Large-scale Machine Learning for Metagenomics Sequence Classification - Supplementary Materials

Kévin Vervier, Pierre Mahé , Maud Tournoud,
Jean-Baptiste Veyrieras, Jean-Philippe Vert

October 26, 2015

Contents

1 Details about the data used

Supplementary Table 1 shows the list of species present in the *small* database.

Acetobacter pasteurianus
Acinetobacter baumannii
Bacillus amyloliquefaciens
Bacillus anthracis
Bacillus subtilis
Bacillus thuringiensis
Bifidobacterium bifidum
Bifidobacterium longum
Borrelia burgdorferi
Brucella abortus
Brucella melitensis
Buchnera aphidicola
Burkholderia mallei
Burkholderia pseudomallei
Campylobacter jejuni
Corynebacterium pseudotuberculosis
Corynebacterium ulcerans
Coxiella burnetii
Desulfovibrio vulgaris
Enterobacter cloacae
Escherichia coli
Francisella tularensis
Helicobacter pylori
Legionella pneumophila
Leptospira interrogans
Listeria monocytogenes
Methylobacterium extorquens
Mycobacterium bovis
Mycobacterium tuberculosis
Mycoplasma fermentans
Mycoplasma genitalium
Mycoplasma mycoides
Mycoplasma pneumoniae
Neisseria gonorrhoeae
Propionibacterium acnes
Pseudomonas aeruginosa
Pseudomonas stutzeri
Ralstonia solanacearum
Rickettsia rickettsii
Shigella flexneri
Staphylococcus aureus
Streptococcus agalactiae
Streptococcus equi
Streptococcus mutans
Streptococcus pneumoniae
Streptococcus thermophilus
Thermus thermophilus
Treponema pallidum
Yersinia enterocolitica
Yersinia pestis
Yersinia pseudotuberculosis

Supplementary Table 1: List of the 51 microbial species involved in the *small* reference database.

Supplementary Table 2 shows the number of strains involved in the *novel lineage* validation set of the *large database.*

reachable rank	test strains considered	test species	reference taxa represented
genus	584	421	69/126
family	338	146	42/114
order	183	147	18/54
class	143	111	9/52
phylum	97	81	4/16

Supplementary Table 2: **Number of strains involved in the novel-lineages study, per reachable rank.** The first column gives the number of strains considered for each reachable rank. The second column gives the number of species these strains originate from. The last column shows the number of taxa of this rank that they represent in the *large* reference database. For example, 584 strains coming from 421 species are reachable at the genus-level, and represent 69 genera of the 126 genera of the reference database.

2 Calibration procedure

In this section, we provide a detailed description of the procedure used to calibrate the rank-flexible read-classification procedure involved in the experimental studies described in Sections 4.2, 4.3 and 4.4 of the main text, based on the *medium* and *large* databases.

2.1 Definition of the calibration procedure

To calibrate the thresholds involved in the rank-flexible read classification procedure described in Section 2.3 of the main text, we proceed by means of an internal step of validation, that can be summarized as follows :

1. split the reference database into a *calibration database*, obtained by sampling one strain for each species represented by several strains, and a *learning database*, defined from the remaining strains.
2. build rank-specific models from the *learning database*.
3. optimize the thresholds involved in the reject option mechanism using the *calibration database*. This can be done by drawing fragments or simulating reads from calibration genomes, classifying them using the model built at step 2, and optimizing the performance of the model according to the thresholds, as described below.

The final model is then built from the entire reference database, and is ultimately used to make predictions, using the thresholds optimized in step 3.

As described in Section 2.3 of the main text, two types of thresholds enter the definition of the reject option mechanism :

- *credibility* threshold(s) on the maximum score of the linear model, aiming to reject unlikely predictions,
- *confidence* thresholds(s) on the difference of the two largest scores, aiming to reject ambiguous predictions.

These thresholds can be set globally or on a taxon-per-taxon basis, for a given rank-specific model, and can be further optimized for each rank. In this paper, we rely on the following procedure to optimize them:

1. we use the same value of the *credibility* threshold across ranks and taxa. In a rank-flexible context, this threshold can be chosen to reach a user-defined trade-off in terms of the proportions of (i) rejected predictions (i.e., predictions rejected at all ranks), (ii) predictions made at various ranks, (ii) correctness of predictions at various ranks. This is illustrated in Supplementary Figure 1 in a rank-flexible context, where we see in panel A that, as expected, the proportion of predictions made at the species level decreases as the value of the threshold increases, while the proportions of predictions made at upper ranks increase, as well as the reject rate. At the same time, we note from panel B (i) that the error rate decreases, and (ii) that while the rate of correct species-level predictions decreases as well (reflecting the fact that fewer predictions are made at the species level), the rate of predictions made correctly either at the species level or upper ranks remains steady, when the threshold is not too high (i.e., when the reject rate is not to high either). This therefore indicates that this procedure allows to reduce the proportion of misclassified sequences, at the price of unclassified sequences and sequences (correctly) classified above the species level. This global credibility threshold is set to 0 in the experimental studies of Sections 4.2, 4.3 and 4.4 of the main text, a value leading to a reasonable trade-off in terms of error and reject rate.

