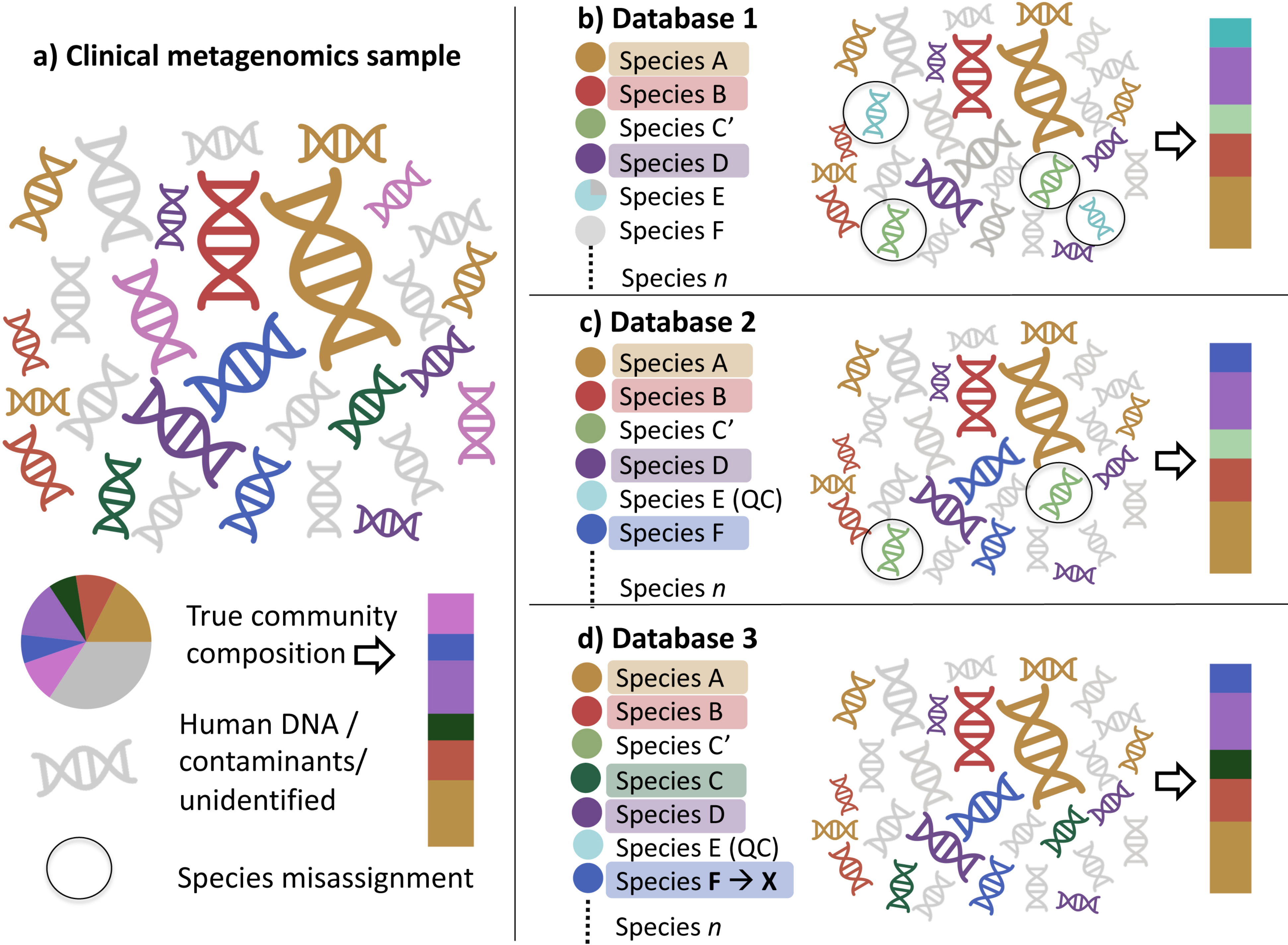
a) Clinical metagenomics sample
True community composition
Human DNA / contaminants/ unidentified
Species misassignment
b) Database 1
Species A
Species B
Species C’
Species D
Species E
Species F
Species *n*
c) Database 2
Species A
Species B
Species C’
Species D
Species E (QC)
Species F
Species *n*
d) Database 3
Species A
Species B
Species C’
Species C
Species D
Species E (QC)
Species **F → X**
Species *n*

Figure 2: The importance of reference database choice, design and versioning in taxonomic profiling of clinical metagenomics samples. a) Schematic representation of a typical clinical metagenomics sample with species assigned as coloured DNA and grey denoting DNA deriving from the host, contaminants, unidentified taxa or taxa sequenced at low depth. The pie chart provides the full metagenomic composition with the bar providing the species composition excluding host and contaminants. **b)** Taxonomic profiling based on Database 1. Species confidently assigned are highlighted by opaque sequence icons with unassigned species shown in grey. Using Database 1, Species A, B and D are correctly assigned. Species are also misassigned as outlined with a circle. In this instance, sequences from Species C are assigned to the closely related Species C' due to the lack of a representative of Species C in the reference database. Additionally, the reference database contains a partially contaminated sequence from Species E, which is misassigned to contaminant sequences in the test clinical metagenomics sample. This impacts the inference of species composition (bar). **c)** The addition of Species F to Database 2 allows assignment of a greater proportion of the species present in the original clinical metagenomics sample. Quality control and improvement of reference Species E, now Species E (QC), removes the spurious assignment of contaminant species. Species C is still misassigned to Species C', its closest representative in the database. **d)** Updating the reference database to include Species C results in the correct assignment of sequences to Species C rather than Species C'. Species F is taxonomically reassigned to Species X, leading to a change in the assigned species name despite no change in the data in the reference or query datasets. In all cases the pink sequences present in the original clinical metagenomics sample (panel a) are not assigned as this species is not present in any of the three reference databases.