Supporting information

Figure S1. **A)** *G. muris* cells propagate as trophozoites in the mice small intestine. The cell trophozoites undergo encystation into cysts and are shed in the feces. The cysts are prompted to excyst as they encounter the conditions of the digestive tract. The excyzoites transition into the vegetative trophozoites. Cell material for genomic DNA was harvested from *G. muris* cysts from infected mice. **B)** Sub-telomeric regions in *G. muris*. Sequence motifs in the 10 unitigs terminating in $(TAGGG)^n$ telomeric sequences (purple). Ribosomal DNA sequences (black). Satellite-like repeated sequences (green).

Figure S2. Chromosome-scale synteny between *G. muris* and *G. intestinalis* WB. *G. muris* chromosomes are color coded. The links between the two genomes representing sequence similarity identified by MUMmer promer[96], are colored after the *G. muris* chromosomal color. Circular plot was drawn with circlize[94].

Figure S3. Sequence logo around start codon (A) and stop codon (B) in *G. muris*. C) *G. muris* gene expression from *in vivo* trophozoites. Genes showing top-ranked FPKM ratios (log2 (Trophs/Cysts)) between *in vivo* derived trophozoites and cysts (cut-off value ≥4). The value is indicated in the teal circle. An orange dot indicates that the ortholog of this gene is upregulated (>2 fold) in *G. intestinalis* at 7 hrs, 12 hrs or 22 hrs into encystation [40].

Figure S4. *Trans*- and *cis*-introns in *G. muris*. A) Aligned *trans*- and *cis*-splice junctions in *G. muris.* **B)** Splice-site associated motifs in *trans*-spliced exons. Base-pairing stems are underlined. Bases in red denote the predicted cleavage motif. **C)** Consensus alignment of the cleavage motifs in *G. muris*. **D-E)** Predicted secondary structures of **D)** *cis*-spliced and **E)** *trans*-spliced genes. The secondary structure was predicted using the mfold webserver using default settings and manually curated. GU-rich stretches and cleavage motifs are boxed.

Figure S5. A) Conservation of the VSP C-terminal in *G. muris* and *G. intestinalis.* Sequence logos of conserved motifs in the C-terminal of VSPs in *G. muris* (n=25) (upper panel) and *G. intestinalis* WB (n=183) (lower panel) were generated using MEME v5.0.5. Transmembrane and cytosolic pentapeptides were predicted by the Phobius webserver. **B) VSP and ψVSP evolutionary dynamics.** The ψVSP arrays (defined as >3 ψVSP genes arrayed) in the *G. muris* genome arranged by chromosome, termini arranged to the right. Arrays of ψVSPs without chromosomal context are shown below as orphan arrays. **C)** VSP cluster mostly together in a phylogeny whereas most ψVSP gene arrays show relaxed clustering. Intact VSP genes are colored red, while ψVSP genes are colored according to their resident ψVSP array in panel A. Genes with black label are not found in arrays. **D) Expression measurements of VSP and ψVSP in *G. muris.*** VSP and ψVSP RNA-seq read counts (FPKM) from trophozoites, cysts and excyzoites. Size of the circle corresponds to expression level.

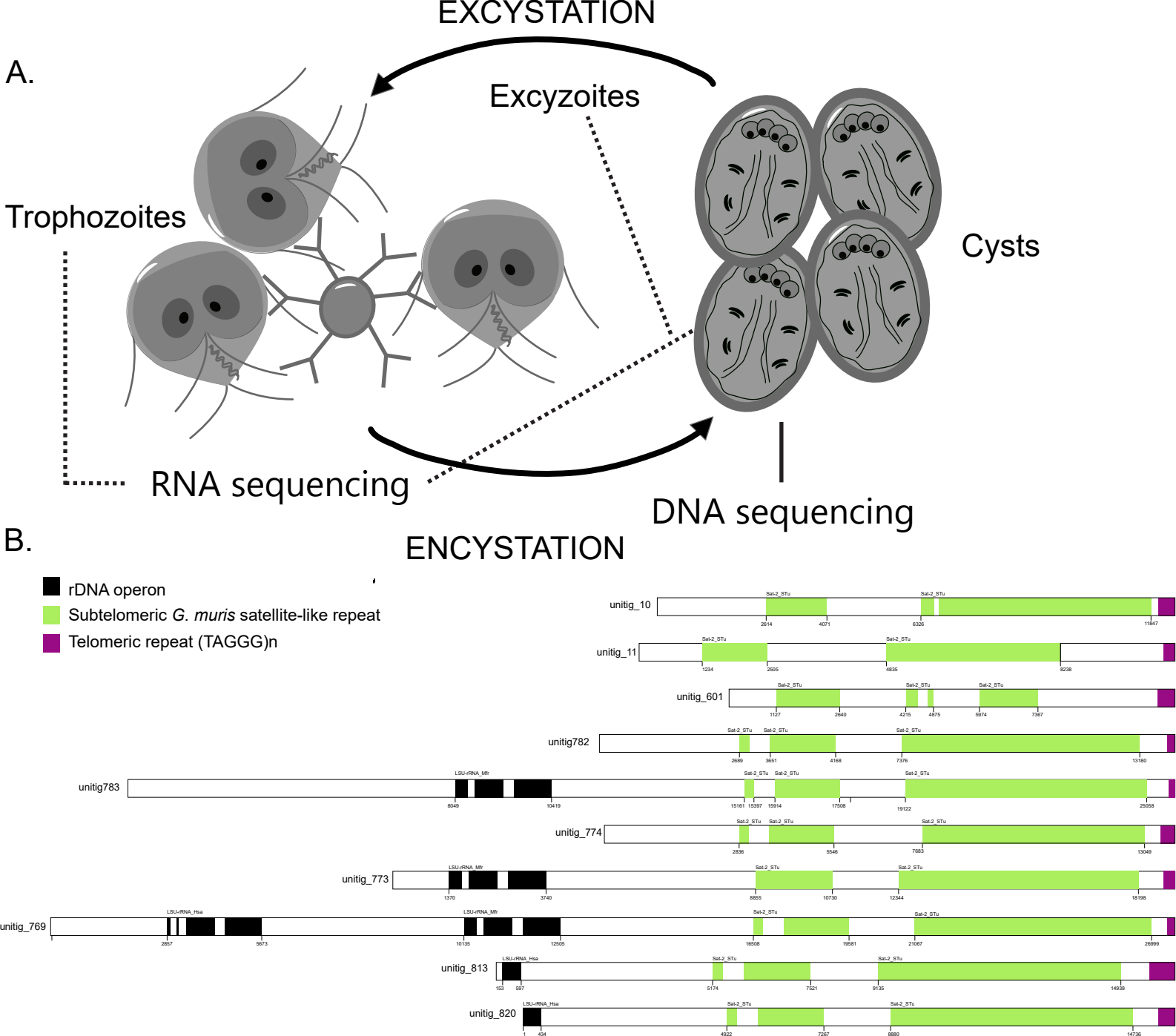
Figure S6. A) Phylogenetic tree of *G. muris* and *G. intestinalis* NEKs. B) Phylogenetic tree of *G. muris* and *G. intestinalis* ARPs. C) Network analysis of *G. muris* and *G. intestinalis* NEKs and ARPs. *G. muris* genes are represented in circle whereas *G. intestinalis* genes are shown in triangle. Blue indicates NEK and pink as ARP. Edges are weighted and scaled by reciprocal BLAST scores with evalue 1e-05 as the cutoff.

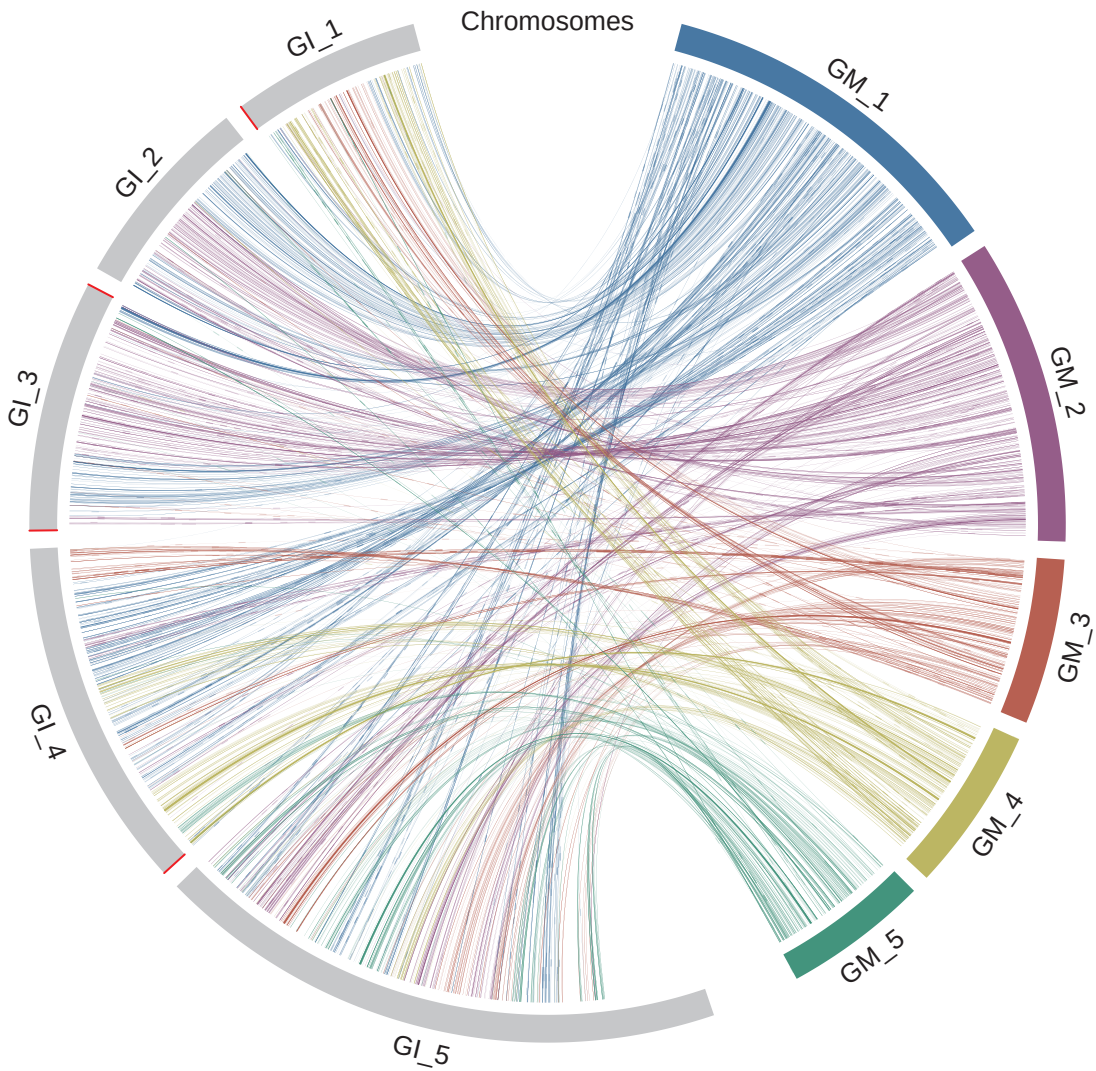
Figure S7. Phylogenetic tree of C1 family cysteine proteases in diplomonads. Black: *G. muris*, Red: *S. salmonicida*, Cyan: *G. intestinalis* GS/M, Green: *G. intestinalis* P15, Purple: *G. intestinalis* WB.

