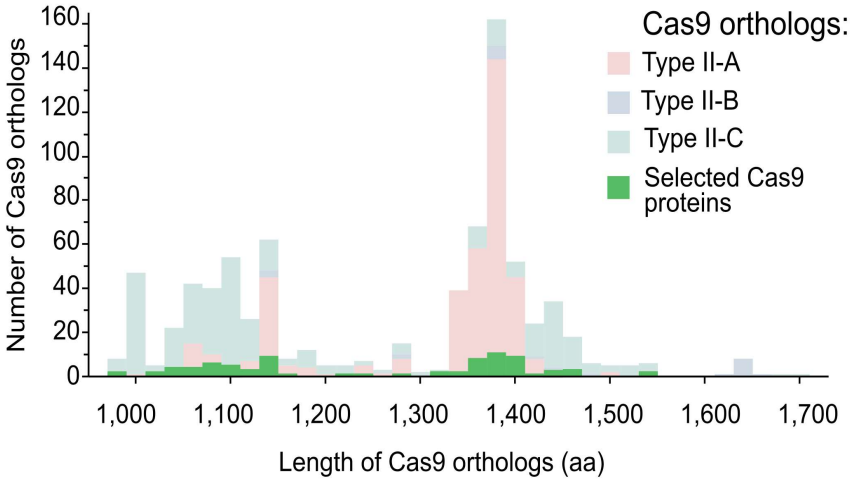
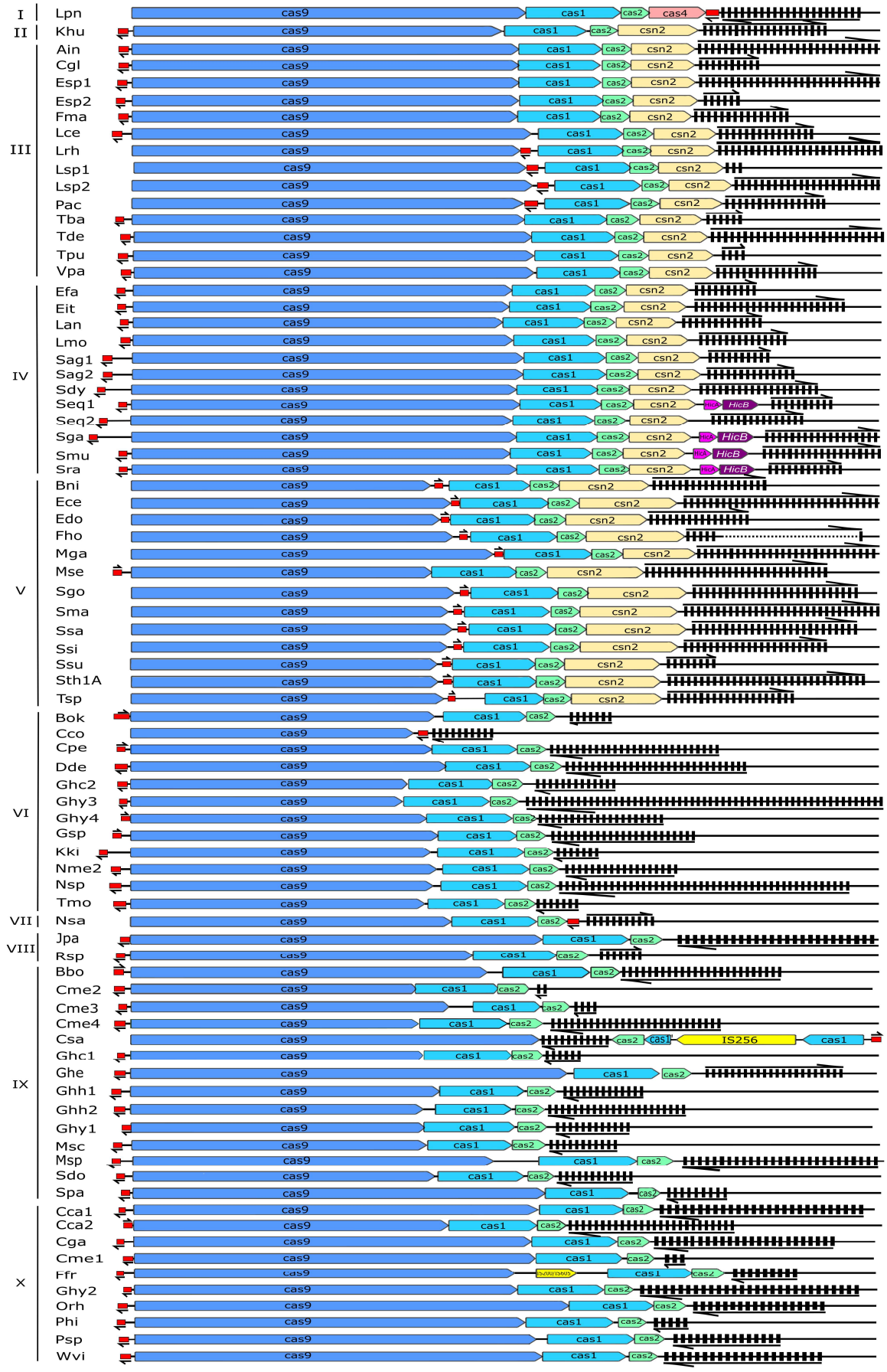
Supplementary Figures

Biochemically diverse CRISPR-Cas9 orthologs

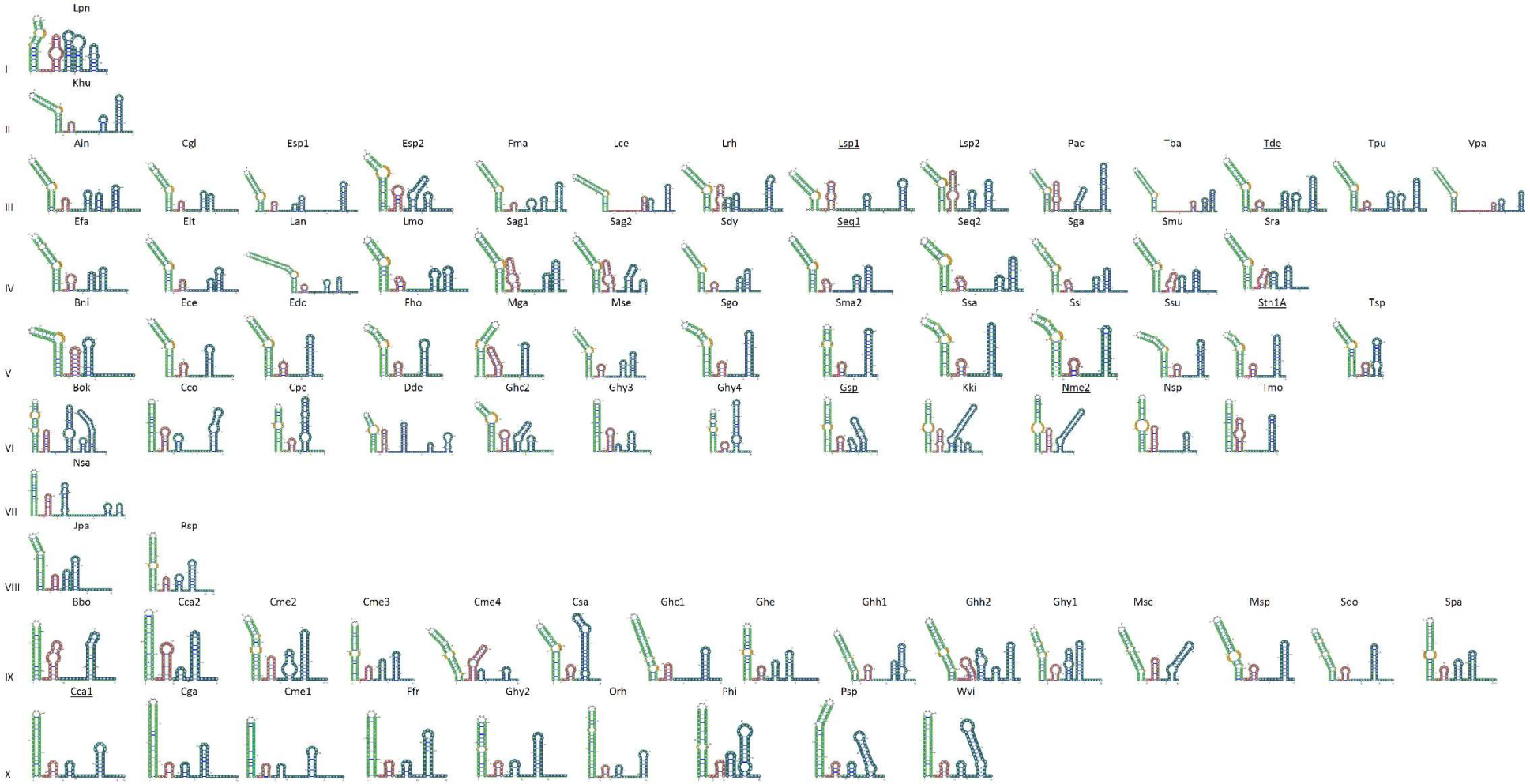
Giedrius Gasiunas, Joshua K. Young, Tautvydas Karvelis, Darius Kazlauskas, Tomas Urbaitis, Monika Jasnauskaite, Mantvyda Grusyte, Sushmitha Paulraj, Po-Hao Wang, Zhenglin Hou, Shane K. Dooley, Mark Cigan, Clara Alarcon, N. Doane Chilcoat, Greta Bigelyte, Jennifer L. Curcuru, Megumu Mabuchi, Zhiyi Sun, Ryan T. Fuchs, Ezra Schildkraut, Peter R. Weigele, William E. Jack, G. Brett Robb, Česlovas Venclovas, and Virginijus Siksnys



