Bioinformatics workflow overview

- Host read elimination
- **Read classification**
- Metagenome assembly
- Contig binning
- Bin QC (completeness)
- Bin classification

Host read elimination

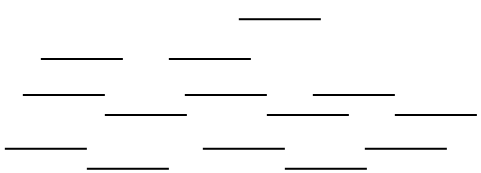
- Identify reads derived from the plant
- Remove them from the collection
- Why?
  - They can complicate assembly
  - They can bloat the data set (affects computer performance)

Read classification

- Which bacteria are present?
- What are their relative proportions?
- Three basic approaches
  - Gene context
  - Read alignment
  - k-mer content

https://doi.org/10.1186/s13059-017-1299-7

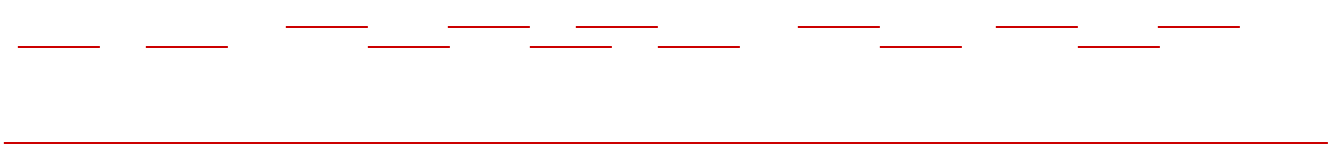
Sequence alignment



Query

Target

Sequence alignment

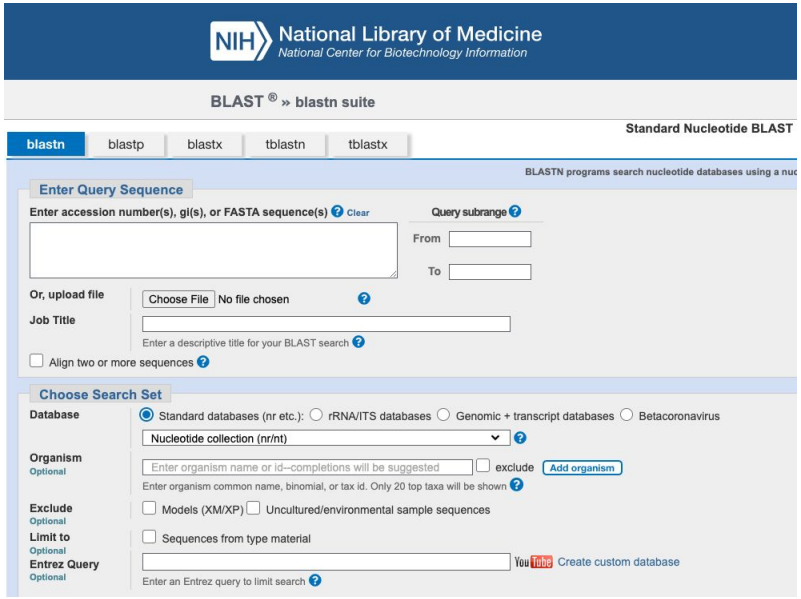


BLAST: Basic Local Alignment Search tool

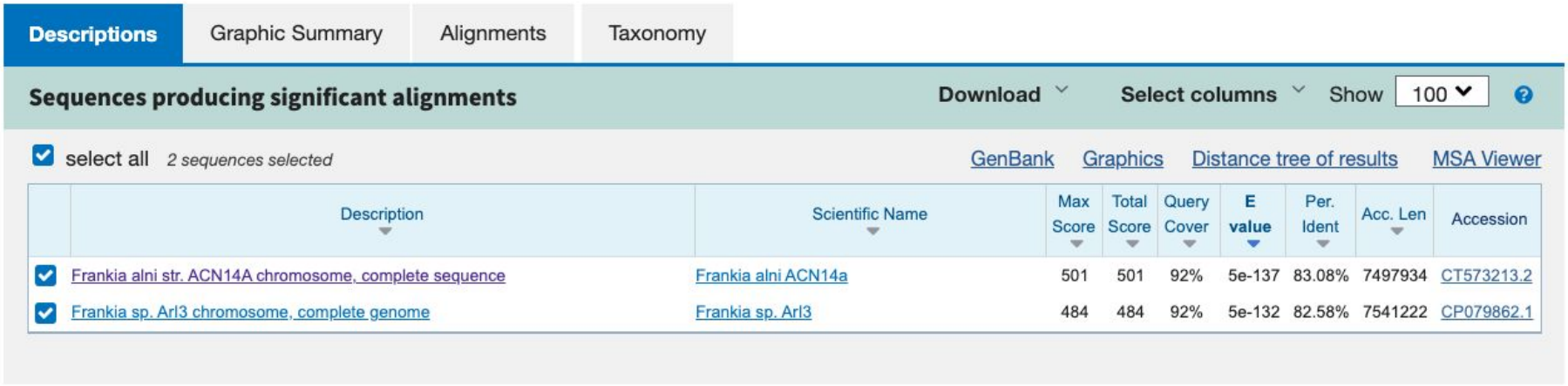
```
>2d828a2d-5b6e-4764-943c-670689c32f55
CGCCAGCGTTCGGTGCGTACTCGGTGGTCGGCTGCAGCCGAACCTGATGGGCGGCAGGCC
AGCAGCAGGAACAGCGCGCTGCGCAGGACCATCTCCCGCCAGCGACGTGGTAACGACTCA
CCGCAGCCGACTGCCGGTAGTTCACCGTCCGGGTGAACAGCGTCCCGCCTCGCCGGTGAT
CAGCGGACGGATCGTCGGTTCCATTCGGCACGATGTCGGCGGTCTCGCCAATACCAGGTC
GAACTCGGCGTGCCTGGCGGTGCCGTCGGCCTGGCGGCAACGCCGTGCGCAGCACGGGTG
CCCGCGACGTTCGAACGCCCCGGCCAGGGACCAGCGTCACCGTGGCGCGCCATAGTCGAA
CCGCGGGTCGCAGCGCAGCGGAACGTCGCCGTCAACGCGGCTCAGGCCTGGCGGACGACG
ACGTAGGGTCTCCGCGACGTCGCTGTCCCTGGGACCGTGAAATCGACGACCTCGCCGGCC
GCCCCGCGGGGGTGAACCTGGTCTCAGGACGTTTGAGTCCGGCGGGTAGAGCTGTTTGGC
CGGTCGGGCGGTGGATGCGAAACGACCCGCCGCGGTCGGCGTCGAGCAGGCTCGAGAAGC
```