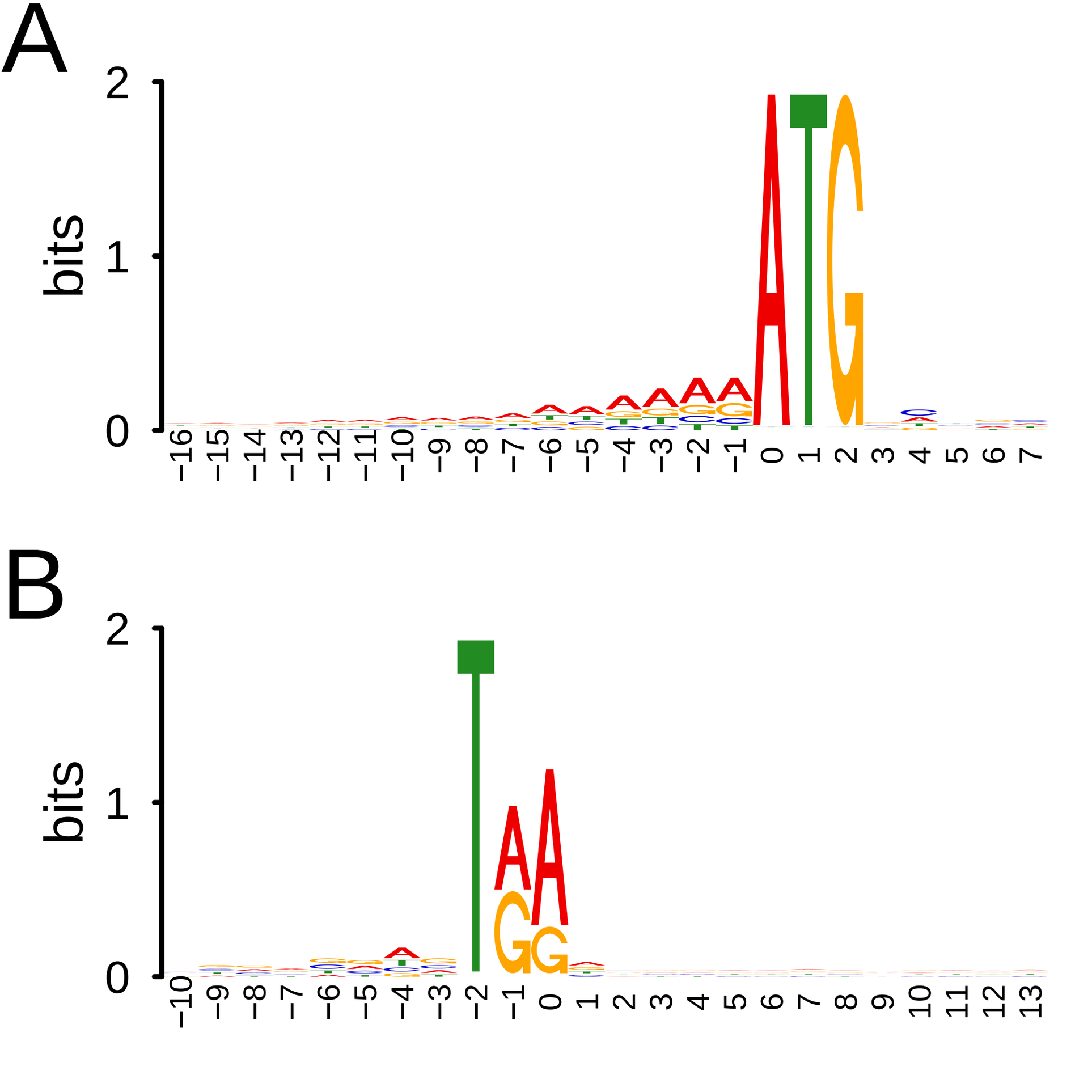
Figure S8. Phylogenetic analyses of HGT genes. A) Fructokinase. **B)** Mannose-6-phosphate isomerase. **C)** Transcriptional regulator/sugar kinase. **D)** Arginase. **E)** 2,5-diketo-D-gluconic acid reductase 1. **F)** 2,5-diketo-D-gluconic acid reductase 2. **G)** Carboxymuconolactone decarboxylase. **H)** Ketosteroid isomerase. **I)** Tryptophanase. **J)** Tae4. **K)** Quorum-quenching N-acyl-homoserine lactonase.

Table S1. A) One-to-one ortholog list between *G. muris* and *G. intestinalis* WB. Orthologs inferred by OrthoMCL. **B)** RNA-Seq expression values in excel file. Sorted according to FPKM values. **C)** VSP and ψVSP gene metadata. '-' in the pentapeptide motif represents lack of the 3' end sequence. **D)** List of ventral disc proteins in *G. muris*. **E)** Proteases detected in *G. muris* and *G. intestinalis*. **F)** Detailed information about hits of unique genes in *G. muris.*

Methods S1. Additional methods section


A.
EXCYSTATION
Excyzoites
Trophozoites
Cysts
RNA sequencing
DNA sequencing
B.
ENCYSTATION
rDNA operon
Subtelomeric G. muris satellite-like repeat
Telomeric repeat (TAGGG)n
unitig_10
unitig_11
unitig_601
unitig782
unitig783
unitig_774
unitig_773
unitig_769
unitig_813
unitig_820
Sat-2_STu
LSU-rRNA_Mfr
LSU-rRNA_Hsa


Chromosomes
GI_1
GI_2
GI_3
GI_4
GI_5
GM_1
GM_2
GM_3
GM_4
GM_5


A
bits
2
1
0
−16 −15 −14 −13 −12 −11 −10 −9 −8 −7 −6 −5 −4 −3 −2 −1 0 1 2 3 4 5 6 7
B
bits
2
1
0
−10 −9 −8 −7 −6 −5 −4 −3 −2 −1 0 1 2 3 4 5 6 7 8 9 10 11 12 13

C

LOG2 Trophs/Cysts	Encyst up 90 min	Encyst up 7h	Encyst up 22h	Gene id
9				GMRT_13982
9		●	●	GMRT_13204
9		●	●	GMRT_10774
9			●	GMRT_12407
8				GMRT_13942
8			●	GMRT_13341
8	●	●	●	GMRT_13400
8		●	●	GMRT_16219
8			●	GMRT_13851
8			●	GMRT_10589
8			●	GMRT_15202
8			●	GMRT_14302
7				GMRT_14494
7		●	●	GMRT_11770
7			●	GMRT_12514
7			●	GMRT_13776
7				GMRT_12207
7		●	●	GMRT_10157
7			●	GMRT_12519
7	●			GMRT_10224
7				GMRT_15543
7				GMRT_15329
7		●	●	GMRT_15984
7				GMRT_10478
7				GMRT_13283
7			●	GMRT_13422
7			●	GMRT_11433
6				GMRT_13315
6			●	GMRT_12667
6				GMRT_12476
6				GMRT_15303
6				GMRT_13901
6	●	●		GMRT_12279
6				GMRT_12528
6				GMRT_15601
6				GMRT_15514
6				GMRT_13605
6	●			GMRT_14361
6				GMRT_14093
6				GMRT_10233

LOG2 Trophs/Cysts	Encyst up 90 min	Encyst up 7h	Encyst up 22h	Gene id
6		●	●	GMRT_10574
6				GMRT_14982
6			●	GMRT_15986
6				GMRT_12453
6			●	GMRT_12189
6			●	GMRT_10312
6				GMRT_12190
6				GMRT_14841
6			●	GMRT_12243
6				GMRT_16336
6			●	GMRT_10833
6		●		GMRT_14208
6				GMRT_11693
6		●	●	GMRT_11117
6			●	GMRT_14265
6			●	GMRT_14586
6				GMRT_10367
6			●	GMRT_11957
6				GMRT_12759
6			●	GMRT_15885
6				GMRT_11905
6			●	GMRT_15547
6	●			GMRT_10756
6			●	GMRT_13994
5				GMRT_10377
5	●		●	GMRT_10123
5			●	GMRT_13780
5		●	●	GMRT_15811
5			●	GMRT_14189
5				GMRT_14883
5			●	GMRT_14997
5				GMRT_12394
5				GMRT_10848
5				GMRT_12419
5				GMRT_10135
5			●	GMRT_15872
5			●	GMRT_16332
5				GMRT_14187
5				GMRT_11145
5				GMRT_12776

LOG2 Trophs/Cysts	Encyst up 90 min	Encyst up 7h	Encyst up 22h	Gene id
5				GMRT_11843
5				GMRT_10541
5				GMRT_15517
5				GMRT_10471
5			●	GMRT_15257
5				GMRT_10501
5				GMRT_12607
5			●	GMRT_14295
5				GMRT_11315
5				GMRT_12559
5				GMRT_13415
5				GMRT_16213
5	●			GMRT_13109
5				GMRT_10567
5			●	GMRT_14575
5				GMRT_12715
5			●	GMRT_15985
5				GMRT_13841
5			●	GMRT_10038
5			●	GMRT_11441
5				GMRT_13609
5			●	GMRT_11004
5		●	●	GMRT_14079
5				GMRT_11698
5		●		GMRT_12381
5			●	GMRT_10030
5	●			GMRT_10585
5			●	GMRT_15240
5				GMRT_14576
5			●	GMRT_10991
5			●	GMRT_10607
5			●	GMRT_11520
5				GMRT_10127
5				GMRT_15829
5		●	●	GMRT_14224
5	●	●	●	GMRT_13401
5				GMRT_12332
5				GMRT_13227
5				GMRT_14385
5			●	GMRT_10590

LOG2 Trophs/Cysts	Encyst up 90 min	Encyst up 7h	Encyst up 22h	Gene id
5			●	GMRT_10989
5				GMRT_12531
5		●	●	GMRT_11758
5				GMRT_15162
5				GMRT_10361
5				GMRT_13783
5	●			GMRT_13387
5			●	GMRT_14264
5			●	GMRT_12874
5			●	GMRT_11847
4				GMRT_13488
4				GMRT_13010
4			●	GMRT_16375
4				GMRT_14214
4				GMRT_12291
4				GMRT_12942
4			●	GMRT_13845
4			●	GMRT_13363
4				GMRT_10998
4				GMRT_13929
4				GMRT_16145
4				GMRT_16374
4			●	GMRT_12385
4			●	GMRT_14301
4				GMRT_14584
4			●	GMRT_11013
4			●	GMRT_10770
4			●	GMRT_13180
4	●			GMRT_15317
4				GMRT_10811
4				GMRT_12821
4				GMRT_12091
4				GMRT_11966
4			●	GMRT_12130
4				GMRT_13902
4				GMRT_16379
4				GMRT_14952
4				GMRT_13245
4	●			GMRT_15524
4				GMRT_15258

LOG2 Trophs/Cysts	Encyst up 90 min	Encyst up 7h	Encyst up 22h	Gene id
4				GMRT_12300
4				GMRT_16311
4				GMRT_15163
4			●	GMRT_12285
4				GMRT_14430
4			●	GMRT_13601
4	●	●		GMRT_12672
4			●	GMRT_15899
4			●	GMRT_15546
4	●			GMRT_13611
4			●	GMRT_14303
4				GMRT_10495
4				GMRT_15159
4		●		GMRT_10188
4				GMRT_11987
4				GMRT_14267
4	●			GMRT_13101
4				GMRT_16305
4				GMRT_12824
4				GMRT_10401
4	●		●	GMRT_14423
4				GMRT_10296
4	●	●	●	GMRT_11676
4				GMRT_13096
4			●	GMRT_13390
4			●	GMRT_15222
4				GMRT_13985
4				GMRT_12620
4	●			GMRT_10467
4		●	●	GMRT_15068
4				GMRT_12343
4		●	●	GMRT_13794
4			●	GMRT_12815
4		●		GMRT_13570
4				GMRT_11953
4				GMRT_14849
4				GMRT_12799
4				GMRT_13282
4				GMRT_10492

A.

Trans-spliced introns

```
GMRT_12388 GTgtatgtgtg.../ /...ggactaacacgcagCT
GMRT_11576 AGgtatgtccg.../ /...tgactaacacgcagCA
GMRT_10332 GGgtatgtggt.../ /...ttactaacacgcagCT
GMRT_11888 TGatatgttgt.../ /...tcactaacacgcagAC
GMRT_22684 TGgtatggccc.../ /...ttactaacacgcagTC
             .****..               ***********
```

Cis-spliced introns

```
GMRT_fx013 GCgtatgtatc    tc    catctaacacgtagAA
GMRT_15227 GGgtatgtgcc (124nt) tcactaacgcgcagAT
GMRT_11066 CCgtatgttca  tatcg  tctctaacacgcagCT
             ******.             .*****.**.**
```

B.

