Supplementary information S3 (box) | **Methods**

# Methods

## Genome weighting

The currently available collection of archaeal and bacterial genomes has a highly biased distribution of isolates across taxa. For example, it includes 46 strains of *Escherichia coli*, whereas entire phyla, such as Nanoarchaeota, Korarchaeota, Chrysiogenetes and others, are represented by a single genome. This extreme bias makes quantitative characterization of genomic features challenging and renders unusable most standard statistical methods that rely on random independent sampling as a null model. A relative genome weighting scheme that assigns low values to members of the densely sampled clades and high values to lone representatives of clades, can be used to mitigate the effects of the sampling bias.

Two notions are central to our model of relative genome weighting: first, closely related genomes should contribute individually less to the total clade weight than their more distant relatives; second, the relative contribution of the clades should reflect the number of independent evolutionary events that occurred in the history of the clade. Using the sum of branch lengths in a (sub)tree allows one to quantify both concepts.

Consider a node in a rooted phylogenetic tree that has several descendant clades, each with the sum of branch lengths (including the length of the branch connecting this subtree to the parent node) $T_i$. If the total weight assigned to this node is set to $W$, then it is distributed between the descendant subtrees as $W_i = WT_i/\Sigma T_i$. The sums of branch lengths for each internal tree node can be easily computed iteratively in the leaf-to-root direction and the total tree weight can be iteratively distributed between clades and leaves in the root-to-leaf direction.

To estimate the genome weights, we used an approximate phylogenetic tree reconstructed from concatenated alignments of ribosomal proteins [1] that was rooted between bacteria and archaea. The subtree encompassing the 1302 *cas*-positive genomes was extracted from the original tree. The weights calculated using this procedure are robust to minor perturbations of tree topology, especially those that involve deep clades and short internal branches.

## CRISPR-*cas* loci identification

An exhaustive search for *cas* genes was performed within the set of protein sequences annotated in 2751 complete archaeal and bacterial genomes that were available at the NCBI as of February 1, 2014. The 185 multiple sequence alignments of Cas proteins that were not available through public databases were constructed and added to the ~29,500 CD, COG and PFAM profiles in the NCBI CDD database [2]. Altogether, 395 profiles represented 93 distinct Cas protein families. Searches were performed using PSI-BLAST [3], with the alignment consensus employed as the master query.

The 93 *cas* genes were classified by sequence similarity into 35 families that belong to 12 distinct functional classes according to the functions of the respective proteins in CRISPR-Cas systems (Supplementary File 1). Of the 35 families of cas genes, 11 constitute the *cas* core and the rest are classified as "ancillary".

The *cas* loci were identified using a two-step procedure. In the first step, PSI-BLAST search results with e-value threshold of $10^{-6}$ were used to annotate all proteins in the set of complete archaeal and bacterial genomes. The highest-scoring profile for all non-overlapping sequence segments were identified. In the second step, gene products from neighborhoods of ±20 genes around all identified *cas* genes were used as queries for the second round of PSI-BLAST search with e-value threshold of 0.01. Additional genes with moderately significant matches to Cas profiles and located in the vicinity of confidently predicted *cas* genes were identified.

Gene neighborhoods of ±5 genes around all identified *cas* genes were extracted; overlapping neighborhoods were merged and trimmed to the first and the last *cas* gene, to form the candidate loci. A locus that contains at least two *cas* genes, of which at least one gene belongs to the *cas* core, was identified as a valid *cas* locus.

### Profile-based CRISPR-*cas* loci classification

A set of Cas sequence profiles was collected over the years since the previous publication on CRISPR-Cas classification [4]. Correspondence between the profiles, gene names and CRISPR-Cas system types and subtypes was reexamined in the course of this work. To assist the assembly of a non-redundant and self-consistent set of Cas protein profiles, the multiple profiles for Cas5, Cas7 and Large Subunit were aligned to each other using HMMER 3.0 [5] and cluster dendrograms were constructed from matrices of relative pairwise scores using UPGMA. The dendrograms were examined for inconsistent annotation of similar profiles; potential discrepancies were investigated on a case by case basis, and annotation was adjusted where required.

