ZAKŁAD INFORMATYKI MEDYCZNEJ
I TELEMEDYCYNY
WARSZAWSKIEGO UNIWERSYTETU MEDYCZNEGO

ITHAKA: A TAXONOMIC CLASSIFIER BASED ON BLOOM FILTERS, AND SPACED SEEDS

MACIEJ SYKULSKI, KAREL BŘINDA, GREGORY KUCHEROV
macieksk@gmail.com, karel.brinda@univ-mlv.fr, gregory.kucherov@univ-mlv.fr

METAGENOMICS & NGS

Metagenomics is a powerful approach to study genetic material contained in environmental samples, which is revolutionized by high-throughput sequencing technologies. Taxonomic classification of metagenomics data sets is a common step in analysis, a step for which computational cost becomes prohibitive with the growth of metagenomic datasets.

Approaches using sequence alignment algorithms, often based on the Burrows-Wheeler transform, such as Kaiju[1], Centrifuge[2], compete successfully with k-mer based alignment-free comparison methods such as Kraken[3], Clark[4]. Improvements to k-mer based approaches include: extending contiguous k-mers with spaced seeds [5][4], using Bloom filters as an underlying data structure for storing k-mers.[6]

SPACED SEEDS

A **spaced seed** is a pattern over alphabet $A = \{\#, -\}$, where
$\#$ matching position, $-$ don't care position.

```
A A C A T T C T
  # # – # – #
A A C C T T C T
```

A seed acts as a mask for comparing short oligonucleotides. The number of #'s in a seed, called *weight*, defines the number k of matching nucleotides. In the above example k = 4, seed span = 6, and the matching (spaced) k-mer is **ACTC** signifying a **hit**.

Previously, we demonstrated that spaced seeds allow for a better classification of NGS reads coming from a genome G between two other genomes G_1 and G_2 of the same genus.[5]

REFERENCES

[1] Menzel, P, Ng Kim Lee, Krogh A Fast and sensitive taxonomic classification for metagenomics with Kaiju. Nat Commun 7, Nature Publishing Group, apr 2016,

[2] Kim D, Song L, Breitwieser F P, Salzberg S Centrifuge: rapid and sensitive classification of metagenomic sequences Genome 1.2: 2.

[3] Wood, D. E. and Salzberg, S. L. Kraken: ultrafast metagenomic sequence classification using exact alignments. In Genome Biol., 15(3), R46. (2014)

[4] Ounit R, Lonardi S, Higher classification sensitivity of short metagenomic reads with CLARK-S bioRxiv 053462; doi: http://dx.doi.org/10.1101/053462

[5] BÅŹinda, K,Sykulski M, Kucherov G. Spaced seeds improve k-mer-based metagenomic classification. Bioinformatics 31.22 (2015): 3584-3592.