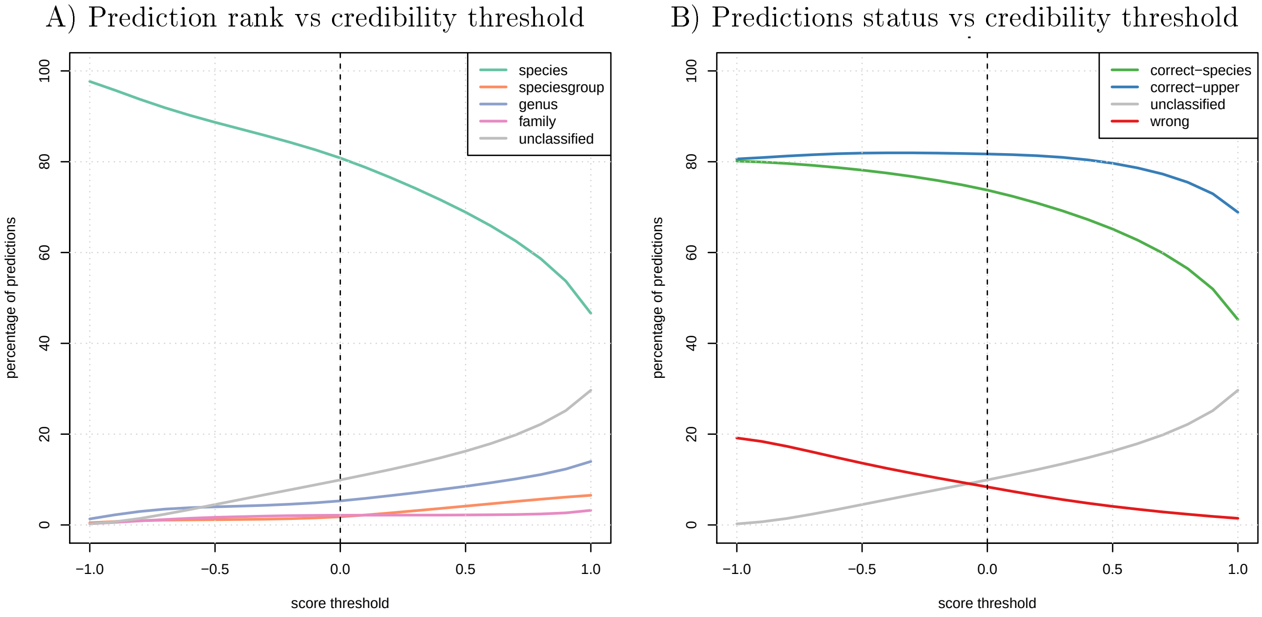
2. we define the *confidence* scores on a taxon-by-taxon basis. Indeed, although the same kind of analysis can be done, as shown in Supplementary Figure 2, we note that species exhibit very different behaviors regarding prediction ambiguity. This is not surprising and simply reflects the fact that the level of genomic proximity can vary across the taxonomy (e.g., it can be more or less important across genera, and even among groups of species of the same genus). A single threshold is therefore unlikely to be optimal for all taxa at once, and has indeed the effect of degrading the performance obtained for some taxa (e.g., rejecting predictions made at the species level that were actually correct), or not rejecting enough ambiguous predictions for others. Supplementary Figure 3 illustrates the four types of behaviors observed while analyzing the effect of this second level of rejection on a species-by-species basis :

    - species for which prediction ambiguity is not an issue. This is for instance the case of the two species shown on the left-hand side, differing by their level of classification performance. In such cases, indeed, confidence-based rejection does not allow to reduce the error rate, and actually has the detrimental effect of deferring to the genus level predictions that were made correctly at the species-level. In both cases, the confidence threshold is set to zero : a prediction involving one of these species made with a positive score (or more generally, higher than the credibility threshold defined in step 1) is always accepted.
    - species for which ambiguity is an issue, and for which this second level of prediction indeed allows to reduce the rate of erroneous predictions. This is for instance the case for the two species shown on the right-hand side, differing by the level of performance that can be attained as we vary the value of the threshold. In the case of *Brucella suis* shown at the top, indeed, this reject procedure allows to reach a target level of classification performance, defined in terms of upper recall (proportion of predictions made correctly at the species level or at a upper rank) represented by the orange curve. This target performance is defined as the average value of the upper recall, taken across species, when this confidence-based rejection was not carried out (shown as the solid horizontal gray line in Supplementary Figure 3)[1]. For such species, the confidence threshold is set to the minimum value allowing to reach the target performance. For *B. suis*, for instance, it is set to 1.5 : a *B. suis* prediction is accepted provided that (i) it is made with a positive score (or more generally, higher than the credibility threshold defined in step 1) and (ii) it is greater than the second largest score of the linear model by 1.5. In the case of *Mycoplasma mycoides* shown at the bottom, on the other hand, although this second level of rejection allows to decrease the error rate, it does not allow to reach the target performance. For such species, the confidence threshold is defined as the smallest value allowing to reach the minimum attainable error rate, up to a tolerance parameter set here to 1 point (i.e., 1% in absolute value).

Last but not least, we note from Supplementary Figure 2, however, that this second level of rejection is essentially relevant at the species-level. Indeed, we note from the left-hand side figures that when this procedure is applied at the species-level (top) or genus-level (bottom), the proportion of predictions made at this rank decreases as the threshold increases, at the benefit of prediction made at upper ranks (and actually mainly at the genus-level in the former case), as expected. We note, however, that while the procedure has a positive impact at the species-level, which can be seen by a decrease of the error rate at the benefit of the rate of correct predictions made at genus or family levels, the error rate does not decrease when it was applied at the genus-level. This therefore means that this second level of rejection is unnecessary and actually to be avoided at the genus level, as it degrades the level of resolution of the prediction, while not reducing its error rate. As a result, we simply apply this procedure at the species-level in the experimental studies shown in Sections 4.2, 4.3 and 4.4 of the main text.

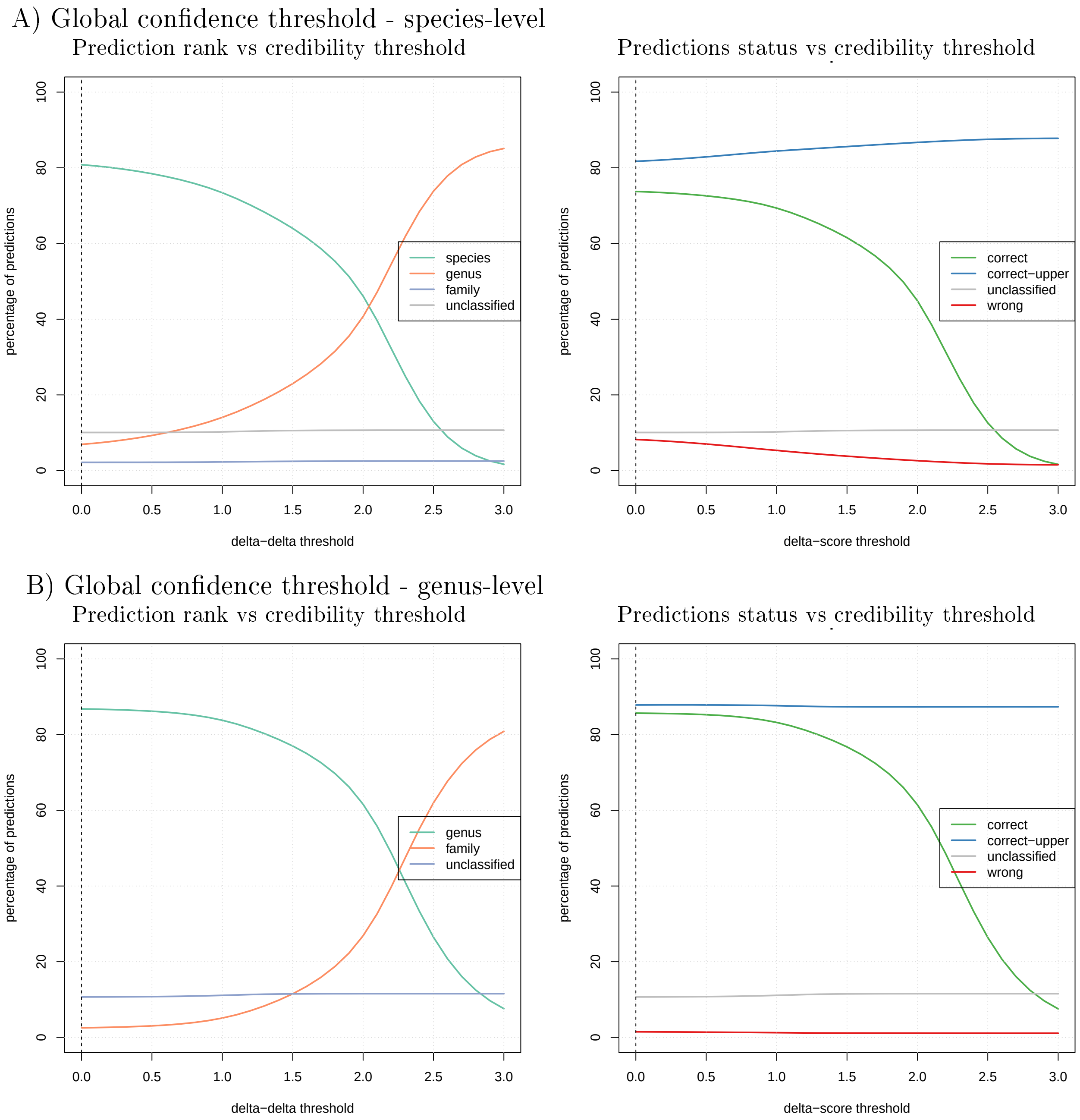
[1] The next section illustrates the trade-off obtained if we consider instead the median upper level recall (shown as the dashed horizontal grey line) as target performance.

