Supplementary Figures

(for “RNA Sequencing and Proteogenomics Reveal the Importance of Leaderless mRNAs in the Radiation-tolerant Bacterium Deinococcus deserti” by de Groot et al.)

Contents:

| Figure | Description                                                                 | Pages |
|--------|-----------------------------------------------------------------------------|-------|
| S1     | Codon usage of leaderless and leadered genes                                | 2     |
| S2     | Amino acid composition of leaderless and leadered gene products             | 3     |
| S3     | Read coverage of a region with leaderless clpP and leadered lon             | 4     |
| S4     | Start codon re-annotations of DNA repair genes in D. deserti               | 5-9   |
| S5     | Examples of predicted start codon corrections in Deinococcal homologs of D. deserti proteins | 10-13 |
| S6     | Detected homologs for peptides and proteins from 17 new leaderless transcripts | 14-23 |
| S7     | New proteins detected by proteomics                                         | 24-29 |
| S8     | Examples of several highly induced genes encoding small proteins           | 30-35 |
| S9     | TSS positions relative to RDRM (radiation-desiccation response motif)       | 36    |
Figure S1. Codon usage of leaderless and leadered genes. Relative synonymous codon usage (RSCU) is shown for the 1174 leaderless genes (339156 codons) and 784 leadered genes (277460 codons). RSCU values are the number of times a particular codon is observed, relative to the number of times that the codon would be observed for a uniform synonymous codon usage.
Figure S2. Amino acid composition of leaderless and leadered gene products. The average amino acid composition, in percentages, is indicated for the products of the 1174 leaderless genes (337982 residues) and 784 leadered genes (276676 residues).
Figure S3. Read coverage of a region with leaderless clpP and leadered lon. The proteases ClpP (Deide_19570, Clp protease proteolytic subunit) and Lon (Deide_19590, Lon protease) are produced from leaderless and leadered mRNA, respectively. Coverage (in blue) of reads that map to the forward genomic DNA strand is shown above the genes (results in non-irradiated and irradiated samples were similar for these genes; only RD19 IR and RD19 IR + TEX samples are shown). Panels B-D are zoomed parts of the region shown in panel A. Transcription start sites (TSSs) for clpP and lon are indicated with arrows in panels B and C, respectively. Panel D is a zoom at the translation initiation codon of lon. Start codons, -10 motifs (upstream of TSSs) and SD sequence (upstream of start codon in leadered lon mRNA) are boxed. Treatment (+) or not (−) of RNA with TEX is indicated.
Figure S4. Start codon re-annotations of DNA repair genes in *D. deserti*. RNA-seq read coverage for *uvrA2* (Deide_2p02060), *recN* (Deide_12310), *rarA* (Deide_04980), *ruvA* (Deide_09360), *ruvC* (Deide_20630) is shown. Coverage (in blue) of reads that map to the forward genomic DNA strand is shown above the genes, and those on the reverse strand below the genes (independent of gene annotation and orientation). Above and below the genes, the order of the samples is: RD19 NI, RD19 NI +TEX, RD19 IR, RD19 IR +TEX. In each figure, the maximum height to show coverage is set at the same value for each sample (= in all 8 “lines”), but this value can be different between the figures. New start codons and -10 motifs are boxed.

**Deide_2p02060 (UvrA2)**

uvrA2, zoom at TSS and initial and new translation start

TSS & new start codon
(Figure S4, continued)

Deide_12310 (recN) & flanking

recN, zoom at TSS and new translation start
(Figure S4, continued)
Deide_09360 (ruvA)

ruvA, zoom at TSS and new translation start
(Figure S4, continued)

Deide_20630 (ruvC) & flanking

ruvC, zoom at TSS and new translation start
Figure S5. Examples of predicted start codon corrections in Deinococcal homologs of *D. deserti* proteins. In each example, the TSS was at the first nucleotide of the new start codon of the *D. deserti* protein. Blast and multiple alignments indicate that start codon re-annotation may also be required in several homologs, mainly from other *Deinococcus* species. Only the N-terminal regions of the proteins are shown. New annotated starts of proteins (*D. deserti*), or possible new starts (others), are in green. Initial wrong starts (*D. deserti*), or possible wrong starts (others), in cyan.

