ECES 450/650 Tutorial 10 - Kaiju Webserver

By: Moaaz Moazzam and Tien Vo

Taxonomic Classification

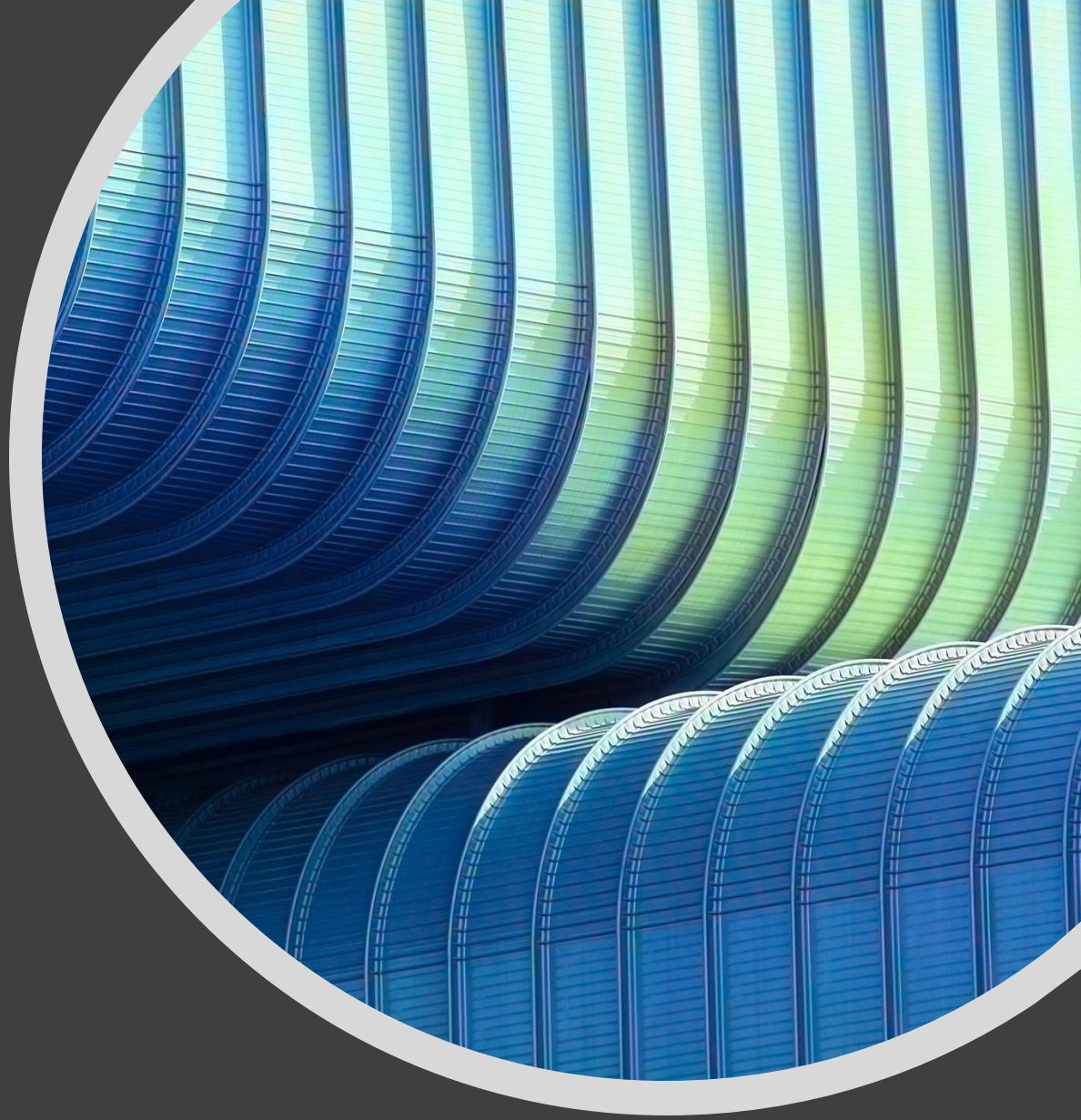
The process of taking in metagenomic reads and classifying those reads into determining the genome.

One method is the use of k-mers and nucleotide sequences. Then searched/compared with a reference database (hash-based).

- Kraken, Clark, LMA
- Issue lies in being only at the DNA level.

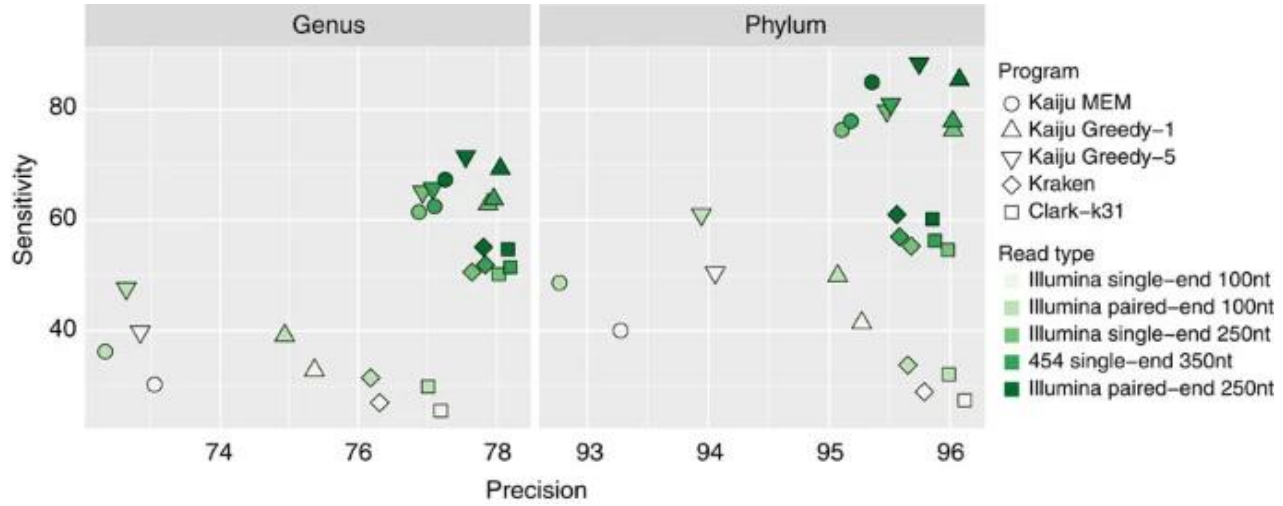
Another method is the use of MEMs and amino-acid sequences then compared with a reference database for classification.

