Supplementary Figure legends

Figure S1: Sequences of new split-inteins, grouped by the type of host protein.

Figure S2: Ou tput of a CD-search (Marchler-Bauer and Bryant 2004) sequence-to-multiple alignments comparison, of the putative DNA-repair exonuclease synthetically joined products (excluding the VSR region) with the SbcD multiple sequence alignment (COG0420). The multiple alignment is represented by its consensus sequence. The joining point of the putative DNA-repair sequence parts includes an "x" and is marked by an arrow.

Figure S3: (A) Annotated multiple sequence alignment of group I introns that include Vsr-like ORFs. The *Bacillus thuringiensis* phage 0305ϕ8-36 (BP0305phi; NCBI accession NC_009760.1) and *Bacillus cereus* AH1134 (bacce-AH1; NCBI accession NZ_ABDA01000203.1). Group I introns of recA genes a re aligned with the *Bacillus anthracis* Sterne recA group I intron (bacan-Ste; NCBI accession AF229167.1). The group I intron regions are indicated by a red line. Their predicted secondary structures (P1 to P10) and conserved sequence motifs (R and S) are annotated according to the *B. anthracis* intron (Ko et al. 2002). Protein coding regions have their amino acids sequence above them. The recA exons are highlighted in yellow, the Vsr-like ORFs are highlighted in green, the putative hairpins at the 5' of the Vsr-like ORFs are double underlined, the conserved Vsr-like motifs (Figure 3) are highlighted in grey, and group I introns conserved core regions are highlighted in cyan. (B) Predicted RNA secondary structure of the group I intron of recA gene from *Bacillus thuringiensis* phage 0305ϕ8-36 (BP0305phi; NCBI accession NC_009760.1). Predicted secondary structures P1 to P10, conserved sequence elements R and S, the G site, and the ORF insertion point and its stop codon are shown according to (A).

Figure S4: Conserved nucleotide motifs in 5' untranslated regions of endonuclease ORFs. Motifs were searched using MEME. The position is o f the distance of the sequence 3' ends from the translation initiation codons. The motifs are shown aligned with their logos (calculated using the WebLogo server at http://weblogo.berkeley.edu/) (A) and in the context of the whole region (B). In the later view, the sequences are aligned by their initiation codons, at their 3' termini, that are also boxed. "…" indicates unknown sequence data.

Table S1
Protein hosts and the GOS reads used as a source for their assembly.

Protein host	GOS location	JCVI reads	JCVI assemblies	Probable origin
gp41-1	Lake Gatun	1095368026018+1095366023924+1095333010713+1093023041874+1095333009945+1093023002130	1097207246556+1097207259355+1101669430723	T4-like viruses
gp41-2	Lake Gatun	1101669426331		
gp41-3	Lake Gatun	1095351007293+1095356001149	1061005454744	
gp41-4	Lake Gatun	093022163129+1093022103376	1097207258370	
gp41-5	Lake Gatun	1091143037864	1097207240963	
gp41-6	Lake Gatun	1095337027236+1095306084054	1097207277771	
gp41-7	Off Nags Head	1093012253551+1092963600993	1097205056197	
gp41-8	Lake Gatun	1091143028835	1097207240798	T4-like viruses
gp41-9	Punta Cormorant	1097156422586		
IMPDH-1	Lake Gatun	1091143056039+1095333020518+1091142153811+1095333021862	1101669394623	Bacterial
IMPDH-2	Lake Gatun	1095351023224+1093022056516+1093022126404+1095349055864	1097207259656	
IMPDH-3	Delaware Bay	1095899230244+1095898151644	1101669181432	
DnaE-1	Rangirora Atoll	1092963363647+1092963253032	1097263593299	Bacterial
DnaE-2	Lake Gatun	1095349061806+1095328030710	1101669426735	
DnaE-3	Gulf Of Mexico	1093022142172+1093022142942	1097205348472	
NrdJ-1	Lake Gatun		1101669428687	Bacterial and dsDNA viruses
NrdJ-2	Lake Gatun		1101669429579	
NrdA-1	Lake Gatun	1095368021404+1095349056558	1101669410105	Bacteria and viruses
NrdA-2	Rangirora Atoll	1092963091065+1093006402449	1097263589660	
NrdA-3	Lake Gatun	1091143060848	1101669394741	
NrdA-4	Lake Gatun	1091143047300+1091142207694	1101669393609	
NrdA-5	Punta Cormorant	1095521346405	1101670305357	
NrdA-6	Lake Gatun	1091142141262+1091143128742	1097207263468	
NrdA-7	Lake Gatun	1097207240466		
DNA ligase	Lake Gatun	1095326019642+1095337013586	1097207283029	viruses
Terminase	Lake Gatun	1095306046644+1095328063912	1101669405733	Myoviridae viruses
Unknown	Off Key West	1097205249970	1091138266427	