As an ending remark we note that the confidence calibration can only be achieved for species represented by at least two strains in the reference database. For the *large* database in particular, only 110 out of the 774 species can be calibrated, and the confidence threshold is set to zero for the 664 remaining ones. This point constitutes an obvious limitation of the method, which will hopefully fade as the amount of sequenced microorganisms grows.

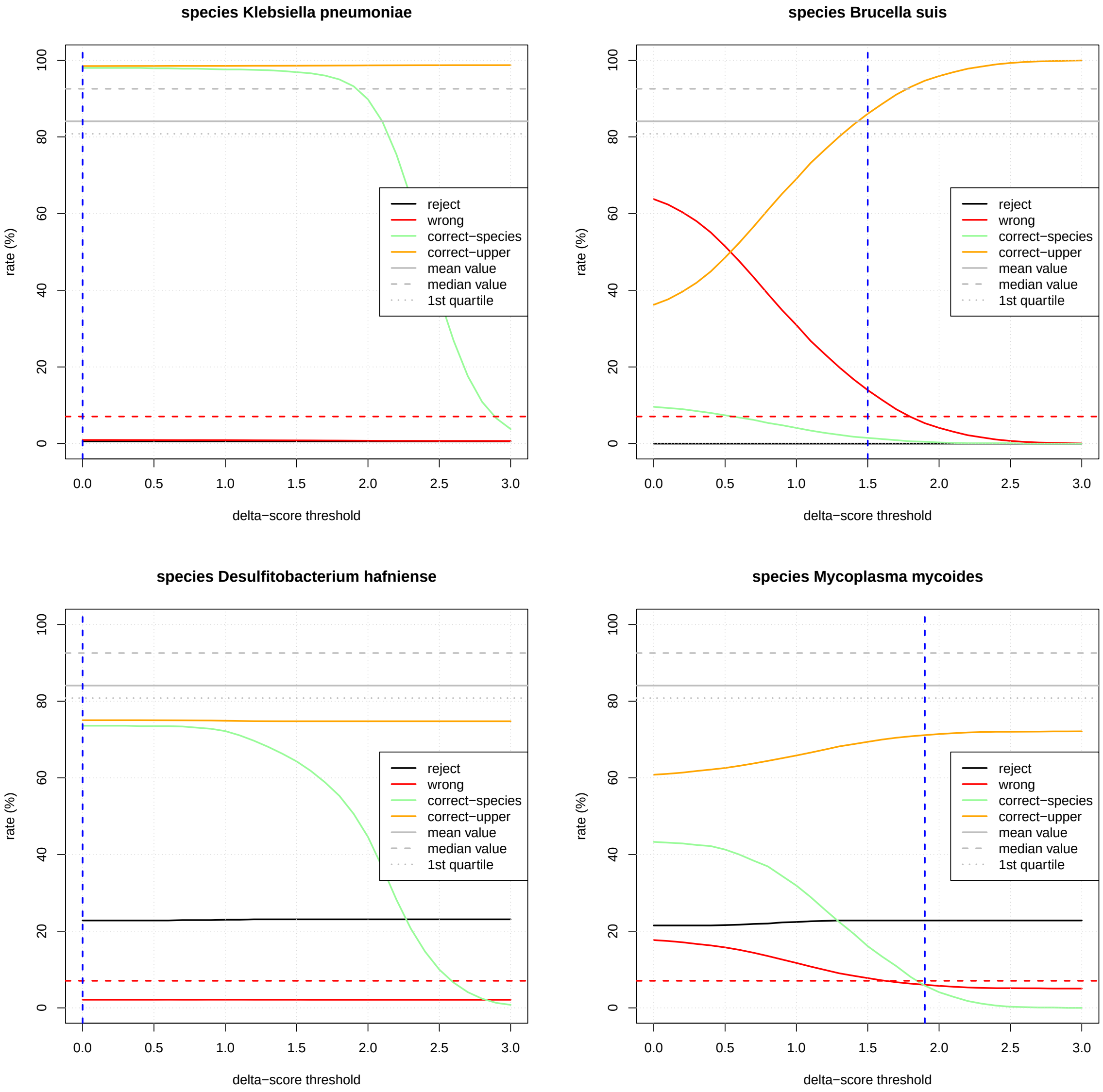

A) Prediction rank vs credibility threshold

B) Predictions status vs credibility threshold

Supplementary Figure 1: **Calibration procedure: impact of a global credibility threshold.** A global credibility threshold taken in $[-1, 1]$ is applied to fragments drawn from the calibration genomes of the *medium* database, in the rank-flexible setting described in Section 2.3 of the main text. Left: evolution of the prediction rank, defined in terms of the proportions of rejected predictions and of predictions made at various ranks. As expected, fewer predictions are made at lower ranks as the threshold increases, at the benefit of rejected predictions and predictions made at upper rank. Right: evolution of the prediction status, defined in terms of the proportions of predictions that are rejected (grey), erroneous (red), correct at the species level (green) and correct at the species level or at a upper rank (blue). This procedure allows to reduce the proportion of misclassified sequences (red), at the price of unclassified sequences (grey) and sequences correctly classified above the species level (reflected by the fact that the green curve decreases while the blue one remains steady).



