**Custom BLAST Databases**

BLAST on the NCBI public server is a fast and widely used tool. It can search a moderate number of query sequences with a single uploaded multi-fasta file, and it can target the search to a specific taxonomic group, genome, or many other interesting data subsets. However, there are situations when a bioinformatician needs to run BLAST on a local computer. If a very large number of query sequences are being tested, the public server will be too slow, or if a custom database is needed a local BLAST must be run. Note that BLAST is faster when it searches a large number of sequences as a database with a smaller number of query sequences.

The local command-line version of BLAST is called blast+ and it is freely available as source code and compiled binaries for Linux, Mac, and Windows systems (<https://www.ncbi.nlm.nih.gov/books/NBK52637/>). The linux system includes the program to make custom databses (makeblastdb), the standard blastn, blastp, blastx, tblastn search tools and the blastdbcmd tool to retrieve sequences from the database.

For this exercise, we will use local BLAST to make a quick search for a drug resistance gene (common in Staph infections) in a shotgun metagenomic data file. This is a model for an infectious disease surveillance system, which could easily be expanded to test a few hundred resistance genes on new samples brought in and sequenced on a daily basis.

The query gene is mecA, which provides methicillin and penicillin resistance to many *Staphylococcus* species. GenBank ID KC243783.1 (download from GenBank yourself)

I have provided a positive control, whole genome sequences from a known MRSA isolate in the file:

*/ifs/home/browns02/Class/MRSA\_wgs.fasta*

You can make your own local Blast database from this file without copying it, just use the full file pathname with the **makeblastdb** program

Like most Linux programs, if you type the program name with **'–h'**, you will get help information:

$ makeblastdb -h

USAGE

makeblastdb [-h] [-help] [-in input\_file] [-input\_type type]

-dbtype molecule\_type [-title database\_title] [-parse\_seqids]

[-hash\_index] [-mask\_data mask\_data\_files] [-gi\_mask]

[-gi\_mask\_name gi\_based\_mask\_names] [-out database\_name]

[-max\_file\_sz number\_of\_bytes] [-taxid TaxID] [-taxid\_map TaxIDMapFile]

[-logfile File\_Name] [-version]

DESCRIPTION

Application to create BLAST databases, version 2.2.28+

Use '-help' to print detailed descriptions of command line arguments

You only need to specify the dbtype (nucleotide = 'nucl'), the input data file in Fasta format, and a name for your new Blast database:

**makeblastdb –dbtype nucl –in /ifs/home/browns02/Class/MRSA\_wgs.fasta –out MRSA**

Now use emacs to create a Fasta formatted file for your mecA query sequence.

You can figure out for yourself how to use blastn to search the MRSA database for mecA.

1. Show your best match result here:
2. I have added an unknown data file, which represents shotgun sequences from a hospital patient who might have Staph infection. Test this data by making another BLAST database and then searching with mecA. What is the e-value of the best match?

1. Try using a protein based search (tblastx). Does this give better e-value? What can you say about using e-value thresholds as cutoff for a sequence search?