Table S2
List of salt bridges across the intein halves as calculated from their models.

gp41-1

```
3  ASP A - 103 ARG B
41 LYS A - 107 ASP B
45 LYS A - 102 GLU B
48 LYS A - 98  GLU B
52 GLU A - 92  LYS B
52 GLU A - 93  LYS B
53 ASP A - 92  LYS B
55 LYS A - 120 ASP B
80 GLU A - 96  LYS B
87 LYS A - 120 ASP B
```

IMPDH-1

```
18 GLU A - 120 LYS B
35 GLU A - 103 LYS B
72 ARG A - 113 GLU B
```

NrdA-2

```
32 ASP A - 101 ARG B
65 HIS A - 104 ASP B
```

NrdJ-1

```
55  ARG A - 128 ASP B
72  LYS A - 140 ASP B
75  ARG A - 124 GLU B
97  ASP A - 116 LYS B
98  ASP A - 114 LYS B
98  ASP A - 116 LYS B
100 GLU A - 114 LYS B
```

gp41-8

```
18 ARG A - 114 GLU B
18 ARG A - 115 GLU B
43 LYS A - 115 GLU B
52 LYS A - 104 GLU B
60 ARG A - 114 GLU B
91 LYS A - 97  GLU B
91 LYS A - 99  GLU B
```

Figure S1

N-inteins:

>gp41-1
CLDLKTQVQTPQGMKEISNIQVGDLVLSNTGYNEVLNVFPKSKKKSYKITLEDGKEIICSEEHLFPTQTGEMNISGGLKEGMC
LYVKE
>gp41-2
CLDLKTQVQTQQGLKDISNIQVGDLVL
>gp41-3
CLDLKTQVQTPQGMKEISNIQVGDLVLSNTGYNEVLNVFPKSKKKS
>gp41-4
CLDLKTQVQTPQGMKEISNIQVGDLVLSNTGYNEVLNVFPKSKKKSYKIT
LEDGKEIICSEEHLFPTQTGEMNISGGLKEGMCLYVKE
>gp41-5
CLDLKTQVQTPQGMKEISNIQVGDLVLSNTGYNEVLNVFPKSKKKSYKIT
LEDGKEIICSEEHLFPTQTGEMNISGGLKEGMCLYVKE
>gp41-6
SYKITLEDGKEIICSEEHLFPTQNGEVNIKGGLKEGMCLYVKE
>gp41-7
CLDLKTQVQTPQGMKELSNIQVGDLVLSNTGYNQVLNVFPKSKKKSYKIT
LEDGKEIICSEEHLFPTQNGEVNIKGGLKEGMCLYVKE

>gp41-8
CLSLDTMVVTNGKAIEIRDVKVGDWLESECGPVQVTEVLPIIKQPVFEIV
LKSGKKIRVSANHKFPTKDGLKTINSGLKVGDFLRSRAK

>IMPDH-1
CFVPGTLVNTENGLKKIEEIKVGDKVFSHTGKLQEVVDTLIFDRDEEIIS
INGIDCTKNHEFYVIDKENANRVNEDNIHLFARWVHAEELDMKKHLLIELE

>DnaE-1
CFTKDTNILTEHGFMDIEELDPKRDGVIYIDKEGKHRINHDYELHYMGRK
EVFEIKTECGKTIKLTSDHEVMTQEGYKKVFELNENDVLIKF
>DnaE-2
ILTNNGEKFINEISCNDQIAYLTNHNSIEYNNDYEIIFQGKKEIFEIILENGTALELTED
HEVMTQNGYKKVKELNDDDSLVNI

