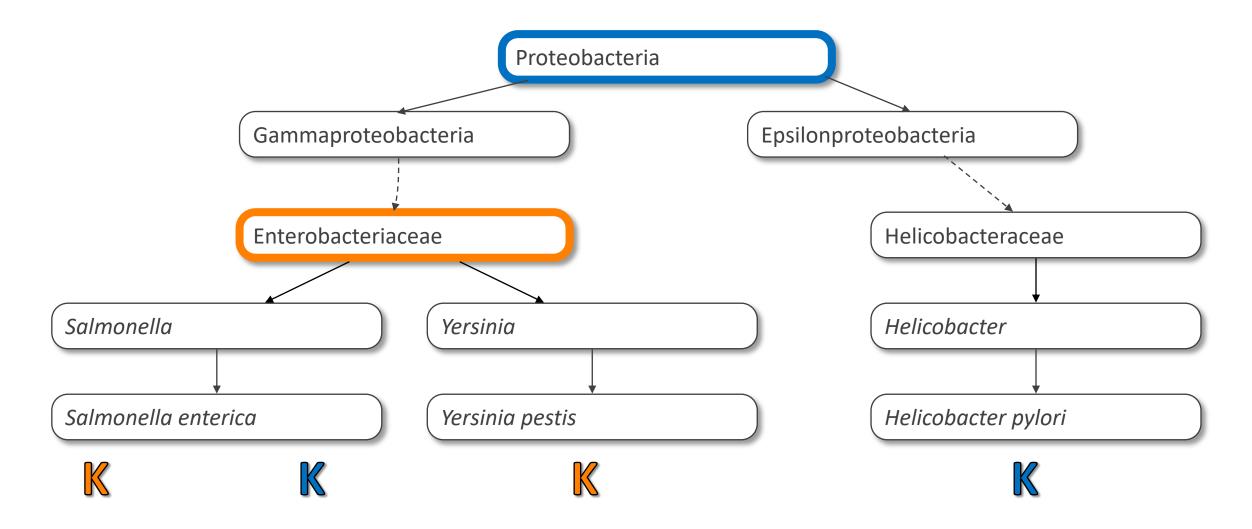
Fast Metagenomic Profiling Methods

JENNIFER LU

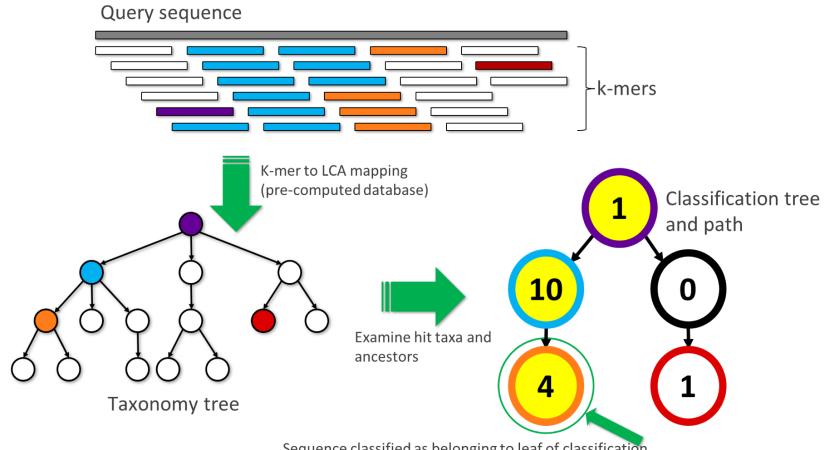
PHD CANDIDATE, BIOMEDICAL ENGINEERING

SALZBERG LAB, JOHNS HOPKINS UNIVERSITY

How does Kraken work?



How does Kraken work?



