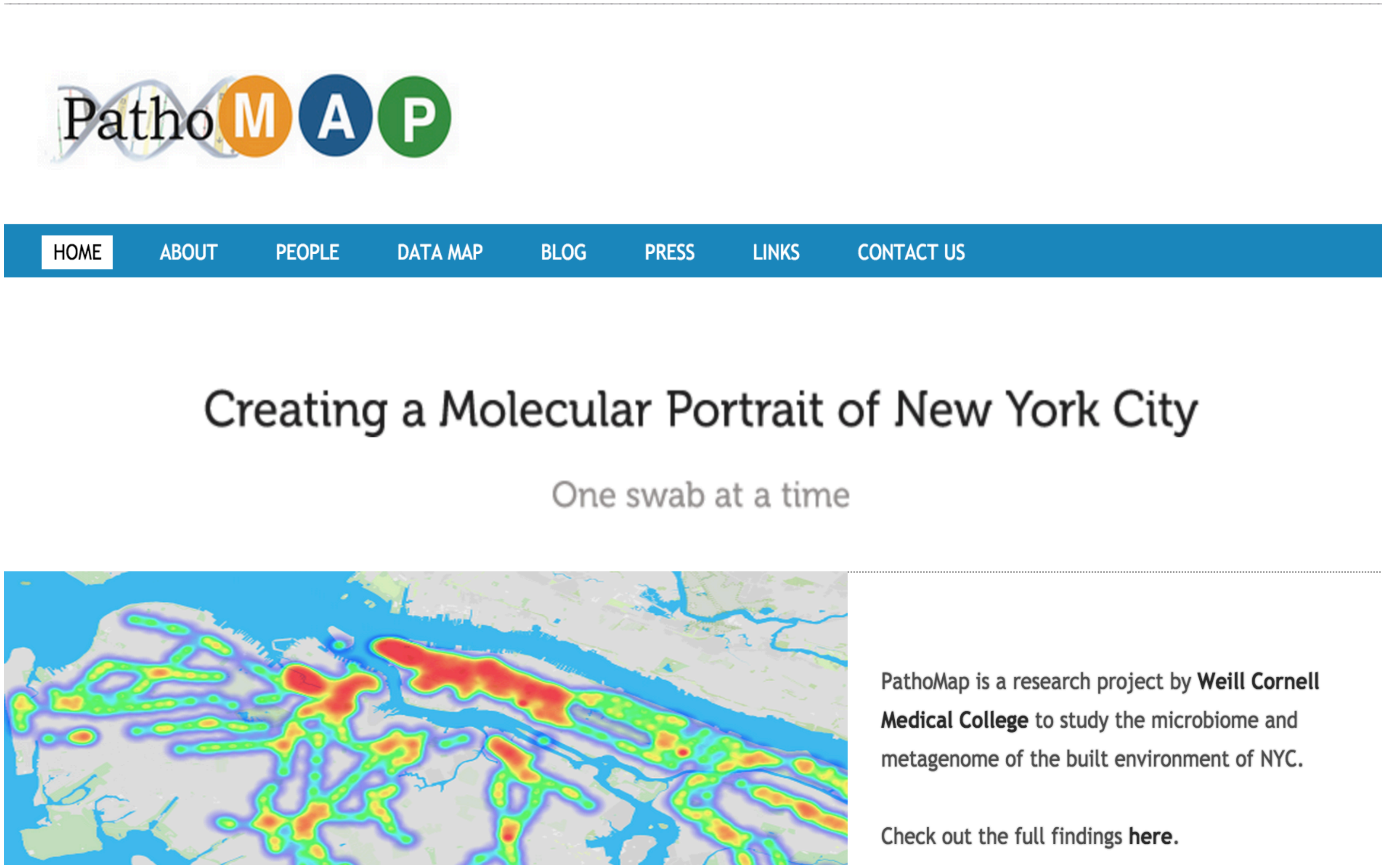
Benchmarking and ensemble approaches for metagenomic classification

Alexa McIntyre
Mason Lab, Weill Cornell Medicine

Tyler Hicks/The New York Times, 2008

PathoMap

New York City Subway



Creating a Molecular Portrait of New York City

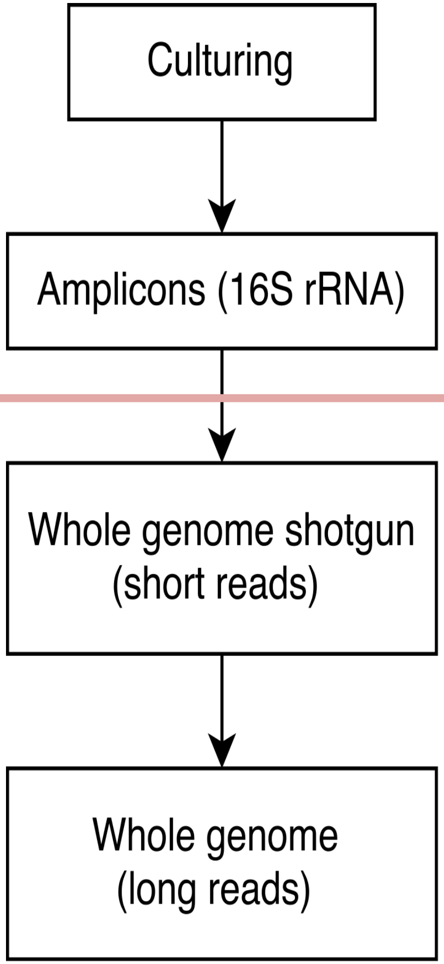
One swab at a time

PathoMap is a research project by **Weill Cornell Medical College** to study the microbiome and metagenome of the built environment of NYC.

Check out the full findings **here**.

Afshinnekoo, et al., 2015

A brief history of microbial genomics



Culturing

Amplicons (16S rRNA)

Whole genome shotgun (short reads)

Whole genome (long reads)

- c. 1960’s onwards [1]
- Est. 1% species culturable [2]

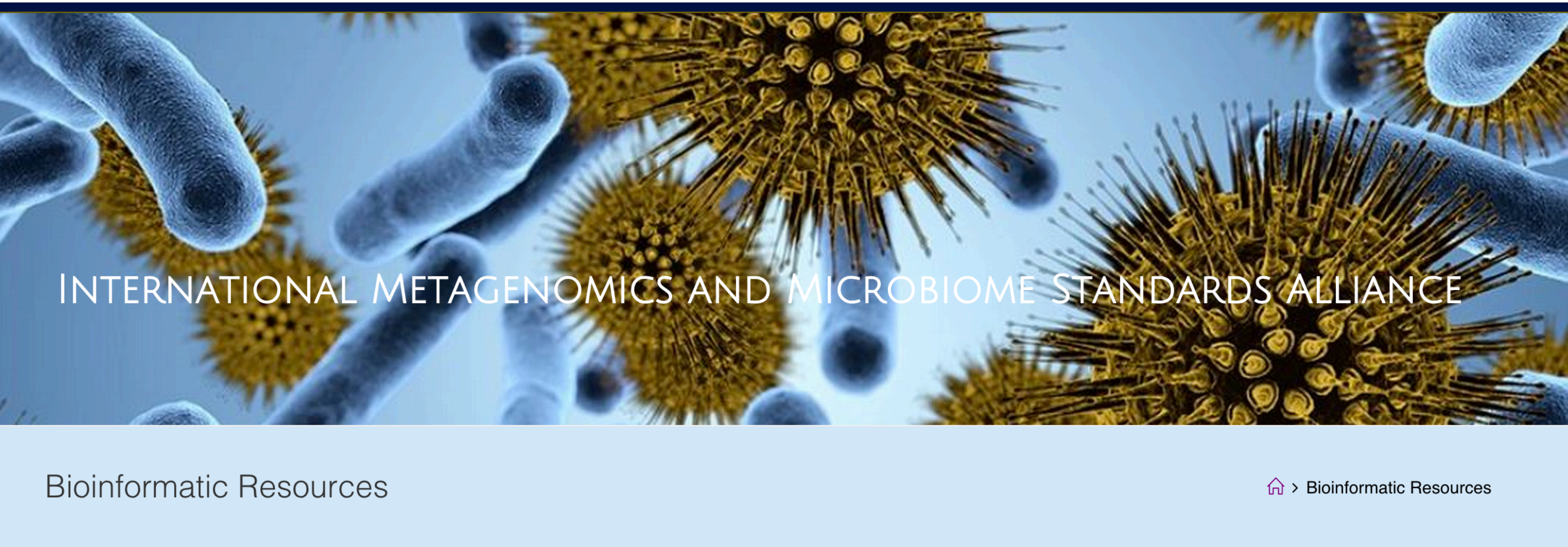
- c. 1990’s onwards [3]
- Lack of truly universal primers, amplification biases [2]
- Low species/strain resolution

- c. 2000’s onwards
- Species ambiguity
- Smaller databases

- Enables the detection of species at lower abundances [4]

[1] Shine and Dalgarno, 1975, [2] Amann et al., 1995, [3] Weisberg et al., 1991, [4] Kuleshov et al., 2016

Too many tools, too few comparisons



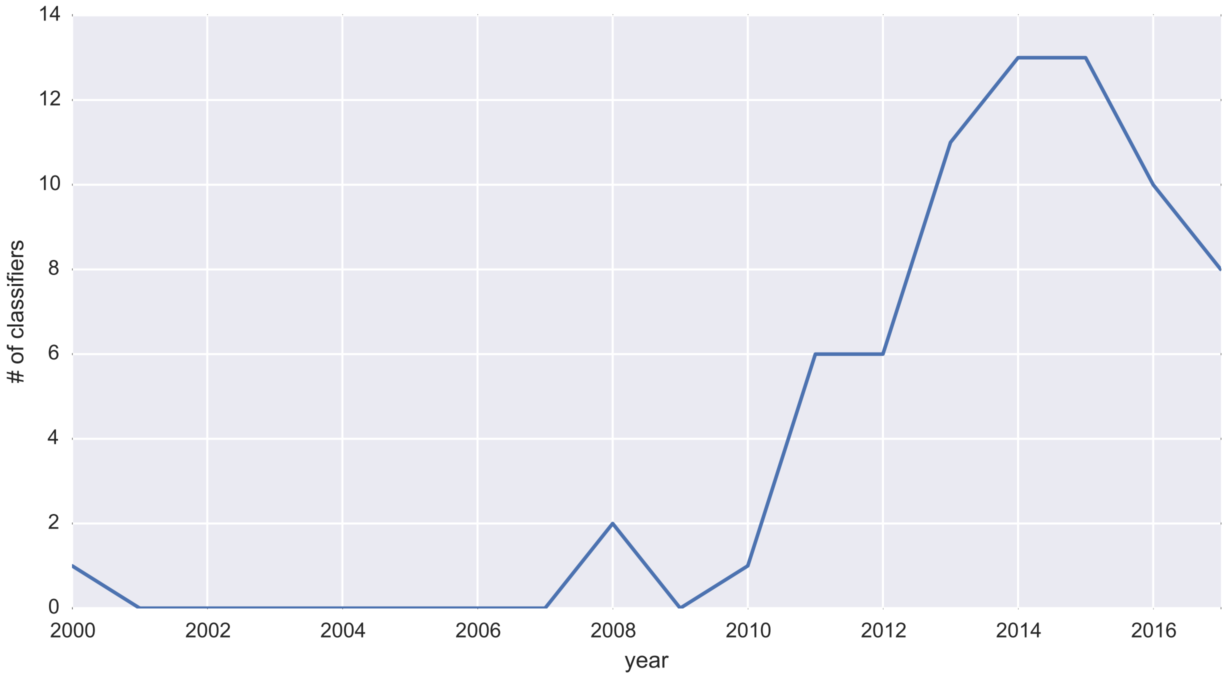
INTERNATIONAL METAGENOMICS AND MICROBIOME STANDARDS ALLIANCE

Bioinformatic Resources

⌂ > Bioinformatic Resources

IMMSA (2017):
At least 71 tools available for profiling microbial communities using WGS
microbialstandards.org/index.php/bioinformatic-resources

Too many tools, too few comparisons



IMMSA (2017):

