A guided tour into genomic contamination detection

Luc Cornet[1,2] and Denis Baurain[2]

[1] BCCM/IHEM, Mycology and Aerobiology, Sciensano, Bruxelles, Belgium
[2] InBioS–PhytoSYSTEMS, Eukaryotic Phylogenomics, University of Liège, Liège, Belgium
`luc.cornet@uliege.be, denis.baurain@uliege.be`

Nowadays, genomes constitute the basis of many research endeavors. This is especially true since the decrease of sequencing cost has led to an explosion of the number of genomes in public repositories. Estimating the amount of contamination, i.e. the inclusion of unwanted DNA in genomic materials (**Fig. 1**), of this deluge of data has become a field in itself, with numerous algorithms now available and an increasing rate of publication over the years. As newly released tools do not simply replace older ones, but have their own scope, it becomes difficult for scientists to efficiently determine which tool to use in their study. Recently, we have published an overview of the main characteristics and applicability of 18 algorithms dedicated to the estimation of genomic contamination (**Fig. 2**). For instance, the conceptual differences between database-free tools and those associated with a reference database have an effect on detection sensitivity, as do the difference between genome-wide and marker-based methods [1]. Beyond this typology of tools, we present here a new analysis designed to compare algorithms on a large simulated dataset derived from empirical data. This protocol, dubbed `CRACOT` for *CRitical Assessment of COntamination detection at multiple Taxonomic levels*, reveals both under- and over-detection by even the most commonly used algorithms, with simulated contamination events ranging from inter-phylum to inter-species.

