

GHOSTX: A Fast Sequence Homology Search Tool for Functional Annotation of Metagenomic Data

Shuji Suzuki, Takashi Ishida, Masahito Ohue, Masanori Kakuta, and Yutaka Akiyama

Abstract

Metagenomic analysis based on whole genome shotgun sequencing data requires fast protein sequence homology searches for predicting the function of proteins coded on metagenome short reads. However, huge amounts of sequence data cause even general homology search analyses using BLASTX to become difficult in terms of computational cost. GHOSTX is a sequence homology search tool specifically developed for functional annotation of metagenome sequences. The tool is more than 160 times faster than BLASTX and has sufficient search sensitivity for metagenomic analysis. Using this tool, user can perform functional annotation of metagenomic data within a short time and infer metabolic pathways within an environment.

Keywords Metagenomic analysis, Sequence homology search, Whole genome shotgun sequencing, Functional annotation, Substitution-score matrix

1 Introduction

Metagenomics is the study of the genomes of uncultured microbes obtained directly from microbial communities in their natural habitats. Such analyses have recently become more popular and important as the throughput of DNA sequencers has increased. Previously, metagenomic analysis was performed based on 16S rRNA data obtained from Sanger-sequencing methods, and the aim was to obtain the phylogenetic profiles of microbial communities from a target environment. However, whole-genome shotgun (WGS) sequencing, carried out using next-generation sequencing (NGS) technologies, produces huge amounts of metagenomic data. This enables us to uncover an abundance of orthologous groups, i.e., the distribution of gene/protein functions, in environmental samples. Based on such information, we can infer metabolic pathways within an environment and compare a metagenomic sample to the others based on its functions or functional

categories. Various metagenomic studies have employed functional annotation techniques, facilitating novel scientific discoveries, such as the existence of enterotypes [1] and the relationship between gut microbes and type II diabetes [2].

To perform functional annotations of metagenomic data, it is necessary to determine the function of DNA short reads obtained from environmental samples. However, such metagenome sequences generally include DNA sequences from many different species, and closely related reference genome sequences are often unavailable. Thus, DNA short reads are translated into protein-coding sequences, and the functions of proteins are then predicted for more sensitive identification of novel genes. The BLASTX [3] program and well-annotated protein sequence databases, such as KEGG [4] and COG [5], have been used for protein function prediction with metagenomic data [6]. However, the computing speed of BLASTX is insufficient for analyzing the large quantities of metagenomic data produced by current sequencers, such as the Illumina HiSeq2500, which can produce several hundred billion base pairs of sequence data in a single run. Thus, a special tool is required for functional prediction of metagenome sequences.

GHOSTX is a sequence homology search tool specifically developed for functional annotation of metagenome sequences [7]. The GHOSTX algorithm uses a seed search method that relies on a score-based optimal seed length. In this method, only subsequences with scores greater than or equal to a threshold are searched, based on a given score matrix. Thus, the algorithm can effectively exclude seeds with sufficient length but insufficient match scores. In addition, the program accelerates its seed search process by using suffix arrays of both queries and database sequences. As a result, GHOSTX can achieve approximately 160 times greater speeds than BLASTX searches at similar levels of sensitivity. Several other tools, such as RAPSearch2 [8] and DIAMOND [9], have also been developed for functional annotation of metagenome sequences. The performances of these programs are comparable to that of GHOSTX; however, their algorithms are designed on the premise of the BLOSUM62 score matrix, whereas GHOSTX can use any score matrix and thus can be applied in various types of analyses. The GHOSTX program has been employed for several services, such as GhostKOALA [10], and has been shown to provide reliable results.

2 Materials

In this section, we describe the input data to execute GHOSTX and supplemental tools for the analysis of GHOSTX output data.

