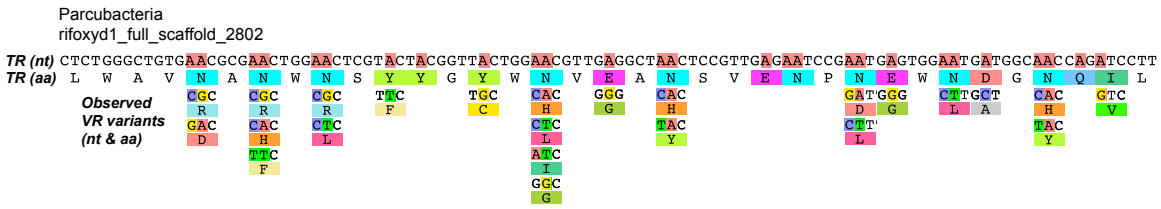
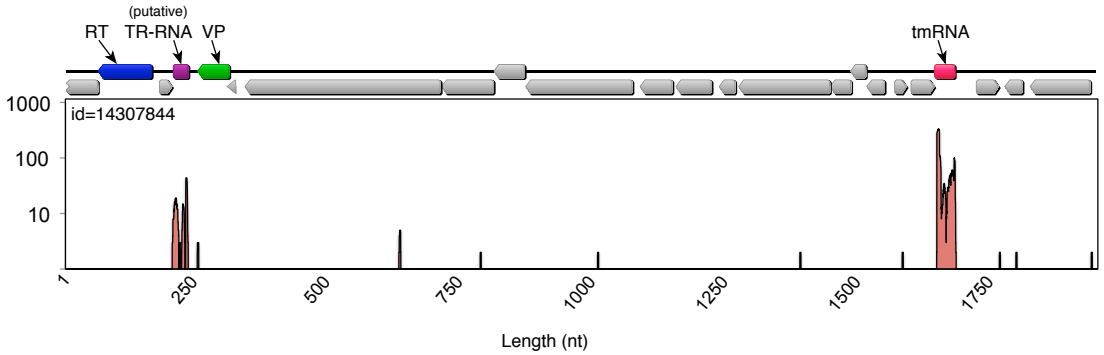
# Retroelement guided protein diversification abounds in vast lineages of bacteria and archaea

