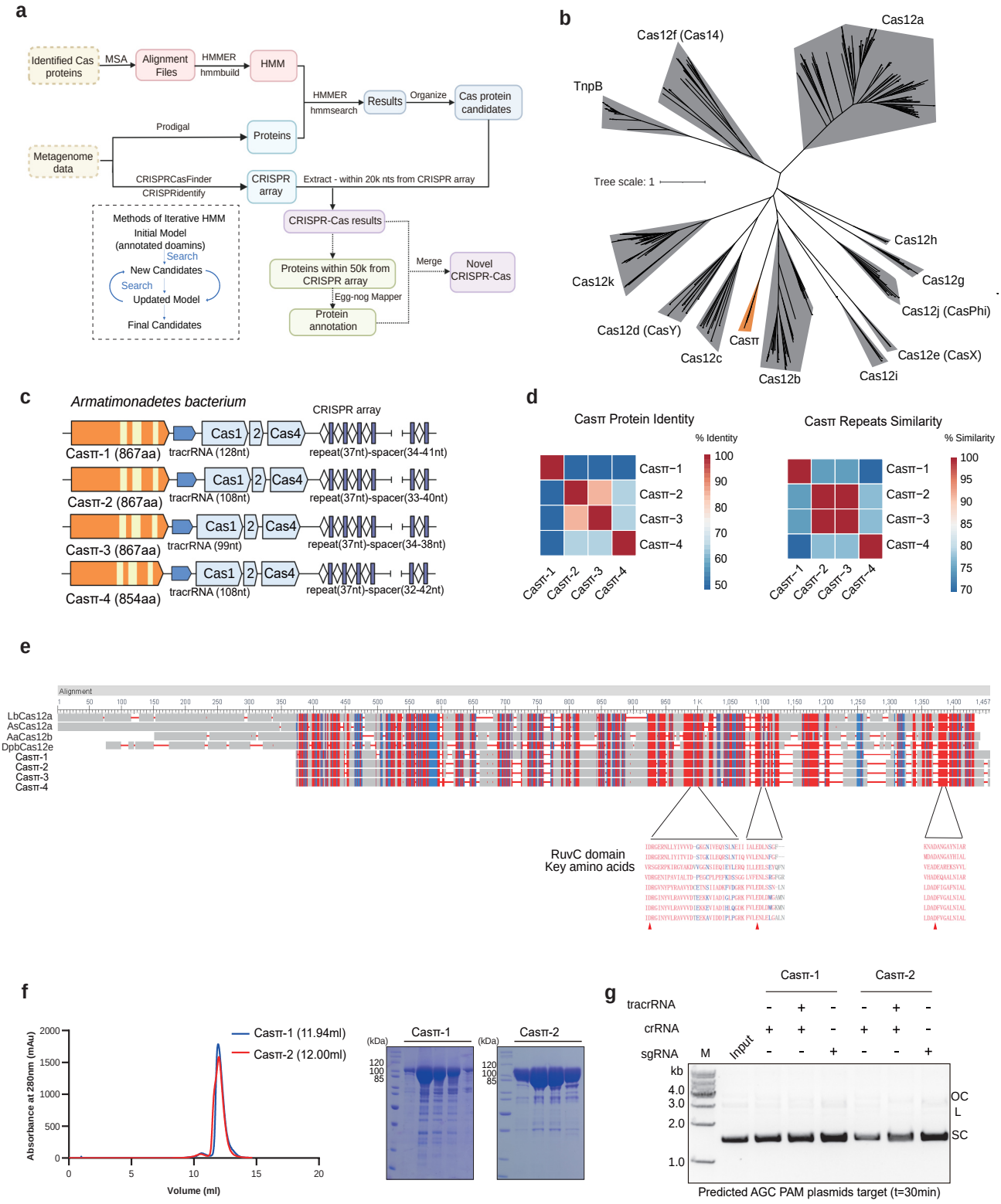
a
Identified Cas proteins
MSA
Alignment Files
HMMER
hmmbuild
HMM
HMMER
hmmsearch
Results
Organize
Cas protein candidates
Prodigal
Proteins
Metagenome data
CRISPRCasFinder
CRISPRidentify
CRISPR array
Extract - within 20k nts from CRISPR array
Methods of Iterative HMM
Initial Model
(annotated doamins)
Search
New Candidates
Search
Updated Model
Final Candidates
CRISPR-Cas results
Proteins within 50k from CRISPR array
Egg-nog Mapper
Protein annotation
Merge
Novel CRISPR-Cas
b
Cas12a
Cas12f (Cas14)
TnpB
Tree scale: 1
Cas12h
Cas12g
Cas12j (CasPhi)
Cas12k
Cas12d (CasY)
Casπ
Cas12c
Cas12b
Cas12e (CasX)
Cas12i
c
Armatimonadetes bacterium
CRISPR array
Cas1
2
Cas4
Casπ-1 (867aa)
tracrRNA (128nt)
repeat(37nt)-spacer(34-41nt)
Casπ-2 (867aa)
tracrRNA (108nt)
repeat(37nt)-spacer(33-40nt)
Casπ-3 (867aa)
tracrRNA (99nt)
repeat(37nt)-spacer(34-38nt)
Casπ-4 (854aa)
tracrRNA (108nt)
repeat(37nt)-spacer(32-42nt)
d
Casπ Protein Identity
% Identity
Casπ-1
Casπ-2
Casπ-3
Casπ-4
Casπ Repeats Similarity
% Similarity
e
Alignment
LbCas12a
AsCas12a
AaCas12b
DpbCas12e
Casπ-1
Casπ-2
Casπ-3
Casπ-4
RuvC domain
Key amino acids
f
Absorbance at 280nm (mAu)
Volume (ml)
Casπ-1 (11.94ml)
Casπ-2 (12.00ml)
(kDa)
Casπ-1
Casπ-2
g
Casπ-1
Casπ-2
tracrRNA
crRNA
sgRNA
M
Input
kb
OC
L
SC
Predicted AGC PAM plasmids target (t=30min)