At least 71 tools available for profiling microbial communities using WGS

microbialstandards.org/index.php/bioinformatic-resources

11 selected tools

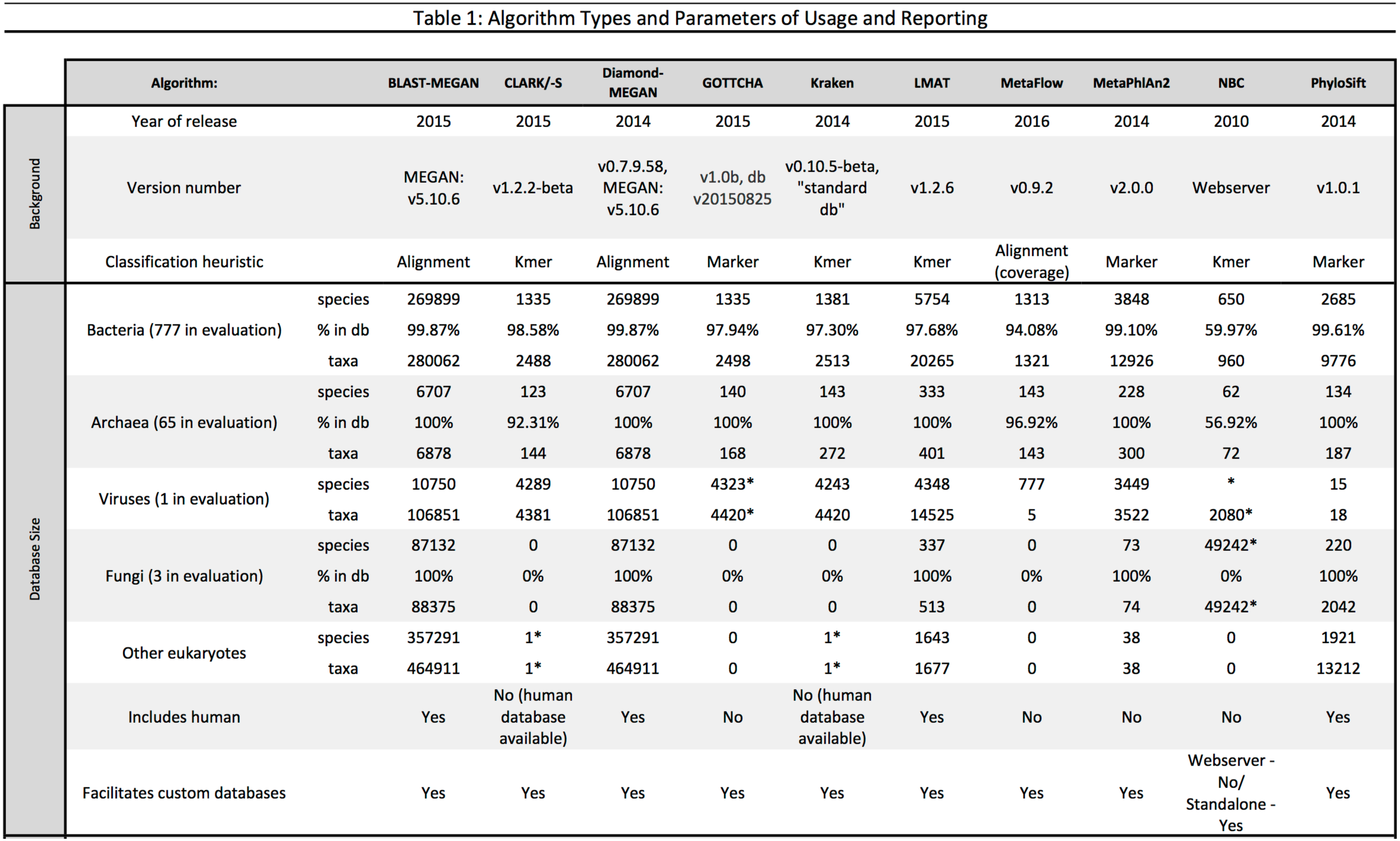
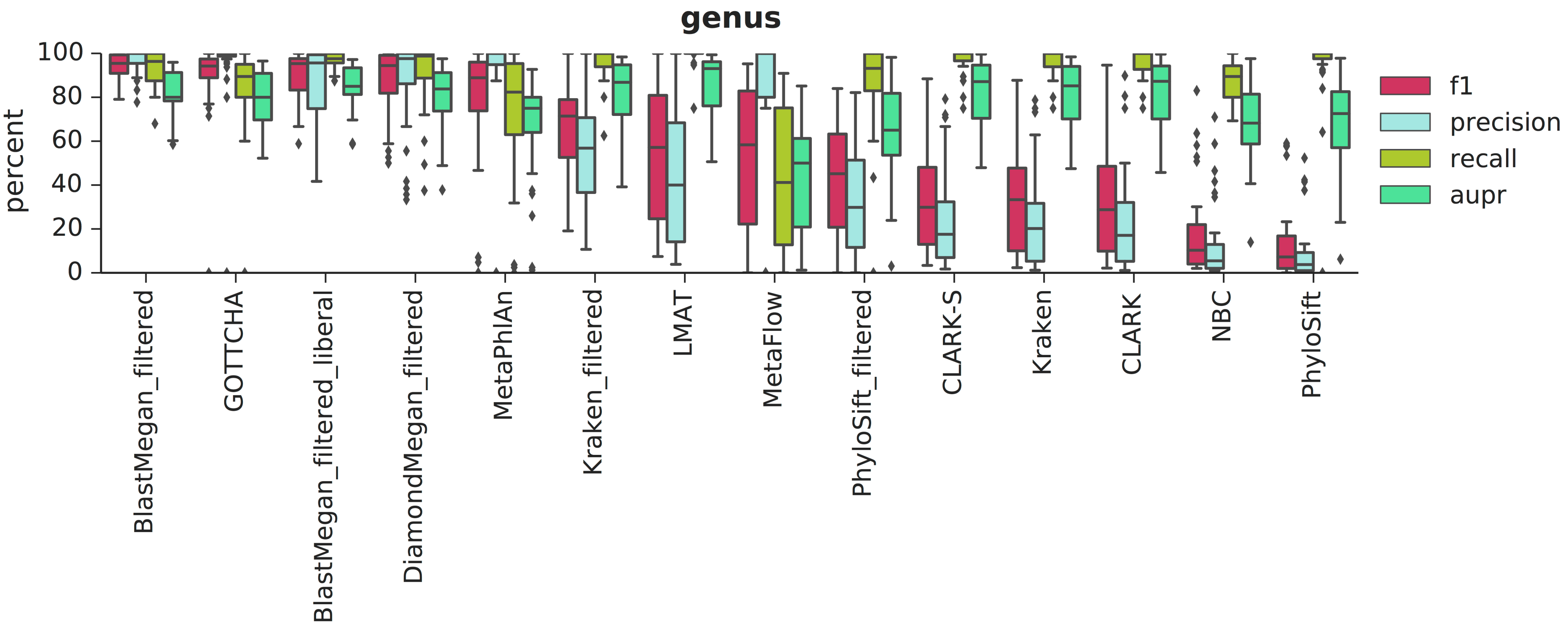


Table 1: Algorithm Types and Parameters of Usage and Reporting

	Algorithm:		BLAST-MEGAN	CLARK/-S	Diamond-MEGAN	GOTTCHA	Kraken	LMAT	MetaFlow	MetaPhlAn2	NBC	PhyloSift
Background	Year of release		2015	2015	2014	2015	2014	2015	2016	2014	2010	2014
	Version number		MEGAN: v5.10.6	v1.2.2-beta	v0.7.9.58, MEGAN: v5.10.6	v1.0b, db v20150825	v0.10.5-beta, "standard db"	v1.2.6	v0.9.2	v2.0.0	Webserver	v1.0.1
	Classification heuristic		Alignment	Kmer	Alignment	Marker	Kmer	Kmer	Alignment (coverage)	Marker	Kmer	Marker
Database Size	Bacteria (777 in evaluation)	species	269899	1335	269899	1335	1381	5754	1313	3848	650	2685
		% in db	99.87%	98.58%	99.87%	97.94%	97.30%	97.68%	94.08%	99.10%	59.97%	99.61%
		taxa	280062	2488	280062	2498	2513	20265	1321	12926	960	9776
	Archaea (65 in evaluation)	species	6707	123	6707	140	143	333	143	228	62	134
		% in db	100%	92.31%	100%	100%	100%	100%	96.92%	100%	56.92%	100%
		taxa	6878	144	6878	168	272	401	143	300	72	187
	Viruses (1 in evaluation)	species	10750	4289	10750	4323*	4243	4348	777	3449	*	15
		taxa	106851	4381	106851	4420*	4420	14525	5	3522	2080*	18
	Fungi (3 in evaluation)	species	87132	0	87132	0	0	337	0	73	49242*	220
		% in db	100%	0%	100%	0%	0%	100%	0%	100%	0%	100%
		taxa	88375	0	88375	0	0	513	0	74	49242*	2042
	Other eukaryotes	species	357291	1*	357291	0	1*	1643	0	38	0	1921
		taxa	464911	1*	464911	0	1*	1677	0	38	0	13212
	Includes human		Yes	No (human database available)	Yes	No	No (human database available)	Yes	No	No	No	Yes
	Facilitates custom databases		Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Webserver - No/ Standalone - Yes	Yes

Performance profiles across 35 datasets



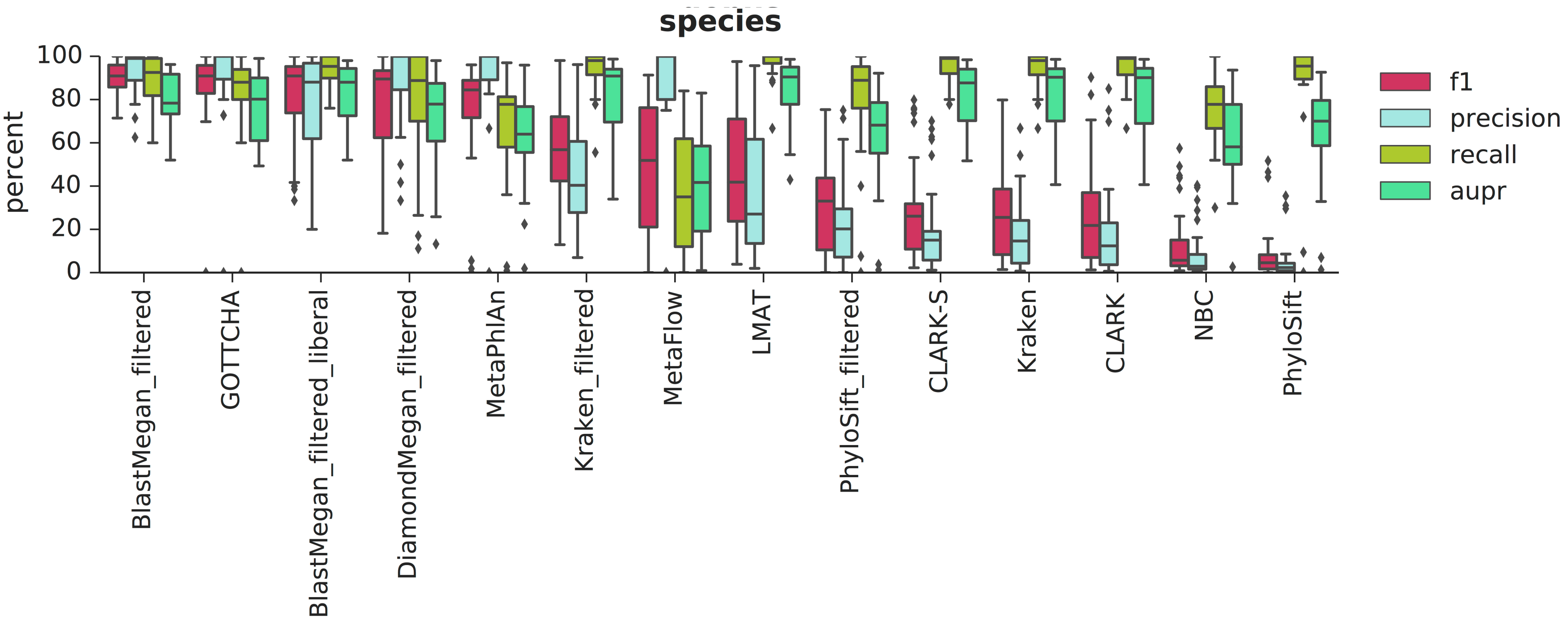
Precision = false positive rate = TP/(TP+FP)

Recall = sensitivity = TP/(TP+FN)

F1 score = 2(precision*recall)/(precision+recall)

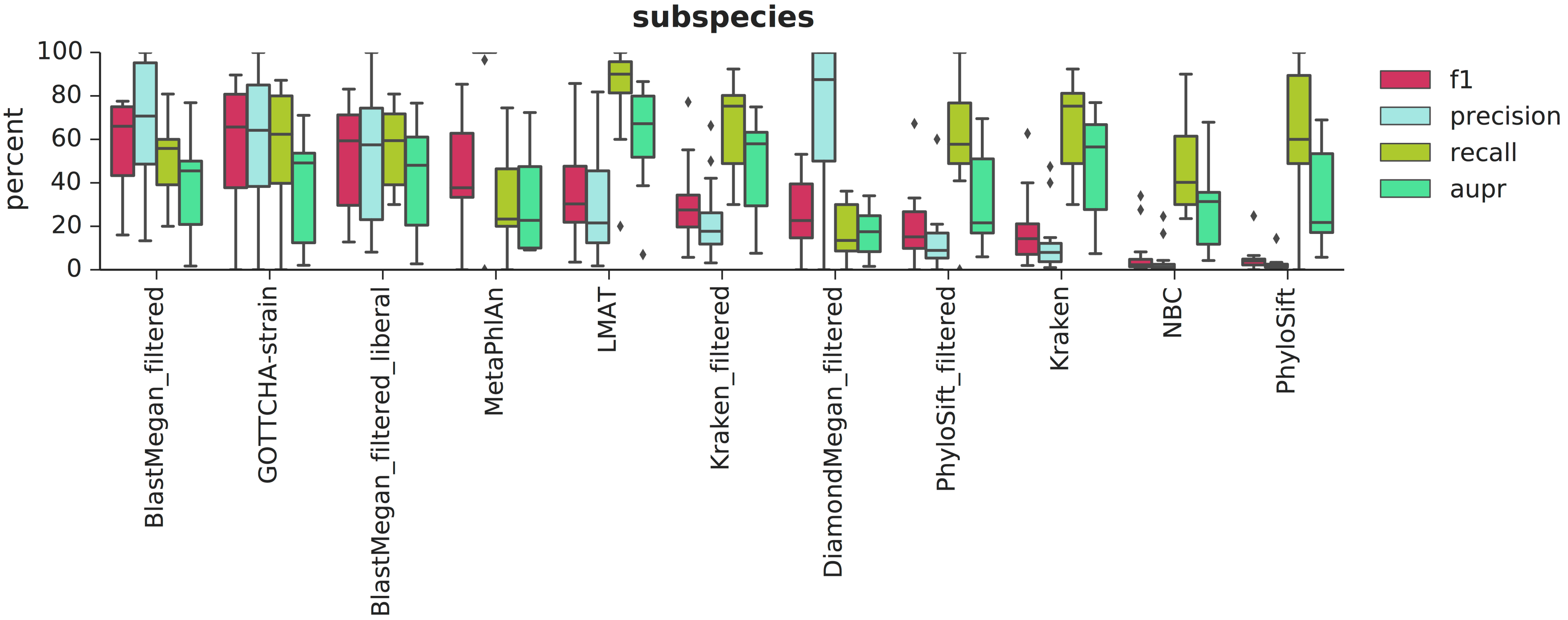
AUPR = area under the precision recall curve

