CONSULT-II:
Taxonomic identification using locality sensitive hashing

Ali Osman Berk Şapcı[1], Eleonora Rachtman[1], and Siavash Mirarab[2]

[1]: Bioinformatics and Systems Biology Graduate Program, University of California, San Diego
[2]: Electrical and Computer Engineering, University of California, San Diego

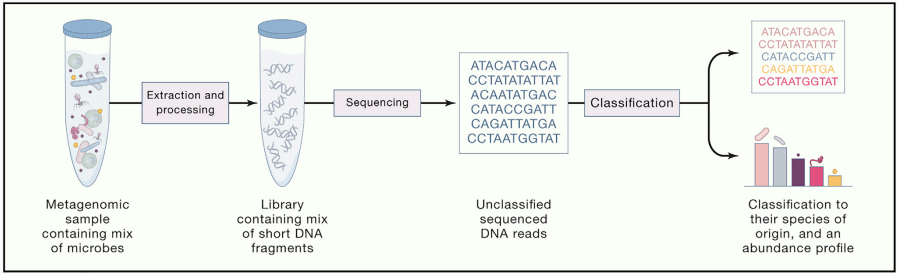
Introduction & Background

Metagenomic analysis

The goal:

Identifying the taxon/taxa present in complex biological and environmental samples.

- Taxonomic classification: drive a single taxon prediction for each sequence.
- Abundance profiling: determine the taxonomic composition of a given sample.



[Ye et al., 2019]

Problem: taxonomic identification

The task:

Classify/profile sequences at the highest resolution (lowest taxonomic rank) possible.

Different approaches:

- *k*-mer based methods, genome-wide alignment, marker-based alignment.
- In essence, matching sequences (e.g., short reads) with genomes from a reference.
- Then, using the matching information in a clever way, perform;
  - taxonomic classification and/or abundance profiling.

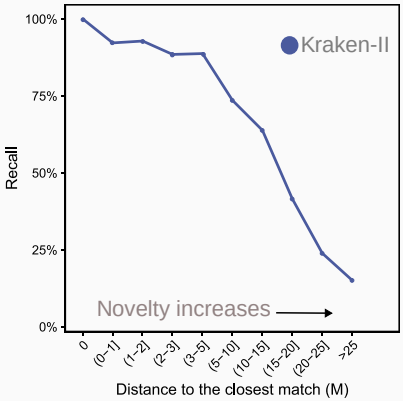
Novel sequences challenge popular methods

Challenge:

Novel sequences, i.e., sequences without a close match in the reference set.

Two *k*-mer-based popular methods:

- CLARK [Ounit et al., 2015] and Kraken-II [Wood et al., 2019].
- They fail frequently for novel genomes [Rachtman et al., 2020, Pachiadaki et al., 2019].



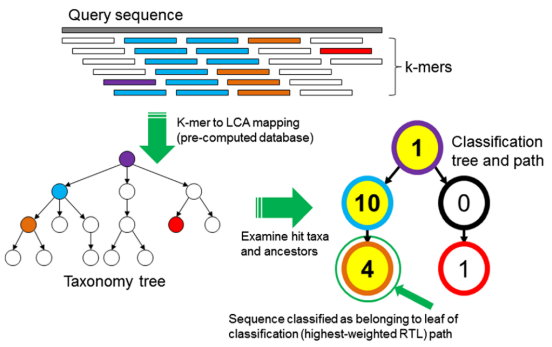
[Rachtman et al., 2020]

Existing k-mer-based metagenomic analysis methods

CLARK and Kraken
Both rely on exact k-mer matches.

Kraken

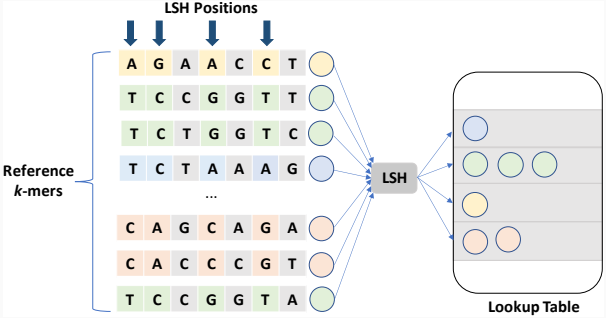
- Each k-mer in the sequence is mapped to least common ancestor of the genomes containing that k-mer.
- Counts mapped k-mers to compute a heighest-weighted root-to-leaf path.



[Wood and Salzberg, 2014]

Locality sensitive hashing and CONSULT

- Orignally designed for contamination detection.
- Uses LSH to partition the k-mers in the reference set into big lookup tables.
  - Each row $\rightarrow$ "similar" k-mers.
- For each k-mer of a query;
  - Is there any reference k-mer with Hamming distance less than some threshold p to the query k-mer?
- Allowing inexact matches is the key feature.



[Rachtman et al., 2021]

Extending CONSULT for taxonomic identification

- CONSULT can not detect which reference species matches with a given read.
- Remembering what reference genomes include each reference *k*-mer?
  - Practically infeasible in terms of memory.
  - Even the *k*-mer encodings and library indices require 120Gb with a modern microbial dataset with 8 billion *k*-mer.

Goals

- Save some taxonomic information with reference *k*-mers, but keep the memory manageable.
- Derive a final taxonomic group from all the exact/inexact matches.