Query

Target



BLAST


Descriptions
Graphic Summary
Alignments
Taxonomy
Sequences producing significant alignments
Download
Select columns
Show 100
select all 2 sequences selected
GenBank
Graphics
Distance tree of results
MSA Viewer
Description
Scientific Name
Max Score
Total Score
Query Cover
E value
Per. Ident
Acc. Len
Accession
Frankia alni str. ACN14A chromosome, complete sequence
Frankia alni ACN14a
501
501
92%
5e-137
83.08%
7497934
CT573213.2
Frankia sp. Arl3 chromosome, complete genome
Frankia sp. Arl3
484
484
92%
5e-132
82.58%
7541222
CP079862.1

BLAST

Frankia alni str. ACN14A chromosome, complete sequence

Sequence ID: CT573213.2 **Length:** 7497934 **Number of Matches:** 1

Range 1: 6829254 to 6829849 GenBank Graphics ▼ Next Match ▲ Previous Match

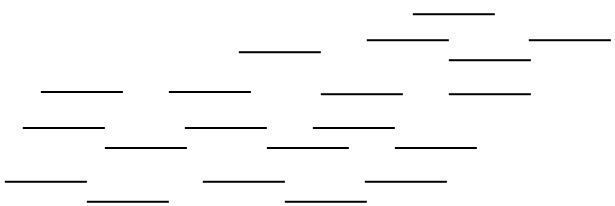
Score	Expect	Identities	Gaps	Strand
501 bits(271)	5e-137	496/597(83%)	46/597(7%)	Plus/Minus

```
Query  48       TGGG-CGGCAGGCCAGCAGC---AGGAACAGCGCGCTGCGCAGGACCATCTCCCGCCAGC  103
                ||||  |||  ||  |||||||||   ||  |  ||||||||||||| ||||||||||||||||||||
Sbjct  6829849  TGGGCCGGTAGACCAGCAGCTTCAGCACCAGCGCGCTGCGTAGGACCATCTCCCGCCAGC  6829790

Query  104      G-ACGTGGTAACGACT-CACCGCAG-C-CGACTGCCGGTAGTTCACCGTCCGGGTGAACA  159
                |  ||||||||||| ||   ||||||| |  || |||||  ||||||||  ||||||||||||||||
Sbjct  6829789  GGCCGTGGTAACGGCTTTGCCGCAGCCACGCCTGCCAGTAGTTCAACGTCCGGGTGAACA  6829730

Query  160      GCGT-CCCGCCTCGCCGGTGATCAGCGGACGGATCGTCG-GTTCCATTCG-GCACGATGT  216
                ||||  ||||||||||||  ||||||||||  ||||||||||  ||||||||||  |||||||||
Sbjct  6829729  GCGTGTCCGCCTCGCCGATGATCAGCGGCCGGATCGTCGAGTTCCATTCGAGCACGATGT  6829670

Query  217      CGGCGGTCTCGCCAAT-ACCAGGTCGAACTCGGCGTGCCTGGCGGTGCCGTCGGCCTG--  273
                ||||||||||||||  ||   ||||||||||||  |||  |  ||||||||||||||  ||  |
Sbjct  6829669  CGGCGGTCTCGCCCATGGGCAGGTCGAACTCCGCGAACACGGCGGTGCCGTCGACCCGCA  6829610

Query  274      GCGGCAACGCCGTGCGCAGCACG--GGTGCCCGCGAC-G-TTCG-A-A-CG-CCCCG-GC  324
                ||||||  ||  ||||||||||||  |||||||  |||  |  ||||  |  |  ||  ||||||  ||
Sbjct  6829609  GCGGCAGCGAGGTGCGCAGCACGAGGGTGCCCTTGACCGATTCGAACACCGCCCCCGAGC  6829550
```

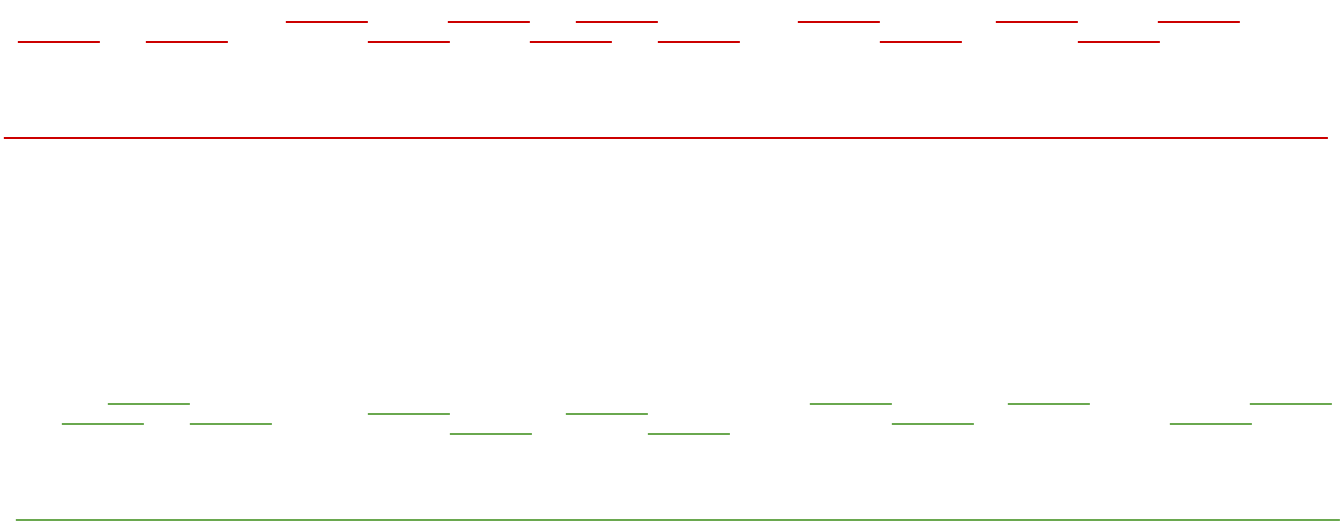
New Mexico INBRE — IDeA Networks of Biomedical Research Excellence