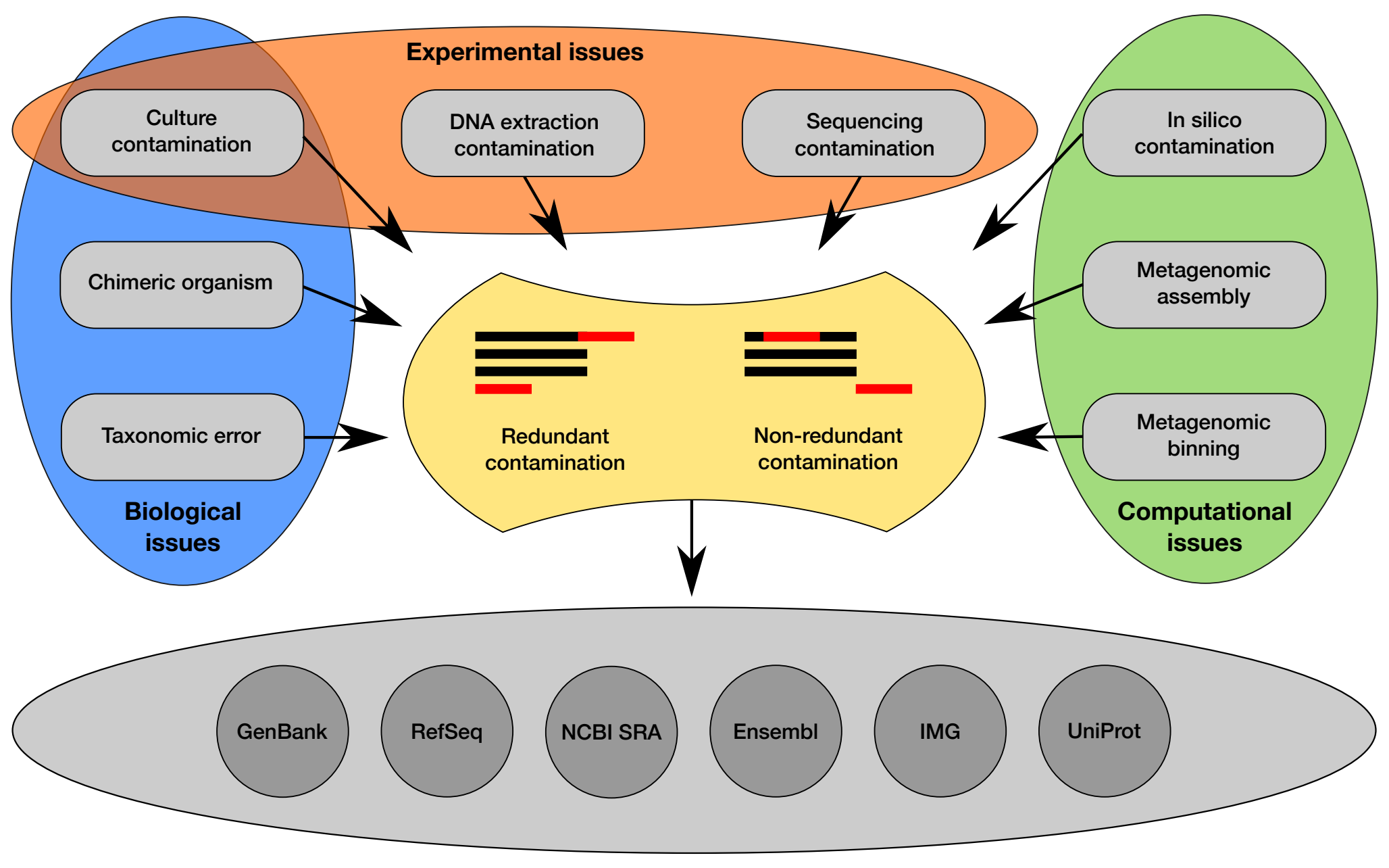

Figure 1: ***Sources and types of genomic contamination.*** *The contamination of "pure" cultures can be due to both experimental and biological causes.* ***Redundant*** *contamination occurs when a genomic segment is present multiple times in a genome.* ***Non-redundant*** *contamination occurs when a genomic region of the expected organism is replaced by the corresponding region of a foreign organism. An extra DNA segment, not part of the expected organism but belonging to a contaminant, is also considered as a non-redundant contamination.*

