Spaced seeds improve *k*-mer-based metagenomic classification

Karel Břinda, Maciej Sykulski, Gregory Kucherov

Laboratoire d'Informatique Gaspard-Monge
Université Paris-Est

cf poster and preprint arXiv:1502.06256

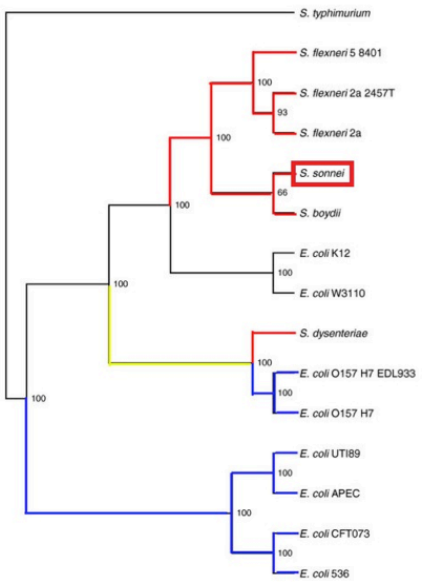
Metagenomics in the era of NGS

- Powerful approach to study genetic material in environmental samples
- Before NGS:
  - involves cloning
  - often based on *marker genes* (16S rRNA)
- Modern metagenomics (HMP, …)
  - Large sets of NGS reads (metagenome)
  - Metagenome is matched against large databases of genomes
- *Main goal*: metagenomic classification (binning)
- (Partial) assembly: a fragile approach
- *Recent solutions*: direct mapping of NGS reads to reference database

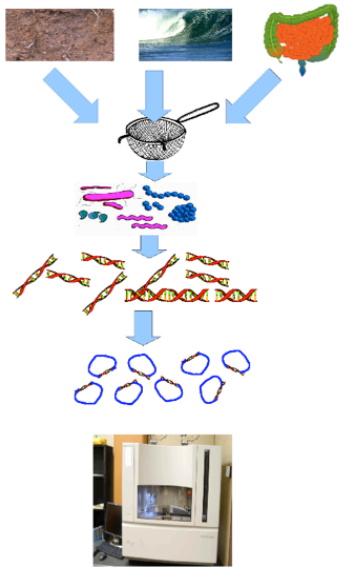
(Taxonomy-dependent) metagenomic classificaton

INPUT

Database of reference genomes + Taxonomic tree



NGS reads (metagenome)



TASK

For each read:

Assign it to a node of the tree

or

Classify it as a novel unit

Typical size of a database: thousands of species
Typical counts of reads: hundreds of millions

Traditional methods: alignment-based

Aligning metagenome reads to each reference genome of the database

- based on **BLAST-like tools** (e.g. Megan, PhymmBL)
  - Good accuracy
  - Very slow
- based on **NGS read mappers** (e.g. Genometa based on Bowtie)
  - Faster (but speed still insufficient)
  - Heuristics only for alignments with high similarity
  - *Example:* BWA index for NCBI's NT database requires 100 GB of RAM

With large data sets produced nowadays, **alignment-free methods are usually used**

Alignment-free methods

- *alignment-free* (or *composition-based*) methods compare sequences by comparing their composition in words, usually *k*-mers

- alignment-free methods do not allow gene/function identification but make it possible to scale the classification to the dimensions of modern metagenomic datasets

- frequencies (multiplicities) of *k*-mers are often taken into account; however, storing frequency information is still too demanding for modern metagenomic applications
- ⇒ in this work we do ***not*** take frequencies into account

