

SUPPLEMENTAL MATERIAL

Prevalence and implications of contamination in public genomic resources: a case study of 43 reference arthropod assemblies

Clementine M. Francois^{1,2*}, Faustine Durand¹, Emeric Figuet¹, Nicolas Galtier¹

1. UMR 5554 - Institut des Sciences de l'Evolution; CNRS – University of Montpellier – IRD – EPHE; Place E. Bataillon – CC064; F-34095 Montpellier, France

2. Present address: Univ Lyon, Université Claude Bernard Lyon 1, CNRS, ENTPE, UMR5023 LEHNA, F-69622, Villeurbanne, France

* Corresponding author: clementine.francois@univ-lyon1.fr

Table S1: Genomic features of the 43 arthropod species from EnsemblMetazoa investigated in this study.

The N50 metric is used to describe the fragmentation level of the genome assembly, as 50% of the assembly is contained in scaffolds equal to or longer than this value. N_{CDS} corresponds to the number of coding sequences (CDS) annotated in EnsemblMetazoa for each species, while ‘N_{CDS} with tax’ refers to the subset of CDS for which a taxonomy can be reliably assigned during the first similarity-based step of the pipeline. ‘% CDS mapped’ refers to the proportion of foreign CDS candidates (i.e. assigned to a non-metazoan group) and ‘confident-arthropod’ CDS which were successfully mapped onto the genomic scaffolds during the second synteny-based step of the pipeline. Regarding contaminant and HGT candidates, the percentages indicated in this table were calculated based on the number of CDS with a reliable taxonomy (‘N_{CDS} with tax’), and not based on the total number of CDS. All metrics were calculated after filtering of the short scaffolds (< 200 bp) and CDS (< 150 bp).

Table S2: Composition of the custom reference database.

The number of CDS for each species is given after redundancy and length filtering. MBGD stands for the MicroBial Genome Database for comparative analysis (Uchiyama et al. 2014).

Table S3: Detailed categorization of all CDS in the 43 arthropod genomes.

This table summarizes the output of our pipeline when applied to the 43 arthropod assemblies.

First similarity-based step of the pipeline (DIAMOND BLASTP): ‘no reliable taxonomic assignment’ refers to the CDS without any hit or with too few reliable hits (orphan genes). Otherwise, CDS are either assigned to Metazoa (i.e. classical vertical descent) which is split between Arthropoda and ‘other metazoa’ (i.e. non-arthropod metazoa), or assigned to one of the five considered non-metazoan groups (eubacteria, archaea, fungi, viridiplantae and ‘protists’).

Second synteny-based step of the pipeline (GMAP): Foreign CDS candidates (i.e. those assigned to a non-metazoan group) are either categorized as contaminant candidates, HGT candidates or ‘uncertain’. (See Methods and Fig. 1 for details).

Table S4: Inferred function and potential donor for the six validated HGT families in the pea aphid assembly.

The broad taxonomy of the donor was inferred based on the best BLAST hits. The genus of the donor (indicated in brackets) was inferred based on the closest neighbour in the phylogenetic tree.

Figure S1: Correlation between the log-transformed N50 of each genome assembly and the percentage of foreign CDS candidates initially identified in the 1st similarity-based step of the pipeline which were subsequently considered as uncertain in the 2nd synteny-based step.

Each dot represents one species (n=43). The fitted regression line describes a linear relationship between both variables.