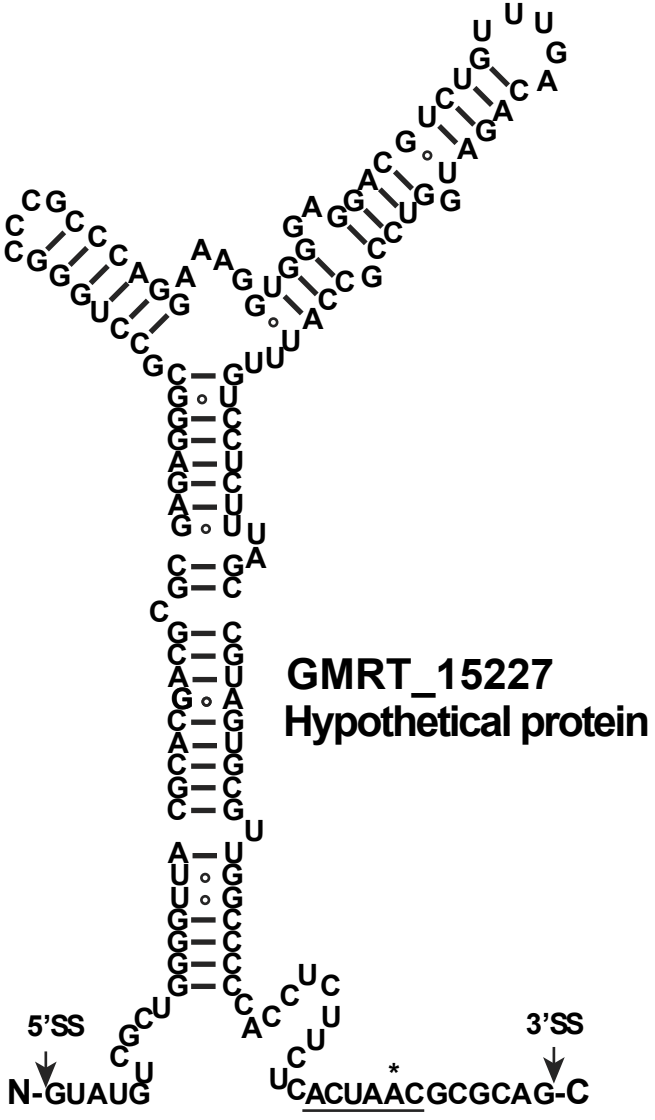
```
GMRT_12388-1     gtatgtgtgtgcgtcagccgctgtgtgttgtgtgtttccttttactcaatgacggtctc

GMRT_12388-1d    gtatgtgtgtgcgtcagccgctgtgtgttgtgtgtttccttttacttaaatggttc

GMRT_12388-2     aaaaatcagcggctgacgcgcaggaccatctggactaacacgcag


GMRT_11576-1     gtatgtccgctgtgtgcgagcatgtgtgttgtgtgtctcctttactcaacttccgcgg

GMRT_11576-1d    gtatgtccgctgtgtgcgagcatgtgtgttgtgtgtctcctttactcaatatcaaattctccattgtatttcat

GMRT_11576-2     aaaaaacctcaagtactggagggaagaaaagctcgcaccagcagaatcttgactaacacgcag

GMRT_11576-2-gt  gtatgtggtgttccgcttgctgtgtgttgtgtgttccttcacttaagttctgtc

GMRT_11576-3     aaaaagagacacgaggacgatcggcaagcggaacacacagtcctttactaacacgcag


GMRT_11888-1     atatgttgtgcgtttcaccccgttgtgtgttgtgtgttccttttctcaagcattggct

GMRT_11888-2     aaaacggggtgaacgtgcagtcctcactaacacgcag


GMRT_22684-1     gtatggcccctcctccttcactcaagccttct

GMRT_22684-2     aaaggagggaccaatcttactaacacgcag
```

C.

Cleavage motif

```
tcctttactcaa
tcctttacttaa
tcctttactcaa
tccttcacttaa
tcctttctcaa
tccttcactcaa
*****..**.**
```

D.



GMRT_15227
Hypothetical protein

5'SS 3'SS

N-GUAUG ACUAACGCGCAG-C

GMRT_fx013
Hypothetical protein

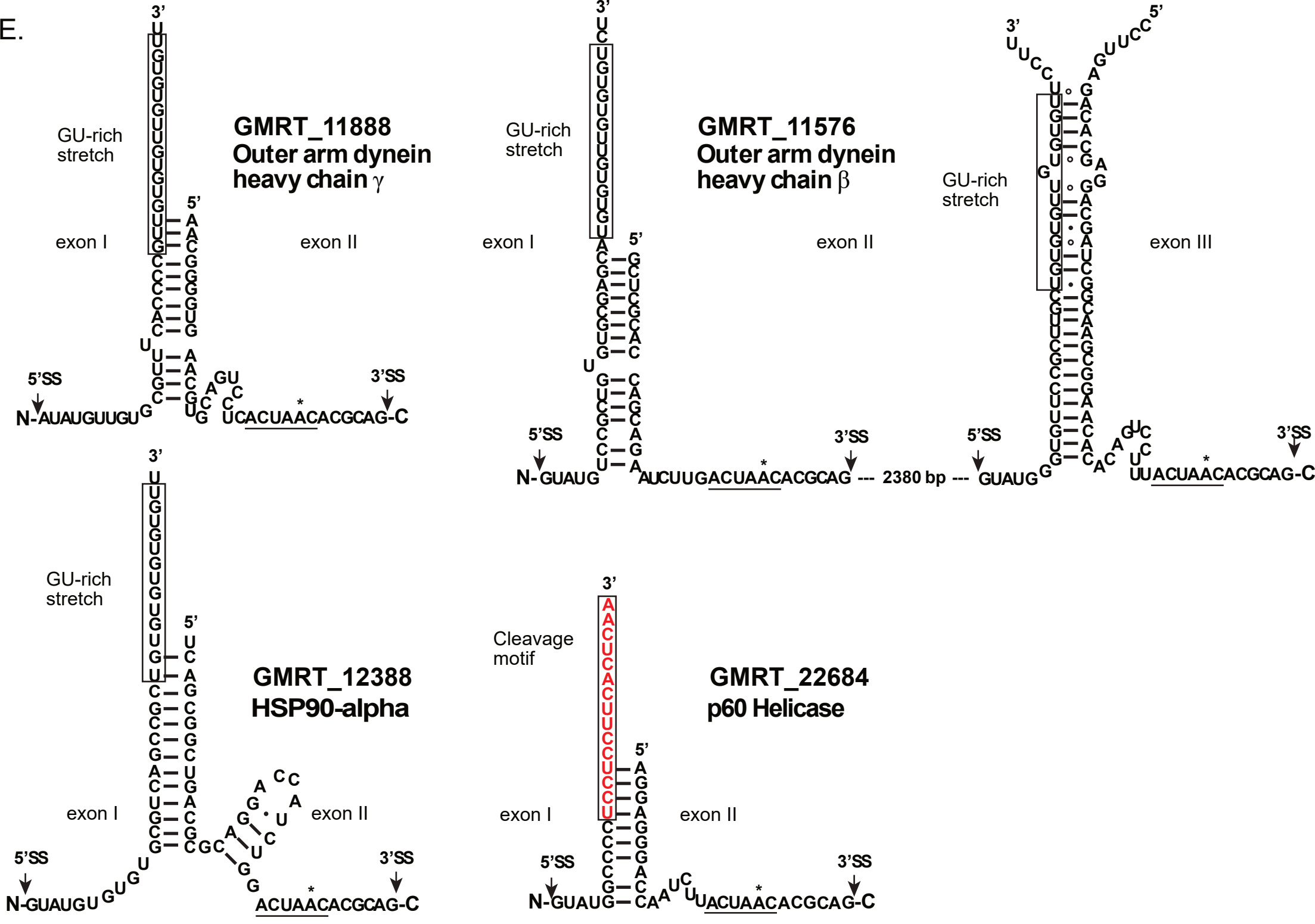
5'SS 3'SS

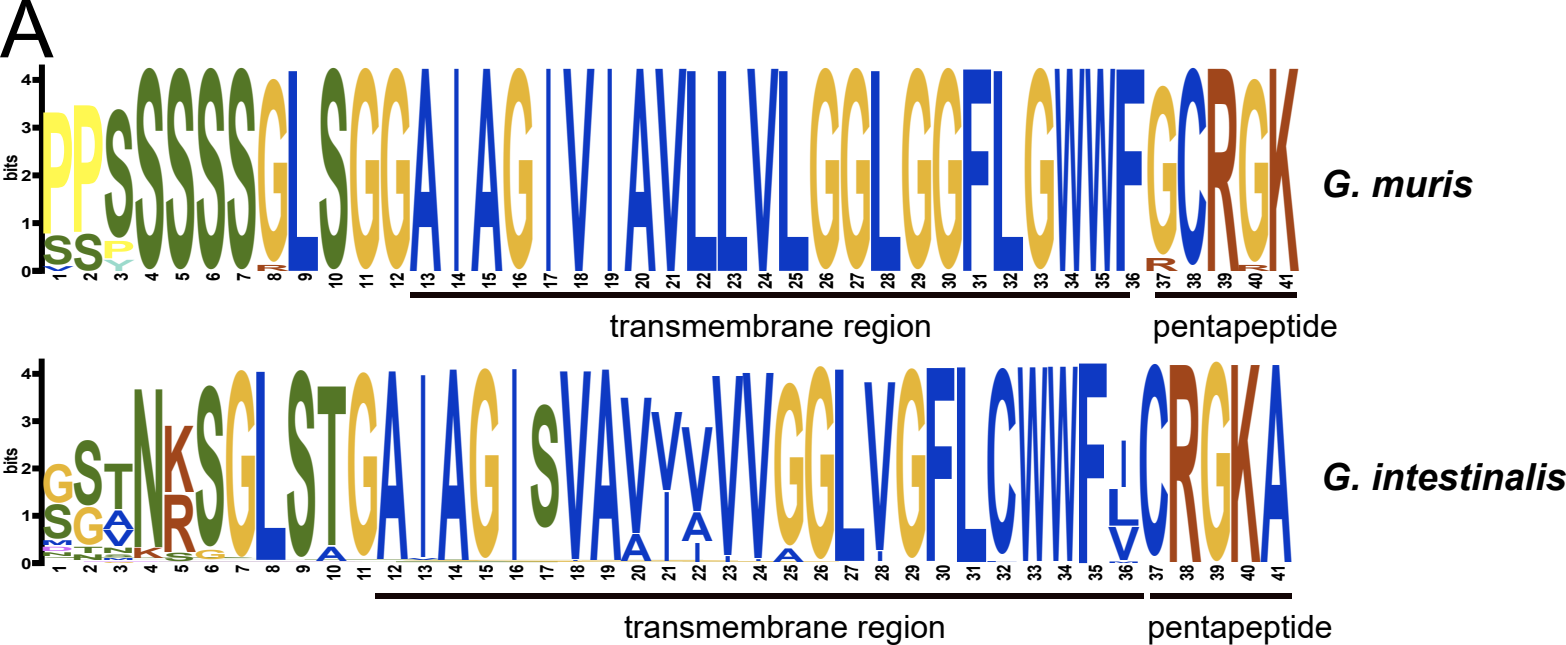
N-GUAUGUAUCUCCAUCUAACACGUAG-C

GMRT_11066
Hypothetical protein

5'SS 3'SS

N-GUAUGUUCAUAUCGUCUCUAACACGCAG-C


E.
GMRT_11888
Outer arm dynein
heavy chain γ
GU-rich
stretch
exon I
exon II
5'SS
3'SS
N-AUAUGUUGU
UCACUAACACGCAG-C
GMRT_11576
Outer arm dynein
heavy chain β
GU-rich
stretch
exon I
exon II
exon III
5'SS
3'SS
N-GUAUG
AUCUUGACUAACACGCAG
2380 bp
GUAUG
UUACUAACACGCAG-C
GMRT_12388
HSP90-alpha
GU-rich
stretch
exon I
exon II
5'SS
3'SS
N-GUAUGU
ACUAACACGCAG-C
GMRT_22684
p60 Helicase
Cleavage
motif
exon I
exon II
5'SS
3'SS
N-GUAUG
UACUAACACGCAG-C


A
bits
G. muris
transmembrane region
pentapeptide
bits
G. intestinalis
transmembrane region
pentapeptide

B

Chromosome 1 array 2

Chromosome 2 array 1

Chromosome 2 array 2

Chromosome 3 array 1

Chromosome 3 array 2

Chromosome 4 array 1

Chromosome 4 array 2

Chromosome 5 array 1

Chromosome 5 array 2

Chromosome 5 array 3 (internal)

