. Formats for UniProt, NCBI and PDB IDs; FASTA Headers for Option C

A. Formats for UniProt IDs, NCBI IDs, and PDB IDs

UniProt IDs

UniProtKB ID is 6 or 10 alphanumerical characters in the following formats:

1	2	3	4	5	6	7	8	9	10
[O,P,Q]	[0-9]	[A-Z,0-9]	[A-Z,0-9]	[A-Z,0-9]	[0-9]				
[A-N,R-Z]	[0-9]	[A-Z]	[A-Z,0-9]	[A-Z,0-9]	[0-9]				
[A-N,R-Z]	[0-9]	[A-Z]	[A-Z,0-9]	[A-Z,0-9]	[0-9]	[A-Z]	[A-Z,0-9]	[A-Z,0-9]	[0-9]

For example:

```
P11444
T2HDW6
A0A0A7PVN6
```

NCBI RefSeq IDs

An NCBI RefSeq ID is 2 letters followed by an underscore followed by a series of digits, a period, and one or more digits for the sequence version number, e.g.,

```
WP_016501748.1
NP_708575.1
YP_002409124.1
```

NCBI UniProt/Swiss-Prot IDs

An NCBI UniProt/Swiss-Prot ID is the UniProt ID followed by a period and one or more digits for the sequence version number, e.g.,

```
Q31XL1.1
B7LEJ8.1
C4ZZT2.1
```

NCBI GenBank IDs

The format for NCBI GenBank IDs is 3 letters followed by five digits, a period, and one or more digits for the sequence version number, e.g.,

```
BAN56663.1
AAC15504.1
BAM38409.1
```

PDB IDs

The format for PDB IDs is one digit followed by two letters and a digit/letter:

```
1MDL
1MRA
3UXL
```

NCBI GI Numbers

An NCBI GI number (now retired) is a series of digits.

B. Formats for FASTA headers for Option C

UniProt (TrEMBL and SwissProt, respectively; from UniProt BLAST)

>tr|**R9RJF1**|R9RJF1_PSEAI Mandelate racemase OS=Pseudomonas aeruginosa PE=4 SV=1

>sp|**P11444**|MANR_PSEPU Mandelate racemase OS=Pseudomonas putida GN=mdlA PE=1 SV=1

NCBI RefSeq (from NCBI BLAST)

>**WP_016501748.1** mandelate racemase [Pseudomonas putida]

NCBI UniProt/Swiss-Prot ID (from NCBI BLAST)

>**Q0TE80.1** RecName: Full=Enolase; AltName: Full=2-phospho-D-glycerate hydro-lyase; AltName: Full=2-phosphoglycerate dehydratase

NCBI GenBank ID (from NCBI BLAST)

>**AAA25887.1** mandelate racemase (EC 5.1.2.2) [Pseudomonas putida]

NCBI PDB ID (from NCBI BLAST)

>pdb|**1MDR**|A Chain A, The Role Of Lysine 166 In The Mechanism Of Mandelate Racemase From Pseudomonas Putida: Mechanistic And Crystallographic Evidence For Stereospecific Alkylation By (r)-alpha-phenylglycidate

NCBI GI Number (from NCBI BLAST; now retired)

>gi|**347012980**| 4-O-methyl-glucuronoyl methylesterase [Myceliophthora thermophila ATCC 42464]

Option C also accepts FASTA headers in which the IDs (formats described in Option D) immediately follow the ">" symbol, e.g., the following headers abbreviated from those shown above:

UniProt

>**R9RJF1**

>**P11444**

NCBI RefSeq

>**WP_016501748.1**

NCBI UniProt/Swiss-Prot ID

>**Q0TE80.1**

NCBI GenBank ID

>**AAA25887.1**

NCBI PDB ID

>**1MDR**

NCBI GI Number (now retired)