Fig. S1 Casπ contains an active RuvC domain for DNA cleavage.

(**a**) Bioinformatics pipeline of identifying novel CRISPR-Cas systems. In general, the prediction of CRISPR-Cas systems is based on the identification of CRISPR arrays and Cas proteins. The architecture of CRISPR array is relatively conserved and can be predicted by public software, such as CRISPRCasFinder and CRISPRidentify. Meanwhile, Cas proteins can be predicted by canonical protein homology search tools, such as HMMER and Pfam based on Hidden Markov models (HMM). However, these existed tools can only identify 'new' proteins which share relatively high sequence similarity to annotated Cas proteins. Therefore, we develop an unbiased workflow using iterative HMM to identify novel Cas candidates. For example, to identify novel type V Cas nucleases, RuvC domains from annotated Cas12a nucleases in the Pfam database (Pfam accession number: PF18516) were used to build an initial HMM for the first round searching of RuvC-containing proteins in all accessible databases. 1,349 protein candidates within 300 to 1600 amino acids (un-annotated in Pfam) were found, and all 'RuvC' domains within those candidates are used to build an improved HMM for the next round search of new candidates in all accessible databases. Using this iterative pipeline, the HMM was continuously updated with new candidates found in the current round and yielded to an updated pool of Cas-nuclease candidates from the next round search. A typical 2~10 rounds of iterative HMM and search were used for generating the final pool of novel type V Cas nucleases. Further, we filtered the candidates in the final pool by their functions annotated by eggNOG mapper, similarities to the final HMM (E-value), sizes and the confidence of putative CRISPR arrays for manual screening. Particularly in this study, Casπ showed up in the candidate pool with three rounds of iterative HMM and search.

(**b**) Maximum likelihood phylogenic analysis of Casπ with type V subtypes a-k. Casπ proteins are outlined in orange, with other subtypes in grey. Bootstrap=1500, Casπ protein sequences are shown in Table S1.

(**c**) Genomic architectures of all four CRISPR-Casπ systems.

(**d**) Similarity matrix built and visualized using heatmap for Casπ protein sequences (left) and repeat sequences (right).

(**e**) Multiple sequence alignment of Casπ with LbCas12a, AsCas12a, AaCas12b and DpbCas12e. Key amino acids of RuvC domain are marked by red arrows.

(**f**) Purification of both Casπ proteins. The SEC curves were aligned referring to the injection volume. The peak fractions were analyzed by SDS-PAGE.

(**g**) *In vitro* cleavage of plasmids containing predicted AGC PAM by crRNA only, tracrRNA and crRNA pair or sgRNA by Casπ effectors (SC, supercoiled plasmids; L, linearized plasmids; OC, open-circle plasmids).