Kaiju

Fast and sensitive taxonomic classification for metagenomics

Peter Menzel

pmenzel@gmail.com ptr_menzel

Metagenomics

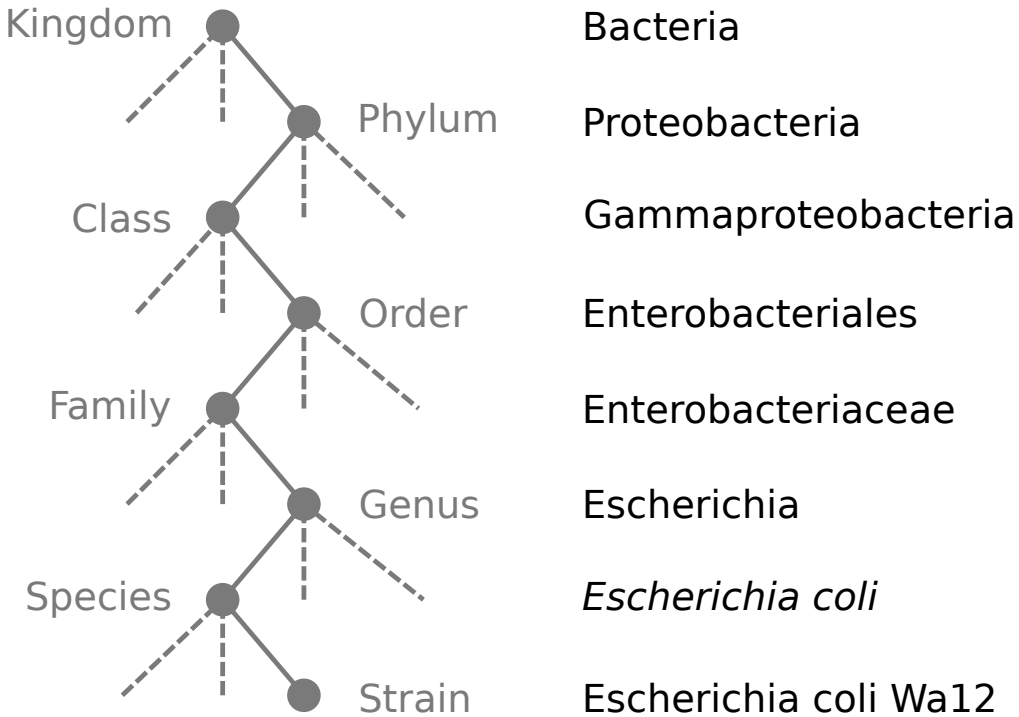
Biological Question

What is the composition of the microbial community in a specific habitat?
For example: drinking water, compost, human skin or gut, bioreactor, . . .

Metagenomics

Biological Question

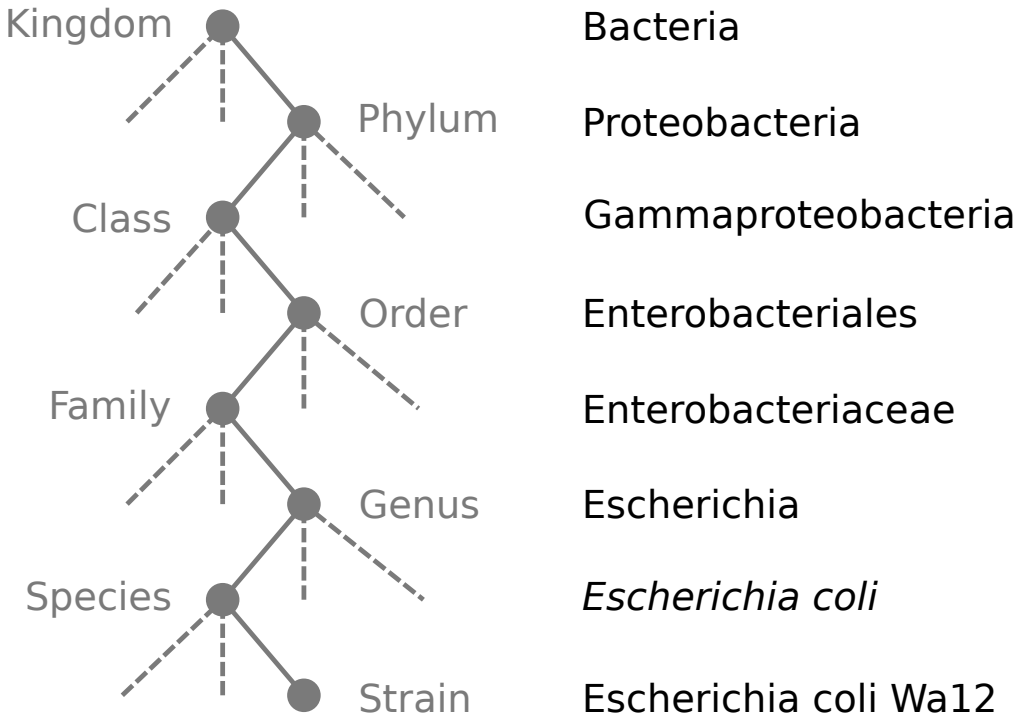
What is the composition of the microbial community in a specific habitat?
For example: drinking water, compost, human skin or gut, bioreactor, . . .



Metagenomics

Biological Question

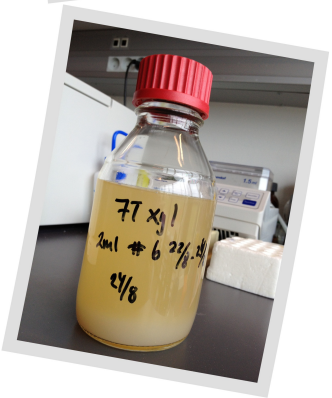
What is the composition of the microbial community in a specific habitat?
For example: drinking water, compost, human skin or gut, bioreactor, . . .

