# MCRL –

## METAGENOME CLUSTERING BY REFERENCE LIBRARY

# User Guide

VERSION 2.1.6 (beta)

Arbel D. Tadmor[1,2]

*[1]TRON - Translational Oncology at the University Medical Center of the Johannes Gutenberg, University Mainz, 55131, Mainz, Germany, [2]Department of Biochemistry and Molecular Biophysics, California Institute of Technology, Pasadena, CA 91125, USA*

April 2019

# Table of contents

# 1. Installation instructions

## 1.1 System requirements

1. Matlab 2014a or later version
2. Bioinformatics toolbox
3. Optional: Parallel Computing toolbox v4.2 or higher
4. Installation of MCRL requires an internet connection

Installation is platform independent. Utilizing more than one processor requires the Parallel Computing toolbox. Analysis of a ~25Mb metagenome using a single processor can take up to 24h run time.

A license agreement is provided with the installation of this program (see `License.txt`).

Additional requirements:

- Please avoid the "+", "-" characters in the definition line of metagenome FASTA records.
- MCRL upon installation downloads and assembles the most recent viral RefSeq database as a reference library of known genes. MCRL can be executed, however, using any reference library and is not limited to viral reference libraries. The only requirement is that the genes in the reference library be translated into amino acids.

## 1.2 Installation procedure

1. Download the compressed sources for MCRL and extract them locally
2. Start up Matlab
3. Change directories in Matlab to the `bin` folder of MCRL
4. Run MCRL by typing `MCRL_EXE` in the Matlab command prompt
5. Click “Automatic installation (recommended)”
6. The program automatically downloads and installs BLAST version 2.2.22+ from NCBI for your specific platform[1]. This process may take several minutes. Note: you may be

[1]If blast 2.2.22+ is already installed on your computer this step is automatically skipped. Other blast versions that may be installed on your computer are ignored by MCRL.

promoted by your operating system to allow access for installation. You must manually accept in order for the installation to proceed. Once the installation of BLAST starts click 'next' and accept all of the default entries. Once blast is locally installed the program proceeds to download and assemble the latest release of the viral RefSeq database as a reference library. Installation takes approximately 10 minutes.

7. Once the main interface of MCRL loads you may start using MCRL.

To check that the installation was successful and to get familiar with MCRL try running MCRL on the default demo metagenome by clicking the "Run BLAST and MCRL" button.

### 1.3 Troubleshooting

1. It is recommended to have administrator/root privileges when installing MCRL.
2. If there is a problem downloading the `blast` application then make sure the firewall in the background does not block the ftp port (e.g., temporarily disable the firewall). Alternatively blast v2.2.22+ can be downloaded manually from the NCBI website: ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/2.2.22/, selecting the software version appropriate for your OS. Installation instructions for blast can be found here: ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/2.2.22/user_manual.pdf
   Once blast is installed, start MCRL, click "Locate sources on computer" and locate the `bin` folder of the blast installation.
3. If there is a problem running blast there may be a previous installation of blast from another utility which is interfering with MCRL. In you are running Windows and Windows is installed on the c:\ drive, check in the Windows folder on your computer if there is a file called `ncbi.ini`. If there is temporarily rename it.

### 1.4 Assembling a RefSeq reference library (optional)

The RefSeq database is continuously updated. MCRL automatically downloads and assembles the most up-to-date viral RefSeq database from the NCBI ftp server during installation. RefSeq files (FASTA and GenPept files) and release notes are saved to the `RefSeq_database` folder. The viral RefSeq database serves as the default reference

library for MCRL. If this database suits your requirements you do not need to proceed reading this section and can begin using MCRL.

To update the viral RefSeq database at any time after MCRL has been installed to the most recent release erase the `RefSeq_file_name_header.mat` file in the `msrc` folder and run MCRL. It is possible to manually download from NCBI other taxonomic or logical groups or previous versions of the viral RefSeq database as a reference library. This section explains how to download and assemble from the NCBI ftp site a RefSeq database.