Supplementary Figure 2: **Calibration procedure: impact of a global confidence threshold.** A global confidence threshold taken in $[0, 3]$ is applied to fragments drawn from the calibration genomes of the *medium* database, in the rank-flexible setting described in Section 2.3 of the main text, for the models defined at the species (top) and genus (bottom) levels. Left-hand side figures: evolution of the prediction rank as the threshold increases. As expected, a larger proportion of predictions are made at the next upper rank (i.e., genus or family) when the threshold increases. Right-hand side figures : evolution of the prediction status as the threshold increases. When applied at the species-level, this procedure has the positive impact of reducing the error rate (red) at the benefit of the prediction made correctly at an upper rank (genus in this case, as indicated by the corresponding left-hand side figure). This therefore indicates that the procedure successfully manages the ambiguity issue by deferring to the genus level uncertain predictions. This is not the case at the genus level where more predictions are made at the family level, without further decreasing the error rate. This therefore indicates that these predictions are correctly made at the genus-level, hence that this confidence-based rejection has the sole detrimental effect of degrading the level of resolution of the prediction.



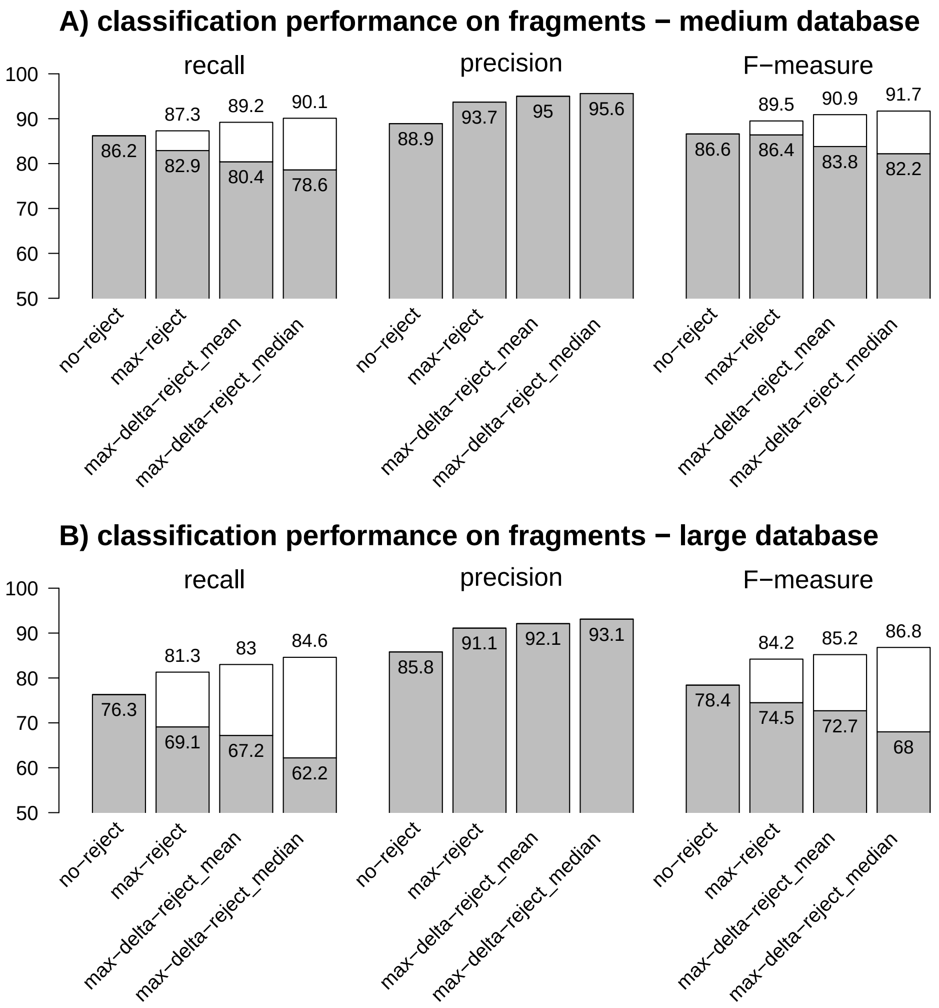
Supplementary Figure 3: **Calibration procedure: taxon-by-taxon definition of the confidence threshold.** Four types of behaviors are observed while analyzing the effect of the confidence-based rejection on a species-by-species basis. Left-hand side: species for which prediction ambiguity is not an issue, hence for which confidence-based rejection does not allow to reduce the error rate. Some of these species show a good classification performance, and in particular higher than a pre-defined target level of performance (shown in gray), as *Klebsiella pneumoniae* shown on the top. Other show a lesser level of performance, like *Desulfitobacterium hafniense* shown on the bottom. In both cases, the confidence threshold is set to zero (meaning that it has not effect) for such species, as shown by the vertical blue lines. Right-hand side: species for which ambiguity is an issue and for which this confidence-based rejection has a positive impact. This is for instance the case of *Brucella suis* shown at the top. For this species, indeed, sufficiently increasing the threshold allows to reach a target performance, defined as the mean upper recall (proportion of prediction made correctly at the species-level or an upper rank), taken across species, when this confidence-based rejection is not applied (show as the solid horizontal gray line). For such species, the confidence threshold is set as the minimum value allowing to reach the target performance (1.5 for *B. suis*). For other species like *Mycoplasma mycoides* shown at the bottom, on the other hand, although this second level of rejection allows to decrease the error rate, it does not allow to reach the target performance. In such cases, the confidence threshold is defined as the smallest value allowing to reach the minimum attainable error rate, up to a tolerance parameter set here to 1 point.

2.2 Impact of the calibration

Figure 4 illustrates the trade-offs that can be achieved by means of the calibration procedure, considering the four following variants:

1. `no-reject` : the rank-specific strategy defined at the species-level, with no rejection mechanism.
2. `max-reject` : the rank-flexible strategy based only on a global credibility threshold (set to zero).
3. `max-delta-reject_mean` : the rank-flexible strategy based on the combination of a global credibility threshold (set to zero) and species-specific confidence thresholds, defined according to the procedure described in the previous section, in order to reach a target performance defined as the average upper recall (across species).
4. `max-delta-reject_mean` : the same strategy, where the target performance considered to optimize the species-specific confidence thresholds is defined as the median upper recall (across species).

