Lecture 20

Metagenomic data analysis

MCB 416A/516A
Statistical Bioinformatics and Genomic Analysis

Prof. Lingling An
Univ of Arizona

Metagenomic data analysis pipeline

- Sequence data to count data
- Downstream analysis
  - Visualization (among samples)
  - Differential abundant analysis
  - Network analysis (among features)
  - …

From sequence data to count data

- Metagenomic sequence data
  - whole genome
  - marker genes (e.g., 16s rRNA, or 18S rRNA)
- Download datasets
  - NCBI -SRA
  - MG-RAST
  - Papers
  - …
- Mapping sequence to reference database
  - blast or blastx (for whole genome data)
  - Use RDP classifier (for 16S data)
  - …
- In R
  - Read alignment results into R and make count table

Download 16S rRNA dataset

www.ncbi.nlm.nih.gov/sra/?term=metagenomics

NCBI Resources How To Sign in to NCBI

SRA SRA metagenomics Search

Create alert Advanced Help

Access
Controlled (8,085)
Public (233,196)

Source
DNA (224,255)
RNA (6,558)
metagenomic (152,943)

Type
exome (324)
genome (32,590)

Other
aligned data (905)

Clear all

Show additional filters

Summary 20 per page Send to

Search results

Items: 1 to 20 of 234329 << First < Prev Page 1 of 11717 Next > Last >>

1. Caciocavallo cheeses and swabs environmental microbiota from the dairy plant
   1 LS454 (454 GS Junior) run: 4,735 spots, 2.5M bases, 5.2Mb downloads
   Accession: SRX1704838
2. Caciocavallo cheeses and swabs environmental microbiota from the dairy plant
   1 LS454 (454 GS Junior) run: 4,424 spots, 2.3M bases, 4.9Mb downloads
   Accession: SRX1704836
3. Caciocavallo cheeses and swabs environmental microbiota from the dairy plant
   1 LS454 (454 GS Junior) run: 5,903 spots, 3.2M bases, 6.9Mb downloads
   Accession: SRX1704835
4. Caciocavallo cheeses and swabs environmental microbiota from the dairy plant
   1 LS454 (454 GS Junior) run: 11,060 spots, 6M bases, 12.5Mb downloads
   Accession: SRX1704834
5. Caciocavallo cheeses and swabs environmental microbiota from the dairy plant
   1 LS454 (454 GS Junior) run: 9,184 spots, 4.9M bases, 10.2Mb downloads
   Accession: SRX1704833
6. Caciocavallo cheeses and swabs environmental microbiota from the dairy plant
   1 LS454 (454 GS Junior) run: 3,688 spots, 1.9M bases, 4.1Mb downloads
   Accession: SRX1704832
7. Caciocavallo cheeses and swabs environmental microbiota from the dairy plant
   1 LS454 (454 GS Junior) run: 6,933 spots, 3.7M bases, 7.7Mb downloads

Filters: Manage Filters

Results by taxon

Top Organisms [Tree]
- human gut metagenome *(31795)*
- soil metagenome *(31450)*
- human metagenome *(24076)*
- gut metagenome *(15044)*
- mouse gut metagenome *(10544)*
- marine metagenome *(10413)*
- metagenome *(5223)*
- Homo sapiens *(4498)*
- freshwater metagenome *(4426)*
- aquatic metagenome *(3960)*
- human oral metagenome *(3898)*
- bovine gut metagenome *(3545)*
- rhizosphere metagenome *(3460)*
- human skin metagenome *(3336)*
- sediment metagenome *(3023)*
- terrestrial metagenome *(3003)*
- human lung metagenome *(2876)*
- unclassified Bacteria (miscellaneous) *(2248)*
- food metagenome *(2176)*
- wastewater metagenome *(2135)*
- All other taxa *(63205)*

Less...

