BMC Evolutionary Biology

RESEARCH ARTICLE Open Access

Phylogenomics of Cas4 family nucleases

Sanjarbek Hudaiberdiev[1], Sergey Shmakov[1,2], Yuri I. Wolf[1], Michael P. Terns[3], Kira S. Makarova[1] and Eugene V. Koonin[1*]

Abstract

Background: The Cas4 family endonuclease is a component of the adaptation module in many variants of CRISPR-Cas adaptive immunity systems. Unlike most of the other Cas proteins, Cas4 is often encoded outside CRISPR-*cas* loci (solo-Cas4) and is also found in mobile genetic elements (MGE-Cas4).

Results: As part of our ongoing investigation of CRISPR-Cas evolution, we explored the phylogenomics of the Cas4 family. About 90% of the archaeal genomes encode Cas4 compared to only about 20% of the bacterial genomes. Many archaea encode both the CRISPR-associated form (CAS-Cas4) and solo-Cas4, whereas in bacteria, this combination is extremely rare. The solo-*cas4* genes are over-represented in environmental bacteria and archaea with small genomes that typically lack CRISPR-Cas, suggesting that Cas4 could perform uncharacterized defense or repair functions in these microbes. Phylogenomic analysis indicates that both the CRISPR-associated *cas4* genes are often transferred horizontally but almost exclusively, as part of the adaptation module. The evolutionary integrity of the adaptation module sharply contrasts the rampant shuffling of CRISPR-*cas* modules whereby a given variant of the adaptation module can combine with virtually any effector module. The solo-*cas4* genes evolve primarily via vertical inheritance and are subject only to occasional horizontal transfer. The selection pressure on *cas4* genes does not substantially differ between CAS-Cas4 and solo-*cas4*, and is close to the genomic median. Thus, *cas4* genes, similarly to *cas1* and *cas2*, evolve similarly to 'regular' microbial genes involved in various cellular functions, showing no evidence of direct involvement in virus-host arms races. A notable feature of the Cas4 family evolution is the frequent recruitment of *cas4* genes by various mobile genetic elements (MGE), particularly, archaeal viruses. The functions of Cas4 in these elements are unknown and potentially might involve anti-defense roles.

Conclusions: Unlike most of the other Cas proteins, Cas4 family members are as often encoded by stand-alone genes as they are incorporated in CRISPR-Cas systems. In addition, *cas4* genes were repeatedly recruited by MGE, perhaps, for anti-defense functions. Experimental characterization of the solo and MGE-encoded Cas4 nucleases is expected to reveal currently uncharacterized defense and anti-defense systems and their interactions with CRISPR-Cas systems.

Background

Cas4 is one of the core CRISPR-associated (Cas) proteins, which is implicated in the adaptation phase in several subtypes of CRISPR-Cas systems [1–4]. During this phase of the CRISPR response, a segment of invading DNA (protospacer), usually from a virus or plasmid, is selected, typically, based on the presence of a protospacer adjacent motif (PAM) [5, 6]. The protospacer then is incorporated by the Cas1-Cas2 adaptation complex into a CRISPR array, most often, next to the first, 5′-terminal repeat. The array is transcribed into a single, long pre-crRNA, which is subsequently processed into mature CRISPR (cr)RNAs and is incorporated into the effector complex or a single effector protein, which scans invading DNA and, once a match is found, recruits nuclease domain-containing Cas proteins to cleave the target DNA [7–10]. So far the *cas4* gene has been identified in the subtypes I-A, B, C, D, U of Class 1 CRISPR-Cas systems and subtypes II-B and V-A, B, E of Class 2. In each of the respective loci, the *cas4* gene is adjacent to *cas1* and *cas2*, and in some cases, fused with *cas1*, which is compatible with the involvement of Cas4 in adaptation [11, 12]. Furthermore, it has been shown that in subtype I-A loci, *cas1*, *cas2* and *cas4* genes form a single operon [1], and for subtype I-B, the requirement of Cas4 for adaptation has been demonstrated in a direct experiment [3]. A recent study on the cas gene requirement for

* Correspondence: koonin@ncbi.nlm.nih.gov
[1]National Center for Biotechnology Information, National Institutes of Health, Bethesda, MD, USA
Full list of author information is available at the end of the article

adaptation by the type I-A system of the crenarchaeon *Sulfolobus islandicus* has shown that two paralogous *cas4* genes (one denoted *csa1*) in the *cas* operon are both essential, suggesting that these proteins could form a complex in which the two paralogs would perform unique functions [13].

Cas4 is a DNA exonuclease that contains a nuclease domain and a Fe-S cluster-binding module [2, 14–16] and is homologous to the nuclease moieties of well-characterized proteins involved in recombination and repair in bacteria (RecB, AddB) and eukaryotes (Dna2) [17–21]. The Cas4 nuclease belongs to the expansive PD-DEXK phosphodiesterase superfamily, named after the conserved catalytic motif present in most members [21]. Similar to the AddB nuclease, Cas4 is thought to form recombinogenic 3′ ssDNA overhangs in the protospacers, thus facilitating their subsequent incorporation into the CRISPR array as dsDNA spacers [16]. In the structures that have been solved for Cas4 proteins from *Pyrobaculum calidifontis* (Pcal_0546; pdb: 4R5Q) [15] and *Sulfolobus solfataricus* (SSO0001, pdb: 4IC1) [2], Pcal_0546 is a monomer, whereas SSO0001 forms a toroidal decamer composed of 5 tightly packed dimers. The Pcal_0546 protein has been shown to exhibit a metal-dependent 5′ to 3′ exonuclease activity against ssDNA substrates [15], whereas the Cas4 protein SSO1391 from *S. solfataricus* has been reported to cleave ssDNA in both the 5′ to 3′ and the 3′ to 5′ directions [2]. However, in an independent study, only the 5′-3′ exonuclease activity has been detected for SSO1391 as well [16].

Although the *cas4* gene is usually located next to *cas1* or *cas2* genes and, in some CRISPR-Cas systems, is fused to *cas1* (eg. subtypes I-U and V-B), many bacterial and archaeal genomes encompass an additional *cas4* gene that is not associated with CRISPR-*cas* loci (hereinafter, solo-Cas4). Notably, solo-cas4 genes are present in many archaeal and bacterial genomes that lack CRISPR-*cas* loci [11].

The recent discovery of Cas4 homologs encoded in bacteriophage genomes has added a new twist to the functional repertoire of the Cas4 family. It has been shown that a *Campylobacter* phage that encodes a Cas4-like protein, employs this protein to stimulate acquisition of host-derived spacers by the *Campylobacter* type II-C CRISPR-Cas system (which lacks *cas4*). The phage appears to use this mechanism to evade host immunity via a mechanism that remains to be characterized [22, 23].

Cas4 homologs have been identified also in archaeal viruses [24] and casposons, self-synthesizing transposons that employ Cas1 homologs as recombinases and are thought to be the ancestors of the CRISPR-Cas adaptation modules [25, 26]. One archaeal virus, Thermoproteus tenax virus 1 (TTV1), encodes an inactivated derivative of a Cas4-like nuclease that became a component of the viral nucleocapsid [24].

Despite the current keen interest in the mechanisms, diversity and evolution of the CRISPR-Cas systems, no comprehensive analysis of the Cas4 family has been reported so far. The aim of this work is to present such an analysis with the focus on *cas4* gene neighborhoods and comparison of the (predicted) functions and evolutionary regimes of CRISPR-associated Cas4 and solo-Cas4 subfamilies.

Results and discussion

Phylogenomic analysis of the Cas4 protein family

As reported previously, the majority of the Cas4 proteins belong to two families, namely, COG1468 and COG4343; the latter is specific for subtype I-A CRISPR-Cas systems and is also known as Csa1 (these two COGs correspond to pfam01930/DUF83 and pfam06023/ DUF911, respectively) [11]. In addition, Cas4-like nucleases of pfam12705/ PDDEXK_1 are occasionally found in association with *cas* loci and have been shown to affect spacer adaptation by certain CRISPR-Cas systems [22, 23]. In the present analysis, we focused primarily on these three major families of Cas4 homologs and disregarded a few other, less common PD-DExK families that also have been reported in the vicinity of several CRISPR-*cas* loci [11].

To collect protein sequences of the Cas4 family, we screened 48,599 prokaryotic genomes (Genbank, March 2016) using PSI-BLAST search [27] with the sequence profiles related to the three families of Cas4 homologs. The results of this search were filtered to exclude sequences that produced better scores to other profiles in the CDD database (See Material and Methods for details). Additionally, we searched 2996 phage genomes from the PHANTOM database [28]. In total, 7060 protein sequences were retrieved, of which 883 were from complete bacterial and archaeal genomes, and 272 were from bacterial or archaeal viruses.

Inspection of the multiple alignment of Cas4 proteins (see Materials and Methods) shows conservation of the catalytic motifs of the PD-DEXK phosphodiesterase superfamily, suggesting that all Cas4 proteins are active nucleases. All members of the Cas4 family contain a Fe-S-cluster-binding module with 4 conserved cysteines. It has been shown that different Cas4 protein bind either [2Fe-2S] or [4Fe-4S] clusters [2, 15]; however, no consistent difference in the patterns of sequence conservation was detected between proteins binding these distinct ligands.

We then constructed a sequence similarity dendrogram (see Materials and Method for details) for the 7060 identified Cas4 family proteins and mapped the gene context to each branch, including assignments to CRISPR-Cas subtypes (Additional file 1: Figure S1). As expected, the sequences that belong to distinct pfam entries and COGs do not overlap in the tree, and pfam12705 contained no major branches associated with CRISPR-Cas systems; thus, hereinafter, we only briefly discuss this family. The remaining Cas4 homologs could be classified into three major groups based on the genomic context: 1) CRISPR-Cas-associated Cas4 (CAS-Cas4), 2) Cas4 associated with mobile genetic elements (MGE-Cas4), and 3) solo-Cas4 (i.e. *cas4* genes located outside functionally characterized genomic contexts). The CAS-*cas4* genes are embedded in several typical arrangements of the adaptation module that also include the *cas1* and *cas2* genes, and are often shared between several CRISPR-Cas subtypes (Fig. 1). As reported previously, Cas1 and Cas4 proteins are fused in some systems (I-B, I-U and V-A), whereas I-A is the only CRISPR-Cas subtype that typically contains two *cas4-like* genes from two distinct subfamilies within the adaptation module (Fig. 1). One of the conserved arrangements shared by some of I-B and I-D systems also includes the *cas6* gene, which is not known to be involved in adaptation, but has been identified in several fusions with Cas1 and reverse transcriptase domains in some type III CRISPR-Cas systems (Fig. 1) [29, 30].