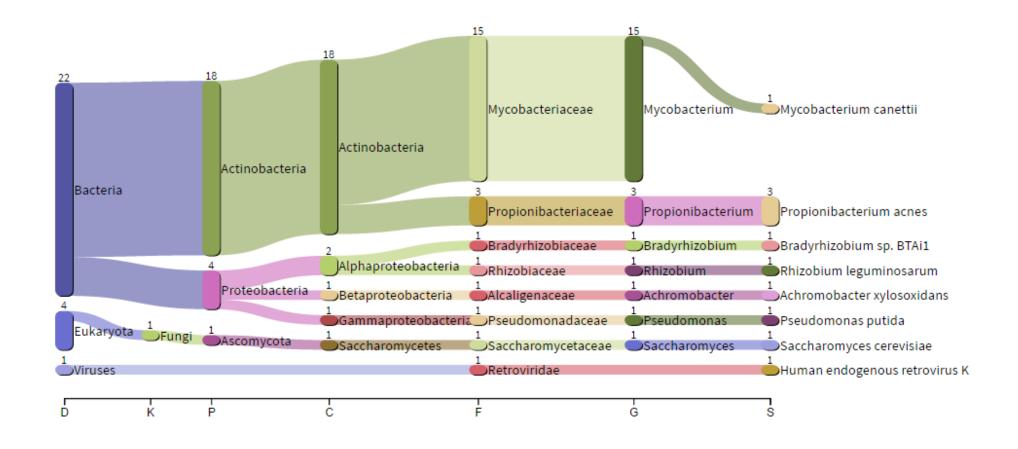
Sequence classified as belonging to leaf of classification (highest-weighted RTL) path

D. Wood and S.L. Salzberg, Genome Biology 2014

Kraken Output: Text Report

27.45	7319764	1567	Р	1239	Firmicutes
20.93	5581780	2613	С	91061	Bacilli
16.04	4276646	1127	0	1385	Bacillales
10.82	2884169	830	F	186817	Bacillaceae
10.81	2883248	1623	G	1386	Bacillus
7.74	2063734	635215	-	86661	Bacillus cereus group
4.08	1086754	28739	S	1396	Bacillus cereus
0.7	187882	20115	S	1428	Bacillus thuringiensis
0.57	151306	151097	S	1392	Bacillus anthracis
0.01	1934	0	S	86662	Bacillus weihenstephanensis
0	643	0	S	580165	Bacillus cytotoxicus
1.06	281618	0	S	79880	Bacillus clausii
1.01	269485	0	S	86665	Bacillus halodurans
1	266768	1270	-	653685	Bacillus subtilis group
1	265462	0	S	1423	Bacillus subtilis
0	23	0	S	492670	Bacillus methylotrophicus
0	13	0	S	1402	Bacillus licheniformis
0	20	0	S	1408	Bacillus pumilus