Orphan array 1

Orphan array 2

Orphan array 3

Orphan array 4

Orphan array 5

Orphan array 6

Orphan array 7

C

VSP cluster 2

Internal pVSP array

VSP cluster 3

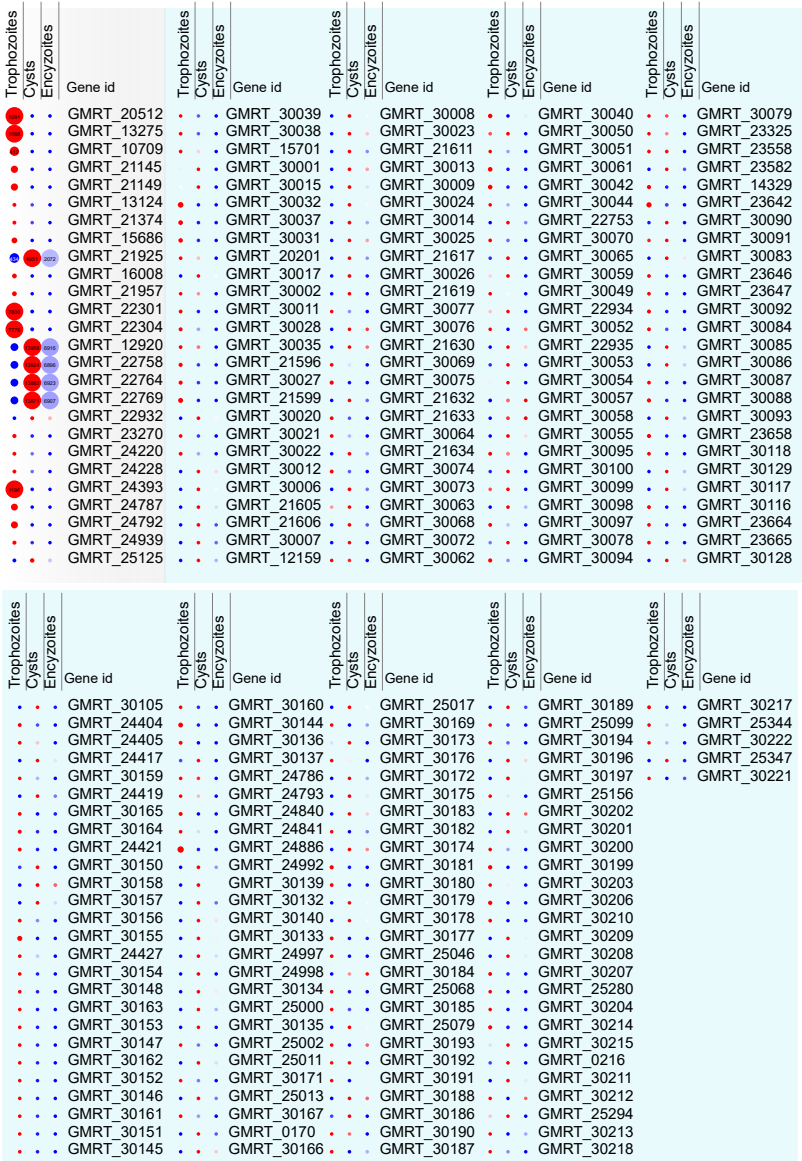
VSP cluster 1

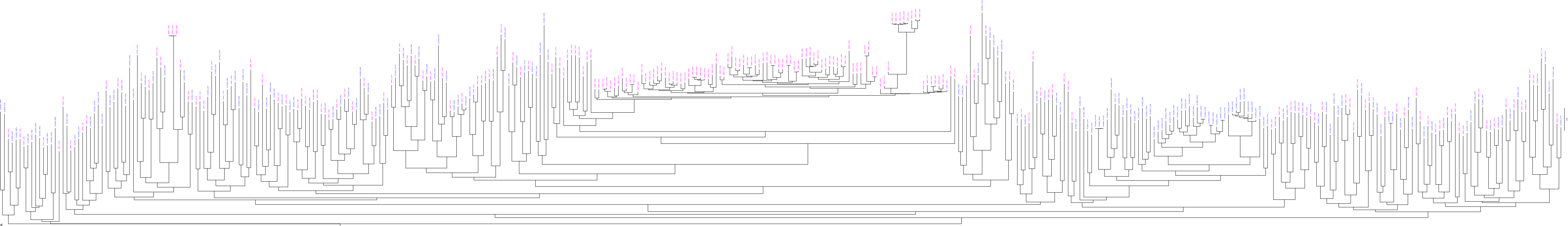
0.3

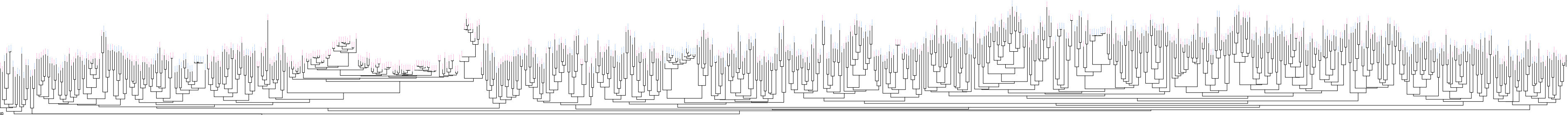
VSP directly adjacent to a pVSP

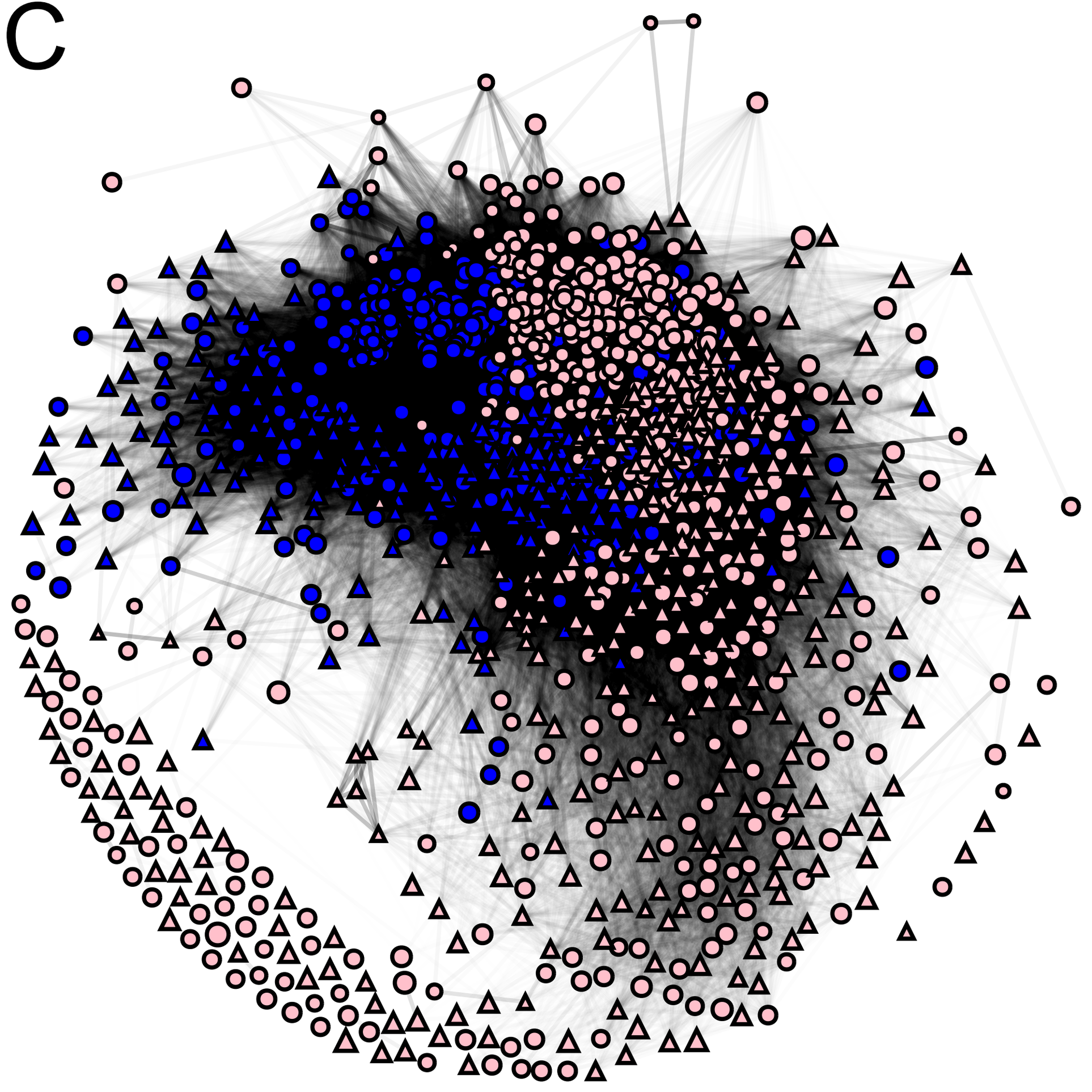
Part of a twinned pVSP array

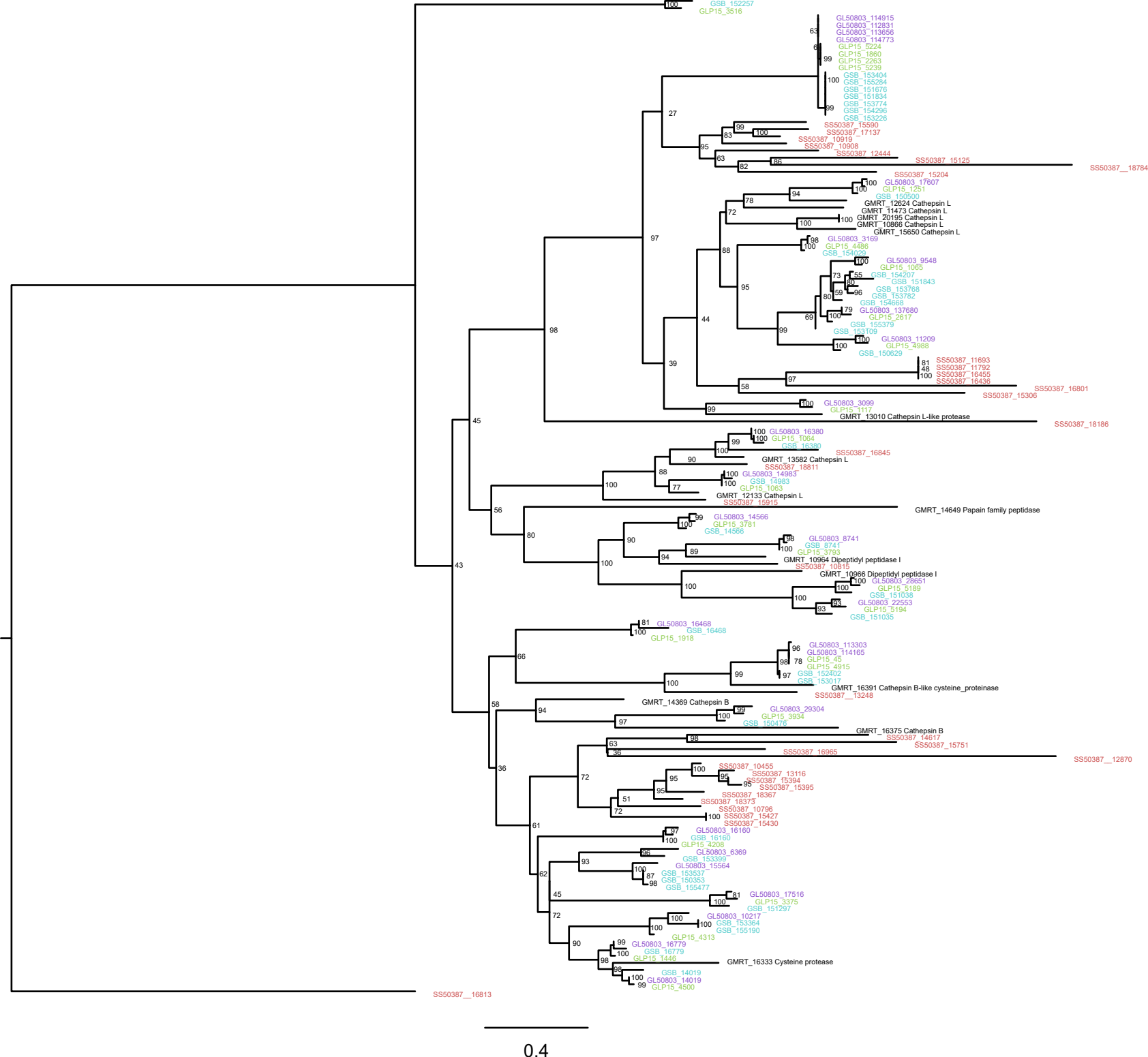
D


VSPs
pVSPs
Trophozoites
Cysts
Encyzoites
Gene id
GMRT_20512
GMRT_13275
GMRT_10709
GMRT_21145
GMRT_21149
GMRT_13124
GMRT_21374
GMRT_15686
GMRT_21925
GMRT_16008
GMRT_21957
GMRT_22301
GMRT_22304
GMRT_12920
GMRT_22758
GMRT_22764
GMRT_22769
GMRT_22932
GMRT_23270
GMRT_24220
GMRT_24228
GMRT_24393
GMRT_24787
GMRT_24792
GMRT_24939
GMRT_25125
GMRT_30039
GMRT_30038
GMRT_15701
GMRT_30001
GMRT_30015
GMRT_30032
GMRT_30037
GMRT_30031
GMRT_20201
GMRT_30017
GMRT_30002
GMRT_30011
GMRT_30028
GMRT_30035
GMRT_21596
GMRT_30027
GMRT_21599
GMRT_30020
GMRT_30021
GMRT_30022
GMRT_30012
GMRT_30006
GMRT_21605
GMRT_21606
GMRT_30007
GMRT_12159
GMRT_30008
GMRT_30023
GMRT_21611
GMRT_30013
GMRT_30009
GMRT_30024
GMRT_30014
GMRT_30025
GMRT_21617
GMRT_30026
GMRT_21619
GMRT_30077
GMRT_30076
GMRT_21630
GMRT_30069
GMRT_30075
GMRT_21632
GMRT_21633
GMRT_30064
GMRT_21634
GMRT_30074
GMRT_30073
GMRT_30063
GMRT_30068
GMRT_30072
GMRT_30062
GMRT_30040
GMRT_30050
GMRT_30051
GMRT_30061
GMRT_30042
GMRT_30044
GMRT_22753
GMRT_30070
GMRT_30065
GMRT_30059
GMRT_30049
GMRT_22934
GMRT_30052
GMRT_22935
GMRT_30053
GMRT_30054
GMRT_30057
GMRT_30058
GMRT_30055
GMRT_30095
GMRT_30100
GMRT_30099
GMRT_30098
GMRT_30097
GMRT_30078
GMRT_30094
GMRT_30079
GMRT_23325
GMRT_23558
GMRT_23582
GMRT_14329
GMRT_23642
GMRT_30090
GMRT_30091
GMRT_30083
GMRT_23646
GMRT_23647
GMRT_30092
GMRT_30084
GMRT_30085
GMRT_30086
GMRT_30087
GMRT_30088
GMRT_30093
GMRT_23658
GMRT_30118
GMRT_30129
GMRT_30117
GMRT_30116
GMRT_23664
GMRT_23665
GMRT_30128
GMRT_30115
GMRT_30127
GMRT_30114
GMRT_30113
GMRT_30112
GMRT_30111
GMRT_30125
GMRT_30124
GMRT_23922
GMRT_23982
GMRT_30106
GMRT_24386
GMRT_30126
GMRT_30110
GMRT_30121
GMRT_24390
GMRT_30109
GMRT_30120
GMRT_24395
GMRT_30104
GMRT_24397
GMRT_24398
GMRT_30108
GMRT_24400
GMRT_24401
GMRT_24402
GMRT_30105
GMRT_24404
GMRT_24405
GMRT_24417
GMRT_30159
GMRT_24419
GMRT_30165
GMRT_30164
GMRT_24421
GMRT_30150
GMRT_30158
GMRT_30157
GMRT_30156
GMRT_30155
GMRT_24427
GMRT_30154
GMRT_30148
GMRT_30163
GMRT_30153
GMRT_30147
GMRT_30162
GMRT_30152
GMRT_30146
GMRT_30161
GMRT_30151
GMRT_30145
GMRT_30160
GMRT_30144
GMRT_30136
GMRT_30137
GMRT_24786
GMRT_24793
GMRT_24840
GMRT_24841
GMRT_24886
GMRT_24992
GMRT_30139
GMRT_30132
GMRT_30140
GMRT_30133
GMRT_24997
GMRT_24998
GMRT_30134
GMRT_25000
GMRT_30135
GMRT_25002
GMRT_25011
GMRT_30171
GMRT_25013
GMRT_30167
GMRT_0170
GMRT_30166
GMRT_25017
GMRT_30169
GMRT_30173
GMRT_30176
GMRT_30172
GMRT_30175
GMRT_30183
GMRT_30182
GMRT_30174
GMRT_30181
GMRT_30180
GMRT_30179
GMRT_30178
GMRT_30177
GMRT_25046
GMRT_30184
GMRT_25068
GMRT_30185
GMRT_25079
GMRT_30193
GMRT_30192
GMRT_30191
GMRT_30188
GMRT_30186
GMRT_30190
GMRT_30187
GMRT_30189
GMRT_25099
GMRT_30194
GMRT_30196
GMRT_30197
GMRT_25156
GMRT_30202
GMRT_30201
GMRT_30200
GMRT_30199
GMRT_30203
GMRT_30206
GMRT_30210
GMRT_30209
GMRT_30208
GMRT_30207
GMRT_25280
GMRT_30204
GMRT_30214
GMRT_30215
GMRT_0216
GMRT_30211
GMRT_30212
GMRT_25294
GMRT_30213
GMRT_30218
GMRT_30217
GMRT_25344
GMRT_30222
GMRT_25347
GMRT_30221
row max
row min