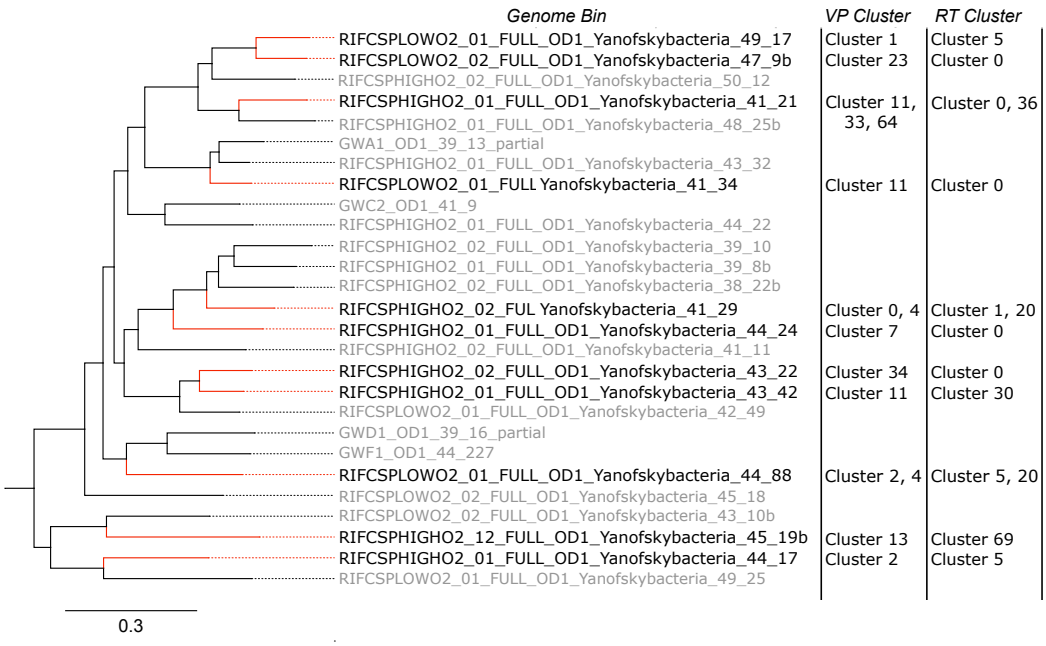
Supplementary Information



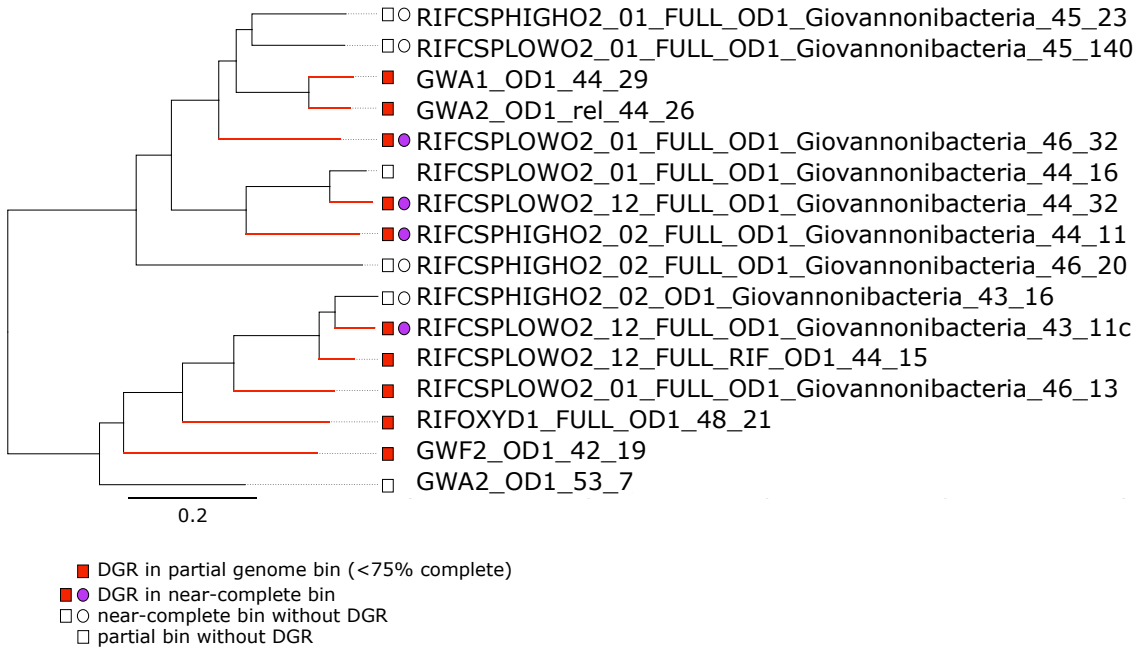

**Supplementary Figure 1. Adenine-specific substitutions compared with a TR sequence.** An example of the VR variants (nt, nucleotide; aa, amino acid) observed by read-mapping analysis of a Parcubacteria genome DGR.



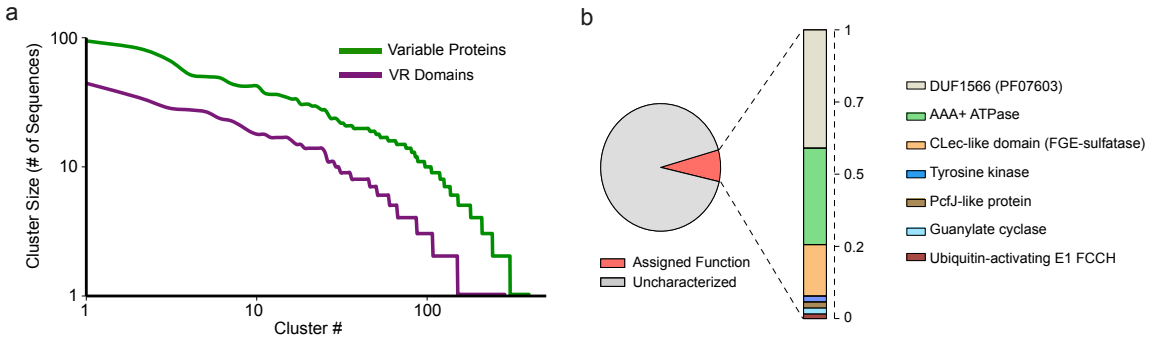

**Supplementary Figure 2. Stringent transcript mapping to DGR-containing features.** Showing an example sequence based on high coverage for a DGR feature, where transcripts are mapped to the TR-RNA coding region (purple boxes), as well as to tmRNA encoding sequence (pink boxes). The variable protein gene is labeled “VP”.