- Kaiju



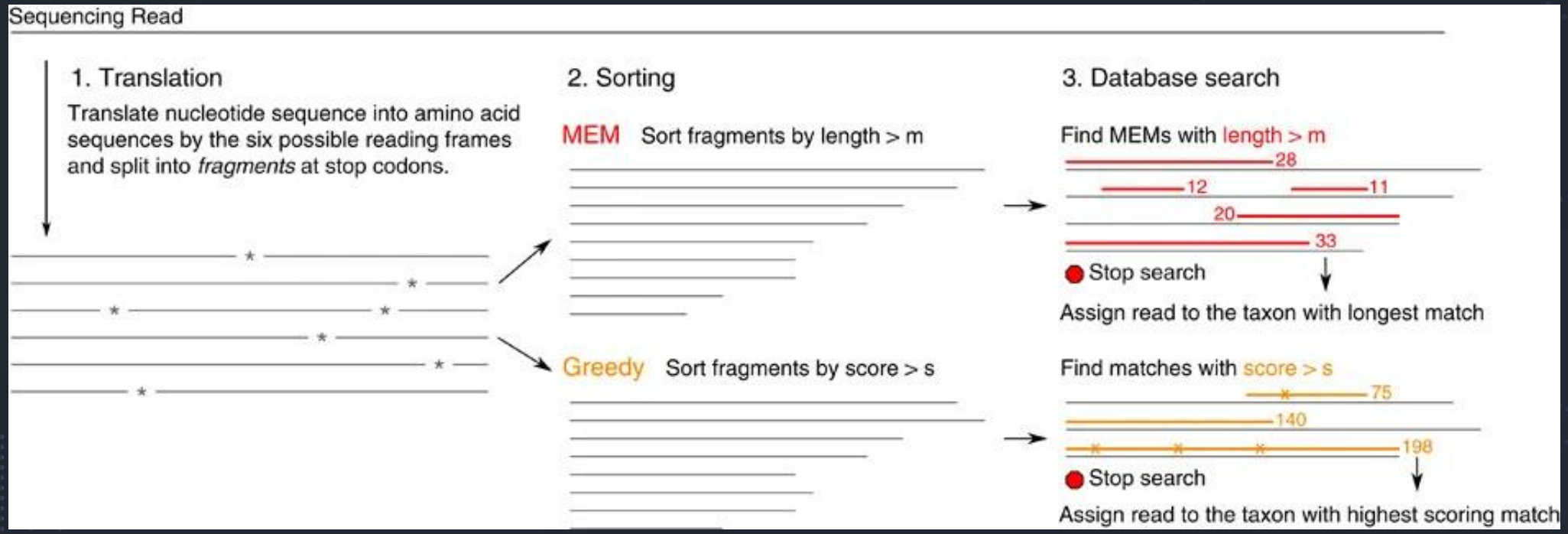
Kaiju vs. Kraken and Clark

- Image below demonstrates the comparison between sensitivity and precision of the few classification methods.



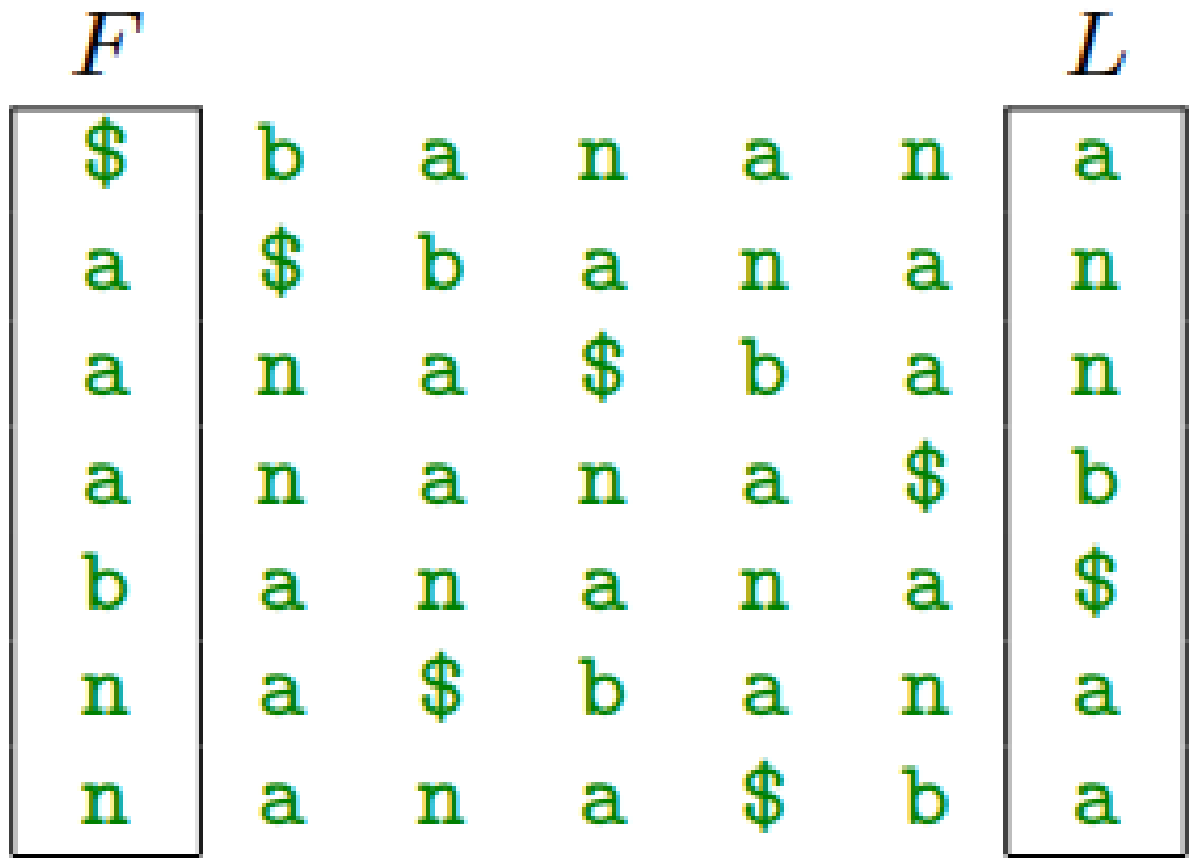
Kaiju Method

- Uses metagenomic reads that are translated into amino-acid sequences.
- From there, a reference database is used to determine matches between the reads and the database.
- Using the MEM or Greedy process, matches are then found after sorting and are either set as:
  - Classified
  - Unclassified



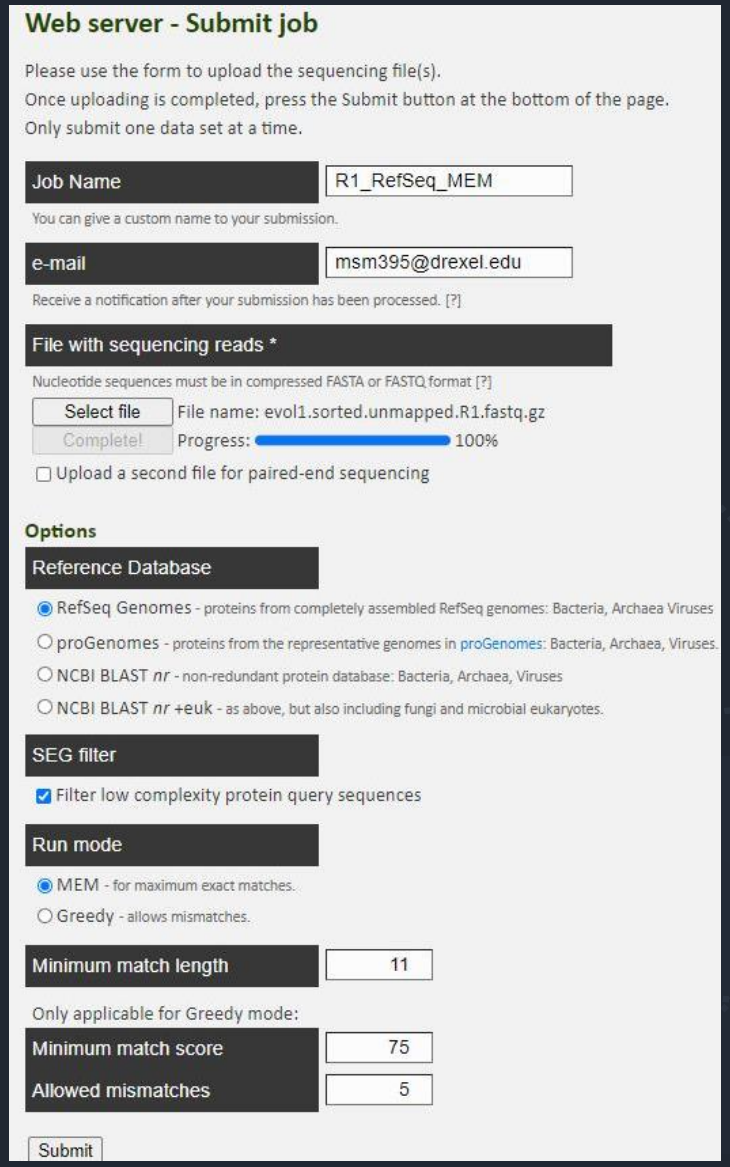
Burrows-Wheeler Transform

- The Burrows-Wheeler Transform (BWT) allows for easier searching between the reference and the sequence read.
- Converts the reference sequence into a compressed form that can alleviate some of the difficulties in determining matches.
- BWT primarily takes in a text string and produces various alterations of the string.
- Paired with MEM or Greedy.