FASTA files on the NCBI website have an `*.faa` extension and GenPept files have a `*.gpff` extension. Although MCRL can run with just a FASTA file, to get the most out of MCRL it is recommended to download both the FASTA file and its corresponding GenPept file. Since the databases are large these files are typically separated into smaller files. Downloading RefSeq files can be performed manually or using an ftp software by logging to the ftp site ftp://ftp.ncbi.nih.gov as an anonymous user and browsing to the refseq/release/ directory. Combining all RefSeq files into a single file can be performed as follows:

1. Confirm that MCRL's `Combine_RefSeq_files` folder contains only the file `util_concat_all_files_in_folder.m`
2. Copy all the files you wish to combine (and only these files) into the above folder (e.g., all `.faa` files)
3. In Matlab change directories to the `Combine_RefSeq_files` folder
4. Run the source code `util_concat_all_files_in_folder` in Matlab

This utility will combine all the files in the given directory expect for the Matlab source into a single file named `combined_all`. After the program finishes you may rename this file and give it the appropriate extension (e.g., `.faa` or `.gpff`).

### 1.5 MCRL folders

The following folders are installed with MCRL:

| | |
|---|---|
| `bin` | source that runs MCRL |
| `blast` | BLAST installation folder |
| `Combine_RefSeq_files` | MCRL utility (see §1.4) |
| `data` | folder to store metagenome FASTA files |
| `msrc` | MCRL source code |
| `output` | folder to which output files are written (see §2.2) |
| `RefSeq_database` | Reference library files (e.g., `.faa` and `.gpff` RefSeq files) |
| `tmp` | temporary files generated by MCRL |

### 1.6 Known problems

NCBI blast 2.2.22+ appears to have an intrinsic bug where BLASTing a single FASTA record against a large database may result in a slightly different output compared to the case where the FASTA record is embedded in a FASTA file. This can lead to minor differences in the BLAST output depending on the number of cpus used, as the parallelization requires splitting the RefSeq FASTA file into smaller files. This bug appears to be rare, affecting only about 1 out of 20000 records.

# 2. The MCRL algorithm

## 2.1 Definitions of terms

We define certain mathematical terms in the context of the MCRL algorithm that are useful for explaining the algorithm. A gene in the reference library provided to MCRL is referred to as a "known reference gene" (KRG). Records in the metagenome provided to MCRL are referred to as "metagenome gene objects" (MGOs). A "signature" of a KRG in a metagenome is defined as the list of all MGOs yielding E values $<10^{-3}$ when BLASTed against the given KRG. Two KRGs are defined to be "related" if the overlap between the members of their "signatures" exceeds 50%. Finally, the "prevalence" of a KRG in the given metagenome is defined as the number members in its "signature".