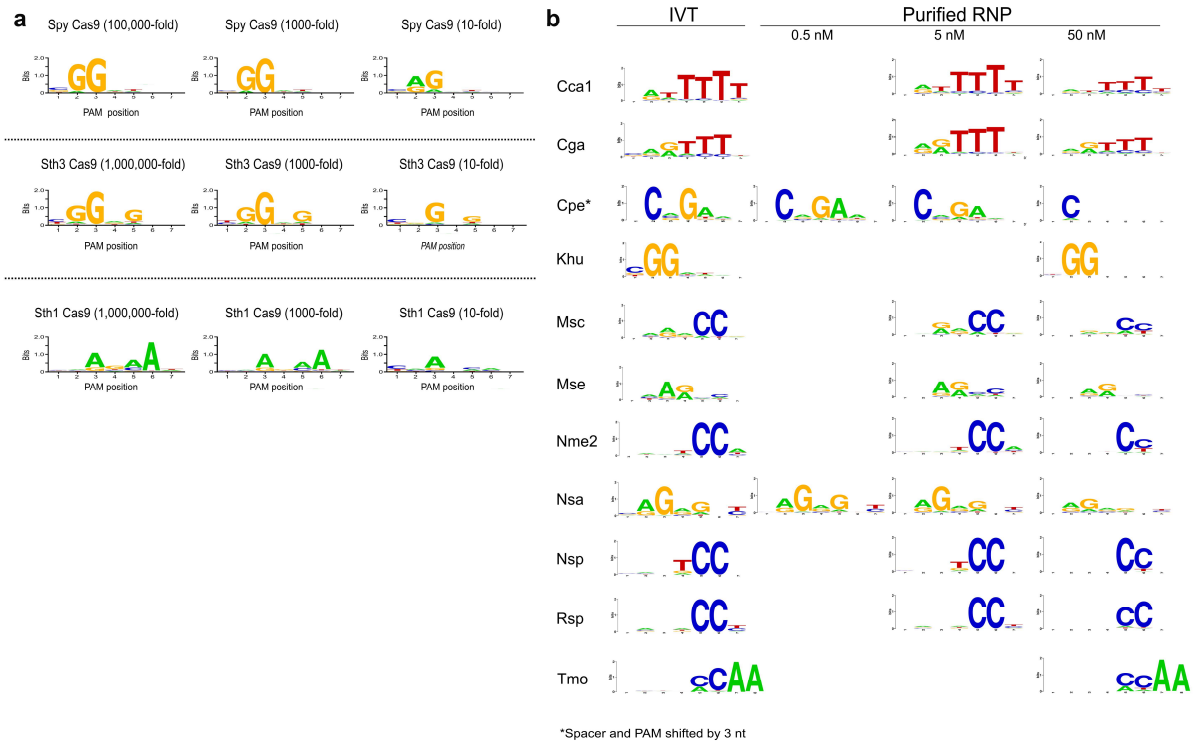
Supplementary Fig. 1. Cas9 protein size distribution according to subtype. Type II-A, B, and C systems are color-coded, red, blue, and turquoise, respectively. The numbers of analysed Cas9s in each group are indicated in green.


I
II
III
IV
V
VI
VII
VIII
IX
X
Lpn
Khu
Ain
Cgl
Esp1
Esp2
Fma
Lce
Lrh
Lsp1
Lsp2
Pac
Tba
Tde
Tpu
Vpa
Efa
Eit
Lan
Lmo
Sag1
Sag2
Sdy
Seq1
Seq2
Sga
Smu
Sra
Bni
Ece
Edo
Fho
Mga
Mse
Sgo
Sma
Ssa
Ssi
Ssu
Sth1A
Tsp
Bok
Cco
Cpe
Dde
Ghc2
Ghy3
Ghy4
Gsp
Kki
Nme2
Nsp
Tmo
Nsa
Jpa
Rsp
Bbo
Cme2
Cme3
Cme4
Csa
Ghc1
Ghe
Ghh1
Ghh2
Ghy1
Msc
Msp
Sdo
Spa
Cca1
Cca2
Cga
Cme1
Ffr
Ghy2
Orh
Phi
Psp
Wvi
cas9
cas1
cas2
cas4
csn2
HicA
HicB
IS256

Supplementary Fig. 2. Schematic representation of the CRISPR-Cas loci encoding the Cas9 orthologs characterized in this study. Cas9 and tracrRNA encoding sequences are colored in blue and red accordingly. Predicted transcriptional directions of the tracrRNA and CRISPR arrays are indicated by arrows. Clade membership is indicated with roman numerals. Additional genes, *cas1*, *cas2*, *csn2* and *cas4*, are color-coded in cyan, green, yellow, and pink, respectively.



Supplementary Fig. 3. **Predicted guide RNA secondary structures for the Cas9 orthologs described in this study**. Phylogenetic clade membership is indicated with roman numerals. The single guide RNA (sgRNA) structures can be classified into at least 6 groups based on structural prediction and exemplified by Tde, Seq1 (Spy-like), Sth1A, Nme2, Nsa and Cca1 (names are underlined). Main structural elements are color-coded: repeat:anti-repeat duplex is indicated in green, the nexus in purple and 3' tracrRNA hairpin(s) in blue. Secondary structures of tracrRNA visualized with Varna v 3-93[1]



Supplementary Fig.4. Verification of IVT-based methods for characterizing Cas9 PAM recognition. **a** PAM preferences for Spy, Sth3 and Sth1 Cas9 proteins using different dilutions of *in vitro* translated (IVT) ribonucleoprotein (RNP) mixtures. **b** PAM recognition for 11 new orthologs assayed using both IVT and purified RNP.

a Query C7_ID46__WP_148224960.1__ADV46720.1
Match_columns 306
No_of_seqs 35 out of 39
Neff 3.64315
Searched_HMMs 49849
Date Fri Jan 24 13:02:17 2020
Command hhsearch -cpu 8 -i ../results/full.a3m -d /cluster/toolkit/production/databases/hh-suite/mmcif70/pdb70 -o ../results/9327865_5.hhr -oa3m ../results/9327865_5.a3m -p 20 -Z 250 -loc -z 1 -b 1 -B 250 -ssm 2 -sc 1 -seq 1 -dbstrlen 10000 -norealign -maxres 32000 -contxt /cluster/toolkit/production/bioprogs/tools/hh-suite-build/data/context_data.crf

No	Hit	Prob	E-value	P-value	Score	SS	Cols	Query HMM	Template HMM
1	5CZZ_A CRISPR-associated endon	96.9	0.00033	6.7E-09	76.1	12.0	226	5-274	774-1041(1056)
2	6KC8_A CRISPR-associated endon	96.5	0.0038	7.7E-08	68.4	15.2	200	8-274	830-1070(1083)
3	6RJA_C AcrIIA6, CRISPR-associa	96.2	0.0088	1.8E-07	66.0	14.2	250	5-274	810-1107(1121)
4	6JOO_A CRISPR-associated prote	87.5	1.2	2.3E-05	48.9	6.1	119	4-180	645-770 (927)
5	5X2H_A CRISPR-associated endon	86.8	1.7	3.4E-05	47.1	6.8	79	193-274	737-822 (835)
6	5TGY_A PS1; 4-helix bundle, co	39.6	2.1E+02	0.0043	24.1	6.6	70	54-125	31-100 (109)

...