Top Bioprojects

Protistan V9 method *(19)*
Functional genomics project ... *(6)*

Search in related databases

NCBI Resources How To Sign in to NCBI

BioProject

BioProject ((metagenomics) AND "human gut metagenome"[orgn:__txid408170]) AND bioproject_sra[filter] N Search

Create alert Advanced Help

BioProject

Project Types
Primary submission (12)

Data Types
Metagenome (7)
Other (4)
Targeted locus (1)

Project Data
Nucleotide (1)
Protein (1)
Assembly (1)
SRA (12)

Scope
Monoisolate (1)
Multi-species (1)
Environmental (7)
Other (3)

Clear all

Show additional filters

Display Settings: Summary, 20 per page, Sorted by Default order

Send to:

Filters: Manage Filters

Search results

Items: 12

1. human gut metagenome
   16S amplicon sequence analysis of stools and surface of CRC tissues from patients
   Taxonomy: *human gut metagenome*
   Project data type: Targeted Locus (Loci)
   Scope: Multispecies
   Laboratory of **Metagenomics**, Graduate School of Frontier Sciences, The University of Tokyo
   Accession: PRJDB4636 ID: 317260

2. human gut metagenome
   Metagenomics of Japanese gut microbiomes
   Taxonomy: *human gut metagenome*
   Project data type: Metagenome
   Scope: Environment
   The University of Tokyo
   Accession: PRJDB3601 ID: 314752

3. Effect of Saccharomyces boulardii and mode of delivery on the early development of the gut microbial community in preterm infants
   Project data type: Other
   Scope: Other
   Centrum Medyczne Kształcenia Podyplomowego
   Accession: PRJEB9898 ID: 313762

4. Colorectal cancer and the human gut microbiome
   Reproducibility of associations between the human gut microbiome and colorectal cancer assessed in a patient population from Washington, DC, USA
   Project data type: Other
   Scope: Other
   EMBL
   Accession: PRJEB12449 ID: 310722

Find related data

Database: Select

Find items

Search details

```
(metagenomics[All Fields] AND
"human gut metagenome"[orgn]) AND
bioproject_sra[filter] NOT
bioproject_gap[filter]
```

Search

See more...

Recent activity

Turn Off Clear

((metagenomics) AND "human gut metagenome"[orgn]) AND bioproje BioProject

(metagenomics) AND "human gut metagenome"[orgn] (31797) SRA

((metagenomics) AND "human gut metagenome"[orgn]) AND biosam BioSample

metagenomics (234329) SRA

SRA Handbook

See more...

ww.ncbi.nlm.nih.gov/bioproject

RDP classifier for 16S rRNA data

https://rdp.cme.msu.edu/classifier/classifier.jsp

BROWSERS | CLASSIFIER | LIBCOMPARE | SEQMATCH | PROBE MATCH | FUNGENE | RDPIPELINE | SEQCART | TAXOMATIC | TREE BUILDER | ASSIGNGEN

Classifier - Start

[NEW procedural tutorials for CLASSIFIER | video tutorial | help]

Introduction

Classifier now provides gene copy number adjustment for 16S gene sequences. The 16S gene copy number data is provided by rrnDB website.

We are pleased to release two new Fungal ITS training sets to classify fungal ITS sequences. Warcup is an version from an active curatorial effort kindly provided by Paul Greenfield, Vinita Deshpande and colleagues of the Australian CSIRO [V. Deshpande, Q. Wang, P. Greenfield, M. Charleston, A. Porras-Alfaro, C. R. Kuske, J. R. Cole, D. J. Midgley, and N. Tran-Dinh. 2015. Fungal identification using a Bayesian Classifier and the 'Warcup' training set of Internal Transcribed Spacer sequences. Mycologia (In press)]. UNITE is a set consisting of UNITE core sequences for each dynamic species hypothesis provided by Kessy Abarenkov of UNITE. See RDP's technical report Comparison of Three Fugal ITS Reference Sets for detailed analysis.

How to cite Classifier? Wang, Q, G. M. Garrity, J. M. Tiedje, and J. R. Cole. 2007. Naïve Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy. Appl Environ Microbiol. 73(16):5261-7. The RDP Classifier publication has been selected by Essential Science Indicators as the most-cited paper in a highlighted research area of microbiology. It's featured in December 2011 Science Watch.