Constructing the Reference Library

Library construction: saving a taxon ID per *k*-mer

Idea

- Instead of keeping all species-level IDs of each genome with a given *k*-mer, compute and save the ID of the LCA taxon.
  - 2 bytes for each LCA taxon ID, 16Gb in total for 8 billion *k*-mers.

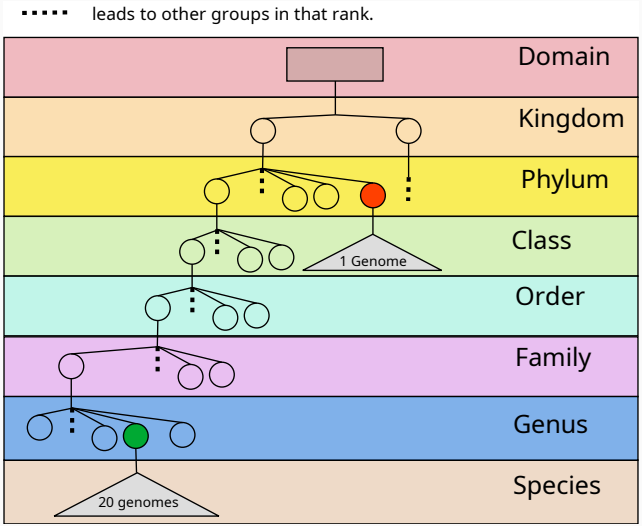
Problem

- Pushing up taxonomic identifications due to errors in the reference.
- Saving kingdom level IDs is not very useful, we want to be as specific as possible.

Example - Library construction: saving a taxon ID per *k*-mer

Example

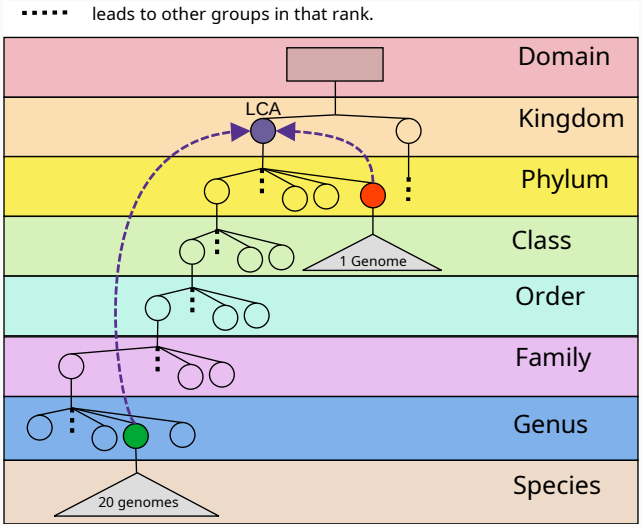
- 20 genomes from the same genus and 1 genome from an incorrect phylum.
- Kraken would push the LCA to the kingdom rank.
- Saving kingdom-level IDs is not very useful.



Example - Library construction: saving a taxon ID per *k*-mer

Example

- 20 genomes from the same genus and 1 genome from an incorrect phylum.
- Kraken would push the LCA to the kingdom rank.
- Saving kingdom-level IDs is not very useful.



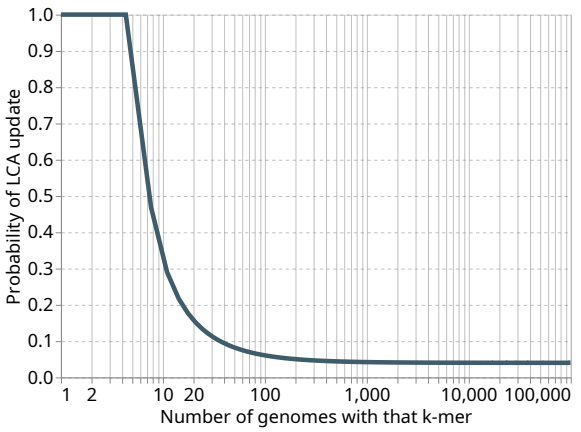
Probabilistic LCA taxon computation

- N_i: number of genomes including k-mer i.
- For each genome having the k-mer i, update the LCA taxon with probability $p_u(N_i)$.

Intuition

- A k-mer should appear sufficiently many times in a group to affect the LCA taxon.
  - Frequent k-mers $\rightarrow$ many times.
  - Rare k-mers $\rightarrow$ a few would be enough.

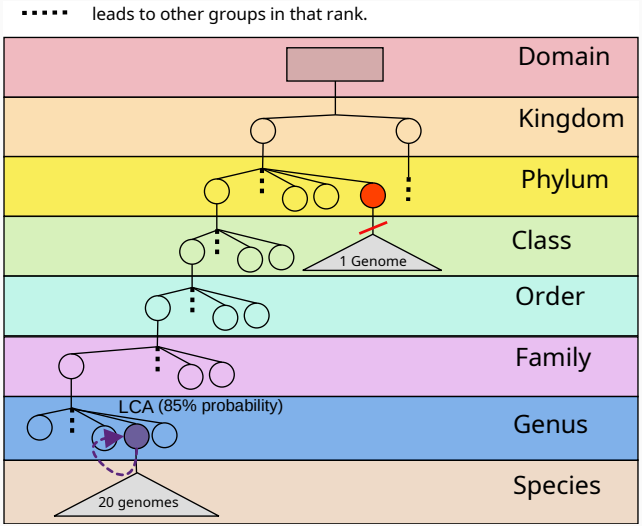
- Two parameters; rate of decrease and the offset of the function p_u.