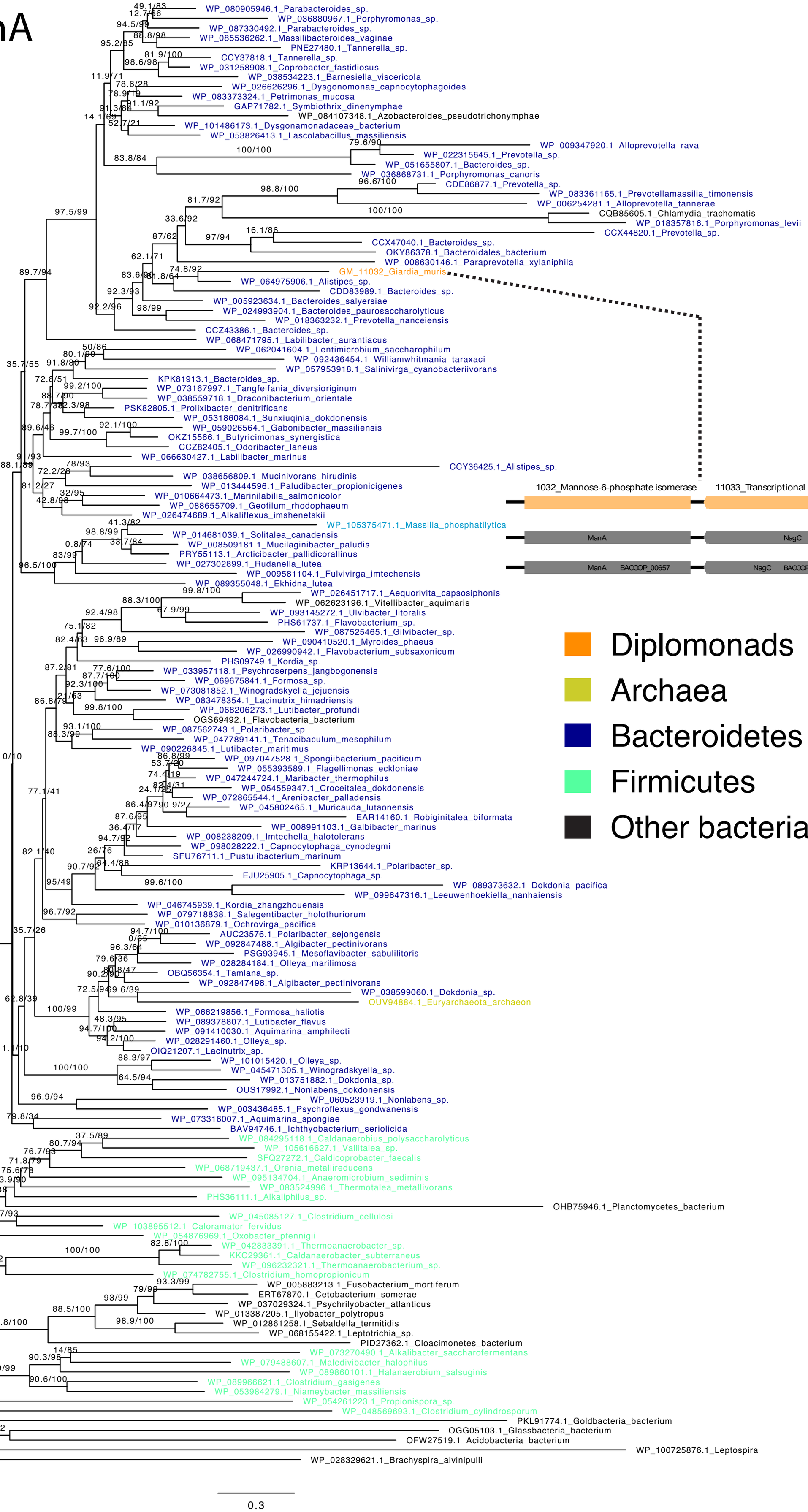



C

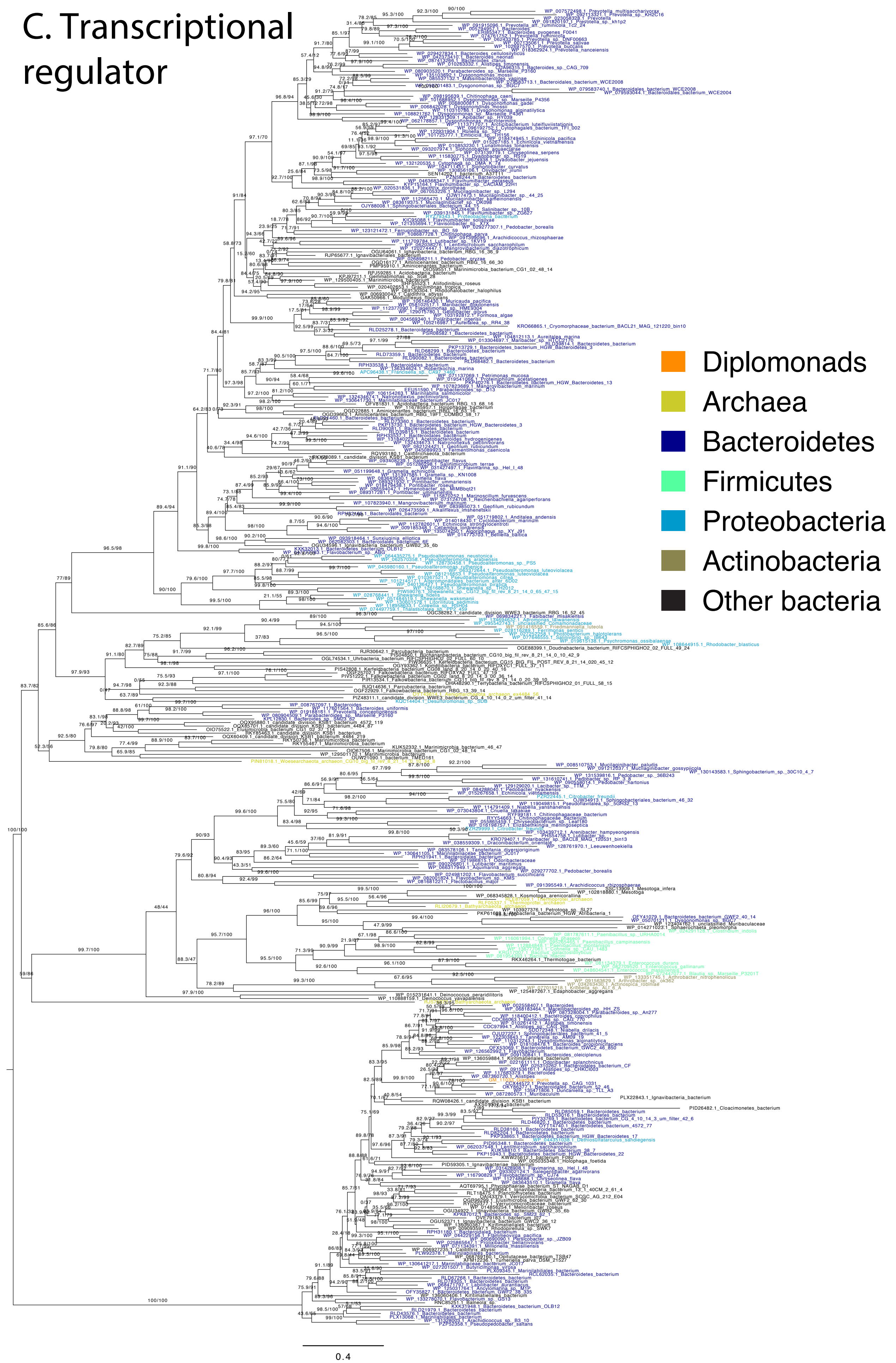

GSB_152257
GLP15_3516
GL50803_114915
GL50803_112831
GL50803_113656
GL50803_114773
GLP15_5224
GLP15_1860
GLP15_2263
GLP15_5239
GSB_153404
GSB_155284
GSB_151676
GSB_151834
GSB_153774
GSB_154296
GSB_153226
SS50387_15590
SS50387_17137
SS50387_10919
SS50387_10908
SS50387_12444
SS50387_15125
SS50387__18784
SS50387_15204
GL50803_17607
GLP15_1251
GSB_150500
GMRT_12624 Cathepsin L
GMRT_11473 Cathepsin L
GMRT_20195 Cathepsin L
GMRT_10866 Cathepsin L
GMRT_15650 Cathepsin L
GL50803_3169
GLP15_4486
GSB_154029
GL50803_9548
GLP15_1065
GSB_154207
GSB_151843
GSB_153768
GSB_153782
GSB_154668
GL50803_137680
GLP15_2617
GSB_155379
GSB_153109
GL50803_11209
GLP15_4988
GSB_150629
SS50387_11693
SS50387_11792
SS50387_16455
SS50387_16436
SS50387_16801
SS50387_15306
GL50803_3099
GLP15_1117
GMRT_13010 Cathepsin L-like protease
SS50387_18186
GL50803_16380
GLP15_1064
GSB_16380
SS50387_16845
GMRT_13582 Cathepsin L
SS50387_18811
GL50803_14983
GSB_14983
GLP15_1063
GMRT_12133 Cathepsin L
SS50387_15915
GMRT_14649 Papain family peptidase
GL50803_14566
GLP15_3781
GSB_14566
GL50803_8741
GSB_8741
GLP15_3793
GMRT_10964 Dipeptidyl peptidase I
SS50387_10815
GMRT_10966 Dipeptidyl peptidase I
GL50803_28651
GLP15_5189
GSB_151038
GL50803_22553
GLP15_5194
GSB_151035
GL50803_16468
GSB_16468
GLP15_1918
GL50803_113303
GL50803_114165
GLP15_45
GLP15_4915
GSB_152402
GSB_153017
GMRT_16391 Cathepsin B-like cysteine_proteinase
SS50387__13248
GMRT_14369 Cathepsin B
GL50803_29304
GLP15_3934
GSB_150476
GMRT_16375 Cathepsin B
SS50387_14617
SS50387_15751
SS50387_16965
SS50387__12870
SS50387_10455
SS50387_13116
SS50387_15394
SS50387_15395
SS50387_18367
SS50387_16373
SS50387_10796
SS50387_15427
SS50387_15430
GL50803_16160
GSB_16160
GLP15_4208
GL50803_6369
GSB_153398
GL50803_15564
GSB_153537
GSB_150353
GSB_155477
GL50803_17516
GLP15_3375
GSB_151297
GL50803_10217
GSB_153364
GSB_155190
GLP15_4313
GL50803_16779
GSB_16779
GLP15_1446
GMRT_16333 Cysteine protease
GSB_14019
GL50803_14019
GLP15_4500
SS50387__16813
0.4

A. Fructokinase

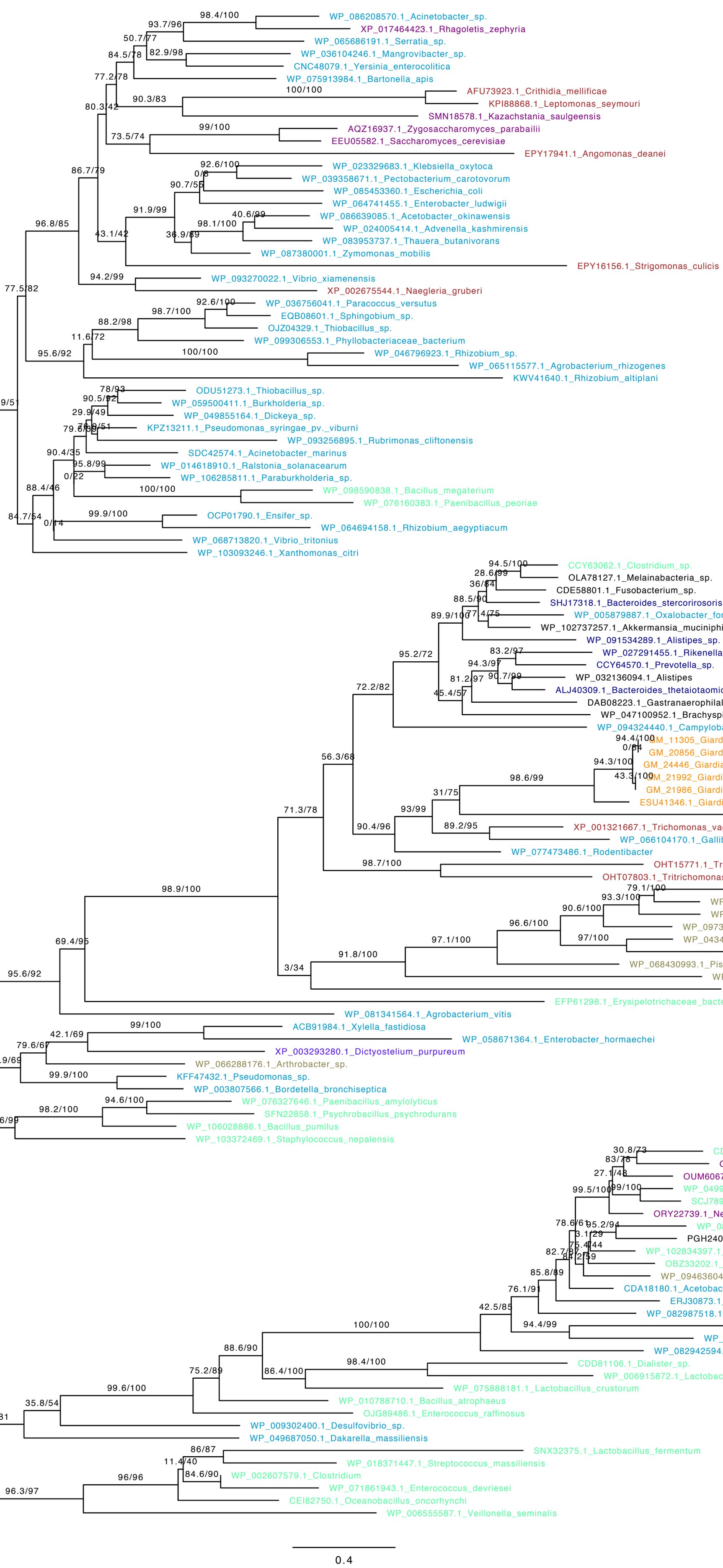
B. ManA


WP_080905946.1_Parabacteroides_sp.
WP_036880967.1_Porphyromonas_sp.
WP_087330492.1_Parabacteroides_sp.
WP_085536262.1_Massilibacteroides_vaginae
PNE27480.1_Tannerella_sp.
CCY37818.1_Tannerella_sp.
WP_031258908.1_Coprobacter_fastidiosus
WP_038534223.1_Barnesiella_viscericola
WP_026626296.1_Dysgonomonas_capnocytophagoides
WP_083373324.1_Petrimonas_mucosa
GAP71782.1_Symbiothrix_dinenymphae
WP_084107348.1_Azobacteroides_pseudotrichonymphae
WP_101486173.1_Dysgonamonadaceae_bacterium
WP_053826413.1_Lascolabacillus_massiliensis
WP_009347920.1_Alloprevotella_rava
WP_022315645.1_Prevotella_sp.
WP_051655807.1_Bacteroides_sp.
WP_036868731.1_Porphyromonas_canoris
CDE86877.1_Prevotella_sp.
WP_083361165.1_Prevotellamassilia_timonensis
WP_006254281.1_Alloprevotella_tannerae
CQB85605.1_Chlamydia_trachomatis
WP_018357816.1_Porphyromonas_levii
CCX44820.1_Prevotella_sp.
CCX47040.1_Bacteroides_sp.
OKY86378.1_Bacteroidales_bacterium
WP_008630146.1_Paraprevotella_xylaniphila
GM_11032_Giardia_muris
WP_064975906.1_Alistipes_sp.
CDD83989.1_Bacteroides_sp.
WP_005923634.1_Bacteroides_salyersiae
WP_024993904.1_Bacteroides_paurosaccharolyticus
WP_018363232.1_Prevotella_nanceiensis
CCZ43386.1_Bacteroides_sp.
WP_068471795.1_Labilibacter_aurantiacus
WP_062041604.1_Lentimicrobium_saccharophilum
WP_092436454.1_Williamwhitmania_taraxaci
WP_057953918.1_Salinivirga_cyanobacteriivorans
KPK81913.1_Bacteroides_sp.
WP_073167997.1_Tangfeifania_diversioriginum
WP_038559718.1_Draconibacterium_orientale
PSK82805.1_Prolixibacter_denitrificans
WP_053186084.1_Sunxiuqinia_dokdonensis
WP_059026564.1_Gabonibacter_massiliensis
OKZ15566.1_Butyricimonas_synergistica
CCZ82405.1_Odoribacter_laneus
WP_066630427.1_Labilibacter_marinus
CCY36425.1_Alistipes_sp.
WP_038656809.1_Mucinivorans_hirudinis