NOTE: The classifier requires a sequence with at least 50 bases to get a good classification result. The number of query sequences can be submitted online is limited to ***100000***. For larger dataset, use the command-line version available as part of RDPTools package from github.com/rdpstaff or SourceForge.

Help topics: Partial sequences with length shorter than 250 bps should use bootstrap cutoff 50%.

Please enter your sequences:

Running Jobs: 0; **Pending Jobs:** 0

Choose a gene: 16S rRNA training set 14

Did you know you can select sequences from *myRDP* and Hierarchy Browser to do classification?

Choose a file (unaligned format) to upload: Choose File No file chosen

Cut and paste sequence(s) (in Fasta, GenBank, or EMBL format):

http://rdp.cme.msu.edu/classifier/classifier.jsp

Submit a .fastq dataset and it's running…

https://rdp.cme.msu.edu/classifier/cl_status.jsp

Classifier :: Query Sequences Status

Running Jobs: 1
Pending Jobs: 0

Status: running

Current Time: Sun Apr 17 03:14:53 EDT 2016

Progress: 13% completed

refresh cancel

Download the alignment result

BROWSERS | CLASSIFIER | LIBCOMPARE | SEQMATCH | PROBE MATCH | FUNGENE | RDPIPELINE | SEQCART | TAXOMATIC | TREE BUILDER | ASSIGNGEN

Classifier :: Assignment Detail

[start over | hierarchy view | help]

Classifier: RDP Naive Bayesian rRNA Classifier Version 2.10, October 2014
Taxonomical Hierarchy: RDP 16S rRNA training set 14
Query File: ERR345422.fastq
Query Submit Date: Sun Apr 17 03:16:39 EDT 2016

Lineage *(click to return to particular node)*:
Root (30442)

Assignment Detail (for Root with Confidence threshold: 80%): download allrank result | download fixrank result

ERR345422.1	Root[100%] Bacteria[99%] "Bacteroidetes"[96%] "Bacteroidia"[92%] "Bacteroidales"[92%] "Porphyromonadaceae"[88%] Butyricimonas[46%]
ERR345422.2	Root[100%] Bacteria[100%] Firmicutes[100%] Bacilli[100%] Lactobacillales[100%] Lactobacillaceae[96%] Lactobacillus[71%]
ERR345422.3	Root[100%] Bacteria[100%] "Bacteroidetes"[91%] "Bacteroidia"[59%] "Bacteroidales"[59%] "Porphyromonadaceae"[37%] Barnesiella[9%]
ERR345422.4	Root[100%] Bacteria[95%] "Bacteroidetes"[89%] "Bacteroidia"[67%] "Bacteroidales"[67%] "Porphyromonadaceae"[48%] Dysgonomonas[11%]
ERR345422.5	Root[100%] Bacteria[99%] "Bacteroidetes"[95%] "Bacteroidia"[61%] "Bacteroidales"[61%] "Porphyromonadaceae"[57%] Barnesiella[8%]
ERR345422.6	Root[100%] Bacteria[100%] "Bacteroidetes"[91%] "Bacteroidia"[60%] "Bacteroidales"[60%] "Porphyromonadaceae"[38%] Barnesiella[9%]
ERR345422.7	Root[100%] Bacteria[100%] "Bacteroidetes"[88%] "Bacteroidia"[69%] "Bacteroidales"[69%] "Porphyromonadaceae"[50%] Barnesiella[12%]
ERR345422.8	Root[100%] Bacteria[100%] Firmicutes[98%] Bacilli[98%] Lactobacillales[98%] Lactobacillaceae[87%] Lactobacillus[84%]
ERR345422.9	Root[100%] Bacteria[97%] "Bacteroidetes"[82%] "Bacteroidia"[70%] "Bacteroidales"[70%] "Porphyromonadaceae"[54%] Barnesiella[16%]
ERR345422.10	Root[100%] Bacteria[99%] "Bacteroidetes"[94%] "Bacteroidia"[88%] "Bacteroidales"[88%] "Porphyromonadaceae"[67%] Dysgonomonas[19%]
ERR345422.11	Root[100%] Bacteria[100%] "Bacteroidetes"[99%] "Bacteroidia"[81%] "Bacteroidales"[81%] "Porphyromonadaceae"[72%] Dysgonomonas[32%]
ERR345422.12	Root[100%] Bacteria[100%] Firmicutes[100%] Clostridia[100%] Clostridiales[100%] Lachnospiraceae[100%] Lactonifactor[54%]
ERR345422.13	Root[100%] Bacteria[100%] "Bacteroidetes"[87%] "Bacteroidia"[74%] "Bacteroidales"[74%] "Porphyromonadaceae"[54%] Barnesiella[14%]