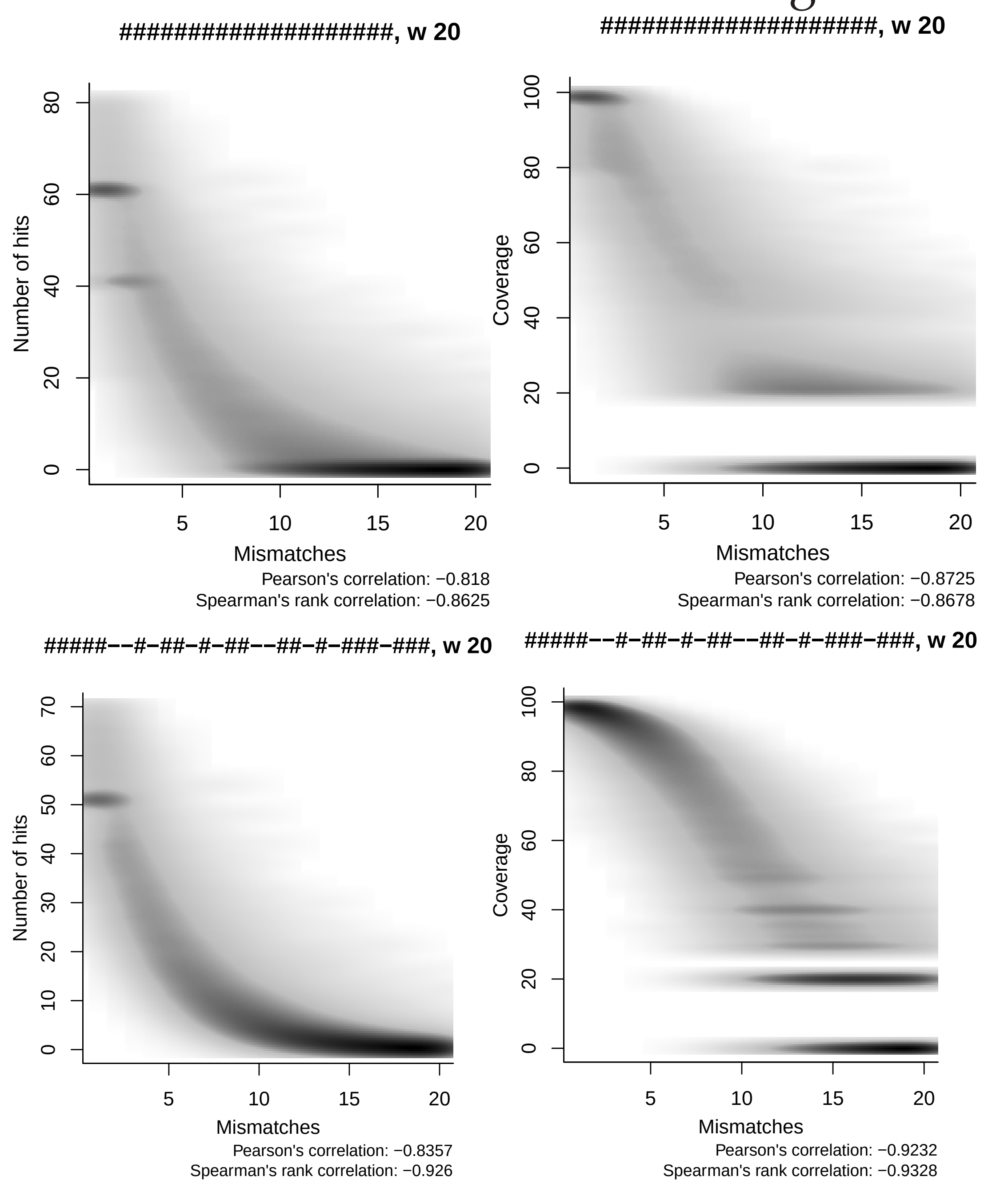
[6] Holley G, Wittler R, Stoye J. Bloom Filter Trie: an alignment-free and reference-free data structure for pan-genome storage. Algorithms for Molecular Biology. 2016;11: 3.

https://github.com/gregorykucherov/ithaka
http://seed-kraken.readthedocs.org

SCORES ON REAL GENOMES

We generated a set of ILLUMINA-like single-end reads: we've selected random substrings of *M.tuberculosis* genome of length $L = 100$ and introduced k mismatch errors, with k random between 1 and 20. For each read, we computed: **number of hits** and **coverage** to the genome under a given seed. A typical plot error vs score (seed weight 20).