**Shorter *D. deserti* proteins:**

| Deide_04020 (7aa) | conserved protein of unknown function |
|-------------------|-------------------------------------|
| **Blastn on D. radiodurans:** DR_1392 should be longer |
| **Query 1** MTD S WQPAPAGFKHVVSVLGGSKRNAREEINVLGQPFVLERIGT | **Sbjct 1396664** MTD S WQPAPAGFKHVVSVLGGSKRNAREEINVLGQPFVLERIGT |

| Deide_04980 (17aa) | RarA | TSS at GTG; V in *D. geothermalis* is also GTG |
|-------------------|------|------------------------------------------|
| **Blastn on D. radiodurans:** DR_1392 should be longer |
| **Query 1** MTLFEPPAPLRLRPVFTEVQGQLLGG | **Sbjct 1396766** MTLFEPPAPLRLRPVFTEVQGQLLGG |

| Deide_08770 (3aa) | SsrA-binding protein [smpB] | TSS at GTG (others also GTG at V); start of E8U8U7_DEIML is also GTG |
|-------------------|-----------------|------------------------------------------|
| **Blastn on D. radiodurans:** DR_1392 should be longer |
| **Query 1** MIYTLNRARRAEYEHYELLERFAAGLSLTGSEVKSIRAGGVDFRDAFARLNG | **Sbjct 1396766** MIYTLNRARRAEYEHYELLERFAAGLSLTGSEVKSIRAGGVDFRDAFARLNG |

| Deide_09750 (17aa) | Deinococcus+Truepera-specific |
|-------------------|---------------------------|
| **Blastn on D. radiodurans:** DR_1392 should be longer |
| **Query 1** MLRRQSRPQGQLRPG | **Sbjct 1396766** MLRRQSRPQGQLRPG |
(Figure S5, continued)

**Deide_09860** (16aa) Skp (OmpH)

Deide_09860: 4SCFIRLRKADTVTVMNKAVPLAPLAVAAGFLGTVPASSQTPAQKGIVGDVQLLLS
Degeo_0715: 4KGFIRLRLKNRsVPAMNKAALAPLALVAAGFLGTVTAPAQTGPQKGIVDVGQLLS
DR_0989: 4TCFIRLRLKNASVAVMNKALAPLAVLAAGFLGTTAPAQTPAQKGIVGDVQLLS
Deima_1344: --------------------------MVNKQMLPVAVAAAFVATLPASPQKGIVVDVQLLLS
Deppr_1168: --------------------------MNKAAKVLPLASAVAAVATSAQTPAQKGIVVVDVQLLLS

**Deide_10430** (8aa) acetylglutamate kinase

Deide_10430: 8SYAKVRLIIIVKVGGSGAGIDYDAVCDLAAARKAGERLRLV
Degeo_0678: 8PALLFCYLLIVKVGGSGAGIDYDAVCDLAAARKAGERLRLV
DR_1420: 8LRSQDFCPTARRSFPLVCRIRVMIVKVGGSGAGIDYDAVCDLAAARKAGERLRLV
Deima_1346: 8MNQGDSFPTARRSFPLVCRIRVMIVKVGGSGAGIDYDAVCDLAAARKAGERLRLV
Trad_1399: 8MNQGDSFPTARRSFPLVCRIRVMIVKVGGSGAGIDYDAVCDLAAARKAGERLRLV
Mrub_2721: 8MNQGDSFPTARRSFPLVCRIRVMIVKVGGSGAGIDYDAVCDLAAARKAGERLRLV
Ocepr_1796: 8MNQGDSFPTARRSFPLVCRIRVMIVKVGGSGAGIDYDAVCDLAAARKAGERLRLV
Messil_0435: 8MNQGDSFPTARRSFPLVCRIRVMIVKVGGSGAGIDYDAVCDLAAARKAGERLRLV

**Deide_11030** (8aa) *Deinococcus*-specific

Deide_11030: 8MVIALLGVLVLVVFYILLSATLIALSFQPIKTAENIRVIRMPFAVQYQLALLLAL
Degeo_1349: 8MVIALLGVLVLVVFYILLSATLIALSFQPIKTAENIRVIRMPFAVQYQLALLLAL
Deima_2680: 8MVIALLGVLVLVVFYILLSATLIALSFQPIKTAENIRVIRMPFAVQYQLALLLAL
DR_1429: 8MVIALLGVLVLVVFYILLSATLIALSFQPIKTAENIRVIRMPFAVQYQLALLLAL
Deppr_0258: 8MVIALLGVLVLVVFYILLSATLIALSFQPIKTAENIRVIRMPFAVQYQLALLLAL

**Deide_14250** (3aa) RecF, TSS at GTG

Start codon in RecF from DEIGI, DEIPD, TRUR, MARHT, MEIRD and MEISD is GTG, and ** in RecF from DEIGI, DEIPD, DEIMP, DEIMEL is GTG; RecF start in OCEP & THET is ATG.