b Query C8_WP_087368840.1
Match_columns 299
No_of_seqs 103 out of 122
Neff 5.0316
Searched_HMMs 49849
Date Fri Jan 24 12:53:48 2020
Command hhsearch -cpu 8 -i ../results/full.a3m -d /cluster/toolkit/production/databases/hh-suite/mmcif70/pdb70 -o ../results/9327865_2.hhr -oa3m ../results/9327865_2.a3m -p 20 -Z 250 -loc -z 1 -b 1 -B 250 -ssm 2 -sc 1 -seq 1 -dbstrlen 10000 -norealign -maxres 32000 -contxt /cluster/toolkit/production/bioprogs/tools/hh-suite-build/data/context_data.crf

No	Hit	Prob	E-value	P-value	Score	SS	Cols	Query HMM	Template HMM
1	5CZZ_A CRISPR-associated endon	99.8	8.5E-23	1.7E-27	220.1	16.6	230	46-296	768-1055(1056)
2	6KC8_A CRISPR-associated endon	99.6	2E-17	4E-22	179.3	18.7	264	1-283	772-1071(1083)
3	4OGE_A HNH endonuclease domain	99.1	7.2E-13	1.4E-17	144.6	10.1	226	37-293	776-1097(1101)
4	6JOO_A CRISPR-associated prote	97.9	1.3E-06	2.7E-11	94.6	12.9	239	27-293	622-923 (927)
5	5X2H_A CRISPR-associated endon	97.8	2.3E-06	4.7E-11	91.7	13.2	177	42-283	616-823 (835)
6	6RJA_C AcrIIA6, CRISPR-associa	97.2	0.00038	7.6E-09	77.2	18.1	213	42-283	798-1108(1121)
7	3BID_H UPF0339 protein NMB1088	35.9	78	0.0016	23.0	3.3	18	174-191	2-19 (64)

...

c Query C9_NCBI_ASWY01000120:2294253111_16
Match_columns 232
No_of_seqs 10 out of 12
Neff 3.60828
Searched_HMMs 49849
Date Fri Jan 24 12:54:45 2020
Command hhsearch -cpu 8 -i ../results/full.a3m -d /cluster/toolkit/production/databases/hh-suite/mmcif70/pdb70 -o ../results/9327865_3.hhr -oa3m ../results/9327865_3.a3m -p 20 -Z 250 -loc -z 1 -b 1 -B 250 -ssm 2 -sc 1 -seq 1 -dbstrlen 10000 -norealign -maxres 32000 -contxt /cluster/toolkit/production/bioprogs/tools/hh-suite-build/data/context_data.crf

No	Hit	Prob	E-value	P-value	Score	SS	Cols	Query HMM	Template HMM
1	6KC8_A CRISPR-associated endon	95.0	0.059	1.2E-06	56.9	10.7	174	18-211	845-1033(1083)
2	5CZZ_A CRISPR-associated endon	94.3	0.093	1.9E-06	55.2	9.0	161	18-184	788-980 (1056)
3	2ZDO_B Heme-degrading monooxyg	73.1	35	0.0007	23.6	7.2	61	98-160	6-68 (109)
4	1X7V_A PA3566 protein; structu	71.9	21	0.00041	23.4	5.3	62	100-162	10-71 (99)
5	6DS7_A Heme-degrading monooxyg	70.8	28	0.00055	23.1	5.8	64	98-162	3-66 (99)

...

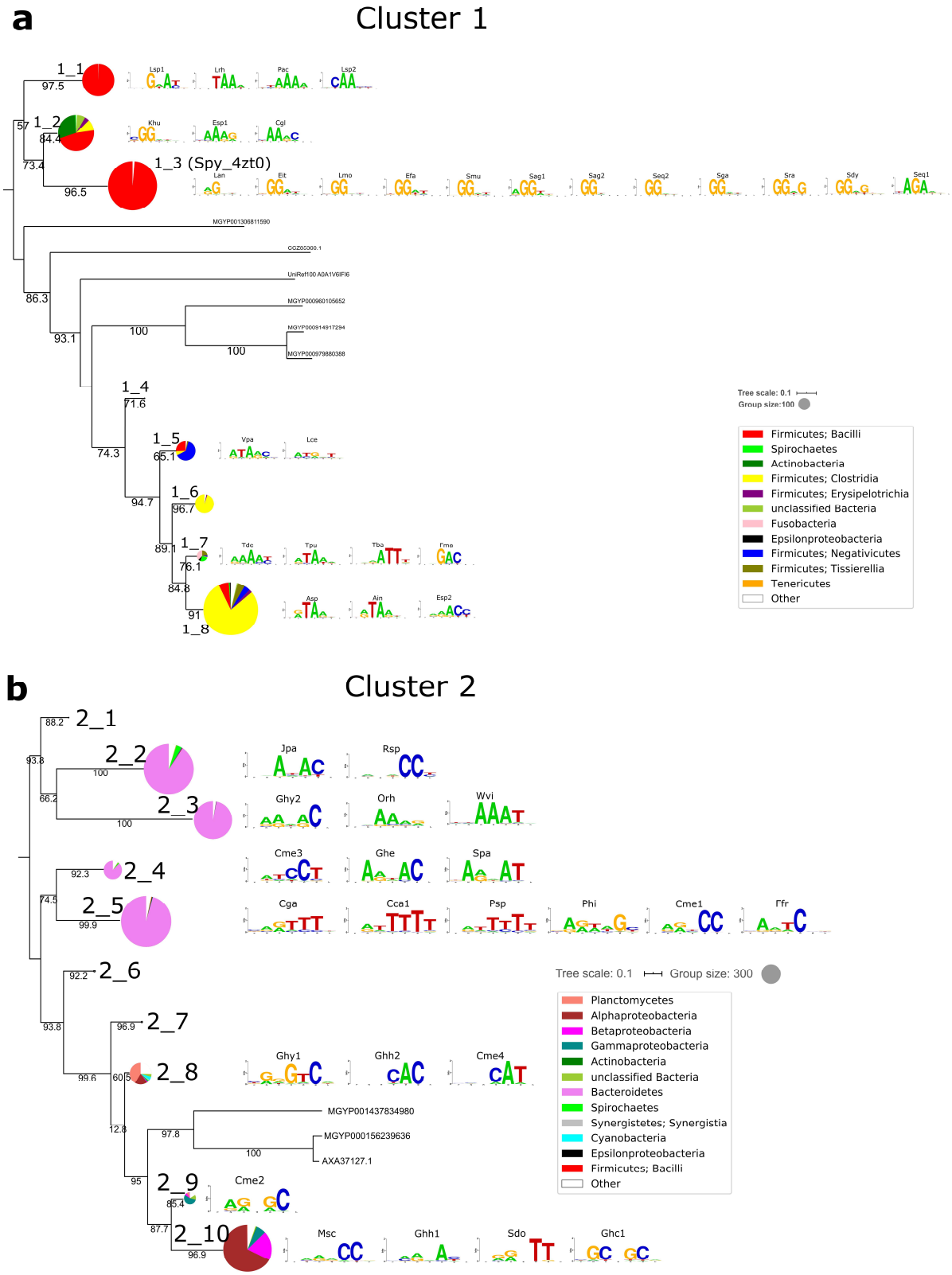
d Query C10_WP_009541330.1__ID16
Match_columns 415
No_of_seqs 4 out of 6
Neff 1.64058
Searched_HMMs 80583
Date Fri Jan 24 13:06:48 2020
Command hhsearch -cpu 8 -i ../results/full.a3m -d /cluster/toolkit/production/databases/hh-suite/mmcif70/pdb70 -d /cluster/toolkit/production/databases/hh-suite/pfama/pfama -d /cluster/toolkit/production/databases/hh-suite/NCBI_CD/NCBI_CD -o ../results/9327865_7.hhr -oa3m ../results/9327865_7.a3m -p 20 -Z 250 -loc -z 1 -b 1 -B 250 -ssm 2 -sc 1 -seq 1 -dbstrlen 10000 -norealign -maxres 32000 -contxt /cluster/toolkit/production/bioprogs/tools/hh-suite-build/data/context_data.crf