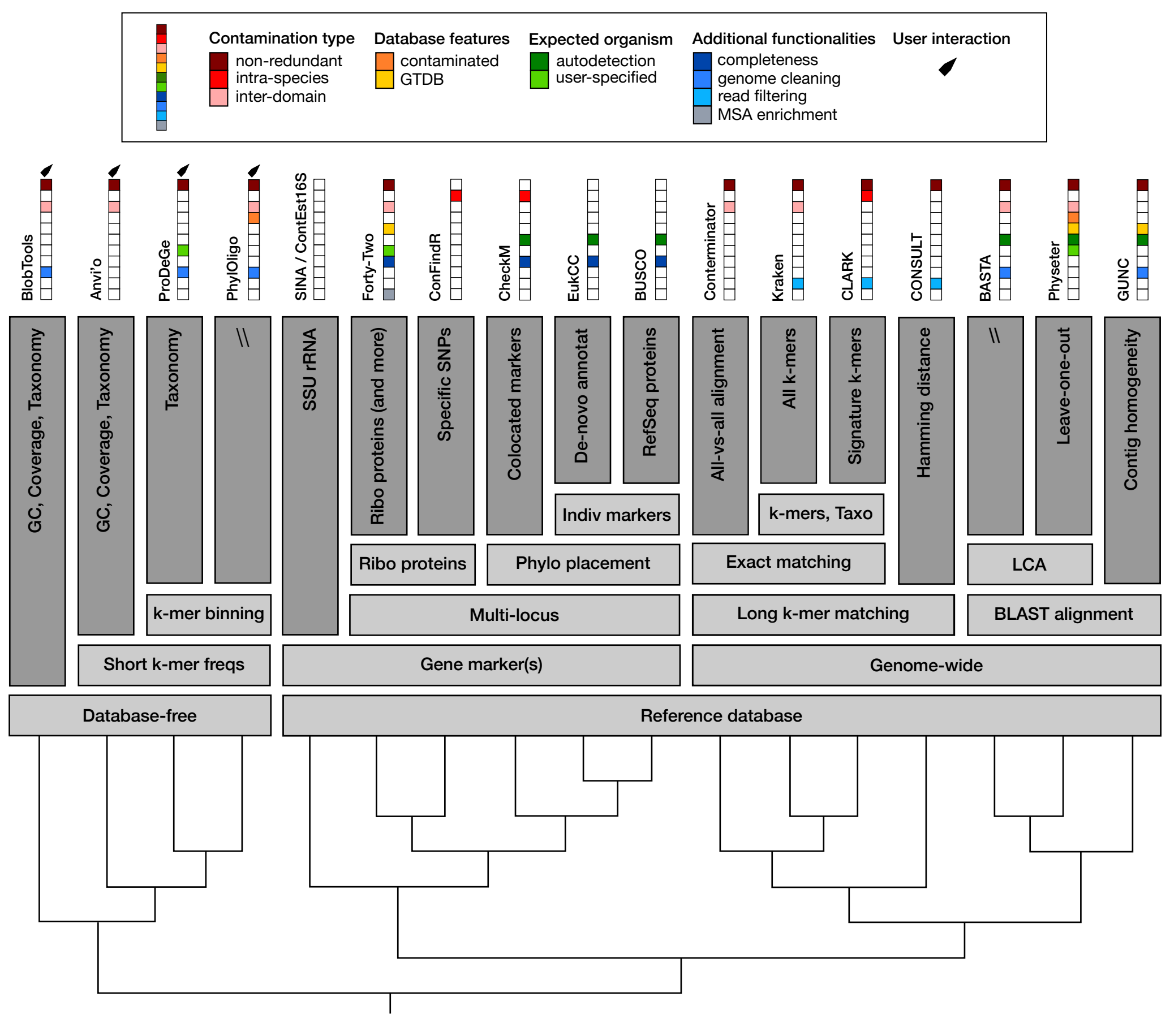

Figure 2: ***Overview of available algorithms for the evaluation of genomic contamination. Non-redundant*** *means that the software can detect contaminant genes without equivalent in the surveyed genome.* ***Intra-species*** *means that the algorithm can detect contamination at the species level.* ***Inter-domain*** *means that the algorithm can detect prokaryotic and eukaryotic contamination simultaneously.* ***Database features*** *show that the algorithm can use the GTDB Taxonomy and/or a moderately contaminated reference database.* ***Expected organism*** *indicates whether the algorithm can detect the main organism by itself and/or if the user can specify it.*

Methods

The `CRACOT` flowchart is summarized in **Fig. 3**. Briefly, 816 high-quality Firmicutes genomes, either belonging to Lactobacillus or Clostridia class, were collected as input. These were selected on two criteria: 1) number of contigs ≤ 5 (metric of assembly quality) and 2) clade separation score ≤ 0.01 (`GUNC` [2] metric of taxonomic chimerism). The median `CheckM` [3] contamination value of these 816 contigs was 0.24%. The genome pairing step creates random pairwise associations at multiple taxonomic levels. At a given level, the two genomes must be different at the lower taxonomic level (i.e., at the phylum level, Firmicutes, if one genome belongs to Lactobacillus, the other genome belongs to Clostridia). Hence, 100 chimeric genomes were created for each level. As plasmids carry sequences involved in HGT, these were removed from all the genomes with `PlasmidPicker` (`https://github.com/haradama/PlasmidPicker`). After protein prediction, orthologous genes (OGs) were inferred with `OrthoFinder` [4]. Three types of chimerism were created at the final step of `CRACOT` – **Replacements**, **Duplications** and **Singletons** (**Table 1**). Common OGs were used as pools for gene replacement and duplication events in the *Master* genomes, by inserting sequences from the *Contaminant* genomes. Gene replacements were performed in place within the chromosome whereas duplicated and singleton genes were concatenated at the end of the last contig.