Hits Coverage

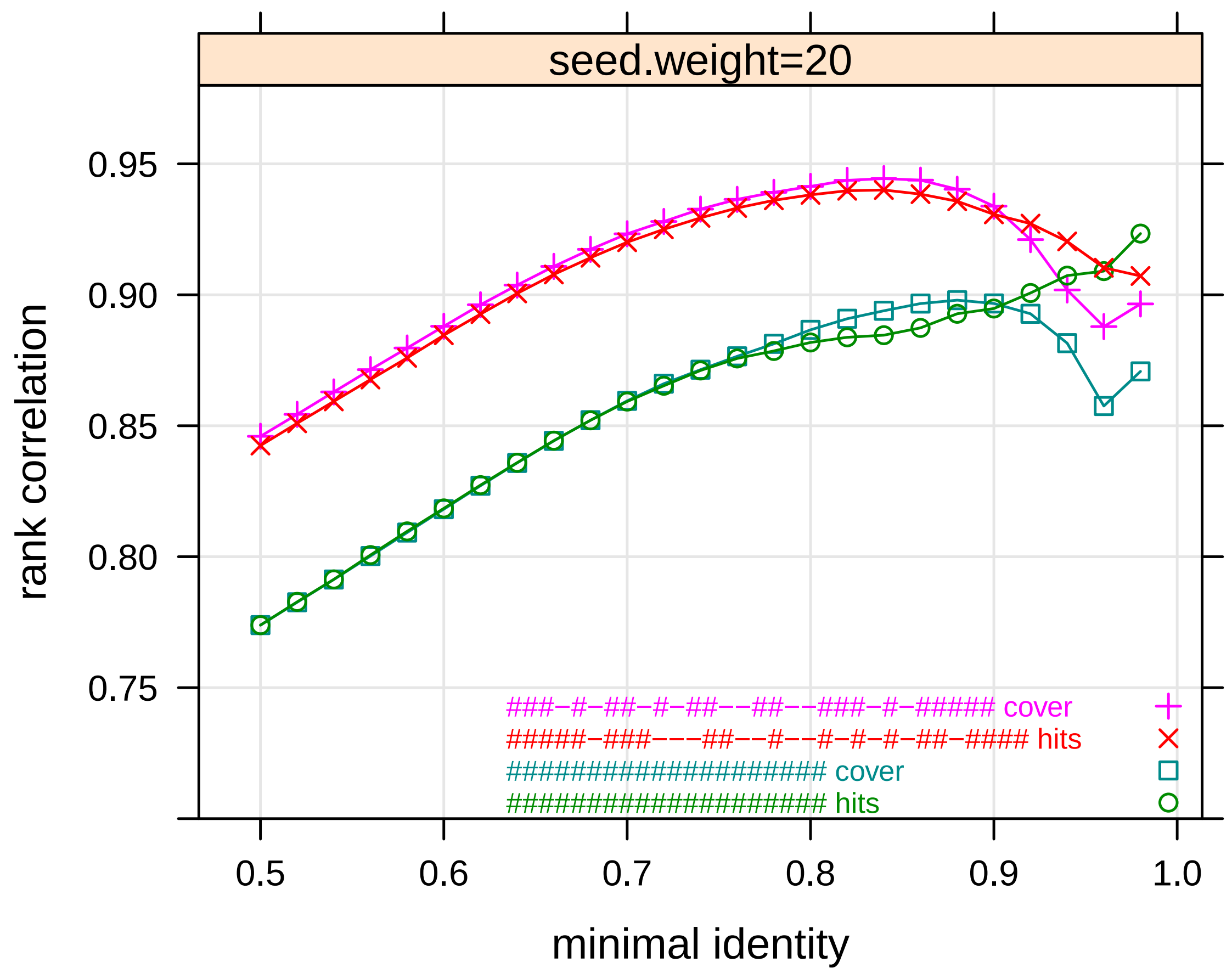


Spaced seeds exhibit a better correlation between errors and score, while *contiguous seeds* plots are more blurred.

SIMULATED ALIGNMENTS

An accurate mapping of a read to a corresponding clade requires estimating its distances to each of the genomes.

For a fixed minimal identity rate p_{id}, we randomly sampled gapless alignments of length 100 with identity rate from interval $[p_{id}..1]$, and collected pairs (number of mismatches, score), where 'score' stands for either **number of hits**, or **coverage** of a given seed. For these data, we plot Spearman's rank correlation.