2.1 Input Data

The input data for GHOSTX are metagenome sequences produced from a DNA sequencer as query and protein sequences as a

```

>61JGYAAXX100510:7:115:4328:19392/1
ACGTTCCCTAAGTCCAATGCTTTGAACGCTTGCCATATTGCTCAAGCACTTGTGATTCT
TTTGATTGTGTAATCACCG
>61JGYAAXX100510:7:116:12157:10767/1
TGTCAGTTCCTTCAAAGGTATCCAAGTCAAAGGCATGCATTGGTTGACCAAAGTAGAGCA
GGATGTAGTTGGTCACGTCCACTACGTTATTGATGGGACGA
>61JGYAAXX100510:7:116:19272:8295/1
ATCTCCAAAGCAGCAAGTGTCGCAAGTGATATCCACGCTTCAGGATCAGTAATCCATTCA
AACATAATTTTCTATAATAGAGTTAAATAAATCCGAAAG
>61JGYAAXX100510:7:116:9375:5125/1
CAACAATACTATGTCATTGTGAAAGCGAAAGAACTTCCAGTTGTACCAACAAAACAGTA
AATAAAAATAATGAACCAGTAGATCCAACAGATCCAACAAC

```

Fig. 1 An example of a FASTA-formatted file. The description line is distinguished from the sequence data by a greater-than (“>”) symbol at the beginning of a line. By concatenating multiple single-sequence FASTA files, a multiple sequence FASTA formatted-file can be generated. The figure is a part of deep WGS sequencing data from the Human Microbiome Project (HMP) [11], with sequences from a buccal mucosa sample (SRS011090)

database. As a query input, GHOSTX requires the DNA sequence data in a FASTA format (Fig. 1). A multiple sequence FASTA format is acceptable. If the format of a query file is FASTQ, the user has to use an external tool, such as FASTX-Toolkit [12], to convert the format. The `fastq_to_fasta` command of FASTX-Toolkit converts the FASTQ-formatted file into a FASTA-formatted file.

```
$ fastq_to_fasta -i query.fastq -o query.fasta
```

The database sequences of GHOSTX are annotated protein sequences in the FASTA format. The database must be indexed preliminarily. The steps required to index the database sequences are described in Subheading 3. The user can use any protein sequence database, such as NCBI nr or Uniprot. However, well-annotated databases, such as COG and EggNOG, are recommended for functional prediction and analysis. In addition, the KEGG GENES database [4] is preferable because the post-analysis tool KEGG Analyzer, developed by our laboratory, is compatible only with the KEGG database.

2.2 Programs in the Software Packages

GHOSTX is available as open-source free software under the terms of the BSD 2-Clause license in source code form. GHOSTX is written in C++. It can be compiled and run on a wide variety of UNIX platforms and similar systems (including FreeBSD and Linux). The core of the GHOSTX is `ghostx` binary, generated after source codes are compiled. The KEGG Analyzer program for phylogenetic analysis and functional analysis is also available as open-source free software.

2.3 Web Sites

GHOSTX (version 1.3.7) is available at <http://www.bi.cs.titech.ac.jp/ghostx/>, and the previous version is still downloadable at <http://www.bi.cs.titech.ac.jp/ghostx/releases/>. The KEGG analyzer program is also available at <http://www.bi.cs.titech.ac.jp/ghostx/kegg/>.

3 Methods

3.1 Database Construction

Similar to NCBI BLAST, database sequences have to be indexed before GHOSTX sequence homology search. To construct the indices of database sequences, the `ghostx db` command is used.

```
$ ghostx db -i database.fasta -o exdb
```

The `ghostx db` command converts a FASTA file into GHOSTX format database files. The input format of `ghostx db` is as follows:

```
$ ghostx db -i database_fasta_file -o database_name [-l chunk_size] [-t p|d]
```

where `-i` and `-o` options are required. `-l` specifies the chunk size of a database (default as 1 GByte = “-l 1073741824”). GHOSTX requires a considerable amount of memory during its homology search. The required memory size depends on the database size. And the indexing technique requires larger memory size than storing original sequences. However, current computing systems generally have relatively small memory sizes compared with database sizes. Therefore, GHOSTX divides a database into several chunks, each of whose size is specified on the `-l` option, before it constructs its indices. For example, when using the KEGG GENES database (3.9 GByte), with the chunk sizes set to 512 Mbyte, 1 GByte, and 2 GByte, the total memory sizes required for storing indexed database and performing homology searches are 4.6, 9.2, and 18.2 GByte, respectively. The relative computation speeds (based on a 2 GByte chunk) are 0.8×, 0.9×, and 1.0× when the chunk sizes are set to 512 Mbyte, 1 GByte, and 2 GByte, respectively [7]. The `-t` option can designate the database sequence type as protein “p” or DNA “d.” Protein “p” is chosen as default. Searching the query sequence on a DNA database can be executed by “-t d.”