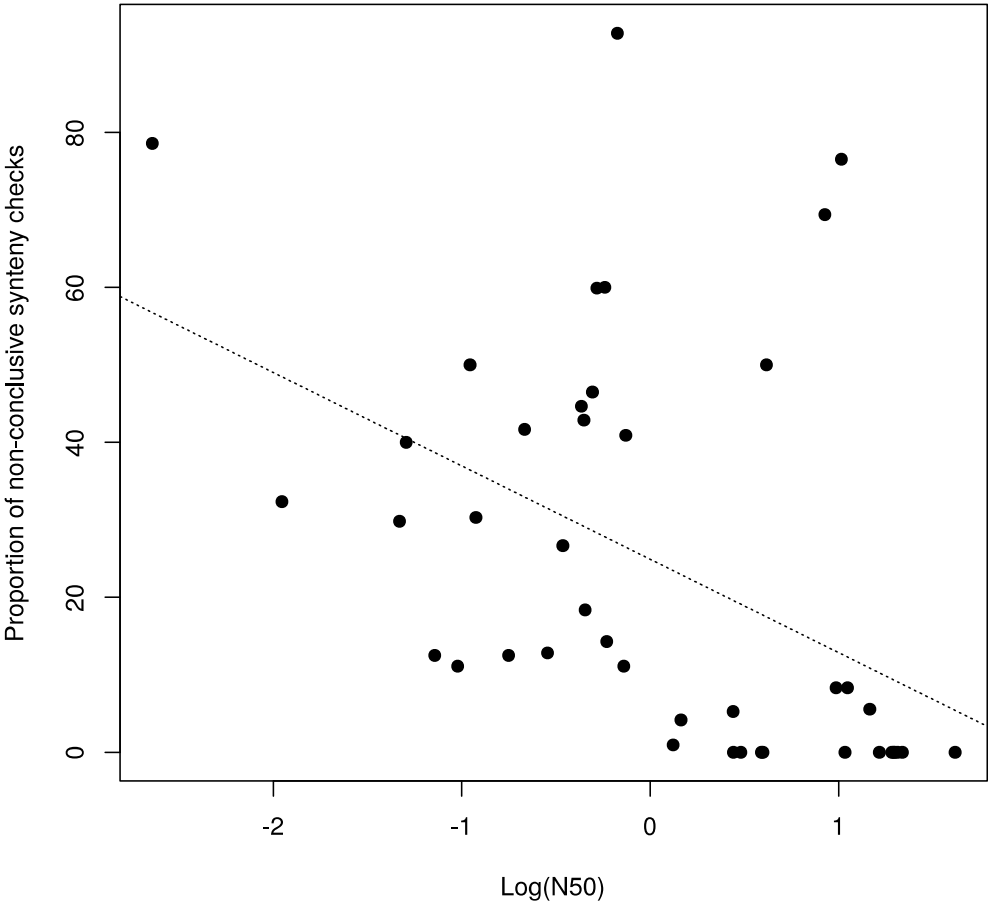


Figure S2: Number of contaminant (red circles) and HGT (black squares) candidates detected in each of the 43 arthropod genomes, according to the assembly N50.