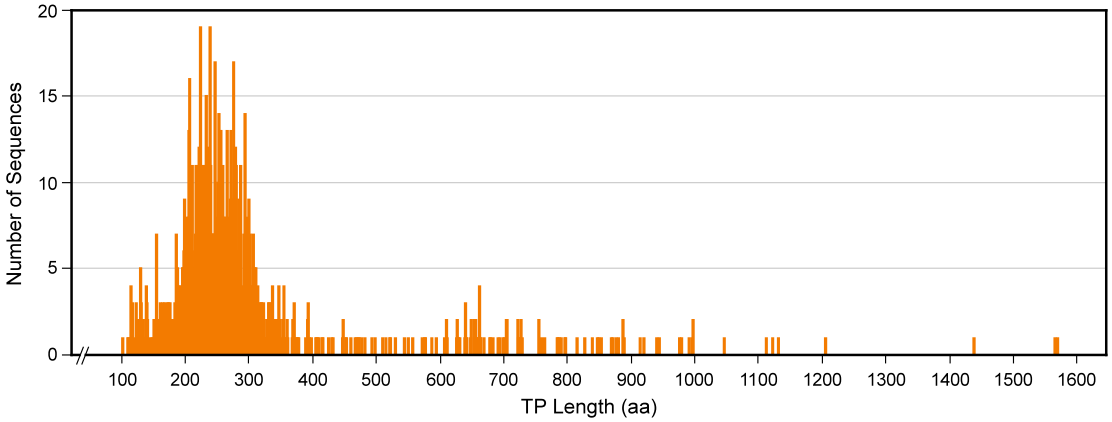

**Supplementary Figure 3. Phylogenomic distribution of DGR-containing Yanofskybacteria genomes**. A concatenated ribosomal tree from Hug et al. 2016 was used to overlay DGR occurrence (red; black text) versus absence (grey text). Variable protein and reverse transcriptase cluster affiliations are indicated to the right of genomes that contain a DGR. The scale shows substitutions per site.



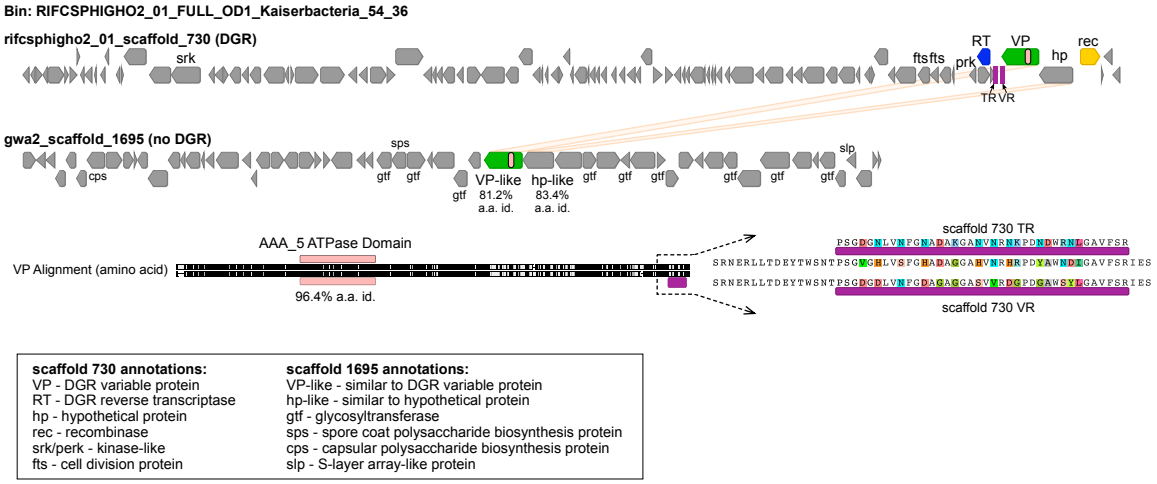

**Supplementary Figure 4. Phylogenomic distribution of DGRs in near-complete and partial Giovannonibacteria genomes**. A concatenated ribosomal tree from Hug et al. 2016 was used to overlay DGR occurrence (red branches and square/circle symbols). The scale shows substitutions per site.