Example - Probabilistic LCA taxon computation

Example

- 20 genomes from the same genus and 1 genome from an incorrect phylum.
- 85% probability $\rightarrow$ LCA taxon is the correct genus.



Reference library construction

CONSULT-II reference library[1]:

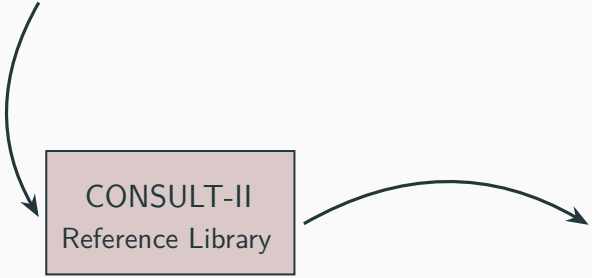
- Tree of Life (ToL) (Zu et al., 2019) microbial genomic dataset.
  - 10,470 microbial species in total (after removing query genomes for testing).
  - 11,920 taxa in total.
- All unique canonical 35-mers from all genomes minimized down to 32-mers.

[1] Extending the library constructed by original CONSULT [Rachtman et al., 2021]

Given a read;

`@READ-ID.X-XXX`

`ATACGATTACAGGGGAGATT…`



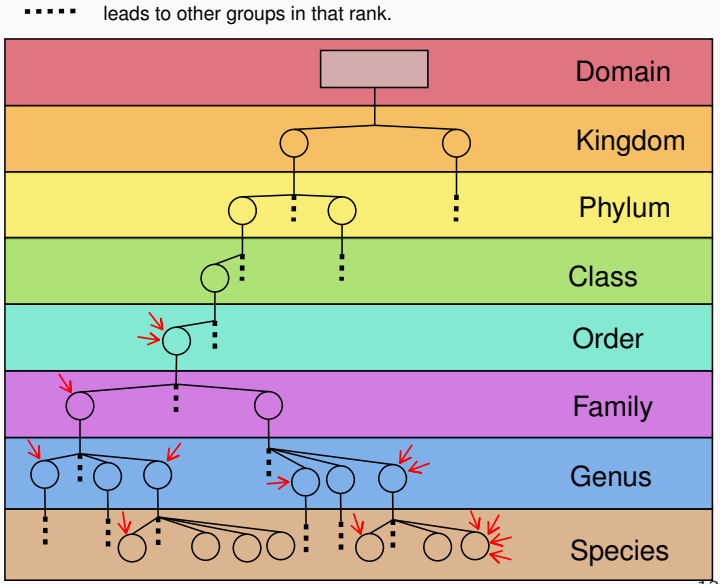
A list of

`TaxonomicID:HammingDistance`

`@READ-ID.X-XXX`

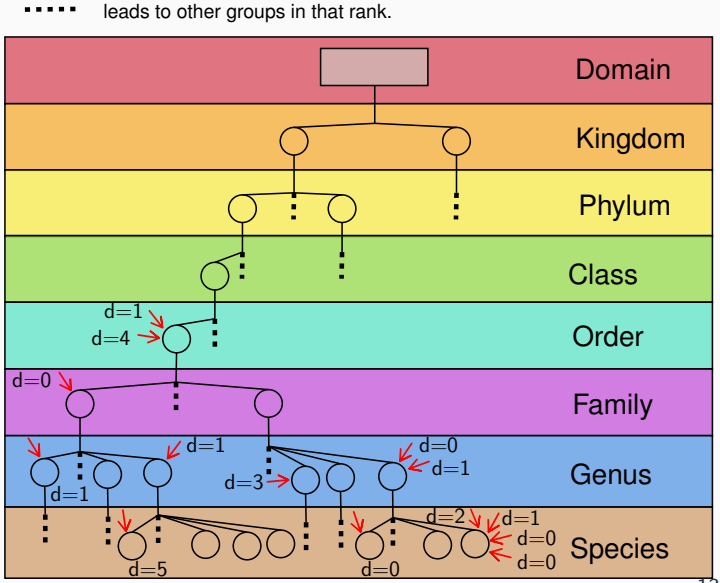
- `8770:0`
- `8770:2`
- `8770:0`
- `8770:1`
- `8770:1`
- `3030:0`
- `3030:0`
- `...`

```
@READ-ID.X-XXX
- 8770:0
- 8770:2
- 8770:0
- 8770:1
- 8770:1
- 3030:0
- 3030:0
- ...
```



