

# Wrap-up Exercise 6

# MGMapper input

## MGmapper 2.4 (MetaGenomics mapper)

Click to 'Show' available databases:

### Mapping options:

Mapping mode

Trimming of reads via cutadapt

Trimming parameters:

### Database selection:

Database id's for Best-mode mapping  Comma separated list of database id's NOT already selected for Full-mode mapping

Database id's for Full-mode mapping  Comma separated list of database id's NOT already selected for Best-mode mapping

Contig sequence assembly:

Fastq read alignment criteria:

Clade level post-processing:  Failed hits available for download as text tab separated files (negative.\*) and excel file (Results.negative.xlsx)

Abundance cutoff (%)  Minimum size normalized abundance: 100\*ReadCount/Size\*2 in PE-mode or 100\*ReadCount/Size in SE-mode

Unique reads ratio  Minimum ratio: Reads\_uniq/Reads

Max mismatch ratio  Maximum ratio: Edit\_distance/Nucleotides

Min read count  Minimum number of reads

Databases: 2 (bacteria), 3 (MetaHitAssembly), 4 (HumanMicrobiome), 5 (bacteria\_draft), 8 (virus)

Upload fastq file(s) as Single-end or Paired-end reads. In paired-end mode, only properly paired reads will be used and fastq files can NOT be interleaved. Files can be text or compressed as .gz or .Z files. Multiple files can be uploaded in both Single-end and Paired-end mode. Jobs are run in parallel using 4 cores. Max runtime 72 hours.

Below press 'Isolate File' to browse for FASTQ files to be uploaded.

Name	Size	Progress	Status

Max mismatch ratio = 0.1

# MGMapper output

[Home](#)[Services](#)[Instructions](#)[Output](#)

## Running time in single-end mode:

Job start	Job end
Mon Sep 18 10:48:55 2017	Mon Sep 18 11:03:50 2017

Table 1. Start and end times for running MGmapper.

## File statistics:

Counter	FileName	Reads in
1	outbreak_matrix_pathogen.fastq	109804
1	/home/data1/services/MGmapper/MGmapper-2.4/IO/2_18_9_2017_103_864_173355/uploads/outbreak_matrix_pathogen.fastq	109804

Table 2. Statistics for input fastq files. Last line sums up number of fastq files and total number of reads.

## Summary file statistics:

Biological relevant reads = 109804

Fastq files	Reads read	After cleaning
1	109804	109804

Table 3. The number 'After cleaning' represents reads with a biological origin, after removing potential Illumina PhiX control reads ([http://www.illumina.com/products/phix\\_control\\_v3.html](http://www.illumina.com/products/phix_control_v3.html)).

## Reference sequence databases:

Database	Version	No of sequence	No of nucleotides
Bacteria	20151124	6969	19137675153
MetaHitAssembly	20140111	41126	908004652
HumanMicrobiome	20140409	42928	2831377980
Bacteria_draft	20151124	137980	7972498520
Virus	20151124	6640	196732536

Table 4. Reference database statistics. 'Version' is the date at which reference sequences was updated.

# MGMapper output

## Overall mapping results before post-processing:

Mapping mode	Database	Percent mapped	No reads
Fullmode	notPhiX	100.00	109804
Bestmode	Bacteria	8.808	9671
Bestmode	MetaHitAssembly	0.008	9
Bestmode	HumanMicrobiome	0.027	30
Bestmode	Bacteria_draft	58.751	64511
Bestmode	Virus	0.049	54
-	Unmapped	32.357	35529

Table 5. The percentages are in relation to the number of reads ('After cleaning') as shown in Table 3..