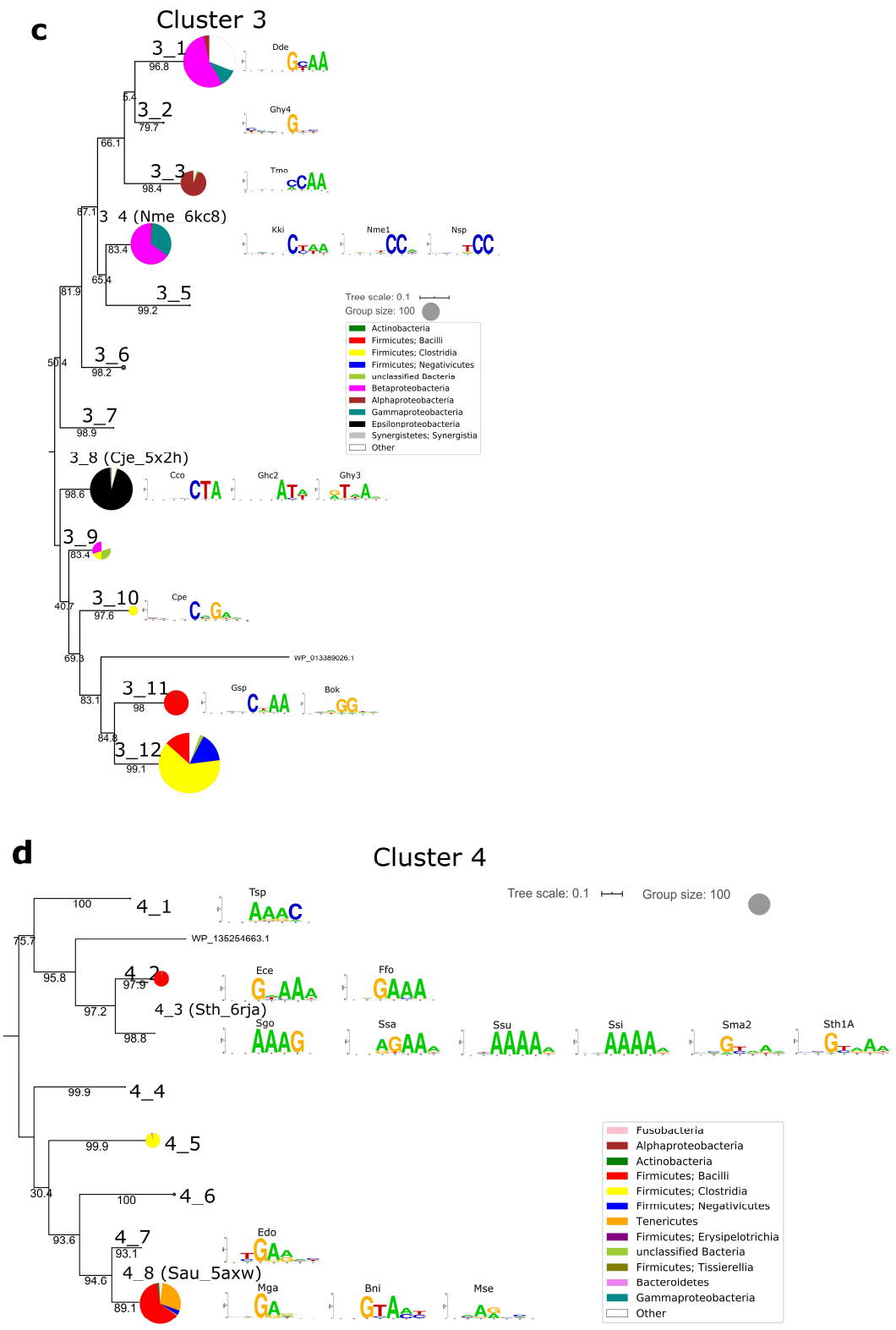
No	Hit	Prob	E-value	P-value	Score	SS	Cols	Query HMM	Template HMM
1	PF13856.6 ; Gifsy-2 ; ATP-bind	71.9	12	0.00015	28.9	3.4	27	355-382	70-96 (99)
2	PF05354.11 ; Phage_attach ; Ph	71.0	12	0.00015	30.3	3.3	27	354-380	74-100 (117)
3	1K0H_A gpFII; twisted beta-san	64.8	15	0.00018	29.9	2.6	18	354-371	74-91 (117)

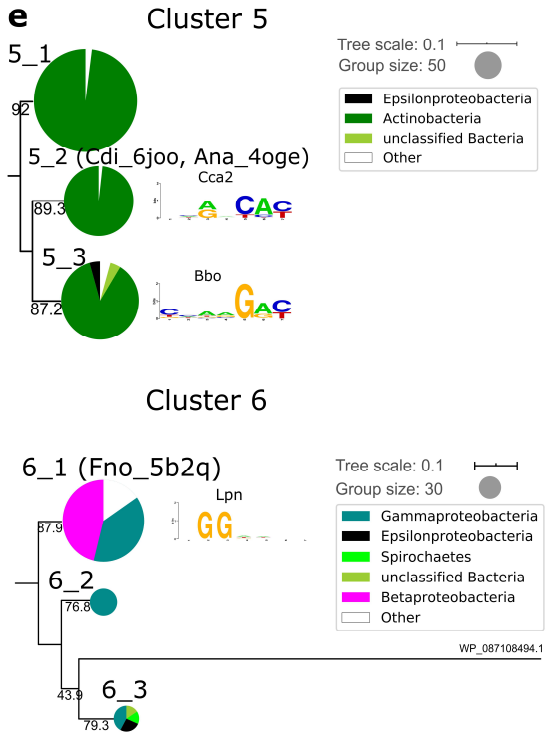
...

Cluster3 Cluster4

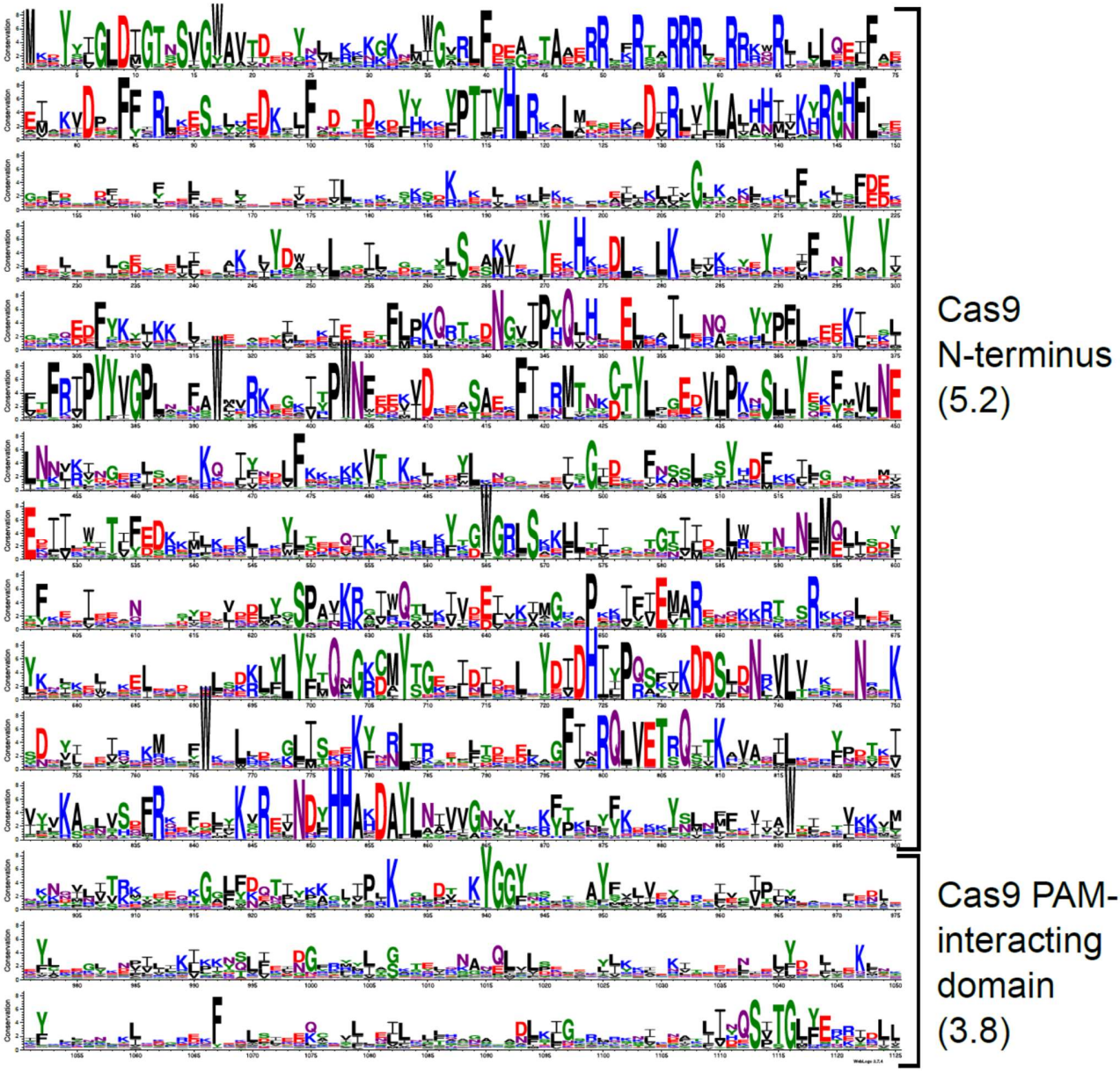
Supplementary Fig. 5. HHpred results of searches using representative members from clusters 7 -a, 8 - b, 9 -c and 10 - d.