NCGR — National Center for Genome Resources

Sequence alignment



Sequence alignment



Gene content

- e.g. Phylosift
- https://doi.org/10.7717/peerj.243
- Aligns input reads to a set of 37 gene families
  - Present in nearly all bacteria
  - Single copy
- Makes a multiple sequence alignment
- Places each input sequence in a phylogenetic tree
- Positions each output in the taxonomy

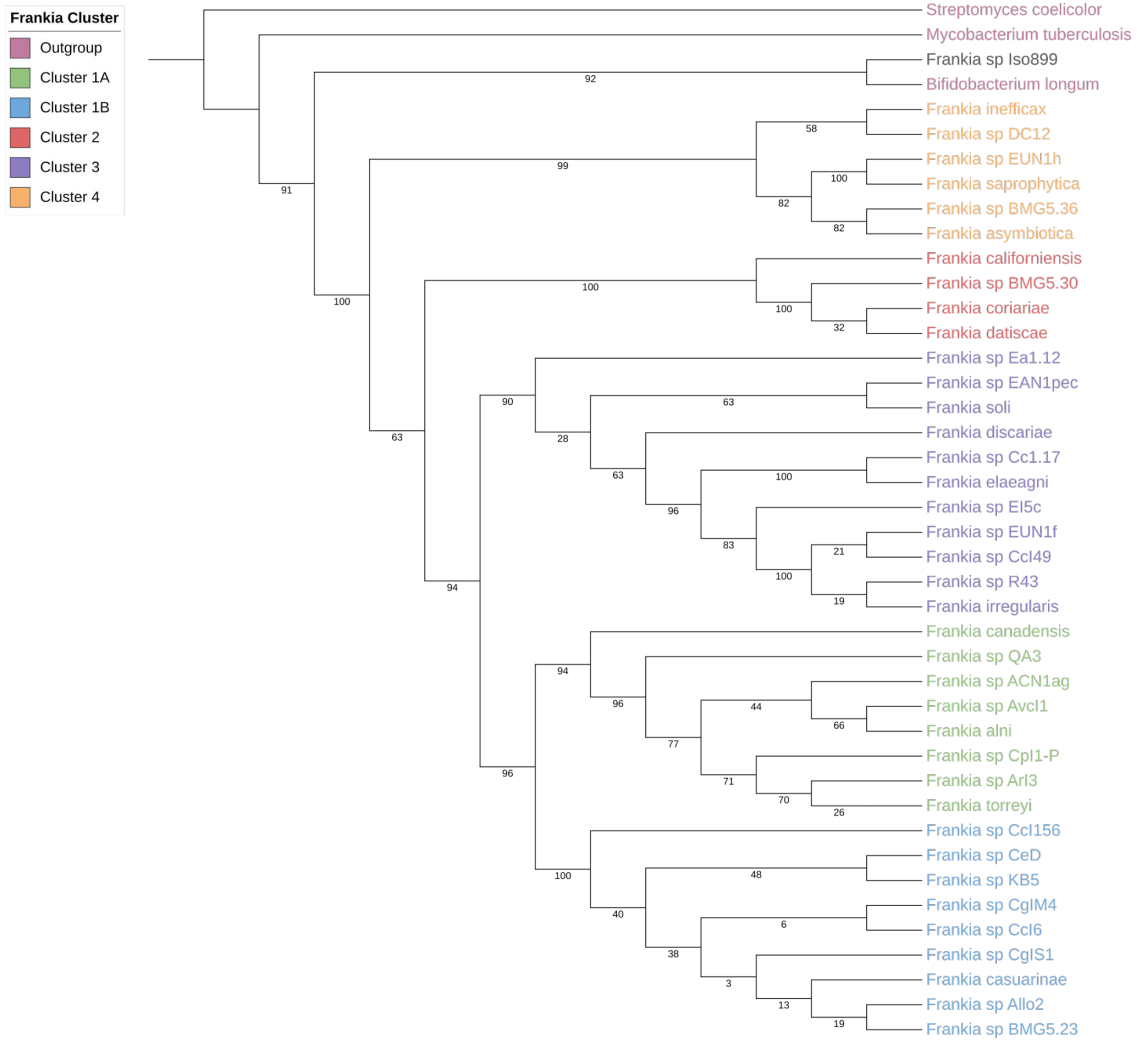
glnA1 (protein) MSA

```
B.longum_glnA1                                  PSSLAEAMDALEEDHDFLTAGDVFTDDLIETWIGLKR-DEIDQARLSPTPLEYELYFHI-
Frankia_sp_CeD_glnA1                            PGSLEKVLDALEADNEFLREGDVFTADLIETWLDYKRVNEVDAIRLRPHPYEFTLYYDI*
Frankia_sp_CgIM4_glnA1                          PGSLEKVLDALEADNEFLREGDVFTADLIETWLDYKRVNEVDAIRLRPHPYEFTLYYDI*
Frankia_sp_CgIS1_glnA1                          PGSLEKVLDALEADNEFLREGDVFTADLIETWLDYKRVNEVDAIRLRPHPYEFTLYYDI*
Frankia_sp_CcI156_glnA1                         PGSLEKVXDALEADNEFLREGDVFTADLIETWLDYKRVNEVDAIRLRPHPYEFTLYYDI*
Frankia_sp_Allo2_glnA1                          PGSLEKVLDALEADNEFLREGDVFTADLIETWLDYKRVNEVDAIRLRPHPYEFTLYYDI*
Frankia_sp_Iso899_glnA1                         PGSLGEVLTELERDHDYLLEGGVFTEDLIATWIDYKRVNEVDPVRLRPHPYEFDLYYNI*
Frankia_sp_BMG5.23_glnA1                        PGSLEKVLDALEADNEFLREGDVFTADLIETWLDYKRVNEVDAIRLRPHPYEFTLYYDI*
Frankia_sp_CcI6_glnA1                           PGSLEKVLDALEADNEFLREGDVFTADLIETWLDYKRVNEVDAIRLRPHPYEFTLYYDI*
Frankia_coriariae_glnA1                         PGSLEKVLDALEADHEFLTEGNVFTPDLIETWIDYKRKNEVDAIRLRPHPYEFQLYYDI*
Frankia_sp_BMG5.30_glnA1                        PGSLEKVLDALEADHEFLTEGNVFTPDLIETWIDYKRKNEVDAIRLRPHPYEFQLYYDI*
Frankia_sp_EI5c_glnA1                           PGSLEKVLDALEADNDFLREGNVFTGDLIETWIDYKRVNEVDAIRLRPHPYEFSLYYDI*
Frankia_sp_DC12_glnA1                           PGSLEKVLEHLEADNEFLTAGDVFTPDLIETWLDYKRTNEVDAIRLRPHPFEFELYYDI*
Frankia_casuarinae_ASM1334v1_glnA1              PGSLEKVLDALEADNEFLREGDVFTADLIETWLDYKRVNEVDAIRLRPHPYEFTLYYDI*
Frankia_canadensis_glnA1                        PGSLEKVLDALEADHEFLREGDVFTTDLIETWLDYKRRNEVDAIRLRPHPYEFTLYYDI*
Frankia_sp_CpI1-P_glnA1                         PGSLEKVLDALEADNDFLREGDVFTTDLIETWLEYKRLNEVDAIRLRPHPYEFTLYYDI*
Frankia_sp_AvcI1_glnA1                          PGSLEKVLDALEADNDFLREGDVFTTDLIETWLEYKRLNEVDAIRLRPHPYEFTLYYDI*
Frankia_sp_ACN1ag_glnA1                         PGSLEKVLDALEADNDFLRXGDVFTTDLIETWLEYKRLNEVDAIRLRPHPYEFTLYYDI*