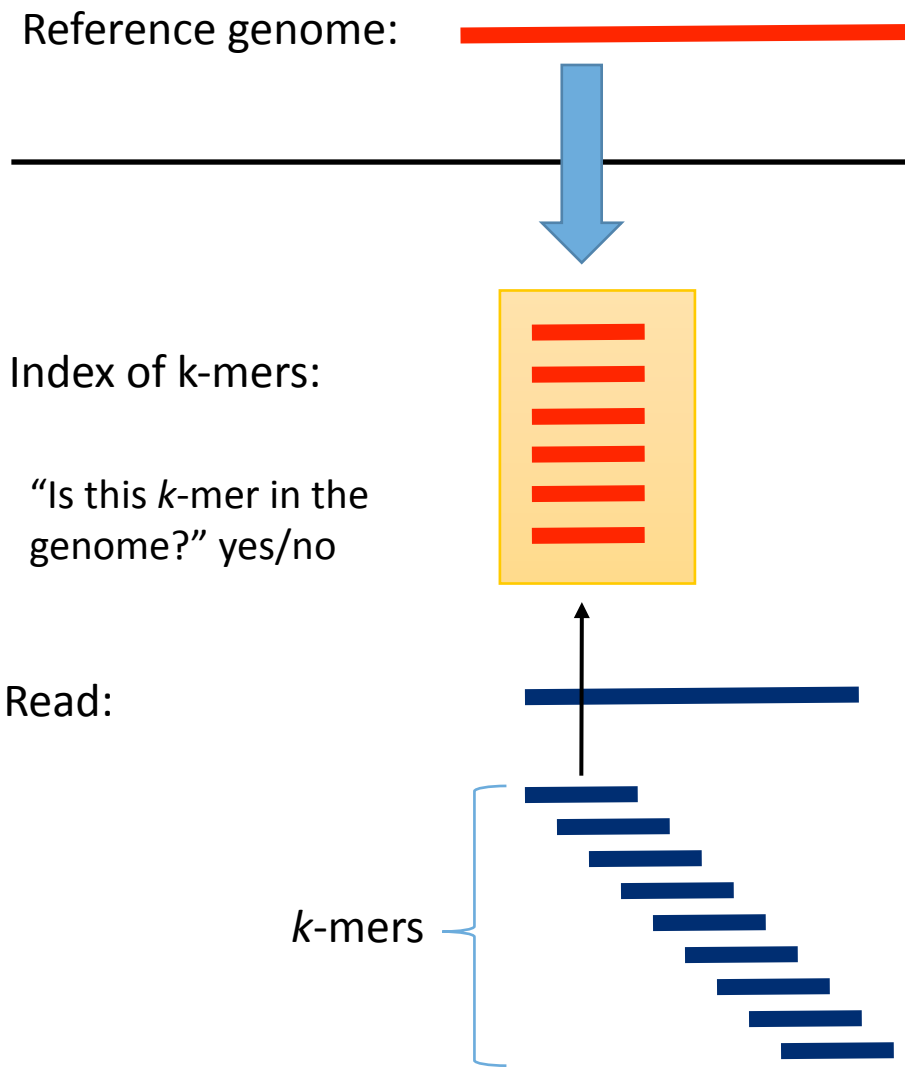
Alignment-free metagenomic classifiers

- LMAT: Ames, *et al.*, Scalable metagenomic taxonomy classification usng a reference genome database, *Bioinformatics* 29, 2013
- Kraken: Wood, Salzberg, Kraken: ultrafast metagenomic sequence classification using exact alignments, *Genome Biology* 15, 2014

More continue to appear …

- R.Ounit, S.Wanamaker, T.Close, S.Lonardi, CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers, BMC Genomics (Jan 2015), e16:236
- J.Kawulok, S.Deorowicz, CoMeta: Classification of Metagenomes Using k-mers, PLOS ONE (April 17, 2015)

Kernel of alignment-free classification



Goal: estimate the "likelihood" of a read to belong to the genome (assign a score)

Example (Kraken): number of occurring *k*-mers (*hit number*)

Spaced seeds

- **Patterns** over **#** (match) and **–** ("don't care") viewed as a mask
  - *Example*: **##–#––#**

```
AGTAGAGGTGAAC
##-#--#          pos 0: AGAG
 ##-#--#         pos 1: GTGG
  ##-#--#        pos 2: TAAT
```

spaced *k*-mers

  - *weight* of a seed = number of **#**'s

- introduced in ~2002 to improve Blast-like seed-and-extend alignment heuristics
  - shown to significantly improve the 'sensitivity-selectivity' trade-off compared to contiguous seeds
  - implemented in many alignment programs: PatternHunter, YASS, ZOOM, LAST, Bfast, SHRiMP, SToRM, …
  - many extensions studied in the 00's

Applying spaced seeds to metagenomics

- recently, spaced seeds have been shown to improve some alignment-free methods as well, e.g. [Morgenstern, *et al*. *BMC Algorithms for Mol. Bio.* **10**, 2015], [Noé, Martin, *J. of Comp. Bio.* **21,** 2014]

In metagenomics:

- comparing a short sequence (*read*) against a long sequence (*genome*)
- data sets are **huge** (both reads and references genomes), which calls for very "minimalistic" techniques