We further compared the phyletic distributions of the CAS-Cas4 and solo-Cas4 in complete microbial genomes (Fig. 2a). This analysis revealed pronounced differences between the distributions of the Cas4 family members in archaea and bacteria. The Cas4 family is represented in about 90% of the archaeal genomes, and about 50% of these encode both CAS-Cas4 and solo-Cas4 (Additional file 4: Information File S2). We further compared the representation of solo-Cas4 and distinct CAS-Cas4 contexts for the complete genome set, including draft genomes, counting one strain per species for each category to avoid sequencing bias (Fig. 2b). Bacteria show a patchy distribution of the Cas4 family genes, with a smaller fraction of solo-Cas4 compared to archaea (Fig. 2b). Numerous archaeal genomes encode both CAS-Cas4 and solo-Cas4 but the frequency of this combination is close to the product of individual frequencies of each form, thus showing no evidence of a functional connection. Notably, most members of the DPANN superphylum (a recently discovered major, apparently monophyletic group of archaea that includes *Aenigmarchaeota, Diapherotrites, Micrarchaeota, Nanohaloarchaeota, Pacearchaeota, Parvarchaeota*, and *Woesearchaeota*), largely symbiotic archaea with small genomes [31, 32], encode solo-Cas4, but not CAS-Cas4 (Additional file 7: Table S1). Within the entire genomic set, which includes draft genomes, CAS-Cas4 is most often present in type I-B and I-C CRISPR-Cas loci (Fig. 2b). We also observed many cases of CAS-Cas4 being a component of stand-alone (not associated with effector modules) adaptation modules, which are present in about 12% of the archaea and 4% of the bacteria that possess at least one *cas4* gene (Fig. 2b). Conceivably, the high prevalence of Cas4 in archaea is due to the dominance of I-B and I-C systems as opposed to I-E and I-F that are most

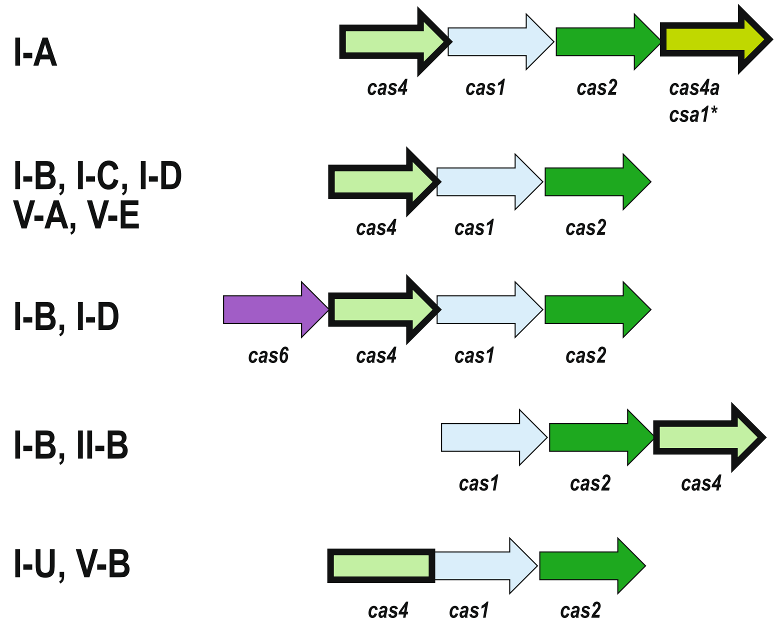

Fig. 1 Typical organization of adaptation modules that include *cas4* genes. Most common organizations of adaptation modules containing *cas4* gene are shown. Genes are shown by block arrows according to the transcription direction and are not to scale. Homologous genes are color-coded and identified by a systematic name and a legacy name (indicated by an asterisk). Cas4 family genes are indicated by thick outline. CRISPR-Cas system subtypes found in association with each module are indicated. Notably, the adaptation modules of the I-B and I-D loci include *cas6* genes

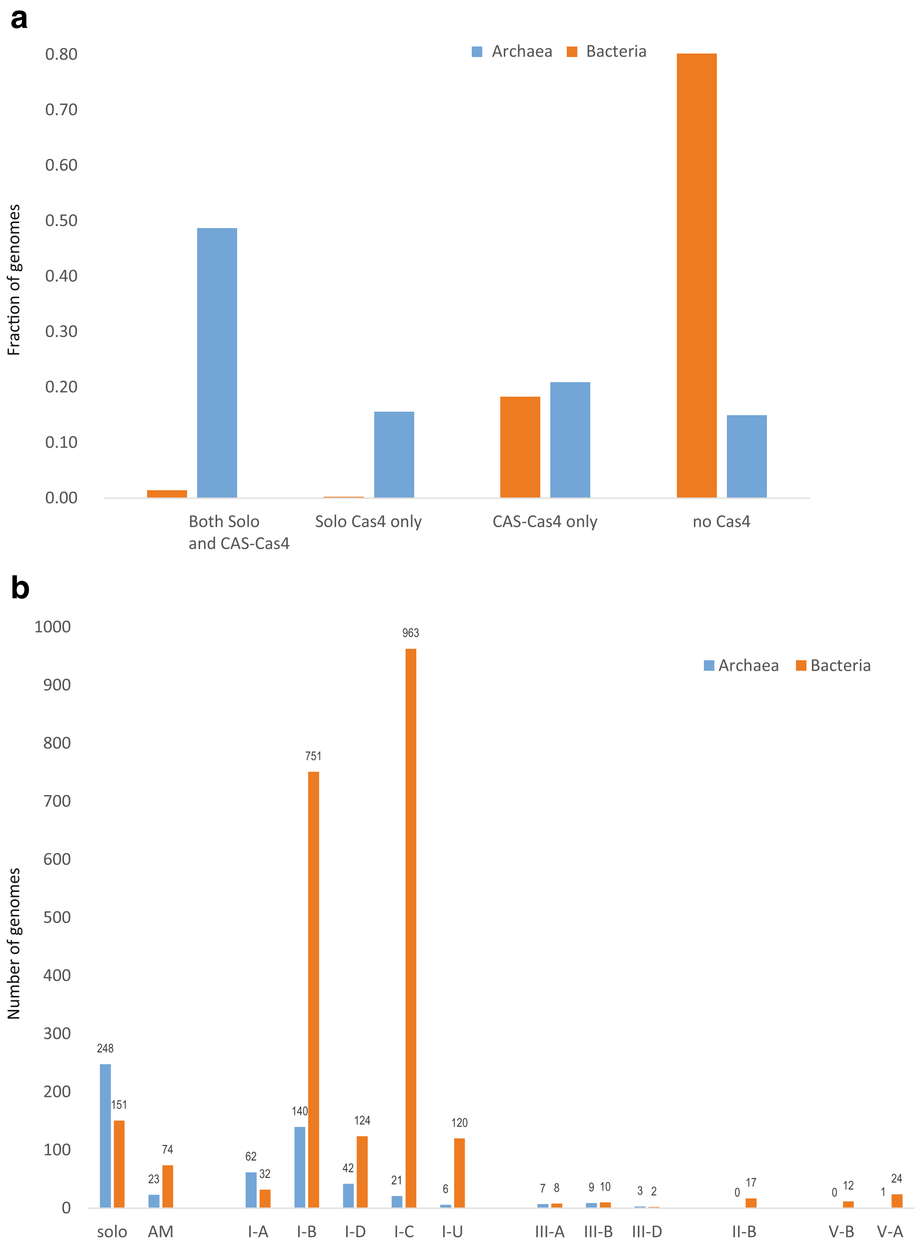

Fig. 2 Representation of different Cas4 groups in archaea and bacteria. **a**. Presence and absence of CAS-Cas4 and Cas4 solo in completely sequenced genomes of archaea and bacteria. Individual genomes were assigned weights inversely proportional to the number of genomes from each species; the presence fraction was calculated with genome weights taken into account (see Additional file 4: File S2). **b**. Numbers of species in the complete data set containing Cas4 from different CRISPR-Cas4 subtypes and Cas4 solo. The counts were obtained from the complete set of CAS-Cas4 and Cas4-solo (5571 and 702, respectively). Only one strain per species was counted for each respective category of Cas4 assignments to avoid sequencing bias. See Additional file 7: Table S1 for details. Abbreviation: AM, stand-alone adaptation module

common in bacteria and lack the *cas4* gene [11]. The biological underpinnings of this distinct distribution of Cas4-containing CRISPR-Cas remains obscure.

Phylogenetic analysis indicates that the majority of the CAS-Cas4 proteins form clades that are compatible with the classification of the CRISPR-Cas subtypes delineated through comparison of effector gene sets. However, there are many exceptions to this congruence, so that, in the phylogenetic tree of Cas4, only subtypes I-C, II-B, and V-A are both monophyletic and compact (i.e. do not include representatives of other subtypes) (Fig. 3, Additional file 1: Figure S1, and Additional file 3: Information File S1). In agreement with the previous phylogenetic analysis of Cas1, the subtype II-B-associated Cas4 is nested within the I-B/I-A branch, whereas the subtype V-B-associated Cas4 proteins, although not forming a clade, fall within the I-U branch (Additional file 1: Figure S1, Additional file 3: Information File S1) [33] (Fig. 3a, Additional file 1: Figure S1, Additional file 3: Information File S1). Thus, the Cas4-containing adaptation modules in subtypes II-B and V-B apparently were acquired from the respective type I systems. In contrast, the subtype V-A-associated Cas4 is distinct and does not show clear affinity to any known CRISPR-Cas systems types (Additional file 1: Figure S1, Additional file 3: Information File S1). This observation is in agreement with the long, deep V-A branch in the respective Cas1 phylogeny [33].

Subtypes I-A, I-B, I-D and to a lesser extent I-U, are scattered among many strongly supported Cas4 branches suggesting extensive shuffling of the adaptation and effector modules (Fig. 3b-f, Additional file 1: Figure S1, Additional file 3: Information File S1). This evidence of module shuffling is consistent with previous observations from phylogenetic analysis of Cas1, although this effect has been underestimated and was not carefully explored. With many more genomes now available for analysis, it becomes apparent that adaptation modules of subtypes I-A, I-B and I-D can function with (almost) any of the respective effector modules (Fig. 3, Additional file 1: Figure S1, Additional file 3: Information File S1). Furthermore, these adaptation modules apparently can combine also with type III and type IV effector modules (Fig. 3, Additional file 1: Figure S1, Additional file 3: Information File S1) because some branches within well-supported Cas4 clades correspond to these CRISPR-Cas types (Fig. 3, Additional file 1: Figure S1, Additional file 3: Information File S1).

We examined in greater detail several cases where highly similar *cas4* genes were associated with different CRISPR-Cas system (sub)types, especially in closely related species or strains, where *cas4* genes could be expected to be vertically inherited (Fig. 4). Analysis of these cases suggests that adaptation module genes remain untouched by recombination, whereas effector genes replace the ancestral effector modules. This inference of the preferred directionality of recombination events is supported, in particular, by the presence of flanking syntenic regions (Fig. 4). These observations prompted us to search for similar exchanges within the same CRISPR-Cas subtype. To this end, we clustered the large subunits of the respective systems (Cas8 for type I-A, I-U and I-B and Cas10d for I-D) with permissive cutoffs and mapped these clusters on the tree in order to identify cases where closely related *cas4* genes are co-located with substantially different groups of large subunits of the same subtype. This procedure allowed us to identify additional cases where effector module genes are replaced, whereas the adaptation module remained unaffected (Fig. 5). In many of these cases, flanking syntenic regions allow identification of the precise boundaries of the effector gene replacement sequences, even on the nucleotide level (Additional file 2: Figure S2). We also noticed some cases when the effector genes are deleted, resulting in emergence of stand-alone adaptation modules that are usually located inside the branch for the respective subtypes and were identified in many archaea and bacteria (Fig. 2b, Additional file 1: Figure S1, Additional file 3: Information File S1).