Performance profiles across 35 datasets



Precision = false positive rate = TP/(TP+FP)
Recall = sensitivity = TP/(TP+FN)
F1 score = 2(precision*recall)/(precision+recall)
AUPR = area under the precision recall curve

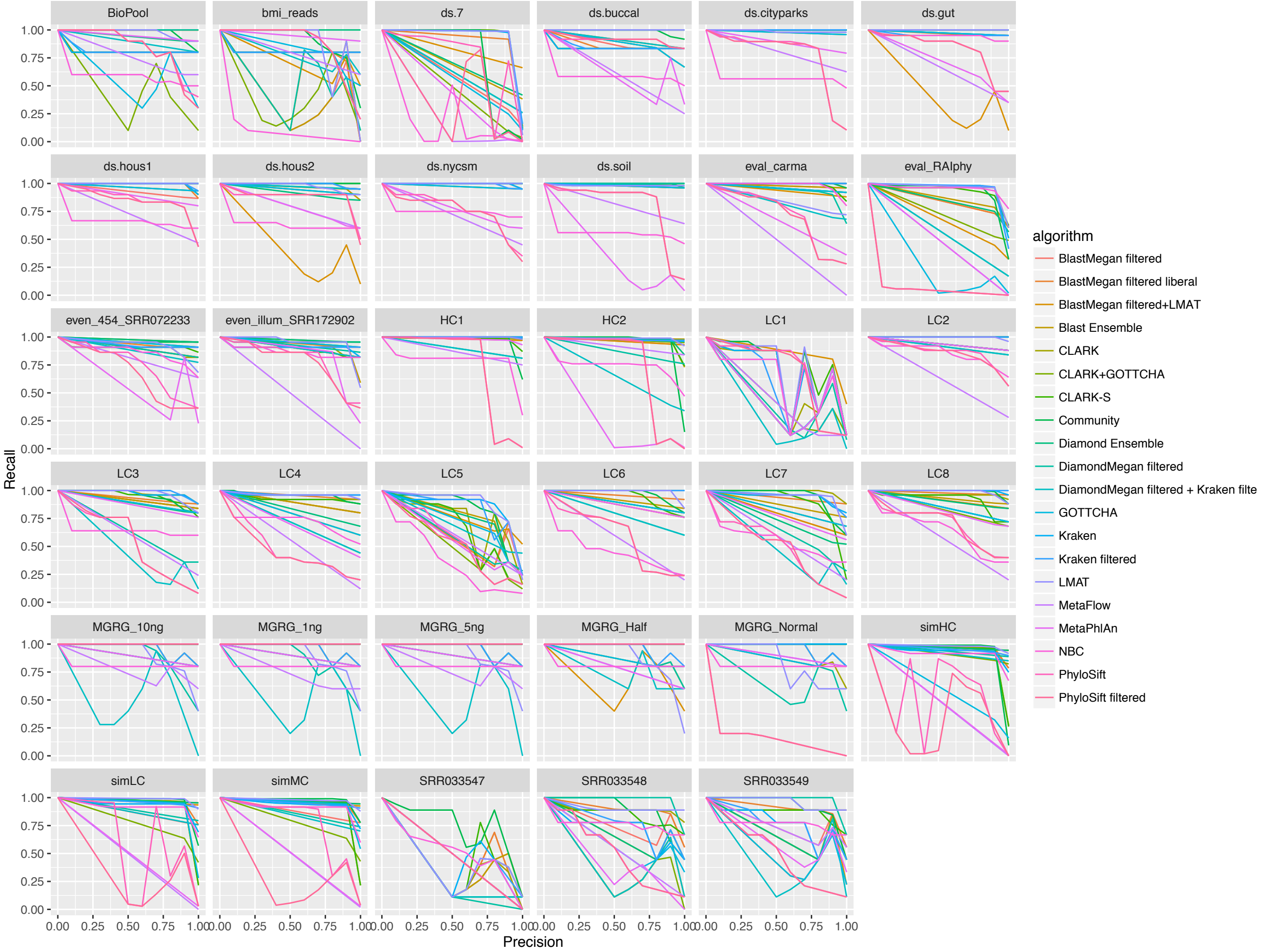
Performance profiles across 16 datasets



Precision = false positive rate = TP/(TP+FP)
Recall = sensitivity = TP/(TP+FN)
F1 score = 2(precision*recall)/(precision+recall)
AUPR = area under the precision recall curve

Precision-recall curves

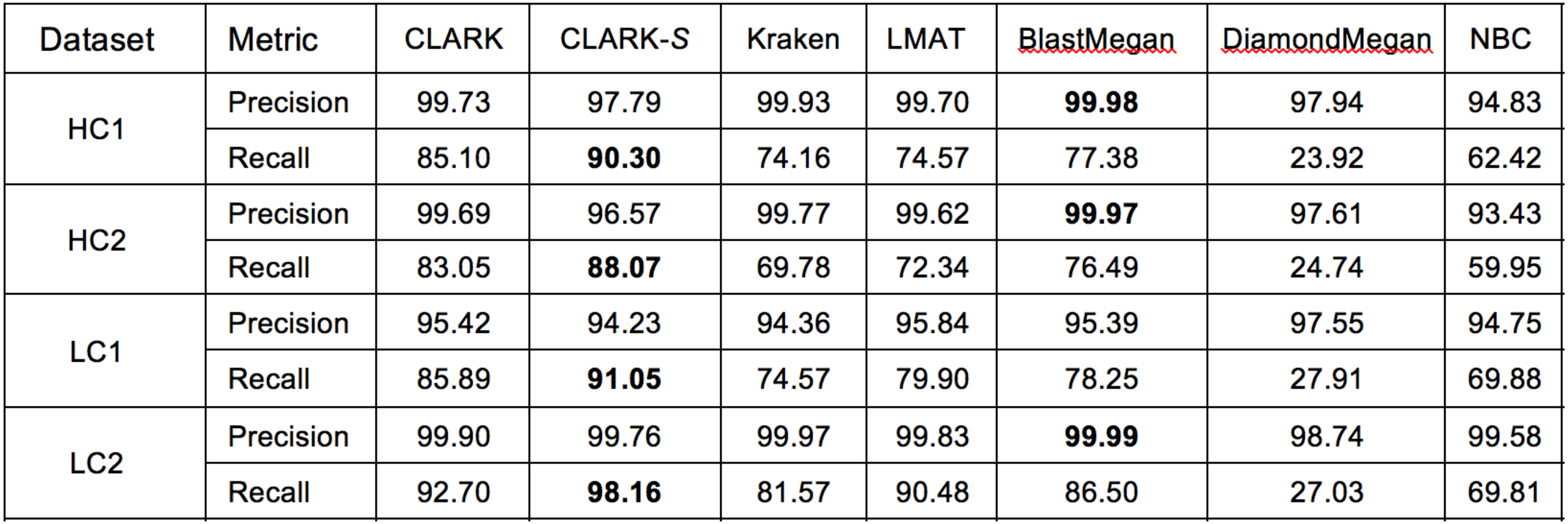
Supplementary Figure 3



Precision-recall curves for tools on individual samples.

Read-level classification

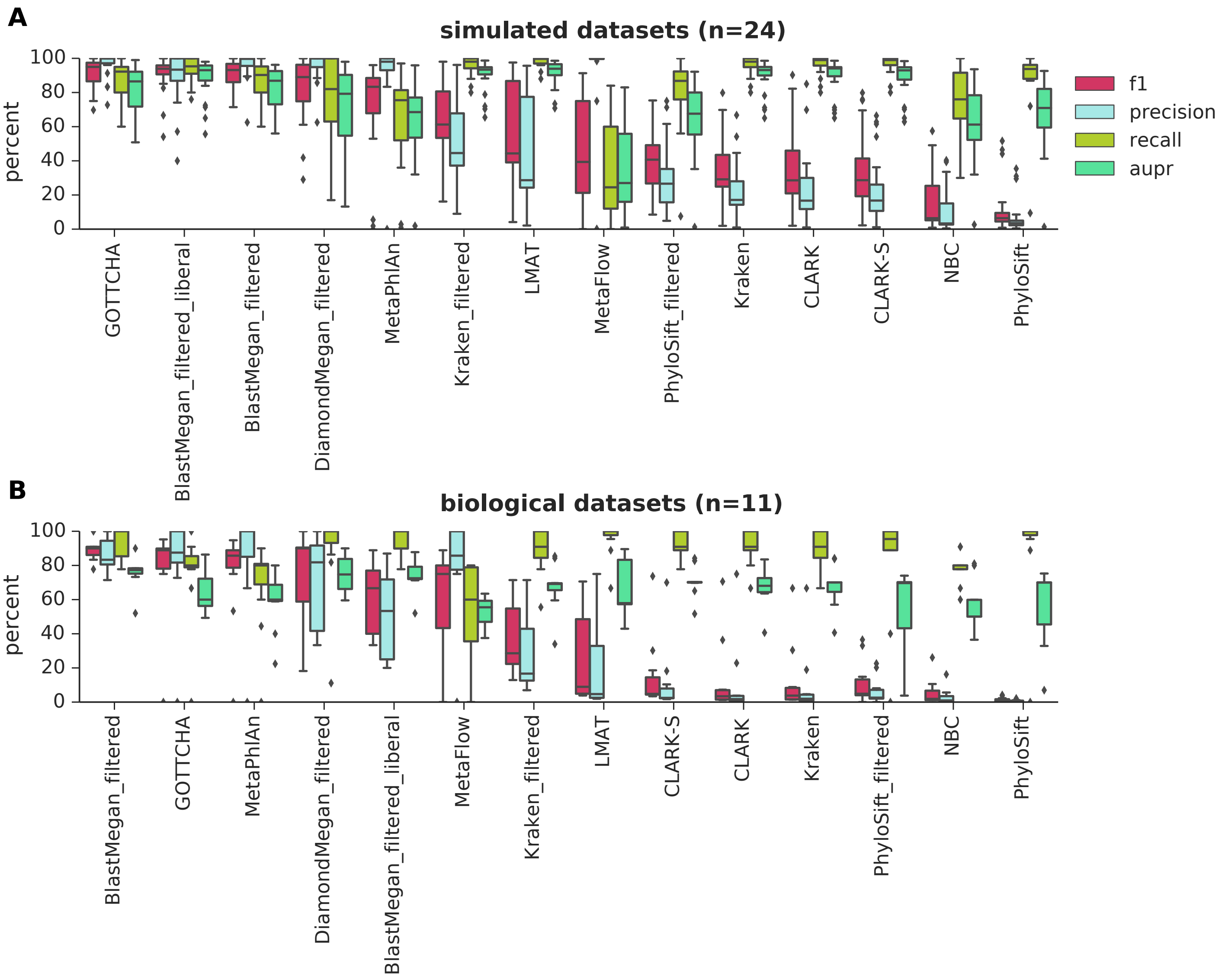
- Classification precision increased for *k*-mer-based tools when calculated at the read-level compared to the organismal level



Dataset	Metric	CLARK	CLARK-*S*	Kraken	LMAT	BlastMegan	DiamondMegan	NBC
HC1	Precision	99.73	97.79	99.93	99.70	**99.98**	97.94	94.83
	Recall	85.10	**90.30**	74.16	74.57	77.38	23.92	62.42
HC2	Precision	99.69	96.57	99.77	99.62	**99.97**	97.61	93.43
	Recall	83.05	**88.07**	69.78	72.34	76.49	24.74	59.95
LC1	Precision	95.42	94.23	94.36	95.84	95.39	97.55	94.75
	Recall	85.89	**91.05**	74.57	79.90	78.25	27.91	69.88
LC2	Precision	99.90	99.76	99.97	99.83	**99.99**	98.74	99.58
	Recall	92.70	**98.16**	81.57	90.48	86.50	27.03	69.81