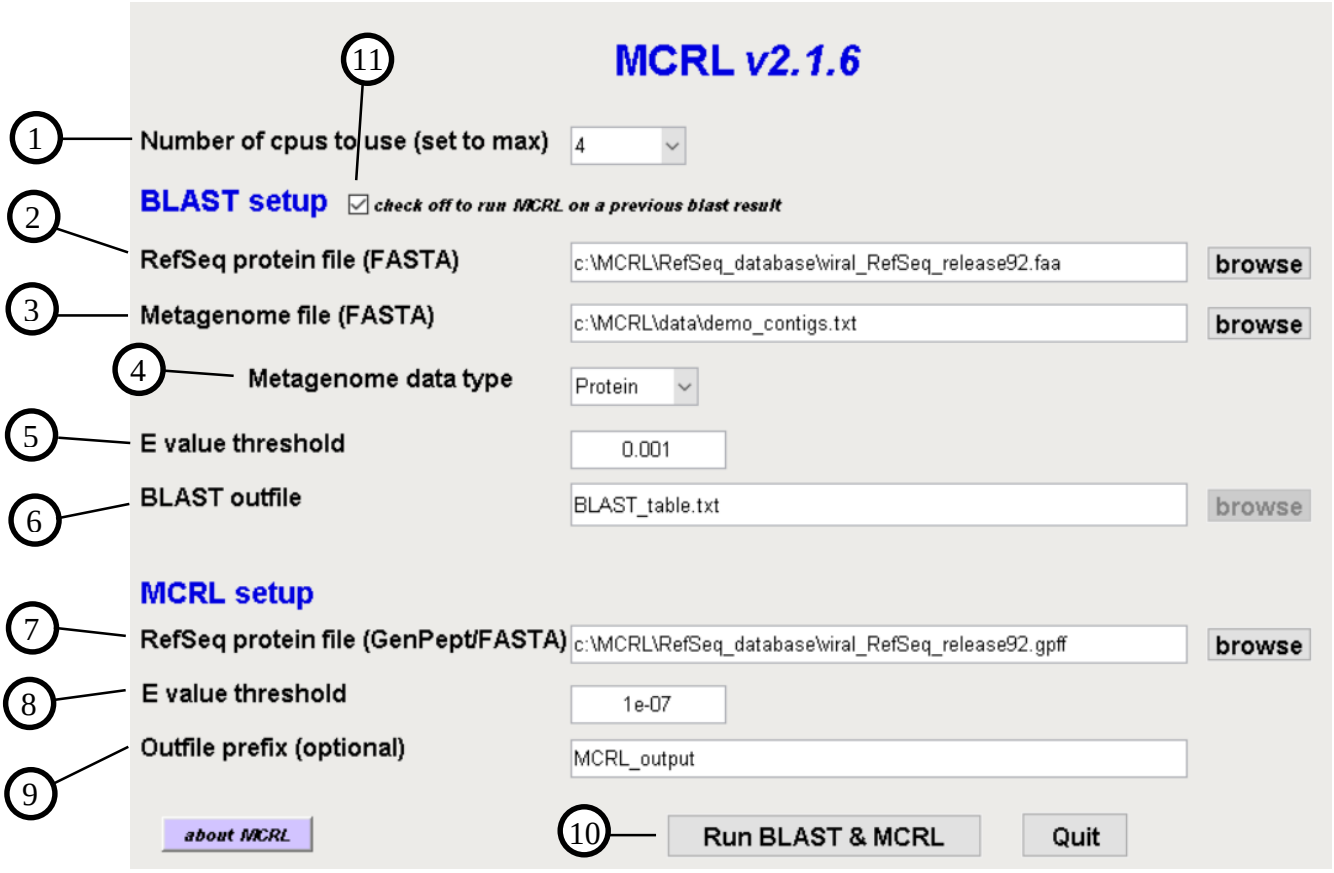
## 2.2 Overview of the MCRL algorithm

MCRL applies an iterative compression algorithm to remove "related" KRGs with respect to a given metagenome. Initially MCRL will report all KRGs in the reference library. In each subsequent iteration step MCRL determines for each KRG the group of "related" KRGs, and the KRG with the lowest E value of this group will be declared to represent this group. The number of declared KRGs reported by MCRL will be reduced from one iteration to the next when two or more different KRGs declare the same KRG, in which case this KRG is declared only once. The algorithm terminates when each KRG declares itself (when there are no other related KRGs the given KRG always declares itself because by definition each KRG is always "related" to itself). The final output of MCRL is this non-redundant list of KRGs (`output5_NonredundantTable.txt` output file). Only KRGs whose lowest E value when BLASTed against the metagenome is below a default E value threshold of $E_{th} = 10^{-7}$ are considered. Each KRG reported by MCRL represents a "gene group", which is defined as the list of all KRGs "related" to the KRG reported by MCRL.

## 3. Software operation

### 3.1 First time run

**Figure 1.** Main interface of MCRL.




**Computing resource parameter:**

**(1)** If the pull-down menu is activated this indicates that the Matlab Parallel Computing toolbox is installed on your computer. Select the number of cpus to be used by MCRL. The default is set to the number of cpus detected on your computer. If the Parallel Computing toolbox is not installed the pull-down menu is inactivated and MCRL will utilize by default a single cpu.

**BLAST input parameters:**

**(2)** Path and name of reference library (a protein FASTA file). Records in the reference library are called known reference genes (KRGs). For demonstration purposes the `RefSeq_database` folder upon installation contains the most recent viral RefSeq FASTA file (release XX): `viral_RefSeq_releaseXX.faa`.

**(3)** Path and name of metagenome (a FASTA file). For demonstration purposes the file `demo_contigs.txt` is provided in the `data` folder.