Solo-Cas4 can be clearly differentiated from CAS-Cas4 in the phylogenetic tree where they form several distinct branches that are mostly consistent with the order-level taxonomy of archaea, suggesting predominant vertical inheritance (Additional file 1: Figure S1, Additional file 3: Information File S1). Despite the general trend to retain solo-Cas4 in the genome, in some archaeal lineages, this gene apparently evolved fast, so that, for example, two groups of *Thermoproteales, Pyrobaculum/Thermoproteus* and *Vulcanisaeta*, encode dissimilar solo-Cas4 sequences that do not cluster with each other or with homologs from other archaea. Several members of the archaeal DPANN superphylum that mostly lack CRISPR-Cas systems, encode two distinct solo-Cas4 proteins (Additional file 1: Figure S1, Additional file 3: Information File S1). In bacteria, solo-Cas4 shows a patchy distribution, with some notable exceptions. First, solo-Cas4 is present and monophyletic in the majority of *Thermoanaerobacteriales* and many *Chloroflexi.* Second, solo-Cas4 is encoded in many genomes of the Candidate Phyla Radiation (CPR), a recently described assemblage of bacteria with small genomes and many unusual features, such as *Parcubacteria, Nomurabacteria, Roizmanbacteria* and others [34, 35]. Some of the solo-Cas4 proteins from the CPR confidently group with homologs from the archaeal DPANN superphylum (Additional file 1: Figure S1, Additional file 3: Information File S1). Thus, solo-Cas4 is clearly over-represented in environmental bacteria and archaea with small genomes and could perform important biological functions in these organisms that remain to be characterized. Furthermore, dissemination of solo-Cas4 genes via HGT between bacteria and archaea appears likely. In several of the DPANN archaeal genomes, the gene context, namely the adjacency to genes coding for restriction-modification systems, implies that solo-Cas4 functions in so far uncharacterized defense pathways (Additional file 7: Table S1).

Cas4 in mobile genetic elements (MGE-Cas4)

Phylogenetic analysis of the Cas4 family identified several distinct branches that are associated with MGE and viruses, especially, archaeal ones; to our knowledge, the spread of Cas4 homologs in MGE has not been systematically described previously. In particular, a *cas4* gene is present in most of the known viruses of Desulfurococcales and Sulfolobales (Additional file 1:

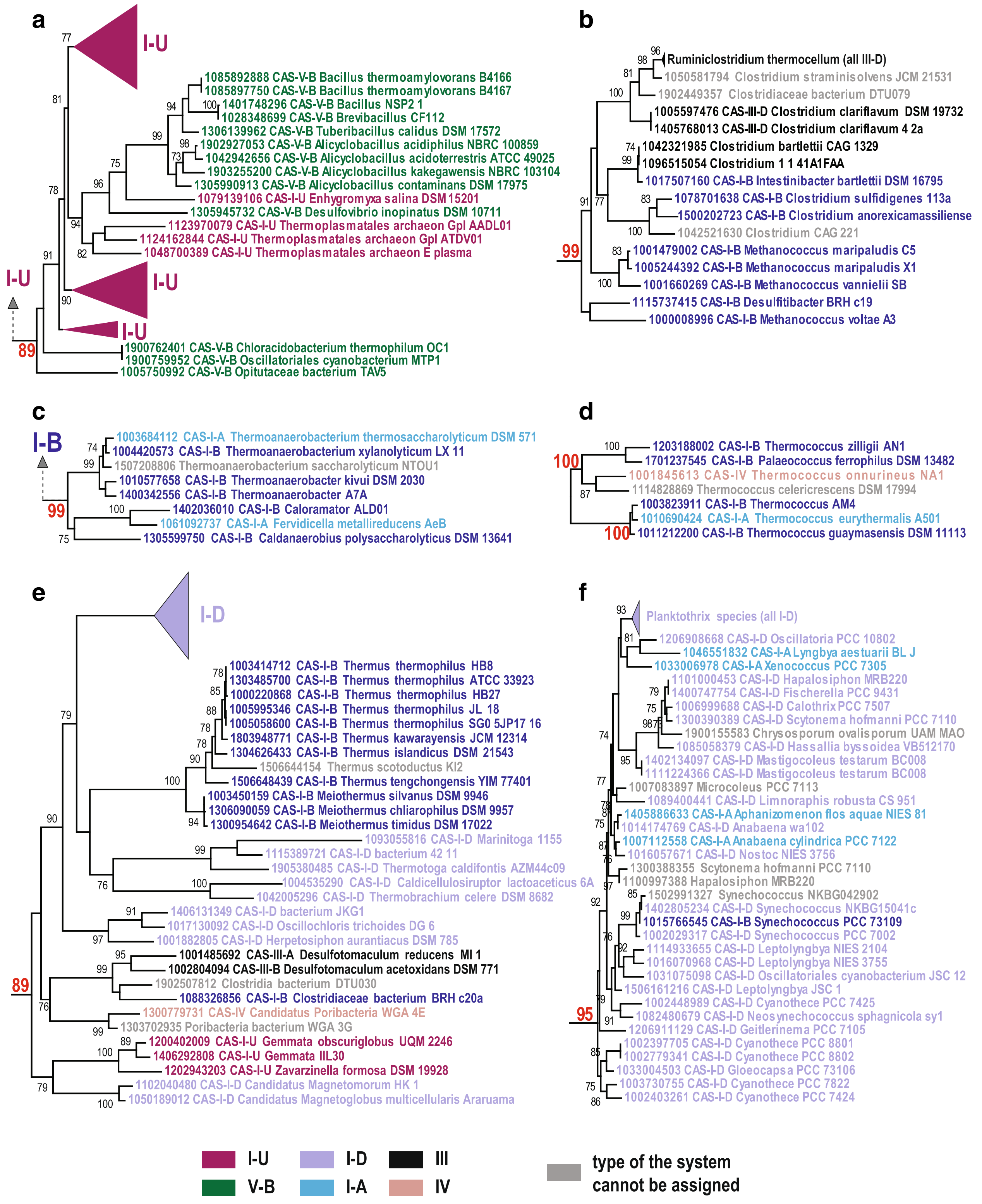
a
I-U
1085892888 CAS-V-B Bacillus thermoamylovorans B4166
1085897750 CAS-V-B Bacillus thermoamylovorans B4167
1401748296 CAS-V-B Bacillus NSP2 1
1028348699 CAS-V-B Brevibacillus CF112
1306139962 CAS-V-B Tuberibacillus calidus DSM 17572
1902927053 CAS-V-B Alicyclobacillus acidiphilus NBRC 100859
1042942656 CAS-V-B Alicyclobacillus acidoterrestris ATCC 49025
1903255200 CAS-V-B Alicyclobacillus kakegawensis NBRC 103104
1305990913 CAS-V-B Alicyclobacillus contaminans DSM 17975
1079139106 CAS-I-U Enhygromyxa salina DSM15201
1305945732 CAS-V-B Desulfovibrio inopinatus DSM 10711
1123970079 CAS-I-U Thermoplasmatales archaeon Gpl AADL01
1124162844 CAS-I-U Thermoplasmatales archaeon Gpl ATDV01
1048700389 CAS-I-U Thermoplasmatales archaeon E plasma
I-U
I-U
I-U
89
1900762401 CAS-V-B Chloracidobacterium thermophilum OC1
1900759952 CAS-V-B Oscillatoriales cyanobacterium MTP1
1005750992 CAS-V-B Opitutaceae bacterium TAV5
b
Ruminiclostridium thermocellum (all III-D)
1050581794 Clostridium straminisolvens JCM 21531
1902449357 Clostridiaceae bacterium DTU079
1005597476 CAS-III-D Clostridium clariflavum DSM 19732
1405768013 CAS-III-D Clostridium clariflavum 4 2a
1042321985 Clostridium bartlettii CAG 1329
1096515054 Clostridium 1 1 41A1FAA
1017507160 CAS-I-B Intestinibacter bartlettii DSM 16795
1078701638 CAS-I-B Clostridium sulfidigenes 113a
1500202723 CAS-I-B Clostridium anorexicamassiliense
1042521630 Clostridium CAG 221
99
1001479002 CAS-I-B Methanococcus maripaludis C5
1005244392 CAS-I-B Methanococcus maripaludis X1
1001660269 CAS-I-B Methanococcus vannielii SB
1115737415 CAS-I-B Desulfitibacter BRH c19
1000008996 CAS-I-B Methanococcus voltae A3
c
I-B
1003684112 CAS-I-A Thermoanaerobacterium thermosaccharolyticum DSM 571
1004420573 CAS-I-B Thermoanaerobacterium xylanolyticum LX 11
1507208806 Thermoanaerobacterium saccharolyticum NTOU1
1010577658 CAS-I-B Thermoanaerobacter kivui DSM 2030
1400342556 CAS-I-B Thermoanaerobacter A7A
99
1402036010 CAS-I-B Caloramator ALD01
1061092737 CAS-I-A Fervidicella metallireducens AeB
1305599750 CAS-I-B Caldanaerobius polysaccharolyticus DSM 13641
d
1203188002 CAS-I-B Thermococcus zilligii AN1
100
1701237545 CAS-I-B Palaeococcus ferrophilus DSM 13482
1001845613 CAS-IV Thermococcus onnurineus NA1
1114828869 Thermococcus celericrescens DSM 17994
1003823911 CAS-I-B Thermococcus AM4
1010690424 CAS-I-A Thermococcus eurythermalis A501
100
1011212200 CAS-I-B Thermococcus guaymasensis DSM 11113
e
I-D
1003414712 CAS-I-B Thermus thermophilus HB8
1303485700 CAS-I-B Thermus thermophilus ATCC 33923
1000220868 CAS-I-B Thermus thermophilus HB27
1005995346 CAS-I-B Thermus thermophilus JL 18
1005058600 CAS-I-B Thermus thermophilus SG0 5JP17 16
1803948771 CAS-I-B Thermus kawarayensis JCM 12314
1304626433 CAS-I-B Thermus islandicus DSM 21543
1506644154 Thermus scotoductus KI2
1506648439 CAS-I-B Thermus tengchongensis YIM 77401
1003450159 CAS-I-B Meiothermus silvanus DSM 9946
1306090059 CAS-I-B Meiothermus chliarophilus DSM 9957
1300954642 CAS-I-B Meiothermus timidus DSM 17022
1093055816 CAS-I-D Marinitoga 1155
1115389721 CAS-I-D bacterium 42 11
1905380485 CAS-I-D Thermotoga caldifontis AZM44c09
1004535290 CAS-I-D Caldicellulosiruptor lactoaceticus 6A
1042005296 CAS-I-D Thermobrachium celere DSM 8682
1406131349 CAS-I-D bacterium JKG1
1017130092 CAS-I-D Oscillochloris trichoides DG 6
1001882805 CAS-I-D Herpetosiphon aurantiacus DSM 785
1001485692 CAS-III-A Desulfotomaculum reducens MI 1
1002804094 CAS-III-B Desulfotomaculum acetoxidans DSM 771
89
1902507812 Clostridia bacterium DTU030
1088326856 CAS-I-B Clostridiaceae bacterium BRH c20a
1300779731 CAS-IV Candidatus Poribacteria WGA 4E
1303702935 Poribacteria bacterium WGA 3G
1200402009 CAS-I-U Gemmata obscuriglobus UQM 2246
1406292808 CAS-I-U Gemmata IIL30
1202943203 CAS-I-U Zavarzinella formosa DSM 19928
1102040480 CAS-I-D Candidatus Magnetomorum HK 1
1050189012 CAS-I-D Candidatus Magnetoglobus multicellularis Araruama
f
Planktothrix species (all I-D)
1206908668 CAS-I-D Oscillatoria PCC 10802
1046551832 CAS-I-A Lyngbya aestuarii BL J
1033006978 CAS-I-A Xenococcus PCC 7305
1101000453 CAS-I-D Hapalosiphon MRB220
1400747754 CAS-I-D Fischerella PCC 9431
1006999688 CAS-I-D Calothrix PCC 7507
1300390389 CAS-I-D Scytonema hofmanni PCC 7110
1900155583 Chrysosporum ovalisporum UAM MAO
1085058379 CAS-I-D Hassallia byssoidea VB512170
1402134097 CAS-I-D Mastigocoleus testarum BC008
1111224366 CAS-I-D Mastigocoleus testarum BC008
1007083897 Microcoleus PCC 7113
1089400441 CAS-I-D Limnoraphis robusta CS 951
1405886633 CAS-I-A Aphanizomenon flos aquae NIES 81
1014174769 CAS-I-D Anabaena wa102
1007112558 CAS-I-A Anabaena cylindrica PCC 7122
1016057671 CAS-I-D Nostoc NIES 3756
1300388355 Scytonema hofmanni PCC 7110
1100997388 Hapalosiphon MRB220
1502991327 Synechococcus NKBG042902
1402805234 CAS-I-D Synechococcus NKBG15041c
1015766545 CAS-I-B Synechococcus PCC 73109
1002029317 CAS-I-D Synechococcus PCC 7002
1114933655 CAS-I-D Leptolyngbya NIES 2104
1016070968 CAS-I-D Leptolyngbya NIES 3755
1031075098 CAS-I-D Oscillatoriales cyanobacterium JSC 12
1506161216 CAS-I-D Leptolyngbya JSC 1
1002448989 CAS-I-D Cyanothece PCC 7425
95
1082480679 CAS-I-D Neosynechococcus sphagnicola sy1
1206911129 CAS-I-D Geitlerinema PCC 7105
1002397705 CAS-I-D Cyanothece PCC 8801
1002779341 CAS-I-D Cyanothece PCC 8802
1033004503 CAS-I-D Gloeocapsa PCC 73106
1003730755 CAS-I-D Cyanothece PCC 7822
1002403261 CAS-I-D Cyanothece PCC 7424
I-U
V-B
I-D
I-A
III
IV
type of the system
cannot be assigned