**Deide_16430** (12aa) (uncharacterized *Deinococcus*-specific)

Deide_16430: 12CPAASPAAYAANVTRHFSDTRNTGVRFLQLGVRVLAEVGESQWQH
Degeo_1997: 12PQDILLI110TAPLTSYPAYAERNVTRHFSDTRTGVRFLQLGVRVLAEVGPPWQLE
DR_0600: 12NVTRHFSDTRTGVRFLQLGVRVLAEVGPPWQLE
DGeo_CA0379: 12NSARYARGVTRHFSDTRTGVRFLQLGVRVLAEVGPPWQLE
Deppr_0112: 12MLQILPYAVWANVTRHFSDTRTGVRFLQLGVRVLAEVGPPWQLE
Deima_1143: 12MRFLDSRTDGRVPLIDASVLLARHTWAQH

**Acetylglutamate kinase**

**Deinococcus*-specific**

**RecF, TSS at GTG**

**Deinococcus*-specific**
Deide_19140 (19aa) aminomethyltransferase or folate-binding protein YgfZ
Deide_19140
DR_0538
Dgeo_Ca2386
Deima_0616
Deipr_1428

Deide_20630 (4aa) RuvC
tr|C1CYQ3|C1CYQ3_DEIDV
sp|Q1J1K4|RUVC_DEIGD
tr|H8GUN8|H8GUN8_DEIGI
sp|Q9RX75|RUVC_DEIRA
tr|E8U2Z2|E8U2Z2_DEIML
tr|F0RMA9|F0RMA9_DEIPM

Deide_22690 (4aa) Deinococcus-specific
Deide_22690
Dgeo_2325
DR_2558
Deima_0019
Deipr_0036

Longer D. deserti proteins:
Deide_04180 (8aa) ferredoxin; D. radiodurans probably longer, too.
Deide_04180
tr|F0RM49|F0RM49_DEIPM
tr|Q9RSP9|Q9RSP9_DEIRA
tr|E8U4J5|E8U4J5_DEIML
tr|F0RQR3|F0RQR3_DEIPM
tr|F0RR58|F0RR58_DEIPM

Deide_17480 (30aa) (uvrD/rep helicase)
Deide_17480
DR_0358
Dgeo_0358
Deima_0616
Deipr_1428

Deide_2p02060 (15aa) uvrA2; D. radiodurans homolog is probably shorter than annotated, and that of D. gobiensis longer.
Deide_2p02060
Deipr_2182
Deipe_1722
DGo_Ca0556
DR_A0188
Deima_0190

(tblastn on D. gobiensis (H8GW29_9DEIO):)
Query 1
Sbjct 491664

Deide_2p02060 (15aa) uvrA2; D. radiodurans homolog is probably shorter than annotated, and that of D. gobiensis longer.
Deide_2p02060
Deipr_2182
Deipe_1722
DGo_Ca0556
DR_A0188
Deima_0190
(Figure S5, continued)