a
Cluster 1
1_1
97.5
1_2
84.4
57
73.4
1_3 (Spy_4zt0)
96.5
MGYP001306811590
OCZ80360.1
UniRef100 A0A1V6IFI6
MGYP000960105652
MGYP000914917294
MGYP000979880388
86.3
93.1
100
100
1_4
71.6
74.3
1_5
65.1
94.7
1_6
96.7
89.1
1_7
76.1
84.8
91
1_8
Tree scale: 0.1
Group size:100
Firmicutes; Bacilli
Spirochaetes
Actinobacteria
Firmicutes; Clostridia
Firmicutes; Erysipelotrichia
unclassified Bacteria
Fusobacteria
Epsilonproteobacteria
Firmicutes; Negativicutes
Firmicutes; Tissierellia
Tenericutes
Other
b
Cluster 2
2_1
88.2
93.8
2_2
100
66.2
2_3
100
92.3
2_4
74.5
2_5
99.9
92.2
2_6
93.8
96.9
2_7
99.6
60.5
2_8
12.8
97.8
MGYP001437834980
MGYP000156239636
100
AXA37127.1
95
2_9
85.4
87.7
2_10
96.9
Tree scale: 0.1
Group size: 300
Planctomycetes
Alphaproteobacteria
Betaproteobacteria
Gammaproteobacteria
Actinobacteria
unclassified Bacteria
Bacteroidetes
Spirochaetes
Synergistetes; Synergistia
Cyanobacteria
Epsilonproteobacteria
Firmicutes; Bacilli
Other

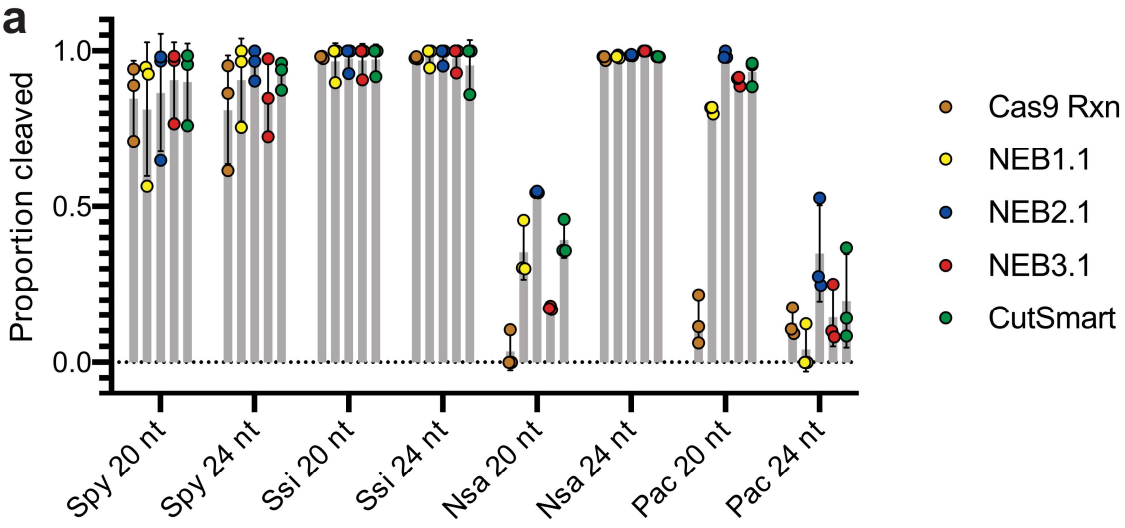

c
Cluster 3
3_1
96.8
Dde
3_2
79.7
Ghy4
3_3
98.4
Tmo
3_4 (Nme_6kc8)
83.4
Kki
Nme1
Nsp
3_5
99.2
3_6
98.2
3_7
98.9
3_8 (Cje_5x2h)
98.6
Cco
Ghc2
Ghy3
3_9
83.4
3_10
97.6
Cpe
WP_013389026.1
3_11
98
Gsp
Bok
3_12
99.1
Tree scale: 0.1
Group size: 100
Actinobacteria
Firmicutes; Bacilli
Firmicutes; Clostridia
Firmicutes; Negativicutes
unclassified Bacteria
Betaproteobacteria
Alphaproteobacteria
Gammaproteobacteria
Epsilonproteobacteria
Synergistetes; Synergistia
Other
d
Cluster 4
4_1
100
Tsp
WP_135254663.1
4_2
97.9
Ece
Ffo
4_3 (Sth_6rja)
98.8
Sgo
Ssa
Ssu
Ssi
Sma2
Sth1A
4_4
99.9
4_5
99.9
4_6
100
4_7
93.1
Edo
4_8 (Sau_5axw)
89.1
Mga
Bni
Mse
Tree scale: 0.1
Group size: 100
Fusobacteria
Alphaproteobacteria
Actinobacteria
Firmicutes; Bacilli
Firmicutes; Clostridia
Firmicutes; Negativicutes
Tenericutes
Firmicutes; Erysipelotrichia
unclassified Bacteria
Firmicutes; Tissierellia
Bacteroidetes
Gammaproteobacteria
Other

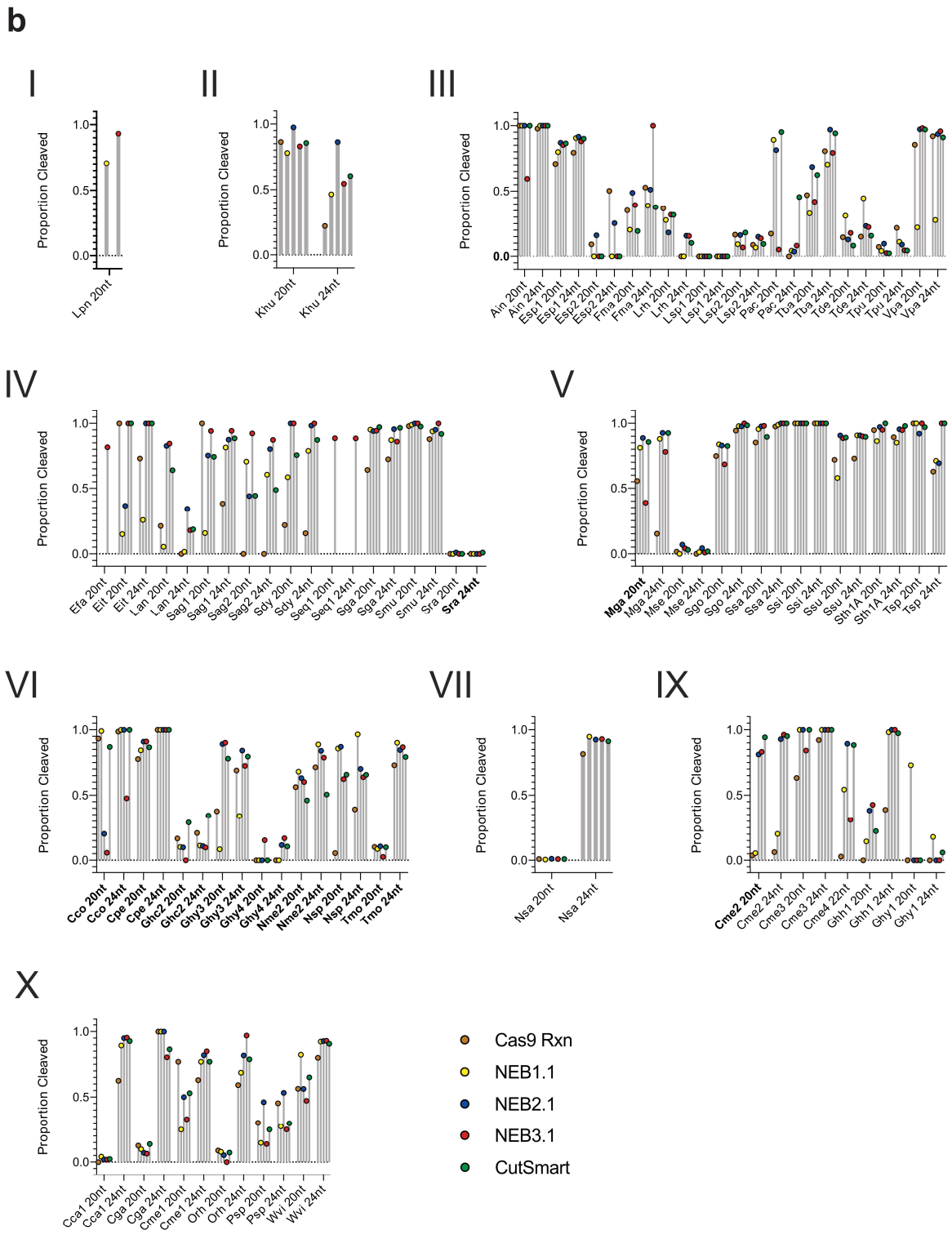


Supplementary Fig. 6. Phylogenetic tree of PI domains from clusters 1 (a), 2 (b), 3 (c), 4 (d) and 5&6 (e). SH-aLRT bootstrap values are shown at the branches. Pie charts show taxonomic diversity of groups. Color codes for taxa and scale of a pie chart are provided in each panel.

