KH-domain proteins: another family of bacterial RNA matchmakers?

Mikolaj Olejniczak[1,*], Xiaofang Jiang[2], Maciej M. Basczok[1], Gisela Storz[3,]**

[1]*Institute of Molecular Biology and Biotechnology, Faculty of Biology, Adam Mickiewicz University, Uniwersytetu Poznańskiego 6, 61-614 Poznań, Poland.*

[2]*Intramural Research Program, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA.*

[3]*Division of Molecular and Cellular Biology, Eunice Kennedy Shriver National Institutes of Child Health and Human Development, Bethesda, MD 20892-4417, USA.*

For correspondences. [*]E-mail mol@amu.edu.pl and [**]E-mail storzg@mail.nih.gov

Supplementary Information

Table S1. Sequences and taxonomic information on KhpA and KhpB homologs identified in Genome Taxonomy Database (GTDB) (release 202).

Protein families and domains were identified using InterProScan with default parameters. Proteins that contain only the conserved KH domain (PTHR34654 or MF_00088) were identified as KhpA. Three conserved domains were used to identify KhpB genes: the Jag-N domain (PF14804, SM01245 or G3DSA:3.30.30.80), the KH domain (cd02414, G3DSA:3.30.300.20 or PF13083), and the R3H domain (PS51061, SM00393, PF01424, cd02644, G3DSA:3.30.1370.50 or SSF82708). Manual curation together with synteny information was used to improve the predictions. Some possible contaminants are noted in the .xlsx table.