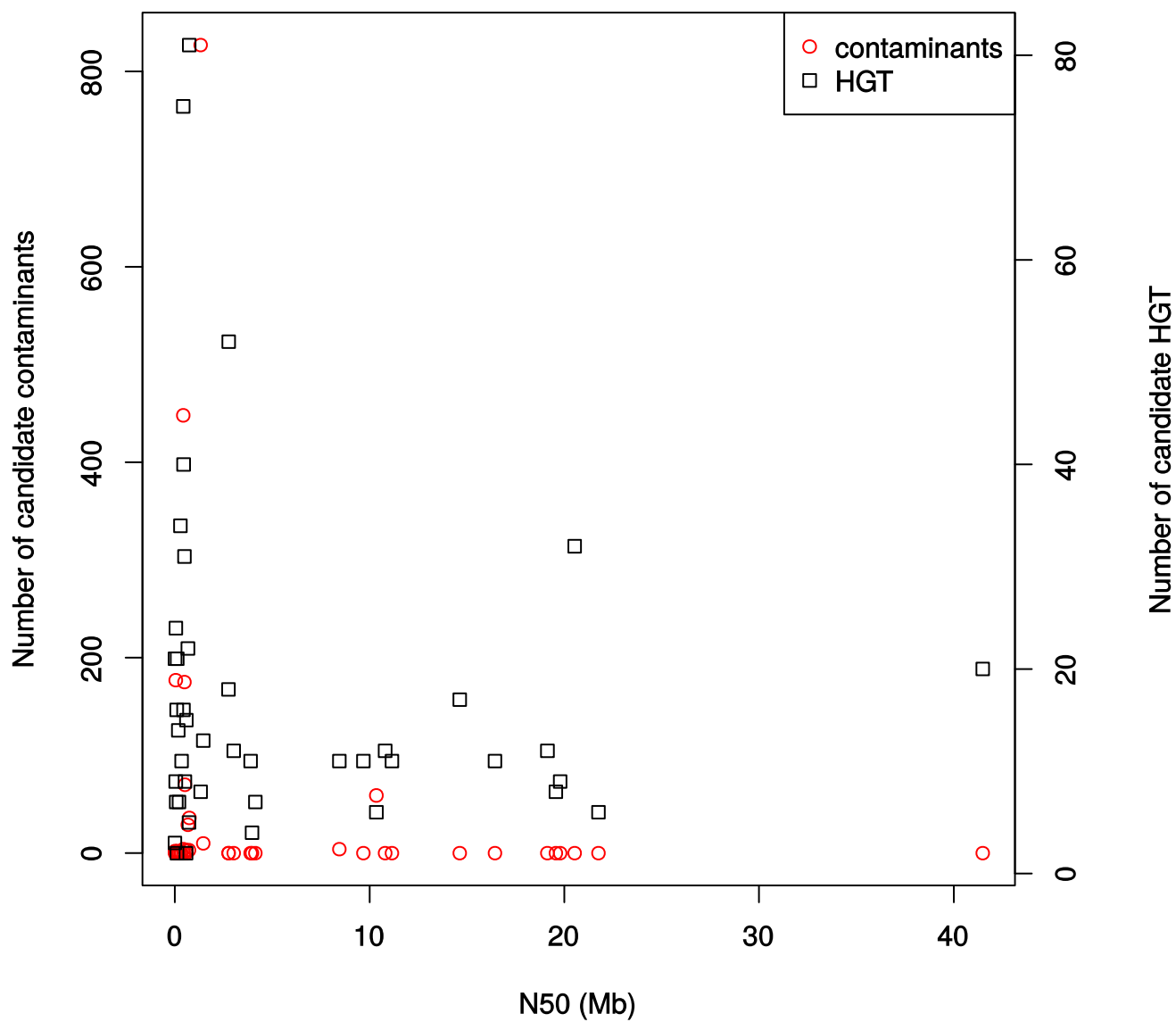
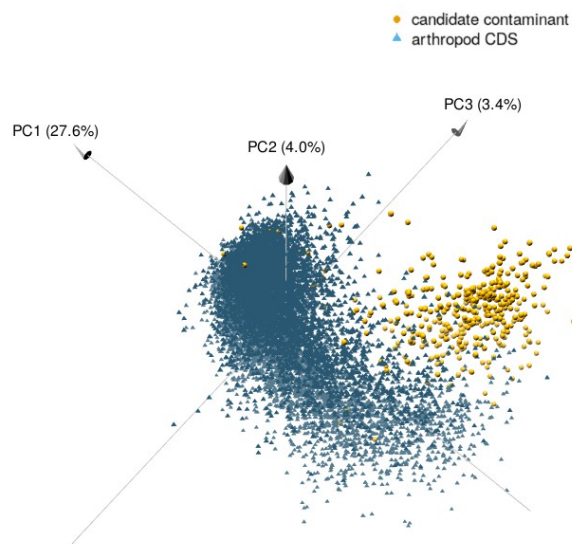


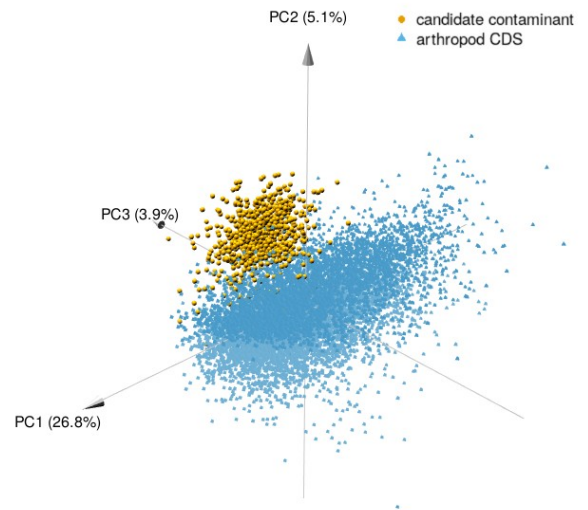
Figure S3: Principal Components Analysis on tetranucleotide frequencies differentiates between contaminant candidates and genuine arthropod genes.

(a) in the pea aphid (*Acyrtosiphon pisum*) assembly (n=448 contaminant CDS and n=11,927 “confident-arthropod” CDS). The first three principal components captured 27.6 %, 4.0 % and 3.4 % of the variance, respectively.

(b) in the bumblebee (*Bombus impatiens*) assembly (n=827 contaminant CDS and n=9,531 “confident-arthropod” CDS). The first three principal components captured 26.8 %, 5.1 % and 3.9 % of the variance, respectively.