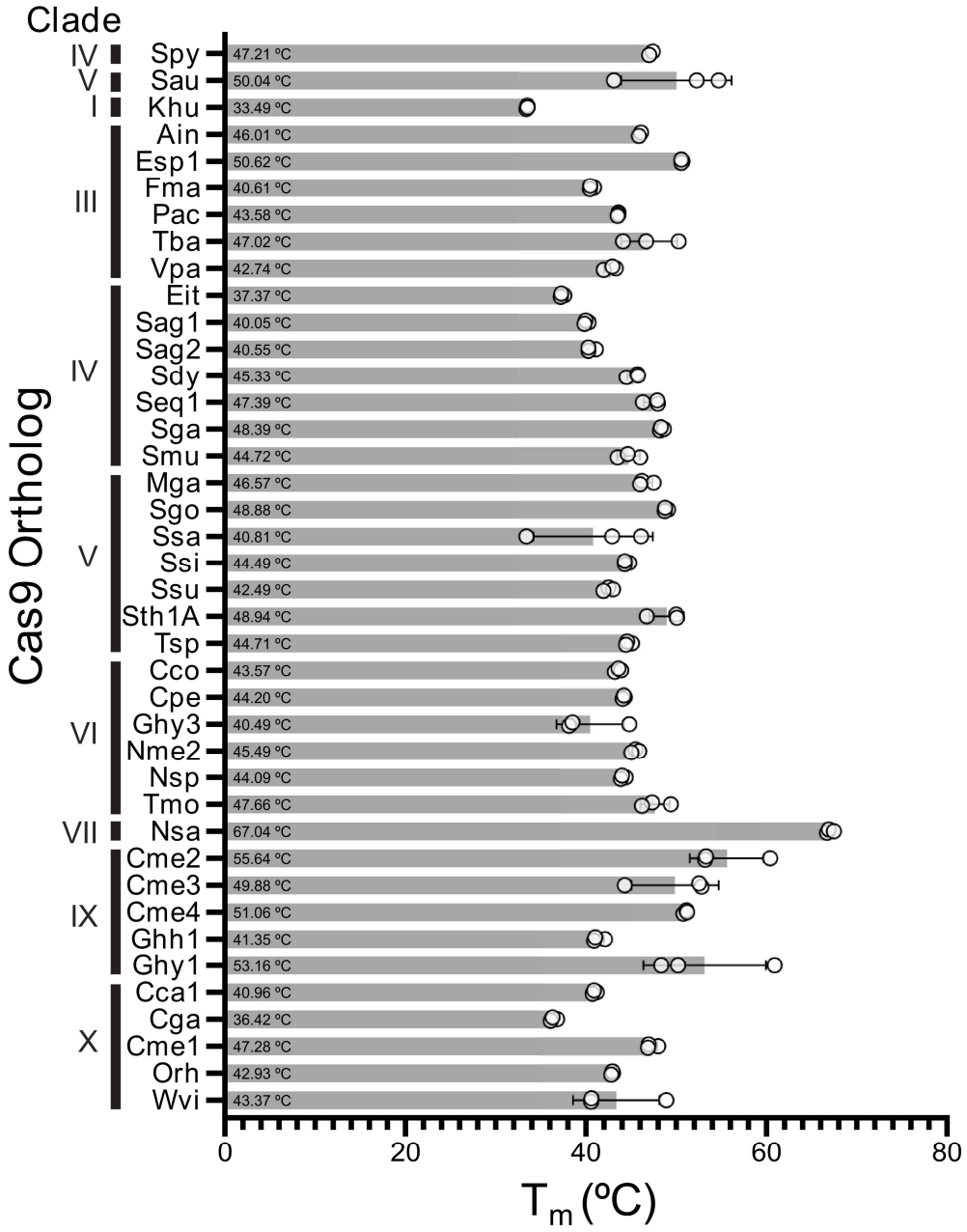


Supplementary Fig. 7. Weblogo of Spy Cas9 homologs. Only positions having Jalview conservation score of 1 and higher are shown. Average conservation score is shown in parenthesis. Sequences having all domains characteristic to Spy Cas9 from Cluster 1 (Supplementary Data 3) were used to build the MAFFT alignment.

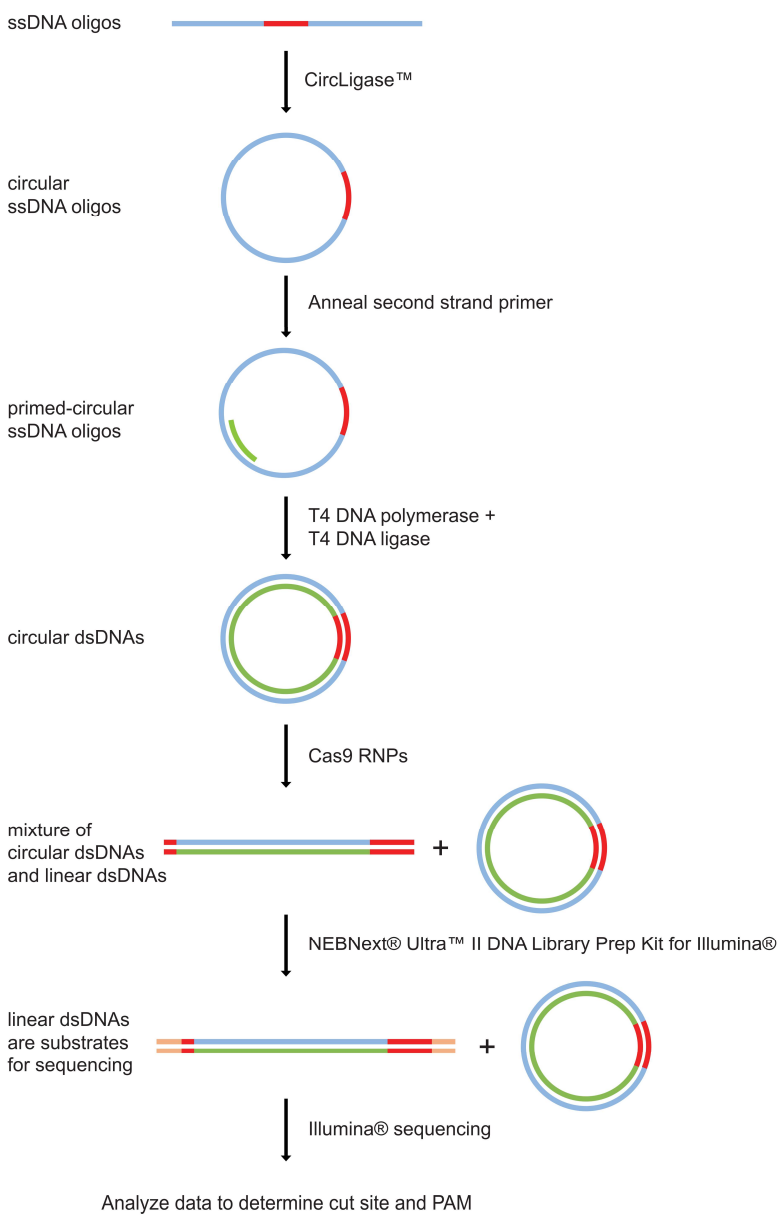

a
Proportion cleaved
1.0
0.5
0.0
Spy 20 nt
Spy 24 nt
Ssi 20 nt
Ssi 24 nt
Nsa 20 nt
Nsa 24 nt
Pac 20 nt
Pac 24 nt
Cas9 Rxn
NEB1.1
NEB2.1
NEB3.1
CutSmart