Loci were classified using the correspondence table between Cas sequence profiles and CRISPR-Cas (sub)types (Supplementary File 1). The classification procedure consisted of two steps. First, a gene group annotation was used to identify genes of the effector module (*cas5*-like, *cas7*-like and Large Subunit), *cas9* and *cpf1*. A genomic segment containing either each of the major effector module genes or one *cas9* gene or one *cpf1* gene was considered a complete CRISPR-Cas system unit of type I/III/IV, type II or type V, respectively. Loci that contained neither the full complement of effector module genes nor *cas9* or *cpf1* were classified as partial.

At the second step, each locus unit or single-unit locus was analyzed separately. Each Cas profile within the unit contributed a "vote" for the type and subtype that this profile corresponded to. Contributions from profiles with multiple affinities (such as, for example, *cas5* pfam09704 profile that does not discriminate between subtypes of type I) were equally divided between the corresponding (sub)types. The "votes" were tallied across the unit; if the dominant (sub)type accounted for at least 2/3 of the total, the locus (unit) was assigned to the respective (sub)type. If no type or subtype received the qualified majority, the locus or unit was considered to be ambiguously classified.

## Sequence-based phylogenetic reconstruction

Multiple sequence alignments were constructed using a combination of MUSCLE [6] to align closely related sequences and MAFFT [7] to merge these alignments. Sites with of gap character fraction >0.5 and homogeneity <0.1 [8] were removed from the alignment. Phylogenetic trees were reconstructed using the FastTree program [9] with the WAG evolutionary model and the discrete gamma model with 20 rate categories. The same program was used for bootstrap value calculation.

For the phylogenetic analysis of the Cas1 family, 1418 Cas1 protein sequences were used. A filtered Cas1 alignment (306 positions) was used for tree reconstruction. For the phylogenetic analysis of the Cas3 family, 1093 protein sequences were used, and the filtered Cas3 alignment for tree reconstruction included 283 positions. For the Cas10 family, three alignments for three distinct families were constructed and used for phylogenetic analysis. These families consisted of 443, 36 and 10 protein sequences, and the respective filtered alignments included 427, 910 and 652 positions.