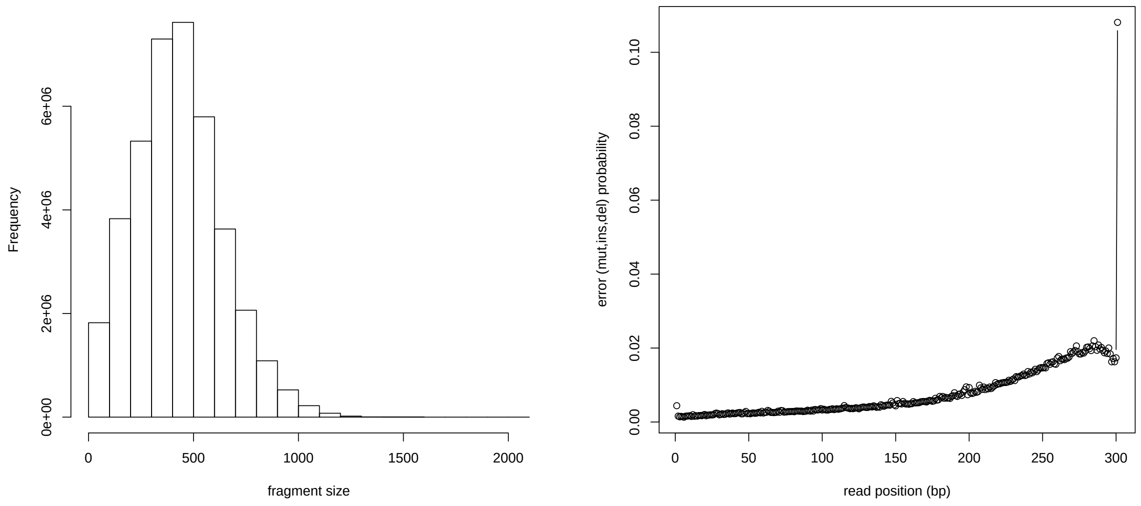
The first and third strategies were considered in the main text. As could be expected, the average species-level recall gradually decreases from the first to the fourth strategy, while the upper recall gradually increases, together with the species-level precision.



Supplementary Figure 4: **Illustration of the trade-offs that can be achieved by the calibration procedure.** Top : *medium* reference database ; Bottom : *large* reference database. This figure shows the classification performance measured on genomic fragments in terms of average species-level recall, precision and F-measure, for the four classification strategies defined in Supplementary Section 2.2

3 Simulation of Illumina MiSeq reads

We developed a MiSeq reads simulator to reproduce the sequencing error profile observed on actual MiSeq runs carried out internally. The MiSeq reads simulator proceeds in two steps: i) draw fragments from a multi-fasta file, and ii) get 300 bp paired-end reads at both ends of each fragment, to which we add sequencing errors. For the first step, we need to set-up the abundance of each individual genome in the final metagenomic sample, and the distribution of fragment length. Supplementary Figure 5 (left) presents the fragment length distribution estimated from MiSeq sequencing runs with the V3 chemistry. Each fragment length is estimated from the alignment with BWA-MEM of the raw sequencing reads on the contigs obtained after de novo assembly. For the second step, we need to define the mutation rate distribution (probability of mutation at each position in the read), the substitution matrix probability (the cell i, j of the substitution matrix is the substitution probability from base $i = \{A, T, C, G\}$ to base $j = \{A, T, C, G\}$), the insertion rate distribution (probability of insertion at each position in the read), the distribution of the insert size, the deletion rate distribution (probability of deletion at each position in the read), and the distribution of the deletion size. Supplementary Figure 5 (right) presents the overall error probability (i.e., of encountering a mutation, an insertion or a deletion) along the read. These error rate distributions are also estimated from real V3 MiSeq sequencing runs, by remapping of reads against the contigs obtained after de-novo alignment.



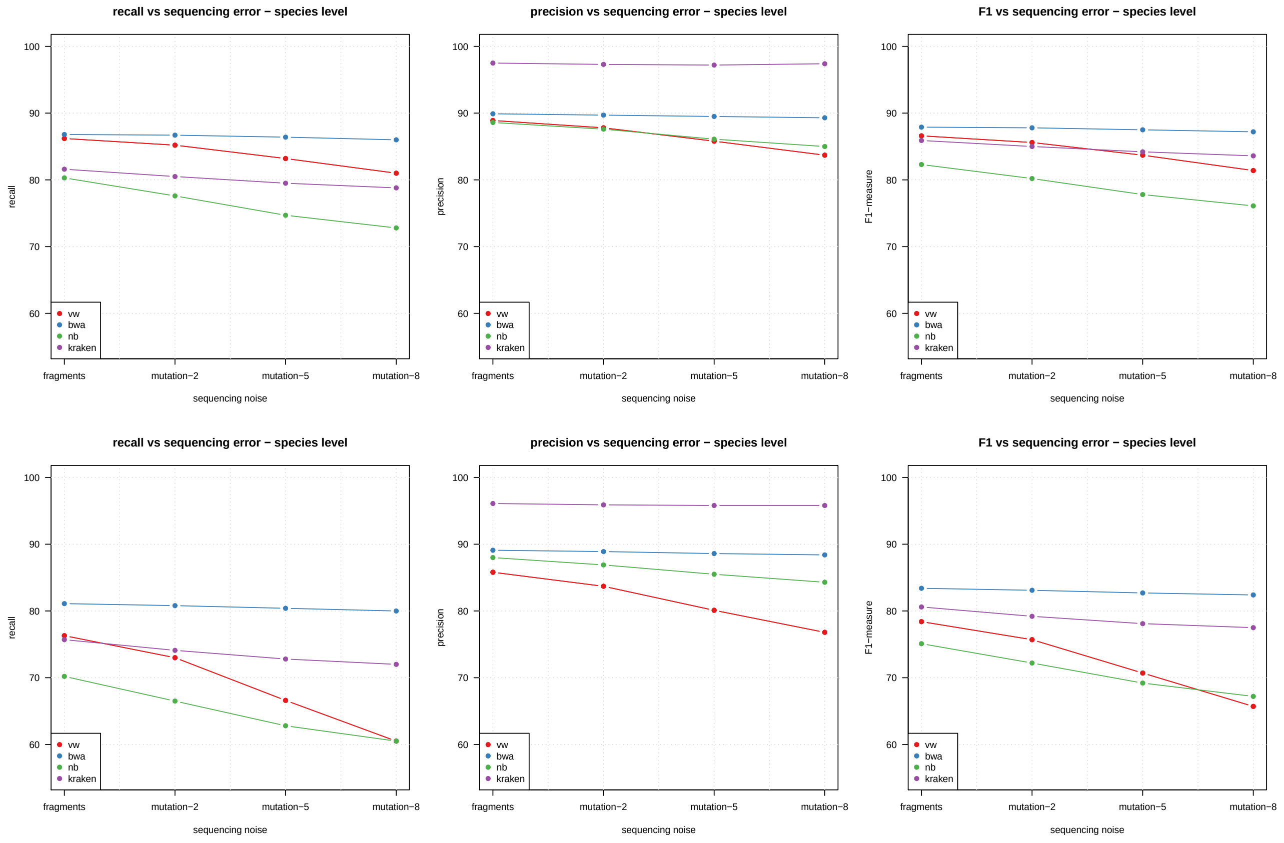
Supplementary Figure 5: **Simulation of MiSeq reads.** Top: fragment size (bp) distribution. Right: overall error probability (i.e., of encountering a mutation, an insertion or a deletion) along the read. Both distributions are estimated from MiSeq V3 chemistry runs, and used to simulate MiSeq test reads.