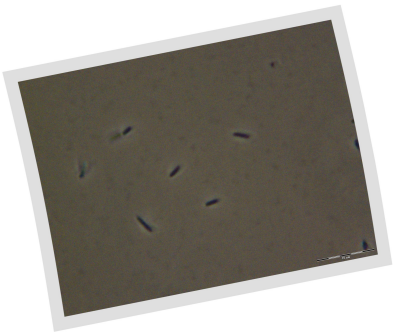


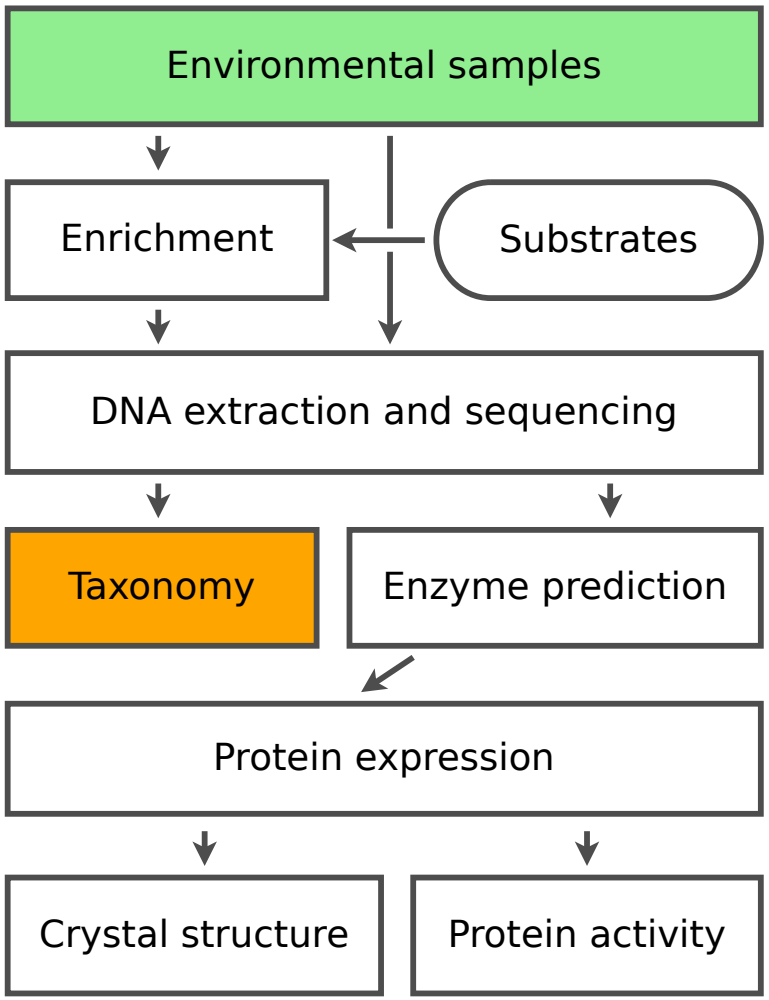
Metagenomic Sequencing

Sequencing of total genomic DNA.
Estimate relative abundances of taxa at a given taxonomic level.

Hotzyme: Discovery of novel hydrolases from hot springs







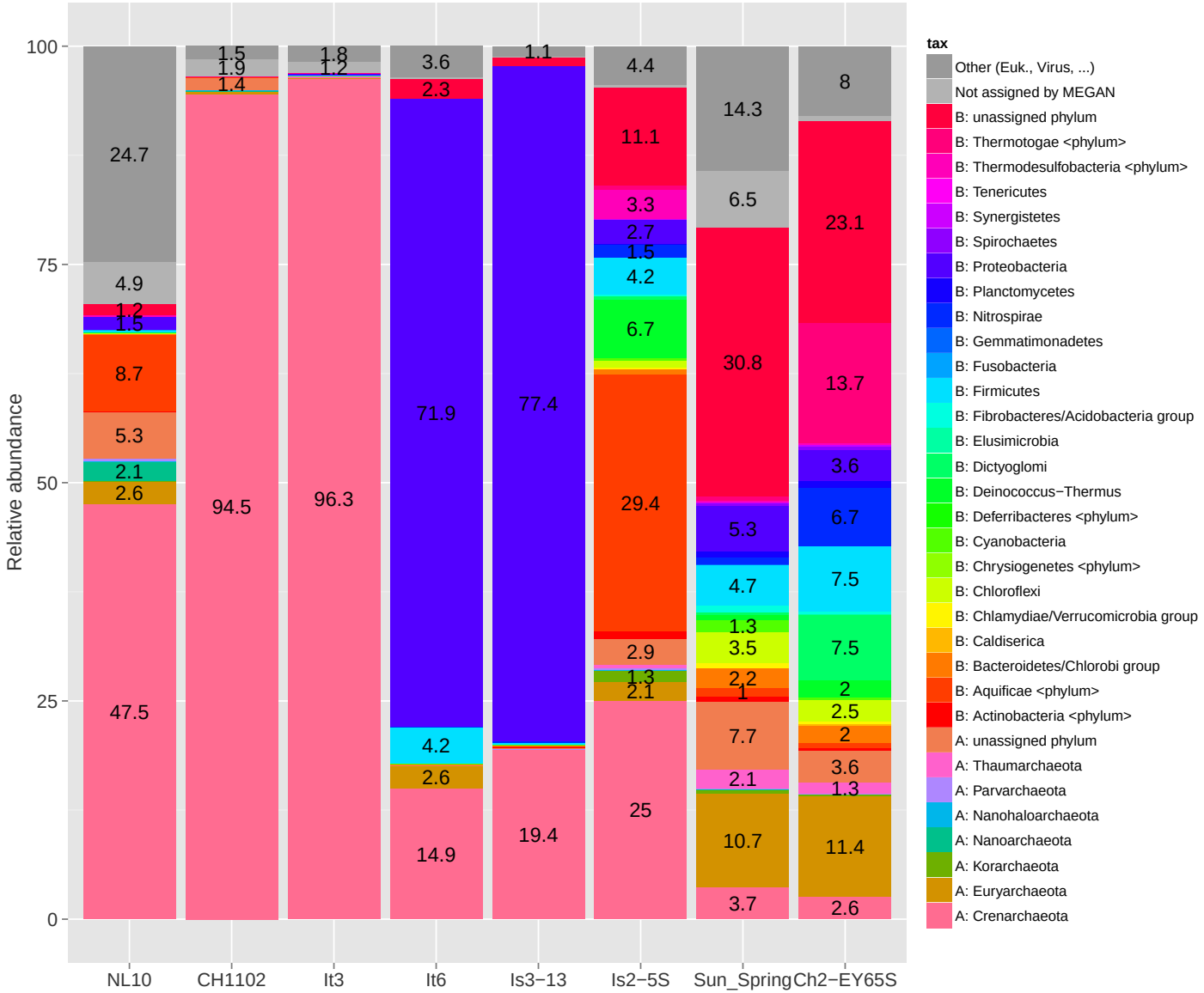
The Hotzyme Project - Metagenomics

Whole genome sequencing of total DNA from samples

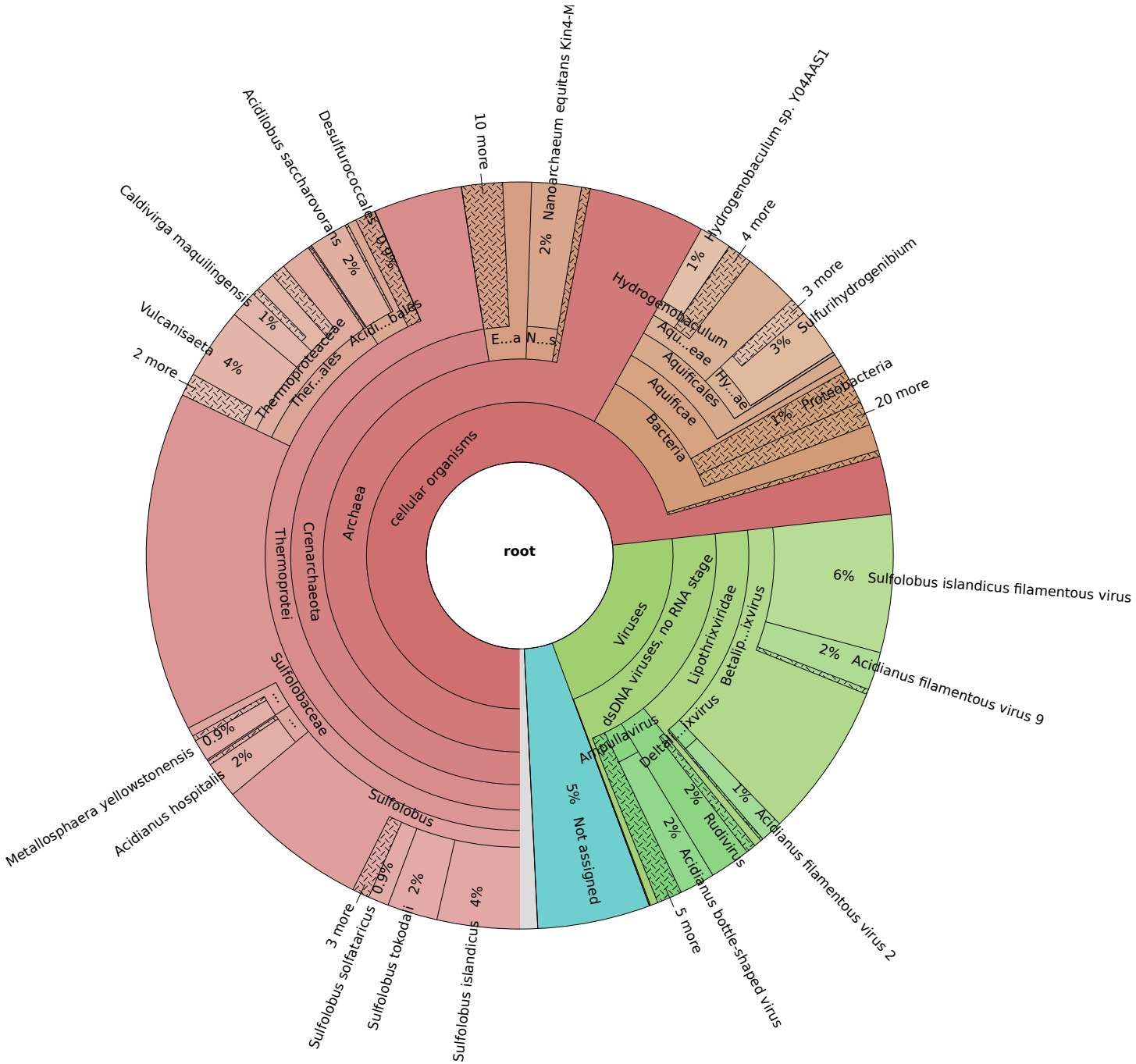
Sample Name	T in °C	pH	Site
Ch2-EY65S	65	7.0	Eryuan, China
Sun Spring	61-64	5.8-6.0	Uzon Caldera, Russia
It6	76	3.0	Pozzuoli, Italy
It3	86	5.5	Pisciarelli, Italy
Is2-5S	85-90	5.0	Grensdalur, Iceland
Is3-13	90	3.5-4.0	Krísuvík, Iceland
CH1102	79	1.8	YNP, USA
NL10	92	3.0-4.0	YNP, USA