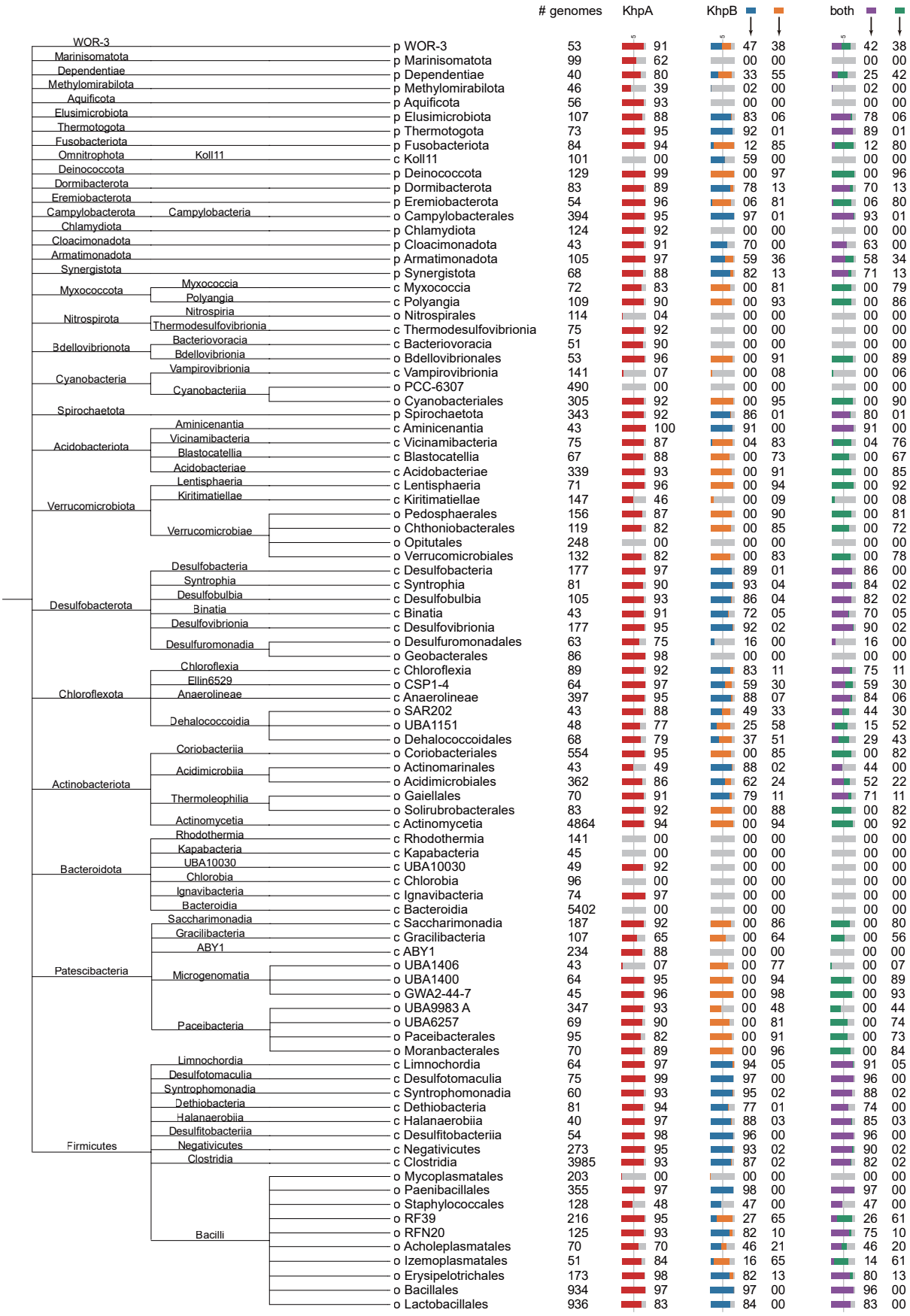
WOR-3
Marinisomatota
Dependentiae
Methylomirabilota
Aquificota
Elusimicrobiota
Thermotogota
Fusobacteriota
Omnitrophota
Deinococcota
Dormibacterota
Eremiobacterota
Campylobacterota
Chlamydiota
Cloacimonadota
Armatimonadota
Synergistota
Myxococcota
Nitrospirota
Bdellovibrionota
Cyanobacteria
Spirochaetota
Acidobacteriota
Verrucomicrobiota
Desulfobacterota
Chloroflexota
Actinobacteriota
Bacteroidota
Patescibacteria
Firmicutes
Koll11
Campylobacteria
Myxococcia
Polyangia
Nitrospiria
Thermodesulfovibrionia
Bacteriovoracia
Bdellovibrionia
Vampirovibrionia
Cyanobacteriia
Aminicenantia
Vicinamibacteria
Blastocatellia
Acidobacteriae
Lentisphaeria
Kiritimatiellae
Verrucomicrobiae
Desulfobacteria
Syntrophia
Desulfobulbia
Binatia
Desulfovibrionia
Desulfuromonadia
Chloroflexia
Ellin6529
Anaerolineae
Dehalococcoidia
Coriobacteriia
Acidimicrobiia
Thermoleophilia
Actinomycetia
Rhodothermia
Kapabacteria
UBA10030
Chlorobia
Ignavibacteria
Bacteroidia
Saccharimonadia
Gracilibacteria
ABY1
Microgenomatia
Paceibacteria
Limnochordia
Desulfotomaculia
Syntrophomonadia
Dethiobacteria
Halanaerobiia
Desulfitobacteriia
Negativicutes
Clostridia
Bacilli
p WOR-3 53 91 47 38 42 38
p Marinisomatota 99 62 00 00 00 00
p Dependentiae 40 80 33 55 25 42
p Methylomirabilota 46 39 02 00 02 00
p Aquificota 56 93 00 00 00 00
p Elusimicrobiota 107 88 83 06 78 06
p Thermotogota 73 95 92 01 89 01
p Fusobacteriota 84 94 12 85 12 80
c Koll11 101 00 59 00 00 00
p Deinococcota 129 99 00 97 00 96
p Dormibacterota 83 89 78 13 70 13
p Eremiobacterota 54 96 06 81 06 80
o Campylobacterales 394 95 97 01 93 01
p Chlamydiota 124 92 00 00 00 00
p Cloacimonadota 43 91 70 00 63 00
p Armatimonadota 105 97 59 36 58 34
p Synergistota 68 88 82 13 71 13
c Myxococcia 72 83 00 81 00 79
c Polyangia 109 90 00 93 00 86
o Nitrospirales 114 04 00 00 00 00
c Thermodesulfovibrionia 75 92 00 00 00 00
c Bacteriovoracia 51 90 00 00 00 00
o Bdellovibrionales 53 96 00 91 00 89
c Vampirovibrionia 141 07 00 08 00 06
o PCC-6307 490 00 00 00 00 00
o Cyanobacteriales 305 92 00 95 00 90
p Spirochaetota 343 92 86 01 80 01
c Aminicenantia 43 100 91 00 91 00
c Vicinamibacteria 75 87 04 83 04 76
c Blastocatellia 67 88 00 73 00 67
c Acidobacteriae 339 93 00 91 00 85
c Lentisphaeria 71 96 00 94 00 92
c Kiritimatiellae 147 46 00 09 00 08
o Pedosphaerales 156 87 00 90 00 81
o Chthoniobacterales 119 82 00 85 00 72
o Opitutales 248 00 00 00 00 00
o Verrucomicrobiales 132 82 00 83 00 78
c Desulfobacteria 177 97 89 01 86 00
c Syntrophia 81 90 93 04 84 02
c Desulfobulbia 105 93 86 04 82 02
c Binatia 43 91 72 05 70 05
c Desulfovibrionia 177 95 92 02 90 02
o Desulfuromonadales 63 75 16 00 16 00
o Geobacterales 86 98 00 00 00 00
c Chloroflexia 89 92 83 11 75 11
o CSP1-4 64 97 59 30 59 30