4 Impact of high-level of sequencing noise

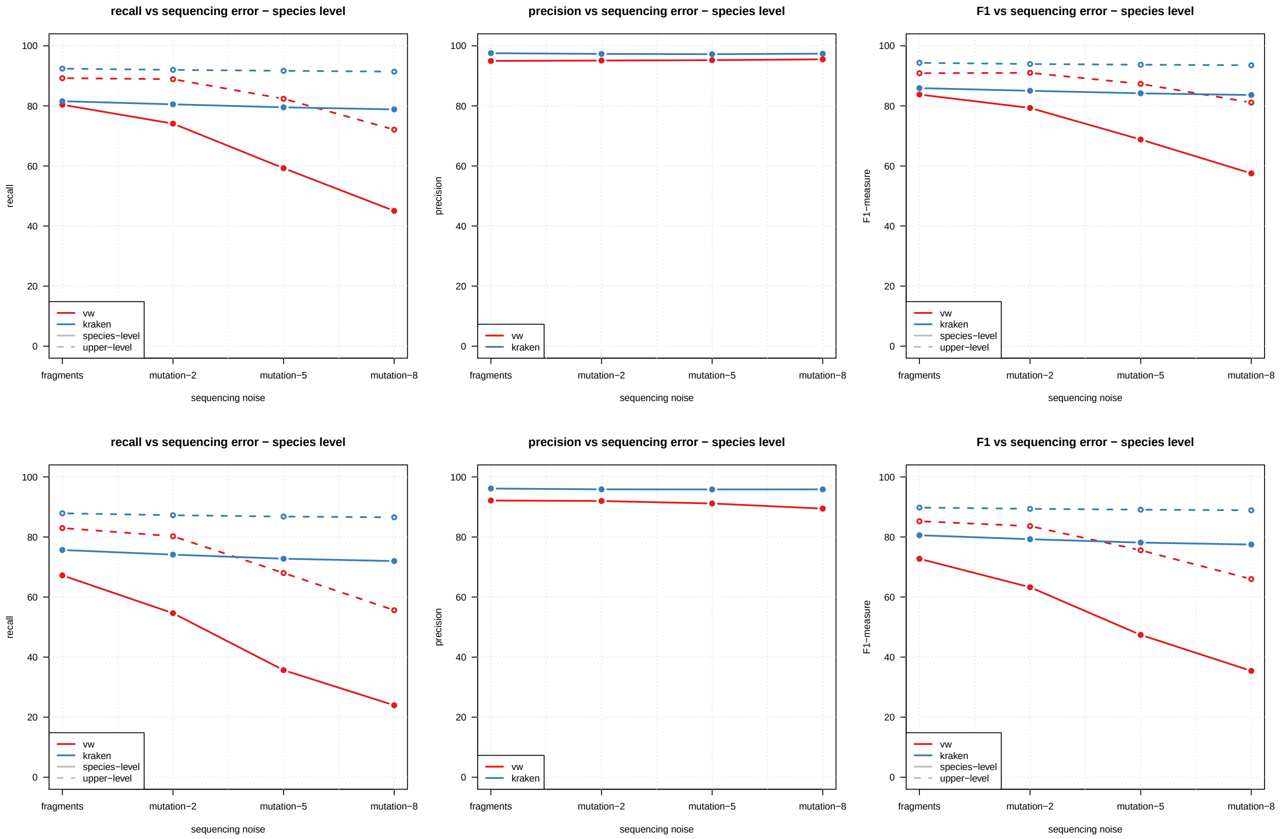
In this section, we analyze the robustness of read classification procedures to high levels of sequencing noise. For this purpose, we consider the mutation error model of Grinder, and generate additional test datasets comprising a median mutation rate of 5% and 8%, a level of noise much higher than can be expected from current next-generation sequencing technologies. Supplementary Figures 6 and 7 respectively show the results obtained by rank-specific and rank-flexible procedures.

We can note from Supplementary Figure 6 that BWA and Kraken are much less impacted than VW and NB by these levels of sequencing noise, and that the impact is particularly severe for VW on the *large* reference database. While both VW and Kraken are based on k-mers, they are based on completely different algorithms. Kraken indeed bases its prediction on the detection of (at least one) k-mer(s) of length 31, while VW relies on the overall k-mer profile of the sequence, thereby taking into account every k-mer observed. We postulate that the robustness of Kraken to these levels of sequencing noise lies in the fact that mutations accumulate in the end of the reads with this error model (in agreement to what is observed in actual NGS data). The beginning of the sequence may therefore not be altered, which may allow Kraken to detect a sufficient number of specific k-mers. On the other hand, because it considers the entire k-mer profile, VW suffers from every modification of the sequence, and the results shown here suggest that a mutation rate of 5% or more is sufficient to severely disrupt the k-mer profile.

As shown in Supplementary Figure 7, however, the rank-flexible VW strategy manages to maintain a level of precision comparable to that obtained from fragments. This comes at the expense of a drastic decrease in terms of recall (either at the species or upper level). This is actually due to a a serious increase in the proportion of rejected predictions, especially with the *large* reference database (from 8% on fragments to 13, 26 and 38% in average per species, for mutation rates of 2, 5 and 8% respectively), which tends to confirm the above intuition that the k-mer profile is strongly disrupted by mutation rates of 5 and 8%.



Supplementary Figure 6: **Impact of high levels of mutation sequencing noise - rank-specific procedures.** Top : *medium* database ; Bottom : *large* database. These figure show the evolution of the species-level recall (left), precision (middle) and F-measure (right) as the mutation rate of the sequencing error model increases, for the rank-specific strategies considered in the main text.