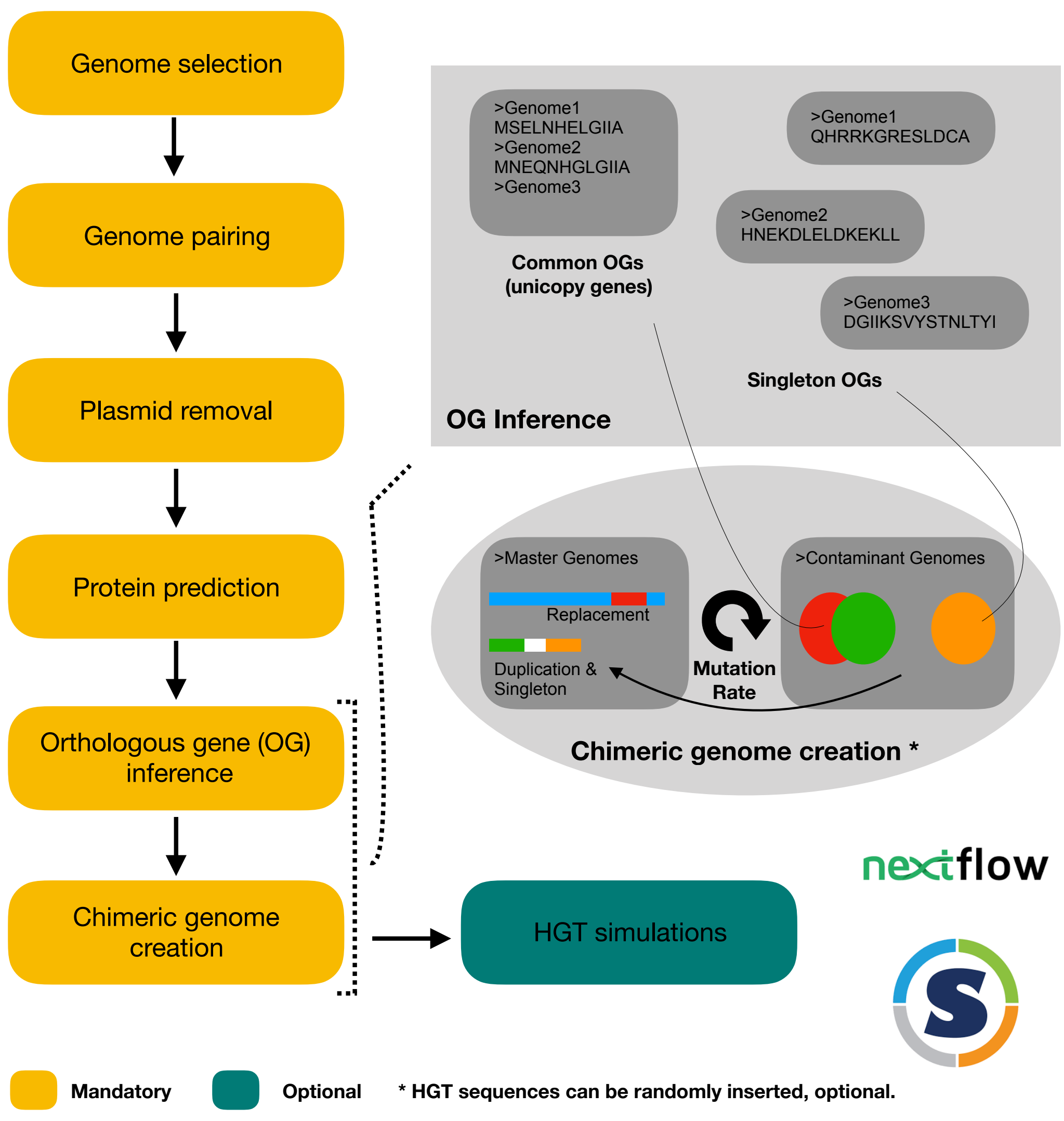

Figure 3: **`CRACOT`** ***flowchart.*** *The protocol was designed to simulate realistic contamination events at multiple taxonomic levels.*

Table 1: ***Specified and actual percentages of genomic chimerism in*** **`CRACOT`** ***simulations.*** *Introduced* ***Replacements, Duplications*** *and* ***Singletons*** *are expressed in gene proportions while* ***Chimerism*** *is a median value based on sequence length.*

Level	Replac.	Duplic.	Singlet.	Chimer.
Phylum	2.0	2.0	4.0	4.86
Class	2.0	2.0	4.0	4.91
Order	2.0	2.0	4.0	5.00
Family	3.0	3.0	4.0	6.17
Genus	4.0	4.0	3.0	6.58
Species	4.0	4.0	2.0	6.08

The contamination rate in the simulated chimeric genomes was estimated with six tools: 1) `CheckM` [3] with the `lineage_wf` option and the provided database, 2) `BUSCO` [5] in the `auto-lineage` mode and with the provided database, 3) `GUNC` [2] with default settings and the `proGenomes2` [6] database, 4) `Physeter` [7] with the `auto-detect` option, a `DIAMOND blastx` engine [8] and the database used in [7], 5) `Kraken2` [9] with default settings and the `PLUSFP` database downloaded from `https://benlangmead.github.io/aws-indexes/k2`, and 6) `CheckM2` [10] with default settings and the provided database.

Results

To the exception of `Kraken2`, all tools fail to recover the actual contamination percentage of the simulated genomes (**Fig. 4**). While `CheckM` (the field standard) and `GUNC` underestimate the chimerism, `BUSCO` and `Physeter` rather overestimate it. `CheckM2` also belongs to the latter category.

Perspectives

Even if the performance of `Kraken2` is impressive, its algorithm is based on exact matching of long *k*-mers and requires an exhaustive database of potential contaminants. Therefore, it is likely to decline once simulations will allow for mutational changes (**Fig. 3**, bottom) in order to emulate HGT in addition to mere technical contamination.