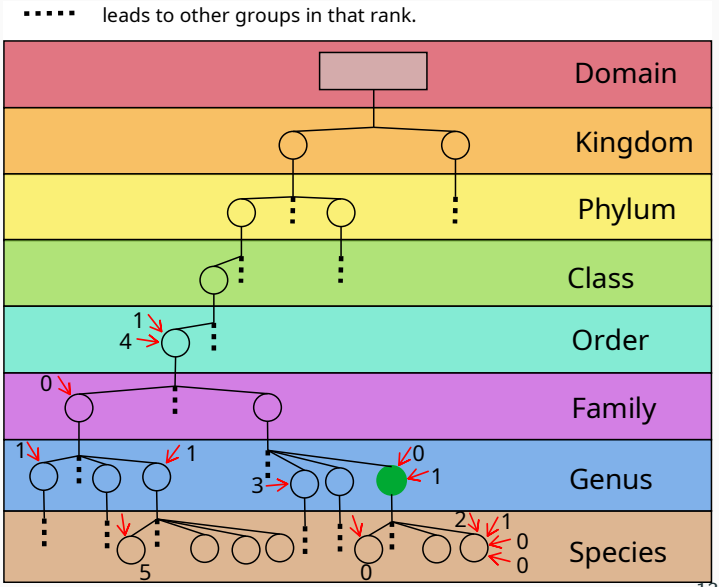
@READ-ID.X-XXX

- 8770:0
- 8770:2
- 8770:0
- 8770:1
- 8770:1
- 3030:0
- 3030:0
- ...



@READ-ID.X-XXX
- 8770:0
- 8770:2
- 8770:0
- 8770:1
- 8770:1
- 3030:0
- 3030:0
- ...

- Classifying under the green genus seems to be correct.
- How would we model this algorithmically?



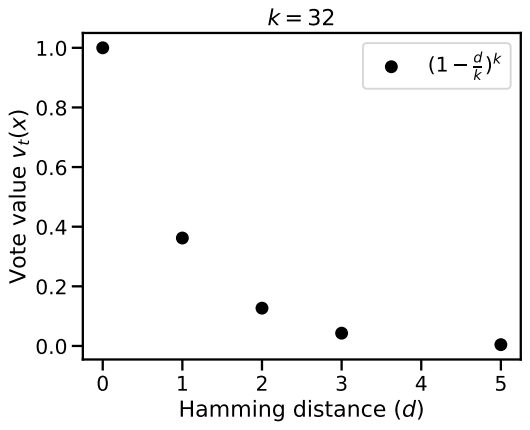
Taxonomic Classification Algorithm

A vote-based taxonomic identification approach

- Consider each k-mer match as a vote to the corresponding taxon, weighted by its distance.

$$\mathrm{v}_t(x) = \left(1 - \frac{d}{k}\right)^k \mathbb{1}\{d \leq d_{\max}\}$$

where x is a k-mer, d is the Hamming distance between x and its closest k-mer in the reference, and t is the taxon.

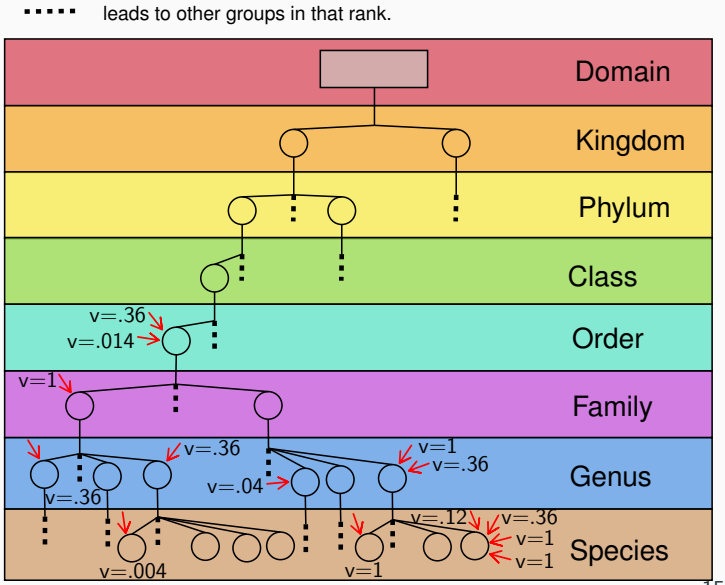


- Vote values drops close to exponentially w.r.t. distance d.

Vote of k-mer x to the taxon t:

$$\mathrm{v}_t(x) = \left(1 - \frac{d}{k}\right)^k \mathbb{1}\{d \leq d_{\max}\}$$

- d=0 → v=1
- d=1 → v=0.36
- d=2 → v=0.12
- d=3 → v=0.04
- d=4 → v=0.014
- d=5 → v=0.004



Aggregating votes to derive a single taxon ID

- We have the taxonomic tree.
  - We can incorporate the hierarchical relationships between taxa.
- To aggregate vote values, recursively sum up individual votes contributed by each k-mer in a bottom-up manner;

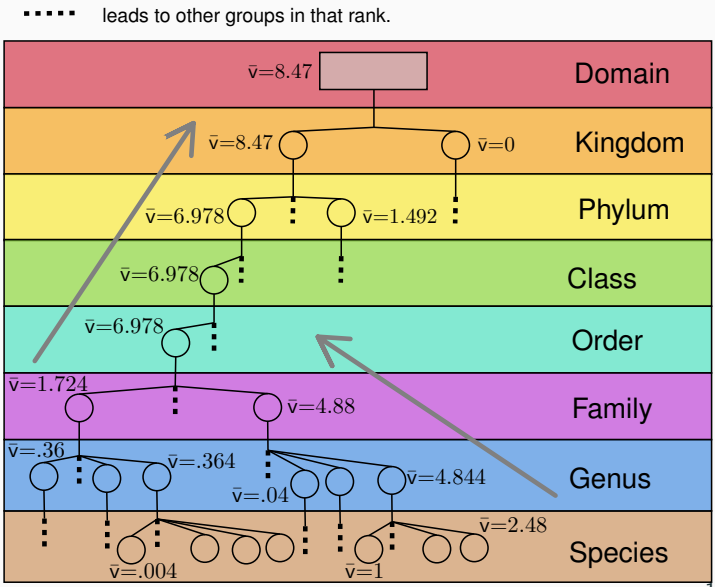
$$\bar{\mathrm{v}}(t) = \sum_{x \in \mathcal{R}} \mathrm{v}_t(x) + \sum_{t' \in \mathrm{C}(t)} \bar{\mathrm{v}}\left(t'\right)$$

where $\mathrm{C}(t)$ is the set of children of the taxon t in the taxonomic tree.