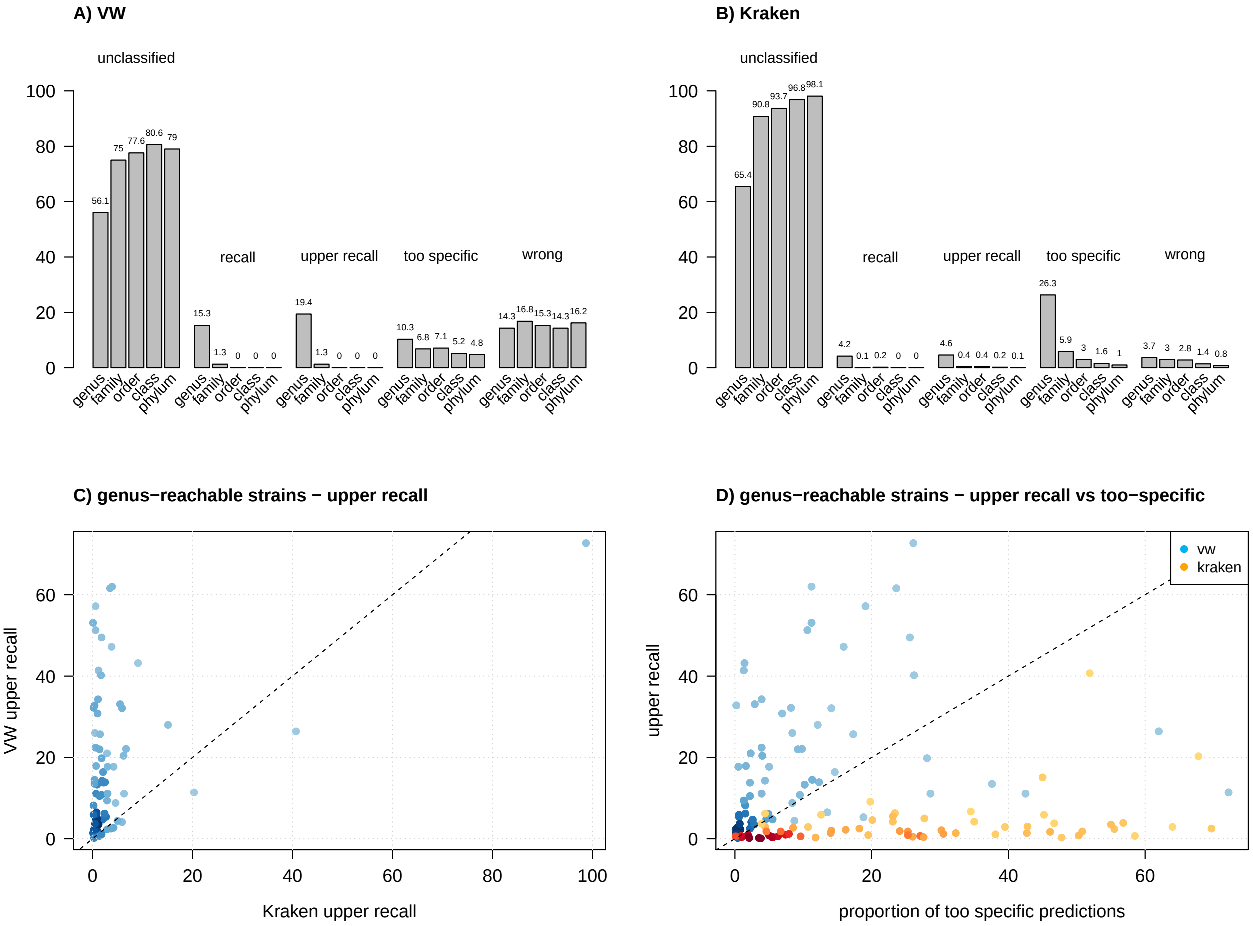


Supplementary Figure 7: **Impact of high levels of mutation sequencing noise - rank-flexible procedures.** Top : *medium* database ; Bottom : *large* database. These figures show the evolution of the species-level recall (left), precision (middle) and F-measure (right), and their upper level counterpart, as the mutation rate of the sequencing error model increases, for the VW rank-flexible strategy and Kraken.

5 Detailed results of novel-lineages study

In this section we provide further details related to the novel-lineages study :

- Supplementary Figure 8 shows the results obtained for all reachable ranks considered. Moreover, panel D illustrates the different trade-offs achieved by VW (in its rank-flexible setting) and Kraken in terms of upper recall and proportion of too specific predictions.
- Supplementary Table 3 provides the values of the performance indicators obtained on strains reachable at the genus level, on a genus-per-genus basis.



Supplementary Figure 8: **VW and Kraken performances on novel lineages**. Panels A and B are the same as those shown in Figure 5 of the main text, with the addition of the results obtained at the class and phylum levels. Panel C is identical to the one shown in Figure 5. Panel D plots the upper recall versus the proportion of too specific predictions obtained for each reachable genus, for VW (blue) and Kraken (orange).