Kraken Output: Visualized



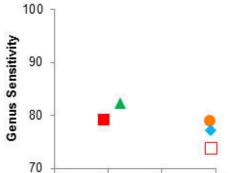
Zor i companison manch in other

classifiers

Tests on Hi-Seq Generated Data

90

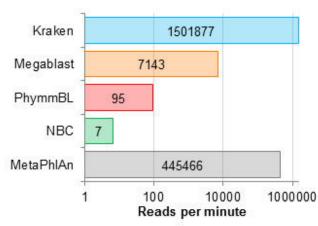
100



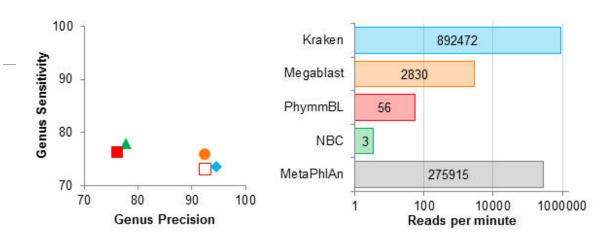
80

Genus Precision

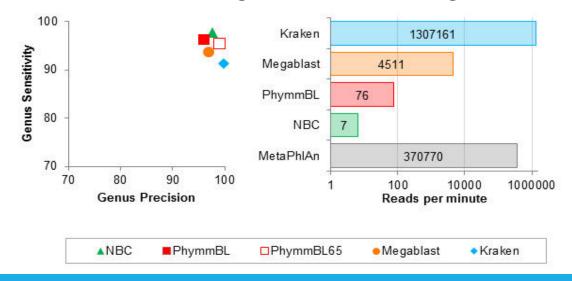
70



Tests on Mi-Seq Generated Data

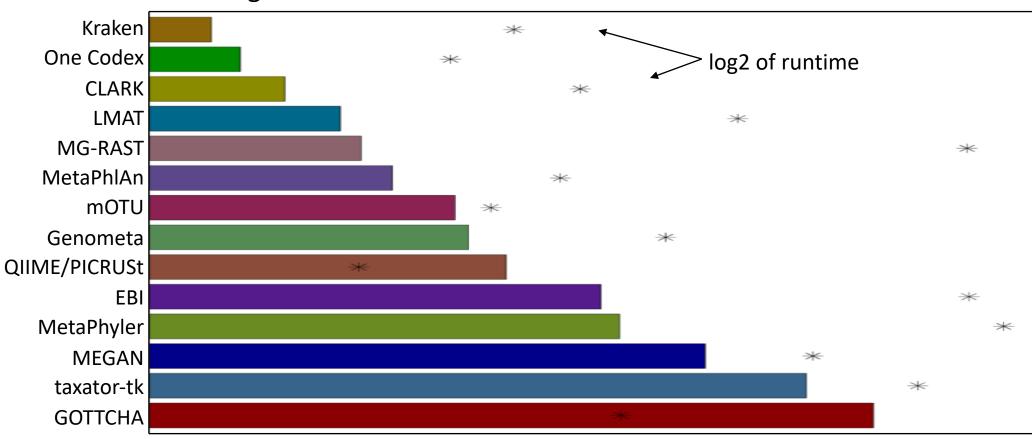


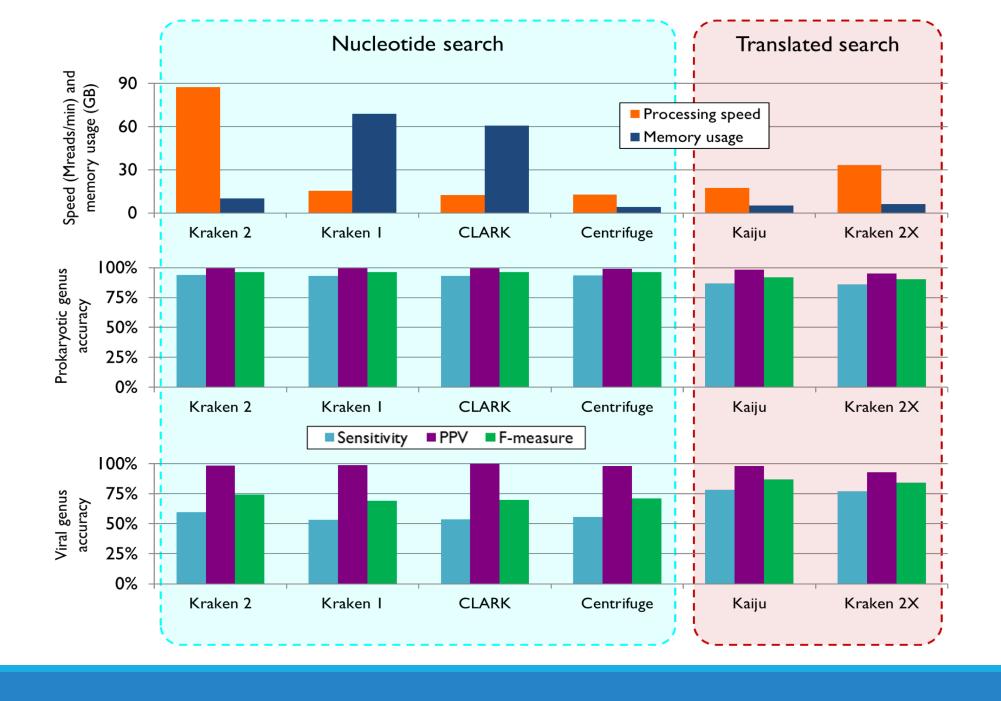
Tests on Simulated Metagenomic Data with high error rates



The first "unbiased, comprehensive benchmark of metagenome analysis tools in which the authors are not involved in any of the tools tested."