fixrank_ERR345422.fastq_classified.txt

```
Classifier: RDP Naive Bayesian rRNA Classifier Version 2.10, October 2014
Taxonomical Hierarchy: RDP 16S rRNA training set 14
Query File: ERR345422.fastq
Submit Date: Thu Oct 29 17:56:13 EDT 2015
Confidence threshold (for classification to Root ONLY): 80%
Symbol +/- indicates predicted sequence orientation

ERR345422.1;+;Bacteria;99%;"Bacteroidetes";96%;"Bacteroidia";92%;"Bacteroidales";92%;"Porphyromonadaceae";88%;Butyricimonas;46%
ERR345422.2;+;Bacteria;100%;Firmicutes;100%;Bacilli;100%;Lactobacillales;100%;Lactobacillaceae;96%;Lactobacillus;71%
ERR345422.3;+;Bacteria;100%;"Bacteroidetes";91%;"Bacteroidia";59%;"Bacteroidales";59%;"Porphyromonadaceae";36%;Barnesiella;8%
ERR345422.4;+;Bacteria;95%;"Bacteroidetes";89%;"Bacteroidia";67%;"Bacteroidales";67%;"Porphyromonadaceae";48%;Dysgonomonas;11%
ERR345422.5;+;Bacteria;99%;"Bacteroidetes";95%;"Bacteroidia";61%;"Bacteroidales";61%;"Porphyromonadaceae";57%;Barnesiella;8%
ERR345422.6;+;Bacteria;100%;"Bacteroidetes";91%;"Bacteroidia";59%;"Bacteroidales";59%;"Porphyromonadaceae";37%;Barnesiella;9%
ERR345422.7;+;Bacteria;100%;"Bacteroidetes";88%;"Bacteroidia";69%;"Bacteroidales";69%;"Porphyromonadaceae";50%;Barnesiella;12%
ERR345422.8;+;Bacteria;100%;Firmicutes;98%;Bacilli;98%;Lactobacillales;98%;Lactobacillaceae;87%;Lactobacillus;84%
ERR345422.9;+;Bacteria;97%;"Bacteroidetes";82%;"Bacteroidia";71%;"Bacteroidales";71%;"Porphyromonadaceae";55%;Barnesiella;15%
ERR345422.10;+;Bacteria;99%;"Bacteroidetes";94%;"Bacteroidia";87%;"Bacteroidales";87%;"Porphyromonadaceae";64%;Dysgonomonas;18%
ERR345422.11;+;Bacteria;100%;"Bacteroidetes";99%;"Bacteroidia";81%;"Bacteroidales";81%;"Porphyromonadaceae";72%;Dysgonomonas;32%
ERR345422.12;+;Bacteria;100%;Firmicutes;100%;Clostridia;100%;Clostridiales;100%;Lachnospiraceae;100%;Lactonifactor;53%
ERR345422.13;+;Bacteria;100%;"Bacteroidetes";87%;"Bacteroidia";74%;"Bacteroidales";74%;"Porphyromonadaceae";53%;Barnesiella;14%
ERR345422.14;+;Bacteria;100%;Firmicutes;100%;Bacilli;100%;Lactobacillales;100%;Lactobacillaceae;96%;Lactobacillus;71%
ERR345422.15;+;Bacteria;95%;Firmicutes;92%;Clostridia;92%;Clostridiales;92%;Lachnospiraceae;91%;Dorea;15%
ERR345422.16;+;Bacteria;100%;Firmicutes;94%;Clostridia;94%;Clostridiales;94%;Lachnospiraceae;92%;Dorea;5%