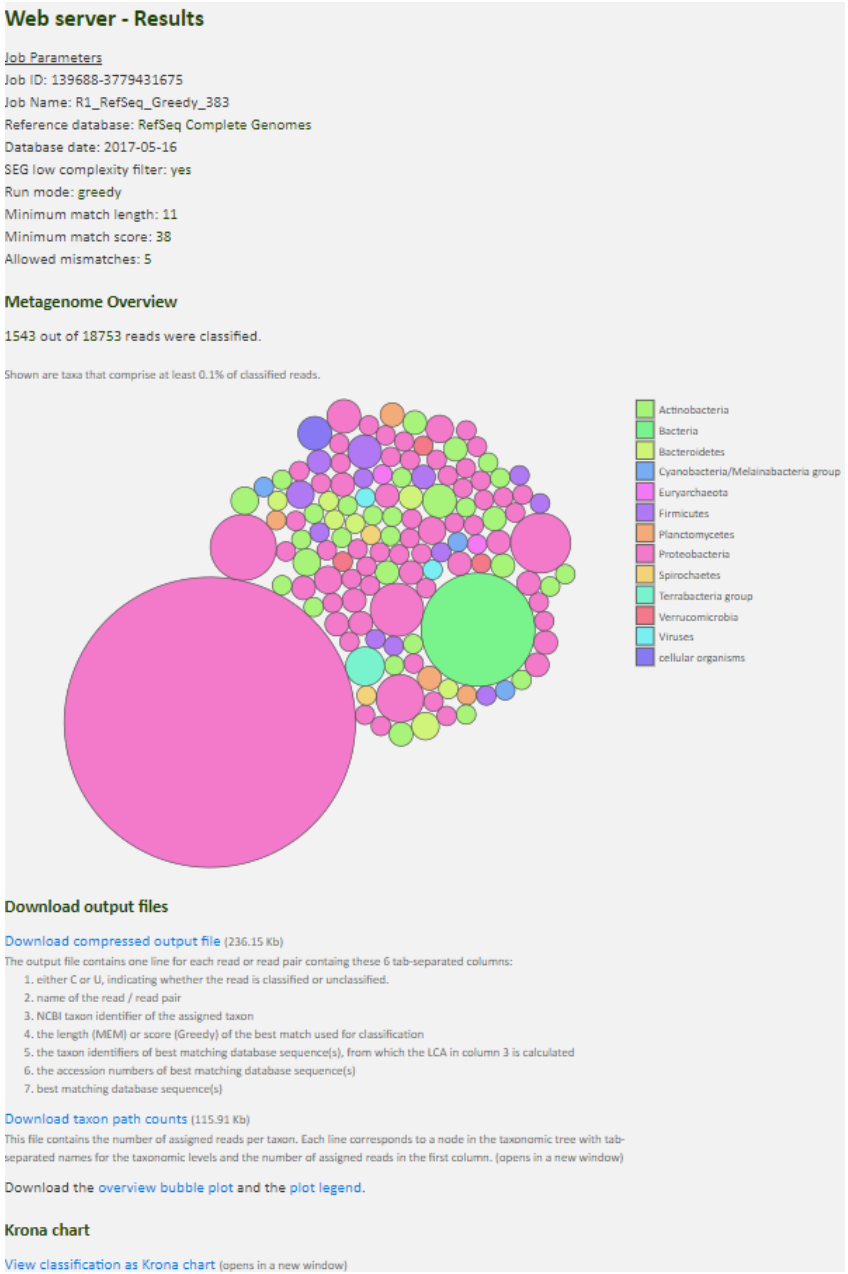
Using the Kaiju Web Server

1. Input Task Name and Email.
2. Provide FASTA/FASTQ File in Zip Form.
3. Adjust Parameters and Run. Should Receive an Email with Result.



Results after Processing

- Taxa Path Listing in a Text File.
- Bubble Chart displaying graphically the taxa and abundances.
- A Classified/Unclassified Krona Chart showcasing percentages.
- Output file showcasing results of each read in a line-based format.



Web server - Results

Job Parameters
Job ID: 139688-3779431675
Job Name: R1_RefSeq_Greedy_383
Reference database: RefSeq Complete Genomes
Database date: 2017-05-16
SEG low complexity filter: yes
Run mode: greedy
Minimum match length: 11
Minimum match score: 38
Allowed mismatches: 5

Metagenome Overview

1543 out of 18753 reads were classified.

Shown are taxa that comprise at least 0.1% of classified reads.

Download output files

Download compressed output file (236.15 Kb)
The output file contains one line for each read or read pair containg these 6 tab-separated columns:
1. either C or U, indicating whether the read is classified or unclassified.
2. name of the read / read pair
3. NCBI taxon identifier of the assigned taxon
4. the length (MEM) or score (Greedy) of the best match used for classification
5. the taxon identifiers of best matching database sequence(s), from which the LCA in column 3 is calculated
6. the accession numbers of best matching database sequence(s)
7. best matching database sequence(s)

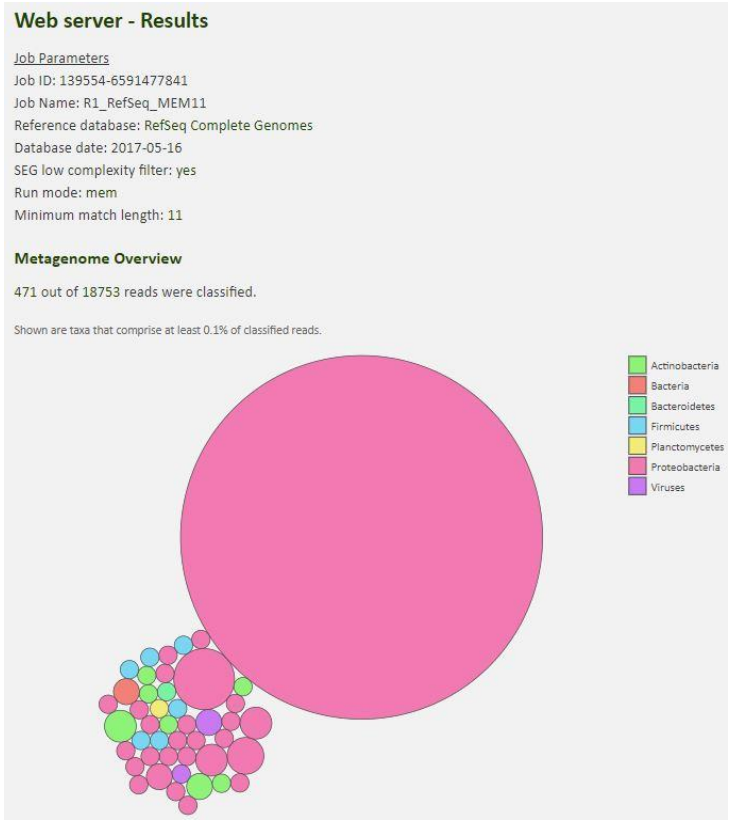
Download taxon path counts (115.91 Kb)
This file contains the number of assigned reads per taxon. Each line corresponds to a node in the taxonomic tree with tab-separated names for the taxonomic levels and the number of assigned reads in the first column. (opens in a new window)