>NrdA-1
CVAGDTKIKIKYPESVGDQYGTWYWNVLEKEIQIEDLEDYIIMRECEIYDSNAPQIEVLSYNIETGEQEWKPITAFAQTSPKA
KVMKITDEESGKSIVVTPEHQVFTKNRGYVMAKDLIETDEPIIVNKDMNF
>NrdA-2
CLTGDAKIDVLIDNIPISQISLEEVVNLFNEGKEIYVLSYNIDTKEVEYKEISDAGLISESAEVLEIIDEETGQKIVCTPDHK
VYTLNRGYVSAKDLKEDDELVFS
>NrdA-4
CLAGDTTVTVLEGDIVFEMTLENLVSLYKNVFSVSVLSFNPETQKQEFKPVTNAALMNPESKVLKITDSDTGKSIVCTPDHKV
FTKNRGYVIASELNAEDILEIK
>NrdA-5
HTETVRRVGTITAFAQTSPKSKVMKITDEESGNSIVVTPEHKVFTKNRGYVMAKNLVETDELVIN
>NrdA-6
YVCSRDDTTGFKLICTPDHMIYTKNRGYIMAKYLKEDDELLINEIHLPT

>NrdJ-1
CLVGSSEIITRNYGKTTIKEVVEIFDNDKNIQVLAFNTHTDNIEWAPIKAAQLTRPNAELVELEIDTLHGVKTIRCTPDHPVY
TKNRGYVRADELTDDDELVVAI
>NrdJ-2
CLVGSSEIITRNYGKTTIKEVVEIFDNDKNIQVLAFNTHTDNIEWAPIKAAQLTRPNAELVELEINTLHGVKTIRCTPDHPVY
TKNRDYVRADELTDDDELVVAI

C-inteins:

>gp41-1
MMLKKILKIEELDERELIDIEVSGNHLFYANDILTHNS
>gp41-2
MMLKKILKIEELDERELIDIEVSGNHLFYANAILTHNS
>gp41-7
MMLKKILKIEELDERELIDIEVSGNH

>gp41-8
MCEIFENEIDWDEIASIEYVGVEETIDINVTNDRLFFANGILTHNS
>gp41-9
MIMKNRERFITEKILNIEEIDDDLTVDIGMDNEDHYFVANDILTHNT

>IMPDH-1
MKFKLKEITSIETKHYKGK-VHDLTVNQDHSYNV-RGTVVHNS
>IMPDH-2
MKFTLEPITKIDSYEVTAEpVYDIEVENDHSFCVeNGFVVHNS
>IMPDH-3
MKFKLVEITSKETFNYSGQ-VHDLTVEDDHSYSI-NNIVVHNS

>DnaE-2
MKNFWRKLLKLLKIKSIKKSRIDNVYDIHHRINYKVFFDEHPNLIAEKIVISNC
>DnaE-3
MNLLGKQQTYDLEVAHHDHQYYLNNGILQSNS

>NrdA-2
MGLKIIKRESKEPVFDITVKDNSNFFANNILVHNC
>NrdA-3
MLKIEYLEEEIPVYDITVEETHNFFANDILIHNC
>NrdA-5
MLKIEYLEEEIPVYDITVEGTHNLAYSL
>NrdA-6
MGIKIRKLEQNRVYDIKVEKIIIFCNNILVHNC
>NrdA-7
MLKIEYLEEEIPVYDITVEKTNNFFANDILVHNC

>NrdJ-1
MEAKTYIGKLKSRKIVSNEDTYDIQTSTHNFFANDILVHNS

>DNA ligase
MQIVRVKKIAKVESRDKFDLEVKKNNNFFANGVLVHNC

>Terminase
MLSDQVERKFTETVEVTDWQVDTDTGWQEVTASNQTIPYAVHELELDNGM
FLSCADTHIVFDQYLNEIFVQDLVPGQQIQTVSGVSAVKSLTATQDQQQMYDLSVNSKDQRYYTGGILSHNS