## Classification comparison and information consistency index

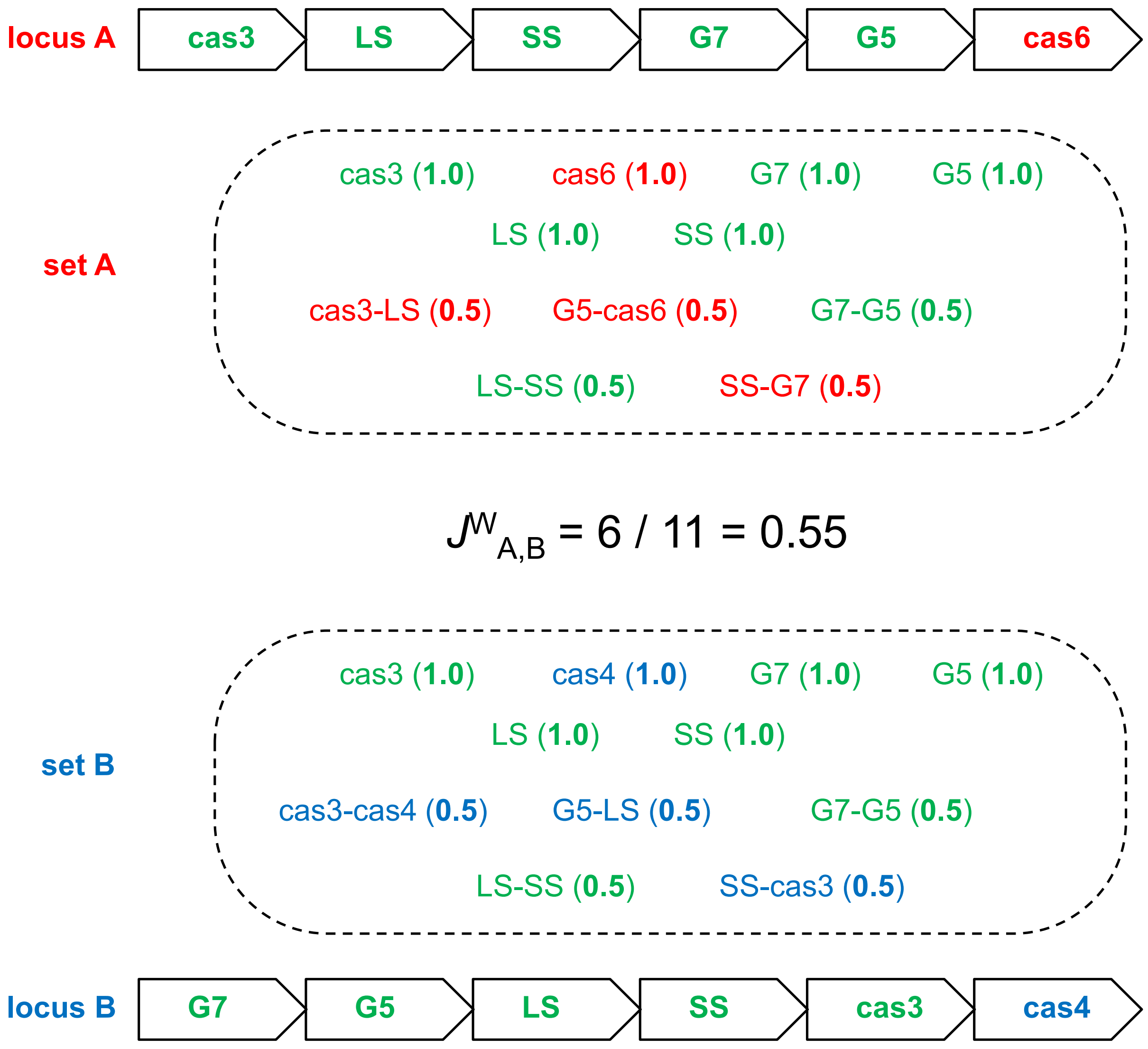
To compare classifications of CRISPR-*cas* loci based on different criteria (e.g. according to the CRISPR-Cas subtypes or according to the sequence classification of repeats in the adjacent CRISPR cassette), we used the Normalized Mutual Information index (mutual information divided by the geometric mean of the entropies of both classifications) [10].

To compare different trees reconstructed for the same set of leaves, the distances between the leaves along the tree branches were computed by summing the branch lengths along the path, connecting the leaves; then, the Spearman rank correlation coefficients between the distances induced by the two trees were calculated.

To quantify the fit between the tree structure and classification of the leaves, the following procedure was used. Within each clade of a rooted tree, the clade entropy was calculated from the distribution of its descendant leaves across the classes. Weighted average of clade entropies was calculated across the tree using clade weights, producing the tree-wide estimate of the classification entropy $E_T$. Then, the tree labels were scrambled, and the procedure was repeated 10 times to estimate the expectation of the tree entropy for the random labeling, $E_R$. The information consistency index then is calculated as $1\text{-}\min(E_T,E_R)/E_R$. A perfect tree that segregates at the root into clades corresponding to pure classes has $E_T$ equal to zero, and therefore, has an information consistency index of 1. The tree with the entropy as high as that of a tree with random leaf labeling (or higher) has the information consistency index of 0.