**Deide_2p02180 (231aa) putative transcriptional regulator**

| Gene   | Sequence                                                                                           |
|--------|----------------------------------------------------------------------------------------------------|
| Deide_2p02180 | DTLPARLRGRLHEGLRVVAPVEVGAQPALRKRWAARASUVTLDLPGP=GARRWLPY                        |
| DGo_PB0322  | MVYLPARVRALAVTVPVVVPVEVGAQFDALLQWATKGYTVTRDLGD=ETGFLTWVP                      |
| Deima_0584  | MSLSPSALELLATAPVVAPVGGFAQTDLAEWAHKTRRILRPDAEFSLAPLTLIP                          |
| Mrub_0345   |                                                                                                   |
| Deide_2p02180 | ATHTDLRSLAG---EHYAILILSMDLSPLPDDEDWNAALSASGPQWRQTFAAFGWPAAM                   |
| DGo_PB0322  | RYSQEVAWAAGHAAKVLILSDGEHLDDLGAWEAAAG---DSGWGRFTAQSSGKPFAAL                   |
| Deima_0584  | QRSSDLRLNS--SDPRDFLFLRESDELLFHFDEWQQAVS-----TTQTYAEITGWPEAL                   |
| Mrub_0345   |                                                                                                   |
| Deide_2p02180 | ELLARLLAQGQTenLPVEELHRHPMSLVAQPYQSGSLRAAAVQALAAALVTPAVADG                     |
| DGo_PB0322  | DFRPRIQGRAAGD-----WAAMQQLAALAPLDP--DVEQNDYQLARTPLVTPFQAL                      |
| Deima_0584  | PLAAALADPATO-----LAAPLAPALLGCPPMPLRSAFERLAPPLTVPVVARL                       |
| Mrub_0345   | ALLQRIVLQPGQES------LVRHFLCVARLGSLLPKDQPREILKAAQGFLPLELYGL                   |
| Deide_2p02180 | LDVERHRHLETLSDEGWLWPSQGNAWFPELLRRTLAPVPDPRRAIRAAQALQAAGHPEAL                 |
| DGo_PB0322  | LGVDGAAALATLADGWLWPAQGVRFAPALLRRLLVAPALYNLSAQVAALASSAGHVGEAL                 |
| Deima_0584  | LGTASDVOARALSGGWLTPFPGWNAAPTLRLHLLAALAPQARTAQIAAIRALHDSAGHTDAAL             |
| Mrub_0345   | LGLDDTSVAEYDQLRDGQLQMSLAQKLLRLYLRSIPAEVARFIEITLALSGLHVTAVL                  |
| tblastn on D. gobiensis:                                                                                   |
| Query 1     | DTLAPARLRGRLHEGLRVVAPVEVGAQPALRKRWAARASUVTLDLPGP=GARRWLPY                        |
| Sbjct 259075 | SEVQRLRQRLESGNPGWAGIAAPVEVGAQGPELLAQAQGRPTRAAPVHGTRGCVW 258896           |
Figure S6. Detected homologs for peptides and proteins from 17 new leaderless transcripts. New gene labels are mentioned. The TSS for these 17 D. deserti genes is at the first nucleotide of the start codon. Results of BLASTP, and TBLASTN (if any), and alignments are shown. Non-annotated homologs in D. radiodurans, D. geothermalis and/or D. gobiensis were found for Deide_07364, Deide_14766, Deide_15148 and Deide_23068. Amino acid composition and domains found by SMART are shown for several proteins (horizontal red and pink bars represent signal peptide and low complexity region, respectively; vertical blue bar represents transmembrane helix).

**Deide_00694** conserved protein of unknown function (63aa)

MTDKGNEAEQMQEAYAQEQEQTGKTSAGGAGSTGTPGNQHTGETTEENDNGPRTGPTEN

Blastp: only one homolog (Deipe_2139)
Tblastn: no more hits

```
Deide_00694  MTDKGNEAEQMQEAYAQEQEQTGKTSAGGAGSTGTPGNQHTGETTEENDNGPRTGPTEN
Deipe_2139   MTDNKNEAEHEQMQEAYAQEQEQTGKTSAGGAGSTGTPGNQHTGETTEENDNGPRTGPTEN
```

Number of amino acids: 63
Molecular weight: 6570.7
Theoretical pI: 4.26
Amino acid composition:

| Amino Acid | Number | Percentage |
|------------|--------|------------|
| Ala (A)    | 6      | 9.5%       |
| Arg (R)    | 2      | 3.2%       |
| Asn (N)    | 5      | 7.9%       |
| Asp (D)    | 2      | 3.2%       |
| Cys (C)    | 0      | 0.0%       |
| Gln (Q)    | 6      | 9.5%       |
| Glu (E)    | 10     | 15.9%      |
| His (H)    | 1      | 1.6%       |
| Ile (I)    | 0      | 0.0%       |
| Leu (L)    | 0      | 0.0%       |
| Lys (K)    | 2      | 3.2%       |
| Met (M)    | 2      | 3.2%       |
| Phe (F)    | 0      | 0.0%       |
| Pro (P)    | 3      | 4.8%       |
| Ser (S)    | 3      | 4.8%       |
| Thr (T)    | 10     | 15.9%      |
| Trp (W)    | 0      | 0.0%       |
| Tyr (Y)    | 1      | 1.6%       |
| Val (V)    | 0      | 0.0%       |

**Deide_02488** conserved protein of unknown function (85aa)

MILILDGKYQQVQQNKRLTLAEAGHLPKGLQSDIDALHDCCAQHGRCDVQVNTQHGLMQTLVEKPKLFSLWQFEGHLSF

Blastp: homolog only of other new leaderless gene Deide_11736 and DGo_CA2102 (86aa)
Tblastn: no more hits

```
Deide_02488  MILILDGKYQQVQQNKRLTLAEAGHLPKGLQSDIDALHDCCAQHGRCDVQVNTQHGLMQ
Deide_11736  MILILDGKYQQVQQNKRLTLAEAGHLPKGLQSDIDALHDCCAQHGRCDVQVNTQHGLMQ
DGo_CA2102   MILILDGKYQQVQQNKRLTLAEAGHLPKGLQSDIDALHDCCAQHGRCDVQVNTQHGLMQ
```

