SUPPLEMENTARY DATA

Supplemental Figure 1. Scatter plot showing the relationship between %G+C content and sequence length of chromosomal and plasmid sequences from different phyla. No correlation was found between %G+C content and sequence length. X-axis: sequence length (kb); Y-axis: mean %G+C content.

Supplemental Figure 2. Principal Component Analysis (PCA) plot showing the variation between the PlasFlow- and cBar- predicted plasmid, chromosomal, and unclassified sequence bins in terms of level1 subsystems abundance. Samples from uranium (KOW) and gold (ZS) mine microbial mats were assembled using SPAdes 3.9.1 and classified using PlasFlow. Obtained sequence bins were annotated using MG-RAST and analyzed using PCA to identify functions mostly contributing to the variance between classification bins. Individual bins. Colored symbols correspond to individual classification bins for each sample, where color code for a type of classification and shape for the source of sample. Vectors indicate the direction and strength of each subsystem to the overall distribution. 1- Phages, Prophages, Transposable Elements, Plasmids, 2 – DNA Metabolism, 3 – Respiration, 4 – Protein Metabolism, 5 – Carbohydrates, 6 – Amino Acids and Derivatives.

Supplemental Figure 3. Comparison of level3 subsystems abundance in the plasmid, chromosomal, and unclassified sequence bins (identified by PlasFlow) of samples from uranium (KOW) and gold (ZS) mine microbial mats. Samples were assembled using SPAdes 3.9.1 and classified using PlasFlow. Obtained sequence bins were annotated using MG-RAST and analyzed using DESeq2. A heatmap was drawn for subsystems with the highest variance using the complete-linkage method for column clustering.

Supplemental Figure 4. Comparison of level3 subsystems abundance in the plasmid, chromosomal, and unclassified sequence bins (identified by PlasFlow or cBar) of samples from uranium (KOW) and gold (ZS) mine microbial mats. Samples were assembled using SPAdes 3.9.1 and classified using PlasFlow. Obtained sequence bins were annotated using MG-RAST and analyzed using DESeq2. A heatmap was drawn for subsystems with the highest variance using the complete-linkage method for column clustering.

Supplemental Table 1: A list of plasmid and chromosome sequences, downloaded from the Refseq database, used in the neural network training and evaluation. For each sequence its phylogenetic classification and calculated %G+C content is provided.

Supplemental Table 2: Performance of individual neural network classifiers tested during the development of PlasFlow. As the input to the network, kmer frequencies were calculated for 5, 10, and 50 kb sequence fragments using kmer lengths in range k={3,4,5,6,7}. Hidden layer configurations tested included one-layer design with 20 or 30 neurons or a two-layer design with 10 or 20 neurons in each layer. For each of tested model accuracy, f1 score, precision and recall were calculated in the 1-fold cross-validation procedure. Classifiers with the best accuracies for 10 kb sequence fragments, included in the PlasFlow, are marked in green. Classifiers with the best accuracies for 5 and 50 kb sequence fragments, used for benchmarking, are marked in red.

Supplemental Table 3: Performance of PlasFlow classification on fragmented RefSeq genomes and plasmids (fragments of sequences listed in Supplemental Table 1) using different probability thresholds. For every run accuracy of plasmid or phylum prediction was calculated as well as the number of false predictions. Respective data for cBar shown for comparison. Values for classifiers run with the optimal (0.7) probability threshold are marked in red.

Supplemental Table 4: PlasFlow evaluation on public plasmidome data. For each assembled plasmidome dataset, classification was performed using PlasFlow (with a 0.7 probability threshold). cBar and PlasmidFinder. For each dataset the number of sequences assigned to chromosome, plasmid or unclassified class is reported as well as the cumulative length of all sequences belonging to each class. For PlasFlow and cBar classifications the plasmid to chromosome ratio was calculated, both by the mean of number of sequences and cumulative length, to show the enrichment of plasmids in the datasets.

Supplemental Table 5: Results of PlasFlow and cBar classification of the Aeromonas sp. 023A draft assembly scaffolds. PlasFlow was run using default probability threshold (0.7). BLASTn was used to confirm which scaffolds came from plasmids.

Supplemental Table 6: Comparison of PlasFlow and cBar performance on the dataset containing 42 bacterial genomes with various quantity of plasmids. For each genome the Short Read Archive (SRA) and reference genome assembly accessions are given. For both the PlasFlow (with 0.7 probability threshold) and cBar the calculated precision and recall values are reported.

Supplemental Table 7: PlasFlow classification results of the RefSeq contigs dataset. 101454 sequences from the Refseq bacterial genomic draft assemblies (as described in Materials and Methods) were classified using PlasFlow with default (0.7) probability threshold. For all resultant datasets a presence of plasmid-related annotations was checked as well as accuracy of phylum prediction.

Supplemental Table 8: Assembly summary of KOW and ZS samples. 3 independent samples from each environment were assembled with SPAdes 3.9.1 (raw reads are deposited in MG-RAST – accessions provided in the table).

Supplemental Table 9: PlasFlow, cBar, PlasmidFinder and Recycler classification results of assembled KOW and ZS samples. For each sample number of sequences assigned to each class (plasmid, chromosome, unclassified) is shown, as well as cumulative and mean length of sequences in each sequence bin.

Supplemental Table 10. Subsystem annotation data, represented as the raw counts, for all obtained classification bins. The same counts (after required grouping) were used as the input for PCA and DESeq2 analyses. Data for PlasFlow, cBar, and Recycler classifications are shown. PlasmidFinder as well as ZS2 recycler annotations are not shown, since the MG-RAST is only able to analyze those datasets that consist of more than 100 sequences.

All supplemental tables are available in the separate Excel workbook file.