This work: spaced seeds applied to metagenomic classification

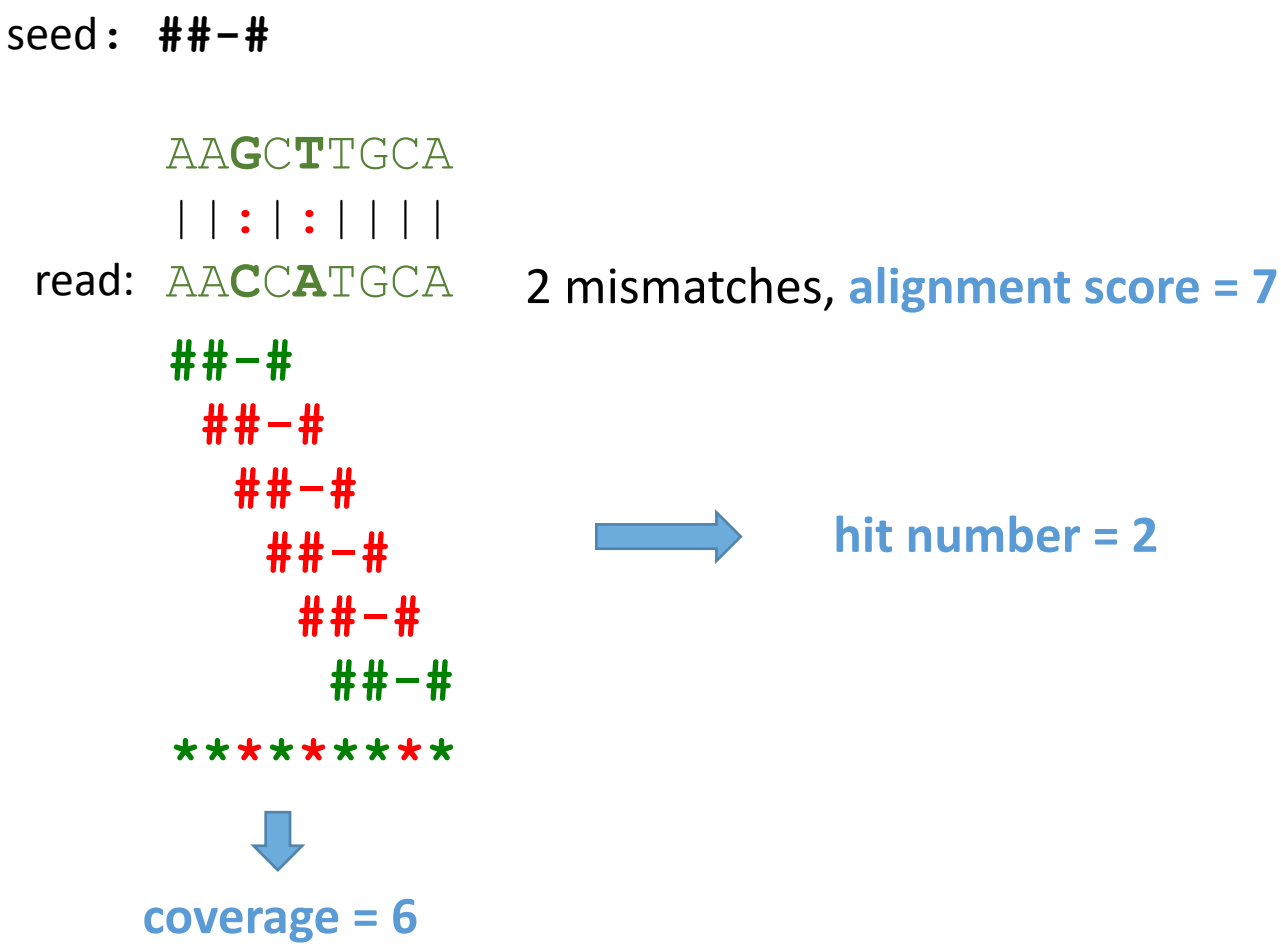
Hit number and coverage estimators

seed: **##-#**

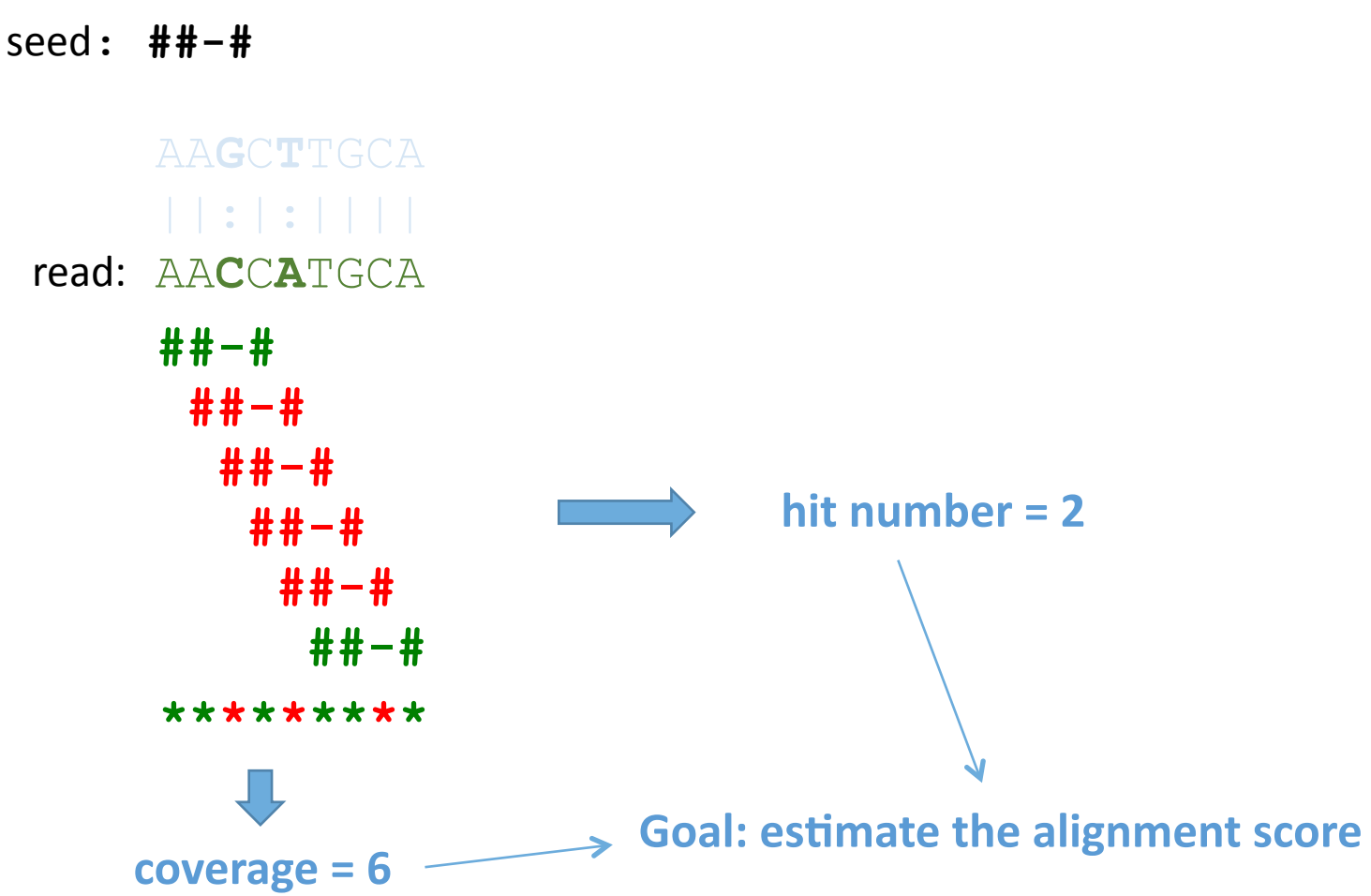
```
        AAGCTTGCA
        ||:|:||||
read:   AACCATGCA
```

2 mismatches, **alignment score = 7**