WP_013444596.1_Paludibacter_propionicigenes
WP_010664473.1_Marinilabilia_salmonicolor
WP_088655709.1_Geofilum_rhodophaeum
WP_026474689.1_Alkaliflexus_imshenetskii
WP_105375471.1_Massilia_phosphatilytica
WP_014681039.1_Solitalea_canadensis
WP_008509181.1_Mucilaginibacter_paludis
PRY55113.1_Arcticibacter_pallidicorallinus
WP_027302899.1_Rudanella_lutea
WP_009581104.1_Fulvivirga_imtechensis
WP_089355048.1_Ekhidna_lutea
WP_026451717.1_Aequorivita_capsosiphonis
WP_062623196.1_Vitellibacter_aquimaris
WP_093145272.1_Ulvibacter_litoralis
PHS61737.1_Flavobacterium_sp.
WP_087525465.1_Gilvibacter_sp.
WP_090410520.1_Myroides_phaeus
WP_026990942.1_Flavobacterium_subsaxonicum
PHS09749.1_Kordia_sp.
WP_033957118.1_Psychroserpens_jangbogonensis
WP_069675841.1_Formosa_sp.
WP_073081852.1_Winogradskyella_jejuensis
WP_083478354.1_Lacinutrix_himadriensis
WP_068206273.1_Lutibacter_profundi
OGS69492.1_Flavobacteria_bacterium
WP_087562743.1_Polaribacter_sp.
WP_047789141.1_Tenacibaculum_mesophilum
WP_090226845.1_Lutibacter_maritimus
WP_097047528.1_Spongiibacterium_pacificum
WP_055393589.1_Flagellimonas_eckloniae
WP_047244724.1_Maribacter_thermophilus
WP_054559347.1_Croceitalea_dokdonensis
WP_072865544.1_Arenibacter_palladensis
WP_045802465.1_Muricauda_lutaonensis
EAR14160.1_Robiginitalea_biformata
WP_008991103.1_Galbibacter_marinus
WP_008238209.1_Imtechella_halotolerans
WP_098028222.1_Capnocytophaga_cynodegmi
SFU76711.1_Pustulibacterium_marinum
KRP13644.1_Polaribacter_sp.
EJU25905.1_Capnocytophaga_sp.
WP_089373632.1_Dokdonia_pacifica
WP_099647316.1_Leeuwenhoekiella_nanhaiensis
WP_046745939.1_Kordia_zhangzhouensis
WP_079718838.1_Salegentibacter_holothuriorum
WP_010136879.1_Ochrovirga_pacifica
AUC23576.1_Polaribacter_sejongensis
WP_092847488.1_Algibacter_pectinivorans
PSG93945.1_Mesoflavibacter_sabulilitoris
WP_028284184.1_Olleya_marilimosa
OBQ56354.1_Tamlana_sp.
WP_092847498.1_Algibacter_pectinivorans
WP_038599060.1_Dokdonia_sp.
OUV94884.1_Euryarchaeota_archaeon
WP_066219856.1_Formosa_haliotis
WP_089378807.1_Lutibacter_flavus
WP_091410030.1_Aquimarina_amphilecti
WP_028291460.1_Olleya_sp.
OIQ21207.1_Lacinutrix_sp.
WP_101015420.1_Olleya_sp.
WP_045471305.1_Winogradskyella_sp.
WP_013751882.1_Dokdonia_sp.
OUS17992.1_Nonlabens_dokdonensis
WP_060523919.1_Nonlabens_sp.
WP_003436485.1_Psychroflexus_gondwanensis
WP_073316007.1_Aquimarina_spongiae
BAV94746.1_Ichthyobacterium_seriolicida
WP_084295118.1_Caldanaerobius_polysaccharolyticus
WP_105616627.1_Vallitalea_sp.
SFQ27272.1_Caldicoprobacter_faecalis
WP_068719437.1_Orenia_metallireducens
WP_095134704.1_Anaeromicrobium_sediminis
WP_083524996.1_Thermotalea_metallivorans
PHS36111.1_Alkaliphilus_sp.
OHB75946.1_Planctomycetes_bacterium
WP_045085127.1_Clostridium_cellulosi
WP_103895512.1_Caloramator_fervidus
WP_054876969.1_Oxobacter_pfennigii
WP_042833391.1_Thermoanaerobacter_sp.
KKC29361.1_Caldanaerobacter_subterraneus
WP_096232321.1_Thermoanaerobacterium_sp.
WP_074782755.1_Clostridium_homopropionicum
WP_005883213.1_Fusobacterium_mortiferum
ERT67870.1_Cetobacterium_somerae
WP_037029324.1_Psychrilyobacter_atlanticus
WP_013387205.1_Ilyobacter_polytropus
WP_012861258.1_Sebaldella_termitidis
WP_068155422.1_Leptotrichia_sp.
PID27362.1_Cloacimonetes_bacterium
WP_073270490.1_Alkalibacter_saccharofermentans
WP_079488607.1_Maledivibacter_halophilus
WP_089860101.1_Halanaerobium_salsuginis
WP_089966621.1_Clostridium_gasigenes
WP_053984279.1_Niameybacter_massiliensis
WP_054261223.1_Propionispora_sp.
WP_048569693.1_Clostridium_cylindrosporum
PKL91774.1_Goldbacteria_bacterium
OGG05103.1_Glassbacteria_bacterium
OFW27519.1_Acidobacteria_bacterium
WP_100725876.1_Leptospira
WP_028329621.1_Brachyspira_alvinipulli
1032_Mannose-6-phosphate isomerase
11033_Transcriptional regulator/sugar kinase
G.mu
ManA
NagC
Alistipes Marseille-P2431
ManA BACCOP_00657
NagC BACCOP_00658
Bacteroides coprocola
Diplomonads
Archaea
Bacteroidetes
Firmicutes
Other bacteria
0.3