3.2 Homology Search

To perform a sequence homology search, the `ghostx aln` command is used.

```
$ ghostx aln -i query.fasta -d exdb -o exout
```

The `ghostx aln` command has several options; the required arguments are `-i` as input query fasta, `-d` as indexed database file, and `-o` as output file name, and the additional options are as follows:

<code>-a</code>	number of CPU threads (default: 1)
<code>-v</code>	maximum number of alignments for each subject
<code>-b</code>	maximum number of the output for a query
<code>-q</code>	query sequence type, protein “p” or DNA “d” (default: “d”)
<code>-t</code>	database sequence type, protein “p” or DNA “d” (default: “p”)
<code>-s</code>	upper mismatch score in seed extension
<code>-T</code>	threshold score of seed search
<code>-G</code>	gap opening penalty (same as BLAST)
<code>-E</code>	gap extension penalty (same as BLAST)
<code>-F</code>	filter query sequence (same as BLAST)
<code>-M</code>	score matrix file

Among these parameters, the upper mismatch score D (`-s`) is the limit of acceptable score difference to determine whether a seed extends or not, and the threshold of seed search T_{seed} (`-T`) is minimum score for a hit in seed search. They regulate the sensitivity and computation speed of searches. The default parameters of “`-s 4 -T 30`” ($D = 4$, $T_{\text{seed}} = 30$) have the best balance between the sensitivity and computation speed. At this time, the sensitivity of GHOSTX is almost the same as that of RAPSearch2 [8]. If a faster calculation is needed, parameters of “`-s 1, -T 30`” ($D = 1$, $T_{\text{seed}} = 30$) with a smaller mismatch allowance are a good option. On the other hand, if higher sensitivity is needed, “`-s 4, -T 24`” ($D = 4$, $T_{\text{seed}} = 24$) with a lower threshold can be used. Refer to Table S1 of [7] for additional information regarding the relationships of these cutoff parameters with the balance between sensitivity and calculation time.

Furthermore, the `-b` option can be used to specify the maximum number of outputs for a query, and the default is 10. The `-v` option can be used to specify the maximum number of alignments for each subject, and the default is 1. The `-F` option is for masking off segments of the query sequence that have low compositional complexity determined by the SEG program [13]. Note that the threshold E -value for saving hits (`-e` of legacy BLAST and `-evalue` of BLAST+) does not exist in GHOSTX; thus, if the user wants to eliminate hits that have higher E -values than a user-defined threshold some text processing is needed after obtaining the GHOSTX output.

1	2	3	4	5	6	7	8	9	10	11	12
613GYAAX100510:7:100:10000:7489/1	smb:smi_0286 hypothetical protein; K07024	0.969697	33	1	0	100	2	79	111	4.09365e-21	102.436
613GYAAX100510:7:100:10000:7489/1	std:SPPN_10260 putative HAD superfamily hydrolase; K07024	0.969697	33	1	0	100	2	79	111	5.49279e-21	102.012
613GYAAX100510:7:100:10000:7489/1	snx:SPNOXC_17560 haloacid dehalogenase-like hydrolase; K07024	0.939394	33	2	0	100	2	79	111	1.83901e-19	97.7704
613GYAAX100510:7:100:10000:7489/1	spv:SPH_2139 Cof family protein; K07024	0.939394	33	2	0	100	2	66	98	1.83901e-19	97.7704
613GYAAX100510:7:100:10000:7489/1	snn:SP70585_2069 Cof family protein; K07024	0.939394	33	2	0	100	2	79	111	1.83901e-19	97.7704
613GYAAX100510:7:100:10001:16821/1	snc:HMPREF0837_11837 hypothetical protein	0.888889	18	2	0	47	100	1	18	1.04419e-07	57.9001
613GYAAX100510:7:100:10001:16821/1	snp:SP_1601 hypothetical protein	0.888889	18	2	0	47	100	1	18	1.04419e-07	57.9001
613GYAAX100510:7:100:10001:16821/1	snn:SP670_1685 hypothetical protein	0.888889	18	2	0	47	100	1	18	1.04419e-07	57.9001
613GYAAX100510:7:100:10001:16821/1	smb:smi_0674 hypothetical protein	0.888889	18	2	0	47	100	1	18	1.04419e-07	57.9001
613GYAAX100510:7:100:10001:16821/1	spd:SPD_1426 hypothetical protein	0.888889	18	2	0	47	100	1	18	1.04419e-07	57.9001