**(4)** Data type of the metagenome: protein or nucleotide.

**(5)** E value threshold for BLAST (default is 0.001). Alignments yielding E values higher than this threshold will be discarded.

**(6)** Name of BLAST output file that will be generated in the `output` folder. This file contains the BLAST hit table generated by the BLAST executable.

**MCRL input parameters:**

**(7)** GenPept file corresponding to the FASTA file entered in (2). If this file is not available the FASTA file entered in (2) can be entered instead. For demonstration purposes the `RefSeq_database` folder contains the GenPept file `viral_RefSeq_releaseXX.gpff`, which corresponds to the RefSeq FASTA file `viral_RefSeq_releaseXX.faa`.

**(8)** E value threshold ($E_{th}$) for the MCRL algorithm (default is $10^{-7}$). RefSeq genes yielding E values higher than this threshold when BLASTed against all MGOs in the metagenome will be ignored.

**(9)** Optional string to add to the file name of all MCRL output files.

**(10)** Click "Run BLAST & MCRL" to run MCRL. Progress can be viewed in the Matlab command window. When MCRL finishes it will open the report (`params`) file.

**(11)** To reanalyze a previous BLAST analysis check off the check box (see §2.4).

### 3.2 Output files generated by MCRL

In addition to the BLAST output table MCRL generates six output files in the `output` folder:

1. `<blast file name><string>_MCRL_output0_params.txt`

2. <blast file name><string>_MCRL_output1_AllGenes.txt
3. <blast file name><string>_MCRL_output2_AllGenesFilt.txt
4. <blast file name><string>_MCRL_output3_RelatedGenes.txt
5. <blast file name><string>_MCRL_output4_ShortTable.txt
6. <blast file name><string>_MCRL_output5_NonredundantTable.txt

A detailed description of the output files is provided in §3. Output files are tab delimited and can be viewed in Excel or a similar application.

**Final MCRL output**

The `NonredundantTable` file lists all KRGs after iterative compression. The list is sorted by the "prevalence" of each KRG such that the first KRG in the list has the highest prevalence in the metagenome.

### 3.3 Post analysis

When KRGs are RefSeq genes and a GenPept file is provided, the biological function of each KRG in the `NonredundantTable` file can be ascertained either from the "RefSeq gene definition" field or from the "GenPept Features" field. The amino acid sequence of the KRG is also provided and can be manually BLASTed against public databases to obtain further information about the KRG.

### 3.4 Subsequent runs of MCRL

MCRL can be executed in two modes – a BLAST analysis followed by a MCRL analysis (the default mode) or only a MCRL analysis (analyzing a previous BLAST analysis). The latter mode is useful for reanalyzing a previous BLAST run using a different $E_{th}$ threshold. To toggle between these two modes click the check box (option 10 in Fig. 1) thereby inactivating all fields related to the BLAST run, and allowing the user to select a previous BLAST output file generated by MCRL.

## 4. Description of output files

**`Output0_params`**

This file contains:

- Statistics on the run such as execution time, date and time of run, version of MCRL used, etc.
- Commands used for BLAST analysis (if appropriate)
- A summary of the parameters/file names used to run MCRL
- List of output files generated by MCRL

**`Output1_AllGenes`**

Information parsed from the BLAST output file. The file lists all KRGs that passed the BLAST E value threshold (default $10^{-3}$). Each KRG is followed by the list of all MGOs that passed the above BLAST E value threshold. The following additional information is provided for the MGO that yielded the lowest E value:

1. **Index**
   Counter of KRG in table.
2. **RefSeq gene**
   KRG name as it appears in the (RefSeq) FASTA file definition line extracted by the BLAST application.
3. **RefSeq gene definition**
   The 'Definition' field for the KRG as it appears in the FASTA file, or if a GenPept file was provided, the KRG definition as it appears in the GenPept file definition line (n/a for the `AllGenes` output file). The definition field is defined as follows: "Brief description of sequence; includes information such as source organism, gene name/protein name, or some description of the sequence's function (if the sequence is non-coding)" [3].
4. **Metagenome gene object ID with lowest E value**
   MGO that yielded the lowest E value for the given KRG.
5. **# of metagenome gene objects similar to this RefSeq gene**

"Prevalence" of KRG in metagenome, defined as the number of MGOs that yielded an E value $\leq 10^{-3}$ when aligned against the given KRG.