>unknown host
MDDLYMLDEDEIVSIELIGEEDTIDITVDDTHMFFANDIYTHNS

Figure S2


Description
Pssmid
Multi-dom
E-value
COG0420, SbcD, DNA repair exonuclease [DNA replication, recombination, and repair]
30769
yes
6e-06
DNA repair exonuclease [DNA replication, recombination, and repair]
CD Length: 390 Bit Score: 47.10 E-value: 6e-06
SbcD-like 11 TEFLKNLADIA-PTFIIAGNHDlNLSNMNRLDSLTPLVNSINHPNLWYLKHSG----------IYTVADTDFAVYSILDE 79
COG0420 65 LEALRRLKDAGiPVVVIAGNHD-SPSRLSEASPLLLLNNLGLHGVVGRLVHEIrppeivaapwLIPGPDPDVVFFLGLNG 143
SbcD-like 80 KELWP----------SHKDCNAKKKVALYHGPVHGAETDAR--YVITNRHVNVSMFDGFDMVLLGDIHKYQVLQERseey 147
COG0420 144 LEKEQfelllhkgllSALDPDDDPSILVLHQSIDALTSGAErdLALGTVDLSLLPKGGFDYVALGHIHKRQVIPKE---- 219
SbcD-like148 ievdakevdsylkkgwtidvxmrkvklkrvKPIIVYPSSTIQQNHGETVKGHGWCLWDM-KNFTHIFKEVPNNFGYYTLE 226
COG0420 220 ----------DPPIVYPGSPERYSFGEEGERKGVVLVEFsGGKLWRFEELFVPLFERLEV 269
SbcD-like227 VKNGKVPVLADVPKNVRMRiftGTLDTTGVKKLVSVLRKQHNIIELSINKSRYNNTEQEKIKSGMELVDVQNVNVQNTLI 306
COG0420 270 DVLNGEDLEDELLKLDREE---DAKDEPVVRLTIEGETDILDERLKERILALYERIERLELLLEERATDLAEEPESSELE 346
SbcD-like307 QDWLERTYDNTISPNLMKKILD 328
COG0420 347 ELLSELDVEEAFEELLADEVLL 368