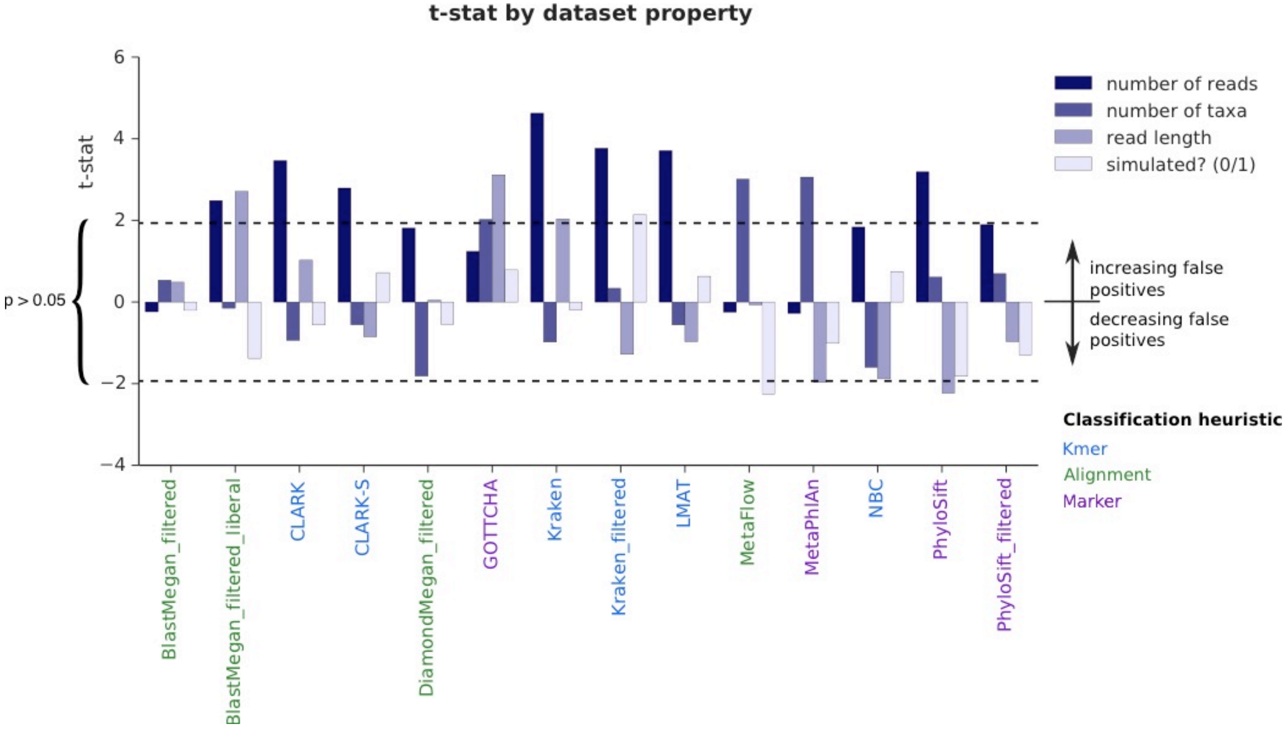
Rachid Ounit

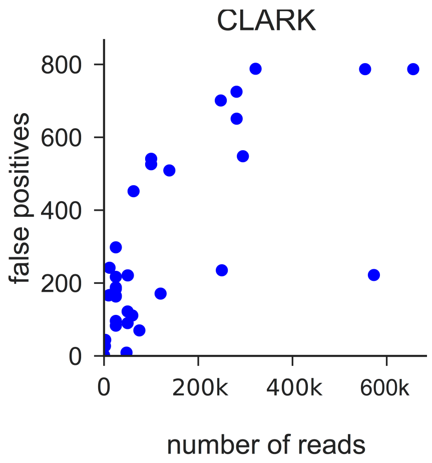
Simulated vs. biological datasets

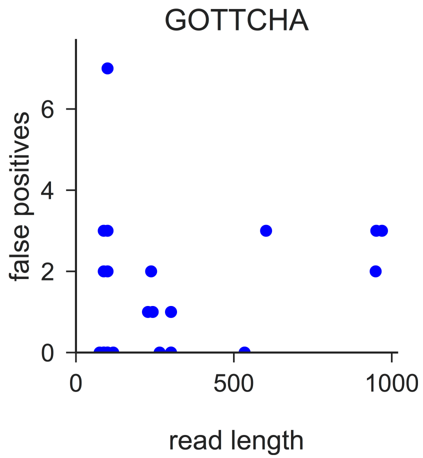


False positives by dataset property

Figure 2







$$\# \text{ FP} \sim \text{Nbin}(\beta_0 + \beta_1(\# \text{ reads}) + \beta_2(\# \text{ taxa}) + \beta_3(\text{read length}) + \beta_4(\text{simulated } 0/1))$$

Accuracy by taxa

Common false positives

- Phyla: Proteobacteria, Firmicutes, Actinobacteria
- Genera: *Lactobacillus*, *Staphylococcus*, *Streptococcus*

Common false negatives

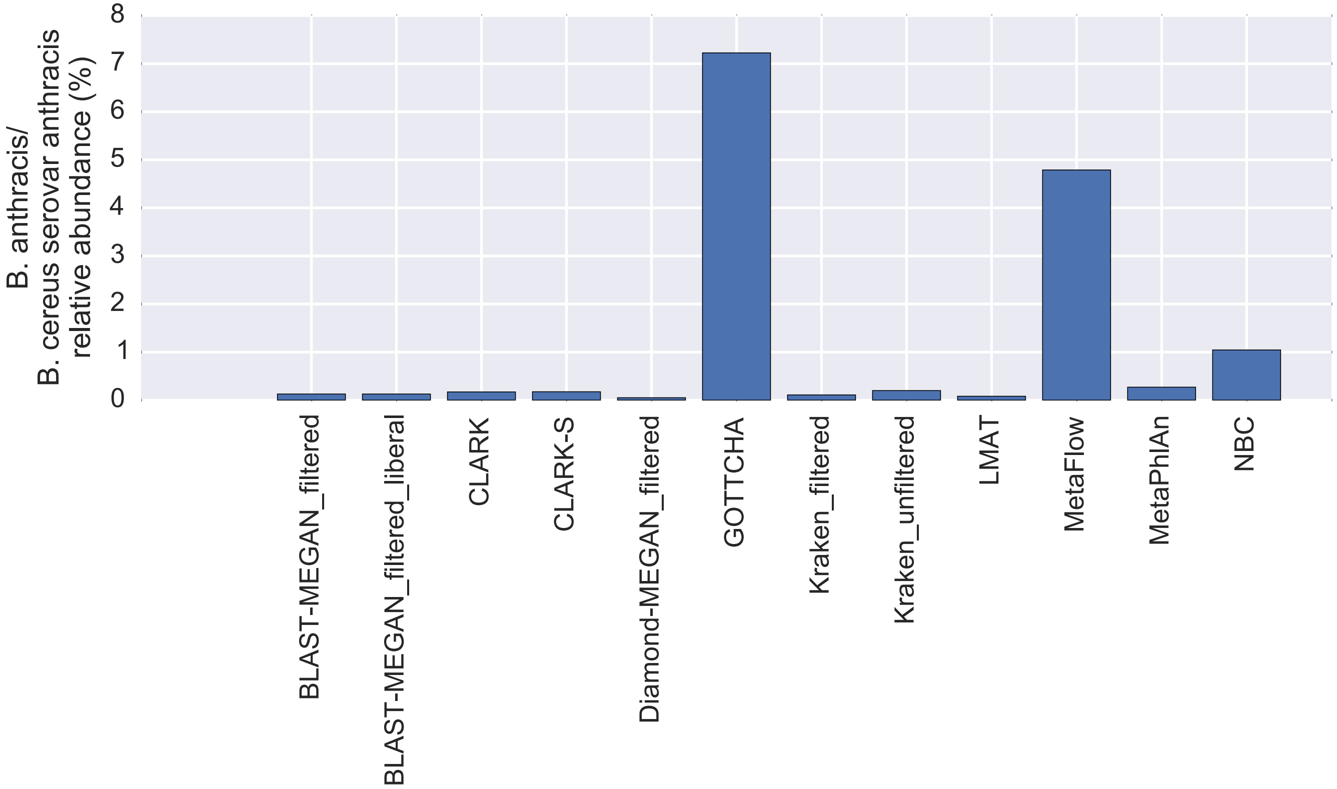
- Genera: *Bacillus*, *Bifidobacteria*, *Shigella*

Negative controls

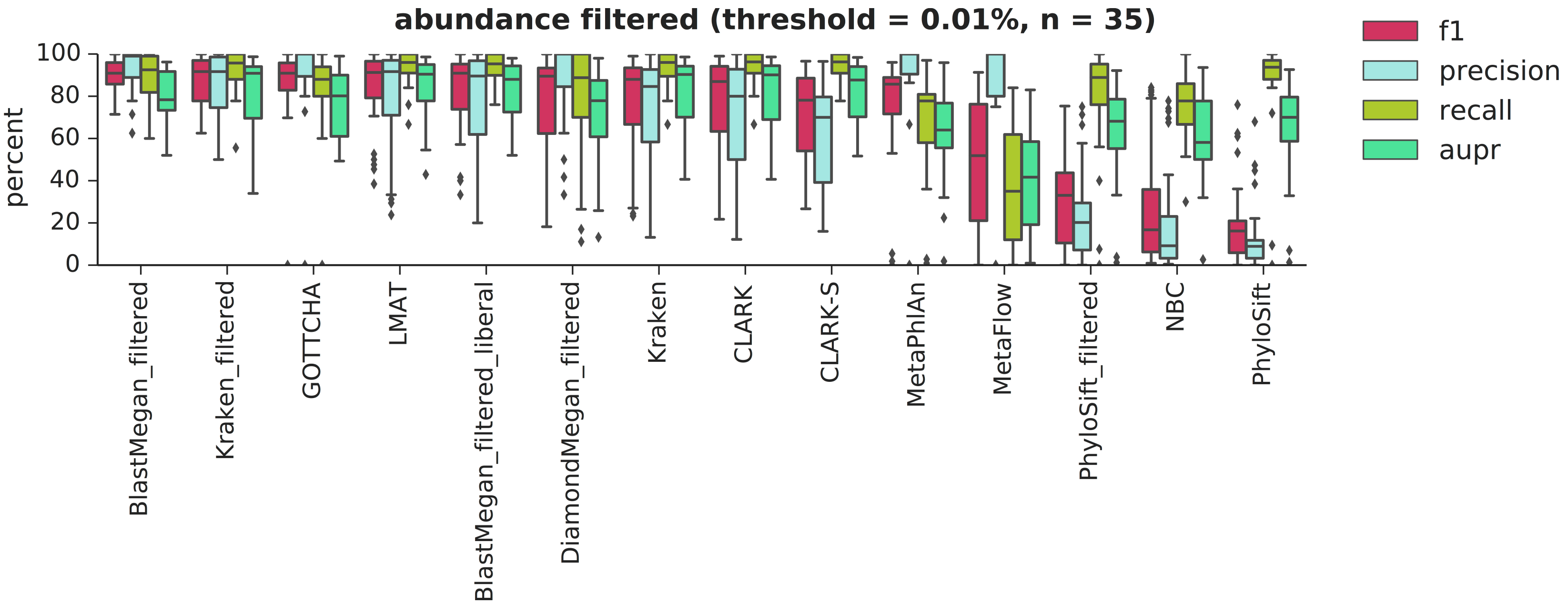
- Human DNA spiked into extraction kit for sequencing contaminants
  - *Escherichia* (*coli*) and *Acinetobacter*

- Nullomers (combined 17-mers that did not match to any known reference sequence)
  - Size and database biases for NBC (*Sorangium cellulosum*, *Escherichia coli*, *Bacillus cereus*), LMAT (human)

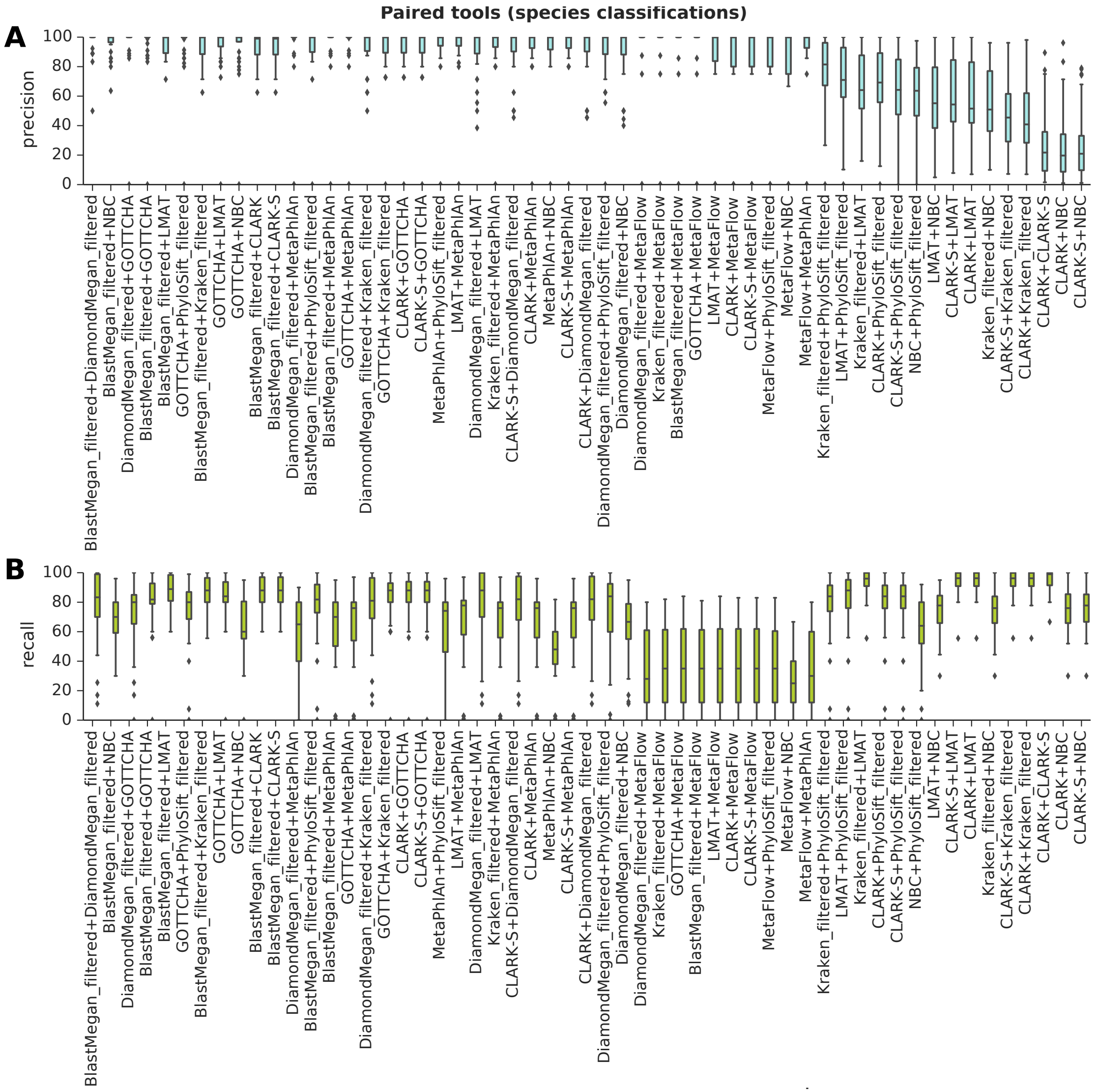
Anthrax on the subway? (No.)