The total vote for the taxon t:

$$\bar{\mathrm{v}}(t) = \sum_{x \in \mathcal{R}} \mathrm{v}_t(x) + \sum_{t' \in \mathrm{C}(t)} \bar{\mathrm{v}}(t')$$

Total vote values increase as we go up in the tree, reaches its maximum at the root.



The total vote for the taxon t:

$$\bar{\mathrm{v}}\left(t\right)=\sum_{x\in\mathcal{R}}\mathrm{v}_t\left(x\right)+\sum_{t'\in\mathrm{C}(t)}\bar{\mathrm{v}}\left(t'\right)$$

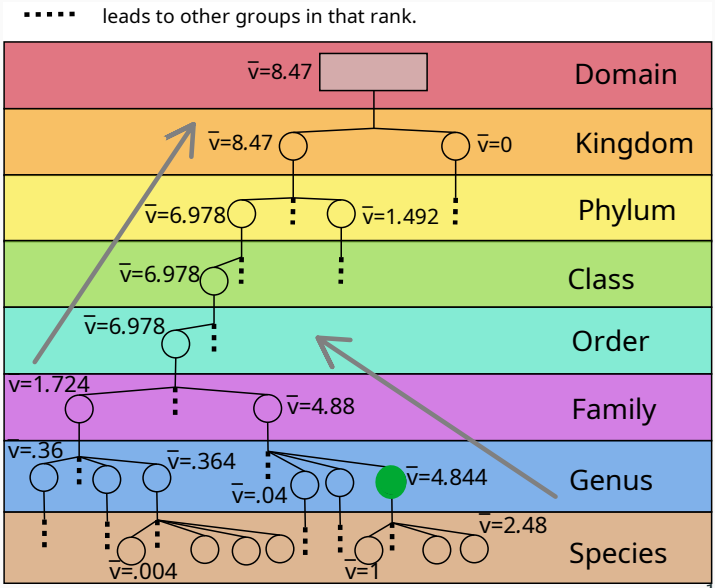
To balance specificity and sensitivity, we require a majority vote by;

$$\tau=0.5\max_{t\in\mathcal{T}}\bar{\mathrm{v}}\left(t\right).$$

Example

$$\bar{\mathrm{v}}\left(t\right)=8.47$$

$$\text{threshold}=4.235$$



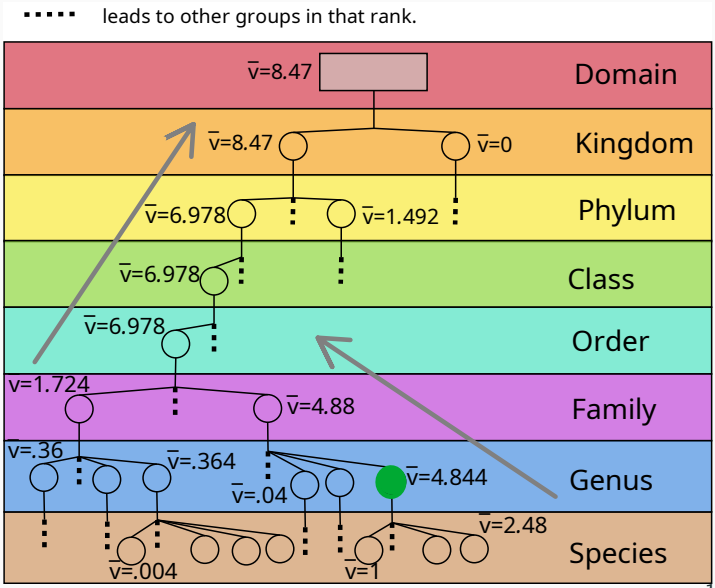
The total vote for the taxon t:

$$\bar{\mathrm{v}}(t) = \sum_{x \in \mathcal{R}} \mathrm{v}_t(x) + \sum_{t' \in \mathrm{C}(t)} \bar{\mathrm{v}}(t')$$

To balance specificity and sensitivity, we require a majority vote by;

$$\tau = 0.5 \max_{t \in \mathcal{T}} \bar{\mathrm{v}}(t).$$

Note that
At a given rank, threshold τ gives a unique taxon.