Fig. 3 (See legend on next page.)

(See figure on previous page.)
Fig. 3 Selected examples of confidently supported subtrees of Cas4 associated with different subtypes of CRISPR-Cas effector complexes. Strongly supported (support values indicated in red) subtrees correspond to branches from the complete Cas4 tree shown in Additional file 1: Figure S1 and available in Newick format as Supplementary Information File 1.Each sequence is denoted by a numeric identifier, matching those in Additional file 7: Table S1, subtype of the CRISPR-Cas system from the respective locus and species name. Several branches were collapsed and are indicated by triangles with CRISPR-Cas system subtype indicated next to respective triangle. The coloring of the sequences corresponds to distinct CRISPR-Cas system types or subtypes according to the color code provided at the bottom

Figure S1, Additional file 3: Information File S1). For example, Cas4 proteins encoded in Acidianus filamentous virus genomes form a lineage that is the sister group of solo-Cas4 from *Sulfolobales* within the archaeal solo-Cas4 branch, suggesting acquisition of these genes from the host by the ancestral virus (Additional file 1: Figure S1, Additional file 3: Information File S1). Sulfolobus spindle shaped virus encodes Cas4 that is most similar to a group of Cas4 proteins associated with type I CRISPR-Cas systems in *Thermococcales* (Additional file 1: Figure S1, Additional file 3: Information File S1). In this case, the ancestral virus seems to have captured a CAS-Cas4 that became solo in the virus genome. Several Cas4 proteins from Sulfolobus islandicus rudiviruses and Acidianus filamentous virus also form a branch, albeit weakly supported. Cas4 proteins from numerous phages form a large, well-supported clade which includes many phages of Cyanobacteria and Proteobacteria, suggesting dissemination of the *cas4* gene among phages. One branch in the Cas4 tree cannot be linked to any known phages but likely corresponds to uncharacterized prophages integrated in genomes of *Campylobacteriales.* Another large clade of Cas4 is apparently associated with conjugative plasmids from *Burkholderiales*. Other Cas4 proteins encoded by bacteriophages and archaeal viruses are scattered over the tree and typically do not show clear affinity to solo-Cas4, CAS-Cas4 or each other (Additional file 1: Figure S1, Additional file 3: Information File S1).

Horizontal gene transfer and selection pressures on different groups of *cas4* genes.

To estimate the relative contributions of the vertical and horizontal components in the evolutionary history of the *cas4* gene family, we compared the tree-induced distances between Cas4 protein sequences and the distances between the 16S rRNAs from the corresponding genomes (proxy for the species tree) as described earlier [11] (see Methods for details); a parallel analysis of Cas1 proteins was used as a reference (see Methods for the link to the data available online).

Both Cas4 and Cas1 only loosely follow the species tree in their history: the Spearman rank correlation coefficients with the 16S rRNA distances for CRISPR-Cas associated Cas4 and Cas1 are 0.15 and 0.22, respectively (see Methods for details). Notably, both solo *cas4* genes and solo *cas1* genes show a stronger vertical evolution trend (correlation coefficients of 0.47 and 0.53, respectively). The substantially greater horizontal component in the evolutionary history of the CRISPR-associated genes reflects the more dynamic nature of the evolution of defense systems compared to other categories of microbial genes [36]. The *cas4* and *cas1* genes encoded in the same locus nearly always evolve (and get transferred) as a single unit (see Methods for the link to the data available online): the correlation between the *cas4* and *cas1* distances for co-located genes is 0.88, and both show closely similar, low correlation coefficients with the rRNA tree (0.14 and 0.13, respectively).

All *cas4* genes seem to evolve under comparable levels of selective pressure, with the medians of the *dN/dS* ratios for the CAS-cas4 and solo-cas4 gene pairs being in the range of 0.06 to 0.18 (Additional file 5: Information file S3). These values are close to the genome-wide median *dN/dS* ratios that are in the range of 0.10 to 0.14 [37]. Thus, in agreement with previous observations [38], all *cas4* genes experience moderate levels of purifying selection that are close to the median values for the respective bacteria and archaea. Furthermore, it has been shown that the characteristic selective pressure on *cas4* gene is only slightly lower than that on *cas1* and *cas2*, the most strongly conserved *cas* genes, which is compatible with the joint involvement of the respective Cas proteins in adaptation [38].

Cas4-like proteins of pfam12705

As mentioned above, Cas4-like proteins that showed the highest similarity to pfam12705 are not associated with CRISPR-Cas systems. However, in addition to the *Campylobacter* phages that encode a Cas4-like nuclease that has been shown to interfere with the host type II-C CRISPR-Cas system [22, 23], members of this family were identified in several other notable contexts. These include a large clade that consists of phages infecting different *Mycobacteria* as well as some *Propionibacterium* and *Bacillus* phages as well as several smaller branches including phages that infect *Bacillus*, *Thermus* and other bacteria (Additional file 1: Figure S1, Additional file 3: Information File S1).

The largest branch in the pfam12705 family includes proteins from many bacteria and mesophilic archaea, in which a Cas4-like nuclease domain is fused to a UvrD-

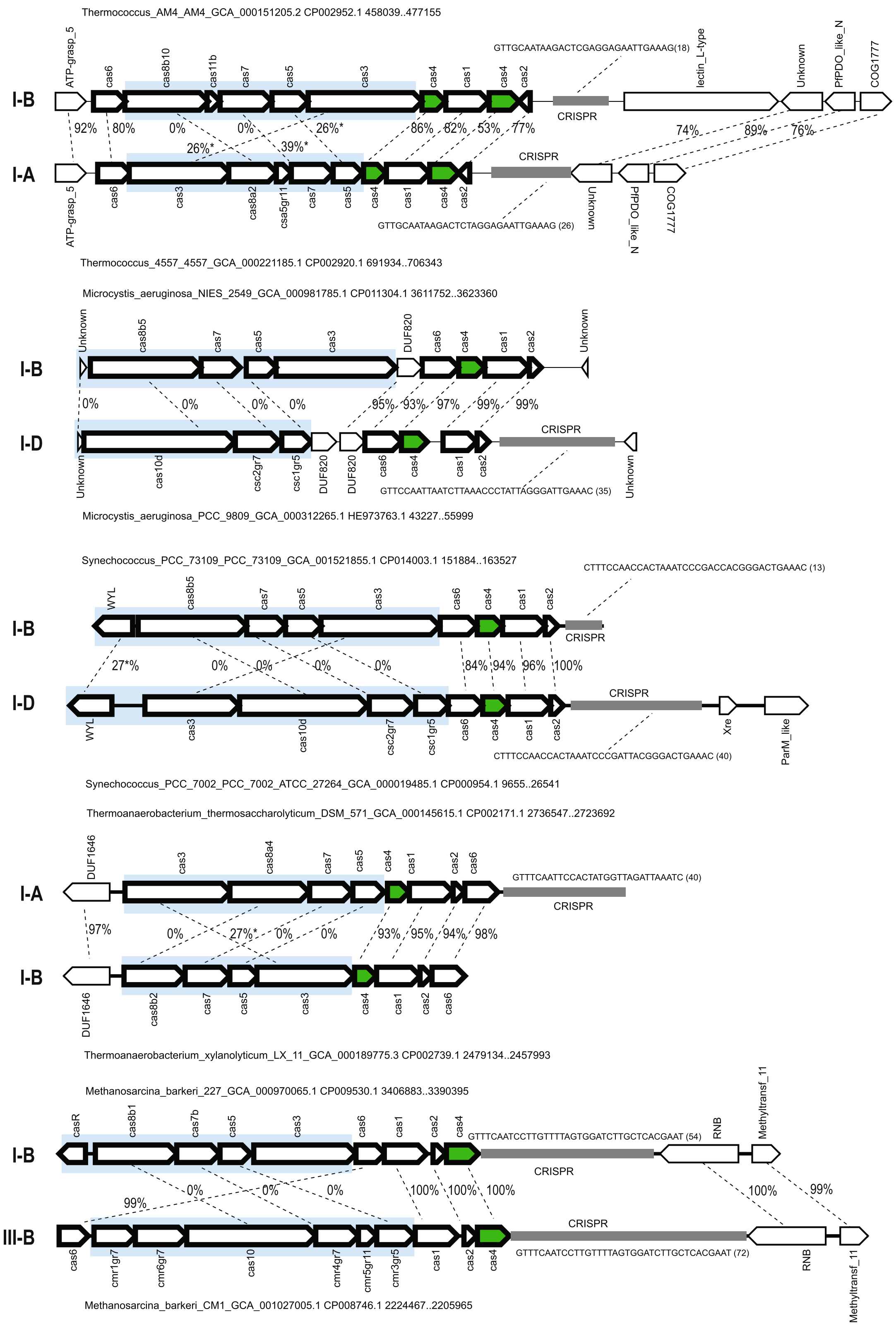
Thermococcus_AM4_AM4_GCA_000151205.2 CP002952.1 458039..477155
I-B
ATP-grasp_5
cas6
cas8b10
cas11b
cas7
cas5
cas3
cas4
cas1
cas4
cas2
GTTGCAATAAGACTCGAGGAGAATTGAAAG(18)
CRISPR
lectin_L-type
Unknown
PfPDO_like_N
COG1777
92%
80%
0%
0%
26%*
86%
82%
53%
77%
26%*
39%*
74%
89%
76%
I-A
ATP-grasp_5
cas6
cas3
cas8a2
csa5gr11
cas7
cas5
cas4
cas1
cas4
cas2
CRISPR
GTTGCAATAAGACTCTAGGAGAATTGAAAG (26)
Unknown
PfPDO_like_N
COG1777
Thermococcus_4557_4557_GCA_000221185.1 CP002920.1 691934..706343
Microcystis_aeruginosa_NIES_2549_GCA_000981785.1 CP011304.1 3611752..3623360
I-B
Unknown
cas8b5
cas7
cas5
cas3
DUF820
cas6
cas4
cas1
cas2
Unknown
0%
0%
0%
0%
95%
93%
97%
99%
99%
I-D
Unknown
cas10d
csc2gr7
csc1gr5
DUF820
DUF820
cas6
cas4
cas1
cas2
CRISPR
GTTCCAATTAATCTTAAACCCTATTAGGGATTGAAAC (35)
Unknown
Microcystis_aeruginosa_PCC_9809_GCA_000312265.1 HE973763.1 43227..55999
Synechococcus_PCC_73109_PCC_73109_GCA_001521855.1 CP014003.1 151884..163527
CTTTCCAACCACTAAATCCCGACCACGGGACTGAAAC (13)
I-B
WYL
cas8b5
cas7
cas5
cas3
cas6
cas4
cas1
cas2
CRISPR
27*%
0%
0%
0%
0%
84%
94%
96%
100%
I-D
WYL
cas3
cas10d
csc2gr7
csc1gr5
cas6
cas4
cas1
cas2
CRISPR
CTTTCCAACCACTAAATCCCGATTACGGGACTGAAAC (40)
Xre
ParM_like
Synechococcus_PCC_7002_PCC_7002_ATCC_27264_GCA_000019485.1 CP000954.1 9655..26541
Thermoanaerobacterium_thermosaccharolyticum_DSM_571_GCA_000145615.1 CP002171.1 2736547..2723692
I-A
DUF1646
cas3
cas8a4
cas7
cas5
cas4
cas1
cas2
cas6
GTTTCAATTCCACTATGGTTAGATTAAATC (40)
CRISPR
97%
0%
27%*
0%
0%
93%
95%
94%
98%
I-B
DUF1646
cas8b2
cas7
cas5
cas3
cas4
cas1
cas2
cas6
Thermoanaerobacterium_xylanolyticum_LX_11_GCA_000189775.3 CP002739.1 2479134..2457993
Methanosarcina_barkeri_227_GCA_000970065.1 CP009530.1 3406883..3390395
I-B
casR
cas8b1
cas7b
cas5
cas3
cas6
cas1
cas2
cas4
GTTTCAATCCTTGTTTTAGTGGATCTTGCTCACGAAT (54)
CRISPR
RNB
Methyltransf_11
99%
0%
0%
0%
100%
100%
100%
100%
99%
III-B
cas6
cmr1gr7
cmr6gr7
cas10
cmr4gr7
cmr5gr11
cmr3gr5
cas1
cas2
cas4
CRISPR
GTTTCAATCCTTGTTTTAGTGGATCTTGCTCACGAAT (72)
RNB
Methyltransf_11
Methanosarcina_barkeri_CM1_GCA_001027005.1 CP008746.1 2224467..2205965

