Supplementary file for

Identification of diversity-generating retroelements in host-associated and environmental genomes: Prevalence, diversity, and roles

This file contains 6 supplementary figures and the description of 13 supplementary tables and 6 data sets. The supplementary tables can be found in four separate table files. The 6 data set files can be found at https://cgm.sjtu.edu.cn/DGRs-MAGs/index.html.

Supplementary Figures.

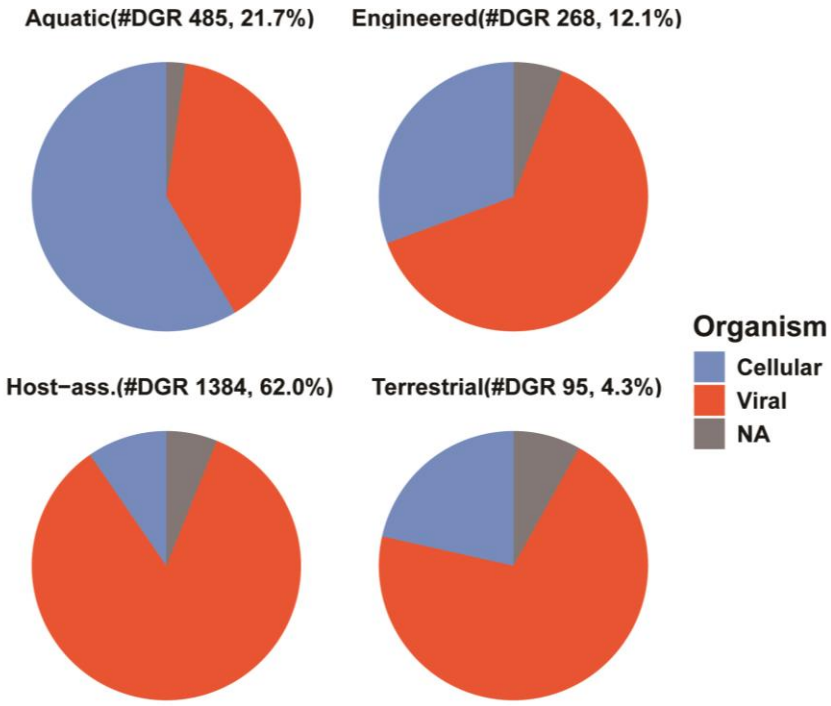

Figure S1. Proportionate distribution of the proportion of viral and cellular forms of the species in which DGR is found in different environments. Four biomes, Aquatic, Engineered, Host-associated, Terrestrial, are shown. Among the host-associated microorganisms, the proportion of viruses is the highest. In the aquatic environment, the proportion of cellular is the highest. Due to the limited data for terrestrial MAGs, these results may be biased.