Fig. 2 An example of a GHOSTX output file. The GHOSTX output is BLAST-like tab-separated format. This example is a search result of a buccal mucosa metagenome sample from the HMP (SRS011090) using the KEGG GENES database. The columns are as follows: (1) name of the query sequence; (2) name of the homolog sequence (subject); (3) sequence identity; (4) alignment length; (5) the number of mismatches in the alignment; (6) the number of gap openings in the alignment; (7) start position of the query in the alignment; (8) end position of the query in the alignment; (9) start position of the subject in the alignment; (10) end position of the subject in the alignment; (11) *E*-value; and (12) normalized score

Figure 2 shows an example of the output from a GHOSTX homology search. The output format of GHOSTX is a tab-separated format, similar to that of BLAST. The format contains 12 columns, as described in the legend of Fig. 2.

3.3 Post-Analysis (KEGG Analyzer)

The KEGG Analyzer can calculate the corrected relative abundance of molecular-level functions based on KEGG Orthology (KO) from the GHOSTX search output with the KEGG GENES database. The KEGG Analyzer can also be used for generating phylogenetic profiles. The tool is effective and easy to use but requires a KEGG subscription because it refers to KEGG-licensed files. For more information, refer to the help file in the KEGG Analyzer program.

The outputs of the KEGG Analyzer include a normalized KO count (`ko.csv`) and normalized phylogenetic profile (`phylogeny.csv`). The statistical analysis of KO-based molecular functions and phylogenetic analysis can be performed using these files. An example of functional analysis with the tool is given in Subheading 4.

4 Case Study

In this section, we show an example of a homology search with GHOSTX and phylogenetic profile analysis and functional analysis from the homology search results. This case study aimed to identify differences in the human oral bacterial flora from WGS metagenomic samples based on phylogeny and gene function. The input data were deep WGS sequencing data from an HMP buccal mucosa sample (SRS011090), including a total of 1,787,927 paired-end reads (file size of FASTQ: 449 MByte). The database used here was `genes.pep` from the KEGG GENES database (file size: approximately 5.8 GByte on February 2, 2016).

4.1 Homology Search with GHOSTX

To perform phylogenetic analysis and functional analysis of metagenomic data, sequence homology searches are required. The user has to create indices of a database using the `ghostx db` command, as shown in the previous section.

```
$ ghostx db -i genes.pep -o kegg.db
```

Then, the user executes the homology search with the `ghostx aln` command using indexed database `kegg.db`.

```
$ ghostx aln -i SRS011090.fasta -d kegg.db
-o SRS011090_out.csv
```

Even using GHOSTX, the search process requires more than 100 h with a workstation. Thus, a cluster system with multiple computing nodes is recommended to execute the process [14]. Finally, the user can obtain the homology search result `SRS011090_out.csv` in a tab-separated format, as shown in Fig. 2.

4.2 Phylogenetic Profile Analysis

Here, we show an example of calculation for relative abundance of phylogeny at the genus and phylum levels from the KEGG Analyzer output. The KEGG Analyzer was executed using the GHOSTX output file and some files from the KEGG database.

```
$ kegg_analyzer genes_list gi_taxid root_map
ko_enzyme usc_g_list SRS011090_out.csv
[ghostx_score_threshold] [ghostx_identity_
threshold] ko.csv usc_g_count.csv normalize_
error.txt unnormalized_out.csv phylogeny.csv
```

where `genes_list` and `root_map` are generated by the scripts from KEGG and NCBI files contained in the KEGG Analyzer package, and `gi_taxid` and `ko_enzyme` are downloaded from KEGG FTP. `usc_g_list` is contained in the KEGG Analyzer package. `ko.csv`, `usc_g_count.csv`, `normalize_error.txt`, `unnormalized_out.csv`, and `phylogeny.csv` are output files. Only `phylogeny.csv` and `ko.csv` are used in this case study.