Fig. 4 (See legend on next page.)

(See figure on previous page.)
Fig. 4 Selected examples of in situ replacement of effector complex genes in CRISPR-Cas systems of different subtypes. For each locus, species name, genome accession number and the respective nucleotide coordinates are indicated. The genes of a representative locus are shown by arrows. The arrow indicates the transcription direction of the respective gene. Genes and CRISPR arrays are shown roughly to scale. Homologous genes are connected by dotted lines and amino acid sequence identity (%) is indicated. Asterisks indicate cases where the overall sequence identity was very low, and the percent identity was taken from a small aligned fragment. Cas4 genes are shown in green. The region of likely in situ replacement is shown by a blue rectangle

like superfamily 1 (SF1) 3′-5′ helicase (Fig. 6). Fusion of RecB family nucleases with SF1 helicases is common, e.g. in RecB and AddB, key proteins in bacterial recombinational repair pathways [39]. The UvrD helicase has been originally identified as a component of the nucleotide excision repair complex along with UvrA and UvrB proteins, but subsequently, has been shown to participate in the regulation of RecA recombinase, mismatch repair, transcription-coupled repair, and chromosome and plasmid replication, as an accessory helicase [39, 40]. Many bacteria encode multiple UvrD (COG0210) paralogs [41].

We constructed a tree of UvrD homologs from all genomes that encode the UvrD-Cas4-like fusion and found that the helicase domains of the fusion proteins are also monophyletic, indicating a single origin and, most likely, a unique function of the fusion proteins (Additional file 6: Information File S4). The UvrD-Cas4 proteins are encoded in two distinct, conserved contexts. The most common genomic neighborhood includes genes for methylation (Mod) and restriction (Res) subunits of Type III Restriction-Modification systems [42–44]. Several of these loci also encompass genes for predicted regulatory proteins containing WYL and helix-turn-helix (HTH) domains as well as Shlaffen-like ATPases fused to an HTH domain that are often encoded in other defense contexts.

Another context of the UvrD-Cas4 fusion includes DNA-cytosine methylase Dcm that is responsible for all 5-methylcytosine modifications in *E. coli* [45]. The physiological role of this methylation remains unknown although it has been shown that Dcm recognition sites correspond to sequence motifs for very short patch repair of T/G mismatches, and the Vsr nucleases involved in this process are often encoded next to the *dcm* gene [45]. Indeed, several *uvrD-cas4* loci also contain a gene for a Vsr nuclease, suggesting that this system retains the same features of recognition sites and functional link between Dcm and Vsr (Fig. 6). The same pair of genes was observed previously in a variety of defense gene contexts, MORC family ATPases in particular [46]. Collectively, these genomic associations strongly suggest that the UvrD-Cas4 helicase-nuclease fusion proteins are components of multiple microbial defense systems.

Considering that, in both cases described above, the UvrD-Cas4 protein appears to belong to the same subfamily and the *uvrD-cas4* gene is either standalone or is found in non-conserved contexts in archaeal and bacterial genomes, it appears that the fusion protein plays an ancillary role in different pathways by facilitating site-specific DNA unwinding and cleavage. The Cas4-like nuclease domain can be predicted to play a role analogous to the nuclease domains of RecB and AddB in promoting DNA recombination pathways [19]. Most of the bacterial genomes that encode the UvrD-Cas4 fusion also encode RecB (a near ubiquitous protein in bacteria) suggesting that the functions of these helicase-nuclease fusion proteins could be partially redundant rather than complementary, a pattern typical of repair pathways [47, 48].

The link between UvrD and Cas4-like protein evolved at least on one other independent occasion. In several bacteria, mostly Clostridiales, the two genes are not fused but are encoded in a predicted operon in a context that does not appear indicative of defense functions, suggesting that the two enzymes could be jointly involved in other cellular processes, such as DNA repair (Fig. 6).

Finally, in several bacteria and mesophilic archaea, we identified another link between a Cas4-like nuclease and a superfamily 2 helicase that belongs to the uncharacterized YprA-like family (COG1205) and is unrelated to UvrD. This helicase contains a characteristic C-terminal metal-binding domain (pfam09369/DUF1998 family) and large inserts with additional putative metal-binding domains (Fig. 6). Helicases of this family are strongly associated with defense islands [49], suggesting that, together with the Cas4-like protein, they participate in so far uncharacterized defense pathways.

Conclusions

Unlike most *cas* genes, Cas4 family endonucleases are about as often encoded outside as they are within CRISPR-*cas* loci and are also found in many MGE, particularly archaeal viruses. Because of this unusual distribution of *cas4* genes in bacterial and archaeal genomes, we sought to explore the phylogenomics of the Cas4 family. The results of this phylogenomic indicate that evolution of those *cas4* genes that belong to CRISPR-Cas adaptation modules (Cas-*cas4*) involves