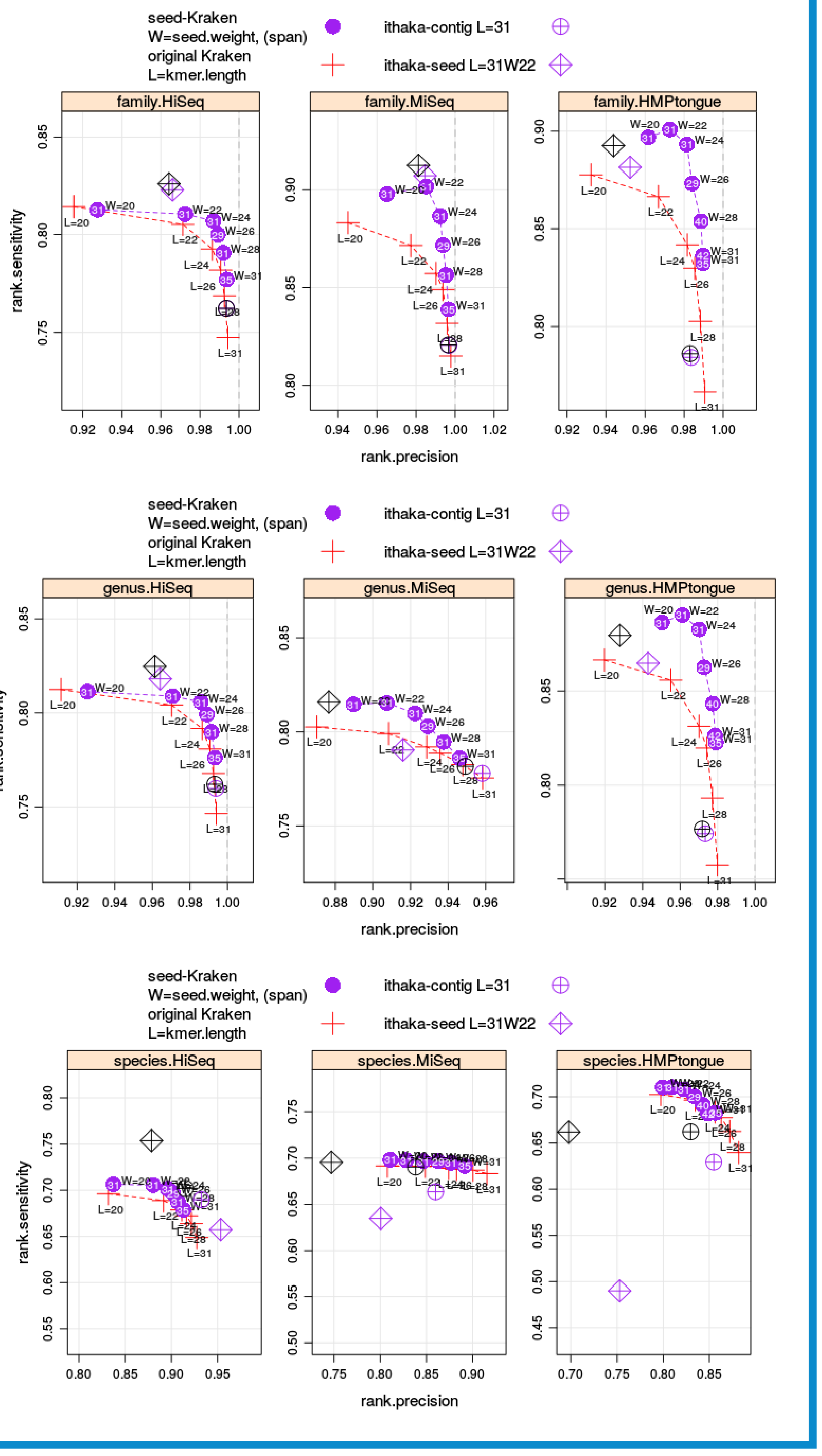


In conclusion, spaced seeds provide a much better distance estimator for alignments whose score ranges over a large interval. For very high-scoring alignments (> 95% of identity), the hit number of contiguous seed becomes a better estimator.

SEED-KRAKEN

Performances of classification of ITHAKA (with spaced seeds), and original KRAKEN, were computed on simulated metagenomes (primarily described and used in [3]): MiSeq (10 bacterial genomes, average error rate), HiSeq, and on 50K subsample of Human Microbiome Project (HMP) Tongue Dorsum wgs sample. Charted are genus precision (positive predictive value) against genus sensitivity (rate of correct assignments). Varying are *k-mer length*, and its spaced seed equivalent *seed weight*, while the *seed span* (not indicated) varies from 31 to 40.

SEED-KRAKEN outperforms original KRAKEN in sensitivity/precision tradeoff (ROC curve characteristics) at the classification levels of genus, and family.


seed-Kraken
W=seed.weight, (span)
original Kraken
L=kmer.length
ithaka-contig L=31
ithaka-seed L=31W22
family.HiSeq
family.MiSeq
family.HMPtongue
genus.HiSeq
genus.MiSeq
genus.HMPtongue
species.HiSeq
species.MiSeq
species.HMPtongue
rank.sensitivity
rank.precision
W=20
W=22
W=24
W=26
W=28
W=31
L=20
L=22
L=24
L=26
L=28
L=31