### Locus architecture dendrogram

To compare the architectures of the *cas* loci, the following procedure was developed. First, the gene order in the loci encoded in the negative strand was inverted. All non-*cas* genes were removed; *cas* genes were classified according to the family classification (Supplementary File 1). Each locus was encoded as a set of weighted components in the following way: all individual genes were included in the set with weights of 1; all ordered pairs of adjacent genes were included in the set with weights of 1/2 (see figure below). The weighted Jaccard similarity index $J^{W}_{A,B}$ for the component sets of loci ***A*** and ***B*** was computed as the sum of weights of the intersection of sets ***A*** and ***B*** divided by the sum of weights of the union of sets ***A*** and ***B***. The distance between the loci ***A*** and ***B*** was computed as $-\ln(J^{W}_{A,B})$. The figure below shows an example of the weighted Jaccard similarity index calculation.



The loci architecture similarity dendrogram was constructed from the pairwise loci distance matrix using the UPGMA method.

### Locus sequence similarity dendrogram

In order to automatically group *cas* loci, we introduce a clustering approach based on protein similarity. Given a *cas* locus, as defined in the Section *CRISPR-cas loci identification*, we select the interference proteins for Type I and Type III and Cas9 for Type II. Becuase some *cas* loci contain multiple effector modules of different types, the effector proteins of each locus were separated according to their types. For each pair of proteins $p_i$ and $p_j$ (belonging to different *cas* loci), the FASTA [11] protein sequence similarity score $S(p_i, p_j)$ was computed. To guarantee appropriate metric properties, the similarity was symmetrized to $S^*(p_i, p_j) = \left(S(p_i, p_j) + S(p_j, p_i)\right)/2$, and the score was normalized to

$$\hat{S}(p_i, p_j) = S^*(p_i, p_j) \Big/ \sqrt{S^*(p_i, p_i) S^*(p_j, p_j)}.$$

The similarity between two *cas* loci $L_m$ and $L_n$ is then defined as the average pairwise similarity between all possible protein pairings:

$$S_L(L_m, L_n) = \frac{1}{|L_m||\ L_n|} \sum_{p_i \in L_m} \sum_{p_j \in L_n} \hat{S}(p_i, p_j).$$

Finally, the dendrogram was generated by manually rooting the unrooted tree obtained by Rapid Neighbor-Joining [12] on the derived pairwise distance, $D_L = 1 - S_L$. The software and instructions for clustering CRISPR-cas loci by protein sequence similarity and automatic subtype assignment are available from http://www.bioinf.uni-freiburg.de/Supplements/NRMmicro_Koonin_2015/.

## References

1 Yutin, N., Puigbo, P., Koonin, E. V. & Wolf, Y. I. Phylogenomics of prokaryotic ribosomal proteins. *PLoS One* **7**, e36972, (2012).

2 Marchler-Bauer, A. *et al.* CDD: specific functional annotation with the Conserved Domain Database. *Nucleic Acids Res* **37**, D205-210 (2009).

3 Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389-3402 (1997).

4 Makarova, K. S. *et al.* Evolution and classification of the CRISPR-Cas systems. *Nat Rev Microbiol* **9**, 467-477, (2011).

5 Finn, R. D., Clements, J. & Eddy, S. R. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res* **39**, W29-37, (2011).

6 Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**, 1792-1797 (2004).

7 Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* **30**, 772-780, (2013).

8 Yutin, N., Makarova, K. S., Mekhedov, S. L., Wolf, Y. I. & Koonin, E. V. The deep archaeal roots of eukaryotes. *Mol Biol Evol* **25**, 1619-1630 (2008).

9 Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One* **5**, e9490, (2010).

10 Strehl, A. & Ghosh, J. Cluster Ensembles - A Knowledge Reuse Framework for Combining Multiple Partitions. *The Journal of Machine Learning Research* **3**, 583-617 (2002).

11 Pearson, W. Finding protein and nucleotide similarities with FASTA. Ch. 3, Unit 3.9 (2004).

12 Simonsen. M & Pedersen. Rapid computation of distance estimators from nucleotide and amino acid alignment. C.N.S. in *SAC2011* (2011).