How to solve a problem like *Bacillus*: abundance filtering

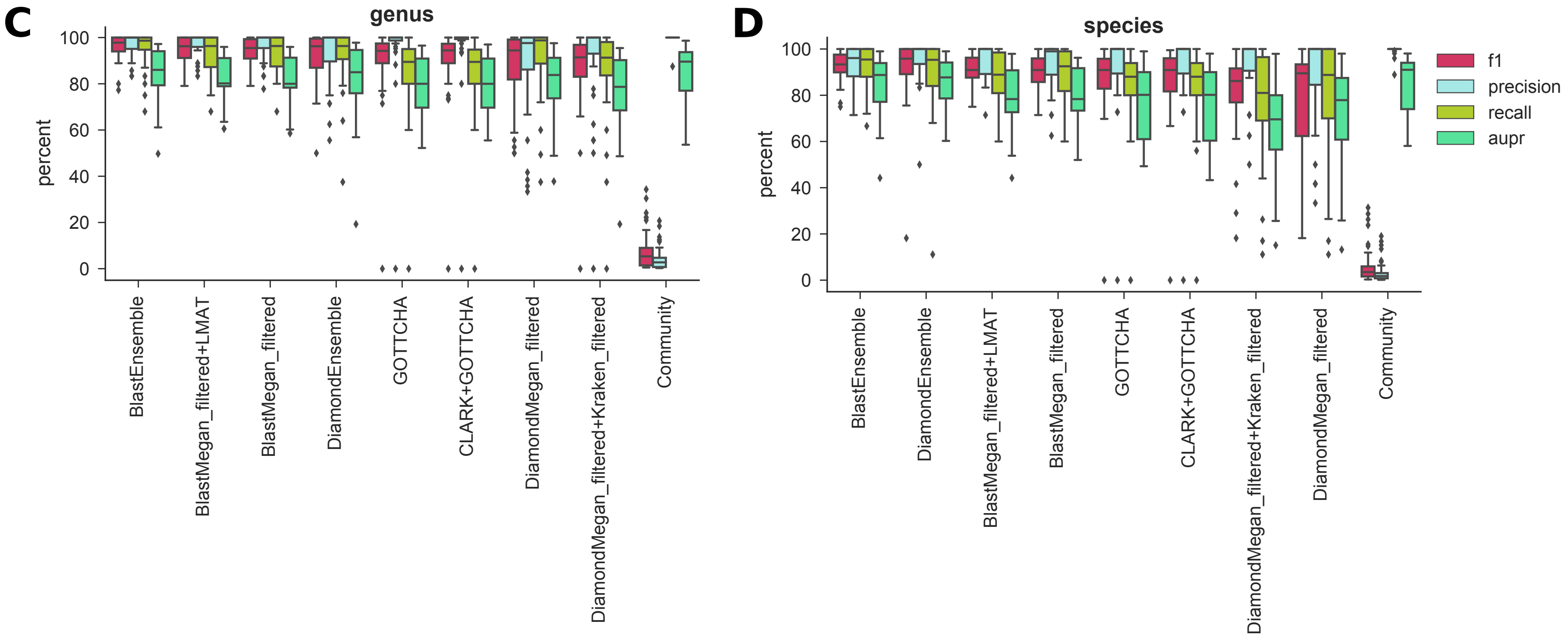


How to solve a problem like *Bacillus*: pairing tools



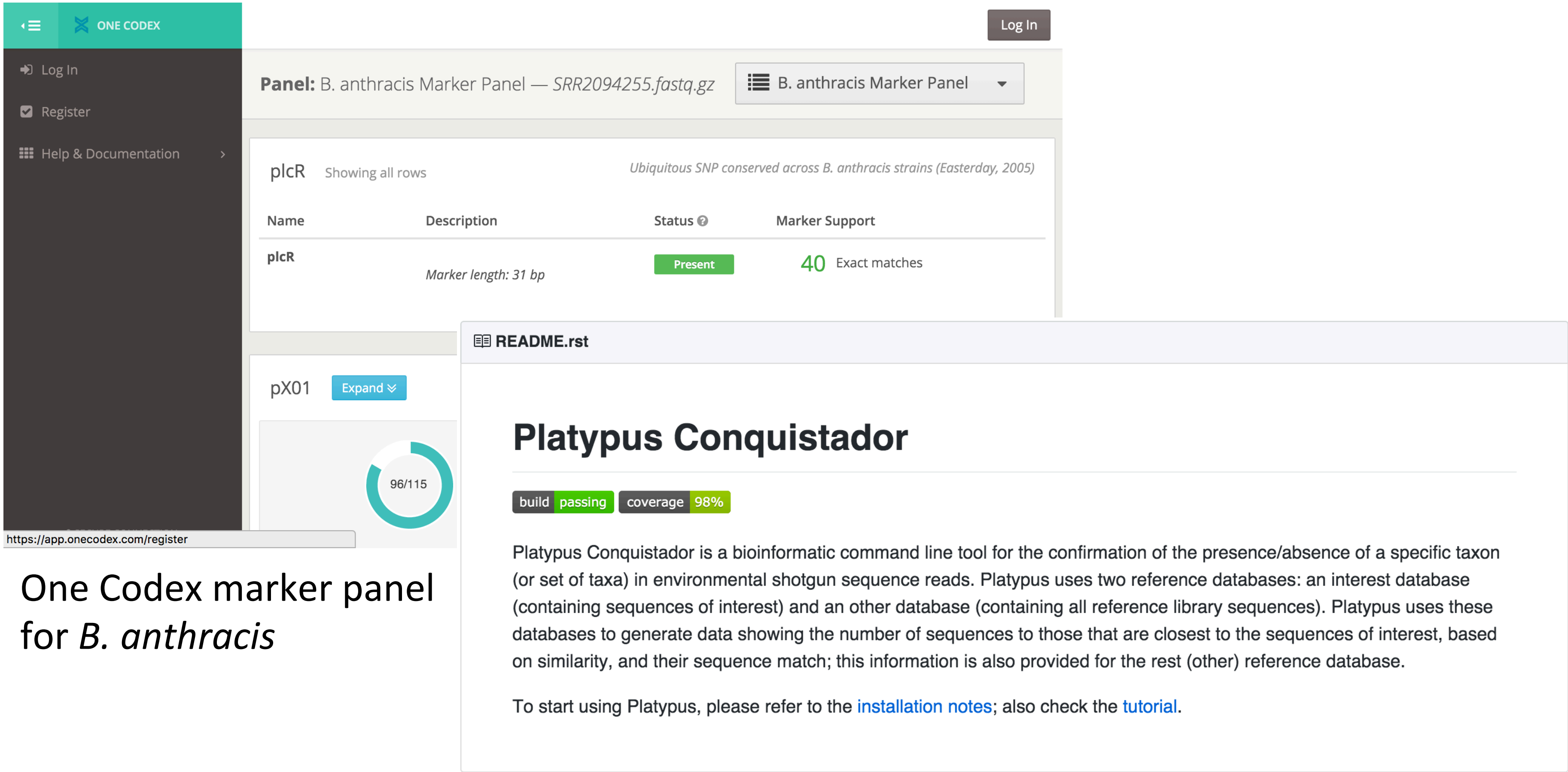
]

How to solve a problem like *Bacillus*: ensemble methods



Quorum ensembles:
Min 2/4 or 3/5 of set of high-precision tools detect a taxon

Specialized tools for pathogen detection



Panel: B. anthracis Marker Panel — *SRR2094255.fastq.gz*

B. anthracis Marker Panel

plcR Showing all rows

Ubiquitous SNP conserved across B. anthracis strains (Easterday, 2005)

Name	Description	Status	Marker Support
plcR	*Marker length: 31 bp*	Present	40 Exact matches

pX01 Expand

96/115

https://app.onecodex.com/register

README.rst

Platypus Conquistador

build passing coverage 98%

Platypus Conquistador is a bioinformatic command line tool for the confirmation of the presence/absence of a specific taxon (or set of taxa) in environmental shotgun sequence reads. Platypus uses two reference databases: an interest database (containing sequences of interest) and an other database (containing all reference library sequences). Platypus uses these databases to generate data showing the number of sequences to those that are closest to the sequences of interest, based on similarity, and their sequence match; this information is also provided for the rest (other) reference database.

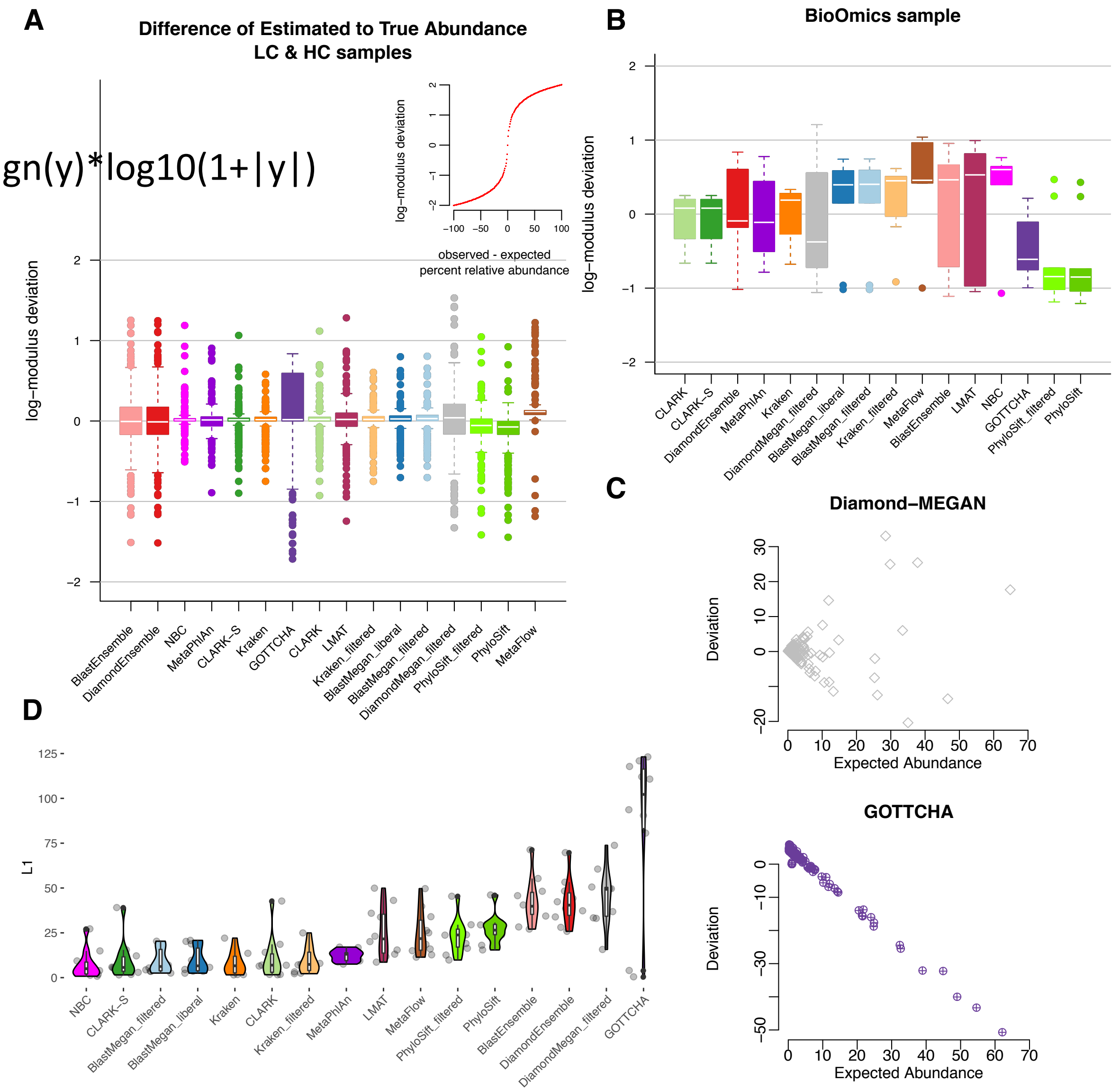
To start using Platypus, please refer to the installation notes; also check the tutorial.