Assembly of reads followed by aligning contigs to GenBank NR.

Diversity on phylum-level



P. Menzel et al., Comparative Metagenomics of Eight Geographically Remote Terrestrial Hot Springs. *Microb Ecol.* (2015)


root
cellular organisms
Archaea
Crenarchaeota
Thermoprotei
Sulfolobaceae
Sulfolobus
Sulfolobus islandicus 4%
Sulfolobus tokodaii 2%
Sulfolobus solfataricus 0.9%
3 more
Metallosphaera yellowstonensis 0.9%
Acidianus hospitalis 2%
2 more
Vulcanisaeta 4%
Caldivirga maquilingensis 1%
Ther...ales
Thermoproteaceae
Acidilobus saccharovorans 2%
Desulfurococcales 0.9%
Aci...bales
10 more
E...a
Bacteria
Aquificae
Aquificales
Aqui...eae
Hydrogenobaculum
Hydrogenobaculum sp. Y04AAS1 1%
4 more
Hy...ae
Sulfurihydrogenibium 3%
3 more
Nanoarchaeum equitans Kin4-M 2%
N...s
Proteobacteria 1%
20 more
Viruses
dsDNA viruses, no RNA stage
Lipothrixviridae
Betalip...ixvirus
Sulfolobus islandicus filamentous virus 6%
Acidianus filamentous virus 9 2%
Ixvirus
Rudivirus
Acidianus filamentous virus 2 1%
Acidianus bottle-shaped virus 2%
5 more
Not assigned 5%

Sequence classification

Computational problem

Assign individual HTS reads to taxa by comparison to a reference database. Fast!

Sequence classification

Computational problem

Assign individual HTS reads to taxa by comparison to a reference database. Fast!

Fast = Alignment-free

- Local alignment methods are too slow for dealing with millions of short reads (blastn, blastx, blat, RapSearch, . . .)
- Comparison of genomic k-mers (exact matches of length k) between reads and database
- LMAT, Kraken, Clark, . . .
- choice of k determines accuracy
- at least one k-mer needed per read
- $k = 31$ in Kraken and Clark

Sequence classification

Computational problem

Assign individual HTS reads to taxa by comparison to a reference database. Fast!

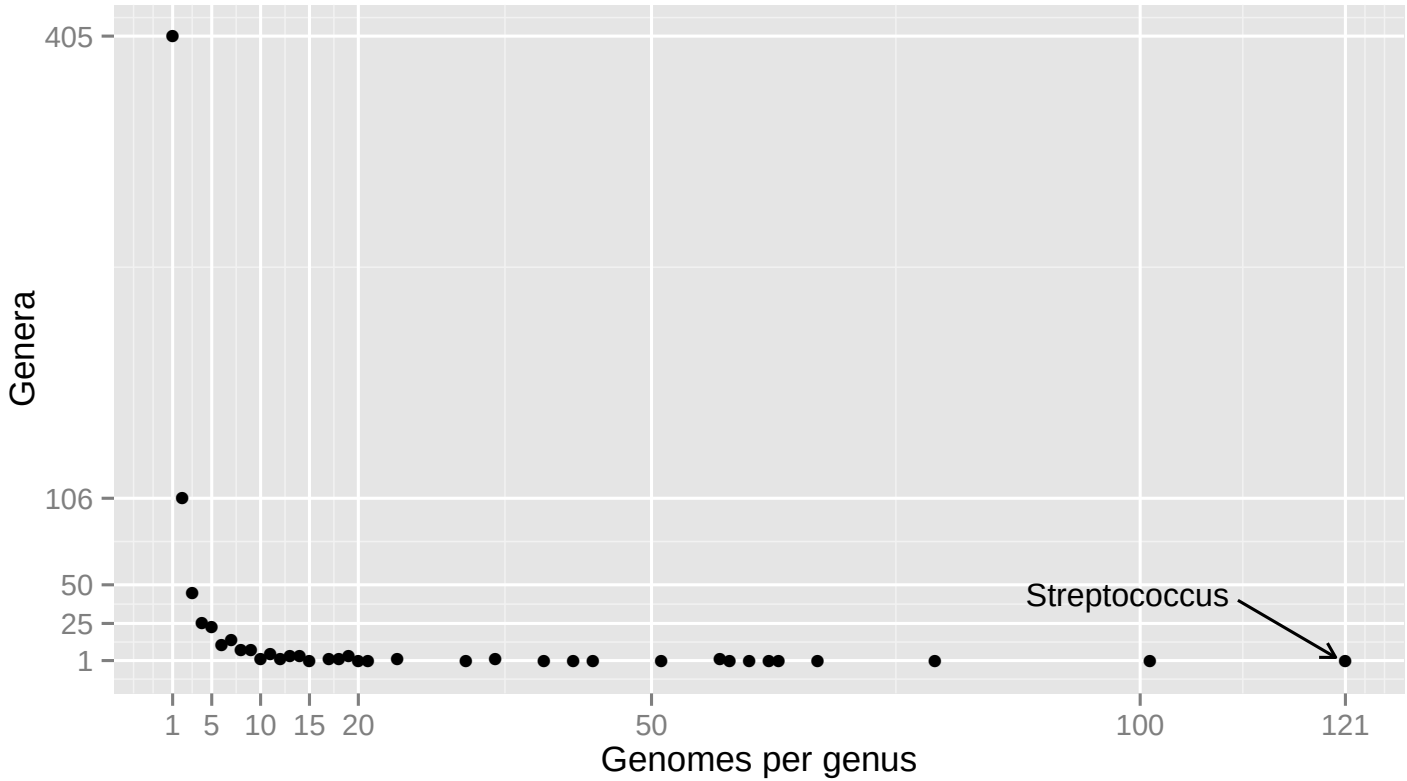
Fast = Alignment-free

- Local alignment methods are too slow for dealing with millions of short reads (blastn, blastx, blat, RapSearch, . . .)
- Comparison of genomic *k*-mers (exact matches of length *k*) between reads and database
- LMAT, Kraken, Clark, . . .
- choice of *k* determines accuracy
- at least one *k*-mer needed per read
- $k = 31$ in Kraken and Clark