>**347012980**

Table S3. Node Attributes for Colored SSNs Generated by EFI-GNT

Node Attribute	Options A, B, C with FASTA header reading, D
Name	Full network - UniProt or UniRef ID; Rep Node network - UniProt or UniRef ID for the longest sequence in the representative node (seed sequence for CD-Hit). For domain SSNs, ID:N-terminus:C-terminus
Shared name	Full network - UniProt or UniRef ID; Rep Node network - UniProt or UniRef ID for the longest sequence in the representative node (seed sequence for CD-Hit). For domain SSNs, ID:N-terminus:C-terminus
UniRef90 Cluster Size	Number of UniProt IDs in UniRef90 cluster
UniRef90 Cluster IDs	List of UniProt IDs in the UniRef90 cluster
UniRef50 Cluster Size	Number of UniProt IDs in UniRef50 cluster
UniRef50 Cluster IDs	List of UniProt IDs in the UniRef50 cluster
Number of IDs in Rep Node	Number of UniProt IDs in the representative node
List of IDs in Rep Node	List of UniProt IDs in the representative node
Sequence Source	Option A, “INPUT” if input sequence, “BLASTHIT” if identified in BLAST, “FAMILY” if from user-specified user-specified Pfam/InterPro family, “USER+BLASTHIT” if from BLAST and family
	Options B, C, and D, “USER” if from user-supplied file, “FAMILY” if from user-specified Pfam/InterPro family, “USER+FAMILY” if from both
Query IDs	Options C and D, Input Query ID(s) that identified a UniProt match in the idmapping file
Other IDs	Option C, headers for FASTA sequences that could not identify a UniProt match in the idmapping file
User IDs in Cluster	Options A, B, and C with UniRef family added and/or rep node SSNs, UniProt IDs for BLASTHITs or user-supplied sequences in metanode
Cluster Number	Number assigned to cluster, in order of decreasing number of sequences in the clusters
Cluster Sequence Count	Number of sequences in the cluster
Node.fillColor	Unique color assigned to cluster, in hexadecimal
Singleton Number	Number assigned to singleton
Present in ENA Database?	“true” if UniProt ID was found in an ENA file (see ENA Database Genome ID); otherwise “false”
Genome Neighbors in ENA Database?	“true” if ENA file has sequences for query plus neighbors; “false” if ENA file has no neighbors; “n/a” if not present in ENA database
ENA Database Genome ID	ENA file used to obtain genome neighbors
Neighbor Pfam Families	Pfam IDs of genome neighborhood proteins in the user-specified window and ≤0% query-neighbor co-occurrence

Neighbor InterPro Families	InterPro IDs of genome neighborhood proteins in the user-specified window and ≤0% query-neighbor co-occurrence
Organism	organism genus/genera and species, from UniProt taxonomy.xml
Taxonomy ID	NCBI taxonomy identifier(s), from UniProt
UniProt Annotation Stastus	SwissProt - manually annotated; TrEMBL - automatically annotated; from UniProt
Description	protein name(s)/annotation(s), from UniProtKB
SwissProt Description	protein name(s)/annotation(s), from UniProtKB for SwissProt reviewed entries
Sequence Length	number(s) of amino acid residues, from UniProt
Gene name	gene name(s)
NCBI IDs	RefSeq/GenBank IDs and GI numbers, from UniProt idmapping
Superkingdom	domain of life of the organism, from UniProt taxonomy.xml
Kingdom	kingdom of the organism, from UniProt taxonomy.xml
Phylum	Phylogenetic phylum of the organism, from UniProt taxonomy.xml
Class	Phylogenetic class of the organism, from UniProt taxonomy.xml
Order	Phylogenetic order of the organism, from UniProt taxonomy.xml
Family	Phylogenetic family of the organism, from UniProt taxonomy.xml
Genus	Phylogenetic genus of the organism, from UniProt taxonomy.xml
Species	Phylogenetic species of the organism, from UniProt taxonomy.xml
EC	EC number, from UniProt
PFAM	Pfam family, from UniProt
PDB	Protein Data Bank entry, from UniProt
InterPro (Domain)	InterPro domain(s), from InterPro
InterPro (Family)	InterPro family(ies), from InterPro
InterPro (Homologous Superfamily)	InterPro homologous superfamily(ies), from InterPro
InterPro (Other)	Other InterPro classes (repeat, site), from InterPro
BRENDA ID	BRENDA Database ID, from UniProt
CAZY Name	Carbohydrate-Active enZYmes (CAZy) family name(s), from UniProt
GO Term	Gene Ontology classification(s), from UniProt
KEGG ID	KEGG Database ID, from UniProt
PATRIC ID	PATRIC Database ID, from UniProt
STRING ID	STRING Database ID, from UniProt
HMP Body Site	location(s) of organism(s) in/on the body, if human microbiome organism, spreadsheet from HMP
HMP Oxygen	oxygen requirement(s), if human microbiome organism, spreadsheet from HMP
P01 gDNA	availability of gDNA(s) at EFI Protein Core, custom

Node Attribute	**Option C without FASTA header reading**
Name	zzznnn, where nnn = number of the sequence in FASTA file
Shared Name	zzznnn, where nnn = number of the sequence in FASTA file
Description	FASTA Header
Sequence Length	Length of sequence in FASTA entry
Present in ENA Database?	“false”
Genome Neighbors in ENA Database?	“n/a”
ENA Database Genome ID	none

Table S4. GNN Node Attributes for SSN Cluster Hub-Nodes and Pfam Family Spoke-Nodes