Frankia_sp_Ea1.12_glnA1                         PGSLEKVLDALEADNEFLREGNVFTPDLIETWIDYKRMNEVDAIRLRPHPYEFSLYYDI*
Frankia_discariae_glnA1                         PGSLEKVLDALEADNEFLREGNVFTPDLIETWIDYKRVNEVDAIRLRPHPYEFSLYYDI*
Frankia_elaeagni_glnA1                          PGSLEKVLDALEADNEFLRAGNVFTPDLIETWIDYKRVNEVDAIRLRPHPYEFSLYYDI*
Frankia_sp_Cc1.17_glnA1                         PGSLEKVLDALEADNEFLRAGNVFTPDLIETWIDYKRVNEVDAIRLRPHPYEFSLYYDI*
Frankia_alni_ACN14a_glnA1                       PGSLEKVLDALEADNDFLREGDVFTTDLIETWLEYKRLNEVDAIRLRPHPYEFTLYYDI*
Frankia_asymbiotica_glnA1                       PGSLEAVLDHLEADNEFLTAGDVFTPDLIETWLDYKRTHEVDAIRLRPHPFEXDLYYDI*
Frankia_inefficax_glnA1                         PGSLEAVLEHLEEDNEFLTAGDVFTPDLIETWLDYKRTNEVDAIRLRPHPFEFELYYDI*
Frankia_sp_EUN1f_glnA1                          PGSLEKVLEALEADNDFLREGNVFTGDLIETWIDYKRVNEVDAIRLRPHPYEFSLYYDI*
Frankia_sp_EUN1h_glnA1                          PGSLEAVLDHLEADNEFLTAGDVFTPDLIETWLDYKRTHEVDAIRLRPHPFEFDLYYDI*
Frankia_irregularis_glnA1                       PGSLEKVLEALEADNDFLREGNVFTGDLIETWIDYKRVNEVDAIRLRPHPYEFSLYYDI*
Frankia_saprophytica_glnA1                      PGSLEAVLDHLEADNEFLTAGDVFTPDLIETWLDYKRTHEVDAIRLRPHPFEFDLYYDI*
Frankia_sp_KB5_glnA1                            PGSLEKVLDALEADNEFLREGDVFTADLIETWLDYKRVNEVDAIRLRPHPYEFTLYYDI*
Frankia_sp_BMG5.36_glnA1                        PGSLEAVLDHLEADNEFLTAGDVFTPDLIETWLDYKRTHEVDAIRLRPHPFEFDLYYDI*
Frankia_sp_CcI49_glnA1                          PGSLEKVLEALEADNDFLREGNVFTGDLIETWIDYKRVNEVDAIRLRPHPYEFSLYYDI*
Frankia_sp_EAN1pec_glnA1                        PGSLEKVLDALEADNEFLREGNVFTPDLIETWIDYKRVNEVDAIRLRPHPYEFSLYYDI*
Frankia_symbiont_of_Datisca_glomerata_glnA1     PGSLEKVLDALEADHEFLTEGNVFTPDLIETWIDYKRKNEVDAIRLRPHPYEFQLYYDI*
Frankia_torreyi_glnA1                           PGSLEKVLDALEADNDFLREGDVFTTDLIETWLEYKRLNEVDAIRLRPHPYEFTLYYDI*
Frankia_sp_NRRL_B-16219_glnA1                   PGSLEKVLDALEADNEFLREGNVFTPDLIETWIDYKRVNEVDAIRLRPHPYEFSLYYDI*
Frankia_sp_R43_glnA1                            PGSLEKVLEALEADNDFLREGNVFTGDLIETWIDYKRVNEVDAIRLRPHPYEFSLYYDI*
Frankia_ArI3_glnA1                              PGSLEKVLDALEADNDFLREGDVFTTDLIETWLEYKRLNEVDAIRLRPHPYEFTLYYDI*
M.tuberculosis_glnA1                            PTQLSDVIDRLEADHEYLTEGGVFTNDLIETWISFKRENEIEPVNIRPHPYEFALYYDV-
S.coelicolor_glnA1                              PTSLGAVLDRLEADHEFLLQGDVFTPDLIETWIDFKRANEIAPLQLRPHPHEFEMYFDV-
                                                * .*  .   ** *:::*  *.*** *** **:  ** .*:   .: * * *  :*:.:
```