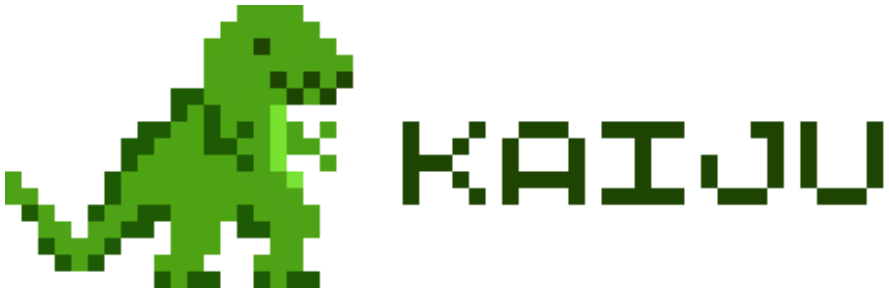
Hotzyme MGs

MG	Kraken
SunSpr	4.3 %
CH2	4.9 %
NL10	9.0 %
It6	14.0 %
CH11	14.7 %
Is3-13	33.5 %
Is2-5S	41.1 %
It3	88.1 %

Reference database bias



- over-representation of certain model organisms, pathogens, . . .
- lack of species that are not possible to culture, extremophiles, . . .
- $>$ 1 500 genera without any sequenced genome
- evolutionary divergence between genomes in database and metagenomes
- requirement of long exact matches (*e.g.* k=31) is very stringent



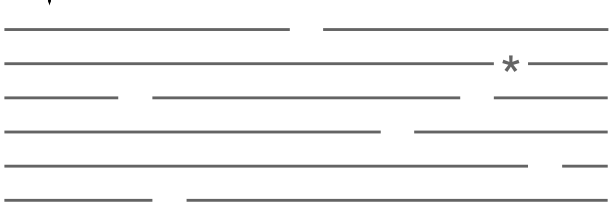
- find maximum exact matches
- compare sequences on protein level
- focus on speed
- Burrows-Wheeler transform + FM-index

Kaiju's algorithm

Sequencing Read

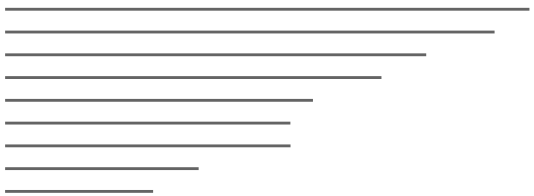
1. Translation

Translate nucleotide sequence into amino acid sequences by the six possible reading frames and split into *fragments* at stop codons.



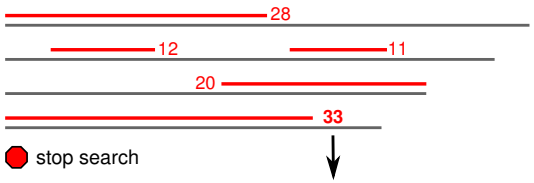
2. Sorting

MEM Sort fragments by length > m



3. Database search

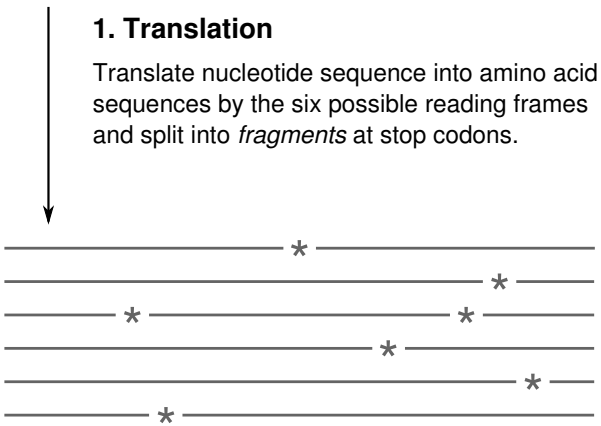
Find MEMs with length > m



assign read to the taxon with longest match

Kaiju's algorithm

Sequencing Read

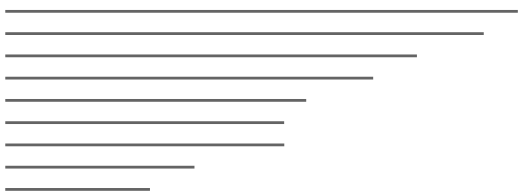


1. Translation

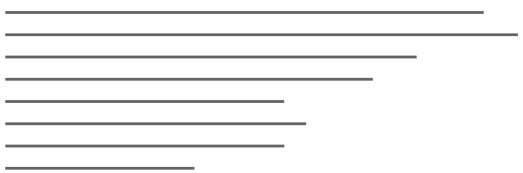
Translate nucleotide sequence into amino acid sequences by the six possible reading frames and split into *fragments* at stop codons.

2. Sorting

MEM Sort fragments by length > m

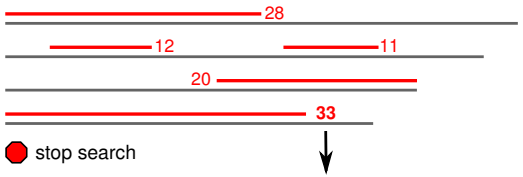


GREEDY Sort fragments by score > s



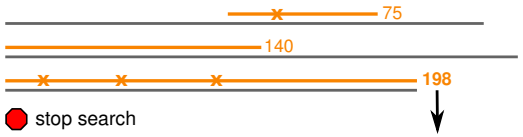
3. Database search

Find MEMs with length > m



assign read to the taxon with longest match

Find matches with score > s



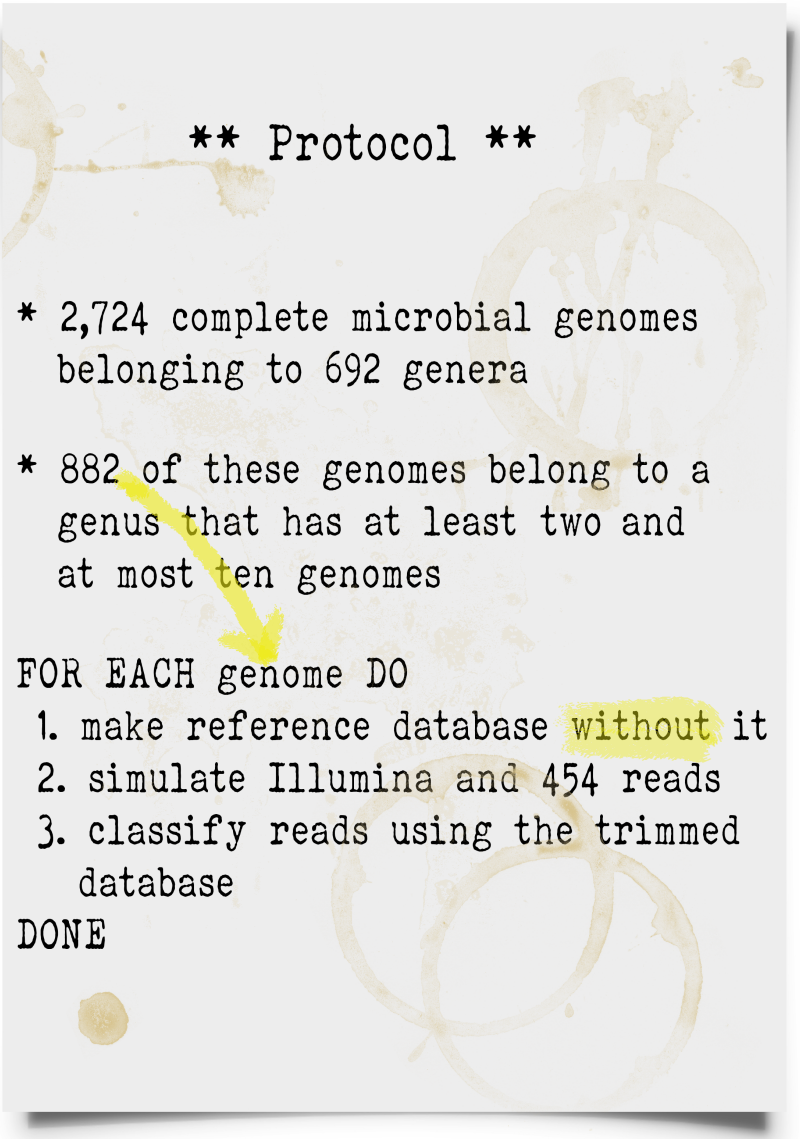
assign read to the taxon with highest scoring match