Figure S3

A


yValSerAspValAlaIleGlnArgTrpArgLysLysArgHisGlyLysPheLysProGlnIleAspThrSerThrHisLeuThrThrProGluArgArg
BP0305phi CGTAAGTGATGTTGCTATCCAACGATGGAGAAAGAAAAGACACGGTAAGTTTAAGCCACAAATTGATACATCGACACATCTAACAACACCTGAACGTCGT
yValSerAspThrCysValAlaLysArgArgLysGluLeuAsnValLysValArgLysLys AsnTyrAspThrLeuLeuGluGlnGln
bacce-AH1 CGTATCCGATACATGTGTTGCTAAGAGGAGAAAAGAGCTTAATGTAAAGGTCAGAAAGAAA----------AATTATGATACTTTACTGGAACAACAG
bacan-Ste ----------
ValLysGluIleLeuAspGluLeuAspIleValTyrPheThrHisHisValValGluGlyTrpAsnValAspPheTyrLeuGlyLysLysLeuAlaIleG
BP0305phi GTAAAAGAAATACTTGATGAATTAGACATCGTGTATTTTACACATCACGTTGTTGAAGGATGGAATGTAGACTTCTATTTAGGTAAGAAACTAGCGATAG
ValGluGlnMetLeuLeuSerLeuAspLeuAlaPheIleLysGlnLysArgIleAspLysTrpSerIleAspPheTyrLeuGlyArgLysTyrCysLeuA
bacce-AH1 GTTGAACAGATGTTGTTAAGTTTAGATCTGGCGTTCATTAAGCAAAAAAGAATCGATAAGTGGTCTATTGATTTCTACTTAGGTAGAAAATATTGTTTAG
bacan-Ste ----------
luValAsnGlyValTyrTrpHisSerLysGlnLysAsnValAsnLysAspLysArgLysLeuSerGluLeuHisSerLysGlyTyrArgValLeuThrIl
BP0305phi AAGTAAACGGTGTATATTGGCACAGTAAACAAAAGAATGTAAATAAGGATAAACGTAAGTTGTCTGAACTACACAGTAAAGGATACCGTGTATTAACAAT
spValHisGlyLysTrpAlaHisSerLeuLysLysIleLysGluArgAspLysArgLysLeuLeuPheMetGluGluGlyCysTyrLysTyrLeuValIl
bacce-AH1 ATGTCCATGGGAAATGGGCACACTCACTCAAAAAGATAAAAGAACGAGATAAAAGAAAACTTTTATTTATGGAAGAGGGCTGCTATAAATATCTTGTTAT
bacan-Ste ----------
eGluAspAspGluLeuAsnAspIleAspLysValLysGlnGlnIleGlnLysPhe TrpValThrHisIleSerAsnGlyMet***
BP0305phi CGAAGACGATGAGTTAAATGATATAGATAAGGTGAAACAACAAATACAAAAGTTT----TGGGTCACCCATATCAGTAATGGTATGTAAAATAACCTTGT
eHisGluGluGluLeuAlaAsnLysGluLysValLeuGlnLysIleLysGluPheThrMetGlyPheProCys***
bacce-AH1 ACATGAAGAAGAGTTAGCAAATAAAGAAAAAGTCTTACAAAAGATTAAGGAGTTTACCATGGGTTTCCCTTGCTAGTGATAGCATGTTAAA-AACCTTGT
bacan-Ste ----------AAAGATTAAGGAGTTTACCATGGGTTTCCCTTGCTAGTGATAGCATGTTAAAAAACCTTGT
----------P10--P1-> <-P2------P2-> <--P3--
GAACGCAAGCAAAAGCGGTGTCGTCGTAAGACG-----GCTAACGGTGGACGGAAGGCTCCCAA-------TACCGTGCCAAGC-CCA-------ATAGG
bacce-AH1 GAACCTTATTGCCTAAG-GTGTA-CTATCCATATAGTTGCTAACGGTGAAGCCCA-CATTATTTAGGGTA-TACCGTGCCAAGCT-CATAGTGATATGGG
bacan-Ste GAACCTTATTGCCTAAGGGTGTAACTATTCATATAGTTGCTAACGGTGAAGCCCACCATTATTCAGGGTAATACCGTGCCAAGCTCCATAGTGATATGGG
P3 > <-P3.1---P3.1-> <-P3.2--------P3.2-> <-P4-- <-P5----------P5-> --P4-><P6 <-P6a----------
GAAGGTGTAGAGACTATCGAAA-GCACATCATACGATGGAAGCGAGTAGAGTAGGATGGAGATTAGCACCATTCGAAGCGCAAGGCTCCTCGATAGAGGA
bacce-AH1 GAAGG-GTAGAGACTAC-GAAA-GGACTTT-AGT-AACTGAGTAGG-TACG
bacan-Ste GAAGGTGTAGAGACTACCGAAAAGGACTTTTAGTTAACTGAGTAGGGTACGGCAGATGATAGGCTACTGCTG--GAAGTGCAAGGCTCTCTGTAAAGAGA
P6a > P6> <-P7--- <-P7.1-------P7.1-> <-P7.2----------P7.2-> <--P3-><-P8----------