(a) pea aphid (*Acyrtosiphon pisum*)



(b) bumblebee (*Bombus impatiens*)

Figure S4: Distribution of the number of contaminant CDS per contaminant scaffold.

Distribution is shown for all arthropod species (n=43), and separately for the pea aphid and for the bumblebee assemblies.

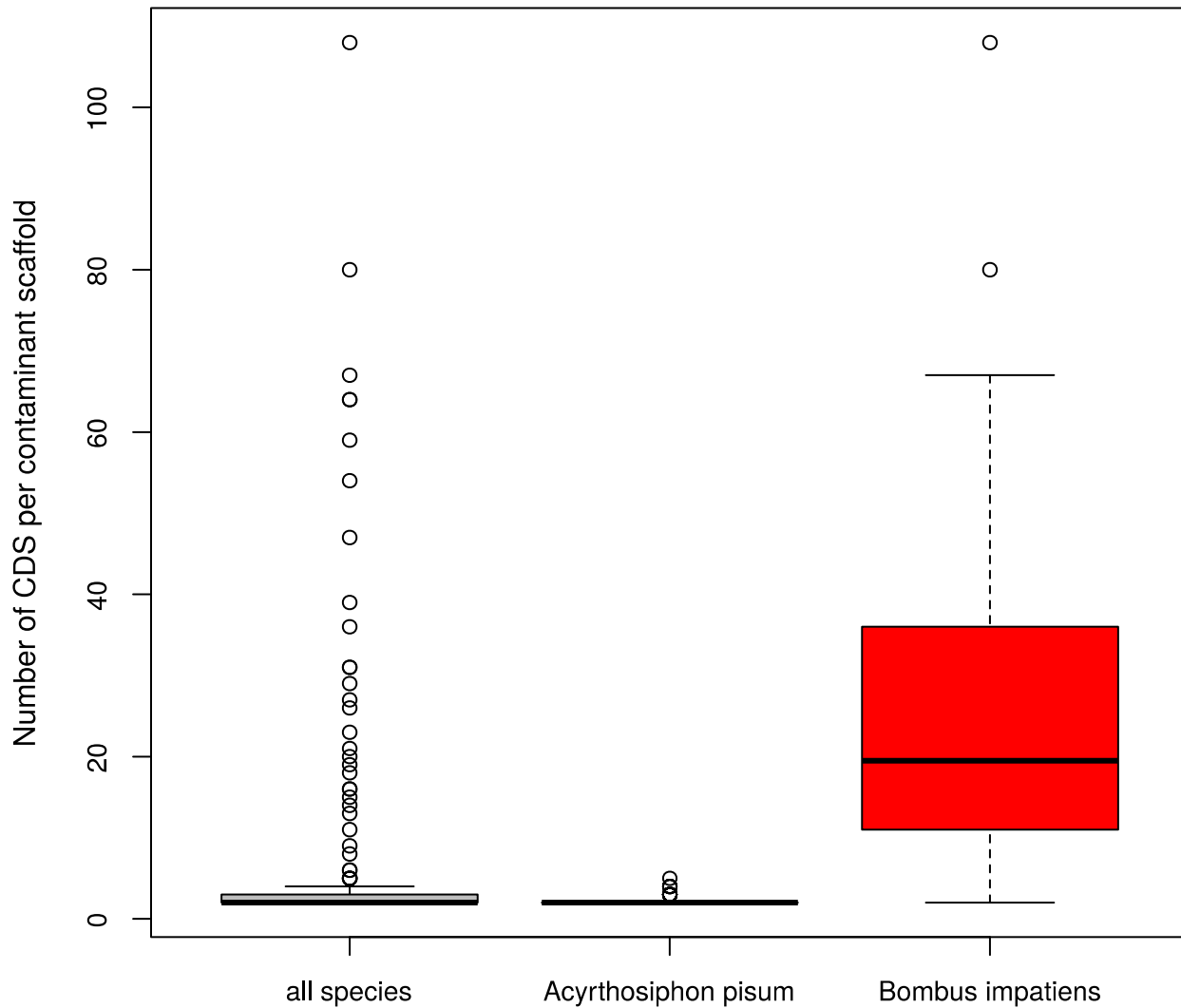


Figure S5: RAxML phylogenies inferred for the six validated HGT families in the pea aphid assembly.

Numbers indicated on the nodes correspond to the bootstrap support (100 replicates).

0.2