Hit number and coverage estimators



Hit number and coverage estimators

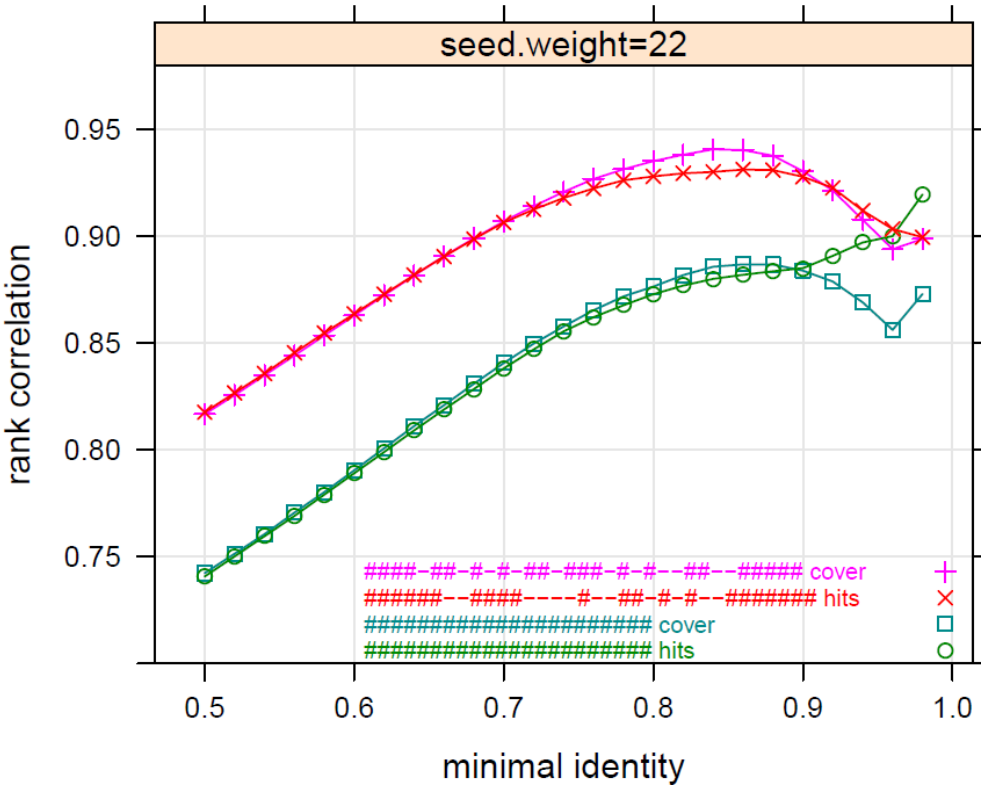


In this talk

- In a series of computational experiments, we study the performance of different estimators using *spaced seeds* and *hit-number*/*coverage* measures

