Ublastx$_{s}tageoneDocumentation$

Release latest

Jan 17, 2018

Contents

We have simplified the running process of ublastx_stage_one. We have made a step by step video about how to use ARGs-OAP platform, hopefully, this video will guid new users to go through the process within ten minutes. The address is: https://www.youtube.com/watch?v=PCr1ctXvZPk

A mirror site was added in Shenzhen China for mainland China users to solve the slow data uploading problem [SUSTC-MIRROR-ARGS-OAP](http://smile.sustc.edu.cn:8080/)

New release of Ublastx Version 1.2

1. adding a method to obtain microbial community structure from the shotgun metagenomics data set. 2. adding copy number correction using Copyrighter database and normalize ARGs abundance by cell number.

Detail introduction of copy number correction can be referred to [Transform ARGs abundance against cell number](https://github.com/biofuture/Ublastx_stageone/wiki/Transform-ARGs-abundance-against-cell-number)

There are some questions raised by users, please refer to the [FAQ](https://github.com/biofuture/Ublastx_stageone/wiki/FAQ) for details.

To run Ublastx, users should download the stage one source code into local computer system (Unix/Linux) and upload the generated files for stage two onto our Galaxy analysis platform (http://smile.hku.hk/SARGs).

CHAPTER 1

What does Ublastx do:

1. Fast environmental searching of antibiotic resistant gene in multiple metagenomics data sets; the ARGs abundance can be normalized to cell number
2. Generate mother table of type and sub-type level ARGs of users' samples and a merged sub-type level mother table
3. Generate a PcoA of users samples with other typical environment samples such as human gut, ocean and sediment to show the relationship of user concerned samples with already sequenced environment.

CHAPTER 2

clone source code into local computer

git clone https://github.com/biofuture/Ublastx_stageone.git

CHAPTER 3

Prepare the meta-data file of your samples

To run the stage one pipeline, users need to prepare relative meta-data.txt file and put all the pair-end fastq file into one directory Example of meta-data file **meta-data.txt** Tips: * You need keep the first and second column's name as SampleID and Name * The SampleID are required to be numbers counting from 1 to 2 to 3 etc. * Category is the classification of your samples into groups and we will colored your samples in PcoA by this informaton * The meta-data table should be separated by tabular for each of the items * The Name of each sample should be the fastq file names for your pair-end Illumina sequencing data, your fastq files will automatically be recognized by Name_1.fq and Name_2.fq, so you need to keep the name consistent with your fq file name. (if you files are end with .fastq or .fasta, you need to change them to end with .fq or .fa)

Please make sure the meta-data file is pure txt format, if you edit the file under windows, using nodepad++ and check the end of each line by cliking View-> Show Symbol -> Show All Characters. If the line is end up with CRLF, please remove the CR by replace r to nothing in the replace dialogue frame

SampleID | Name | Category ———|———|———

1 | STAS | ST 2 | SWHAS104 | SWH

CHAPTER 4

Prepare database and usearch

SARG Database and 32 bit usearch is avaliable in DB/ and bin/ directory, respectively. **Users donot need to download CARD and ARDB anymore!!**

CHAPTER 5

Stage one pipeline

When meta-data.txt and database files are prepared, then put all your fastq files into one directory in your local system (notice the name of your fastq files should be Name_1.fq and Name_2.fq). your can give -h to show the help information. Examples could be found in source directory example, in example directory run test:

nohup ../ublastx_stage_one -i inputfqs -o testoutdir -m meta-data.txt -n 2

> **Usage: ./ublastx_stage_one -i <Fq input dir> -m <Metadata_map.txt> -o <output dir> -n [number of threads] -f [fa|fq] -z**
> -i Input files directory, required -m meta data file, required -o Output files directory, default current directory -n number of threads used for usearch, default 1 -f the format of processed files, default fq -z whether the fq files were .gz format, if -z, then firstly gzip -d, default(none) -c This option fulfill copy number correction by Copywriter database to transfrom 16S information into cell number [direct searching hyper variable region database by usearch; default 1] -h print this help information

This step will search reads against SARG databbase and 16S greengene non-redundant 85 OTUs database to identify potential ARG reads and 16S reads. This step will generate searching results files for each fastq. This step also obtain the microbial community structure information of samples by searching against hyper-variable region database, and then perform copy number correction using Copyrighter copy number database (release date) to finally estimate the cell number of samples.

The results are in testoutdir/, it looks like this:

> extracted.fa STAS_2.16s SWHAS104.16s_hyperout.txt meta_data_online.txt STAS_2.us SWHAS104_1.us STAS_1.16s STAS.extract_1.fa SWHAS104_2.16s STAS.16s_1v6.us STAS.extract_2.fa SWHAS104_2.us STAS.16s_2v6.us SWHAS104_1.16s SWHAS104.extract_1.fa STAS.16s_hvr_community.txt SWHAS104.16s_1v6.us SWHAS104.extract_2.fa STAS.16s_hvr_normal.copy.txt SWHAS104.16s_2v6.us ublastx_bash_Mon-Feb-1-16:20:59-2016.sh STAS.16s_hyperout.txt SWHAS104.16s_hvr_community.txt STAS_1.us SWHAS104.16s_hvr_normal.copy.txt

The **extracted.fa** and **meta_data_online.txt** are two files needed for ublastx_stage_two analysis. The STAS.16s_hvr_community.txt is the microbial community of sample STAS and STAS.16s_hvr_normal.copy.txt is the averagely copy number of the microbial community after CopyRighter database correction.

The meta-data-online.txt looks like this

SampleID | Name | Category | #ofreads | #of16S| **#ofCell** ———|——|————|————-|——-|———

1 | STAS | ST |200000 | 10.1 | 4.9 2 | SWHAS104 | SWH |200000 | 9.7 | 4.1

CHAPTER 6

Stage two pipeline on Galaxy system and download results

Go to http://smile.hku.hk/SARGs and using the module ARG_OAP.

1. Using **ARG_OAP -> Upload Files** module to upload the extracted fasta file and meta_data_online.txt file generated in stage one into Galaxy
2. Click **ARG_OAP** and **Ublast_stagetwo**, select your uploaded files
3. For "Column in Metadata:" chose the column you want to classify your samples (default: 3)

Click **Execute** and you can find four output files for your information

After a while or so, you will notice that their are four files generated for your information.

File 1 and 2: PcoA figures of your samples and other environment samples generated by ARGs abundance matrix normalization to 16s reads number and cell number **File 3 and 4**: Other tabular mother tables which including the profile of ARGs type and sub type information, as long as with other environment samples mother table. File3 results of ARGs abundance normalization aganist 16S reads number; File 4 results of ARGs abundance normalization aganist cell number

This tools only provide the required scripts for ARGs-OAP pipeline (Bioinformatics (2016) doi: 10.1093/bioinformatics/btw136).

This pipeline is distributed in the hope to achieve the aim of management of antibiotic resistant genes in envrionment, but WITHOUT ANY WARRANTY; without even the implied warranty of MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE.This pipeline is only allowed to be used for non-commercial and academic purpose.

The copyrights of the following tools/databases which could be used in this pipeline belong to their original developers. The user of this pipeline should follow the guideline and regulations of these tools/database which could be found at the websites of their developers.

1. Usearch: (http://www.drive5.com/usearch/)
2. Copyrighter: (http://microbiomejournal.biomedcentral.com/articles/10.1186/2049-2618-2-11)
3. Greengenes: (http://greengenes.lbl.gov/cgi-bin/nph-index.cgi)

Please check the above websites for details of these tools/databases.