Kraken	Database	Build time*	Classification Time**	Database Size	MiSeq Genus			HiSeq Genus		
Version		(hours)	(sec/10M reads)	(GB)	Sensitivity	Precision	F1 Score	Sensitivity	Precision	F1 Score
Kraken 1	Standard 1 ^A	11.7	59.2	217.3	88.4%	98.6%	93.2%	88.3%	99.2%	93.5%
	MiniKraken1 ^A 8Gb	2.0	57.9	8.5	80.4%	99.4%	88.9%	71.1%	99.4%	82.9%
	Standard 2 ^B	15.8	63.6	240.8	88.2%	98.5%	93.1%	88.0%	99.3%	93.3%
	MiniKraken2 ^B 8Gb	2.1	58.6	8.5	79.0%	99.4%	88.0%	68.4%	99.4%	81.1%
Kraken 2	Standard 1 ^A	1.1	20.2	31.2	89.4%	97.9%	93.5%	89.6%	99.1%	94.1%
	MiniKraken1 ^A 8Gb	0.5	16.6	8.0	85.4%	98.6%	91.6%	84.0%	99.3%	91.0%
	Standard 2 ^B	3.9	26.0	34.7	89.3%	97.8%	93.3%	89.2%	99.1%	93.9%
	MiniKraken2 ^B 8Gb	0.7	24.3	8.0	84.8%	98.4%	91.1%	82.6%	99.3%	90.2%

^{*}Build time calculated with 32 threads

^{**}Classification time calculated with 16 threads

^AStandard 1 = refseq archaea, bacteria, viral

^BStandard 2 = refseq archaea, bacteria, viral, and human

DBs built in	n 09/2018								
Database		Kra	Kraken 2.0			ŀ	Kraken 1.0		
		Build Time	Build Time hash.k2d		Build Time database.idx		database.kdb		
		[minutes]	[MB]		[minutes]	[(GB]	[GB]	
16S	Greengenes	1.2		92.8		N/A	N/A	N/A	
16S –	Silva	0.8		151.6		N/A	N/A	N/A	
16S	RDP	12.8		222.7		N/A	N/A	N/A	
Database		Kraken 2.0			Kraken 1.0				
		Build Time hash.k2d		Build Time database.idx		database.kdb	refseq DB Contents		
		[hours]	[GB]		[hours]	[(GB]	[GB]	
refseq	Standard 1	1.1		31.2		11.7	8.6	208.7	archaea, bacteria, viral
refseq	MiniKraken1_8Gb	0.5		8.0		2.0	0.5	8.1	archaea, bacteria, viral
refseq	Standard 2	3.9		34.7		15.8	8.6	232.2	archaea, bacteria, viral, human
refseq	MiniKraken2_8Gb	0.7		8.0		2.1	0.5	8.1	archaea, bacteria, viral, human
									archaea, bacteria, fungi, plant, plasmid,
refseq	RefseqMicrobe	3.7		73.0		36.0	8.6	497.7	protozoa, viral, human
									archaea, bacteria, plant, plasmid, viral,
									vertebrate-mammalian, vertebrate-other,
refseq	EupathCleaningDB ^A	9.6		183.0		N/A	27.0	1297.0	human
									archaea, bacteria, fungi, plant, plasmid,
									protozoa, viral, vertebrate-mammalian,
refseq	RefseqAll	11.8		180.8		N/A	N/A	N/A	vertebrate-other, human
nt	nt	11.8		123.0		N/A	N/A	N/A	
protein	nr	19.1		89.0		N/A	N/A	N/A	

AKraken 1 EupathCleaningDB sizes are the combined sizes of three separate databases. A single database requires too much memory for a 1024 GB memory machine

Command Lines

Building:

- > kraken2-build --download-library bacteria --db \${DATABASE} [--use-ftp, default: rsync]
- > kraken2-build --build --db \${DATABASE} --threads 4

Running:

> kraken2 --db \${DATABASE} --report mysample.report --threads 4 mysample.fasta > mysample.kraken

Abundance Estimation: Bracken

> python est_abundance.py –i mysample.report –o mysample.bracken -k database150mers.kmer_dstrib