b
I
II
III
IV
V
VI
VII
IX
X
Proportion Cleaved
1.0
0.5
0.0
Lpn 20nt
Khu 20nt
Khu 24nt
Ain 20nt
Ain 24nt
Esp1 20nt
Esp1 24nt
Esp2 20nt
Esp2 24nt
Fma 20nt
Fma 24nt
Lfh 20nt
Lfh 24nt
Lsp1 20nt
Lsp1 24nt
Lsp2 20nt
Lsp2 24nt
Pac 20nt
Pac 24nt
Tba 20nt
Tba 24nt
Tde 20nt
Tde 24nt
Tpu 20nt
Tpu 24nt
Vpa 20nt
Vpa 24nt
Efa 20nt
Eit 20nt
Eit 24nt
Lan 20nt
Lan 24nt
Sag1 20nt
Sag1 24nt
Sag2 20nt
Sag2 24nt
Sdy 20nt
Sdy 24nt
Seq1 20nt
Seq1 24nt
Sga 20nt
Sga 24nt
Smu 20nt
Smu 24nt
Sra 20nt
Sra 24nt
Mga 20nt
Mga 24nt
Mse 20nt
Mse 24nt
Sgo 20nt
Sgo 24nt
Ssa 20nt
Ssa 24nt
Ssi 20nt
Ssi 24nt
Ssu 20nt
Ssu 24nt
Sth1A 20nt
Sth1A 24nt
Tsp 20nt
Tsp 24nt
Cco 20nt
Cco 24nt
Cpe 20nt
Cpe 24nt
Ghc2 20nt
Ghc2 24nt
Ghy3 20nt
Ghy3 24nt
Ghy4 20nt
Ghy4 24nt
Nme2 20nt
Nme2 24nt
Nsp 20nt
Nsp 24nt
Tmo 20nt
Tmo 24nt
Nsa 20nt
Nsa 24nt
Cme2 20nt
Cme2 24nt
Cme3 20nt
Cme3 24nt
Cme4 22nt
Ghh1 20nt
Ghh1 24nt
Ghy1 20nt
Ghy1 24nt
Cca1 20nt
Cca1 24nt
Cga 20nt
Cga 24nt
Cme1 20nt
Cme1 24nt
Orh 20nt
Orh 24nt
Psp 20nt
Psp 24nt
Wvi 20nt
Wvi 24nt
Cas9 Rxn
NEB1.1
NEB2.1
NEB3.1
CutSmart

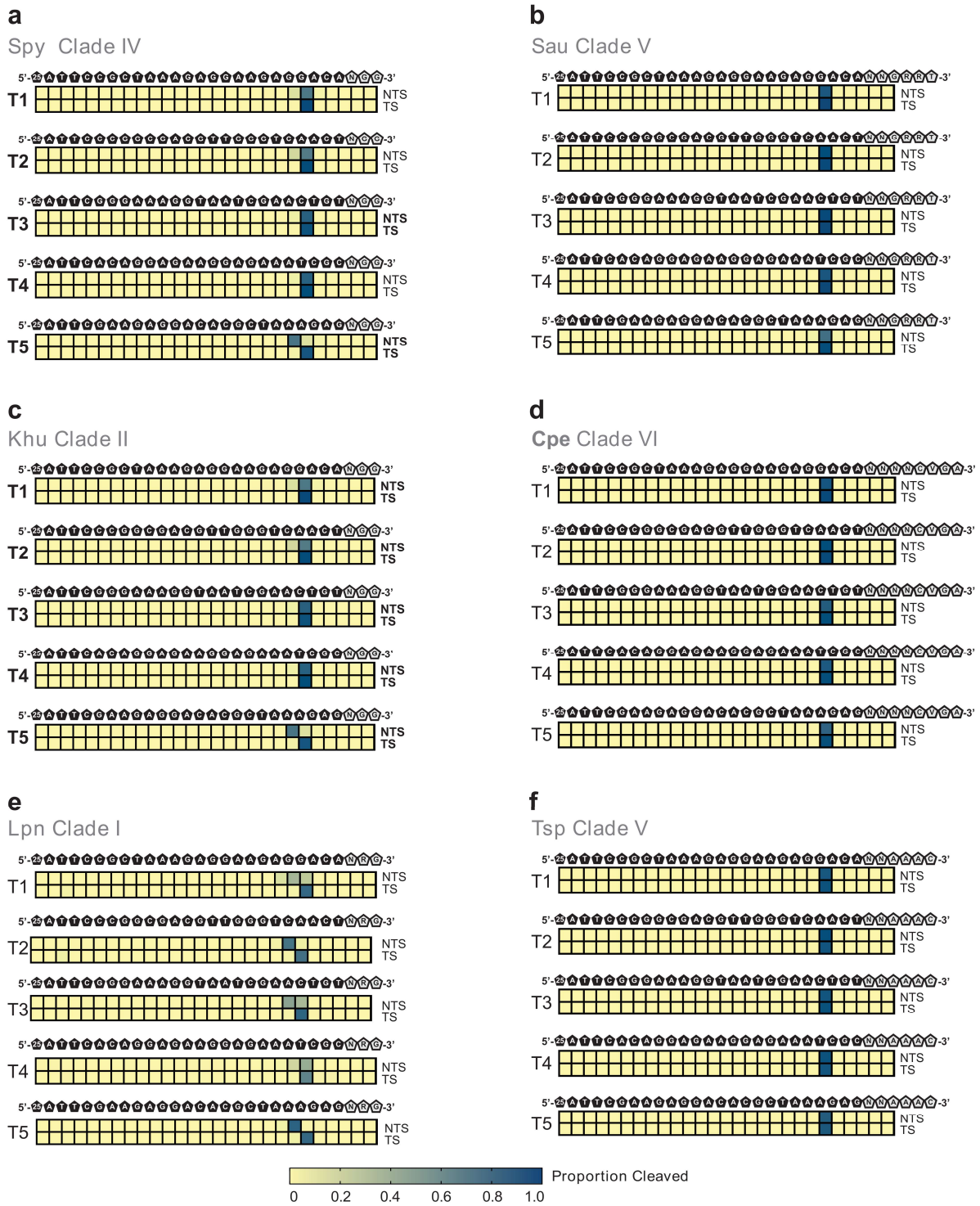
Supplementary Fig. 8. Spacer length preferences for purified Cas9 orthologs. a The cleavage activity of selected Cas9 orthologs was examined using sgRNAs with 20 or 24 nt spacers. *In vitro* DNA cleavage reactions were performed in different buffers at 37ºC (see Methods). Spy and Ssi Cas9 do not have a preference for 20 or 24 nt spacers. Nsa Cas9 has a preference for 24 nt spacers whereas Pac Cas9 has a preference for 20 nt spacers. Data from three independent experiments are shown as points and error bars representing standard deviation The center of the error bars denotes the mean of the measurements. Brown, yellow, blue, red, and green circles represent reactions performed in Cas9 reaction buffer, NEBuffer 1.1, NEBuffer 2.1, NEBuffer 3.1, or CutSmart buffer, respectively **b** The *in vitro* DNA cleavage activity of purified Cas9 proteins was screened using 20 and 24 nt spacers. Each point represents one cleavage determination in multiple buffers conditions as indicated by color. Brown, yellow, blue, red, and green circles represent reactions performed in Cas9 reaction buffer, NEBuffer 1.1, NEBuffer 2.1, NEBuffer 3.1, or CutSmart buffer, respectively. Clade membership is indicated by roman numerals.