Correlation of score and counts for aligned reads

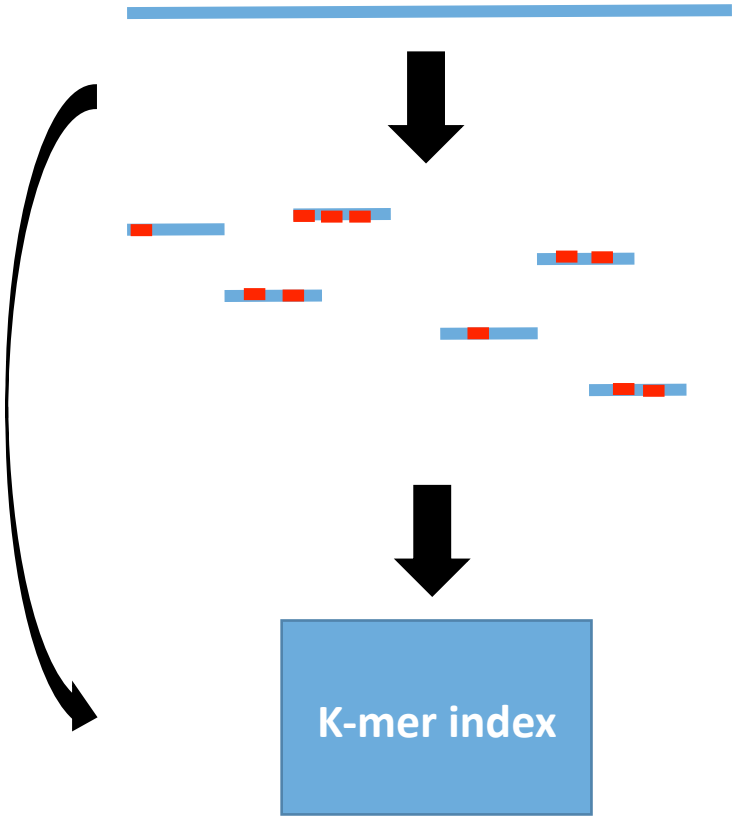
- What estimator correlates best with the alignment quality?
- *Experiment*:
  - fix minimal identity rate p_{id}
  - sample random alignments with random identity rate in [p_{id} .. 1]
  - for these, compute Spearman's rank correlation for
    (id rate, hit count) and
    (id rate, coverage)
- ⇒ Spaced seeds yield a significantly better correlation with id rate, unless *only* high-quality (>95%) alignment are considered



Correlation of score and counts for unaligned reads

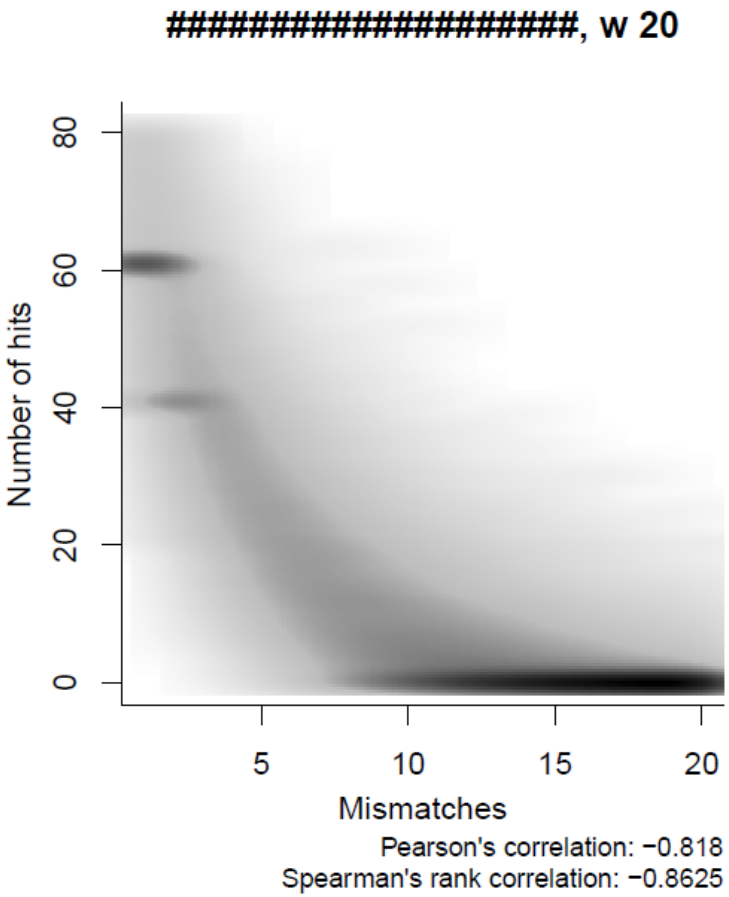
- … and applied to a real genome?
- *Experiment*:
  - pick a genome
  - simulate Illumina-like reads with 0 to 20 mismatches (random)
  - given a seed,
    - construct an index
    - for each read, compute the count (hit number or coverage)
    - plot (# of mismatches, count)