C. Transcriptional regulator

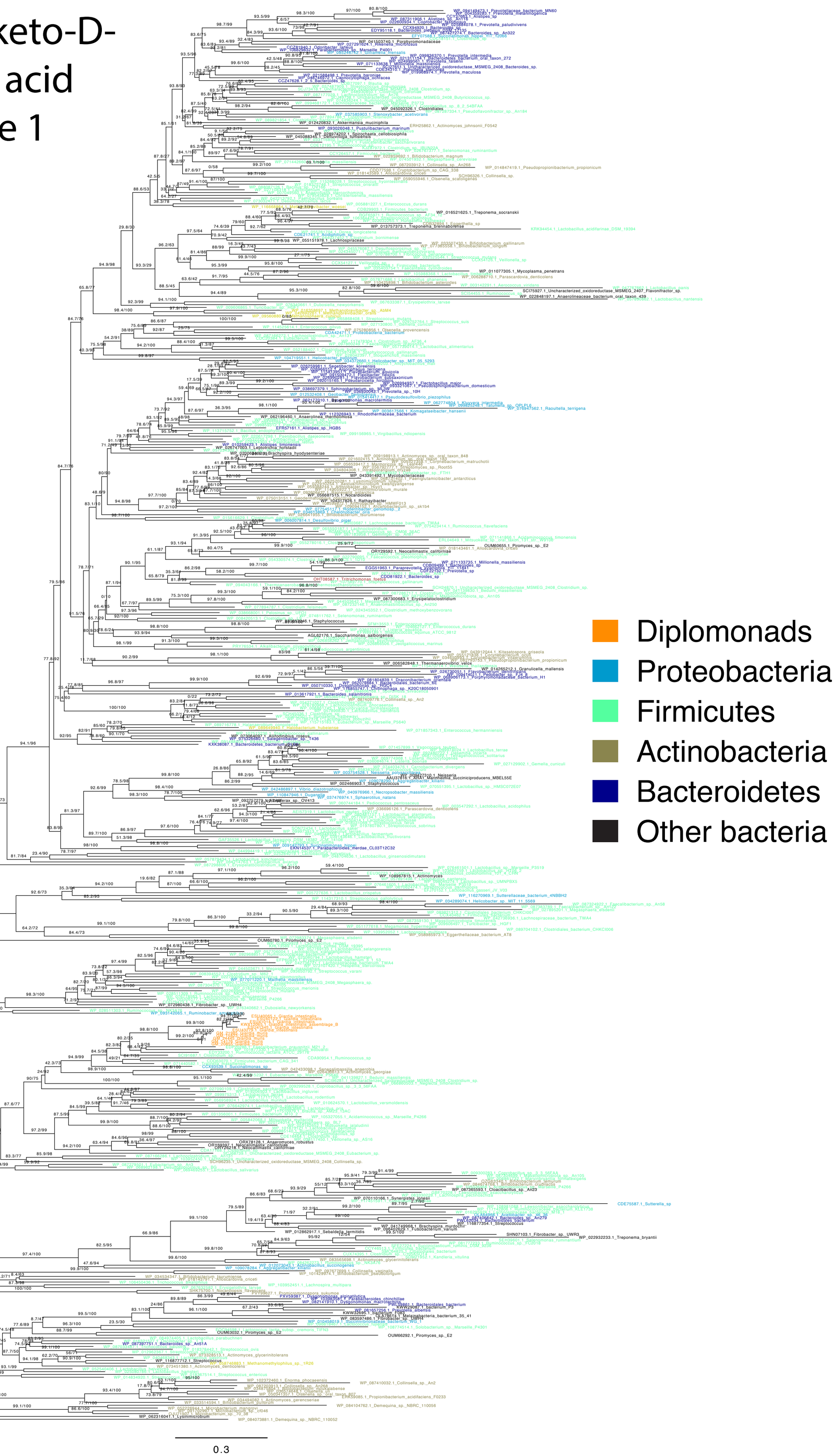


D. Arginase



- Diplomonads
- Excavata
- Amoebozoa
- Opisthokonta
- Proteobacteria
- Firmicutes
- Actinobacteria
- Bacteroidetes
- Other bacteria

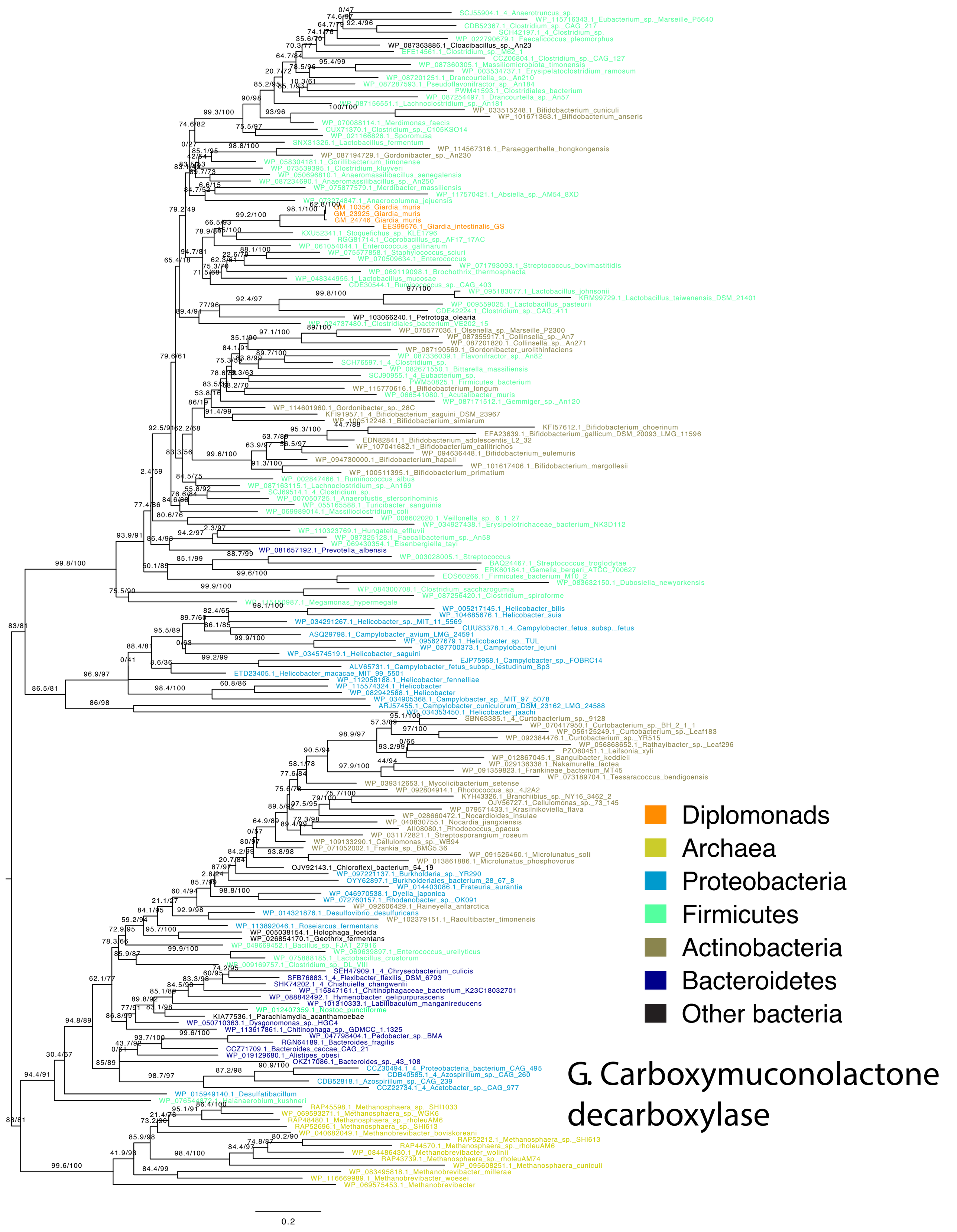
E. 2,5-diketo-D-gluconic acid reductase 1



F. 2,5-diketo-D-gluconic acid reductase 2

- Diplomonads
- Excavata
- SAR
- Archaea
- Proteobacteria
- Firmicutes
- Actinobacteria
- Bacteroidetes
- Other bacteria

0.3


Diplomonads
Archaea
Proteobacteria
Firmicutes
Actinobacteria
Bacteroidetes
Other bacteria
G. Carboxymuconolactone decarboxylase
0.2

Diplomonads

Proteobacteria

Actinobacteria

Bacteroidetes

Cyanobacteria

Other bacteria

H. Ketosteroid isomerase

0.4

I. Tryptophanase

- Diplomonads
- Amoebozoa
- Opisthokonta
- SAR
- Archaea
- Proteobacteria
- Firmicutes
- Actinobacteria
- Bacteroidetes
- Other bacteria

J. Tae4

- Diplomonads
- Opisthokonta
- Proteobacteria
- Bacteroidetes
- Other bacteria

K. Quorum-quenching N-acyl-homoserine lactonase

Methods S1 – Compact genome of *Giardia muris* reveals important steps in the evolution of intestinal protozoan parasites" by Xu et al. (2020)

DNA preparation from *G. muris* for long read sequencing

G. muris excyzoites (~9×10^7 cells) were centrifuged at 500 ×g for five minutes to remove the excystation media and the pellet was dissolved in Lysis Buffer (50 mM EDTA, 1% SDS and 10 mg/ml Proteinase K) by vortexing. The lysed cells were incubated at 56 °C with occasional vortexing for 2 hours followed by addition of 1 µl RNase A (100 mg/µl) and incubation at room temperature for 10-15 minutes. The genomic DNA was extracted by addition of 0.5 volumes each of phenol and CHISAM (chloroform: isoamyl alcohol 24:1). The sample was mixed carefully by vortexing and centrifuged at 13,000 rpm for 10 minutes, the water phase was collected and an equal volume of CHISAM was added to the sample and mixed by vortexing followed by a second centrifugation with same settings. The water phase was collected, and the DNA was precipitated with an equal volume of isopropanol. The precipitate was recovered by immediate centrifugation at 13,000 rpm, 4 °C, 30 min. The supernatant was removed, and the pellet was washed with 70% ice cold ethanol and then centrifuged at 13,000 rpm, 4 °C, 10 minutes. The supernatant was removed, air-dried briefly and resuspended in TE-buffer (10 mM Tris-HCl pH 8.0, 1 mM EDTA). The genomic DNA was further purified using Genomic-tip 20/G (Qiagen) according to the manufacturers instructions.