c Anaerolineae 397 95 88 07 84 06
o SAR202 43 88 49 33 44 30
o UBA1151 48 77 25 58 15 52
o Dehalococcoidales 68 79 37 51 29 43
o Coriobacteriales 554 95 00 85 00 82
o Actinomarinales 43 49 88 02 44 00
o Acidimicrobiales 362 86 62 24 52 22
o Gaiellales 70 91 79 11 71 11
o Solirubrobacterales 83 92 00 88 00 82
c Actinomycetia 4864 94 00 94 00 92
c Rhodothermia 141 00 00 00 00 00
c Kapabacteria 45 00 00 00 00 00
c UBA10030 49 92 00 00 00 00
c Chlorobia 96 00 00 00 00 00
c Ignavibacteria 74 97 00 00 00 00
c Bacteroidia 5402 00 00 00 00 00
c Saccharimonadia 187 92 00 86 00 80
c Gracilibacteria 107 65 00 64 00 56
c ABY1 234 88 00 00 00 00
o UBA1406 43 07 00 77 00 07
o UBA1400 64 95 00 94 00 89
o GWA2-44-7 45 96 00 98 00 93
o UBA9983 A 347 93 00 48 00 44
o UBA6257 69 90 00 81 00 74
o Paceibacterales 95 82 00 91 00 73
o Moranbacterales 70 89 00 96 00 84
c Limnochordia 64 97 94 05 91 05
c Desulfotomaculia 75 99 97 00 96 00
c Syntrophomonadia 60 93 95 02 88 02
c Dethiobacteria 81 94 77 01 74 00
c Halanaerobiia 40 97 88 03 85 03
c Desulfitobacteriia 54 98 96 00 96 00
c Negativicutes 273 95 93 02 90 02
c Clostridia 3985 93 87 02 82 02
o Mycoplasmatales 203 00 00 00 00 00
o Paenibacillales 355 97 98 00 97 00
o Staphylococcales 128 48 47 00 47 00
o RF39 216 95 27 65 26 61
o RFN20 125 93 82 10 75 10
o Acholeplasmatales 70 70 46 21 46 20
o Izemoplasmatales 51 84 16 65 14 61
o Erysipelotrichales 173 98 82 13 80 13
o Bacillales 934 97 97 00 96 00
o Lactobacillales 936 83 84 00 83 00

Fig. S1. The taxonomic distribution of KhpA and KhpB based on the Genome Taxonomy Database (GTDB) (release 202).

Based on Table S1. The numbers to the right of the cladogram represent the number of genomes examined in the corresponding taxa. Only clades with 40 or more genomes were included in the diagram. The red bar represents the percentage of genomes with KhpA genes in the taxa; the blue and orange bars represent the percentage of genomes with KhpB gene with and without the Jag-N domain, respectively; the purple bar represents the percentage of genomes with both KhpA and KhpB with the Jag-N domain; the green bar represents the percentage of genomes with both KhpA and KhpB without the Jag-N domain.

Fig. S2. Gene synteny surrounding *khpA*.

Orthologous gene families surrounding *khpA* genes for organisms shown in Fig. 1 were identified using OrthoFinder (version 2.3.8). Genome synteny was visualized with a custom script. The large triangle indicates the position of a gene encoding the 16S rRNA.

Fig. S3. Gene synteny surrounding *khpB*.

Orthologous gene families surrounding *khpB* genes for organisms shown in Fig. 1 were identified using OrthoFinder (version 2.3.8). Genome synteny was visualized with a custom script. The large triangle indicates the position of a gene encoding the 16S rRNA.