The user can generate a two-column file from `phylogeny.csv` using a one-line command as shown below:

Relative abundance based on genus rank:

```
$ tail -n +7 phylogeny.csv | cut -f 4,5 | awk -F
"\t" '{sum[$1]+=$2}END{for(k in sum){print k
"\t" sum[k]}}'
```

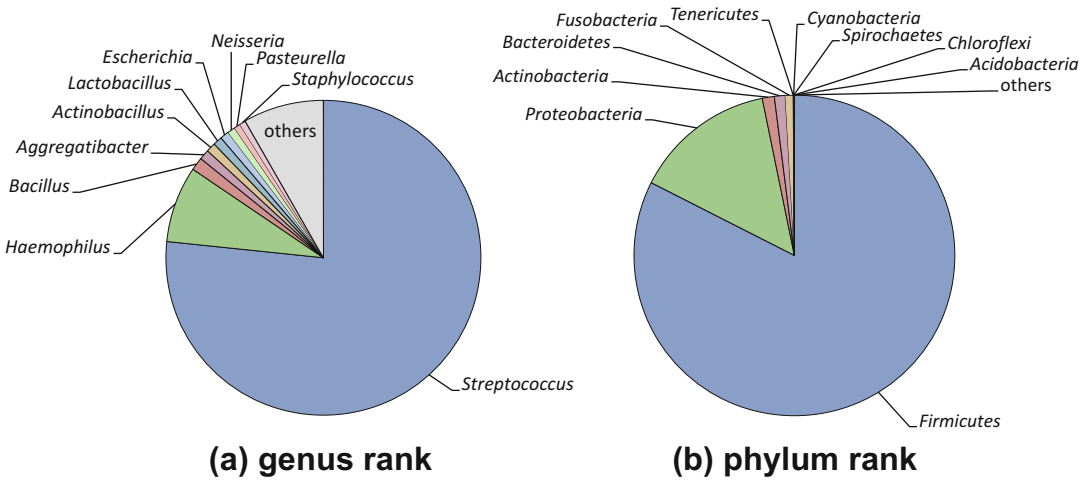


Fig. 3 Phylogenetic profiles of the buccal mucosa HMP metagenome sample (SRS011090). These pie charts show the top 10 phylogenies with higher abundance. The remaining phylogenies are grouped into “others”

Relative abundance based on phylum rank:

```
$ tail -n +7 phylogeny.csv | cut -f 3,5 | awk -F
"\t" '{sum[$1]+=$2}END{for(k in sum){print k
"\t" sum[k]}}'
```

Figure 3 describes the relative abundances as pie charts created using spreadsheet software. From the figure, we can understand that *Firmicutes* is the most major in the phylum rank and most of genera belonging to *Firmicutes* are *Streptococcus* in buccal mucosa.

4.3 Functional Analysis

The abundance and distribution of molecular functions (KOs) of metagenomic data can be understood from a normalized KO count (ko.csv) generated by KEGG Analyzer.

However, it is difficult to understand which pathway is specifically activated or inactivated only from the counts. In such a situation, mapping the information onto pathway maps improves our understanding.

iPATH2 [15] is a web-based tool for the visualization, analysis, and customization of various pathways maps. To assign the information onto the pathway map, the user should first access the iPATH2 website (<http://pathways.embl.de/iPath2.cgi>) and click the “Customize” button. Next, the user should click the “New selection” tab and paste the KO list in “Element selection” box. Then, the user should click the “Submit data and customize maps” button (Fig. 4). Parameters such as line color (# and hex color code), line width (W and integer value), and others can be set using the KO list.