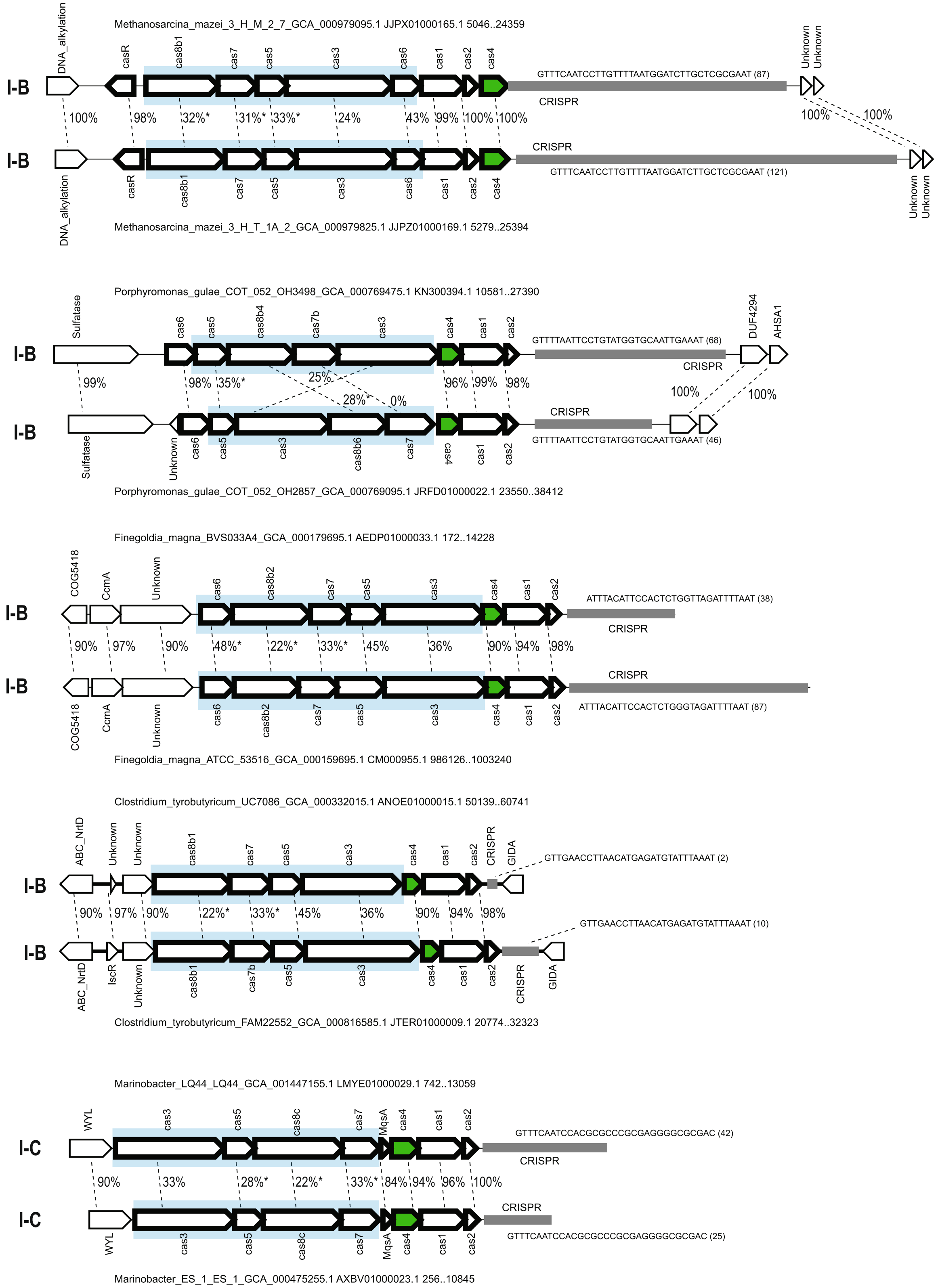
Methanosarcina_mazei_3_H_M_2_7_GCA_000979095.1 JJPX01000165.1 5046..24359
I-B
DNA_alkylation
casR
cas8b1
cas7
cas5
cas3
cas6
cas1
cas2
cas4
GTTTCAATCCTTGTTTTAATGGATCTTGCTCGCGAAT (87)
CRISPR
Unknown
Unknown
100%
98%
32%*
31%*
33%*
24%
43%
99%
100%
100%
100%
100%
I-B
CRISPR
GTTTCAATCCTTGTTTTAATGGATCTTGCTCGCGAAT (121)
DNA_alkylation
casR
cas8b1
cas7
cas5
cas3
cas6
cas1
cas2
cas4
Unknown
Unknown
Methanosarcina_mazei_3_H_T_1A_2_GCA_000979825.1 JJPZ01000169.1 5279..25394
Porphyromonas_gulae_COT_052_OH3498_GCA_000769475.1 KN300394.1 10581..27390
I-B
Sulfatase
cas6
cas5
cas8b4
cas7b
cas3
cas4
cas1
cas2
GTTTTAATTCCTGTATGGTGCAATTGAAAT (68)
CRISPR
DUF4294
AHSA1
99%
98%
35%*
25%
28%*
0%
96%
99%
98%
100%
100%
I-B
CRISPR
GTTTTAATTCCTGTATGGTGCAATTGAAAT (46)
Sulfatase
Unknown
cas6
cas5
cas3
cas8b6
cas7
cas4
cas1
cas2
Porphyromonas_gulae_COT_052_OH2857_GCA_000769095.1 JRFD01000022.1 23550..38412
Finegoldia_magna_BVS033A4_GCA_000179695.1 AEDP01000033.1 172..14228
I-B
COG5418
CcmA
Unknown
cas6
cas8b2
cas7
cas5
cas3
cas4
cas1
cas2
ATTTACATTCCACTCTGGTTAGATTTTAAT (38)
CRISPR
90%
97%
90%
48%*
22%*
33%*
45%
36%
90%
94%
98%
I-B
CRISPR
ATTTACATTCCACTCTGGGTAGATTTTAAT (87)
COG5418
CcmA
Unknown
cas6
cas8b2
cas7
cas5
cas3
cas4
cas1
cas2
Finegoldia_magna_ATCC_53516_GCA_000159695.1 CM000955.1 986126..1003240
Clostridium_tyrobutyricum_UC7086_GCA_000332015.1 ANOE01000015.1 50139..60741
I-B
ABC_NrtD
Unknown
Unknown
cas8b1
cas7
cas5
cas3
cas4
cas1
cas2
CRISPR
GIDA
GTTGAACCTTAACATGAGATGTATTTAAAT (2)
90%
97%
90%
22%*
33%*
45%
36%
90%
94%
98%
I-B
GTTGAACCTTAACATGAGATGTATTTAAAT (10)
ABC_NrtD
IscR
Unknown
cas8b1
cas7b
cas5
cas3
cas4
cas1
cas2
CRISPR
GIDA
Clostridium_tyrobutyricum_FAM22552_GCA_000816585.1 JTER01000009.1 20774..32323
Marinobacter_LQ44_LQ44_GCA_001447155.1 LMYE01000029.1 742..13059
I-C
WYL
cas3
cas5
cas8c
cas7
MqsA
cas4
cas1
cas2
GTTTCAATCCACGCGCCCGCGAGGGGCGCGAC (42)
CRISPR
90%
33%
28%*
22%*
33%*
84%
94%
96%
100%
I-C
CRISPR
GTTTCAATCCACGCGCCCGCGAGGGGCGCGAC (25)
WYL
cas3
cas5
cas8c
cas7
MqsA
cas4
cas1
cas2
Marinobacter_ES_1_ES_1_GCA_000475255.1 AXBV01000023.1 256..10845

Fig. 5 Selected examples of in situ replacement of effector complex genes in CRISPR-Cas loci of the same subtype. Designations are the same as in the Fig. 4

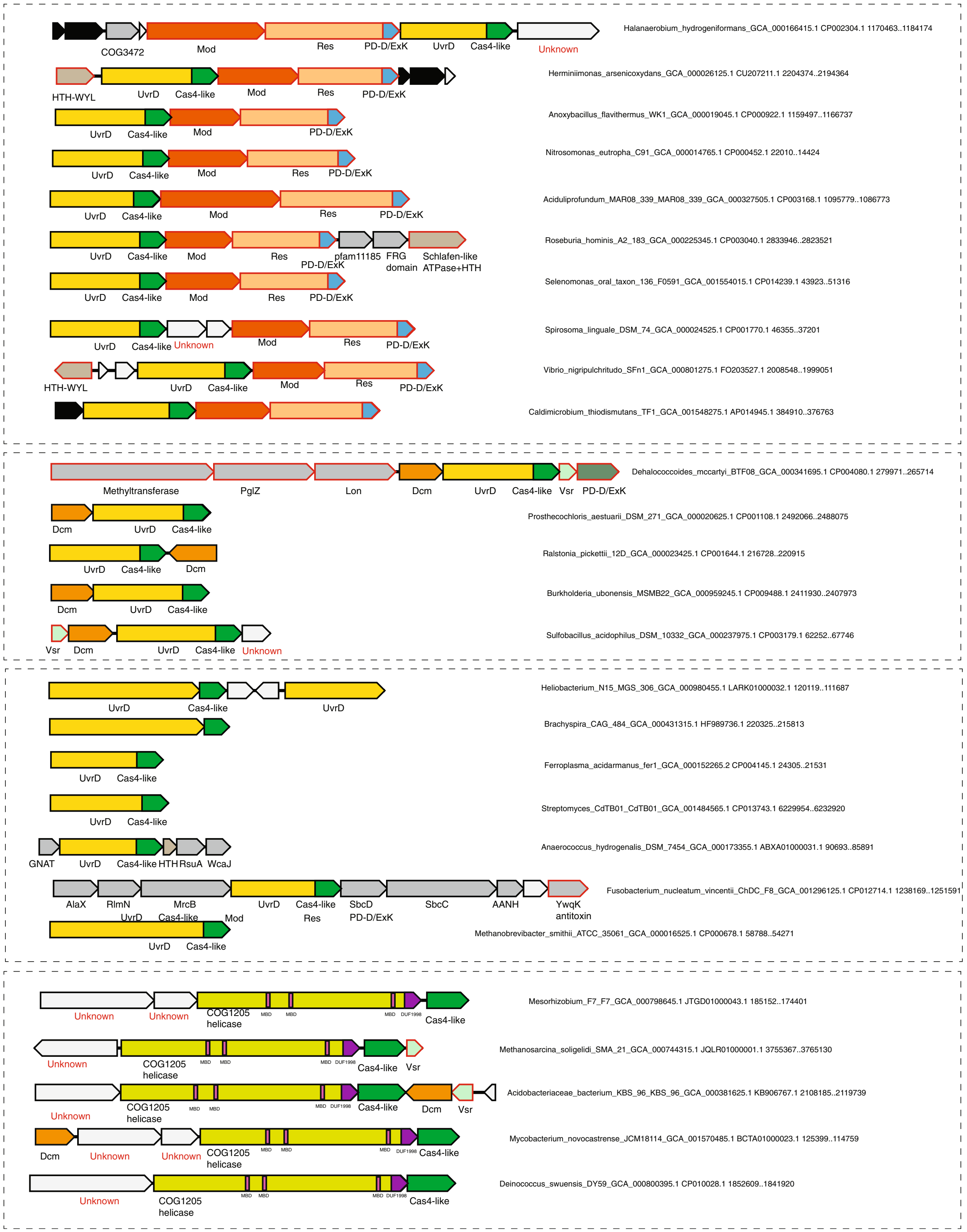
COG3472
Mod
Res
PD-D/ExK
UvrD
Cas4-like
Unknown
Halanaerobium_hydrogeniformans_GCA_000166415.1 CP002304.1 1170463..1184174
HTH-WYL
UvrD
Cas4-like
Mod
Res
PD-D/ExK
Herminiimonas_arsenicoxydans_GCA_000026125.1 CU207211.1 2204374..2194364
UvrD
Cas4-like
Mod
PD-D/ExK
Anoxybacillus_flavithermus_WK1_GCA_000019045.1 CP000922.1 1159497..1166737
UvrD
Cas4-like
Mod
Res
PD-D/ExK
Nitrosomonas_eutropha_C91_GCA_000014765.1 CP000452.1 22010..14424
UvrD
Cas4-like
Mod
Res
PD-D/ExK
Aciduliprofundum_MAR08_339_MAR08_339_GCA_000327505.1 CP003168.1 1095779..1086773
UvrD
Cas4-like
Mod
Res
PD-D/ExK
pfam11185
FRG domain
Schlafen-like ATPase+HTH
Roseburia_hominis_A2_183_GCA_000225345.1 CP003040.1 2833946..2823521
UvrD
Cas4-like
Mod
Res
PD-D/ExK
Selenomonas_oral_taxon_136_F0591_GCA_001554015.1 CP014239.1 43923..51316
UvrD
Cas4-like
Unknown
Mod
Res
PD-D/ExK
Spirosoma_linguale_DSM_74_GCA_000024525.1 CP001770.1 46355..37201
HTH-WYL
UvrD
Cas4-like
Mod
Res
PD-D/ExK
Vibrio_nigripulchritudo_SFn1_GCA_000801275.1 FO203527.1 2008548..1999051
Caldimicrobium_thiodismutans_TF1_GCA_001548275.1 AP014945.1 384910..376763
Methyltransferase
PglZ
Lon
Dcm
UvrD
Cas4-like
Vsr
PD-D/ExK
Dehalococcoides_mccartyi_BTF08_GCA_000341695.1 CP004080.1 279971..265714
Dcm
UvrD
Cas4-like
Prosthecochloris_aestuarii_DSM_271_GCA_000020625.1 CP001108.1 2492066..2488075
UvrD
Cas4-like
Dcm
Ralstonia_pickettii_12D_GCA_000023425.1 CP001644.1 216728..220915
Dcm
UvrD
Cas4-like
Burkholderia_ubonensis_MSMB22_GCA_000959245.1 CP009488.1 2411930..2407973
Vsr
Dcm
UvrD
Cas4-like
Unknown
Sulfobacillus_acidophilus_DSM_10332_GCA_000237975.1 CP003179.1 62252..67746
UvrD
Cas4-like
UvrD
Heliobacterium_N15_MGS_306_GCA_000980455.1 LARK01000032.1 120119..111687
Brachyspira_CAG_484_GCA_000431315.1 HF989736.1 220325..215813
UvrD
Cas4-like
Ferroplasma_acidarmanus_fer1_GCA_000152265.2 CP004145.1 24305..21531
UvrD
Cas4-like
Streptomyces_CdTB01_CdTB01_GCA_001484565.1 CP013743.1 6229954..6232920
GNAT
UvrD
Cas4-like
HTH
RsuA
WcaJ
Anaerococcus_hydrogenalis_DSM_7454_GCA_000173355.1 ABXA01000031.1 90693..85891
AlaX
RlmN
MrcB
UvrD
Cas4-like
SbcD
SbcC
AANH
YwqK antitoxin
Fusobacterium_nucleatum_vincentii_ChDC_F8_GCA_001296125.1 CP012714.1 1238169..1251591
UvrD
Cas4-like
Mod
Res
PD-D/ExK
UvrD
Cas4-like
Methanobrevibacter_smithii_ATCC_35061_GCA_000016525.1 CP000678.1 58788..54271
Unknown
Unknown
COG1205 helicase
MBD
MBD
MBD
DUF1998
Cas4-like
Mesorhizobium_F7_F7_GCA_000798645.1 JTGD01000043.1 185152..174401
Unknown
COG1205 helicase
MBD
MBD
MBD
DUF1998
Cas4-like
Vsr
Methanosarcina_soligelidi_SMA_21_GCA_000744315.1 JQLR01000001.1 3755367..3765130
Unknown
COG1205 helicase
MBD
MBD
MBD
DUF1998
Cas4-like
Dcm
Vsr
Acidobacteriaceae_bacterium_KBS_96_KBS_96_GCA_000381625.1 KB906767.1 2108185..2119739
Dcm
Unknown
Unknown
COG1205 helicase
MBD
MBD
MBD
DUF1998
Cas4-like
Mycobacterium_novocastrense_JCM18114_GCA_001570485.1 BCTA01000023.1 125399..114759
Unknown
COG1205 helicase
MBD
MBD
MBD
DUF1998
Cas4-like
Deinococcus_swuensis_DY59_GCA_000800395.1 CP010028.1 1852609..1841920