New Mexico INBRE
IDeA Networks of Biomedical Research Excellence

NCGR
National Center for Genome Resources

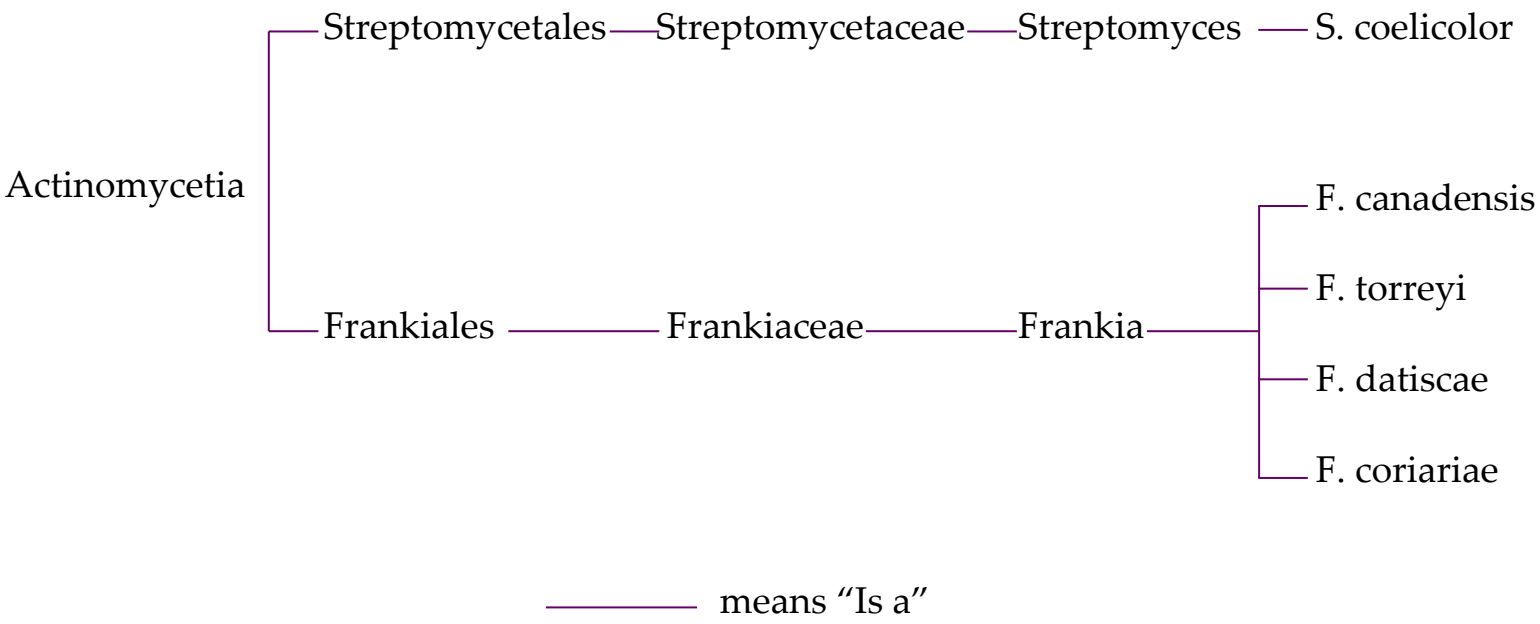
glnA1 (protein) phylogeny



New Mexico INBRE
IDeA Networks of Biomedical Research Excellence

NCGR
National Center for Genome Resources

glnA1 (protein) taxonomy



Read alignment

- e.g. MEGAN
- https://doi.org/10.1101%2Fgr.5969107
- Align all protein or DNA sequences to a large database
- (BLAST or Diamond)
- Assigns taxonomic rank based on LCA (least common ancestor)
- Tends to be slow

k-mer composition

- What is a k-mer?
- k is a natural number, often in the range 20-50
- k-mer is a DNA segment k nucleotides long
- For a DNA string of length *L* there are *L-k+1* k-mers

4-mers of a string

```
mississippi
miss
 issi
  ssis
   siss
    issi
     ssip
      sipp
       ippi
```

k-mer index of a string

```
miss 0
issi 1,4
ssis 2
siss 4
ssip 5
sipp 6
ippi 7
```

k-mer database

- Take all k-mers in a database of bacterial genomes
- each k-mer points to:
  - Each genome it is found in
  - Least common ancestor in the NCBI taxonomy
- Count k-mers (and reads) for each node in the taxonomy

Choosing an algorithm

Research | Open Access | Published: 21 September 2017

Comprehensive benchmarking and ensemble approaches for metagenomic classifiers

Alexa B. R. McIntyre, Rachid Ounit, Ebrahim Afshinnekoo, Robert J. Prill, Elizabeth Hénaff, Noah Alexander, Samuel S. Minot, David Danko, Jonathan Foox, Sofia Ahsanuddin, Scott Tighe, Nur A. Hasan, Poorani Subramanian, Kelly Moffat, Shawn Levy, Stefano Lonardi, Nick Greenfield, Rita R. Colwell, Gail L. Rosen ✉ & Christopher E. Mason ✉