Benchmark

Simulate the problem of classifying a novel species / strain.

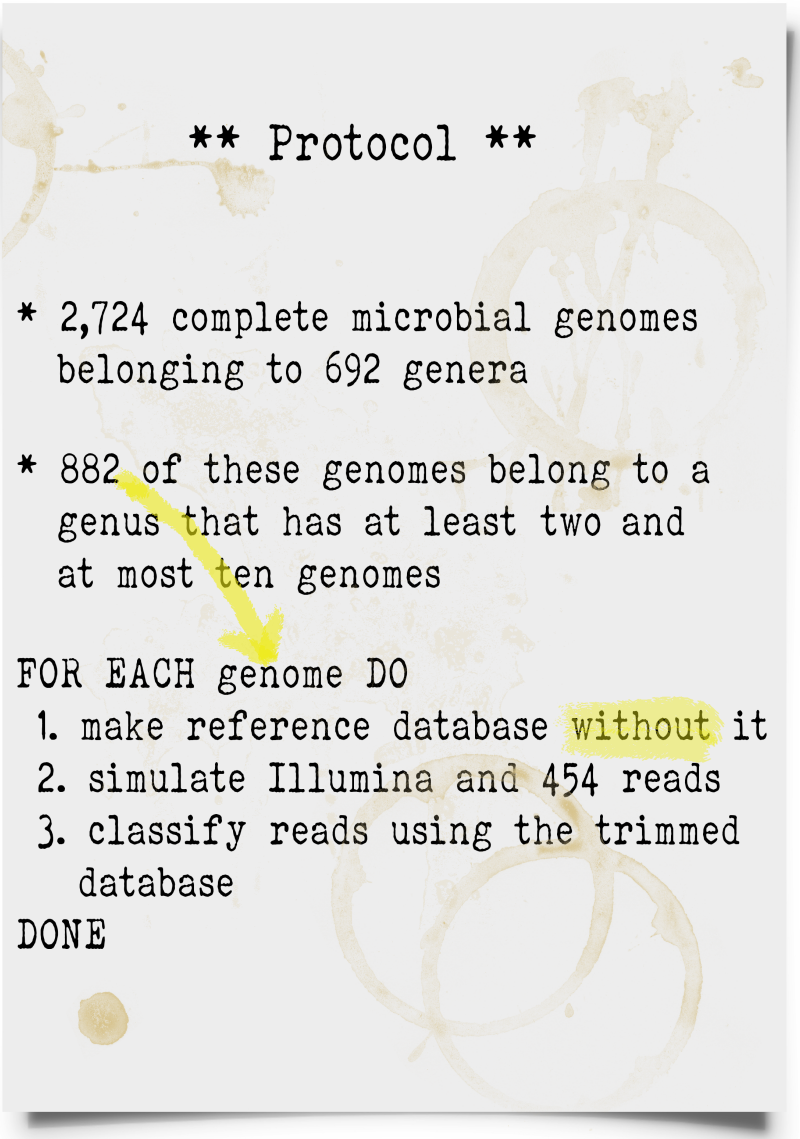
Benchmark

Simulate the problem of classifying a novel species / strain.



Benchmark

Simulate the problem of classifying a novel species / strain.



Programs

Kaiju vs. Kraken vs. Clark

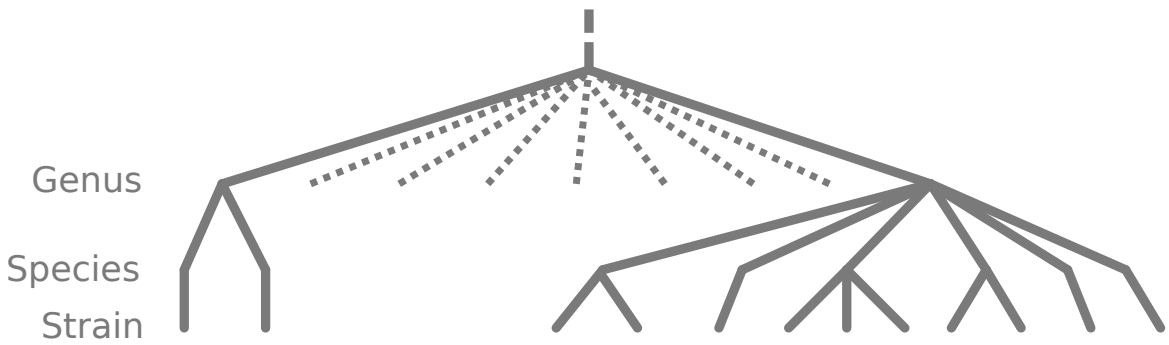
Sensitivity

$$\frac{\text{Number of correct classified}}{\text{Total number of reads}} \times 100$$

Precision

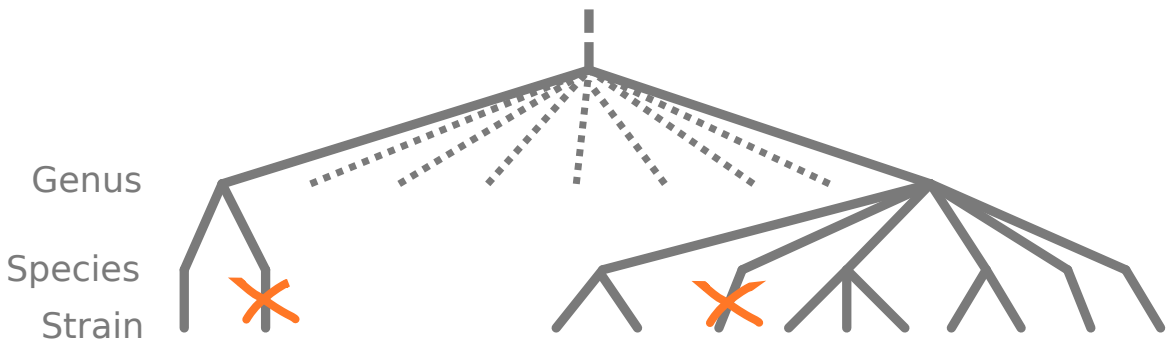
$$\frac{\text{Number of correct classified}}{\text{Number of classified reads}} \times 100$$