ERR345422.17;+;Bacteria;100%;Firmicutes;76%;Clostridia;75%;Clostridiales;75%;Lachnospiraceae;73%;Johnsonella;39%
ERR345422.18;+;Bacteria;100%;Firmicutes;94%;Clostridia;94%;Clostridiales;94%;Lachnospiraceae;92%;Dorea;5%
ERR345422.19;+;Bacteria;100%;Firmicutes;38%;Bacilli;38%;Lactobacillales;35%;Lactobacillaceae;18%;Paralactobacillus;3%
ERR345422.20;+;Bacteria;100%;Firmicutes;98%;Clostridia;98%;Clostridiales;97%;Lachnospiraceae;86%;Lachnobacterium;36%
ERR345422.21;+;Bacteria;100%;"Bacteroidetes";97%;"Bacteroidia";86%;"Bacteroidales";86%;"Porphyromonadaceae";79%;Butyricimonas;39%
ERR345422.22;+;Bacteria;98%;"Bacteroidetes";92%;"Bacteroidia";69%;"Bacteroidales";69%;"Porphyromonadaceae";53%;Parabacteroides;15%
ERR345422.23;+;Bacteria;100%;Firmicutes;93%;Bacilli;91%;Lactobacillales;71%;Lactobacillaceae;51%;Lactobacillus;47%
ERR345422.24;+;Bacteria;98%;"Bacteroidetes";63%;"Bacteroidia";36%;"Bacteroidales";36%;Marinilabiliaceae;8%;Thermophagus;4%
ERR345422.25;+;Bacteria;100%;Firmicutes;100%;Bacilli;100%;Lactobacillales;100%;Lactobacillaceae;99%;Lactobacillus;85%
ERR345422.26;+;Bacteria;100%;Firmicutes;100%;Bacilli;100%;Lactobacillales;98%;Lactobacillaceae;91%;Lactobacillus;85%
ERR345422.27;+;Bacteria;100%;"Bacteroidetes";91%;"Bacteroidia";59%;"Bacteroidales";59%;"Porphyromonadaceae";38%;Barnesiella;9%
ERR345422.28;+;Bacteria;100%;"Bacteroidetes";95%;"Bacteroidia";69%;"Bacteroidales";69%;"Porphyromonadaceae";49%;Dysgonomonas;16%
ERR345422.29;+;Bacteria;100%;"Bacteroidetes";97%;"Bacteroidia";55%;"Bacteroidales";55%;"Porphyromonadaceae";47%;Barnesiella;8%
ERR345422.30;+;Bacteria;100%;"Bacteroidetes";94%;"Bacteroidia";81%;"Bacteroidales";81%;"Porphyromonadaceae";62%;Dysgonomonas;40%
ERR345422.31;+;Bacteria;100%;"Bacteroidetes";89%;"Bacteroidia";71%;"Bacteroidales";71%;"Porphyromonadaceae";67%;Dysgonomonas;41%
ERR345422.32;+;Bacteria;99%;Firmicutes;87%;Clostridia;87%;Clostridiales;87%;Lachnospiraceae;84%;Clostridium XlVa;39%
ERR345422.33;+;Bacteria;99%;"Bacteroidetes";96%;"Bacteroidia";81%;"Bacteroidales";81%;"Porphyromonadaceae";67%;Dysgonomonas;30%
ERR345422.34;+;Bacteria;100%;"Bacteroidetes";95%;"Bacteroidia";89%;"Bacteroidales";89%;"Prevotellaceae";86%;Prevotella;69%
```