	VW					Kraken				
genus	reject	exact	above	below	wrong	reject	exact	above	below	wrong
Acidovorax	63.8	11.1	0	3.9	21.3	65.4	5.2	1.1	23.4	4.9
Acinetobacter	68.4	8.7	2.1	9.5	11.4	72	1.7	0.1	25.3	0.8
Actinobacillus	54.5	8.3	8.1	14.6	14.4	58.5	1.5	0.6	30.2	9.3
Aeromonas	62.6	17.7	0	5	14.7	57.9	3.7	0.5	35	2.9
Agrobacterium	60.5	1.3	16.4	0.5	21.4	79.5	1.9	1.1	4.4	13.2
Arthrobacter	76.1	5.4	0	0.6	17.9	84	2.2	0.5	8.5	4.7
Bacillus	59.7	20.9	5.1	8.4	6	72.6	0.4	0.1	26	0.9
Bacteroides	65.3	13.3	0	10.2	11.2	73.4	0.9	0	25.3	0.3
Bartonella	67.7	21	0	2.3	9	57.3	2.8	0.1	39.5	0.3
Bifidobacterium	64.6	14.3	0	4.4	16.7	73.3	1.8	0.1	24.1	0.7
Bordetella	71.4	4.1	0	2.4	22.1	76.5	5.4	0.5	12.6	4.9
Borrelia	30.7	47.8	5.3	11.2	5.1	13.2	0.1	0	86.6	0.1
Brucella	0	72.7	0	26.1	1.2	0.7	98.7	0	0.6	0
Burkholderia	22	58.1	3.9	11.2	4.8	38.6	3.7	0.2	56.8	0.8
Campylobacter	62.9	18.8	3.6	3.9	10.7	92.4	0.5	0.1	6.3	0.7
Chlamydia	88	0.4	1.5	0.1	10	93.1	0.9	0.1	1.9	4
Chlorobium	80	1.4	0	0.2	18.4	95.4	0.1	0	3.8	0.6
Clostridium	55.4	30.8	0	6.9	6.9	91	1	0	7.4	0.6
Ruminiclostridium	73.1	0.1	0.6	0.1	26.2	82.5	1	0.4	14	2.1
Corynebacterium	70.5	8.8	0	8.4	12.2	67.7	3.9	0.7	20.1	7.6
Cupriavidus	58.2	6.8	11.1	1.6	22.2	97.2	0.5	0.2	0.9	1.2
Desulfitobacterium	86	0.1	2.2	0.1	11.6	50.9	0.3	0	47.8	0.9
Desulfotomaculum	39.2	6.8	4.3	42.5	7.1	70.6	0.5	0.2	27.1	1.5
Desulfovibrio	61.1	13.9	0	12.3	12.7	23.8	1.4	1.1	69.7	3.9
Enterobacter	19.3	28.6	20.9	25.6	5.6	62.8	1.3	0.5	6.7	28.7
Enterococcus	56.1	2.5	0	2.2	39.1	82.7	2.6	1.1	3.9	9.6
Erwinia	55.5	0.8	32	0.2	11.4	96.7	0.2	0.2	1	1.9
Eubacterium	79.8	0.2	0	0.4	19.7	87.2	0.2	0.1	11.8	0.7
Exiguobacterium	70	8.2	0	1.5	20.3	95.9	0.2	0	3.5	0.4
Flavobacterium	82.9	2.4	0	0.3	14.3	62.2	0.1	0.3	27.6	9.8
Glaciecola	61.7	4.8	0.5	18.8	14.3	71.1	1.6	0.9	18.2	8.2
Gordonia	55.8	4.8	0	5.5	33.9	39	0.5	0.2	58.5	1.8
Haemophilus	41.9	9.3	4.2	37.6	7	93	0.3	0.1	5.6	1
Helicobacter	65	20.4	0	4	10.6	56.8	4.1	2.1	4.4	32.7
Klebsiella	31	0.7	40.7	1.3	26.3	89.8	1	0.2	7.9	1.1
Lactobacillus	76.9	9.4	0	1.3	12.4	84.4	2.7	0.2	10.7	1.9
Leuconostoc	77	4.7	0	2.6	15.7	82.4	1.8	0.2	14.1	1.5
Listeria	71.5	4.8	0.1	4.6	19	76.4	0.7	0.2	19.5	3.1
Marinobacter	69.2	3.8	0.6	8.7	17.7	61.8	4.3	0.7	27.7	5.5
Mesorhizobium	63	10.5	0	2.2	24.2	54	1.1	0.3	42.7	1.9
Methylobacterium	53.7	22.1	0	9.8	14.4	55.8	6.1	0.6	34.5	3
Mycobacterium	46	32.2	0	8.2	13.6	97.3	0.1	0.1	2.1	0.4
Mycoplasma	71.6	2.7	0	0.3	25.4	70.8	4	0.2	23.1	1.8
Neisseria	61.7	14.5	0	11.3	12.5	93.5	0.3	0.1	5.4	0.8
Paenibacillus	75.7	6.2	0	1.5	16.7	37	1.3	1.1	55.5	5.2
Pantoea	27.8	15.3	36	10.6	10.4	88.9	0.5	0.1	9.6	0.9
Prevotella	80.2	3.5	0	2.8	13.4	66.4	0.9	0.3	30.5	1.8
Propionibacterium	71.5	6.1	0	4.9	17.5	92.5	0.6	0.2	5	1.7
Pseudoalteromonas	80.1	2.3	0	0.8	16.8	32.1	2.6	0.3	64	1
Pseudomonas	28.9	40.2	0	26.2	4.7	46.7	1.2	0.5	46.1	5.6
Rhizobium	41.4	10.6	9.2	28.1	10.6	43	1.3	0.5	50.8	4.3
Rhodococcus	38.6	20.1	5.6	17.3	18.3	17.3	1.5	0	81.1	0
Rickettsia	16.9	57.2	0	19.1	6.8	73.6	0.5	0.1	0.1	25.7
Ruminococcus	75.5	1.2	2.5	0.7	20.2	94.7	0.2	0.3	2	2.8
Serratia	39.7	10.6	32.6	1.4	15.7	62.6	7.5	1.6	19.8	8.5
Shewanella	54.9	34.3	0	3.9	6.9	60.2	1	0.1	38.1	0.6
Sphingobium	45.1	11.1	0	28.6	15.3	51.8	2.8	0.2	42.8	2.4
Spiroplasma	81.4	1.1	0	0.3	17.1	92.3	1.6	0.2	4.6	1.3
Staphylococcus	70.5	13.8	0	2.2	13.5	80.6	2	0.2	16.2	1.1
Streptococcus	47.9	32.1	0	14.1	5.9	48.6	5.6	0.3	45.2	0.3
Streptomyces	28.8	47.2	0	15.9	8.2	48.3	3.6	0.2	46.7	1.2
Thermoanaerobacter	9.1	26.4	0	62	2.5	6.7	40.6	0.1	51.9	0.7
Thermus	10.3	61.6	0	23.6	4.5	41.2	3.5	0	55	0.3
Thioalkalivibrio	62	6.5	0	13.5	18	46.9	0.4	0.4	50.3	2
Treponema	73.8	3.1	2.8	0.6	19.7	97.2	0.2	0	2.1	0.5
Vibrio	58.8	22	0	9.2	10	65.6	1.3	0.1	32.3	0.7
Wolbachia	6.4	11.1	0.3	72.2	10	11.9	20.3	0	67.8	0
Xanthomonas	41	28	0	12.1	18.9	38	13.5	1.6	45	1.8
Yersinia	53.7	5.4	27.7	2.9	10.3	66.4	4.8	0.7	23.1	5

Supplementary Table 3: **Detailed results obtained on stains reachable at the genus level.** This table provides the values of the performance indicators obtained on strains reachable at the genus level, on a genus-per-genus basis. Reject: proportion of rejected predictions. Exact : proportion of fragments assigned to the appropriate genus (i.e., recall). Above : proportion of fragments assigned to an ancestor taxon (which, together with the Correct statistic, defines the upper recall). Below : proportion of fragments assigned to a sibling species (i.e., to a species of the appropriate genus), which corresponds to too specific predictions. Wrong : proportion of erroneous predictions.

6 Detailed results of time evaluation

Supplementary Table 4 provides the time taken to process each test dataset (fragments and reads) by VW, Kraken and BWA on each reference database. Note that the MiSeq dataset is slightly larger than the 3 other ones since reads can be longer than 200 bp. This explains why the time needed to process it is systematically longer. As a result, the analysis given in Section 4.6 of the main text is based on the 3 other test datasets, which involve reads of 200 bp exactly.

	medium database				*large* database			
	fragments	mutation-2	Balzer	MiSeq	fragments	mutation-2	Balzer	MiSeq
VW	6	4.2	4.4	9.1	9.1	8.8	8.8	12.2
Kraken	2.7	3	2.9	3.7	3.1	3.8	3.3	4.4
BWA	23.1	52	31.4	63.2	21.6	48	39.1	63.5

Supplementary Table 4: **Time evaluation.** Time measured in minutes to process each test dataset (fragments and reads) by VW, Kraken and BWA on each reference database.