Genus categories



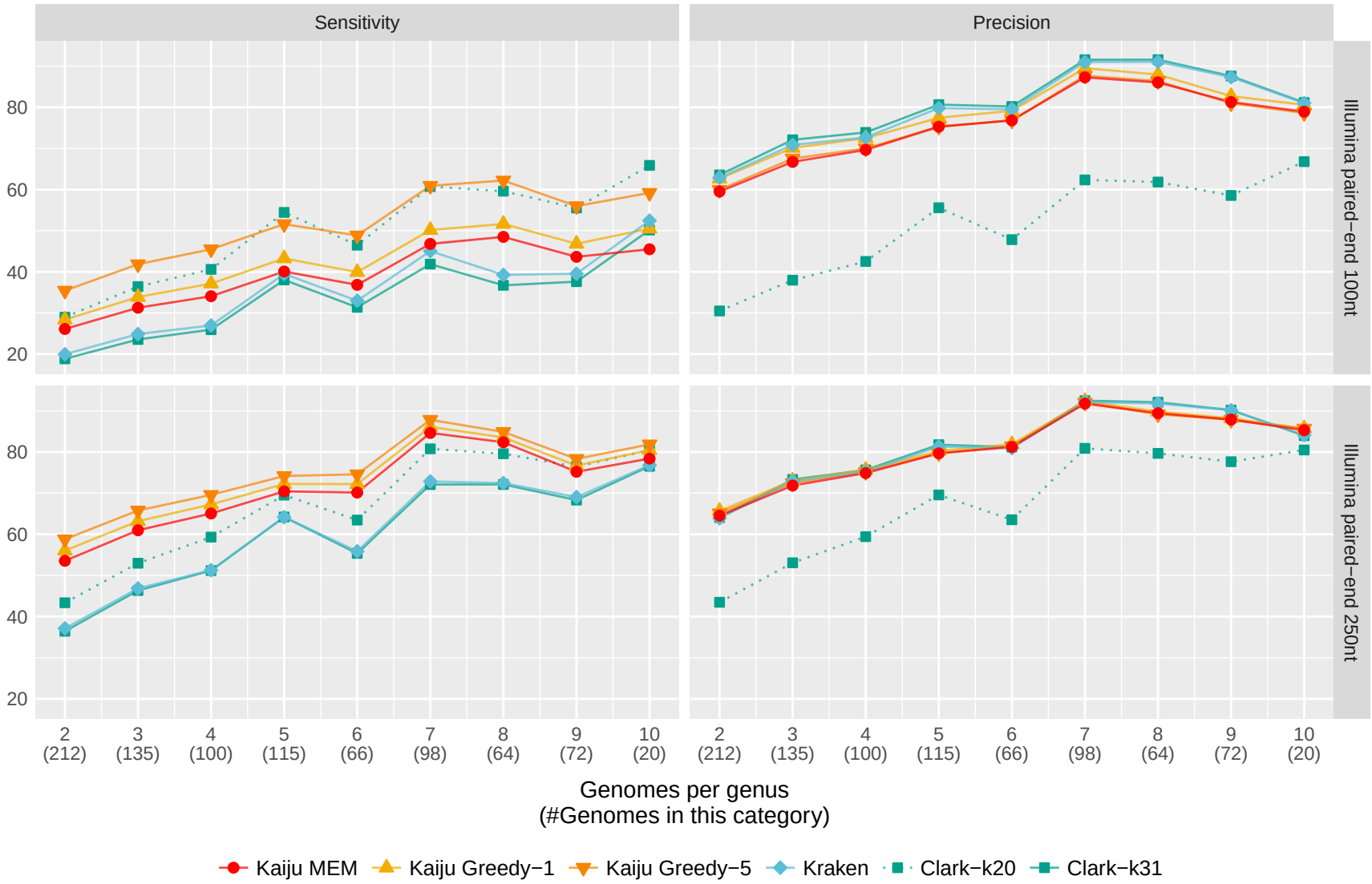
Genomes per genus	2	3	4	5	6	7	8	9	10
(#Genomes in this category)	(212)	(135)	(100)	(115)	(66)	(98)	(64)	(72)	(20)

Genus categories

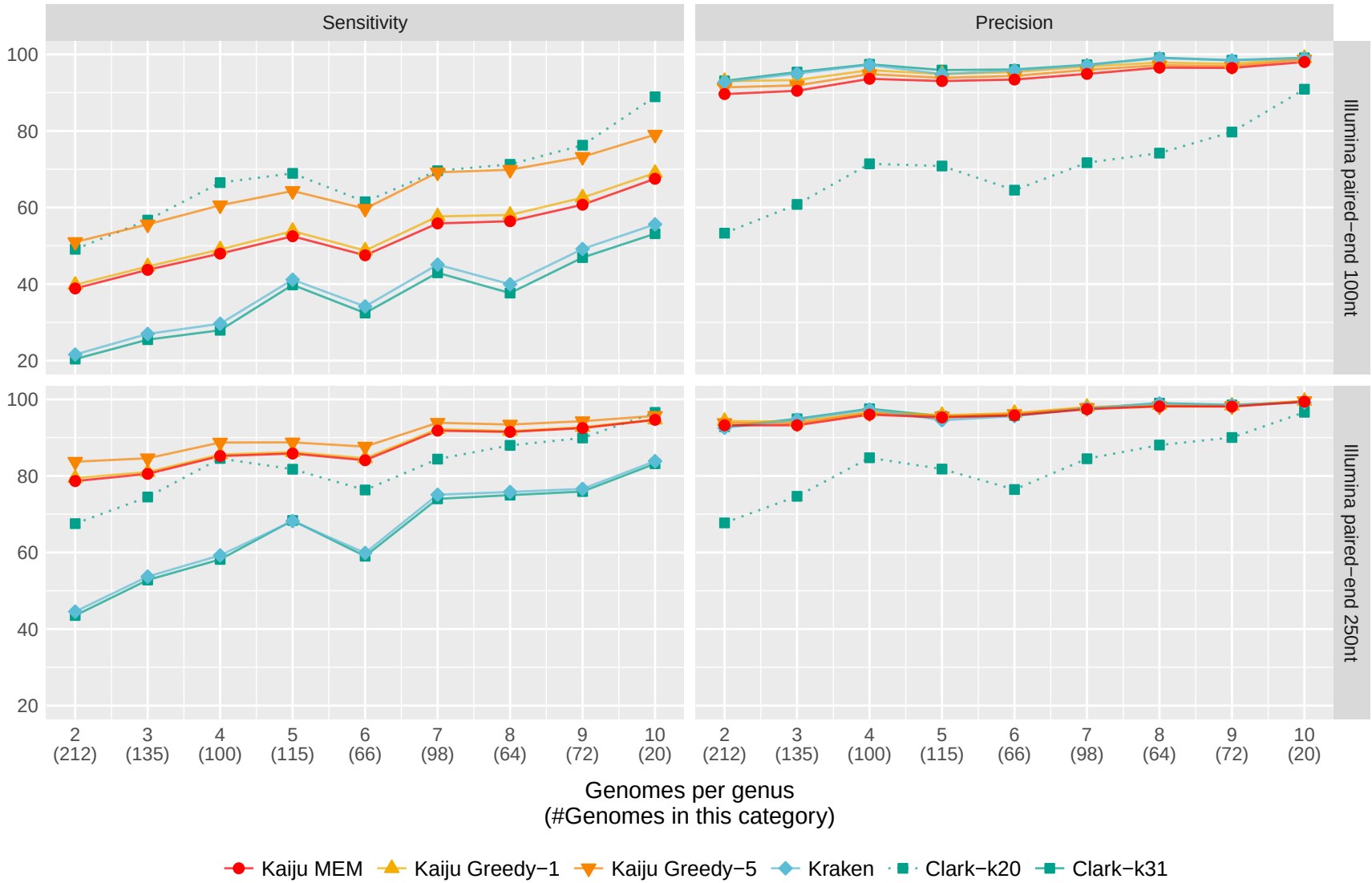


Genomes per genus	2	3	4	5	6	7	8	9	10
(#Genomes in this category)	(212)	(135)	(100)	(115)	(66)	(98)	(64)	(72)	(20)