=R=motif==
ProAsnValThrThrGlyGlyArgAlaLeuAlaPheTyrAlaSerLeu
BP0305phi TGAAGATATAGTCCATTACACTTAAATGTGTAAG----------ccaaacgtgacaactggtggtcgtgcgttagcgttctatgcgtcatta
ProGluThrThrProGlyGlyArgAlaLeuLysPheTyrSerThrVal
bacan-Ste TGAAGAGATAGTCCGGACTATAGAGATGGAAAATCTATAGATAGTG------ccagaaacaactccaggtggtcgtgcgttgaaattctattcaactgtt
P8-> --P7-> <-P9----------P9-> P10
=S=motif====
ArgPheGlyMetAsnLysLeuLysLeuGlnAspSerAspProIleThrAspAspAspGlyLeuLysIleSerAlaArgValAlaLysAsnArgCysValT
BP0305phi cgttttggtatgaataaactgaaactacaagatagtgacccaattacggacgacgatggactgaaaatcagcgctcgtgttgcgaagaaccgatgtgtat
ArgLeuGluValArgArgAlaGluGlnLeuLysGlnGlyAsnAspIleValGlyAsnLysThrLysValLysValValLysAsnLysValAlaPro
bacan-Ste cgtcttgaagtgcgtcgtgcggagcaattaaaacaaggtaacgacatcgttggtaataaaacaaaagtaaaagtagttaaaaataaagtggcacca----
yrAspAsnProTyrLysLysCysGluTyrTyrValIleTyrGlyGluGlyValAspLysLeuThrGluIleMetGluAsnValGluProAlaGlyIleMe
BP0305phi acgacaacccttacaaaaaatgtgaatattacgttatttatggtgagggtgtcgataagcttactgagattatggagaacgttgaacctgctggcatcat
ProPheArgValAlaGluValAspIleMetTyrGlyGluGlyIleSerArgGluGlyGluIleLeuAspMetAlaSerGluLeuAspIleVa
bacan-Ste --------ccattccgtgttgcggaagttgatattatgtacggagaaggtatttcaagagaaggtgaaatcttagatatggcctctgaacttgatatcgt
tArgLysSerGlySerTrpPheTyrTyrGluLysGluAspGlyGluGlnIleGluAlaAlaGluAlaLeuValArgLysLysAspAspProValProGln
BP0305phi gcgtaagtcaggctcatggttctattacgagaaagaagacggtgagcagattgaagcagccgaggcacttgttcgtaagaaggatgatccagtaccgcag
lGlnLysSerGlyAlaTrpTyrSerTyrAsnGluGluArgLeuGlyGlnGlyArgGluAsnSerLysGlnPheLeuLysGluAsnThrAspLeuArgGlu
bacan-Ste tcaaaagagcggtgcttggtactcttataatgaagaacgcttaggacaaggtcgtgaaaattcgaagcagttcttaaaagaaaatacggatttaagagaa
MetAlaLysAsnValProLeuLysPheProGlyLysThrLysPheArgGluPheMetThrAspAsnProTrpPheAlaGluGlnLeuArgGlnGluIleA
BP0305phi atggcgaagaatgtaccattgaagtttccgggtaaaacgaagttccgtgagttcatgacggataatccgtggttcgctgaacagttacgtcaggagattc
GluIleAlaPhePheIleArgGluHisHisGlyIleSerGluAspSerGlyAlaGluGlyMetGluAspProAsnLeuLeuAsp***
bacan-Ste gaaattgccttctttattcgtgagcatcacggaattagcgaagattctggtgctgaaggtatggaagatccaaatcttcttgattaa
rgGlyLysValLysGlnGlyGluLeuLeuProLysTyrGlnAspAlaAspGluMetLysAspIleGluGluLeuGluLysIleGluLysGlnIleAlaAl
BP0305phi gcgggaaagtaaaacaaggtgaactacttccgaagtatcaagatgctgatgaaatgaaagacatcgaagaacttgagaaaatcgagaagcaaattgccgc
aGluGluGluAlaAlaLeuLysLysAlaSerAlaLysLysIleThrThrLysLysProValAlaAspLysLysProAlaThrLysProAlaArgLysThr
BP0305phi agaggaagaagctgctttgaagaaagcgtcggcaaagaagattacaacgaaaaagccggtagcagataagaagccagcgacgaaacctgcacgtaaaaca
ProAlaLysLysProAlaGluLysLysGluThrThrGluIleVal***
BP0305phi ccagcgaagaagccagccgagaagaaagaaacaactgaaattgtataa