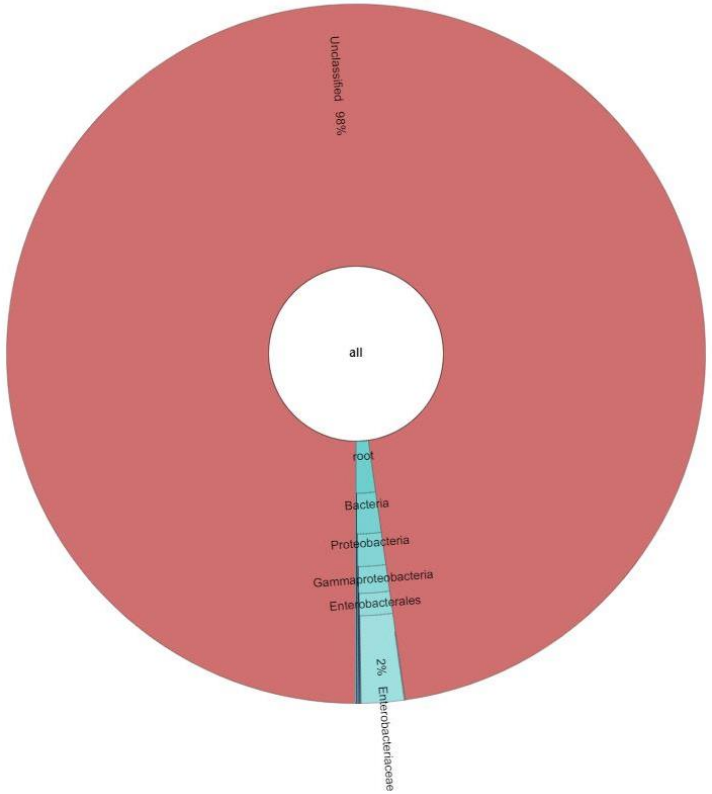
Download the overview bubble plot and the plot legend.

Krona chart

View classification as Krona chart (opens in a new window)

Results with Preset Parameters (R1)





- Ran the R1 data set with the default parameters for both reference databases:
  - Mode 1: MEM with minimum match length of 11.
  - Mode 2: Greedy with minimum match score of 75 and allowed mismatches of 5.
- Images are of the RefSeq Database.

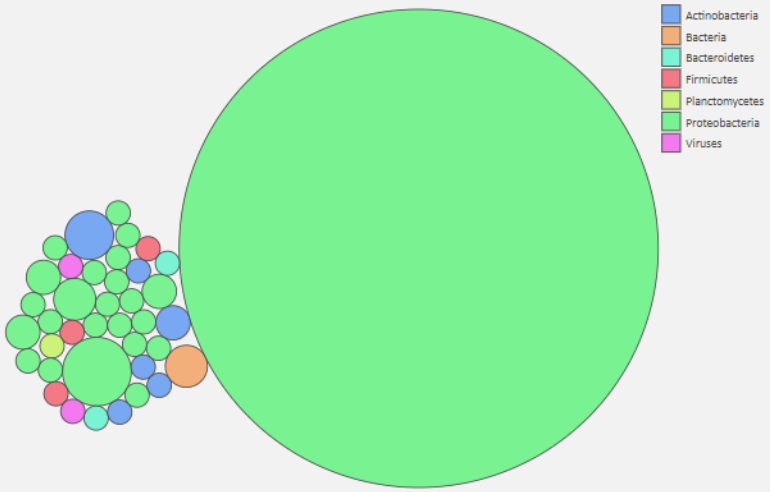
Web server - Results

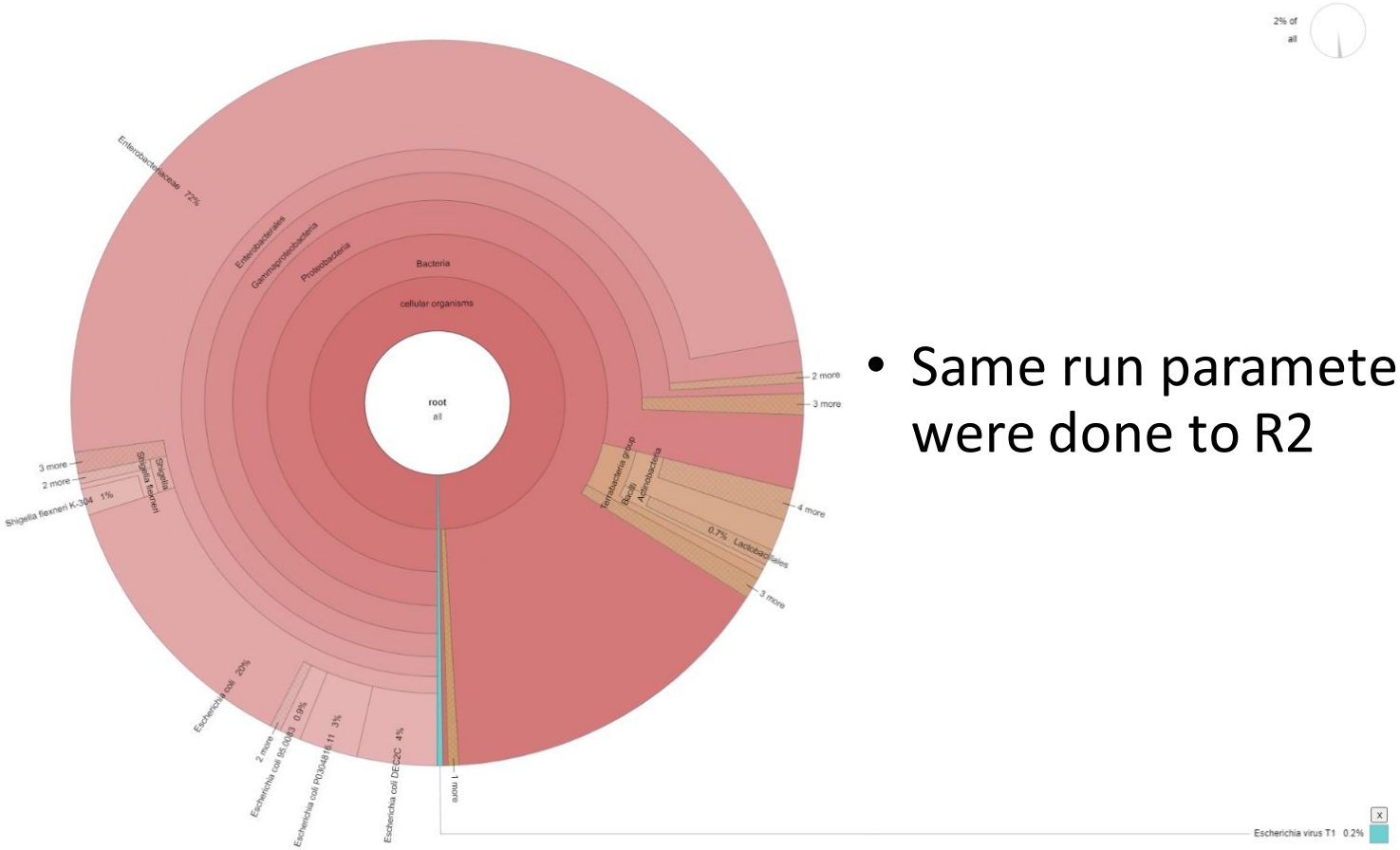
Job Parameters
Job ID: 139560-9410223015
Job Name: R2_RefSeq_MEM
Reference database: RefSeq Complete Genomes
Database date: 2017-05-16
SEG low complexity filter: yes
Run mode: mem
Minimum match length: 11

Metagenome Overview

460 out of 17692 reads were classified.