One Codex marker panel for *B. anthracis*

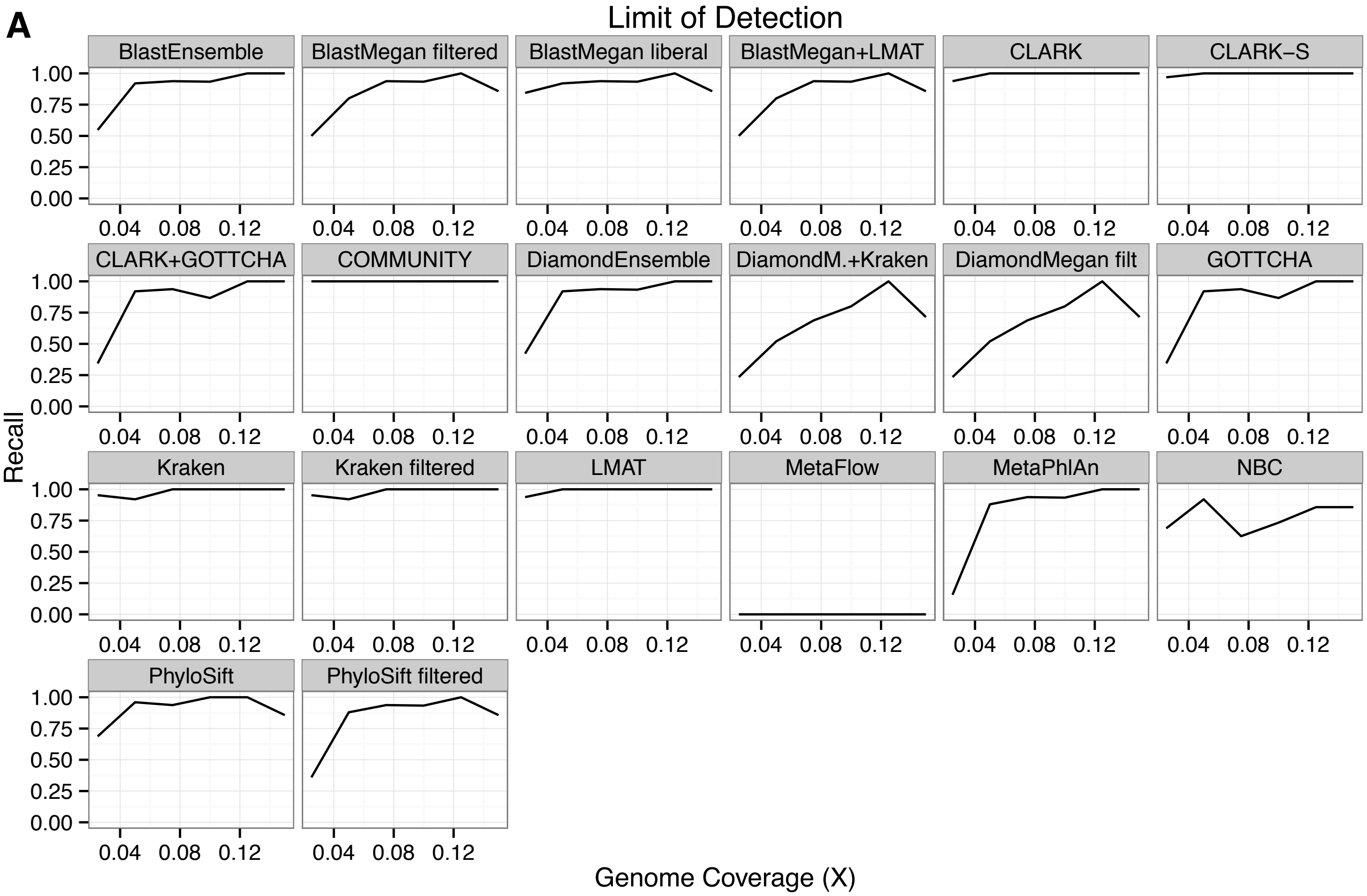
Knight Lab

Abundance estimates

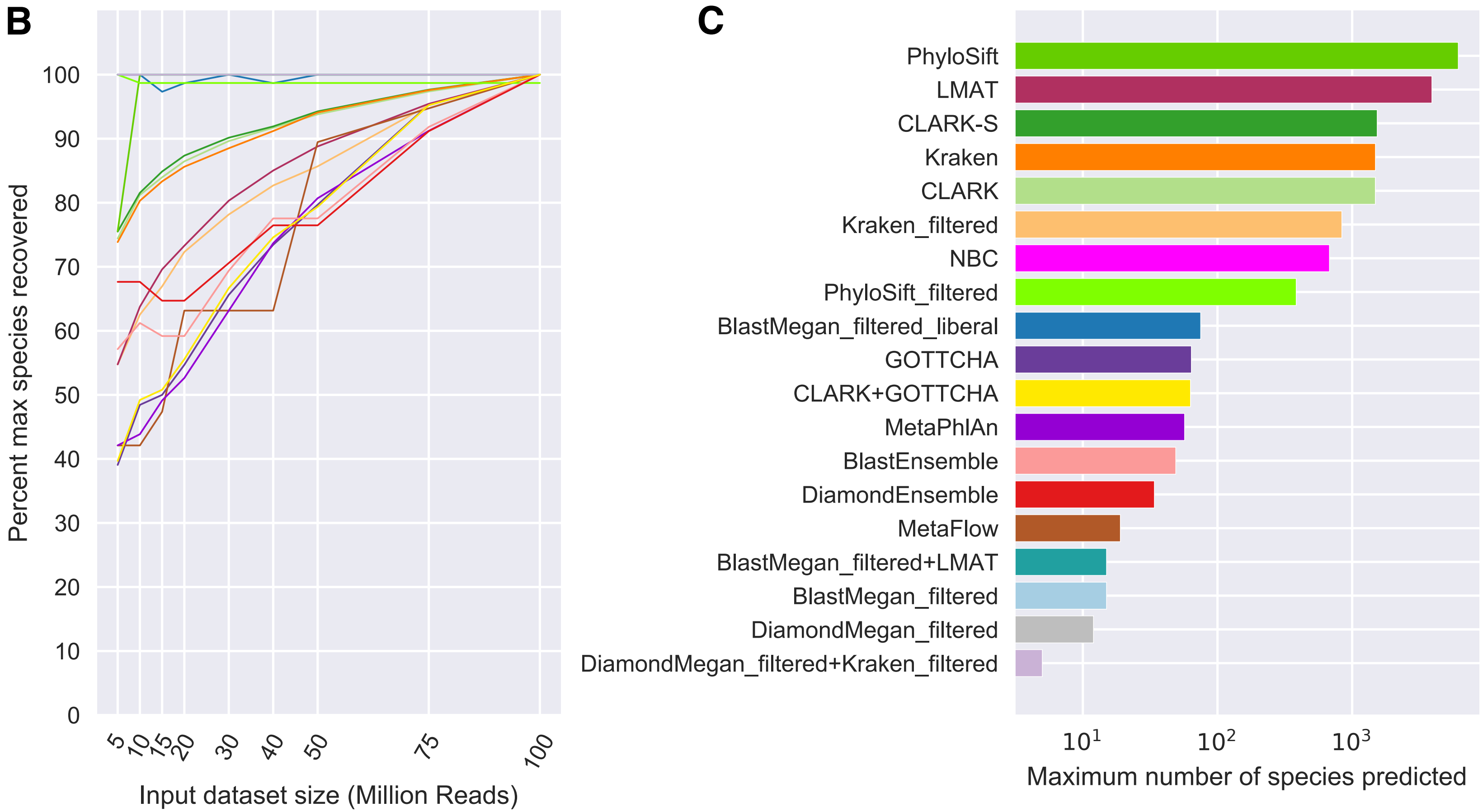
y' = sign(y)*log10(1+|y|)



Precision/recall trade-offs



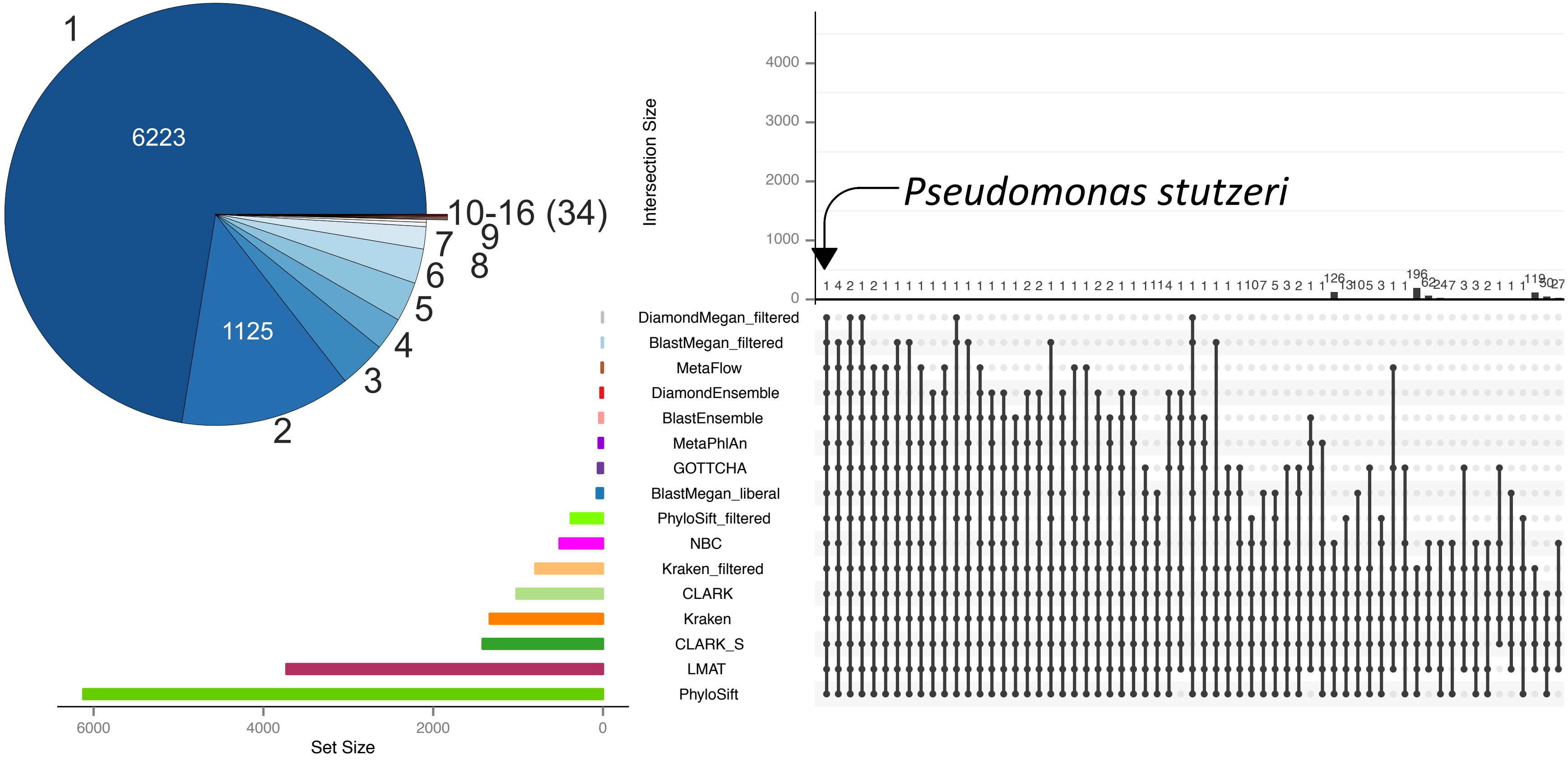
Deep sequenced subway sample



Elizabeth Hénaff

# of species detected by all tools = 1

Tool overlap at 100M reads

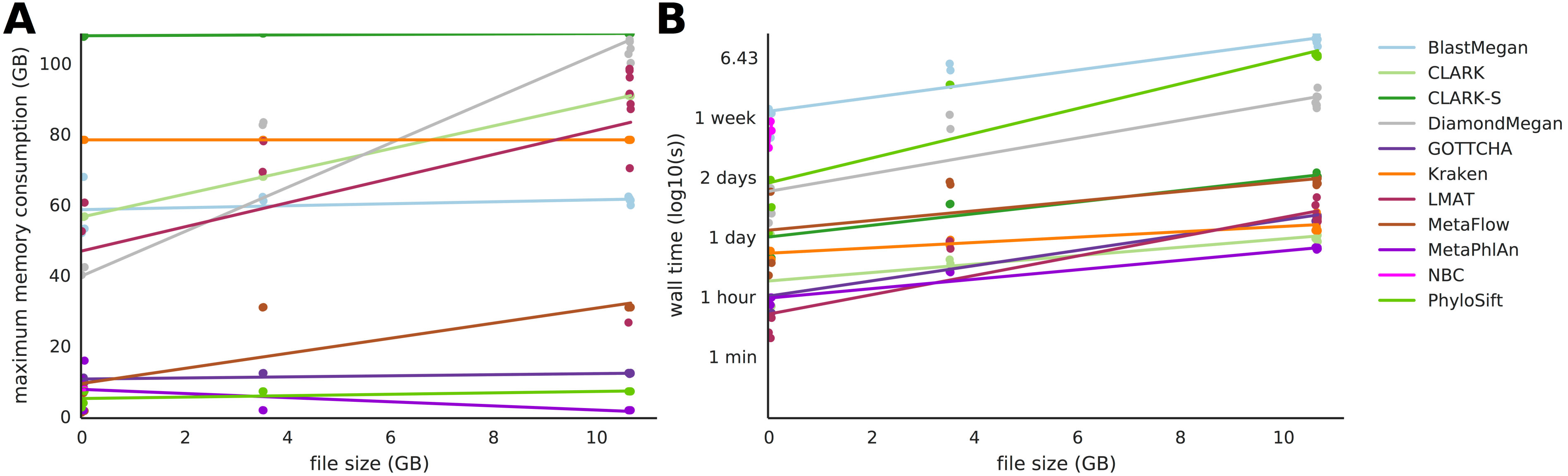


Pseudomonas stutzeri

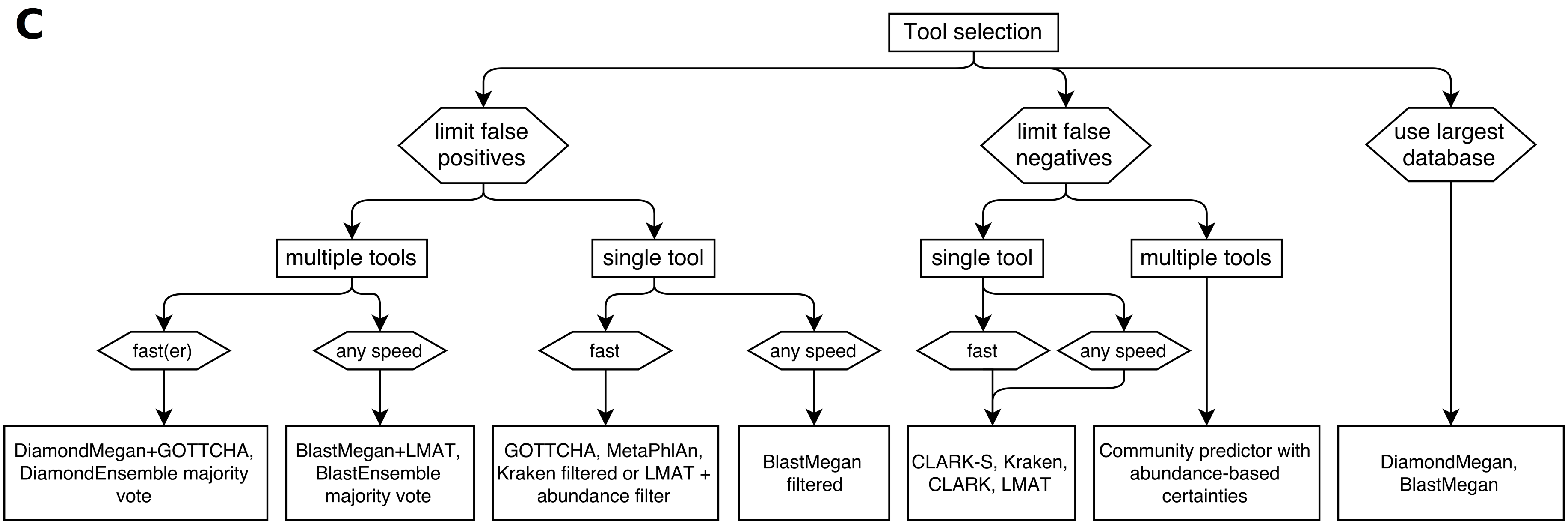
Some species detected by >= 10 tools

- Widespread: *Pseudomonas stutzeri*, *Micrococcus luteus*, *Escherichia coli*, *Cutibacterium acnes*, …
- Soil: *Comamonas testosteroni*, *Bacillus pumilus*
- Wastewater: *Rhodococcus hoagii*
- Cheese: *Glutamicibacter arilaitensis*
- Pathogens: *Bacillus anthracis*

