Supplementary text 1: Distribution of Spacer Hits: Contributions of Microbial Diversity and Diversity of CRISPR-Cas systems

The entire set of identified CRISPR arrays, all the matching protospaces and, separately, the subset of virus protospacers were broken down by the CRISPR-Cas system subtype and by the microbial phyla (Table X1-X3). For all 88 phylum/subtype categories with at least 10 matches, the fraction of spacers with matches and the fraction of virus matches among all matches were calculated. Both of these fractions were analyzed as follows.

The calculated fraction $f_{i,j}$ for the spacers in phylum *i* and belonging to the category *j* was modeled as the product of a phylum-specific factor p_i and a subtype-specific factor s_j. The deviation of the observed and the predicted fractions was assumed to be log-normally distributed with the expectation of 1. Formally,

$$\log f_{i,j} = \log p_i + \log s_j + e_{i,j}$$

where $e_{i,j} \sim \text{Norm}(0,\sigma)$.

Optimization of vectors ***p*** and ***s*** with respect to Σe^2 provides both the phylum- and subtype-specific contribution factors and the residuals $e_{i,j}$ that allow one to estimate the quality of the fit between the model and the observation. Setting one of the vectors to an arbitrary positive value (e.g. 1) and optimizing the other provides a model with phylum- or subtype-only contributions; the null model replaces both vectors with mean($\log f_{i,j}$). The sums of squared residuals can be compared using the Fisher F-test.

Contribution of phylum- and subtype-specific factors to the fraction of spacers with matches

	phylum- and subtype	subtype-	subtype-	None
No. of data points	88	88	88	88
No. of parameters	39	21	18	0
No. of degrees of freedom	47	65	68	85
Σe^2	4.32	7.17	13.02	18.30

Contribution of phylum- and subtype-specific factors to the fraction of spacer matches to virus sequences

	phylum- and subtype	**subtype-**	**subtype-**	**None**
No. of data points	88	88	88	88

No. of parameters	39	21	18	0
No. of degrees of freedom	47	65	68	85
Σe^2	0.24	0.40	0.66	0.79

For both the overall fraction of matches and the fraction of viral matches, both phylum- and subtype-specific factors provide significant and independent contributions to the reduction of residuals (Fisher F-test p-value <0.05 except p=0.07 for the contribution of subtype-specific factors alone to the overall fraction of matches). Taken together, the two factors explain 70-75% of the original variance in the observed fraction.