Fig. 6 (See legend on next page.)

(See figure on previous page.)

Fig. 6 Organization of loci containing UvrD-like, COG1205 and Cas4-like genes. For each locus, species name, genome accession number and the respective nucleotide coordinates are indicated. The genes of a representative locus are shown by arrows. The scale of an arrow is roughly proportional to gene length. The arrow indicates direction of the respective gene. Homologous genes and domains are color-coded. Four panels separated by dotted rectangles show a common theme of locus organization. Abbreviations and gene names: Mod –modification subunit of type III RM system, Res – restriction enzyme of type III RM system; Dcm - DNA-cytosine methylase; HTH- helix turn helix; Vsr – very short patch repair nuclease, PD-DExK – PD-DExK family nuclease; WcaJ-Sugar transferase involved in LPS biosynthesis; RsuA - Pseudouridine synthase; MrcB - Membrane carboxypeptidase; RlmN - Adenine C2-methylase of 23S rRNA A2503 and tRNA A37; AlaX - Ser-tRNA(Ala) deacylase; AANH - Adenine nucleotide alpha hydrolases

extensive horizontal transfer that, however, almost always involves complete adaptation modules. The evolutionary integrity of the adaption modules sharply contrasts the rampant modular shuffling of CRISPR-Cas systems whereby a given variant of the adaptation module can combine with virtually any variety of effector modules. The evolutionary history of solo-*cas4* genes is much better compatible with the microbial species tree and appears to include fewer horizontal gene transfer events. The Cas-*cas4* and solo-*cas4* genes evolve under moderately strong purifying selection that, for both classes of *cas4* genes, is close to the genomic median. Thus, the evolutionary regime of the *cas4* genes, along with *cas1* and *cas2* genes, resembles that of 'regular' microbial genes involved in various cellular functions, with no evidence of direct involvement in virus-host arms races. However, *cas4* genes were recruited by various MGE on many independent occasions, suggesting that Cas4 nucleases might be involved in anti-defense functions. The solo-Cas4 as well as MGE-encoded Cas4 so far have not been studied experimentally. Functional characterization of these proteins is likely to result in identification of novel defense and anti-defense systems, and their connections with CRISPR-Cas systems.

Methods

Data

The prokaryotic genomic dataset was taken from WGS and GenBank repositories at the NCBI as of March 2016 and includes 43,000 genomes. Cas4-like proteins were identified using PSI-BLAST search against sequence profiles corresponding to the following subfamilies: COG1468, COG4343, pfam01930, pfam06023, cd09637, cd09659, cls000170 [11] and pfam12705. To ensure specificity of the extracted set of proteins after the initial identification, all retrieved protein sequences were additionally run against the Conserved Domain Database (CDD) using the RPS-BLAST program with the same parameters and all proteins with higher similarity to any other profiles were discarded. The protein sequences from the *cas4* neighborhoods were extracted from the respective genomes or contigs and annotated using the COGs, Pfam and CD profiles from the CDD and the custom defense gene profiles [11]. The *cas* genes and CRISPR-Cas subtypes were annotated according to the previously described nomenclature [11].

Clustering and Phylogenetic analysis

The NCBI BLASTCLUST program (ftp://ftp.ncbi.nih.gov/blast/documents/blastclust.html) was used to cluster sequences by similarity. For Cas4 clustering, BLASTCLUST was used with the sequence identity threshold of 50% and length coverage threshold of 0.8; for 16S rRNA nucleotide sequences, the program was used with sequence identity threshold of 90% and length coverage threshold of 0.9. For Cas8 and Cas10d clustering, the program was used with sequence identity threshold of 0.5 and length coverage threshold of 80%. Alignments of protein sequences were constructed using MUSCLE [50] and MAFFT [51] programs. Short fragments or disrupted sequences were discarded. Sites with the gap character fraction values >0.5 and homogeneity <0.1 were removed from the alignment. Phylogenetic analysis was performed using the FastTree program [52], with the WAG evolutionary model for amino acid sequences and the GTR evolutionary model for nucleotide sequences. The same program was used to compute bootstrap values.

Relationships within diverse sequence families were established using the following procedure: initial sequence clusters were obtained using UCLUST [53] with the sequence similarity threshold of 0.5; sequences were aligned within clusters using MUSCLE [50]. Then cluster-to-cluster similarity scores were obtained using HHsearch [54] (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 [55]; the procedure was repeated iteratively. At the last step, sequence-based trees were reconstructed from the cluster alignments using FastTree [52], as described above and rooted by midpoint; these trees were grafted onto the tips of the profile similarity-based UPGMA dendrogram.

For the comparison of evolutionary distances between Cas4 and Cas1 protein sequences and 16S rRNA sequences from the corresponding genomes (proxy of the species tree), all pairwise distances between CRISPR-associated or solo proteins were calculated from the corresponding phylogenetic trees; for protein sequences from unaligned clusters, an arbitrary high distance of 12 was assigned to the corresponding pair. If a pair of genomes contained more than one Cas4 or Cas1 sequence in either or both of them, the shortest protein-protein distance was selected to represent this genome pair. Distances between the corresponding 16S rRNA sequences were, likewise, calculated from the 16S rRNA trees or, if the two 16S rRNAs belonged to unaligned clusters, an arbitrary distance high distance of 3 was assigned. Spearman rank correlation between protein-protein and rRNA-rRNA sequence distances for all pairs was reported as a measure of coherence between the protein evolution and the species tree.

All pairwise distances are available for download at ftp://ftp.ncbi.nih.gov/pub/wolf/_suppl/cas4.

Additional files

Additional file 1: Figure S1. Schematic representation of the maximum likelihood phylogenetic tree of Cas4 (7060 sequences all together), available in the Supplementary File 1. Support values are calculated by FastTRee program only for the confidently aligned groups, all other values were assigned to zero automatically. Major well-supported distinct branches are shown by rectangles which are color-coded according to Cas4 assignments. Assignments and other comments are shown next to the each collapsed branch. Individual sequences in the tree are described by a local numeric ID, species name and color-coded according to Cas4 assignment (also provided in the Additional file 7: Table S1). Blue shading shows tree clades that belong to pfam12705 family. (PDF 694 kb)

Additional file 2: Figure S2. Nucleotide sequence comparisons of CRISPR-Cas loci encoded in closely related strains. On the axes, the labels contain the name of the source genome, contig ID and the coordinates of the respective loci. The annotations for CRISPR-Cas loci were taken from the Additional file 7: Table S1, "Loci" worksheet. The cartoons on the axes represent the genes and CRISPR repeats encoded in these loci. The sizes of the cartoons are proportional to the actual sizes of these genes. Colors: black are CRISPR arrays, blue are cas genes, green - cas4 gene, shaded area are the regions which have >70% sequence identity level. Left: Comparison of two I-C systems from Marinobacter strains. Right: Comparison of two I-B systems from Campylobacter strains. (DOCX 606 kb)

Additional file 3: Information File S1. Complete tree for Cas4-like set in Newick format. Cas4 assignments are included in the leaf name. (TXT 463 kb)

Additional file 4: Information File S2. Breakdown of CAS-Cas4 and solo-Cas4 presence in completely sequenced genomes. (XLSX 248 kb)

Additional file 5: Information File S3. Detailed description of the Cas4 *dN/dS* analysis, table with results. (DOCX 51 kb)

Additional file 6: Information File S4. Complete tree for selected representatives of UvrD family in Newick format. The tree is based on the helicase domain alignment. Protein linked to Cas4 are indicated in the respective leaf names. (DOCX 51 kb)

Additional file 7: Table S1 Worksheet "loci" provides detailed information on all cas4 loci in completely sequenced and draft genomes of archaea and bacteria. Annotation for the proteins encoded in the loci is based on Cas protein and CDD assignments using PSI-BLAST program (see Methods for details). Worksheet "tree order and assignments" provides information of the order of the Cas4 in the tree (Supplementary file 1), Cas4 assignments to distinct groups of CAS-Cas4, solo-Cas4 and others. (TXT 28 kb)

Acknowledgements
The authors thank Koonin group members for helpful discussions.

Funding
This work was supported by intramural funds of the US Department of Health and Human Services (the National Library of Medicine, to EVK) and by the NIH grant R35 118,160 (to MPT).

Availability of data and materials
The datasets supporting the conclusions of this article are included within the additional files and available at ftp://ftp.ncbi.nih.gov/pub/wolf/_suppl/cas4

Authors' contributions
Conception and design of the study: KSM, YIW, EVK. Data collection: SH, SS. Data analysis: SH, YIW, MPT, KSM. Manuscript drafting: SH, YIW, KSM. Manuscript revision for critical intellectual content: EVK. All authors read and approved the final manuscript.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details
[1]National Center for Biotechnology Information, National Institutes of Health, Bethesda, MD, USA. [2]Skolkovo Institute of Science and Technology, Skolkovo 143025, Russia. [3]Departments of Biochemistry and Molecular Biology, Genetics, and Microbiology, University of Georgia, Athens, GA, USA.