Shown are taxa that comprise at least 0.1% of classified reads.

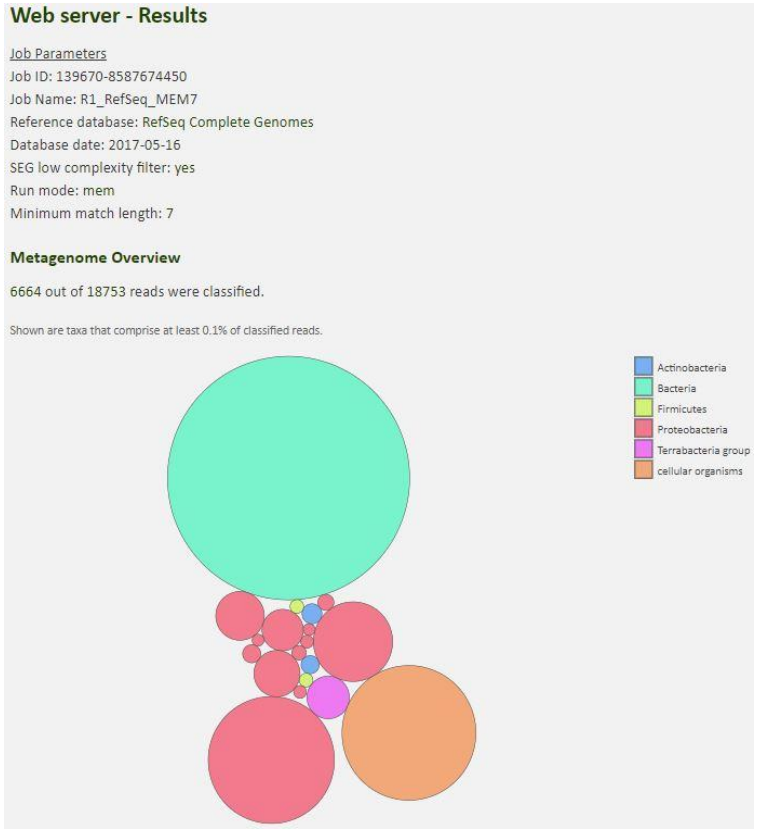


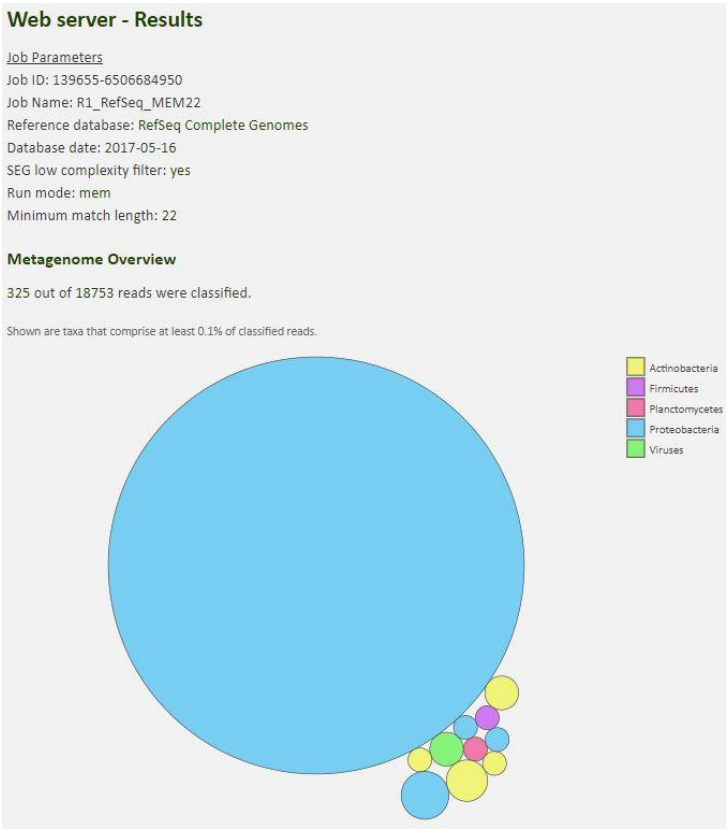


- Same run parameters were done to R2

Results with Preset Parameters (R2)

Adjusting Minimum Match Length – MEM Mode (R1)





- Adjusts the Minimum Match Lengths alters the resultant classified reads.
  - Lower Minimum Match Lengths tend to net higher classified reads.
  - Higher Minimum Match Lengths tend to net lower classified reads.
  - Reduces the exactness of the search with lower match lengths.

Adjusting Minimum Match Length – MEM Mode (R2)

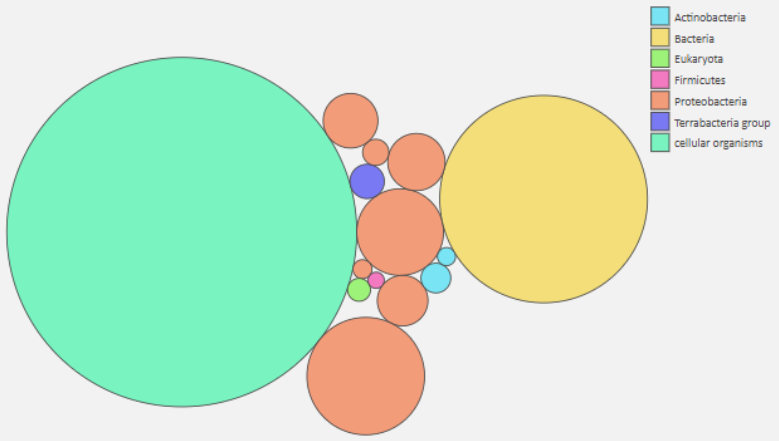
Web server - Results

Job Parameters
Job ID: 139691-7008587735
Job Name: R2_NCBI_EUK_22
Reference database: *nr* +euk
Database date: 2017-05-16
SEG low complexity filter: yes
Run mode: mem
Minimum match length: 7

Metagenome Overview

7291 out of 17692 reads were classified.

Shown are taxa that comprise at least 0.1% of classified reads.