Node Attribute	**SSN cluster hub-nodes**
Shared name	Input SSN cluster number
Name	Input SSN cluster number
Cluster Number	Input SSN cluster number
# of Sequences in SSN Cluster	Total number of sequences in SSN cluster
# of Sequences in SSN Cluster with Neighbors	Number of sequences in SSN cluster with neighbors (queriable sequences)
Hub Queries with Pfam Neighbors	Summary of number of queriable sequences with a neighbor in the Pfam family
Hub Pfam Neighbors	Summary of the total # of Pfam neighbors found by the queriable sequences
Hub Average and Median Distances	Summary of average and median distances between the query and neighbors in each Pfam family
Hub Co-occurrence and Ratio	Summary of the query-neighbor co-occurrence (decimal value) and ratio (fraction) for each Pfam family
Node.fillColor	Hexadecimal color for the SSN cluster in the colored SSN, used by Cytoscape
Node.shape	"hexagon", used by Cytoscape
Node Size	"70.0", used by Cytoscape
Node Attribute	**Pfam family spoke-nodes**
Shared name	Pfam family short name
Name	Pfam family short name
SSN Cluster Number	SSN Cluster that found neighbors in the Pfam family
Pfam	Pfam family number (PFnnnnn)
Pfam description	Pfam family description
# of Queries with Pfam Neighbors	Number of queriable sequences with a neighbor in the Pfam family
# of Pfam Neighbors	Number of Pfam neighbors found by the queriable sequences
Query-Accessions	List of SSN cluster queries that found neighbors in the Pfam family
Query-Neighbor Accessions	Information about query-neighbor pairs in the Pfam family
Query-Neighbor Arrangement	Genome context information for the query-neighbor pairs in the Pfam family
Average Distance	Average distance (in ORFs) between the SSN cluster queries and Pfam neighbors
Median Distance	Median distance (in ORFs) between the SSN cluster queries and Pfam neighbors
Co-occurrence	Decimal value of ratio of queries that found neighbors to queriable sequences
Co-occurrence Ratio	Ratio of queries that found neighbors to queriable sequences
Node.fillColor	#EEEEEE, grey in hexadecimal, used by Cytoscape

Node.shape	"ellipse", "diamond", or "square"; explained in on-line tutorial, used by Cytoscape
Node.size	Co-occurrence * 100, used by Cytoscape

Table S5. GNN Node Attributes for Pfam Family Hub-Nodes and SSN Cluster Spoke-Nodes

Node Attribute	**Pfam family hub-nodes**
Shared name	Pfam family short name
Name	Pfam family short name
Pfam	Pfam family number (PFnnnnn)
Pfam description	Pfam family description
# of Sequences in SSN Cluster	Total number of sequences in SSN cluster
# of Sequences in SSN Cluster with Neighbors	Number of sequences in SSN cluster with neighbors (queriable sequences)
# of Queries with Pfam Neighbors	Number of queriable sequences with a neighbor in the Pfam family
# of Pfam Neighbors	Number of Pfam neighbors found by the queriable sequences
Query-Neighbor Accessions Query-Neighbor Arrangement Hub Average and Median Distances	Information about query-neighbor pairs in the Pfam family
	Genome context information for the query-neighbor pairs in the Pfam family
	Summary of average and median distances between the query and neighbors
Hub Co-occurrence and Ratio	Summary of the query-Pfam family co-occurrence (decimal value) and ratio (fraction)
Node.fillColor	"#FFFFFF", white in hexadecimal, used by Cytoscape
Node.shape	"hexagon", used by Cytoscape
Node.size	"70.0", used by Cytoscape
Node Attribute	**Description - SSN cluster spoke-nodes**
Shared name	Input SSN cluster number
Name	Input SSN cluster number
Cluster Number	Input SSN cluster number
# of Sequences in SSN Cluster	Total number of sequences in SSN cluster
# of Sequences in SSN Cluster with Neighbors	Number of sequences in SSN cluster with neighbors (queriable sequences)
# of Queries with Pfam Neighbors	Number of queriable sequences with a neighbor in the Pfam family
# of Pfam Neighbors	Number of Pfam neighbors found by the queriable sequences
Query-Accessions	List of queries in each SSN cluster that found neigbhors in the Pfam family
Query-Neighbor Accessions	Information about query-neighbor pairs in the Pfam family
Query-Neighbor Arrangement	Genome context information for the query-neighbor pairs in the Pfam family
Average Distance	Average distance (in ORFs) between the SSN cluster queries and Pfam neighbors

Median Distance	Median distance (in ORFs) between the SSN cluster queries and Pfam neighbors
Co-occurrence	Decimal value of ratio of queries that found neighbors to queriable sequences
Co-occurrence Ratio	Ratio of queries that found neighbors to queriable sequences

Table S6. Additional Node Attributes for SSNs Generated by EFI-CGFP

Additional Node Attribute	SSNs with Marker Identification Results
Seed Sequences	ID for the (meta)node that contains the UniProt ID for the seed sequence for a ShortBRED family
Seed Sequence Cluster(s)	ID of ShortBRED family seed sequence to which the (meta)node contributes family members
Marker Types	"true", "quasi", or "junction"
Number of Markers	Number of markers identified for ShortBRED family seed sequence

Additional Node Attribute	SSNs with Metagenome Abundance Quantify Results
Metagenomes Identified by Markers	For ShortBRED family seed sequences, names of metagenome datasets identified by its markers
Metagenomes Identified by CD-HIT Family	For IDs that contribute to ShortBRED (CD-HIT) family seed sequences, names of metagenome datasets identified with the seed sequence markers