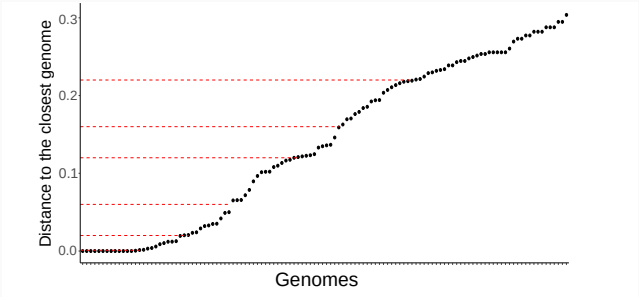


Controlled novelty experiment for taxonomic classification

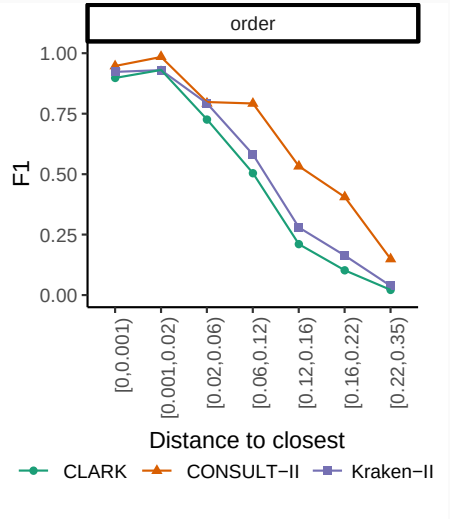
Compared with Kraken-II [Wood et al., 2019] and CLARK [Ounit et al., 2015].

Queries: 120 bacterial genomes.

- Selected with controlled novelty.
- Novelty is defined based on Mash distances to the closest species in the reference.
- Seven categories, with at least 11 genomes in each.



CONSULT-II achieves higher F1-scores in the controlled novelty experiments



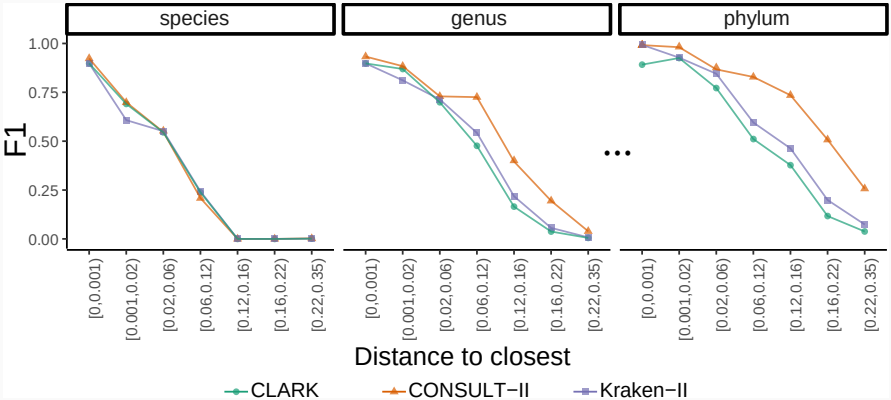
We evaluate each rank separately.

Evaluation

- If classified in the given rank or in a lower rank;
  - **TP:** Classified in the correct lineage.
  - **FP:** Classified in the false lineage.
- If not classified or classified in an upper rank;
  - **TN:** The true taxon is in the reference set.
  - **FN:** The true taxon is not in the reference set.

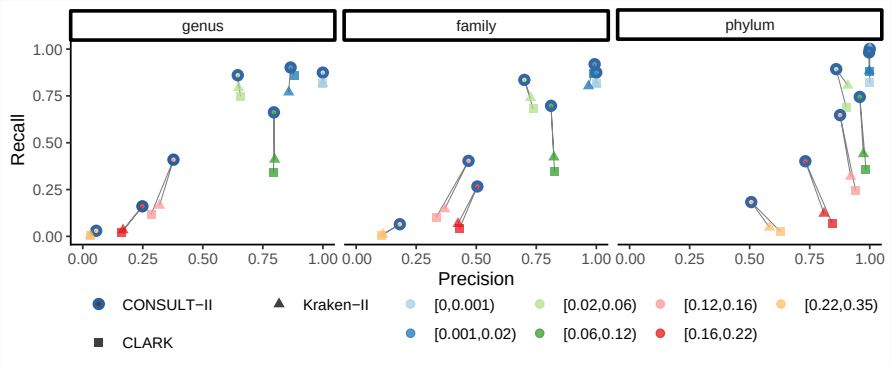
CONSULT-II usually performs much better than other methods for novel genomes.

CONSULT-II achieves higher F1-scores in the controlled novelty experiments



- As queries become more novel, accuracy drops across all ranks for all methods.
- Except at species level, CONSULT-II clearly outperforms for novel genomes.
- Improvements are larger for upper levels, e.g., phylum, class, order.

Precision-recall comparison in controlled novelty experiments



- CONSULT-II has universally higher recall levels.
- All methods have comparable precision levels.
  - Often higher for family and genus.
  - Lower or comparable for phylum.
- Better recall usually comes with no expense of precision.

Computing Abundance Profiles

Utilizing total vote values for abundance profiling

We can further utilize total vote values to derive abundance profiles.

For each rank separately, normalize total vote values to derive a profile vector $\bar{\mathrm{v}}\,(\mathrm{t})$ for $t \in \mathcal{T}_l$ for rank l;

$$p_t^l = \frac{\sqrt{\bar{\mathrm{v}}\,(\mathrm{t})}}{\sum_{t' \in \mathcal{T}_l} \sqrt{\bar{\mathrm{v}}\,(\mathrm{t}')}}$$

where p_t^l is the profile value of taxon t from level l. Then, the abundance profile for rank l is given by $\mathrm{p}^l = \left[p_t^l\right]_{t \in \mathcal{T}_l}$.