Genome Biology **18**, Article number: 182 (2017) | Cite this article

18k Accesses | **132** Citations | **60** Altmetric | Metrics

https://doi.org/10.1186/s13059-017-1299-7

Choosing an algorithm

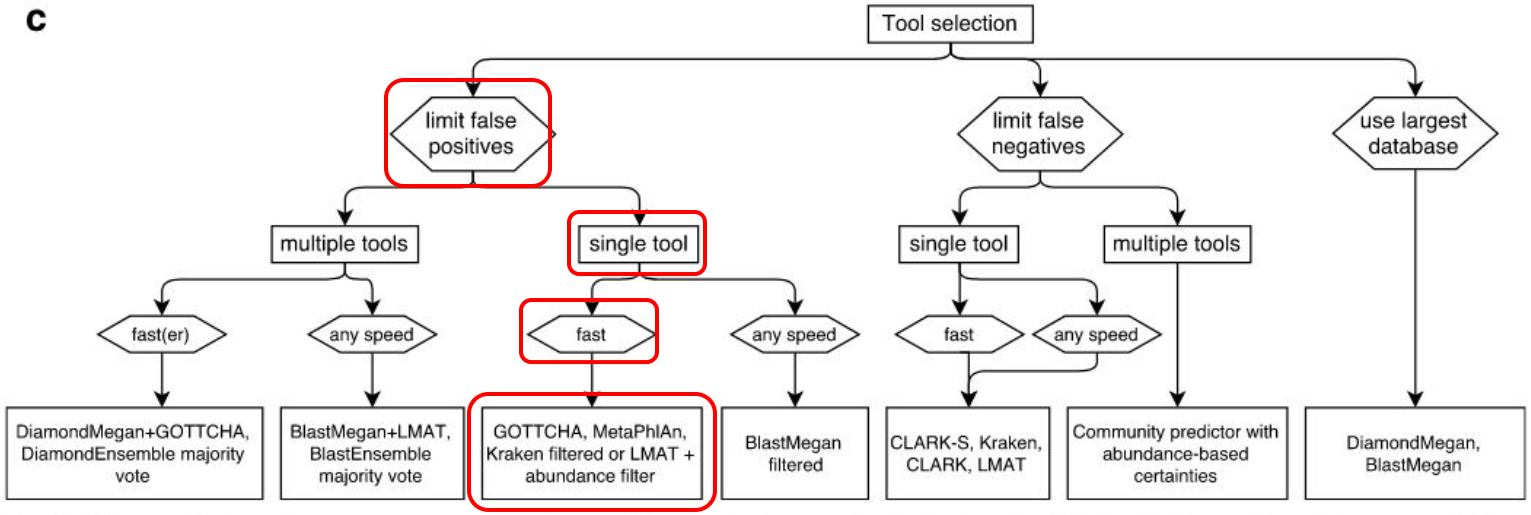


Fig. 7 a Time and (**b**) maximum memory consumption running the tools on a subset of data using 16 threads (where the option was available, except for PhyloSift, which failed to run using more than one thread, and NBC, which was run through the online server using four threads). BLAST, NBC, and PhyloSift were too slow to completely classify the larger datasets, therefore subsamples were taken and time multiplied. **c** A decision tree summary of recommendations based on the results of this analysis

https://doi.org/10.1186/s13059-017-1299-7

Break

Preparing to run krakenuniq

```
cd

mkdir analysis
mkdir analysis/070722
mkdir analysis/070722/minit3

cd analysis/070722/minit3

conda activate seqtools
```

Running Krakenuniq

```
time krakenuniq \
--db /home/cjb/minion/2019/indexes/krakenuniq/db_052422/ \
--threads 32 \
--report-file report \
--unclassified-out unclassified.fna \
--classified-out classified.fna \
/home/cjb/minion/2022/data/minit3/data/fastq_pass/all.fastq \
> output

real 50m3.753s
user 6m28.293s
sys 7m21.626s
```

Test characteristics

- True positive
- False positive
- True negative
- False negative

Confusion Matrix

		Actual Condition	
		Positive	Negative
Test outcome	Positive	TP	FP (Type I error)
	Negative	FN (Type II error)	TN

Precision

		Actual Condition	
		Positive	Negative
Test outcome	Positive	TP	FP (Type I error)
	Negative	FN (Type II error)	TN

- Precision = fraction of positive tests that are actually correct
- Also called the positive predictive value
- TP/(TP+FP)
- High value if false positives are low

Recall

		Actual Condition	
		Positive	Negative
Test outcome	Positive	TP	FP (Type I error)
	Negative	FN (Type II error)	TN

- Recall = How good the test is at detecting positives
- Also called sensitivity
- TP/(TP + FN)
- High value if false negatives are low

Specificity

		Actual Condition	
		Positive	Negative
Test outcome	Positive	TP	FP (Type I error)
	Negative	FN (Type II error)	TN

- Specificity = how good is the test at avoiding false alarms
- TN/(TN+FP)
- High value if false positives are low

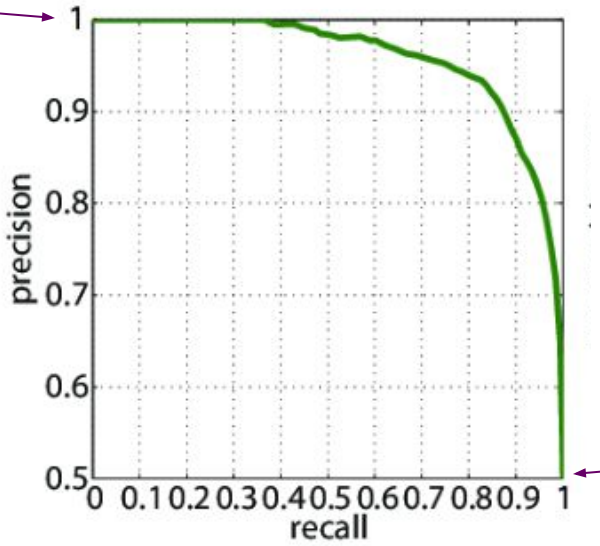
Relevance to metagenomics

		Actual Condition	
		Read belongs to species X	Read does not belong to species X
Test outcome	Read assigned to species X	TP	FP (Type I error)
	Read not assigned to species X	FN (Type II error)	TN

- False positive: we conclude a species is present when it really is not
- By maximizing precision (TP/(TP+FP)) we can minimize the false positive rate

Precision-Recall curve

High threshold
Large number of reads



- TP/(TP+FP)

Low threshold
Small number of reads

- TP/(TP + FN)

https://stackoverflow.com/questions/59519995/roc-curve-and-precision-recall-curve

P and R are evaluated at a range **abundance thresholds**. At each threshold a certain number of reads is required to declare a "positive test."