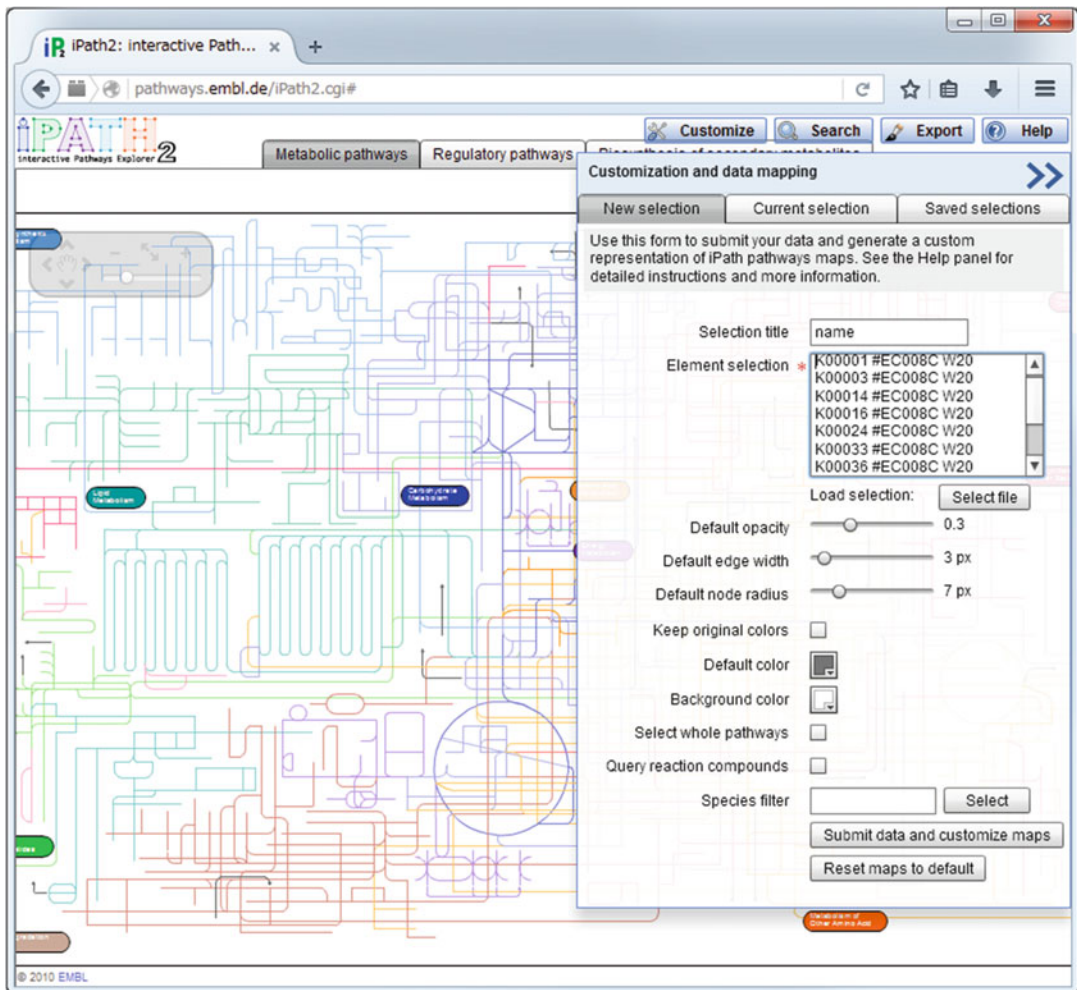


Fig. 4 Screen shot of iPATH2 element selection input. The KO list (color code of magenta [#EC008C] and line width [W20] are shown) is input into the “Element selection” box or “Load selection” dialogue as a text file

Figure 5 shows the KOs with greater than 0.01% relative abundance in the buccal mucosa HMP sample (SRS011090) loaded onto the iPATH2 pathway map. The iPATH2 input file was generated from the KEGG Analyzer output file (ko.csv) by a one-line command:

```
$ tail -n +3 ko.csv | awk -F "\t" ' $4>0.0001 &&
$6!="KO not found"{print substr($6,4) " #EC008C
W20"}' > ipath.in
```

where ipath.in is the input file for iPATH2 mapping.