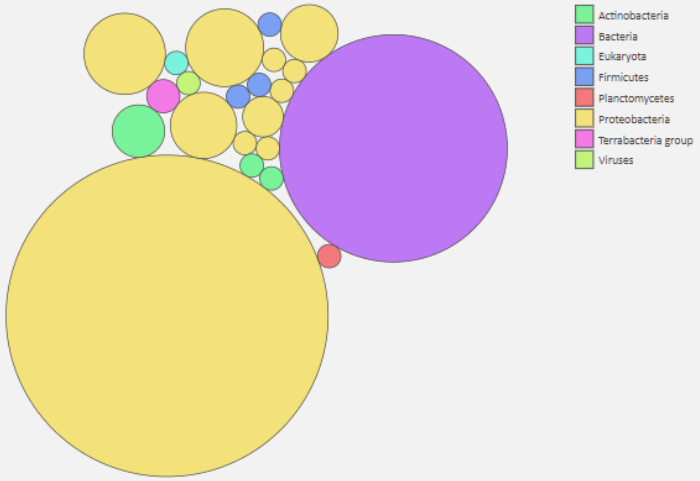
Web server - Results

Job Parameters
Job ID: 139940-3004367237
Job Name: R2_NCBI_EUK_22
Reference database: *nr* +euk
Database date: 2017-05-16
SEG low complexity filter: yes
Run mode: mem
Minimum match length: 22

Metagenome Overview

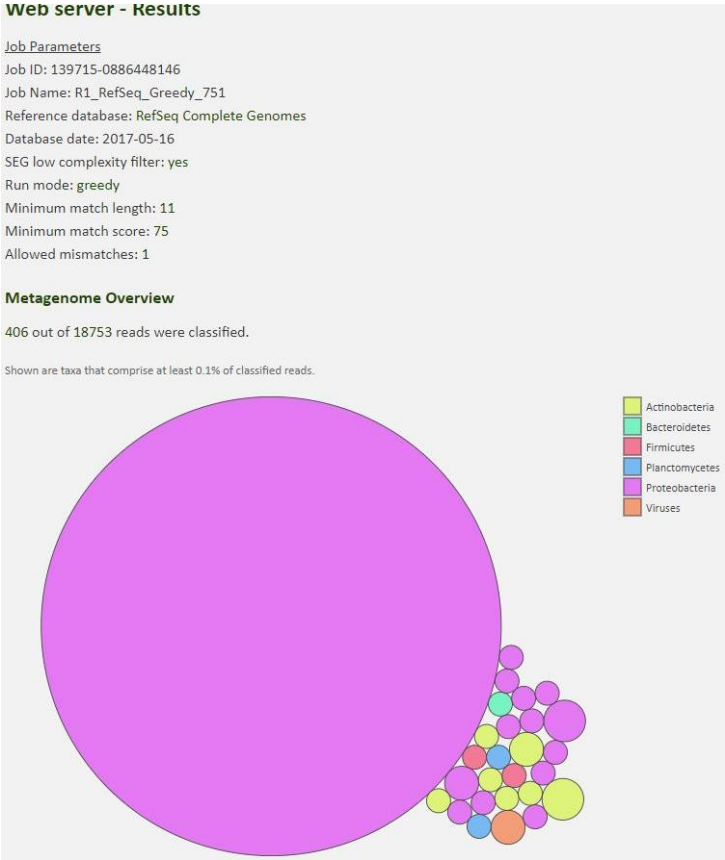
342 out of 17692 reads were classified.

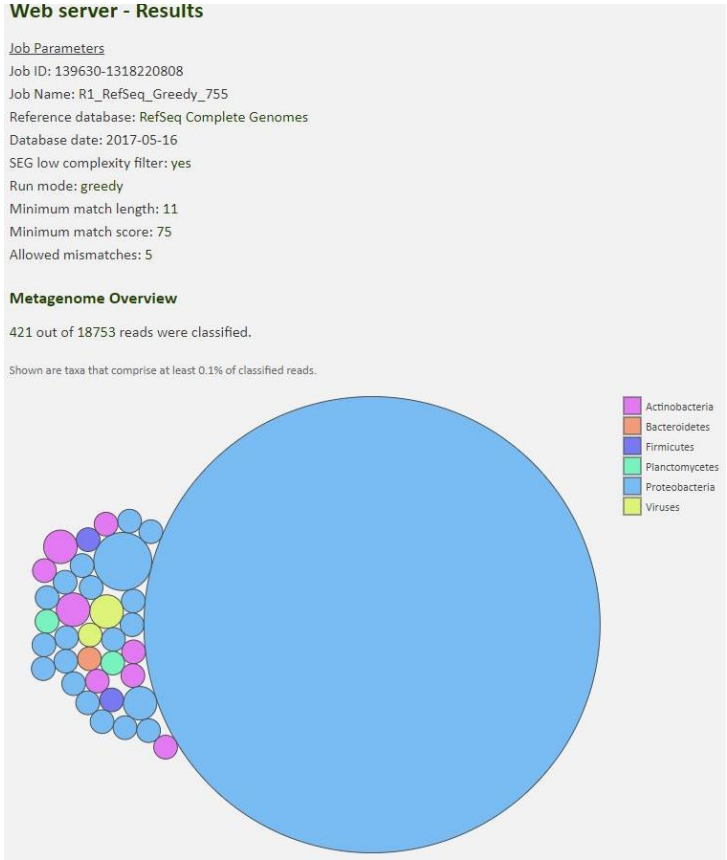
Shown are taxa that comprise at least 0.1% of classified reads.



- Adjusts the Minimum Match Lengths alters the resultant classified reads.
  - Lower Minimum Match Lengths tend to net higher classified reads.
  - Higher Minimum Match Lengths tend to net lower classified reads.
  - Reduces the exactness of the search with lower match lengths.

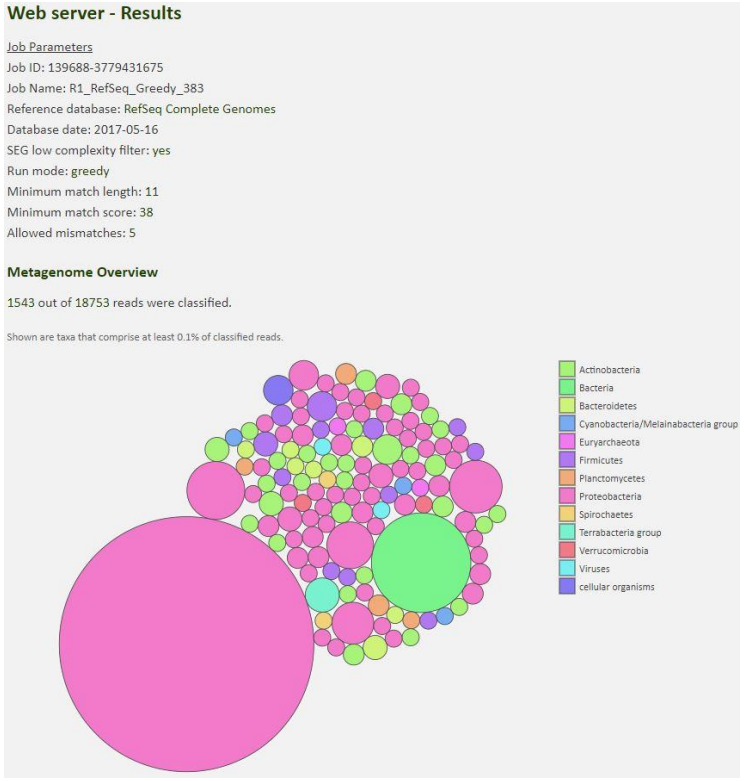
Adjusting Minimum Match Score and Allowed Mismatches - Greedy Mode (R1)





- Similar to adjusting the minimum match length in the MEM mode, the adjustment of the match score and allowed mismatches affects the results.
  - Lower allowed mismatches nets in a more restrictive filter.
  - A lower match score often nets in a less restrictive filter.

Greedy Adjustment (R1) Cont.