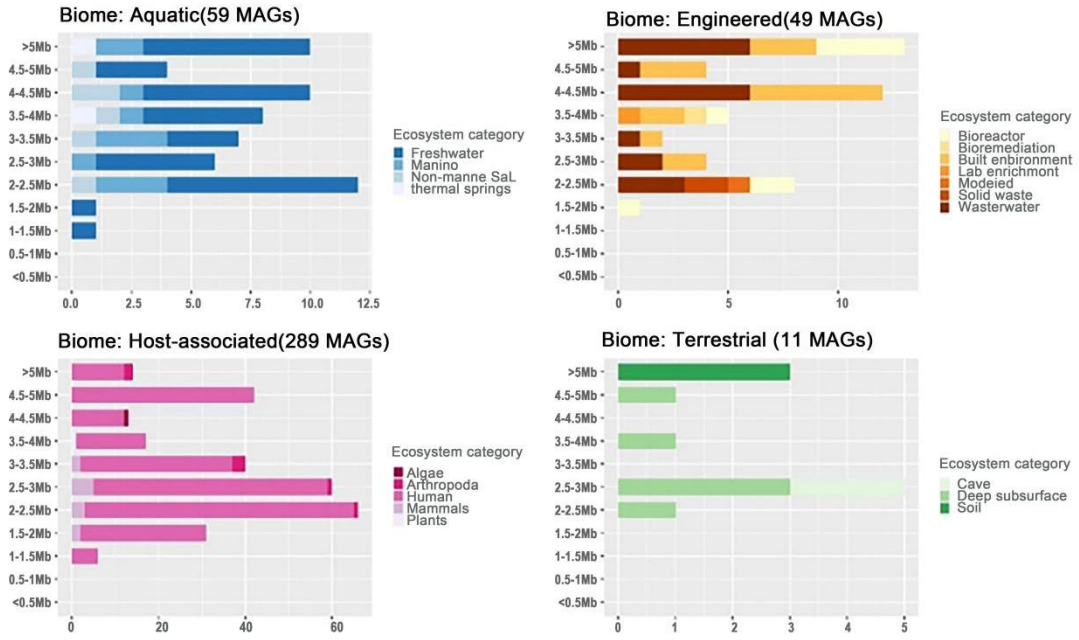

Figure S2. Prevalence of DGRs across different genome sizes. Number of DGR-containing MAGs per different ranges of genome size colored by ecosystem category.

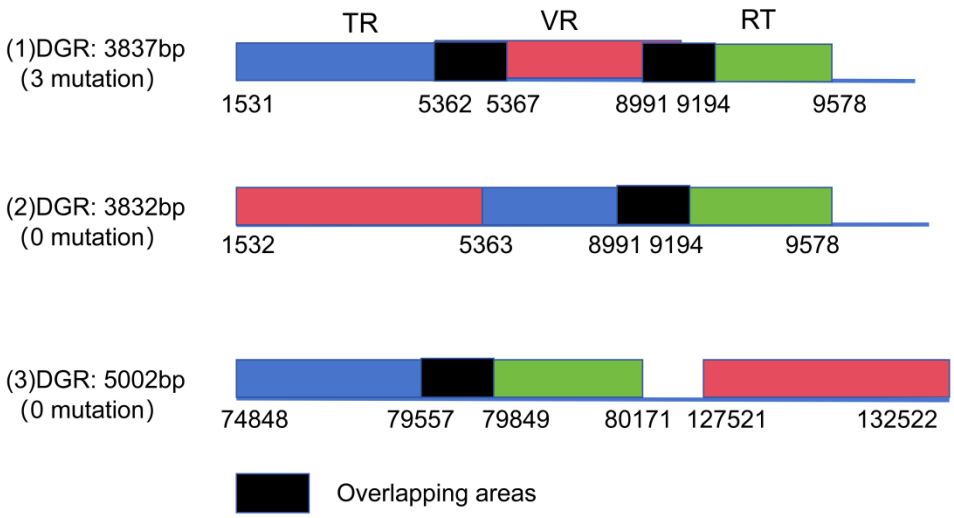

Figure S3. Structural patterns of three atypical-length DGRs. The VR, TR, and RT regions of the DGR structures in the figure exhibit some overlap; therefore, we excluded these three DGRs.