### Statistics - Bestmode mapping (Max top 5 hits are shown below):

Database	Ref Seq	S_Abundance (%)	R_Abundance (%)	Size (bp)	Seq_count	Nucleotides	Covered positions	Coverage	Depth	ReadCount	Read U
Database	Ref Seq	S_Abundance (%)	R_Abundance (%)	Size (bp)	Seq_count	Nucleotides	Covered positions	Coverage	Depth	ReadCount	Read U
Database	Ref Seq	S_Abundance (%)	R_Abundance (%)	Size (bp)	Seq_count	Nucleotides	Covered positions	Coverage	Depth	ReadCount	Read U

Database	Ref Seq	S_Abundance (%)	R_Abundance (%)	Size (bp)	Seq_count	Nucleotides	Covered positions	Coverage	Depth	ReadCount	R
Bacteria_draft	AKVP01000102.1	1559.539	12.953	912	1	4112826	901	0.988	4509.678	14223	14
Bacteria_draft	AKVP01000085.1	1295.931	24.655	2089	1	9656894	1704	0.816	4622.735	27072	21
Bacteria_draft	AKVP01000087.1	818.636	13.122	1760	1	5842575	1760	1.000	3319.645	14408	14
Bacteria_draft	AKVP01000114.1	436.377	2.611	657	1	979948	518	0.788	1491.549	2867	21
Bacteria_draft	AKVP01000128.1	133.778	0.548	450	1	193722	450	1.000	430.493	602	61

			Virus <small>VI</small>	
Database	Ref Seq	S_Abundance (%)	Mismatches	Description
Virus	NC_001489.1	0.722	1969	Hepatitis A virus, complete genome

# Results in Excel

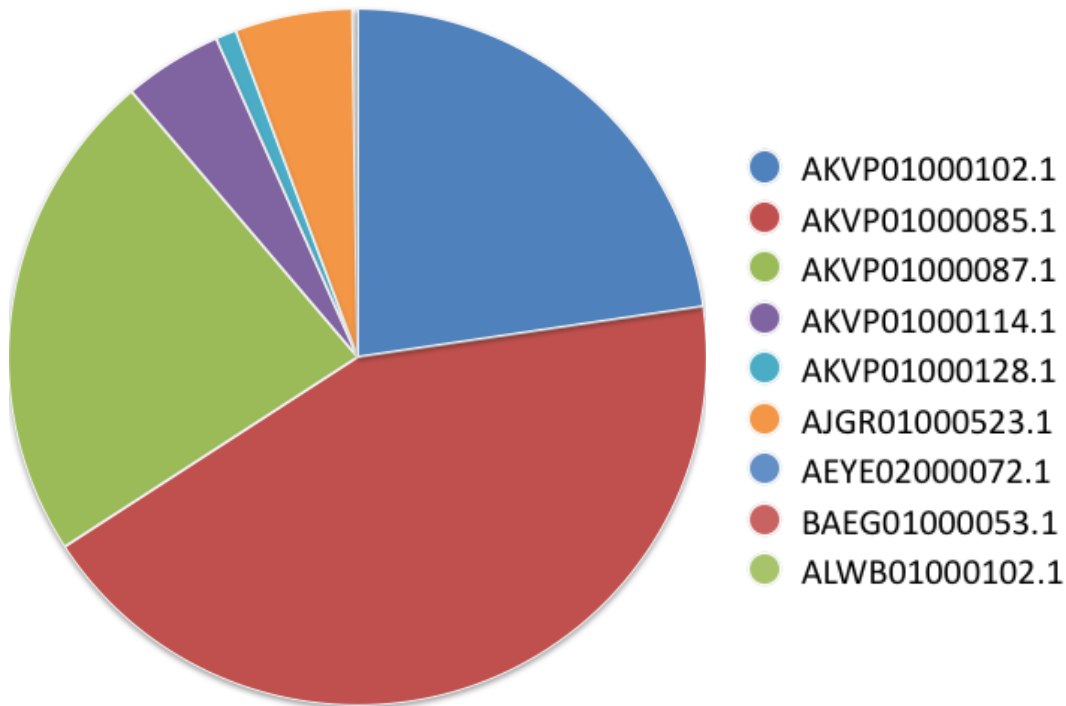
Download excel workbook result files incl graphics:

Results.xlsx

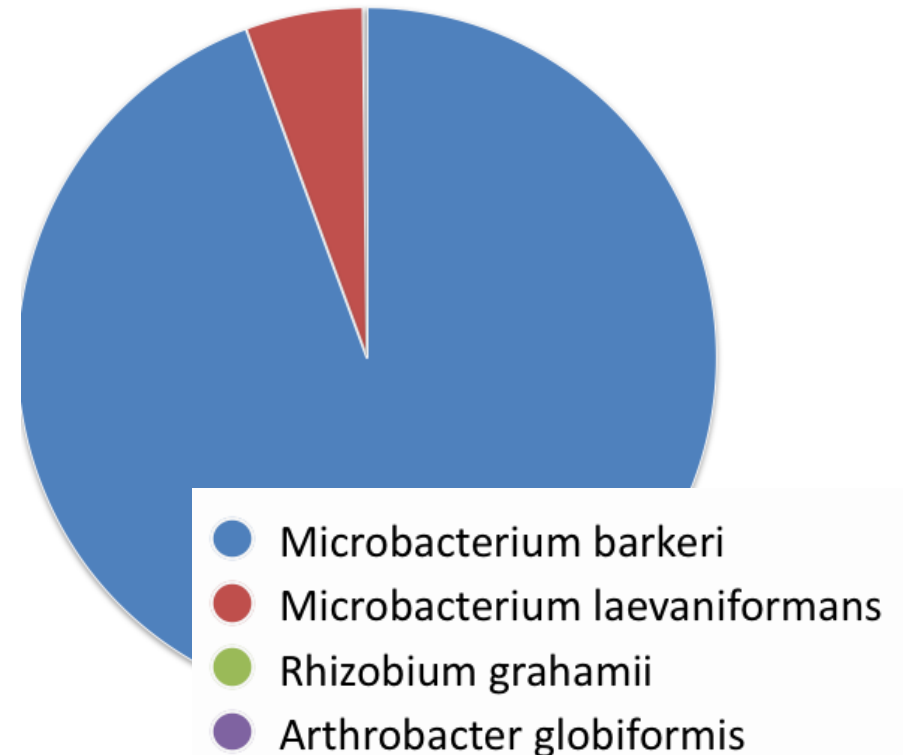
Results.negative.xlsx

Note: Results are both given for individual hits to reference sequences (strain level) and on species-level, where statistics for all references sequences of the same species are lumped together

positive.strain.Bacteria\_draft



positive.species.Bacteria\_draft



# Depth and Coverage

positive.species.HumanMicrobiom			positive.strain.Bacteria_draft		positive.species.Bacteria_draft	
D	E	F	G	H	I	J
Count_Abundant	Size (bp)	Seq_count	Nucleotides	Covered_positions	Coverage	Depth
53,894	7285	6	20786230	5428	0,745	2853,292
3,106	37303	2	1358134	6533	0,175	36,408
0,061	397	1	5159	392	0,987	12,995
0,042	525	1	2102	129	0,246	4,004

Depth = Nucleotides / Size (bp) = 20786230 / 7285 = 2853.292 (depth of coverage)

Coverage = Covered\_positions / Size (bp) = 5428 / 7285 = 0.745

# Viral pathogen

- Which viral pathogen is present in the sample?

Hepatitis A

- How many reads map to the viral pathogen?

54

- Would the viral pathogen have been reported, if Max mismatch ratio = 0.01?

Mismatch ratio = mismatched (edit distance)/ nucleotide =  
1969/21859 = 0.090

Nucleotide	Covered_percentage	Coverage	Depth	Reads	Reads_unique	Edit_distance	Description
21859	6814	0,911	2,923	54	54	1969	Hepatitis A

**positive.strain**