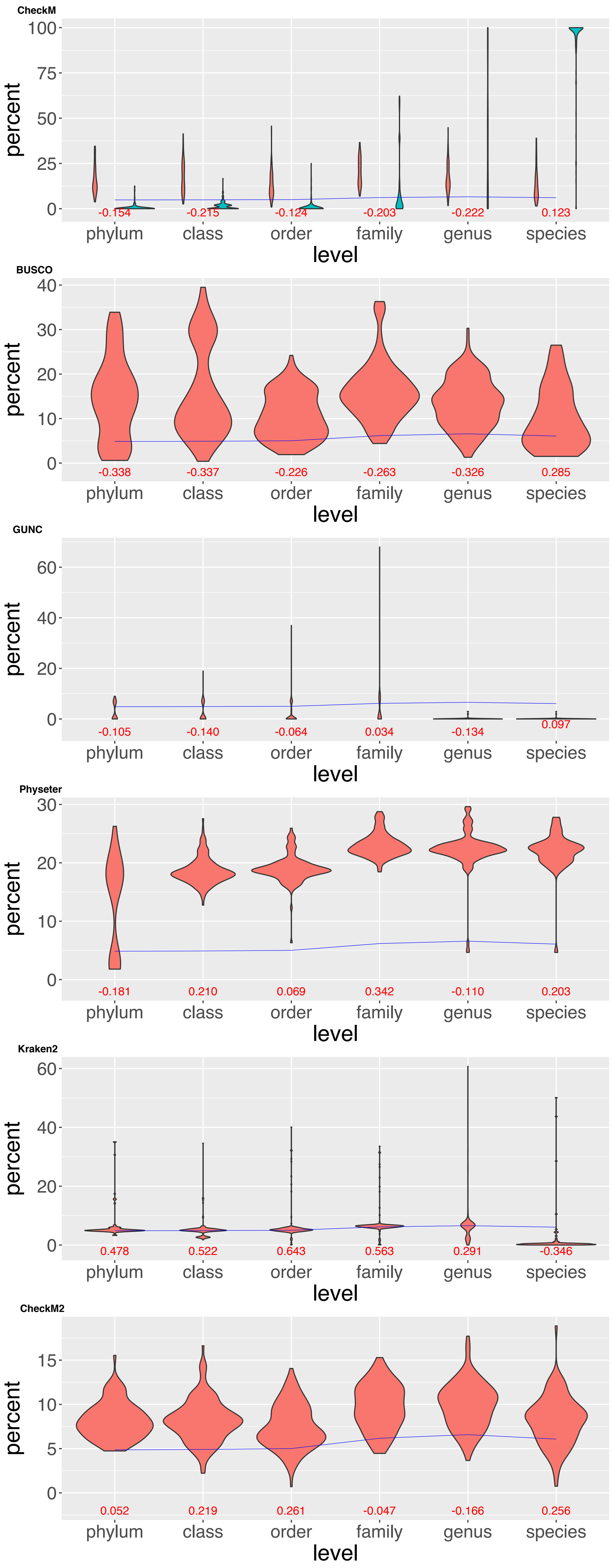

Figure 4: ***Contamination percentages estimated by the six tools*** **(`CheckM`, `BUSCO`, `GUNC`, `Physeter`, `Kraken2`, `CheckM2`)** ***at six taxonomic levels.*** *For* `CheckM`, *the* strain heterogeneity *is further shown in cyan. The blue line corresponds to the median chimerism value, as reported in* ***Table 1****. Numbers indicated in red are Spearman correlations between estimated and actual contamination values over 100 simulated genomes.*

References

1. Cornet L, Baurain D (2022) Contamination detection in genomic data: more is not enough. *Genome Biology* **23**:60.
2. Orakov A et al. (2021) GUNC: detection of chimerism and contamination in prokaryotic genomes. *Genome Biology* **22**:178.
3. Parks DH et al. (2015) CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research* **25**:1043–55.
4. Emms DM, Kelly S (2019) OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biology* **20**:238.
5. Manni M et al. (2021) BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. arXiv:2106.11799.
6. Mende DR et al. (2020) proGenomes2: an improved database for accurate and consistent habitat, taxonomic and functional annotations of prokaryotic genomes. *Nucleic Acids Research* **48**:D621–5.
7. Lupo V et al. (2021) Contamination in Reference Sequence Databases: Time for Divide-and-Rule Tactics. *Frontiers in Microbiology* **12**:3233.
8. Buchfink B et al. (2015) Fast and sensitive protein alignment using DIAMOND. *Nature Methods* **12**:59–60.
9. Wood DE et al. (2019) Improved metagenomic analysis with Kraken 2. *Genome Biology* **20**:257.
10. Chklovski A et al. (2022) CheckM2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning. bioRxiv 2022.07.11.499243.

EMBO | EMBL Symposium: Reconstructing the human past: using ancient and modern genomics, Heidelberg, Sept. 2022