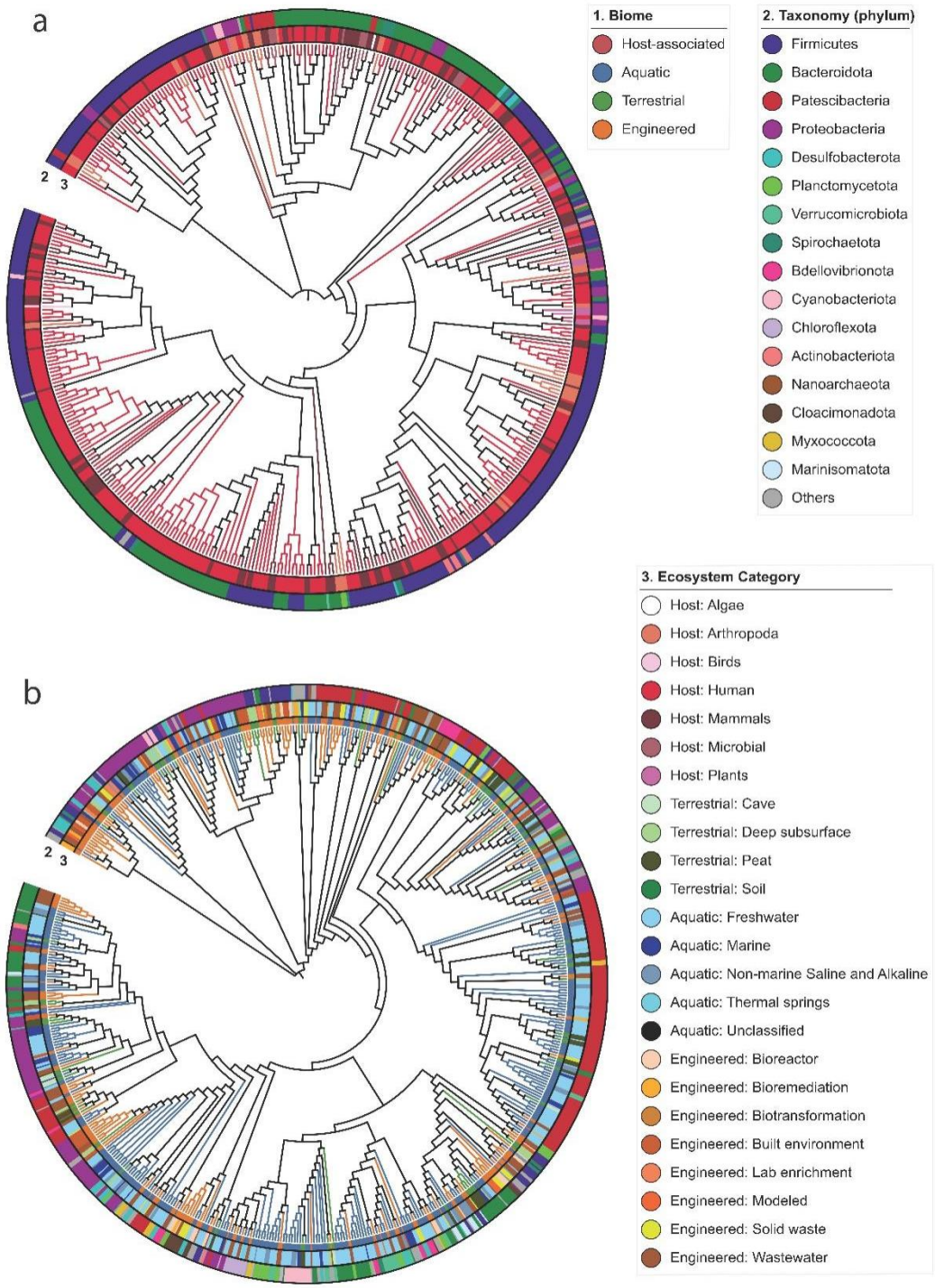

Figure S4. Phylogenetic trees of DGR reverse transcriptases (amino acid sequences) **(a)** for host-associated genomes, **(b)** for environmental genomes (aquatic, terrestrial and engineered). The branches are colored based on the biomes, the inner ring represents the ecosystem categories and the outer ring shows the taxonomic annotation at phylum level.