Constraints



In summary



c

Tool selection

limit false positives

multiple tools

fast(er)

DiamondMegan+GOTTCHA, DiamondEnsemble majority vote

any speed

BlastMegan+LMAT, BlastEnsemble majority vote

single tool

fast

GOTTCHA, MetaPhlAn, Kraken filtered or LMAT + abundance filter

any speed

BlastMegan filtered

limit false negatives

single tool

fast

any speed

CLARK-S, Kraken, CLARK, LMAT

multiple tools

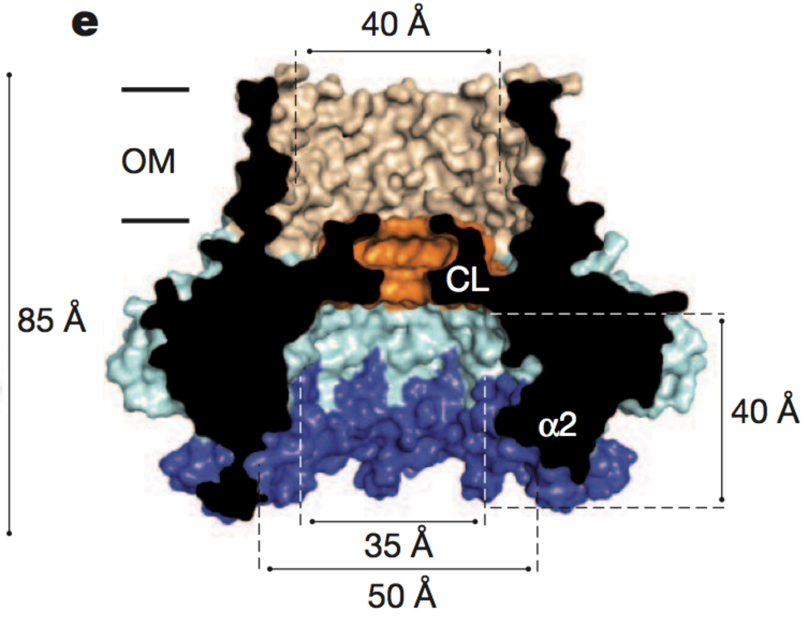
Community predictor with abundance-based certainties