Figure S3

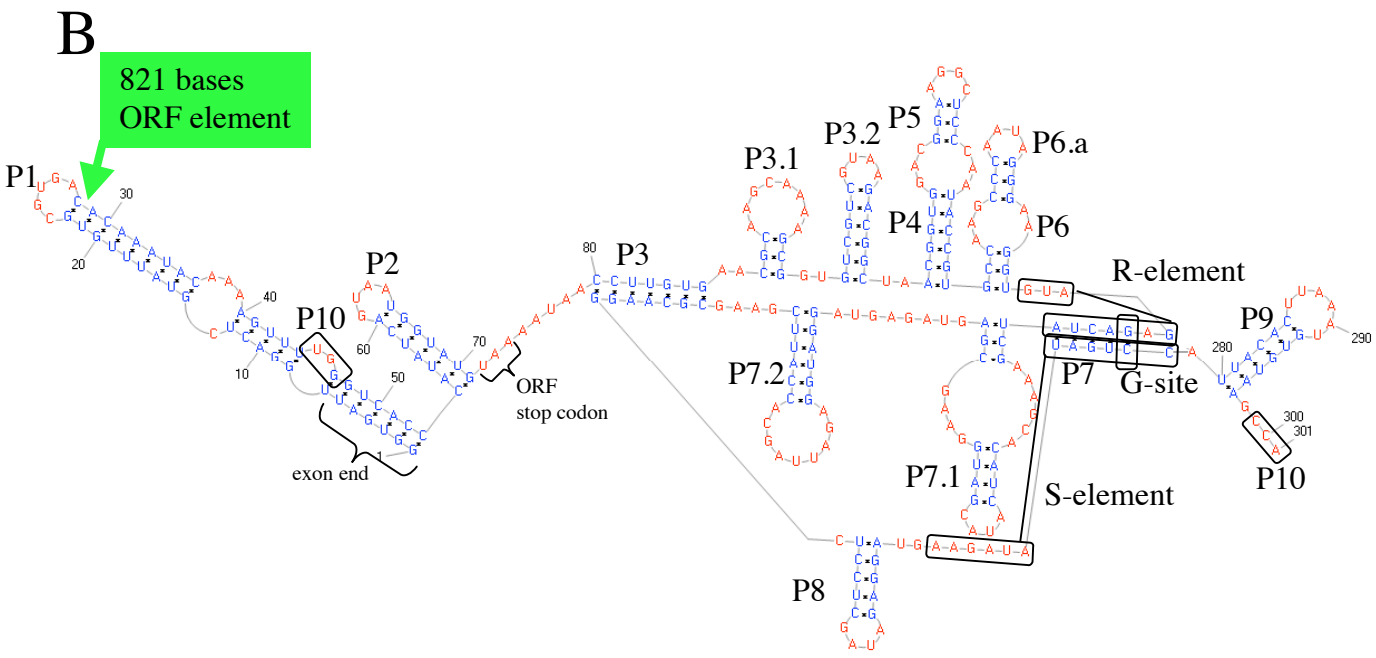

B
821 bases
ORF element
P1
P2
P3
P3.1
P3.2
P4
P5
P6
P6.a
P7
P7.1
P7.2
P8
P9
P10
P10
R-element
G-site
S-element
ORF
stop codon
exon end
1
10
20
30
40
50
60
70
80
280
290
300
301