6. **% identity**

   Percent identity of the KRG and the MGO that yielded the lowest E value.

7. **# of identical amino acids**

   Number of amino acids that were identical when aligning the given KRG and the MGO that yielded the lowest E value.

8. **E value**

   E value of the alignment between the given KRG and the MGO that yielded the lowest E value.

9. **Alignment length (amino acids)**

   Length of alignment in amino acids between the MGO yielding the lowest E value and the given KRG.

10. **RefSeq gene length (amino acids)**

    Number of amino acids encoded by the KRG (n/a for the `AllGenes` output file).

11. **% of RefSeq gene length aligned**

    The percent of the KRG length that appears in the alignment, i.e., the ratio of (9) and (10) multiplied by 100 (n/a for the `AllGenes` output file).

12. **aa sequence**

    The amino acid sequence of the KRG (n/a for the `AllGenes` output file).

**`Output2_AllGenesFilt`**

Same as the `Output1_AllGenes` file but showing only KRGs whose lowest E value is equal to or lower than $E_{th}$.

**`Output3_RelatedGenes`**

List of all KRGs after removing related KRGs at the end of the first iteration of MCRL (i.e., prior to starting the iterative compression algorithm). Following each reported KRG entry is the list of KRGs related to the reported KRG. The list of reported KRGs are sorted by their "prevalence" in the metagenome (highest first). The following additional information is provided for each KRG:

The following additional information is provided for each KRG:

1. **Index** (see above)
2. **RefSeq gene** (see above)
3. **RefSeq gene definition** (see above)
4. **Min % of shared metagenome gene objects**
   The overlap between the MGO list corresponding to the given KRG and the MGO list of the KRG reported by MCRL (in units of percent).
5. **# of metagenome gene objects similar to this RefSeq gene** (see above)
6. **% identity** (see above)
7. **# of identical amino acids** (see above)
8. **E value** (see above)
9. **Alignment length (amino acids)** (see above)
10. **RefSeq gene length (amino acids)** (see above)
11. **% of RefSeq gene length aligned** (see above)
12. **GenPept Features**
    KRG features field as it appears in the GenPept file.

**`Output4_ShortTable`**

Same list of KRGs as appears in `Output3_RelatedGenes` only without showing related KRGs. This file contains the following fields:

1. **Index** (see above)
2. **RefSeq gene** (see above)

3. **Metagenome gene object ID with lowest E value** (see above)
4. **# of metagenome gene objects similar to this RefSeq gene** (see above)
5. **tot # of metagenome gene objects associated with this RefSeq gene group**

   "Prevalence" of the given KRG including all its related KRGs.
6. **# of related RefSeq genes**

   Number of KRGs related to the given KRG.
7. **% identity** (see above)
8. **# of identical amino acids** (see above)
9. **E value** (see above)
10. **Alignment length (amino acids)** (see above)
11. **RefSeq gene length (amino acids)** (see above)
12. **% of RefSeq gene length aligned** (see above)
13. **aa sequence** (see above)
14. **RefSeq gene definition** (see above)
15. **GenPept GenBank division**

    GenBank division field for the given KRG as it appears in the GenPept file.

    The GenBank division field is defined as follows: "The GenBank division to which a record belongs is indicated with a three letter abbreviation. The GenBank database is divided into 18 divisions:

    1. PRI - primate sequences
    2. ROD - rodent sequences
    3. MAM - other mammalian sequences
    4. VRT - other vertebrate sequences
    5. INV - invertebrate sequences
    6. PLN - plant, fungal, and algal sequences
    7. BCT - bacterial sequences
    8. VRL - viral sequences
    9. PHG - bacteriophage sequences
    10. SYN - synthetic sequences
    11. UNA - unannotated sequences
    12. EST - EST sequences (expressed sequence tags)

13. PAT - patent sequences
14. STS - STS sequences (sequence tagged sites)
15. GSS - GSS sequences (genome survey sequences)
16. HTG - HTG sequences (high-throughput genomic sequences)
17. HTC - unfinished high-throughput cDNA sequencing
18. ENV - environmental sampling sequences" [3].