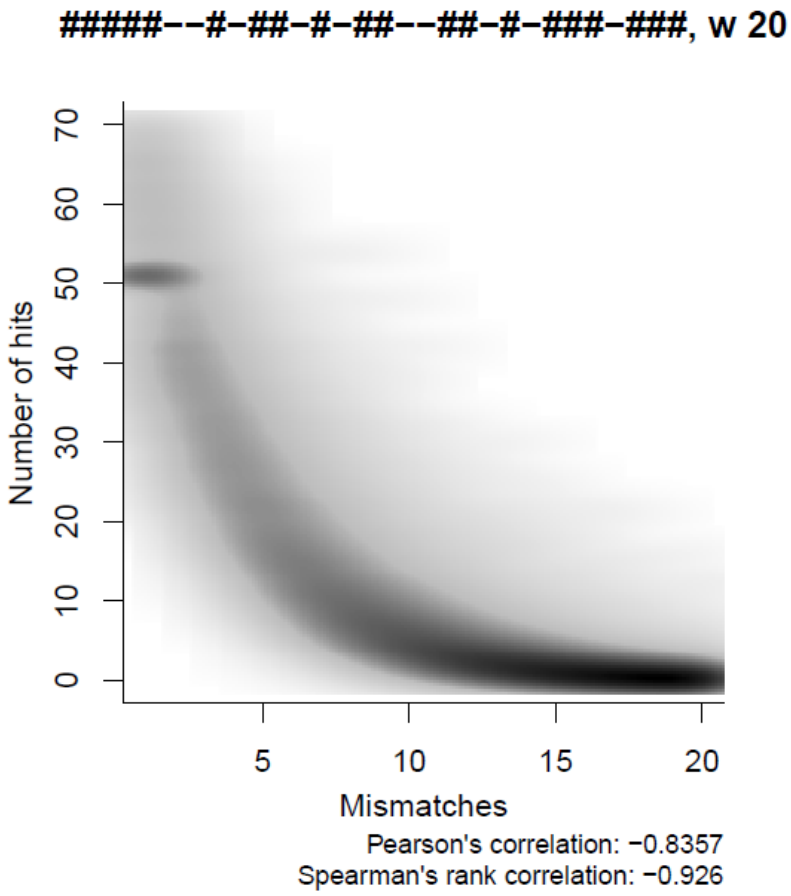
- ⇒ Spaced seeds confirm a significantly better correlation



