Prokaryotic Genome Database

Archaeal and bacterial genome sequences were downloaded from the NCBI FTP site (ftp://ftp.ncbi.nlm.nih.gov/genomes/all/) in March 2016. For incompletely annotated genomes (coding density less than 0.6 CDS per kbp) the existing annotation was discarded and replaced with Meta-GeneMark\(^1\) annotation. Altogether the database includes 4,961 completely sequenced and assembled genomes and 43,599 partially sequenced genomes coding for 1.85x10^8 proteins.

Clustering and Phylogenetic Analysis

To construct a non-redundant, representative sequence set, sequences were clustered using the NCBI BLASTCLUST program\(^2\), (ftp://ftp.ncbi.nih.gov/blast/documents/blastclust.html) with the sequence identity threshold of 90% and length coverage threshold of 0.9. The longest sequence was selected to represent each cluster. Permissive clustering of sequences was performed using UCLUST\(^3\), with sequence similarity threshold of 0.3.

Multiple alignments of protein sequences were constructed using MUSCLE\(^4\) and MAFFT\(^5\) programs. Sites with the gap character fraction values >0.5 and homogeneity <0.1\(^6\) were removed from the alignment. Phylogenetic analysis was performed using the FastTree program\(^7\), with the WAG evolutionary model and the discrete gamma model with 20 rate categories.

Relationships within diverse sequence families were established using the following procedure: initial sequence clusters were obtained using UCLUST\(^3\) with the sequence similarity threshold of 0.5; sequences were aligned within clusters using MUSCLE\(^4\). Then cluster-to-cluster similarity scores were obtained using HHSEARCH\(^8\) (including trivial clusters consisting of a single sequence each); a UPGMA dendrogram was constructed from the pairwise similarity scores. Highly similar clusters (pairwise score to self-score ratio >0.1) were aligned to each other using HHALIGN\(^8\); the procedure was repeated iteratively. At the last step, sequence-based trees were reconstructed from the cluster alignments using the FastTree program\(^7\) as described above and rooted by mid-point; these trees were grafted to the tips of the profile similarity-based UPGMA dendrogram.
Pipeline for CRISPR-Cas loci annotation

The pipeline takes a list of locations (coordinates in the corresponding nucleotide sequence) of the seed features as input. Two types of seeds were used: locations of *cas1* genes in the NCBI WGS database and locations of CRISPR arrays in the WGS database and the prokaryotic genome database. TBLASTN searches with the E-value cutoff of 0.01 and low complexity filtering turned off were run with the Cas1 profiles as queries, resulting in the identification of 20,766 loci. The CRISPRfinder and PILER-CR programs were used with default parameters to identify CRISPR arrays in the WGS database (47,174 loci found) and in the prokaryotic genome database (45,373 loci found). Sequences including up to 10 kbp upstream and downstream of the seed features were extracted.

Open Reading Frame (ORF) annotation was performed using Meta-GeneMark with the standard model MetaGeneMark_v1.mod (Heuristic model for genetic code 11 and GC 30). All ORFs were further annotated using RPS-BLAST searches with 30,953 profiles (COG, pfam, cd) from the NCBI CDD database and 217 custom Cas protein profiles. The CRISPR-Cas system (sub)type identification for all loci was performed using the previously described procedures.

Among all permissive clusters constructed from proteins from the seed loci, potential candidates were selected using the size threshold (> 500aa), the distance to seed (genes closest to the seed were preferred); the selection of candidates was limited to those that were located within 4 genes from the seed; clusters that contained more homologs outside the seed loci than in those loci were discarded. Additional prediction of protein domains was performed using the CD-search and HHpred.

The identified candidates were used as queries for a PSI-BLAST search against the NCBI NR and NCBI WGS databases for the Cas1 seeds, and NCBI WGS and prokaryotic databases for the CRISPR seeds in order to obtain additional loci that were added to the seed list. The evaluation procedure was then repeated until convergence.
Protospacer analysis
The initial pool of 488,437 spacers in the CRISPR arrays was reduced to 268,409 unique sequences. The MEGABLAST program \( \text{word size} \, 18 \) was used to search for protospacers in the virus subset of NR database (TaxID:10239) and the prokaryotic genome database. The maximum number of mismatches for a spacer with length \( l \) was limited to \( \max(0, \sqrt{l - 22}) \). All MEGABLAST hits that target CRISPR arrays as well as all eukaryotic virus sequences were discarded. This procedure produced 63,939 hits to prokaryotic genomes and 5,095 hits to prokaryotic viruses. The 33480 ORFs that contained or intersected with the detected protospacers were used as BLASTP queries to search the virus database. All ORFs with strong hits (e-value <10\(^{-6}\)) were classified as originating from (pro)viruses.

Synteny analysis of subtype V-U loci
Protein sequences encoded by genes in the vicinity (±3 genes) of the Type V-U effector genes were extracted and clustered using UCLUST \(^3\) with the similarity threshold of 0.3. Genes were annotated by the cluster IDs; each locus was represented as a set of genes and unordered gene pairs. Weighted Jackard similarity coefficient was calculated for all pairs of loci as previously described \(^{15}\), a locus similarity graph was constructed with the similarity threshold of 0.61 (e\(^{-0.5}\)), and connected components (subsets of highly similar loci) were extracted.

Analysis of selection in the evolution of Class 2 effector genes
Nucleotide and protein sequences of the effector genes were collected; clusters of identical protein sequences were reduced to a single representative; remaining sequences were clustered using UCLUST \(^3\) with the similarity threshold of 0.67. The sequences from each cluster were aligned, and a phylogenetic tree was constructed as described above and rooted using a modified midpoint procedure. Sub-alignments of protein sequences, corresponding to sub-trees with the mean depth <0.1, were extracted and converted to the nucleotide sequence alignments. Pairwise \( dN \), \( dS \) and \( dN/dS \) values were obtained using the codeml program of the PAML package \(^{18}\). Sequence pairs with \( 0.0002 \leq dN \leq 1.0 \) and \( 0.0002 \leq dS \leq 1.0 \) were selected, and the \( dN/dS \) values were calculated.
References