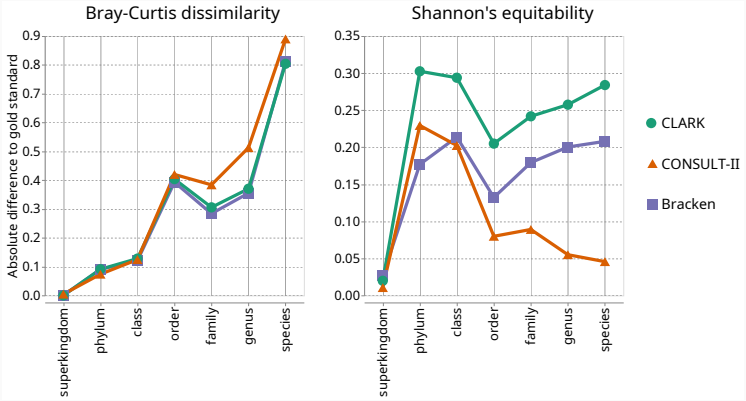
Abundance profiling experiment

Comparison with Bracken [Lu et al., 2017] and CLARK [Ounit et al., 2015].

Queries: CAMI benchmarking challenge.

- CAMI-1 dataset.
- Subsampled the original sample down to 10 million reads.
- Evaluated using two metrics computed by the OPAL tool [Meyer et al., 2019].
  - Shannon's equitability to measure alpha diversity at each rank.
  - Bray-curtis to measure normalized error of abundance estimates at each rank.

Evaluation of profile estimates using different metrics



- Shannon's equitability measures the variety of taxa present in a sample.
  - Outperforms especially in the family, species, and genus levels.
  - May be due to higher recall.

- Bray-Curtis dissimilarity quantifies compositional dissimilarity.
  - Comparable except family and genus levels.

Conclusions

Conclusions

- A promising launching point.
- Controlled novelety experiment shows that LSH-based CONSULT-II identifies novel genomes better.
- Our vote-based approach provides a rich representation for the reads, which can be successfully used in abundance profiling.
- Our heuristics have no theoretical guarantees, but performed well empirically.

Future directions:

- A theoretical framework for the vote and LCA-update probability functions.
- A distance-based phylogenetic placement approach.
- Reducing memory requirements;
  120Gb hash table and k-mer encodings + 16Gb taxonomic IDs + 16Gb k-mer counts $>$ 150Gb

Acknowledgement

Siavash Mirarab & Eleonora Rachtman

References

J. Lu, F. P. Breitwieser, P. Thielen, and S. L. Salzberg. Bracken: estimating species abundance in metagenomics data. *PeerJ Computer Science*, 3:e104, Jan. 2017. ISSN 2376-5992. doi: 10.7717/peerj-cs.104. URL https://doi.org/10.7717/peerj-cs.104.

F. Meyer, A. Bremges, P. Belmann, S. Janssen, A. C. McHardy, and D. Koslicki. Assessing taxonomic metagenome profilers with OPAL. *Genome Biology*, 20(1):51, 2019. ISSN 1474-760X. doi: 10.1186/s13059-019-1646-y. URL https://doi.org/10.1186/s13059-019-1646-y.

R. Ounit, S. Wanamaker, T. J. Close, and S. Lonardi. CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. *BMC Genomics*, 16(1):236, Dec. 2015. ISSN 1471-2164. doi: 10.1186/s12864-015-1419-2. URL http://www.biomedcentral.com/1471-2164/16/236.

M. G. Pachiadaki, J. M. Brown, J. Brown, O. Bezuidt, P. M. Berube, S. J. Biller, N. J. Poulton, M. D. Burkart, J. J. La Clair, S. W. Chisholm, and R. Stepanauskas. Charting the Complexity of the Marine Microbiome through Single-Cell Genomics. *Cell*, 179(7):1623–1635.e11, 2019. ISSN 0092-8674. doi: https://doi.org/10.1016/j.cell.2019.11.017. URL http://www.sciencedirect.com/science/article/pii/S0092867419312735.

E. Rachtman, M. Balaban, V. Bafna, and S. Mirarab. The impact of contaminants on the accuracy of genome skimming and the effectiveness of exclusion read filters. *Molecular Ecology Resources*, 20(3):1755–0998.13135, May 2020. ISSN 1755-098X. doi: 10.1111/1755-0998.13135. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/1755-0998.13135.

E. Rachtman, V. Bafna, and S. Mirarab. CONSULT: accurate contamination removal using locality-sensitive hashing. *NAR Genomics and Bioinformatics*, 3(3): 10.1101/2021.03.18.436035, June 2021. ISSN 2631-9268. doi: 10.1093/nargab/lqab071. URL https://academic.oup.com/nargab/article/doi/10.1093/nargab/lqab071/6342218?guestAccessKey=05db43e1-b54e-4b43-9f04-ee5c92aa0366.

D. E. Wood and S. L. Salzberg. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology*, 15(3):R46, 2014. ISSN 1465-6906. doi: 10.1186/gb-2014-15-3-r46. URL `http://genomebiology.biomedcentral.com/articles/10.1186/gb-2014-15-3-r46`. arXiv: Figures, S., 2010. Supplementary information. Nature, 1(c), pp.1–7. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3006164&tool=pmcentrez&re ISBN: 1465-6914 (Electronic)\r1465-6906 (Linking).