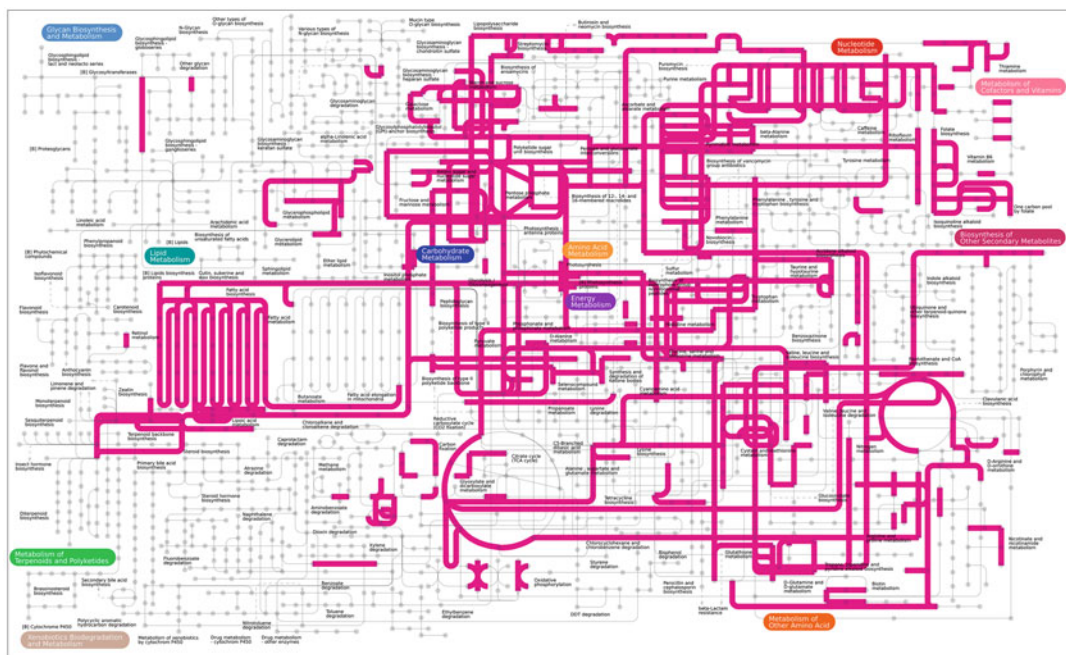


Fig. 5 Example of visualization of the molecular functions of the metagenome sample. In this figure, KOs that have greater than 0.01% relative abundance in the buccal mucosa HMP sample (SRS011090) were mapped onto iPATH2 metabolic pathways

From the figure, we can understand that gene functions in buccal mucosa microbes cover a wide variety of biological pathways, while the distribution of the genera is mostly occupied by a dominant genus, *Streptococcus*.

Acknowledgments

This work was partly supported by the Strategic Programs for Innovative Research (SPIRE) Field 1 Supercomputational Life Science of the Ministry of Education, Culture, Sports, Science and Technology (MEXT) of Japan and Core Research for Evolutional Science and Technology (CREST) “Extreme Big Data” of the Japan Science and Technology Agency (JST).

References

1. Arumugam M, Raes J, Pelletier E et al (2011) Enterotypes of the human gut microbiome. *Nature* 473:174–180
2. Qin J, Li Y, Cai Z et al (2012) A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* 490:55–60
3. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215:403–410
4. Kanehisa M, Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 28:27–30

5. Tatusov RL, Fedorova ND, Jackson JD et al (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* 4:41
6. Kurokawa K, Itoh T, Kuwahara T et al (2007) Comparative metagenomics revealed commonly enriched gene sets in human gut microbiomes. *DNA Res* 14:169–181
7. Suzuki S, Kakuta M, Ishida T, Akiyama Y (2014) GHOSTX: an improved sequence homology search algorithm using a query suffix array and a database suffix array. *PLoS ONE* 9: e103833
8. Zhao Y, Tang H, Ye Y (2012) RAPSearch2: a fast and memory-efficient protein similarity search tool for next-generation sequencing data. *Bioinformatics* 28:125–126
9. Buchfink B, Xie C, Huson DH (2015) Fast and sensitive protein alignment using DIAMOND. *Nat Methods* 12:59–60
10. Kanehisa M, Sato Y, Morishima K (2016) BlastKOALA and GhostKOALA: KEGG tools for functional characterization of genome and metagenome sequences. *J Mol Biol* 428:726–731
11. The Human Microbiome Project Consortium (2012) Structure, function and diversity of the healthy human microbiome. *Nature* 486:207–214
12. Gordon A, Hannon G (2012) FASTX-Toolkit. http://hannonlab.cshl.edu/fastx_toolkit. Accessed 27 Mar 2016
13. Wootton JC, Federhen S (1993) Statistics of local complexity in amino acid sequences and sequence databases. *Comput Chem* 17:149–163
14. Kakuta M, Suzuki S, Ishida T, Akiyama Y. A massively parallel sequence similarity search for metagenomic sequencing data (submitted for publication)
15. Yamada T, Letunic I, Okuda S, Kanehisa M, Bork P (2011) iPath2.0: interactive pathway explorer. *Nucleic Acids Res* 39:W412–W415

Protein Function Prediction

Methods and Protocols

Kihara, D. (Ed.)

2017, X, 239 p. 82 illus., 71 illus. in color., Hardcover

ISBN: 978-1-4939-7013-1

A product of Humana Press