Received: 7 August 2017 Accepted: 16 November 2017
Published online: 28 November 2017

References

1. Plagens A, Tjaden B, Hagemann A, Randau L, Hensel R. Characterization of the CRISPR/Cas subtype I-A system of the hyperthermophilic crenarchaeon Thermoproteus tenax. J Bacteriol. 2012;194(10):2491–500.
2. Lemak S, Beloglazova N, Nocek B, Skarina T, Flick R, Brown G, Popovic A, Joachimiak A, Savchenko A, Yakunin AF. Toroidal structure and DNA cleavage by the CRISPR-associated [4Fe-4S] cluster containing Cas4 nuclease SSO0001 from Sulfolobus solfataricus. J Am Chem Soc. 2013;135(46):17476–87.
3. Li M, Wang R, Zhao D, Xiang H. Adaptation of the Haloarcula hispanica CRISPR-Cas system to a purified virus strictly requires a priming process. Nucleic Acids Res. 2014;42(4):2483–92.
4. Liu T, Li Y, Wang X, Ye Q, Li H, Liang Y, She Q, Peng N. Transcriptional regulator-mediated activation of adaptation genes triggers CRISPR de novo spacer acquisition. Nucleic Acids Res. 2015;43(2):1044–55.
5. Amitai G, Sorek R. CRISPR-Cas adaptation: insights into the mechanism of action. Nat Rev Microbiol. 2016;14(2):67–76.
6. Jackson SA, McKenzie RE, Fagerlund RD, Kieper SN, Fineran PC, Brouns SJ: **CRISPR-Cas: Adapting to change**. Science 2017, **356**(6333).
7. Barrangou R, Horvath P. A decade of discovery: CRISPR functions and applications. Nat Microbiol. 2017;2:17092.
8. Barrangou R, Marraffini LA. CRISPR-Cas systems: prokaryotes upgrade to adaptive immunity. Mol Cell. 2014;54(2):234–44.
9. Marraffini LA. CRISPR-Cas immunity in prokaryotes. Nature. 2015; 526(7571):55–61.
10. Charpentier E, Richter H, van der Oost J, White MF. Biogenesis pathways of RNA guides in archaeal and bacterial CRISPR-Cas adaptive immunity. FEMS Microbiol Rev. 2015;39(3):428–41.
11. Makarova KS, Wolf YI, Alkhnbashi OS, Costa F, Shah SA, Saunders SJ, Barrangou R, Brouns SJ, Charpentier E, Haft DH, et al. An updated evolutionary classification of CRISPR-Cas systems. Nat Rev Microbiol. 2015; 13(11):722–36.

12. Koonin EV, Makarova KS, Zhang F. Diversity, classification and evolution of CRISPR-Cas systems. Curr Opin Microbiol. 2017;37:67–78.
13. Liu T, Liu Z, Ye Q, Pan S, Wang X, Li Y, Peng W, Liang Y, She Q, Peng N. Coupling transcriptional activation of CRISPR-Cas system and DNA repair genes by Csa3a in Sulfolobus islandicus. Nucleic Acids Res. 2017;45(15):8978–92.
14. Makarova KS, Grishin NV, Shabalina SA, Wolf YI, Koonin EV. A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action. Biol Direct. 2006;1:7.
15. Lemak S, Nocek B, Beloglazova N, Skarina T, Flick R, Brown G, Joachimiak A, Savchenko A, Yakunin AF. The CRISPR-associated Cas4 protein Pcal_0546 from Pyrobaculum calidifontis contains a [2Fe-2S] cluster: crystal structure and nuclease activity. Nucleic Acids Res. 2014;42(17):11144–55.
16. Zhang J, Kasciukovic T, White MF. The CRISPR associated protein Cas4 is a 5′ to 3′ DNA exonuclease with an iron-sulfur cluster. PLoS One. 2012;7(10):e47232.
17. Krajewski WW, Fu X, Wilkinson M, Cronin NB, Dillingham MS, Wigley DB. Structural basis for translocation by AddAB helicase-nuclease and its arrest at chi sites. Nature. 2014;508(7496):416–9.
18. Wilkinson M, Chaban Y, Wigley DB. Mechanism for nuclease regulation in RecBCD. elife. 2016;5
19. Wigley DB. Bacterial DNA repair: recent insights into the mechanism of RecBCD, AddAB and AdnAB. *Nat Rev Microbiol*. 2013;11(1):9–13.
20. Zhou C, Pourmal S, Pavletich NP: Dna2 nuclease-helicase structure, mechanism and regulation by Rpa. Elife. 2015;4. doi:10.7554/eLife.09832.
21. Steczkiewicz K, Muszewska A, Knizewski L, Rychlewski L, Ginalski K. Sequence, structure and functional diversity of PD-(D/E)XK phosphodiesterase superfamily. Nucleic Acids Res. 2012;40(15):7016–45.
22. Hooton SP, Brathwaite KJ, Connerton IF. The Bacteriophage carrier state of campylobacter jejuni features changes in host non-coding RNAs and the Acquisition of new Host-Derived CRISPR spacer sequences. Front Microbiol. 2016;7:355.
23. Hooton SP, Connerton IF. Campylobacter jejuni acquire new host-derived CRISPR spacers when in association with bacteriophages harboring a CRISPR-like Cas4 protein. Front Microbiol. 2014;5:744.
24. Krupovic M, Cvirkaite-Krupovic V, Prangishvili D, Koonin EV. Evolution of an archaeal virus nucleocapsid protein from the CRISPR-associated Cas4 nuclease. Biol Direct. 2015;10:65.
25. Krupovic M, Beguin P, Koonin EV: Casposons: the mobile elements that gave rise to the adaptation module of CRISPR-Cas systems. Curr Opin Microbiol 2017, in press.
26. Krupovic M, Makarova KS, Forterre P, Prangishvili D, Koonin EV. Casposons: a new superfamily of self-synthesizing DNA transposons at the origin of prokaryotic CRISPR-Cas immunity. BMC Biol. 2014;12:36.
27. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25(17):3389–402.
28. Aziz RK, Devoid S, Disz T, Edwards RA, Henry CS, Olsen GJ, Olson R, Overbeek R, Parrello B, Pusch GD, et al. SEED servers: high-performance access to the SEED genomes, annotations, and metabolic models. PLoS One. 2012;7(10):e48053.
29. Makarova KS, Koonin EV. Annotation and classification of CRISPR-Cas systems. Methods Mol Biol. 2015;1311:47–75.
30. Silas S, Makarova KS, Shmakov S, Páez-Espino D, Mohr G, Liu Y, Davison M, Roux S, Krishnamurthy SR, Hansen LL et al: On the origin of reverse transcriptase-using CRISPR-Cas systems and their hyper-diverse, enigmatic spacer repertoires. MBio, in press 2017.
31. Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, Cheng JF, Darling A, Malfatti S, Swan BK, Gies EA, et al. Insights into the phylogeny and coding potential of microbial dark matter. Nature. 2013;499(7459):431–7.
32. Castelle CJ, Wrighton KC, Thomas BC, Hug LA, Brown CT, Wilkins MJ, Frischkorn KR, Tringe SG, Singh A, Markillie LM, et al. Genomic expansion of domain archaea highlights roles for organisms from new phyla in anaerobic carbon cycling. Curr Biol. 2015;25(6):690–701.
33. Shmakov S, Abudayyeh OO, Makarova KS, Wolf YI, Gootenberg JS, Semenova E, Minakhin L, Joung J, Konermann S, Severinov K, et al. Discovery and functional characterization of diverse class 2 CRISPR-Cas systems. Mol Cell. 2015;60(3):385–97.
34. Brown CT, Hug LA, Thomas BC, Sharon I, Castelle CJ, Singh A, Wilkins MJ, Wrighton KC, Williams KH, Banfield JF. Unusual biology across a group comprising more than 15% of domain bacteria. Nature. 2015;523(7559):208–11.
35. Burstein D, Sun CL, Brown CT, Sharon I, Anantharaman K, Probst AJ, Thomas BC, Banfield JF. Major bacterial lineages are essentially devoid of CRISPR-Cas viral defence systems. Nat Commun. 2016;7:10613.
36. Puigbo P, Makarova KS, Kristensen DM, Wolf YI, Koonin EV. Reconstruction of the evolution of microbial defense systems. BMC Evol Biol. 2017;17(1):94.
37. Kristensen DM, Wolf YI, Koonin EV. ATGC database and ATGC-COGs: an updated resource for micro- and macro-evolutionary studies of prokaryotic genomes and protein family annotation. Nucleic Acids Res. 2017;45(D1):D210–8.
38. Takeuchi N, Wolf YI, Makarova KS, Koonin EV. Nature and intensity of selection pressure on CRISPR-associated genes. J Bacteriol. 2012;194(5):1216–25.
39. Dillingham MS. Superfamily I helicases as modular components of DNA-processing machines. Biochem Soc Trans. 2011;39(2):413–23.
40. Epshtein V. UvrD helicase: an old dog with a new trick: how one step backward leads to many steps forward. BioEssays. 2015;37(1):12–9.
41. Galperin MY, Makarova KS, Wolf YI, Koonin EV. Expanded microbial genome coverage and improved protein family annotation in the COG database. Nucleic Acids Res. 2015;43(Database issue):D261–9.
42. Schwarz FW, Toth J, van Aelst K, Cui G, Clausing S, Szczelkun MD, Seidel R. The helicase-like domains of type III restriction enzymes trigger long-range diffusion along DNA. Science. 2013;340(6130):353–6.
43. Butterer A, Pernstich C, Smith RM, Sobott F, Szczelkun MD, Toth J. Type III restriction endonucleases are heterotrimeric: comprising one helicase-nuclease subunit and a dimeric methyltransferase that binds only one specific DNA. Nucleic Acids Res. 2014;42(8):5139–50.
44. Gupta YK, Chan SH, Xu SY, Aggarwal AK. Structural basis of asymmetric DNA methylation and ATP-triggered long-range diffusion by EcoP15I. Nat Commun. 2015;6:7363.
45. Marinus MG, Lobner-Olesen A. DNA Methylation. EcoSal Plus. 2014;6(1). doi: 10.1128/ecosalplus.ESP-0003-2013.
46. Iyer LM, Abhiman S, Aravind L. MutL homologs in restriction-modification systems and the origin of eukaryotic MORC ATPases. Biol Direct. 2008;3:8.
47. Aravind L, Walker DR, Koonin EV. Conserved domains in DNA repair proteins and evolution of repair systems. Nucleic Acids Res. 1999;27(5):1223–42.
48. Lenhart JS, Schroeder JW, Walsh BW, Simmons LA. DNA repair and genome maintenance in Bacillus Subtilis. Microbiol Mol Biol Rev. 2012;76(3):530–64.
49. Makarova KS, Wolf YI, Snir S, Koonin EV. Defense islands in bacterial and archaeal genomes and prediction of novel defense systems. J Bacteriol. 2011;193(21):6039–56.
50. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32(5):1792–7.
51. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30(4):772–80.
52. Price MN, Dehal PS, Arkin AP. FastTree 2–approximately maximum-likelihood trees for large alignments. PLoS One. 2010;5(3):e9490.
53. Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinformatics. 2010;26(19):2460–1.
54. Soding J. Protein homology detection by HMM-HMM comparison. Bioinformatics. 2005;21(7):951–60.
55. Soding J, Remmert M, Biegert A, Lupas AN, et al. Nucleic Acids Res. 2006; 34(Web Server issue):W374–8.