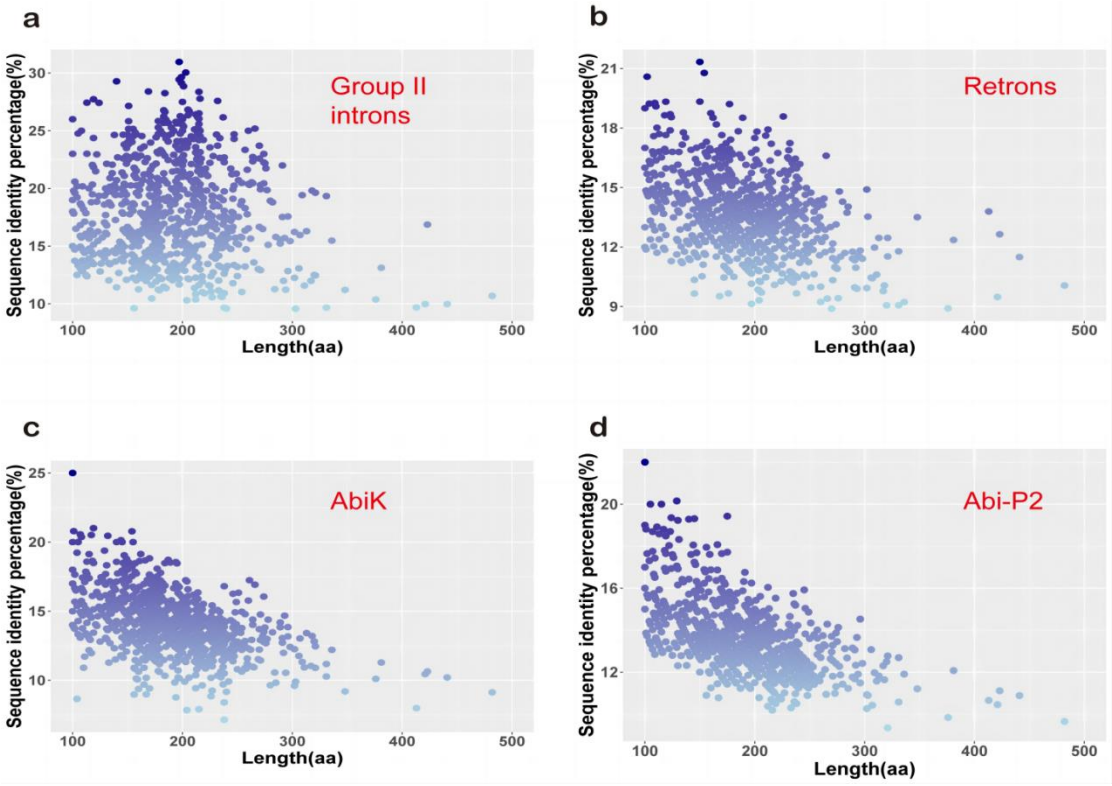

Figure S5. Distribution of sequence comparison scores of DGR-RTs and non-DGR-RTs. (a) DGR-RTs and the RTs of group II introns. (b) DGR-RTs and the RTs of retrons. (c) DGR-RTs and the RTs of AbiK. (d) DGR-RTs and the RTs of Abi-P2. Non-DGR-RTs and DGR-RTs are all less than 30% similar, indicating that the non-DGR-RTs are very different from the DGR-RTs.