D. E. Wood, J. Lu, and B. Langmead. Improved metagenomic analysis with Kraken 2. *Genome Biology*, 20(1):257, 2019. ISSN 1474-760X. doi: 10.1186/s13059-019-1891-0. URL `https://doi.org/10.1186/s13059-019-1891-0`.

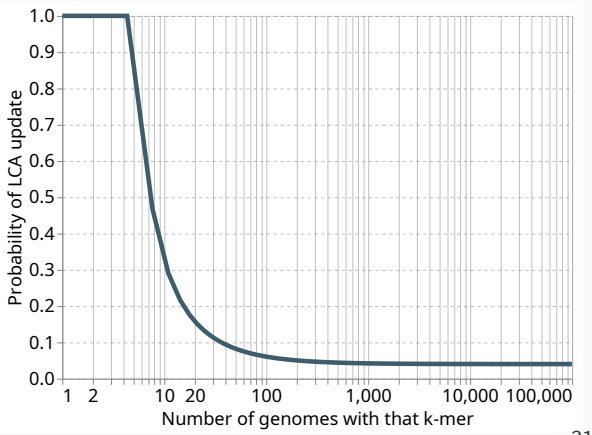
S. H. Ye, K. J. Siddle, D. J. Park, and P. C. Sabeti. Benchmarking Metagenomics Tools for Taxonomic Classification. *Cell*, 178(4):779–794, Aug. 2019. ISSN 1097-4172. doi: 10.1016/j.cell.2019.07.010.

Extra Slides

Probability funciton for LCA computation

$$p_u(N_i) = \min\left\{\frac{w}{\max\{N_i + w - s, w\}} + \frac{1}{s^2}, 1\right\}$$

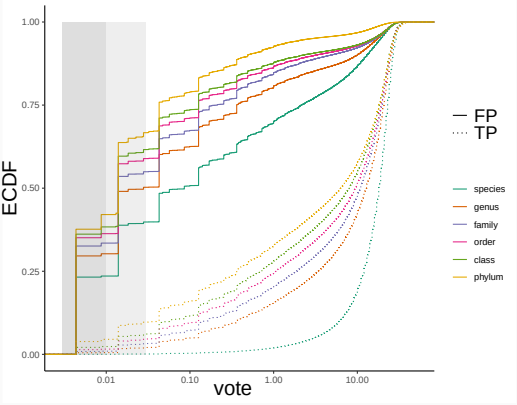
- N_i: number of genomes including k-mer i
- For each genome having the k-mer i, update the LCA taxon with probability $p_u(N_i)$.
  - **w**: rate of decrease
  - **s**: the offset of the function p_u

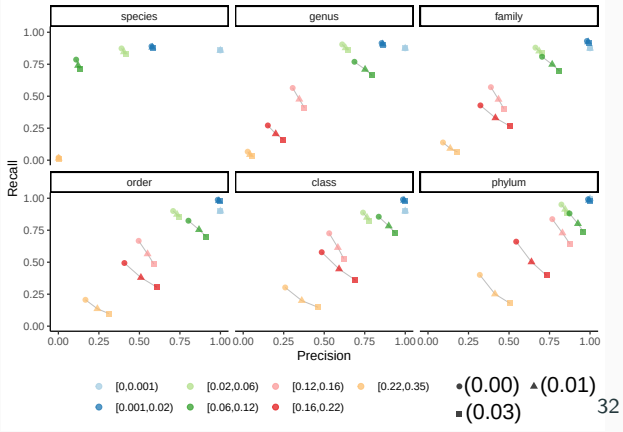


Controlling precision-recall tradeoff with total vote threshold

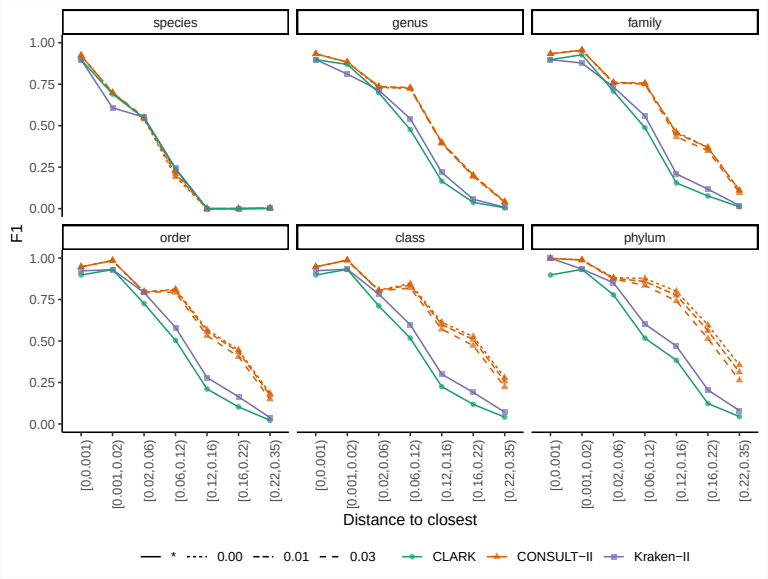
The empirical cumulative distribution of total votes for TP/FP shows the tradeoff.

Removing classifications with low total-vote values increases precision by sacrificing some recall.





CONSULT-II achieves higher F1-scores in the controlled novelty experiments



Precision-recall comparison in controlled novelty experiments