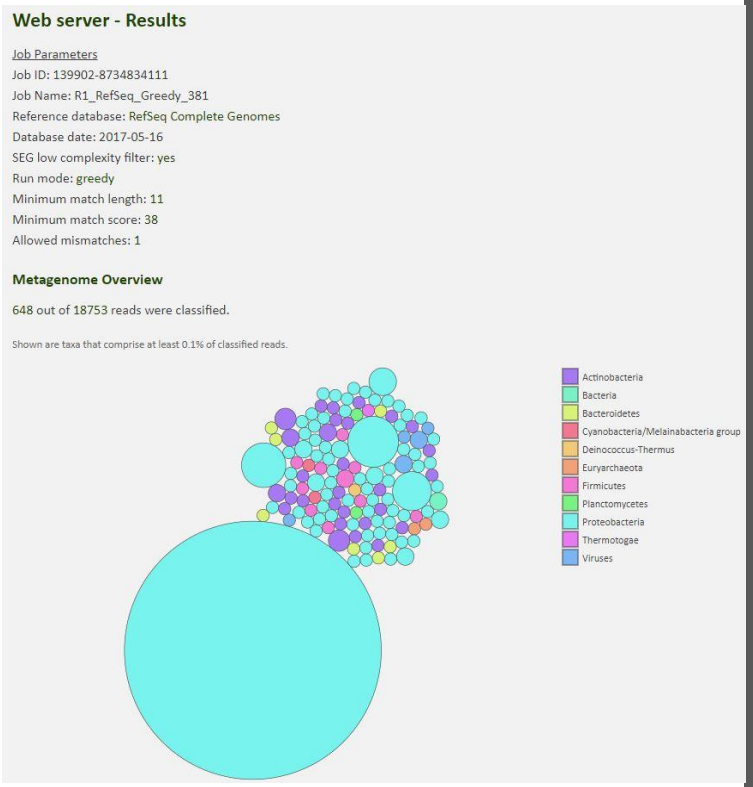
Web server - Results

Job Parameters
Job ID: 139688-3779431675
Job Name: R1_RefSeq_Greedy_383
Reference database: RefSeq Complete Genomes
Database date: 2017-05-16
SEG low complexity filter: yes
Run mode: greedy
Minimum match length: 11
Minimum match score: 38
Allowed mismatches: 5

Metagenome Overview

1543 out of 18753 reads were classified.

Shown are taxa that comprise at least 0.1% of classified reads.



Web server - Results

Job Parameters
Job ID: 139902-8734834111
Job Name: R1_RefSeq_Greedy_381
Reference database: RefSeq Complete Genomes
Database date: 2017-05-16
SEG low complexity filter: yes
Run mode: greedy
Minimum match length: 11
Minimum match score: 38
Allowed mismatches: 1

Metagenome Overview

648 out of 18753 reads were classified.

Shown are taxa that comprise at least 0.1% of classified reads.

- This set of data further demonstrates the adjustment of the parameters to showcase the processing.
- Lower match length and varying allowed mismatches.

Results with Preset Parameters with and without SEG Filter (R2)

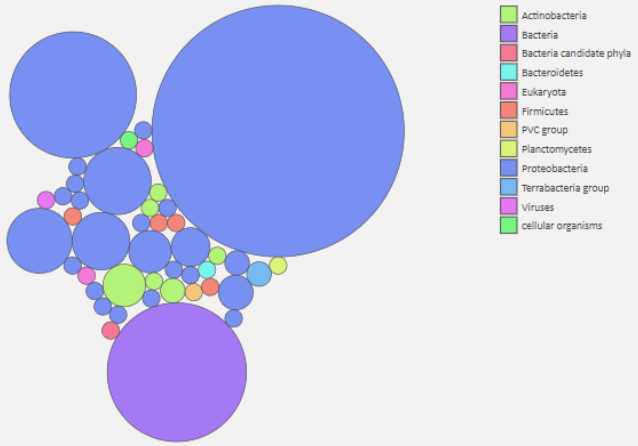
Web server - Results

Job Parameters
Job ID: 139942-7350199146
Job Name: R2_NCBI_EUK_NO_Filter
Reference database: *nr* +euk
Database date: 2017-05-16
SEG low complexity filter: no
Run mode: greedy
Minimum match length: 11
Minimum match score: 75
Allowed mismatches: 5

Metagenome Overview

425 out of 17692 reads were classified.

Shown are taxa that comprise at least 0.1% of classified reads.



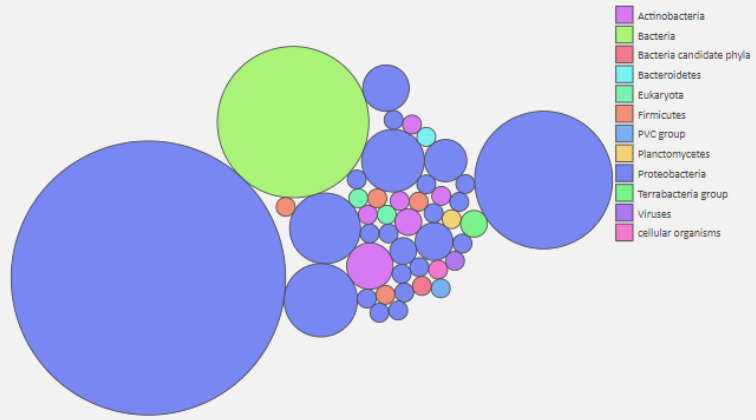
Web server - Results

Job Parameters
Job ID: 139595-8337667093
Job Name: R2_NCBI_EUK_Greed
Reference database: *nr* +euk
Database date: 2017-05-16
SEG low complexity filter: yes
Run mode: greedy
Minimum match length: 11
Minimum match score: 75
Allowed mismatches: 5

Metagenome Overview

425 out of 17692 reads were classified.

Shown are taxa that comprise at least 0.1% of classified reads.