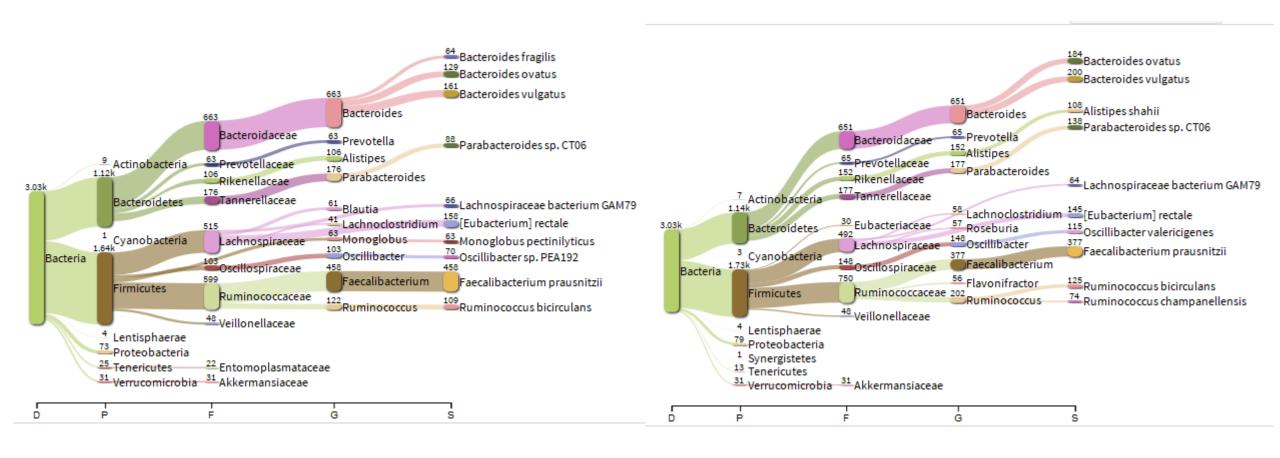
Visualization:

Pavian

MiniKraken Runtimes

```
> kraken --db ../m3taxworkshop/databases/kraken/minikraken1 8GB/ --threads 4 ../m3taxworkshop/data/1-
datasets/hmp/stool sample subset rep set filtered final.fna > hmp.kraken
3032 sequences (1.54 Mbp) processed in 0.632s (287.6 Kseq/m, 145.69 Mbp/m).
 3030 sequences classified (99.93%)
 2 sequences unclassified (0.07%)
> kraken-report -db ../m3taxworkshop/databases/kraken/minikraken1_8GB hmp.kraken > hmp.report
> kraken2 --db ../m3taxworkshop/databases/kraken/minikraken2 8GB/ --threads 4 ../m3taxworkshop/data/1-
datasets/hmp/stool sample subset rep set filtered final.fna > hmp.kraken2
Loading database information... done.
3032 sequences (1.54 Mbp) processed in 0.147s (1239.1 Kseq/m, 627.59 Mbp/m).
 3032 sequences classified (100.00%)
 0 sequences unclassified (0.00%)
```

HMP MiniKraken1 v MiniKraken2





Kraken 2 Databases Supported

Download of Standard Genome Sets Supported:

- Refseq archaea, bacteria, plasmid, viral, fungi, plant, protozoa
- NCBI nr, nt, env_nr, env_nt, UniVec, UniVec_Core
- GRCh38 human

Special Databases: Kraken 2 also provides support for

- 16S databases: SILVA, Greengenes, and RDP
- Protein databases [similar to BLASTX].

Kraken 2 uses a reduced amino acid alphabet for protein databases to represent each amino acid with 4 bits. For classification, Kraken classifies all six frames of translation of the query sequence.

01/09/2019 Jennifer Lu, jlu26@jhmi.edu

Kraken Resources

Kraken: https://ccb.jhu.edu/software/kraken/

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

Pavian: https://github.com/fbreitwieser/pavian

Bracken: https://ccb.jhu.edu/software/bracken/

Bracken Paper: https://peerj.com/articles/cs-104

Kraken-HLL: https://github.com/fbreitwieser/krakenhll

Kraken2: https://ccb.jhu.edu/software/kraken2/

For more information: https://ccb.jhu.edu/software/choosing-a-metagenomics-classifier/

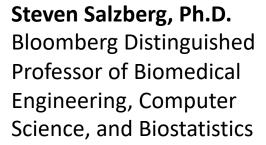
Kraken Team











Principal Investigator (PI)

https://salzberg-lab.org



Derrick Wood, Ph.D.
Principal Bioinformatics
Scientist at Personal
Genome Diagnostics

Developer of Kraken v1/v2



Florian Breitwieser, Ph.D.
Postdoctoral Fellow in the
Center for Computational
Biology at JHUSOM

Developer of Centrifuge, Pavian, and Kraken-HLL

http://ccb.jhu.edu/people/florian/



Jennifer Lu
Ph.D. Candidate in
Biomedical Engineering
at JHU/JHUSOM

Developer of Bracken, Kraken v1/2 Point of Contact

http://ccb.jhu.edu/people/jennifer.lu/