Figure S4

A

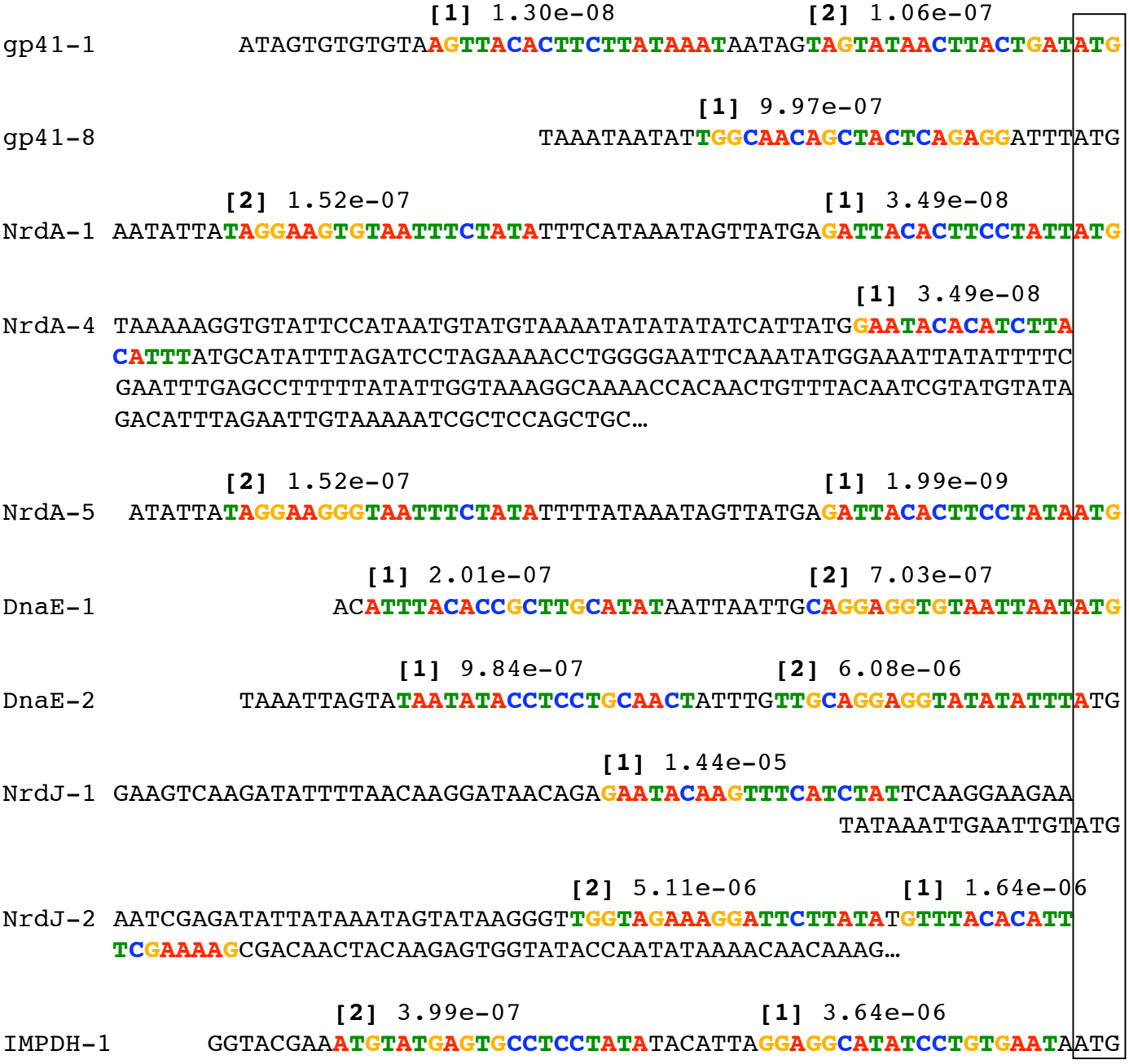
Motif 1 E-value = 1.7e-010

Loci	Position	P-Value	Sequence
gp41-1	22	1.30e-08	AGTTACACTTCTTATAAAT
NrdA-1	0	3.49e-08	GATTACACTTCCTATTATG
NrdA-4	0	3.49e-08	GAATACACATCTTACATTT
NrdA-5	0	1.99e-09	GATTACACTTCCTATAATG
DnaE-1	26	2.01e-07	ATTTACACCGCTTGCATAT
DnaE-2	24	9.84e-07	TAATATACCTCCTGCAACT
NrdJ-1	26	1.44e-05	GAATACAAGTTTCATCTAT
NrdJ-2	38	1.64e-06	GTTTACACATTTCGAAAAG
IMPDH-1	1	3.64e-06	GGAGGCATATCCTGTGAAT

Motif 2 E-value = 9.9e-003

Loci	Position	P-Value	Sequence
gp41-1	0	1.06e-07	TAGTATAACTTACTGATATG
gp41-8	11	9.97e-07	TGGCAACAGCTACTCAGAGG
NrdA-1	34	1.52e-07	TAGGAAGTGTAATTTCTATA
NrdA-5	34	1.52e-07	TAGGAAGGGTAATTTCTATA
DnaE-1	0	7.03e-07	CAGGAGGTGTAATTAATATG
DnaE-2	0	6.08e-06	TTGCAGGAGGTATATATTTA
NrdJ-2	58	5.11e-06	TGGTAGAAAGGATTCTTATA
IMPDH-1	27	3.99e-07	ATGTATGAGTGCCTCCTATA

Figure S4
B


[1] 1.30e-08
[2] 1.06e-07
gp41-1 ATAGTGTGTGTAAGTTACACTTCTTATAAATAATAGTAGTATAACTTACTGATATG
[1] 9.97e-07
gp41-8 TAAATAATATTGGCAACAGCTACTCAGAGGATTTATG
[2] 1.52e-07
[1] 3.49e-08
NrdA-1 AATATTATAGGAAGTGTAATTTCTATATTTCATAAATAGTTATGAGATTACACTTCCTATTATG
[1] 3.49e-08
NrdA-4 TAAAAAGGTGTATTCCATAATGTATGTAAAATATATATATCATTATGGAATACACATCTTA
CATTTATGCATATTTAGATCCTAGAAAACCTGGGGAATTCAAATATGGAAATTATATTTTC
GAATTTGAGCCTTTTTTATATTGGTAAAGGCAAAACCACAACTGTTTACAATCGTATGTATA
GACATTTAGAATTGTAAAAATCGCTCCAGCTGC…
[2] 1.52e-07
[1] 1.99e-09
NrdA-5 ATATTATAGGAAGGGTAATTTCTATATTTTATAAATAGTTATGAGATTACACTTCCTATAATG
[1] 2.01e-07
[2] 7.03e-07
DnaE-1 ACATTTACACCGCTTGCATATAATTAATTGCAGGAGGTGTAATTAATATG
[1] 9.84e-07
[2] 6.08e-06
DnaE-2 TAAATTAGTATAATATACCTCCTGCAACTATTTGTTGCAGGAGGTATATATTTATG
[1] 1.44e-05
NrdJ-1 GAAGTCAAGATATTTTAACAAGGATAACAGAGAATACAAGTTTCATCTATTCAAGGAAGAA
TATAAATTGAATTGTATG
[2] 5.11e-06
[1] 1.64e-06
NrdJ-2 AATCGAGATATTATAAATAGTATAAGGGTTGGTAGAAAGGATTCTTATATGTTTACACATT
TCGAAAAGCGACAACTACAAGAGTGGTATACCAATATAAAACAACAAAG…
[2] 3.99e-07
[1] 3.64e-06
IMPDH-1 GGTACGAAATGTATGAGTGCCTCCTATATACATTAGGAGGCATATCCTGTGAATAATG