Taxa Path File and Other Outputs

- Kaiju also provides a taxa path text file to show a more line by line image

```
390     root    cellular organisms      Bacteria        Proteobacteria  Gammaproteobacteria     Enterobacterales        Enterobacteriaceae
14      root
11      root    cellular organisms      Bacteria        Proteobacteria  Gammaproteobacteria     Enterobacterales
4       root    cellular organisms      Bacteria        Proteobacteria
3       root    cellular organisms      Bacteria        Terrabacteria group     Actinobacteria  Actinobacteria  Propionibacteriales     Propionibacteriaceae    Cutibacterium
3       root    cellular organisms      Bacteria        Proteobacteria  Gammaproteobacteria     Enterobacterales        Enterobacteriaceae      Escherichia     Escherichia coli
3       root    cellular organisms      Bacteria        Proteobacteria  Gammaproteobacteria
2       root    cellular organisms      Bacteria        Terrabacteria group     Actinobacteria  Actinobacteria  Propionibacteriales     Propionibacteriaceae    Cutibacterium   Cutibacterium acnes
2       root    cellular organisms      Bacteria        Proteobacteria  Gammaproteobacteria     Enterobacterales        Enterobacteriaceae      unclassified Enterobacteriaceae unclassified Enterobacteriaceae (miscellaneous) Enterobacteriaceae bacterium strain FGI 57
2       root    cellular organisms      Bacteria
2       root    Viruses dsDNA viruses, no RNA stage     Caudovirales    Siphoviridae    Tunavirinae     T1virus Escherichia virus T1
1       root    cellular organisms      Bacteria        Terrabacteria group     Firmicutes      Clostridia      Clostridiales   Lachnospiraceae Lachnoclostridium       Lachnoclostridium phytofermentans       Lachnoclostridium phytofermentans ISDg
1       root    cellular organisms      Bacteria        Terrabacteria group     Firmicutes      Bacilli Lactobacillales Streptococcaceae        Streptococcus   Streptococcus thermophilus
1       root    cellular organisms      Bacteria        Terrabacteria group     Firmicutes      Bacilli Bacillales      Staphylococcaceae       Staphylococcus  Staphylococcus xylosus
1       root    cellular organisms      Bacteria        Terrabacteria group     Firmicutes      Bacilli Bacillales      Staphylococcaceae       Staphylococcus  Staphylococcus warneri  Staphylococcus warneri SG1
1       root    cellular organisms      Bacteria        Terrabacteria group     Firmicutes      Bacilli Bacillales      Staphylococcaceae       Staphylococcus
1       root    cellular organisms      Bacteria        Terrabacteria group     Firmicutes      Bacilli Bacillales      Bacillaceae     Bacillus        Bacillus oceanisediminis        Bacillus oceanisediminis 2691
1       root    cellular organisms      Bacteria        Terrabacteria group     Actinobacteria  Actinobacteria  Propionibacteriales     Propionibacteriaceae    Cutibacterium   Cutibacterium avidum
1       root    cellular organisms      Bacteria        Terrabacteria group     Actinobacteria  Actinobacteria  Propionibacteriales     Propionibacteriaceae
1       root    cellular organisms      Bacteria        Terrabacteria group     Actinobacteria  Actinobacteria  Micrococcales   Microbacteriaceae       Microbacterium  Microbacterium aurum
1       root    cellular organisms      Bacteria        Terrabacteria group     Actinobacteria  Actinobacteria  Corynebacteriales       unclassified Corynebacteriales  Lawsonella      Lawsonella clevelandensis
1       root    cellular organisms      Bacteria        Terrabacteria group     Actinobacteria  Actinobacteria  Actinomycetales Actinomycetaceae        Arcanobacterium Arcanobacterium haemolyticum    Arcanobacterium haemolyticum DSM 20595
1       root    cellular organisms      Bacteria        Proteobacteria  Gammaproteobacteria     unclassified Gammaproteobacteria        Candidatus Baumannia    Candidatus Baumannia cicadellinicola    Baumannia cicadellinicola str. Hc (Homalodisca coagulata)
1       root    cellular organisms      Bacteria        Proteobacteria  Gammaproteobacteria     Pseudomonadales Moraxellaceae   Acinetobacter   Acinetobacter junii
1       root    cellular organisms      Bacteria        Proteobacteria  Gammaproteobacteria     Pasteurellales  Pasteurellaceae Gallibacterium  Gallibacterium anatis   Gallibacterium anatis UMN179
1       root    cellular organisms      Bacteria        Proteobacteria  Gammaproteobacteria     Enterobacterales        Yersiniaceae    Yersinia        Yersinia pseudotuberculosis complex
1       root    cellular organisms      Bacteria        Proteobacteria  Gammaproteobacteria     Enterobacterales        Enterobacteriaceae      unclassified Enterobacteriaceae ant, tsetse, mealybug, aphid, etc. endosymbionts        ant endosymbionts       Candidatus Blochmannia  Candidatus Blochmannia floridanus
1       root    cellular organisms      Bacteria        Proteobacteria  Gammaproteobacteria     Enterobacterales        Enterobacteriaceae      Shigella        Shigella flexneri       Shigella flexneri 5     Shigella flexneri 5 str. 8401
1       root    cellular organisms      Bacteria        Proteobacteria  Gammaproteobacteria     Enterobacterales        Enterobacteriaceae      Salmonella
1       root    cellular organisms      Bacteria        Proteobacteria  Gammaproteobacteria     Enterobacterales        Enterobacteriaceae      Escherichia     Escherichia coli        Escherichia coli SMS-3-5
1       root    cellular organisms      Bacteria        Proteobacteria  Gammaproteobacteria     Enterobacterales        Enterobacteriaceae      Escherichia
1       root    cellular organisms      Bacteria        Proteobacteria  Gammaproteobacteria     Enterobacterales        Enterobacteriaceae      Enterobacter    Enterobacter cloacae complex    Enterobacter cloacae
1       root    cellular organisms      Bacteria        Proteobacteria  Gammaproteobacteria     Enterobacterales        Enterobacteriaceae      Cronobacter
1       root    cellular organisms      Bacteria        Proteobacteria  Gammaproteobacteria     Enterobacterales        Enterobacteriaceae      Citrobacter     Citrobacter freundii complex    Citrobacter werkmanii
1       root    cellular organisms      Bacteria        Proteobacteria  Gammaproteobacteria     Chromatiales    Chromatiaceae   Marichromatium  Marichromatium purpuratum       Marichromatium purpuratum 984
1       root    cellular organisms      Bacteria        Proteobacteria  Betaproteobacteria      Rhodocyclales   Rhodocyclaceae  Aromatoleum     Aromatoleum aromaticum  Aromatoleum aromaticum EbN1
1       root    cellular organisms      Bacteria        Proteobacteria  Betaproteobacteria      Burkholderiales unclassified Burkholderiales    Burkholderiales Genera incertae sedis
1       root    cellular organisms      Bacteria        Proteobacteria  Betaproteobacteria      Burkholderiales Oxalobacteraceae        Janthinobacterium       Janthinobacterium sp. B9-8
1       root    cellular organisms      Bacteria        Proteobacteria  Betaproteobacteria      Burkholderiales Comamonadaceae  Delftia
1       root    cellular organisms      Bacteria        Proteobacteria  Betaproteobacteria      Burkholderiales Comamonadaceae  Acidovorax      Acidovorax citrulli     Acidovorax citrulli AAC00-1
1       root    cellular organisms      Bacteria        Proteobacteria  Betaproteobacteria      Burkholderiales Comamonadaceae
1       root    cellular organisms      Bacteria        Proteobacteria  Alphaproteobacteria     Rickettsiales   Anaplasmataceae Wolbachieae     Wolbachia       Wolbachia endosymbiont of Drosophila simulans   Wolbachia endosymbiont of Drosophila simulans wNo
1       root    cellular organisms      Bacteria        Proteobacteria  Alphaproteobacteria
1       root    cellular organisms      Bacteria        PVC group       Planctomycetes  Planctomycetia  Planctomycetales        Gemmataceae     Gemmata Gemmata sp. SH-PL17
1       root    cellular organisms      Bacteria        FCB group       Bacteroidetes/Chlorobi group    Bacteroidetes
1       root    Viruses dsDNA viruses, no RNA stage     Poxviridae      Chordopoxvirinae        unclassified Chordopoxvirinae   BeAn 58058 virus
18282   Unclassified
```

Processing Time Benchmarks

Data	MEM Preset	MEM with min 7	MEM with min 22
R1 with refseq	4 Hrs 58 Mins	4 Mins	4 Mins
R1 with ncbi+euk	3 Hrs 3 Mins	17 Mins	19 Mins
R2 with refseq	5 Hrs 15 Mins	7 Mins	8 Mins
R2 with ncbi+euk	3 Hrs 52 Mins	10 Mins	12 Mins

- The processing times for the various runs differed throughout the several runs and didn't follow a strict pattern.
  - Generally, the Greedy mode ran at a faster time than MEM mode.

Processing Time Benchmarks Cont.

Data	Greed Preset	Greed with Minimum Score of 75 (misses = 1)	Greed with Minimum Score of 38 (misses = 5)	Greed with Minimum Score of 38 (misses = 1)
R1 with refseq	2 Hrs 7 Mins	2 Hrs 10 Mins	2 Mins	1 Hr 2 Mins
R1 with ncbi+euk	2 Hrs 4 Mins	9 Mins	17 Mins	11 Mins
R2 with refseq	2 Hrs 2 Mins	7 Mins	12 Mins	2 Mins
R2 with ncbi+euk	1 Hr 52 Mins	9 Mins	7 Mins	15 Mins

References

- http://kaiju.binf.ku.dk
- Menzel, P., Ng, K. & Krogh, A. Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nat Commun* **7,** 11257 (2016). https://doi.org/10.1038/ncomms11257