Sensitivity vs precision on genus-level



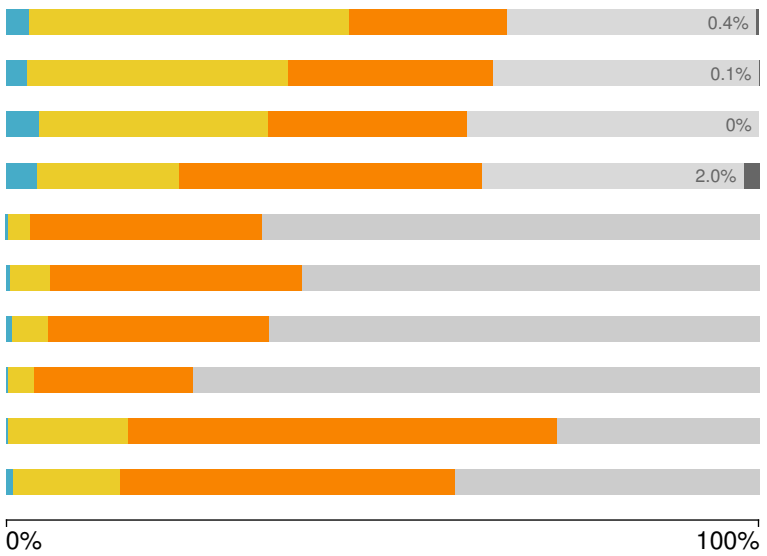
Sensitivity vs precision on phylum-level



Real metagenomes

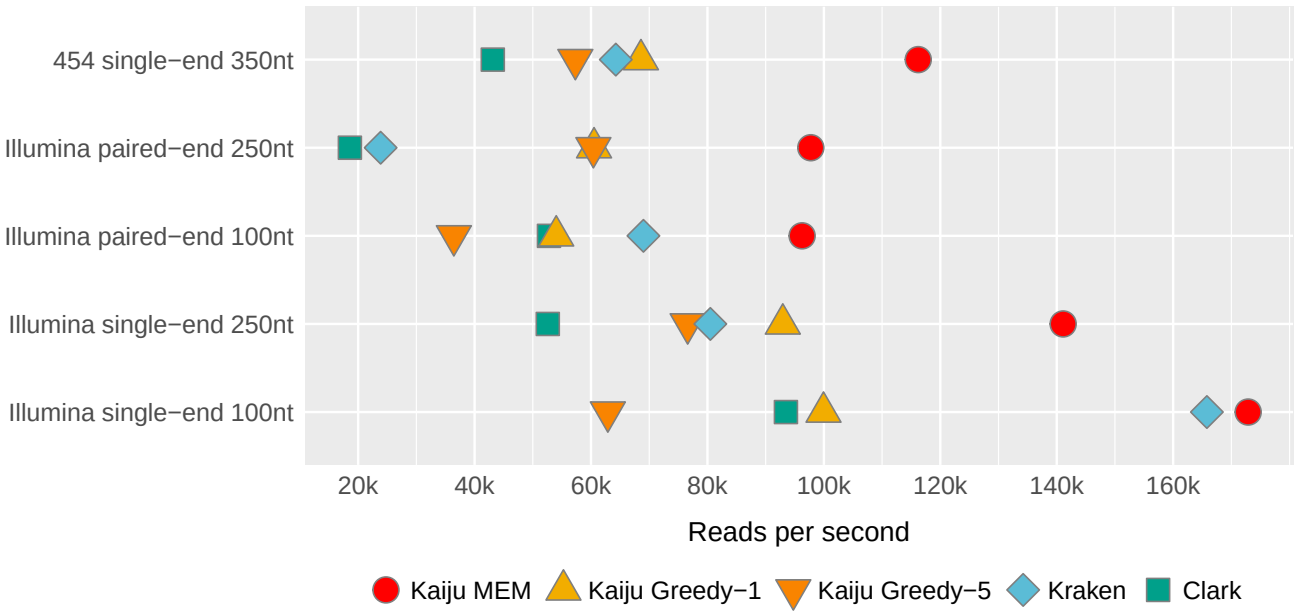
Percentage of classified reads from random metagenomic samples

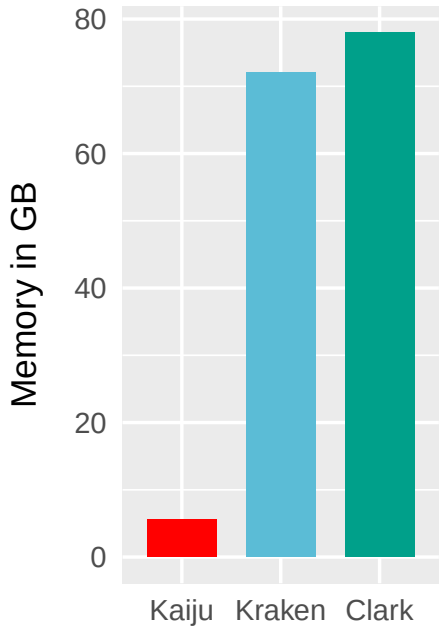
	MEM	Greedy	Kraken	Merged
Human Vagina	58.6	63.3	45.5	66.4
Human Saliva	56.1	61.8	37.4	64.6
Human Gut	51.6	56.7	34.7	61.1
Cat Gut	51.8	58.9	23.0	63.1
Lake	26.5	33.7	3.1	34.0
River Plume	31.8	38.8	5.8	39.3
Baltic Sea water	25.3	33.9	5.6	34.8
Desert Soil	19.3	24.4	3.8	24.7
Bioreactor Sediment	65.0	72.8	16.2	73.1
Bioreactor Compost	50.3	58.5	15.1	59.5
Mean	43.6	50.3	19.0	52.1