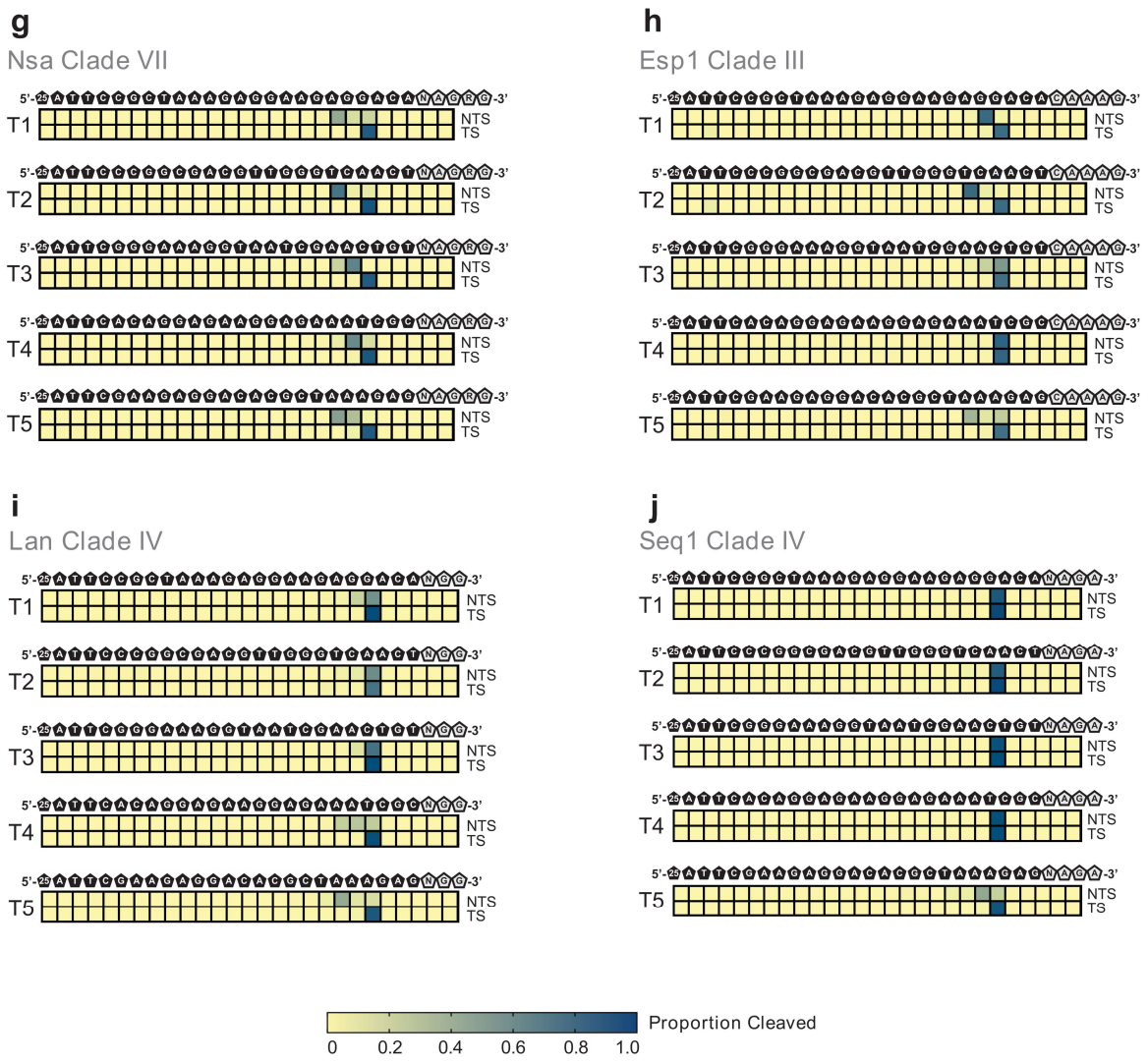


Supplementary Fig. 9. Thermostability of purified Cas9 proteins. Melting temperatures of purified Cas9 proteins was determined by nanoDSF (see Methods). Tm indicated in the histograms represent the mean of 3 independent determinations plotted as open circles. Error bars indicate standard deviation with the center of each denoting the mean of the determinations.

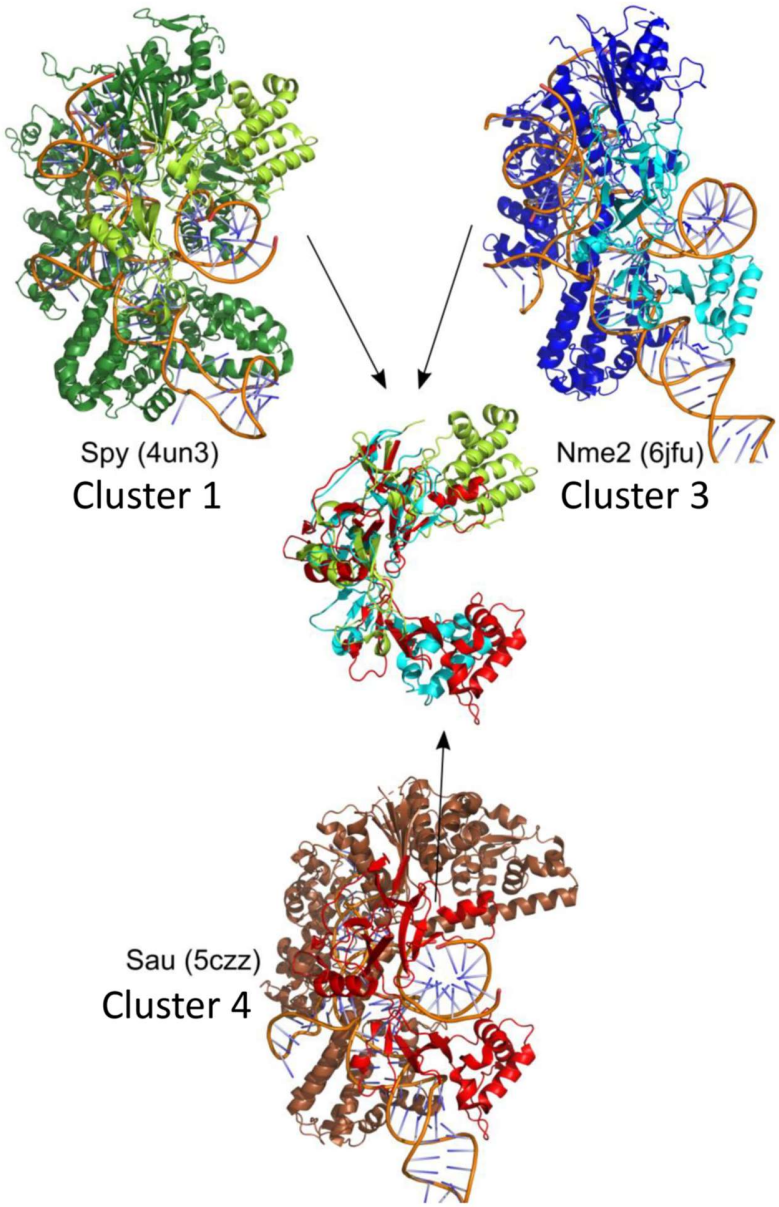


Supplementary Fig. 10. PAM discovery and cleavage site mapping using purified Cas9 RNP and dsDNA minicircles. Diagram of workflow to make small circular substrates for PAM and cut site determination assays. Pools of single stranded DNA oligonucleotides (blue) containing a 5' PO_4 ends and 10 nt of randomized sequence (red) were circularized and converted to double stranded circles after annealing a second strand primer (green) and extending it with polymerase in the presence of a ligase. DNA circles are then exposed to Cas9 RNPs containing guide RNA complementary to the DNA sequence flanking the 10 nt randomized sequence region. DNA circles that are cut by the Cas9 are substrates for ligation of adapters (orange) for Illumina sequencing. Custom bioinformatic analysis was used to determine cut site pattern and PAM preference. Control reactions were carried out where each pool of circles was cut with the restriction enzyme BstXI to determine the baseline frequency of nucleotides at every position in the randomized region.


a
Spy Clade IV
T1
T2
T3
T4
T5
b
Sau Clade V
c
Khu Clade II
d
Cpe Clade VI
e
Lpn Clade I
f
Tsp Clade V
NTS
TS
Proportion Cleaved
0
0.2
0.4
0.6
0.8
1.0



Supplementary Fig. 11. Target-dependent cleavage patterns. Diverse cleavage sites over multiple targets for a single Cas9 ortholog are depicted as heatmaps depicting mapped cleavage ends at each position in a single DNA target. Sequence of the DNA non-target strand and PAM is depicted above the heatmaps. Intensity of the blue color indicates the proportion of mapped cleavage ends. NTS indicates non-target strand; TS indicates target strand. **a** Spy Cas9 cleavage patterns. **b** Sau Cas9 cleavage patterns. **c** Khu Cas9 cleavage patterns. **d** Cpe Cas9 cleavage patterns. **e** Lpn cleavage patterns. **f** Tsp Cas9 cleavage patterns. **g** Nsa cleavage patterns. **h** Esp1 Cas9 cleavage patterns. **i** Lan Cas9 cleavage patterns. **j** Seq1 Cas9 cleavage patterns. Source data are provided in Supplementary Data 5.



Supplementary Fig. 12. Comparison of PAM Interacting (PI) domain structures from 3 major clusters (Fig. 3): Spy (4zto), Nmen (6kc8), Sau (5axw). There are no solved structures of Cas9 proteins with a PI domain associated with cluster 2.

Supplementary References

1. Darty, K., Denise, A. & Ponty, Y. VARNA: Interactive drawing and editing of the RNA secondary structure. *Bioinformatics* **25**, 1974–1975 (2009).