Kraken algorithm

- https://doi.org/10.1186/gb-2014-15-3-r46

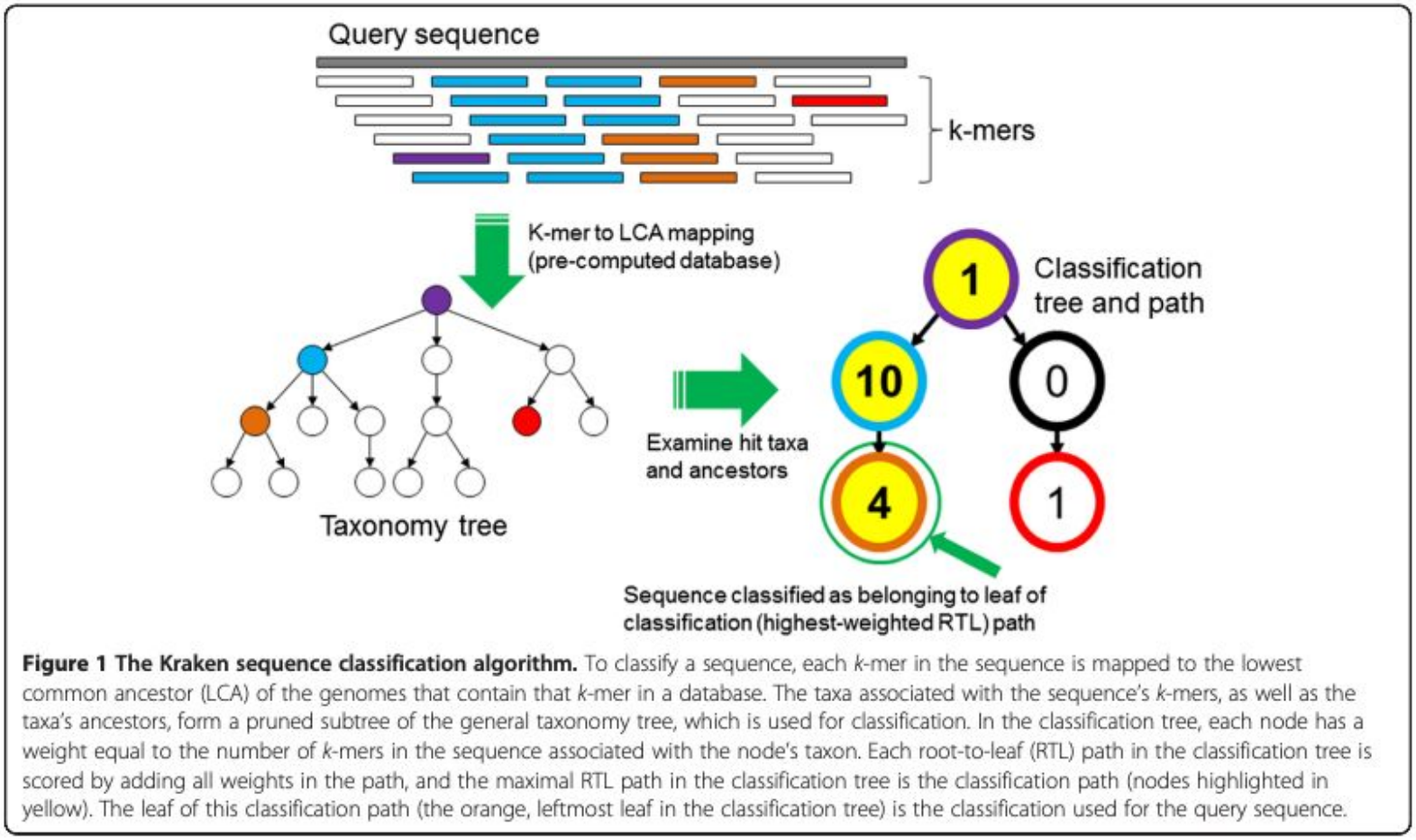


Figure 1 The Kraken sequence classification algorithm. To classify a sequence, each *k*-mer in the sequence is mapped to the lowest common ancestor (LCA) of the genomes that contain that *k*-mer in a database. The taxa associated with the sequence's *k*-mers, as well as the taxa's ancestors, form a pruned subtree of the general taxonomy tree, which is used for classification. In the classification tree, each node has a weight equal to the number of *k*-mers in the sequence associated with the node's taxon. Each root-to-leaf (RTL) path in the classification tree is scored by adding all weights in the path, and the maximal RTL path in the classification tree is the classification path (nodes highlighted in yellow). The leaf of this classification path (the orange, leftmost leaf in the classification tree) is the classification used for the query sequence.

Kraken variants

Kraken	High memory requirement	
Kraken2	Faster and uses less memory	Higher false positive rate. Cost in specificity and accuracy
KrakenUniq	Not much more memory but still a lot (350 Gbyte)	Better at eliminating false positives. Best precision and recall among competitors

https://doi.org/10.1186/gb-2014-15-3-r46

https://doi.org/10.1186/s13059-019-1891-0

https://doi.org/10.1186/s13059-018-1568-0

Krakenuniq

KrakenUniq: confident and fast metagenomics classification using unique *k*-mer counts

F. P. Breitwieser ✉, D. N. Baker & S. L. Salzberg ✉

Genome Biology **19**, Article number: 198 (2018) | Cite this article

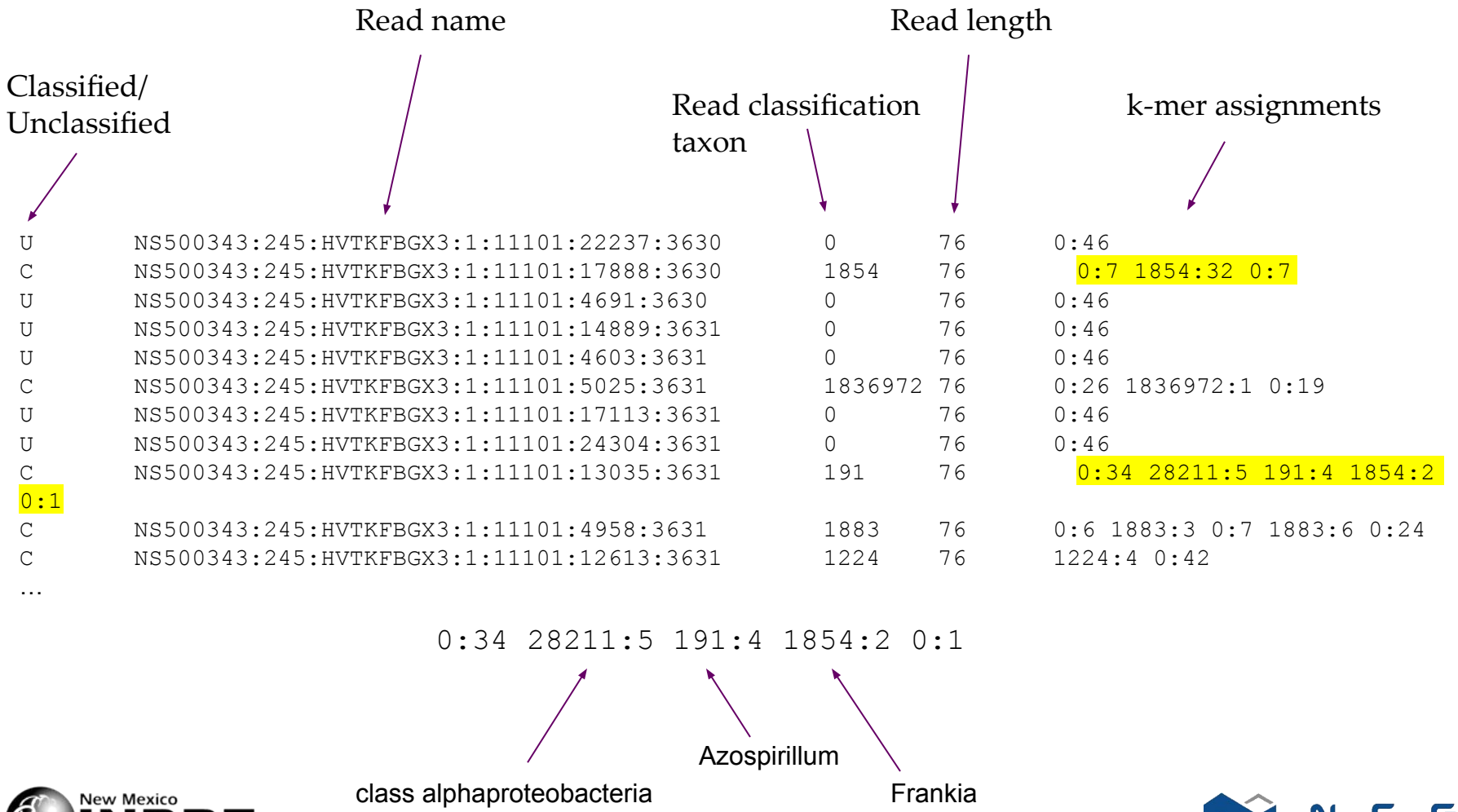
18k Accesses | **92** Citations | **83** Altmetric | Metrics