use largest database

DiamondMegan, BlastMegan

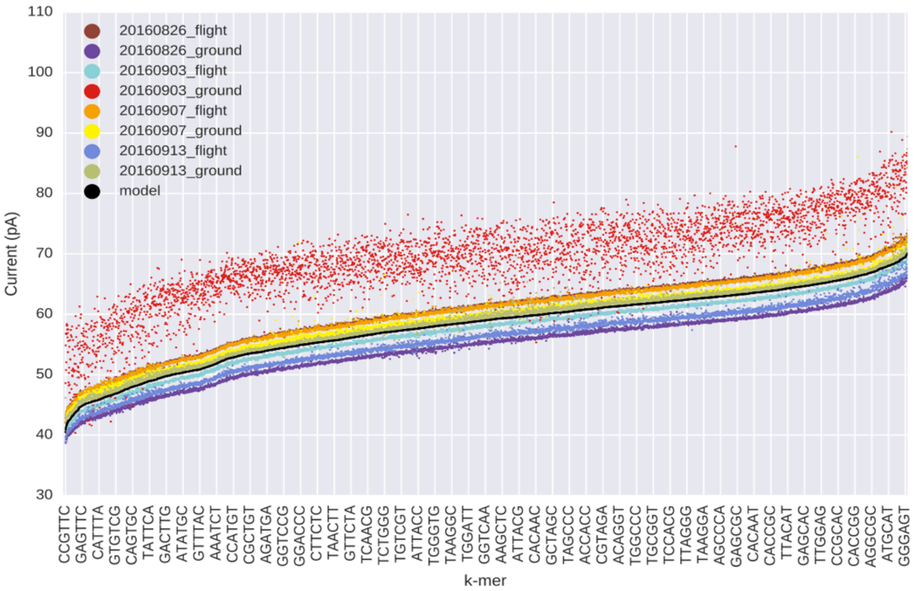
Nanopore sequencing

Pores to currents

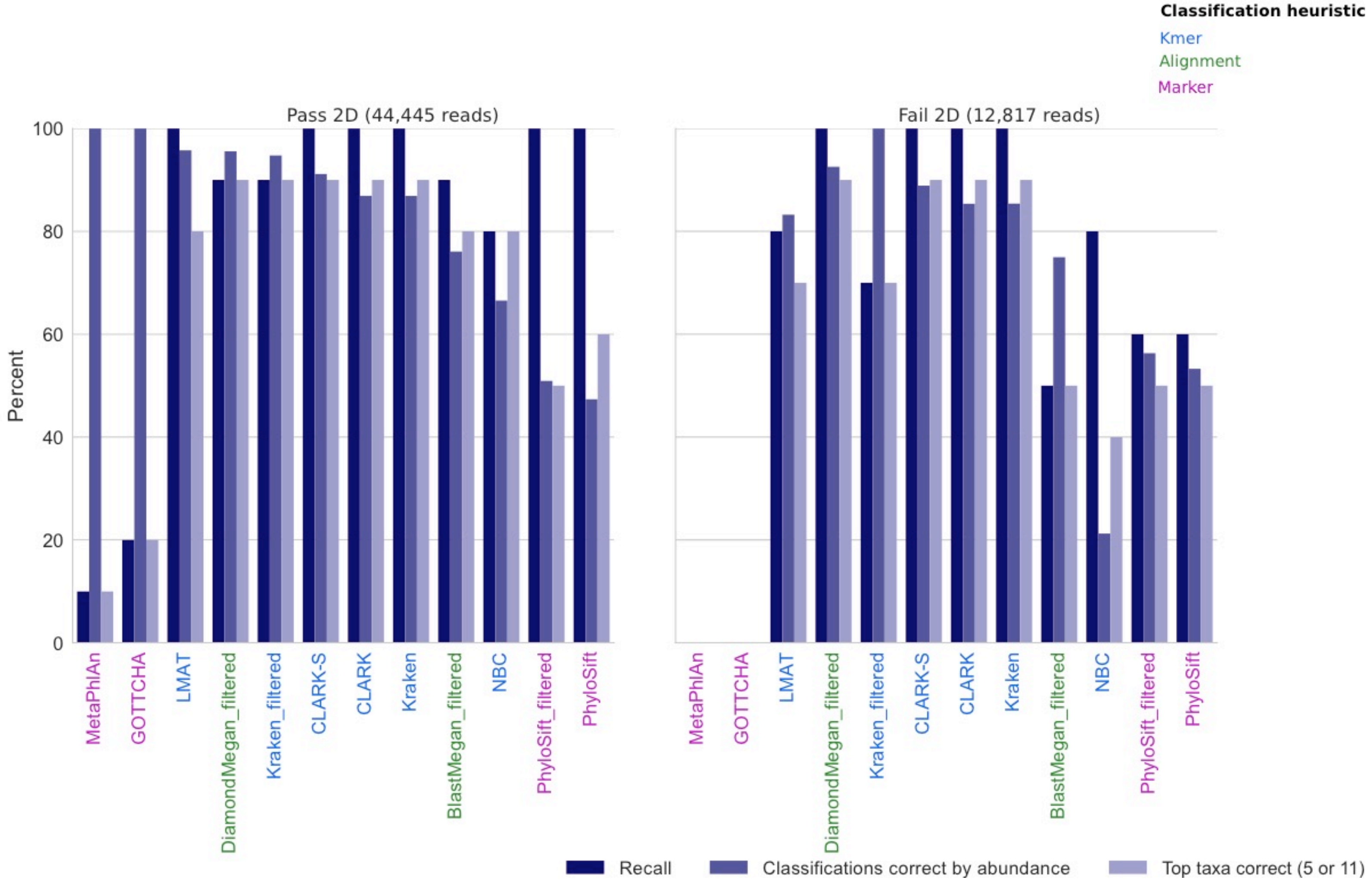


Goyal et al., 2014

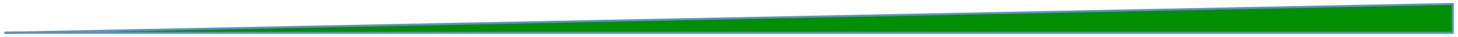
Currents to *k*-mers



The long and the short of it



Bacterial epigenomics: *N*6-methyladenine

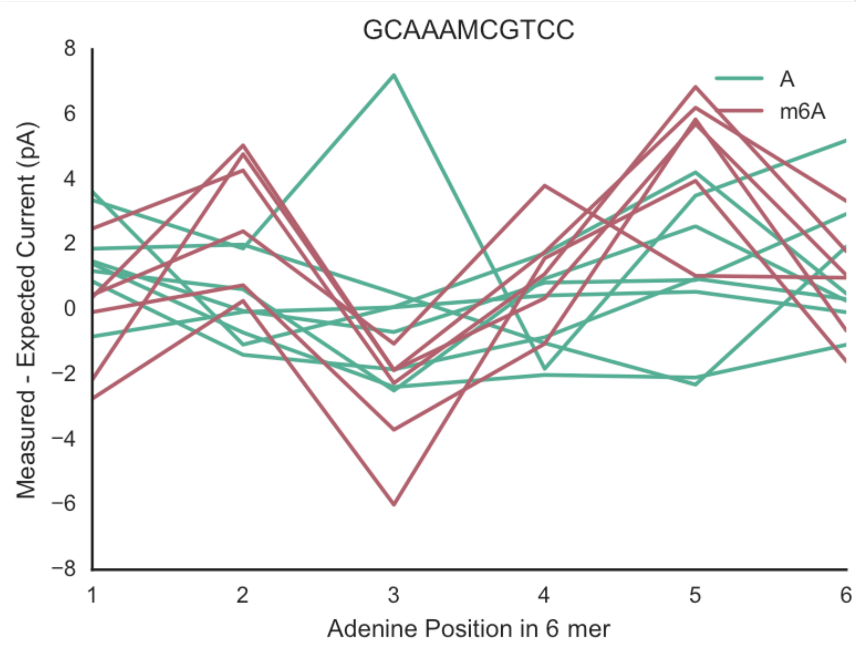


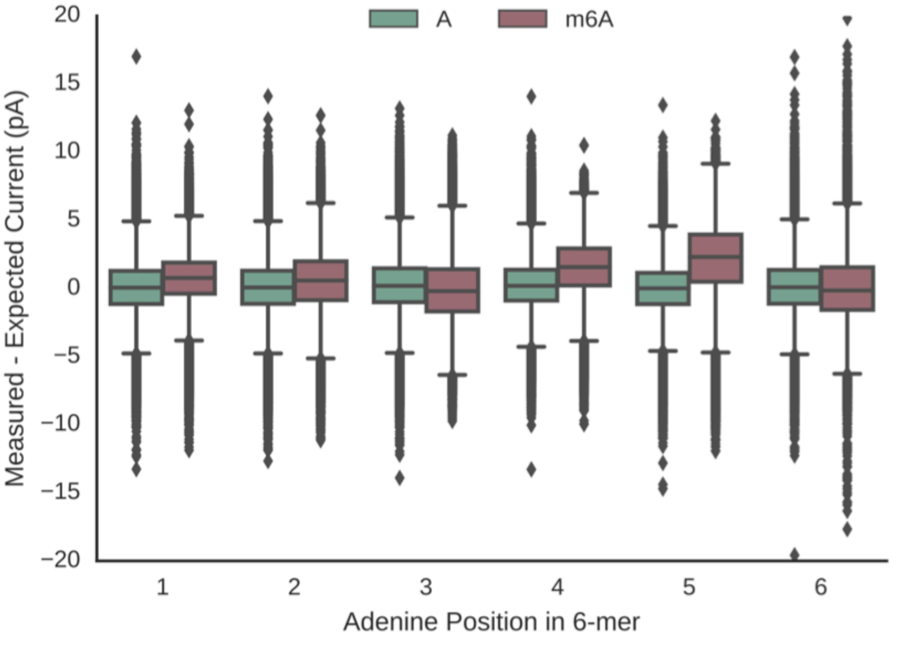
0.00009% m^{6}A

10-20% m^{6}A

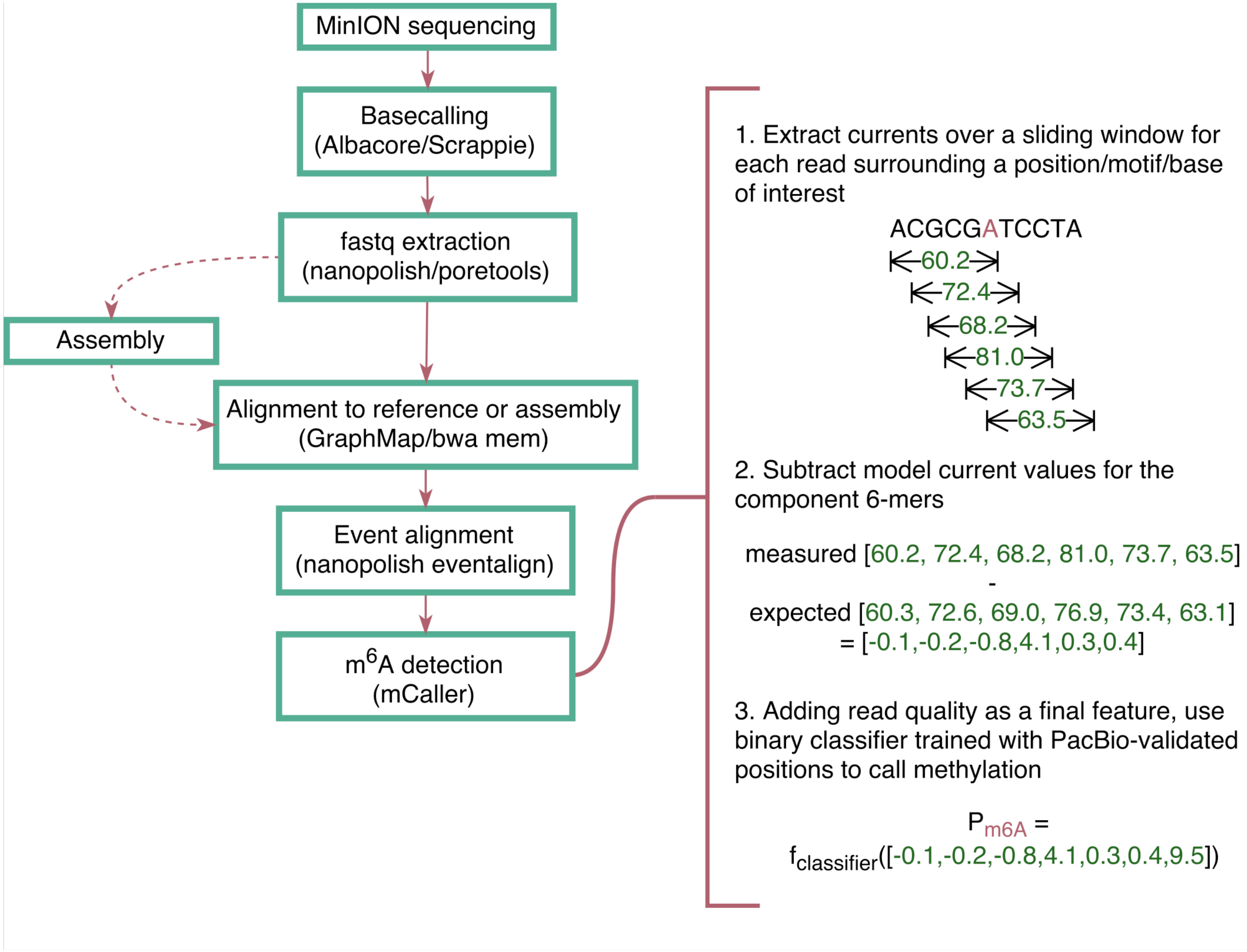
- Eukaryotes
  - Rare in most
  - Roles:
    - Fertility (*C. elegans*), nucleosome positioning (green algae), unknown (vertebrates)
- Prokaryotes
  - Most common base modification
  - Roles:
    - Defense against foreign DNA, replication repair, pathogenicity

Bacterial epigenomics





Bacterial epigenomics



MinION sequencing

Basecalling (Albacore/Scrappie)

fastq extraction (nanopolish/poretools)

Assembly

Alignment to reference or assembly (GraphMap/bwa mem)

Event alignment (nanopolish eventalign)

m^{6}A detection (mCaller)

1. Extract currents over a sliding window for each read surrounding a position/motif/base of interest

ACGCGATCCTA

60.2

72.4

68.2

81.0

73.7

63.5

2. Subtract model current values for the component 6-mers

measured [60.2, 72.4, 68.2, 81.0, 73.7, 63.5]

-

expected [60.3, 72.6, 69.0, 76.9, 73.4, 63.1]

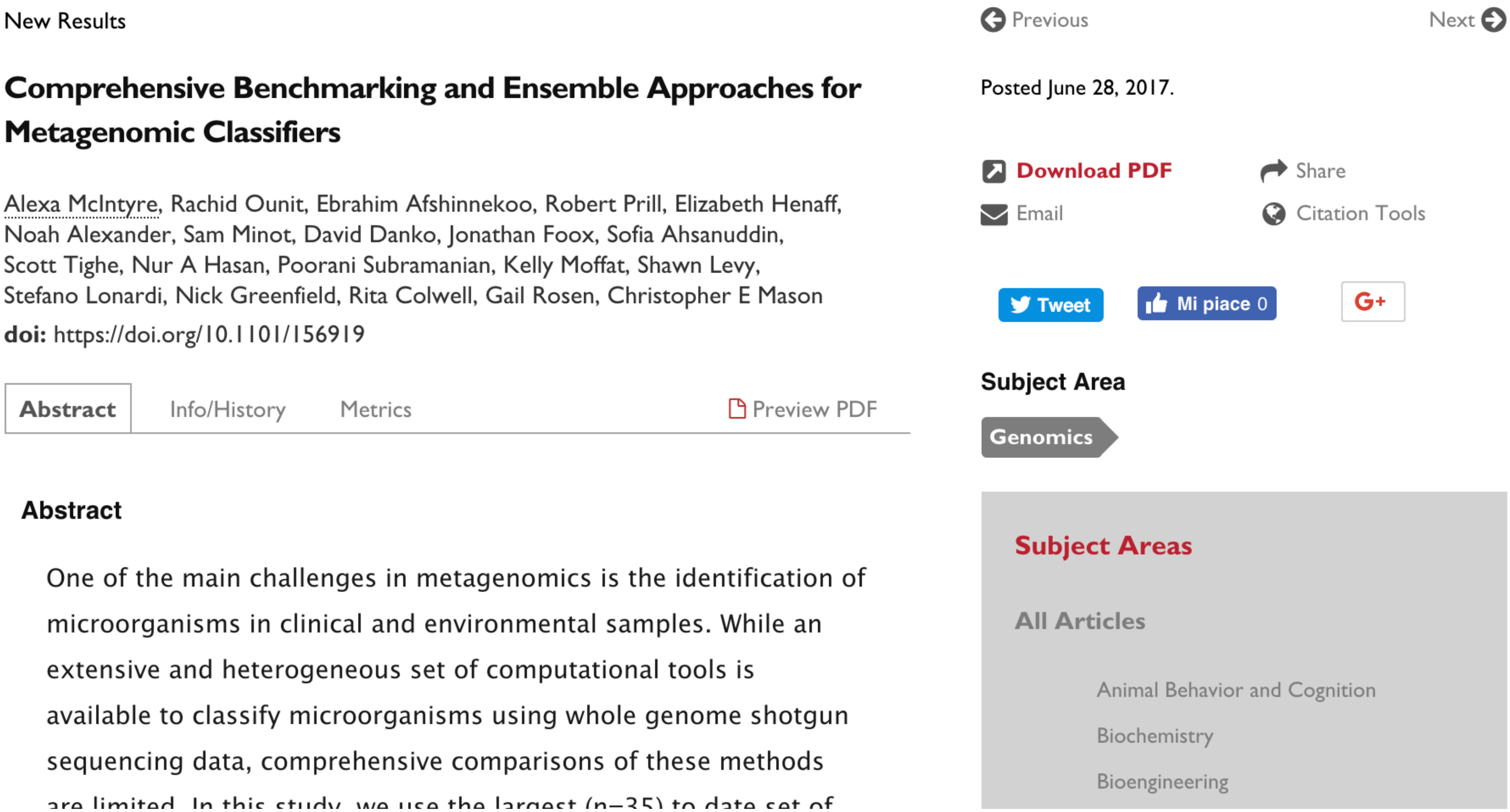
= [-0.1,-0.2,-0.8,4.1,0.3,0.4]

3. Adding read quality as a final feature, use binary classifier trained with PacBio-validated positions to call methylation

$$P_{m6A} = f_{classifier}([-0.1,-0.2,-0.8,4.1,0.3,0.4,9.5])$$

Benchmarking bioRxiv

- https://ftp-private.ncbi.nlm.nih.gov/nist-immsa/IMMSA/



Comprehensive Benchmarking and Ensemble Approaches for Metagenomic Classifiers

Posted June 28, 2017.

Alexa McIntyre, Rachid Ounit, Ebrahim Afshinnekoo, Robert Prill, Elizabeth Henaff, Noah Alexander, Sam Minot, David Danko, Jonathan Foox, Sofia Ahsanuddin, Scott Tighe, Nur A Hasan, Poorani Subramanian, Kelly Moffat, Shawn Levy, Stefano Lonardi, Nick Greenfield, Rita Colwell, Gail Rosen, Christopher E Mason

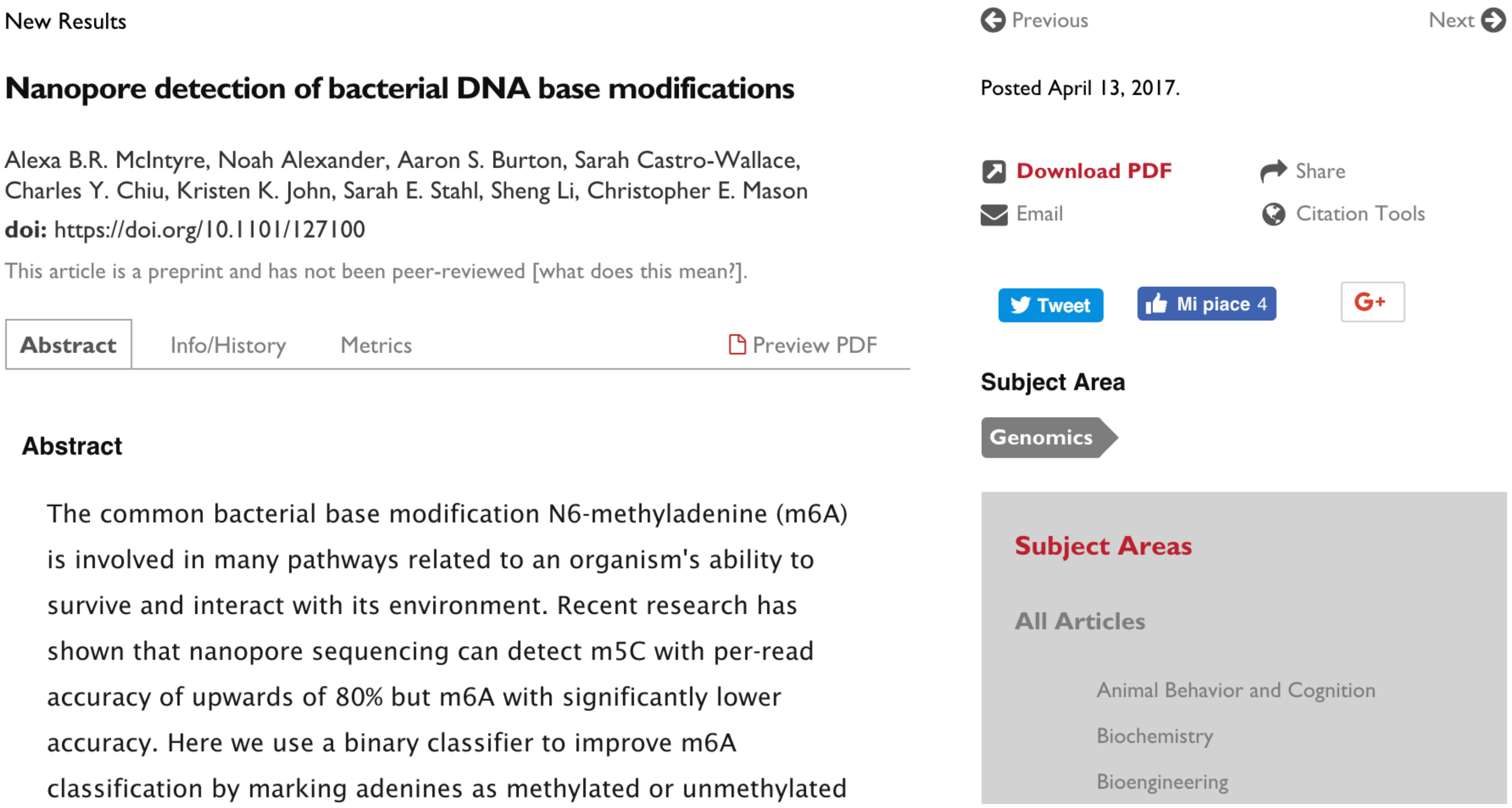
doi: https://doi.org/10.1101/156919

Subject Area

Abstract

One of the main challenges in metagenomics is the identification of microorganisms in clinical and environmental samples. While an extensive and heterogeneous set of computational tools is available to classify microorganisms using whole genome shotgun sequencing data, comprehensive comparisons of these methods are limited. In this study, we use the largest (n=35) to date set of

mCaller bioRxiv



New Results

Nanopore detection of bacterial DNA base modifications

Alexa B.R. McIntyre, Noah Alexander, Aaron S. Burton, Sarah Castro-Wallace, Charles Y. Chiu, Kristen K. John, Sarah E. Stahl, Sheng Li, Christopher E. Mason

doi: https://doi.org/10.1101/127100

This article is a preprint and has not been peer-reviewed [what does this mean?].

Posted April 13, 2017.

Abstract

The common bacterial base modification N6-methyladenine (m6A) is involved in many pathways related to an organism's ability to survive and interact with its environment. Recent research has shown that nanopore sequencing can detect m5C with per-read accuracy of upwards of 80% but m6A with significantly lower accuracy. Here we use a binary classifier to improve m6A classification by marking adenines as methylated or unmethylated

Subject Area

Genomics

Subject Areas

All Articles

Animal Behavior and Cognition

Biochemistry

Bioengineering

TRI-INSTITUTIONAL PHD PROGRAM
Computational
Biology & Medicine

Acknowledgements

- Mason Lab

  Christopher Mason, Elizabeth Hénaff, Ebrahim Afshinnekoo, David Danko, Jonathan Foox, Noah Alexander, Sofia Ahsanuddin
- Rachid Ounit (Biotia)
- Gail Rosen (Drexel)
- Bobby Prill (IBM)
- CosmosID

  Rita Colwell, Nur Hasan, Poorani Subramanian, Kelly Moffat
- One Codex

  Nick Greenfield, Sam Minot
- Scott Tighe (UVM), Shawn Levy (Hudson Alpha), Stefano Leonardi (UC Riverside), Jonathan Allen (LLNL)
- Biomolecule Sequencer Project (NASA, UCSF)