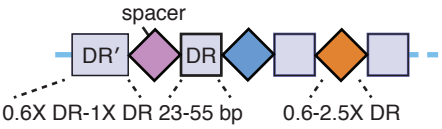
A

CRISPR system identification

Obtain bacterial and archaeal genome assemblies and scaffolds from NCBI WGS database

Identify interspaced repeat sequence locations

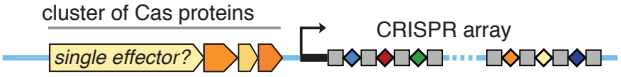


21175 'CRISPRs'

Identify putative proteins in 20kb window around CRISPRs

Eliminate candidate arrays with proteins associated with known CRISPR subtypes

11441 'CRISPRs'



Extract proteins >750 aa length and ± 5 proteins from CRISPR array

CRISPR effector identification

8977 proteins proximal to CRISPR array

all v. all BLASTp
single-linkage hierarchical clustering

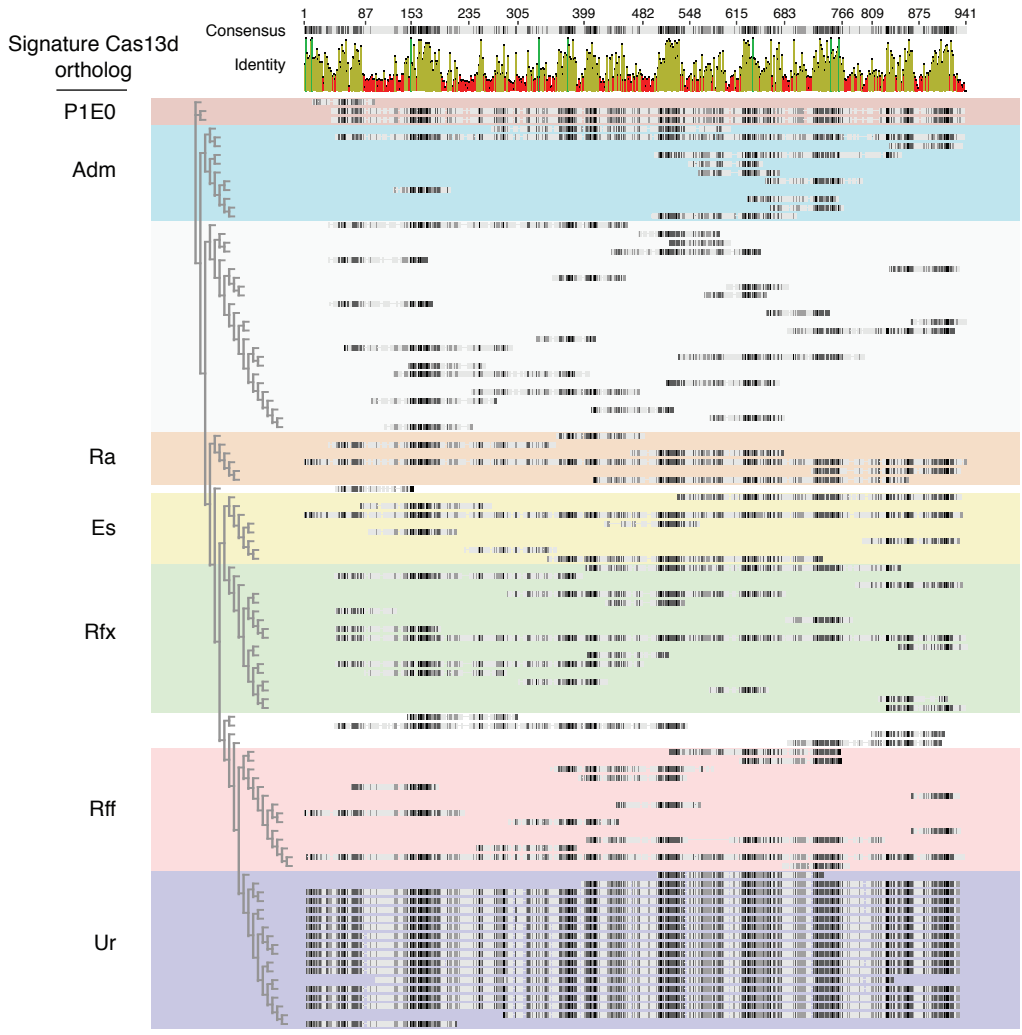
408 clusters

cluster expansion: BLASTp each cluster against all proteins (-nr database)

sort clusters by member protein proximity to CRISPR array (>70% co-occurrence)

candidate novel single effector CRISPR proteins

expand protein family via tblastn of metagenome contigs, custom sequencing of genomic DNA fragments

Type VI Cas13d family

B