**16. GenPept molecule type**

Molecule type corresponding to the given KRG as it appears in the GenPept file. The molecule type field is defined as follows: "The type of molecule that was sequenced. Each GenBank record must contain contiguous sequence data from a single molecule type. The various molecule types are described in the Sequin documentation and can include genomic DNA, genomic RNA, precursor RNA, mRNA (cDNA), ribosomal RNA, transfer RNA, small nuclear RNA, and small cytoplasmic RNA" [3].

**17. GenPept source**

Source field for the given KRG as it appears in the GenPept file. The source field is defined as follows: "Free-format information including an abbreviated form of the organism name, sometimes followed by a molecule type" [3].

**18. GenPept classification**

Organism field for the given KRG as it appears in the GenPept file. The organism filed is defined as follows: "The formal scientific name for the source organism (genus and species, where appropriate) and its lineage, based on the phylogenetic classification scheme used in the NCBI Taxonomy Database. If the complete lineage of an organism is very long, an abbreviated lineage will be shown in the GenBank record and the complete lineage will be available in the Taxonomy Database" [3].

**19. GenPept comments**

Comments field for the given KRG as it appears in the GenPept file. The comments field is defined as follows: "A comment identifying the RefSeq Status is provided for the majority of the RefSeq records. This comment may include information about the RefSeq status, collaborating groups, and the GenBank records(s) from which the RefSeq is derived. The RefSeq comment is not provided comprehensively in this

release. Additional comments are provided for some records to provide information about the sequence function, notes about the aspects of curation, or comments describing transcript variants" [4].

**20. GenPept Features**

Features field for the given KRG as it appears in the GenPept file. The features field is defined as follows**: "**Information about genes and gene products, as well as regions of biological significance reported in the sequence. These can include regions of the sequence that code for proteins and RNA molecules, as well as a number of other features.

**Source:** Mandatory feature in each record that summarizes the length of the sequence, scientific name of the source organism, and Taxon ID number. Can also include other information such as map location, strain, clone, tissue type, etc., if provided by submitter.

**Taxon:** A stable unique identification number for the taxon of the source oganism. A taxonomy ID number is assigned to each taxon (species, genus, family, etc.) in the NCBI Taxonomy Database.

**CDS:** "Coding sequence; region of nucleotides that corresponds with the sequence of amino acids in a protein (location includes start and stop codons." [3]

**Protein Names:** Protein names may be provided by a collaborating group, may be based on the Gene Name, or for some records, the curation process may identify the preferred protein name based on that associated with a specific EC number or based on the literature.

**Protein Products:** Signal peptide and mature peptide annotation is provided by propagation from the GenBank submission that the RefSeq is based on, when provided by a collaborating group, or when determined by the curation process.

**Domains:** Domains are computed by alignment to the NCBI Conserved Domain Database database for human, mouse, rat, zebrafish, nematode, and cow. The best hits are annotated on the RefSeq. For some records, additional functionally significant regions of the protein may be annotated by the curation staff. Domain annotation is not provided comprehensively at this time." [4]

**`Output5_NonredundantTable`**

List of all KRGs reported by MCRL after the iterative compression terminates. Since each reported KRG declares itself the column "# of related RefSeq genes" is always equal to 1 in this file. The list of KRGs is sorted by their "prevalence" in the metagenome (highest first). The format of this file is identical to that of `Output4_ShortTable`.

## 5. Citation

Tadmor A. D., Mahmoudabadi G., Foley H. B., Phillips R., Ubiquitous Phage Markers in Humans, 2019 (submitted)

Tadmor A. D., Ottesen E. A., Leadbetter J. R., Phillips R., Probing individual environmental bacteria for viruses by using microfluidic digital PCR. *Science* 333, 58-62 (2011)

## 6. References


1. Pruitt K, Tatusova T, Maglott D (2005) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Research 33: D501-D504
2. Altschul S, Madden T, Schaffer A, Zhang J, Zhang Z, et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research 25: 3389
3. http://www.ncbi.nlm.nih.gov/sitemap/samplerecord.html
4. RefSeq release notes ftp://ftp.ncbi.nih.gov/refseq/release/release-notes/