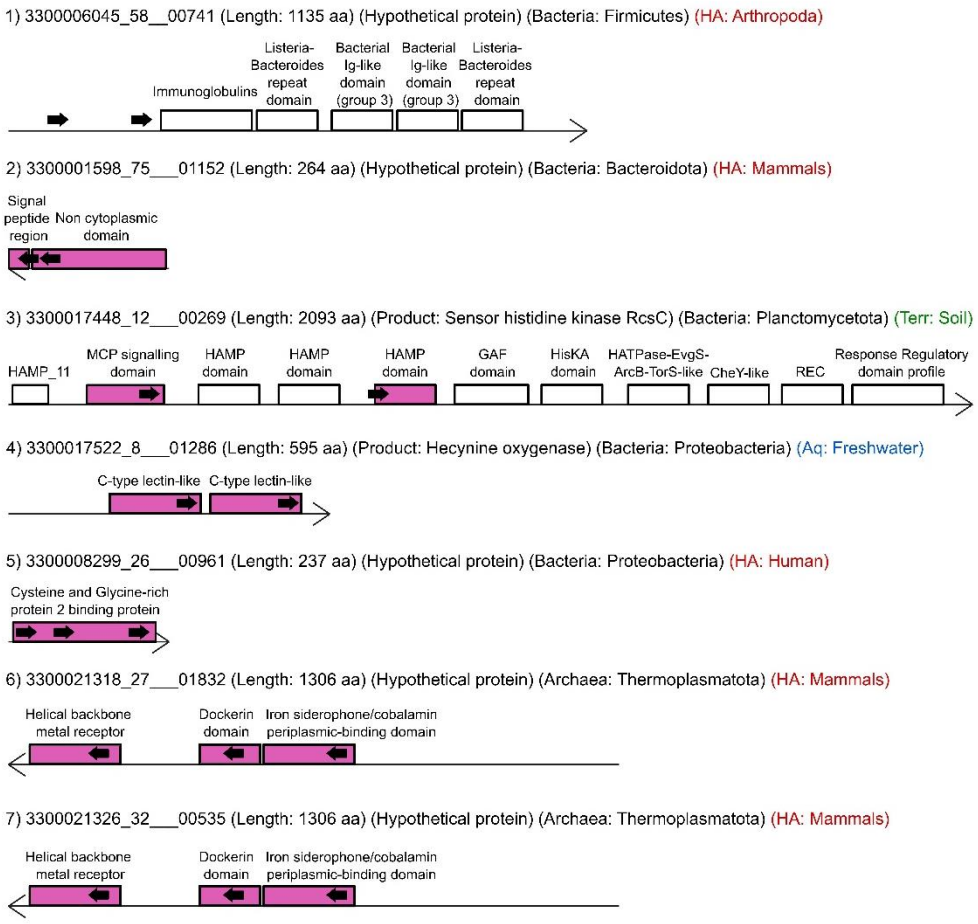

Figure S6. Target proteins with 2 and 3 variable regions. The domains are shown for each of them. The sequence length in amino acids (aa), the product, the phyla and environment to which the proteins belong are shown in parentheses. The regions highlighted in pink represent the domain that contains the VR, which is symbolized by the black arrow. The direction of the arrows indicates whether it was found on the positive (right) or negative (left) DNA strand.

Supplementary Tables （S1-S12）

Additional file 1: Table S1-S8. Metadata of the 2,014 DGR-containing MAGs with the environmental and taxonomic annotations taken from Nayfach *et al*. (2020). Frequency tables (XLSX 8.61 MB)

Additional file 2: Table S9. The numbers of MAGs containing DGRs with different cassette patterns (XLSX 12.1 KB)

Additional file 3: Table S10. Count of products per ecosystem category (XLSX 13.9 KB)

Additional file 4: Table S11. Most common domain annotations (XLSX 10.4 KB)

Additional file 5: Table S12. Final-viral-scores (XLSX 145 KB)

Additional file 6: Table S13. Interpro_RTs_results (XLSX 415 KB)

Supplementary data sets

The following 6 data set files listed below can be found at https://cgm.sjtu.edu.cn/DGRs-MAGs/index.html.

Data S1. Output file from the tool MetaCSST with the detected DGR components per MAG (GTF 2.54 MB)

Data S2. 889 non-redundant RT amino acid sequences (FAA 194 KB)

Data S3. Functional annotation file showing all putative target genes (CSV 772 KB)

Data S4. Folder containing the InterProScan domain annotations for each product in different TSV files, and the overlapped VRs (ZIP 123 KB)

Data S5. The 2014 DGR-containing MAGs (ZIP 1.7GB).

Data S6. The DGRs detected by MetaCSST (ZIP 750KB).