- Specifically addresses the issue of False Positives

Krakenuniq

- Specifically addresses the issue of False Positives
- Determines genome coverage by k-mers
- Can identify probable false positives

Krakenuniq output



New Mexico INBRE
IDeA Networks of Biomedical Research Excellence

NCGR
National Center for Genome Resources

Krakenuniq report

%	reads	taxReads	kmers	dup	cov	taxID	rank	taxName
47.52	173729	173729	203299340		1.48	5.013	0	no rank unclassified
52.48	191865	0	1147101	6.87	3.929e-05		1	no rank root
52.48	191865	0	1147101	6.87	3.929e-05		131567	no rank cellular organisms
52.48	191865	577	1147101	6.87	3.929e-05		2	superkingdom Bacteria
51.39	187862	17	1098468	5.83	9.857e-05		1783272	clade Terrabacteria group
51.28	187481	27	1086873	5.88	0.00018	201174	phylum	Actinobacteria
51.27	187426	2153	1084762	5.88	0.0001842		1760	class Actinomycetia
48.56	177538	0	950921	6	0.00604	85013	order	Frankiales
48.56	177538	0	950921	6	0.00604	74712	family	Frankiaceae
48.56	177538	68346	950921	6	0.00604	1854	genus	Frankia
25.68	93902	2140	220487	6.34	0.003596		2632575	no rank unclassified Frankia
20.44	74712	74712	196028	5.27	0.03234	710111	species	Frankia sp. QA3
0.9669	3535	3535	1444	38.6	0.0003421		298653	species Frankia sp. EAN1pec
0.7262	2655	2655	1012	63	0.01656	1858	species	Frankia sp. ArI3
0.7166	2620	2620	2890	17.7	0.0005019		573497	species Frankia sp. Cc1.17
0.4396	1607	1607	3711	5.72	0.004982		102891	species Frankia sp. ACN1ag
0.3657	1337	1337	332	30.9	0.0001781		269536	species Frankia sp. R43
0.3085	1128	1128	1954	9.25	0.0002126		1834512	species Frankia sp. BMG5.36
0.2828	1034	1034	2452	5.07	0.003522		573496	species Frankia sp. AvcI1
0.1863	681	681	1420	5.3	0.0002448		683316	species Frankia sp. EI5c
0.1813	663	663	918	8.31	0.0004488		573499	species Frankia sp. Ea1.12

…

Parsing krakenuniq report

```
pcregrep "\tgenus\t" report | cut -f 2,9  | sort -nr | more
pcregrep "\tgenus\t" report | cut -f 2,9  | sort -nr > genus_report.txt
perl -pe 's/\s+/,/' genus_report.txt | more
perl -pe 's/\s+/,/' genus_report.txt > genus_report.csv

pcregrep "\tspecies\t" report | cut -f 2,9  | grep Frankia | sort -nr | more
pcregrep "\tspecies\t" report | cut -f 2,9  | grep Frankia | sort -nr > \
frankia_report.txt
perl -pe 's/\s+/,/' frankia_report.txt | more
perl -pe 's/\s+/,/' frankia_report.txt > frankia_report.csv

scp cjb@logrus.training.ncgr.org:/home/cjb/minion/2022/analysis/070622/genus_report.csv .

scp cjb@logrus.training.ncgr.org:/home/cjb/minion/2022/analysis/070622/frankia_report.csv
.
```

Background to krakenuniq

- Download taxonomy
- Download database data
- Prepare custom database
- Build database
  - https://github.com/fbreitwieser/krakenuniq
  - krakenuniq-download
  - krakenuniq-build

Background to krakenuniq

```
d52422/
|-- library
|   `-- bacteria
|         |-- library.fna
|-- seqid2taxid.map
`-- taxonomy
    |-- names.dmp
    `-- nodes.dmp
```

seq2taxid.map

```
GCA_000018005.1 298653
GCA_001636565.1 683316
GCA_000177675.1 102897
GCA_001854645.1 1834515
GCA_900465275.1 573499
GCA_000421445.1 1283283
GCA_002099325.1 683318
GCA_000262465.1 710111
GCA_001306465.1 269536
GCA_000948395.1 1856
...
```

library.fna

```
>NZ_CP013563.1
CAGCCAGAGAGGCGGAGGGCTGAGGCTTTCGCCATCGAGAACCCTCATGATCTTTCGGCG
CGTGTCGCTCAAGCGCTTCTTCCCTATAAAAAATCAAAGATATTTTAAAAGGTTTCTATT
TCTTAGAGTCGGTGTCTATCAAGGATTAAAATCCATCCACTGCCGATGCCATTTGTCGCG
CGCCCGCAACCAAATTCxxxxxxxxxTTTTCCGATTCTCTGAGATTTCCGGAAGAAGACG
AATCCGGAAACGATTCCAAAACCTTGCTCGACAGGTGATTTTCTCCTTCCTGTGGATAGC
TTGTGGACCTGCGTTCGACAAACTGATTTCGTCCCCATCCTCCACAGGGCCGGGCAAAAC
CGACCCGCAAAATTCGAAATGTGGATAAGGCGGCATGTTTTCCCTTGCCCGAGGCAGCCT
…

64663 entries
```

taxonomy files

- nodes.dmp
- names.dmp