Correlation of score and counts for unaligned reads: hit count

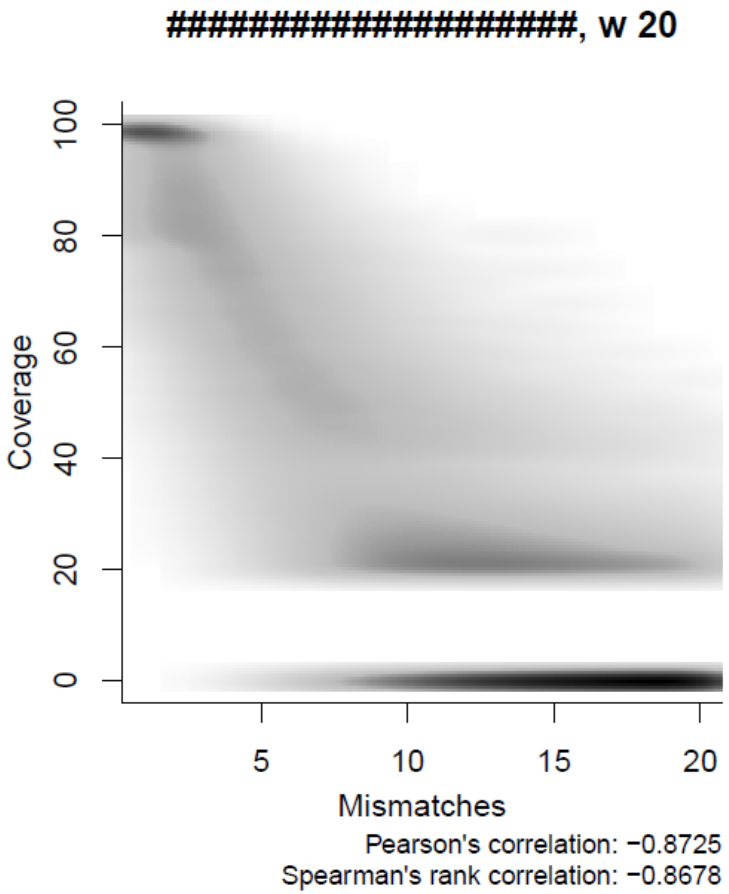
M.tuberculosis

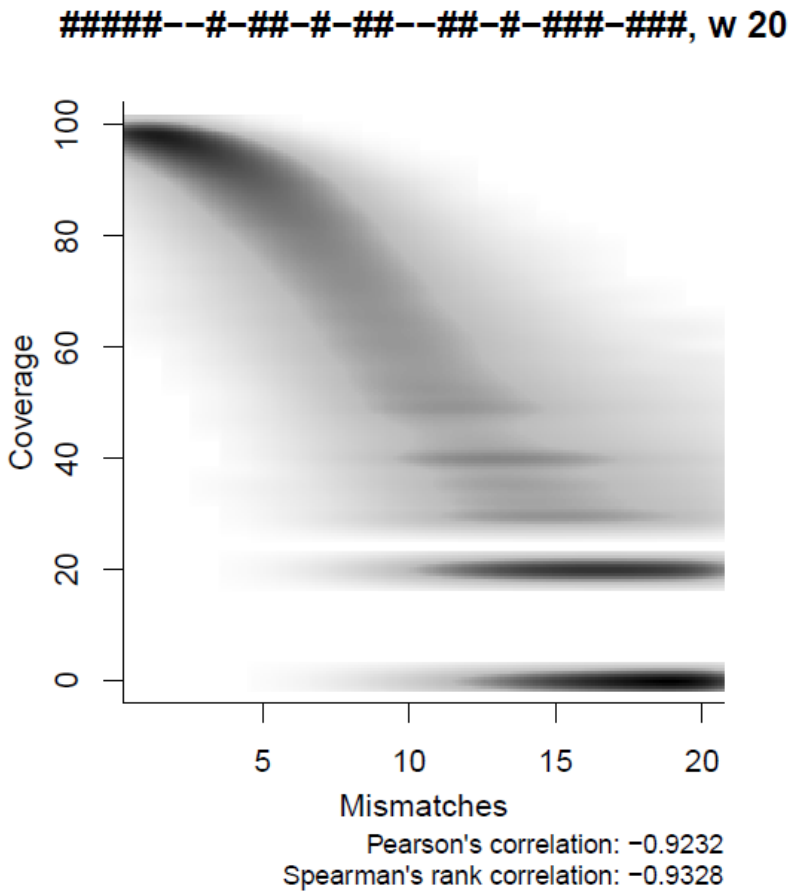




Correlation of score and counts for unaligned reads: coverage

M.tuberculosis



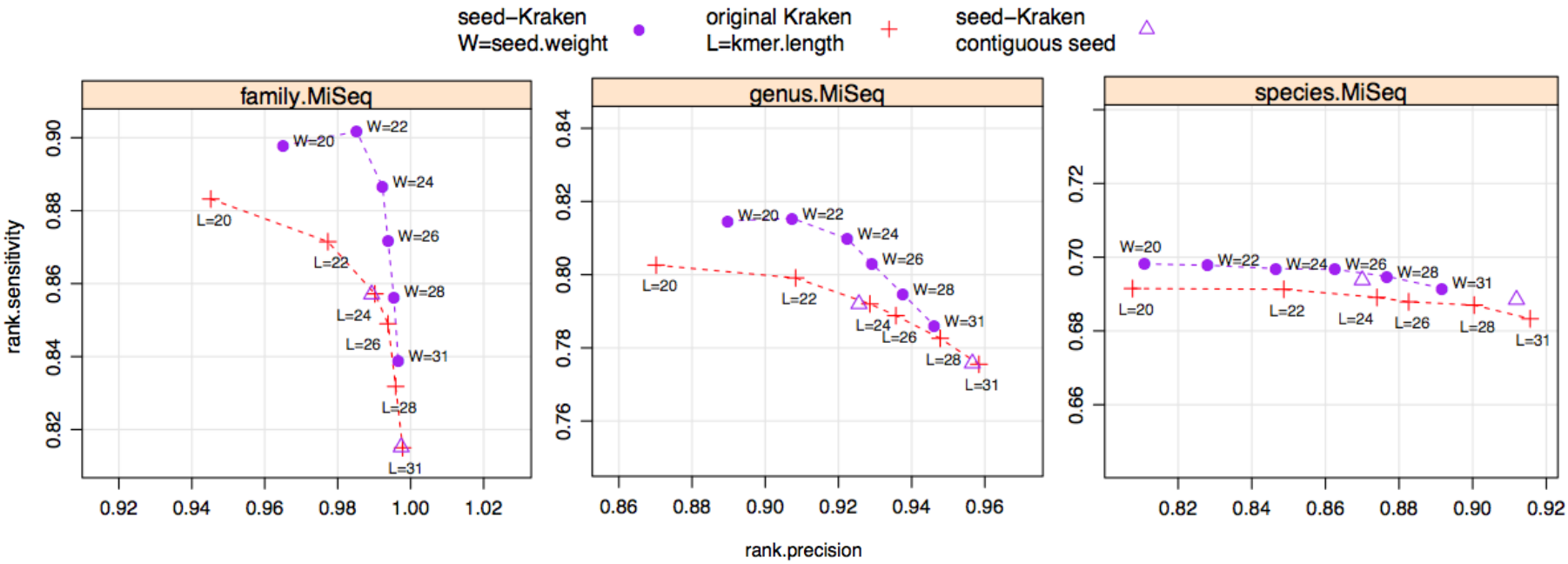


seed-Kraken: modified Kraken for spaced seeds

- seed-Kraken: modified Kraken (and Jellyfish) to work with spaced seeds
- required a minor modification of the assignment algorithm (direct and complementary *k*-mers mapped to distinct index entries); as a result, the index space slightly increased (~5%) and *k*-mer query time doubled
- Kraken default database contains 2,256 genomes and is 70GB; we reduced it (cf below)
- seed-Kraken run on several datasets: MiSeq, HiSeq, simBA-5 (from original Kraken) and HMPtongue (ours)
- for each weight, seed-Kraken has been run on a few seeds and the best was kept

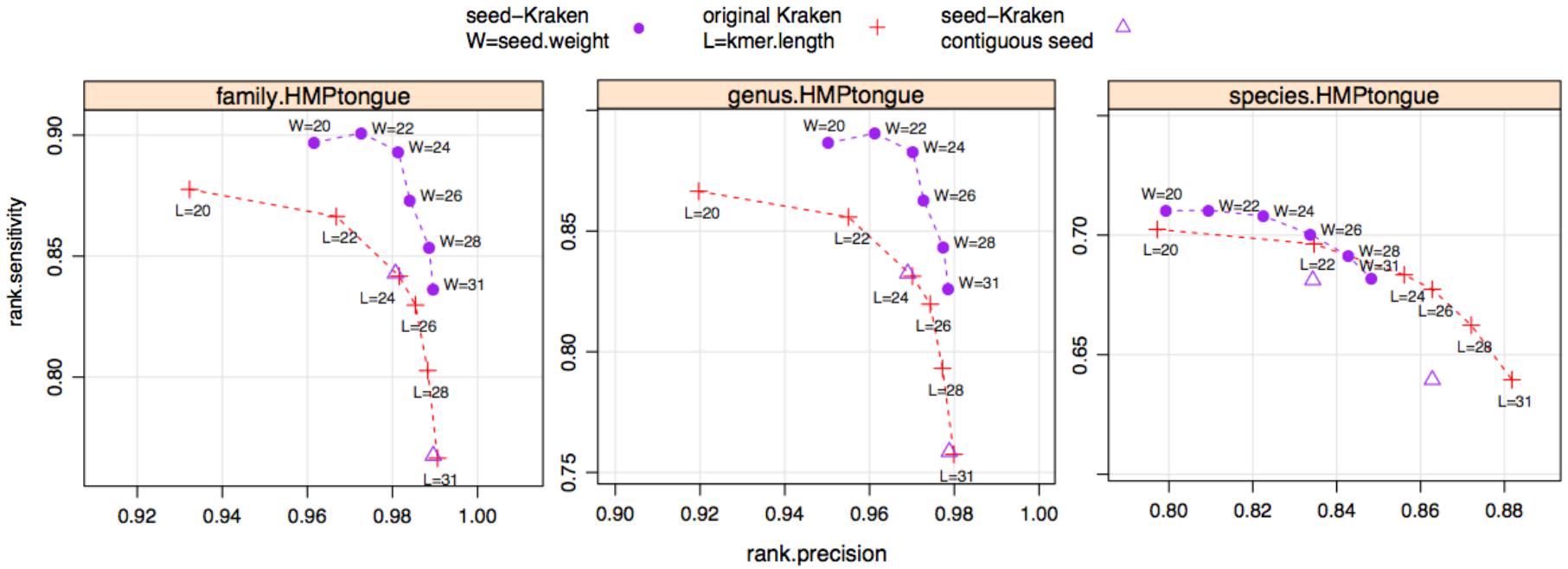
seed-Kraken vs Kraken: MiSeq dataset

- MiSeq: a mix of 10,000 Illumina reads from 10 bacterial genomes
- database selected from Kraken default: one strain per species, 915 genomes, 26GB of memory



seed-Kraken vs Kraken: HMPtongue dataset

- HMPtongue: selection of 50,000 sequences of Tongue dorsum metagenomic sample from Human Microbiome Project
- database augmented by sequences of HMP reference library (~800Mbp)



Conclusions

- *Main message*: spaced seeds improve the classification accuracy
- The improvement is consistent at genus and family levels, more delicate at the species level
- Many further questions:
  - improving the assignment algorithm
  - designing optimal seeds
    - Iedera http://bioinfo.lifl.fr/yass/iedera
    - multiple seeds?
  - designing efficient indexing structures
  - underlying probability mechanisms
- all supplementary information at https://github.com/gregorykucherov/spaced-seeds-for-metagenomics