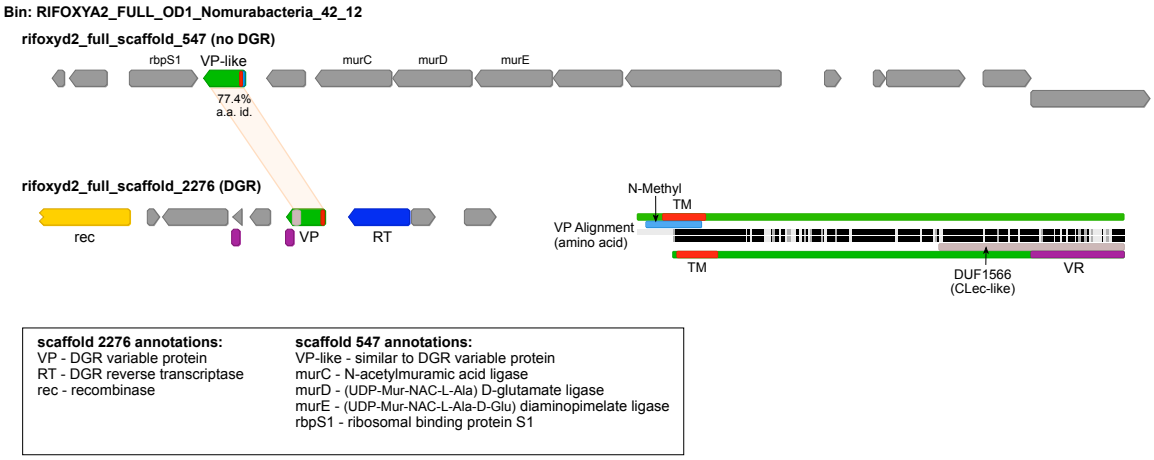

**Supplementary Figure 5.** Variable protein clusters and functional classes for combined binned and unbinned genome fragments. a, Clustering analysis of variable protein sequences (amino acid) with at least 30% intracluster similarity. Distributions of globally aligned variable proteins, and separately, VR domains, are shown on distinct lines as indicated in the legend. b, Relative proportions of characterized versus uncharacterized variable proteins and corresponding putative functional classifications



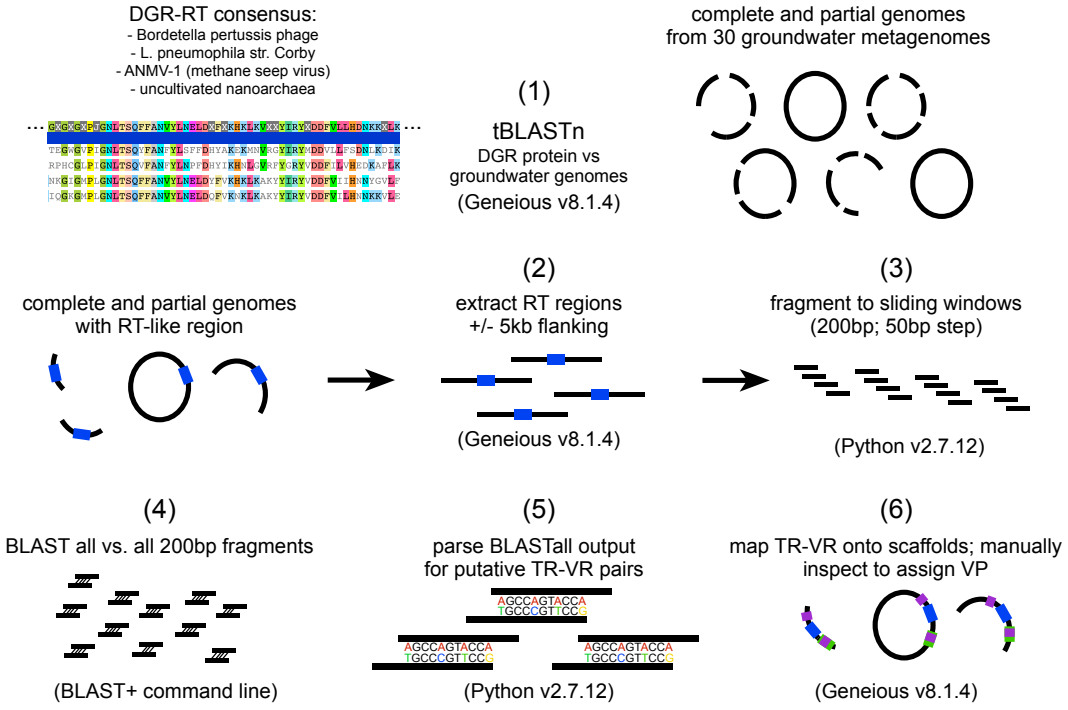

**Supplementary Figure 6. Length distribution of DGR variable proteins.** A non-redundant list of variable proteins was generated using CD-Hit (intracluster global alignment identity >95%). Sequence lengths are indicated in amino acids (aa).




**Supplementary Figure 7. Paralagous variable proteins in DGR and non-DGR loci of a Kaiserbacteria genome.** Sequences with homology are highlighted with a yellow connector and amino acid similarity is shown below each annotation (VP-like and hp-like). Selected neighborhood annotations are indicated by abbreviated gene names. An expanded view depicts the alignment of variable protein and VP-like amino acid sequences. At right, the aligned VR, TR, and VR-like sequences show variable residues.




**Supplementary Figure 8. Paralagous variable proteins in DGR and non-DGR loci of a Nomurabacteria genome.** Sequences with homology are highlighted with a yellow connector and amino acid similarity is shown below the VP-like annotation. Selected neighborhood annotations are indicated by abbreviated gene names. An expanded view depicts the alignment of variable protein and VP-like amino acid sequences. A putative transmembrane region is highlighted in red (TM).




**Supplementary Figure 9. Overview of DGR feature identification.** The schematic depicts steps taken to identify RT-like regions on partial or complete genomes reconstructed from groundwater metagenomes. Software/programs used to conduct each step are listed in parentheses.