Finally, the genomic DNA was purified using phenol-chloroform extraction according to the "Extracting DNA Using Phenol-Chloroform" protocol (https://www.pacb.com/wp-content/uploads/2015/09/Experimental-Protocol-Extracting-DNA-using-Phenol-Chloroform.pdf). The DNA solution was mixed with an equal volume of

the phenol/chloroform/isoamyl alcohol solution (25:24:1). The tube was mixed vigorously for 1 min and spun at high speed for 5 min. The aqueous phase was removed, and the tube was back extracted using EB buffer. The original and back-extracted material was combined and extracted as above using chloroform/isoamyl alcohol (24:1). The aqueous phase was precipitated using 2.5× volume 100% ethanol and NH_4OAc (0.75 M final concentration) while adding glycogen (20 ug /extraction). The precipitate was recovered by high speed centrifugation and washed two times by double volumes of 80% ethanol. The pellet was air-dried and brought up in a small volume of EB buffer. The concentration and quality of the extracted DNA was determined by NanoDrop and agarose gel electrophoresis.

RNA purification

Total RNA from trophozoites were extracted using TriZol according to the method in Franzén et al.[1]. The total RNA was DNase treated using the Turbo DNA-free kit (Applied Biosystems, AM1907) as described in Franzén et al.[1]. RNA from cysts and excyzoites were extracted using TriZol as described above and the resulting RNA was DNase treated (Turbo DNase, 2 U/µl) in a reaction volume of 100 µl for 30 minutes at 37°C. The RNA was subsequently cleaned using the cleanup protocol from the RNeasy Mini Kit (Qiagen, 74104) and eluted in 30 µl of DEPC treated water. The quality and quantity were evaluated using Qubit and BioAnalyzer.

Sequencing

Total genomic DNA was sequenced using both Illumina MiSeq and PacBio RS II sequencers at the SNP&SEQ platform and Uppsala Genome Center respectively at the Science for Life Laboratory (Uppsala University). The trophozoite total RNA was divided in two parts and stranded transcriptome mRNA library was prepared using the appropriate TruSeq kit at GATC Biotech (Konstanz, Germany). The finished libraries were sequenced using the Illumina HiSeq 2000 system at GATC Biotech (Konstanz, Germany). RNA sequencing

libraries from cysts and excyzoites were prepared using the TruSeq stranded mRNA sample preparation kit and sequenced by HiSeq 2500 at the SNP&SEQ platform.

Raw DNA and RNA sequence reads are archived at NCBI Sequence Read Archive (SRA) under accession numbers SRR8858297-SRR8858305.

Biosamples and accession numbers

Biosample accession	**SAMN11231833**		**SAMN11231832**		**SAMN11231834**
Sample description	Cysts		Trophozoites		Excysted trophozoites
DNA/RNA	DNA	RNA	DNA	RNA	RNA
Sequencing instrument	PacBio RS II	Illumina HiSeq 2500	Illumina MiSeq	Illumina HiSeq 2000	Illumina HiSeq 2500
Usage	Genome assembly	RNA-Seq comparison	Assembly base correction	RNA-Seq comparison	RNA-Seq comparison
SRA accession	SRR8858302-SRR8858305	SRR8858298	SRR8858300	SRR8858301	SRR8858299
# Reads	200,601	31,397,561	38,048,422	121,271,107	30,381,809
Read len (bp)	-	125	251*2	51	125
# Total Bbp	2.5	3.9	19.1	6.2	3.8

Repeat detection

RepeatMasker v3.3.0 (http://www.repeatmasker.org/) was run to detect genome repeats. Sequence comparison in RepeatMasker was performed by cross_match v1.080812. RM database v20110920 with RepBase Update 20110920 was used as RepeatMasker library.

Heterozygosity estimation

Samtools mpileup with B flag was used to generate pileup file from MiSeq reads mapped bam file. Sites with at least 20X coverage and 10% of alternative bases were considered as allelic heterozygous sites.

Genome annotation

Two gene prediction programs were used, Prodigal v2.60 and GlimmerHMM v3.0.1. Training set for GlimmerHMM consists of genes predicted from Prodigal which have hits against *G. intestinalis* with evalue <1e-50 and similar length to the hit gene. The union of the two sets of predicted genes was taken for functional annotation. Genes were manually annotated using information from BLASTP results against NR database as well as HMMER (v3.0) search results of domain information against Pfam (v27.0).

tRNA and rRNA

18S, 28S and 5.8S ribosomal RNAs were annotated after BLASTN similarity to the deposited *G. muris* ribosomal RNAs in NCBI (X65063), and the start and stop positions were adjusted to be inline with the record. 5S rRNAs were predicted by searching against Rfam 11.0 using infernal v1.0.2. There are 78 pieces of 28S rRNAs, 73 18S rRNAs, 68 5.8S rRNAs and 10 5S rRNA annotated.

tRNAs were predicted by tRNAScan-SE v1.23 with the most sensitive co-variance model. BLASTN with -task blastn (which is more sensitive for distant matches) of the 40 *G. intestinalis* ncRNAs against *G. muris* genome revealed 10 ncRNAs with evalue <0.1 and not overlapping genes. Alignment of the 10 ncRNAs using MUSCLE v3.8.31 showed a clear motif on the 3' end, CCTTYNHTNAA, which is similar to what was found on *G. intestinalis* (CCTTYNHTHAA). This motif was used to search for other potential ncRNAs in *G. muris* using scan_for_matches (http://blog.theseed.org/servers/2010/07/scan-for-matches.html). Additionally, results were obtained from BLASTN search against miRBase, similarity search against Rfam and prediction of small RNAs based on small RNA-Seq data using Shortstack v1.2.4. All the search and prediction results were combined following the rule that ncRNAs do not overlap (or overlap little) with the annotated genes and with themselves. Manual efforts were also applied. In the end, they are 219 ncRNAs annotated, with 16 from similarity

search against *G. intestinalis* ncRNAs, 8 from Rfam searches, 17 from Shortstack prediction, and 178 from motif search.

Protein kinases

Protein kinases were assigned into group, family and subfamily. Alignment files for different group, family and subfamily were downloaded from the kinase database (http://kinase.com). HMMER 3.0 (http://hmmer.janelia.org) was used to build HMM profiles from the alignment and the protein sequences were searched against the HMM profiles. The search results were then combined with Pfam domain search results where genes contain significant Pfam Pkinase domain (PF00069) (score >25). *G. intestinalis* protein kinase annotations were taken from reference.

Ankyrin repeat proteins

Ankyrin repeat proteins (ARPs) were assigned to genes that contain ankyrin repeat domains (score >=20) but not Pkinase domain. According to if an ARP contains zink finger domains or other domains, they were further divided into 3 subgroups.

Cysteine-rich proteins

Alignments of the proteins with more than 10% cysteines using MAFFT v7.215 revealed that a subset of them contains a conserved pentapeptide motif (GR)CR(GRKE)K at the C-terminus. Cysteine-rich proteins were divided into three subgroups depending on the presence and absence of the transmembrane domains and the pentapeptide motif.

Synteny

MUMmer v3.23 was used to align the draft genomes of *G. muris* and *G. intestinalis*. Promer was used in order to align the genome at protein level. Mummerplot was used to generate the alignment dot plot. Show-coords was used to view a summary of all the alignments produced by promer. Alignment coordinates were used to calculate the percentage of similarity between the two genomes as well as to draw the similarity in Circos plot.

OrthoMCL v2.0.2 was used to cluster genes using match cutoff of 40% and e-value cutoff of 1e-10. The clustering was done within *G. muris*, between *G. muris* and *G. intestinalis*, and among *G. muris*, *G. intestinalis* and *S. salmonicida*.

Phylogenetic trees for NEKs and ARPs

All the NEKs from *G. muris* and *G. intestinalis* were aligned with MAFFT v7.407 (auto setting), and FastTree v2.1.10 No SSE3 was used to construct approximately-maximum-likelihood tree using default WAG+CAT model. ARPs tree was constructed the same way as NEKs tree.

Phylogenetic analysis of lateral transfers

One sequence from *G. muris* was used as a query for each enzyme to retrieve at least 5000 hits with e-value <0.001, using BLASTP against the nr database and the organism-specific proteomes. The resulting datasets were filtered automatically by finding all sequences with greater than 90% sequence identity to another sequence using the software skipredundant (EMBOSS v6.5.7.0 [2]) with default settings. Datasets were aligned using MAFFT v6.603b[3], and regions of ambiguous alignment were automatically trimmed using BMGE v1.12 (blosum30 and block size of 2) [4]. Sequences with a length less than 50% of the multiple alignment were removed, and the datasets were aligned and trimmed again as described above. An initial phylogenetic analyses were performed using FastTree v2.1.8 SSE3, OpenMP [5]. Sequences with a phylogenetic distance < 0.3 were removed in an iterative process to create the final datasets. This final datasets were aligned in forward and reverse orientation using MAFFT and PROBCONS v1.12 [6]. The four resulting alignments were combined with T-COFFEE v11.00.8cbe486 [7] into a maximal consensus alignment, which was trimmed using BMGE as in previous steps. Maximum Likelihood (ML) trees were computed using IQtree v1.6.5 [8] under LG4X substitution model [9]. Branch supports were

assessed using ultrafast bootstrap approximation (UFboot) with 1,000 ultrafast bootstraps [10] and 1,000 replicates for SH-like approximate likelihood ratio test (SH-aLRT) [11].

Promoter motif search

MEME suite v4.10.2 was used for promoter motif analysis. A maximum of 120 bp upstream of annotated genes were used in search of potential promoter. The longest possible sequence was used if the intergenic region is shorter than 120 bp, and the sequence was discarded if the intergenic region is shorter than 8 bp. 4139 sequences in total were used in motif search, and MEME was set to search for 10 most likely motifs with sizes from 6 to 30 bp.

VSP and ψVSP phylogenetic analysis

Nucleotide sequences of *G. muris* VSPs and ψVSP genes were aligned using MAFFT in MAFFT v7.416[3], and regions of ambiguity were trimmed using BMGE v1.12 (DNAPAM100:2, [4]). Maximum Likelihood (ML) trees were computed using IQtree v.1.6.5 [8] under the best model (TIM2+F+R6) with 1,000 ultrafast bootstraps [10] and 1,000 bootstrap replicates for SH-like approximate likelihood ratio test (SH-aLRT)[11].

References

1. Franzén O, Jerlström-Hultqvist J, Einarsson E, Ankarklev J, Ferella M, Andersson B, et al. Transcriptome Profiling of Giardia intestinalis Using Strand-specific RNA-Seq. PLoS Computational Biology. 2013; 9(3):e1003000. doi:10.1371/journal.pcbi.1003000.

2. Rice P, Longden I, Bleasby A. EMBOSS: The European Molecular Biology Open Software Suite. Trends in Genetics. 2000; 16(1):276–277. doi:10.1016/j.cocis.2008.07.002.

3. Katoh K, Misawa K, Kuma K, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic acids research. 2002; 30(14):3059–3066. doi:10.1093/nar/gkf436.

4. Criscuolo A, Gribaldo S. BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. BMC evolutionary biology. 2010; 10(1):210. doi:10.1186/1471-2148-10-210.

5. Price MN, Dehal PS, Arkin AP. FastTree 2--approximately maximum-likelihood trees for large alignments. PLoS ONE. 2010; 5(3):e9490. doi:10.1371/journal.pone.0009490.

6. Do CB, Mahabhashyam MSP, Brudno M, Batzoglou S. ProbCons: Probabilistic consistency-based multiple sequence alignment. Genome research. 2005; 15(2):330–40. doi:10.1101/gr.2821705.

7. Notredame C, Higgins DG, Heringa J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. J. Mol. Biol. 2000; 302(1):205–217. doi:10.1006/jmbi.2000.4042.

8. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Biol. Evol. 2015; 32(1):268–274. doi:10.1093/molbev/msu300.

9. Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat. Methods. 2017; 14(6):587–589. doi:10.1038/nmeth.4285.

10. Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. UFBoot2: Improving the Ultrafast Bootstrap Approximation. Mol. Biol. Evol. 2018; 35(2):518–522. doi:10.1093/molbev/msx281.

11. Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst. Biol. 2010; 59(3):307–321. doi:10.1093/sysbio/syq010.