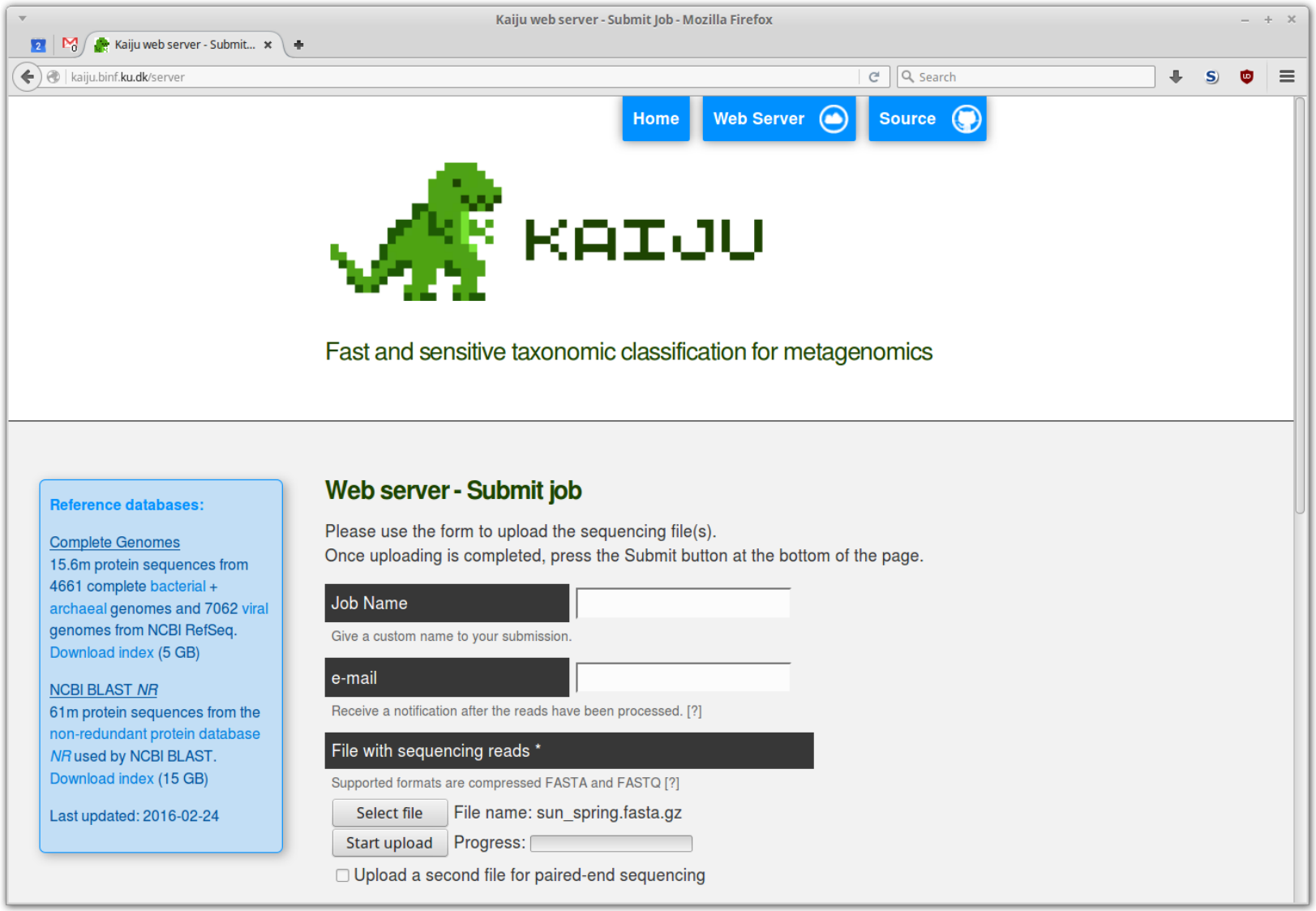
Speed and memory requirement

using a reference database of 2724 complete genomes

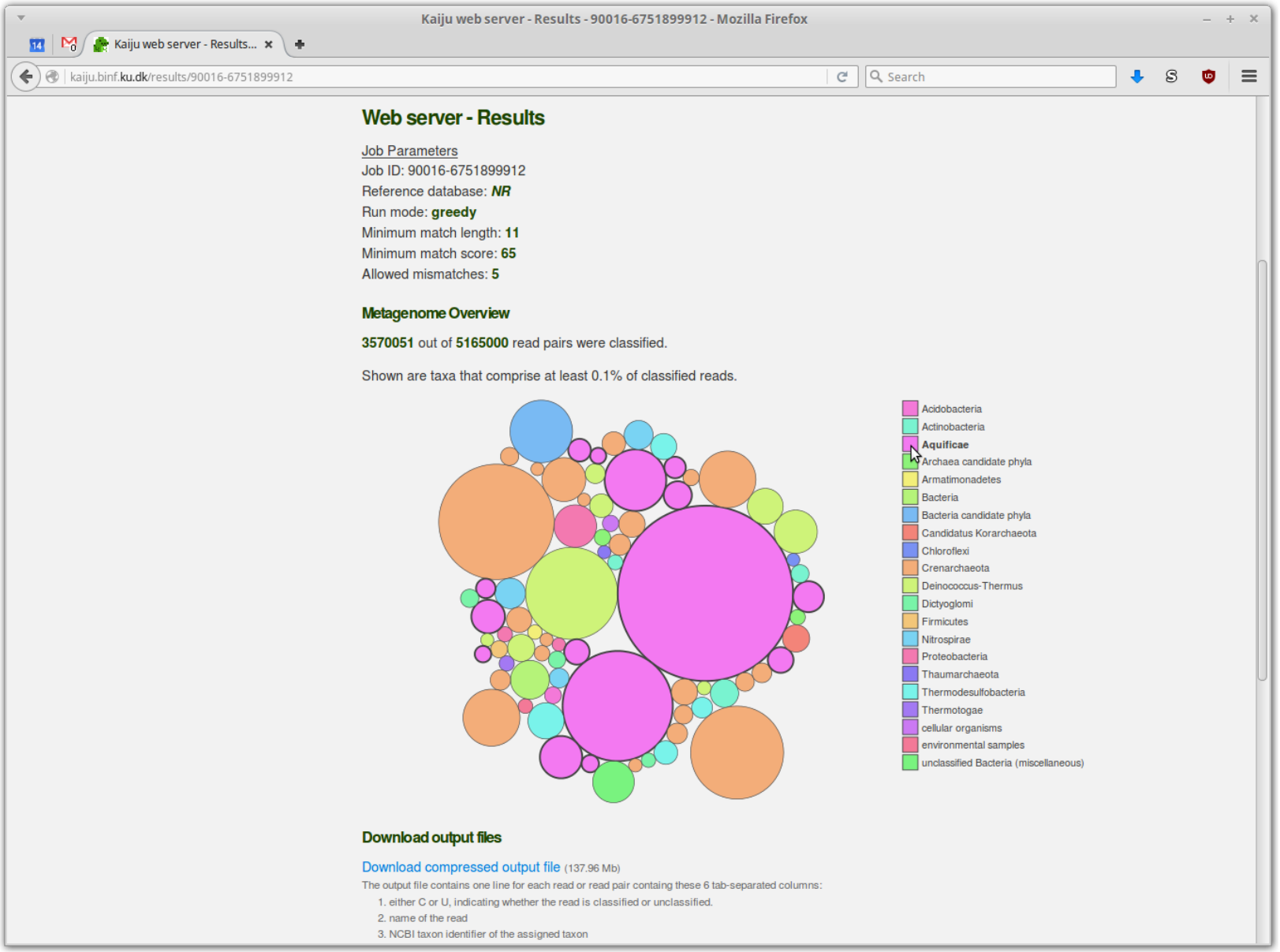




Web server - http://kaiju.binf.ku.dk



Web server - Results page



Conclusion (last slide!)

Hotzyme MGs

MG	Kraken	Kaiju
SunSpr	4.3%	43.1%
CH2	4.9%	31.9%
NL10	9.0%	29.8%
It6	14.0%	49.7%
CH11	14.7%	65.5%
Is3-13	33.5%	63.4%
Is2-5S	41.1%	55.8%
It3	88.1%	92.1%

Conclusion (last slide!)

Hotzyme MGs

MG	Kraken	Kaiju
SunSpr	4.3%	43.1%
CH2	4.9%	31.9%
NL10	9.0%	29.8%
It6	14.0%	49.7%
CH11	14.7%	65.5%
Is3-13	33.5%	63.4%
Is2-5S	41.1%	55.8%
It3	88.1%	92.1%

Summary

- BWT + FM-index can be used with a database of all microbial/viral proteins
- very fast and small index size
- can also use full microbial BLAST *nr*
- Benchmark shows higher sensitivity with similar precision
- Up to ten times more reads classified in real metagenomes

Menzel P., Ng K.L., Krogh A. (2016) Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nature Communications*, 7:11257

Anders Krogh

Kim Ng