Number of amino acids: 85
Molecular weight: 9311.8
Theoretical pI: 4.76
Amino acid composition:

| Amino Acid | Number | Percentage |
|------------|--------|------------|
| Ala (A)    | 6      | 7.1%       |
| Arg (R)    | 2      | 2.4%       |
| Asn (N)    | 5      | 5.9%       |
| Asp (D)    | 2      | 2.4%       |
| Cys (C)    | 0      | 0.0%       |
| Gln (Q)    | 6      | 7.1%       |
| Glu (E)    | 10     | 12.0%      |
| His (H)    | 1      | 1.2%       |
| Ile (I)    | 0      | 0.0%       |
| Leu (L)    | 0      | 0.0%       |
| Lys (K)    | 2      | 2.4%       |
| Met (M)    | 2      | 2.4%       |
| Phe (F)    | 0      | 0.0%       |
| Pro (P)    | 3      | 3.5%       |
| Ser (S)    | 3      | 3.5%       |
| Thr (T)    | 10     | 12.0%      |
| Trp (W)    | 0      | 0.0%       |
| Tyr (Y)    | 1      | 1.2%       |
| Val (V)    | 0      | 0.0%       |
**Deide_11736** conserved protein of unknown function (86aa)
MIILQGTYQVAPTKRLLTILESHGQGKGLATIDALSGKCAQGGKCD1TVTTQHGMPTGTYEKKPRKLSWQFEGHLSFPQRS

Blastp: homolog only of other new leaderless gene Deide_02488 and DGo_CA2102 (86aa)
Tblastn: not more hits

| Deide_02488 | MIILQGTYQVAPTKRLLTILESHGQGKGLATIDALSGKCAQGGKCD1TVTTQHGMPTGTYEKKPRKLSWQFEGHLSFPQRS |
| Deide_11736 | MIILQGTYQVAPTKRLLTILESHGQGKGLATIDALSGKCAQGGKCD1TVTTQHGMPTGTYEKKPRKLSWQFEGHLSFPQRS |
| DGo_CA2102  | MIILQGTYQVAPTKRLLTILESHGQGKGLATIDALSGKCAQGGKCD1TVTTQHGMPTGTYEKKPRKLSWQFEGHLSFPQRS |

**Deide_04802** conserved protein of unknown function (36aa)
MEFLLAGLTIVGLLASIQRHPQ-GRVSVSRKRG

Blastp: only DGo_CA2147
Tblastn: nothing

| Deide_04802 | MEFLLAGLTIVGLLASIQRHPQ-GRVSVSRKRG |
| DGo_CA2147  | MELLLAALATLTALLLASRQAAPRKYARVFVPRHHSRR |

**Deide_11672** protein of unknown function, partial (25aa) (pseudogene together with Deide_11671 and Deide_11670; DegV family protein)
MIAVTLTDSDFSEAPARRGHSTS

Blastp: N-terminal fragment of DegV family protein

| Deide_11672 | MIAVTLTDSDFSEAPARRGHSTS-------------------------------------------- |
| Deipr_0962  | MIAVTLTDSDFSEAPARRGHSTS-------------------------------------------- |
| LJ_1180     | MIAVTLTDSDFSEAPARRGHSTS-------------------------------------------- |
| PF01_00648  | MIAVTLTDSDFSEAPARRGHSTS-------------------------------------------- |
| Deide_12040 | MIAVTLTDSDFSEAPARRGHSTS-------------------------------------------- |
| Deima_2005  | MIAVTLTDSDFSEAPARRGHSTS-------------------------------------------- |

**Deide_1p00954** conserved protein of unknown function (74aa)
MASSLSAVATHVLEFLQGQEHKPRSADELALLQRDRAEVENRRAELEELQAAGLVPEAVSGYGGNDTVWSVTHS

The protein (74aa) has some homology with HTH_11 domain (a Pfam domain. Position: 9 to 63, E-value: 4.9e-05)
Blastp: one good homolog of similar length
Tblastn: not more good hits

| Deide_1p00954 | MASSLSAVATHVLEFLQGQEHKPRSADELALLQRDRAEVENRRAELEELQAAGLVPEAVSGYGGNDTVWSVTHS |
| Deipr_2258    | MASSLSAVATHVLEFLQGQEHKPRSADELALLQRDRAEVENRRAELEELQAAGLVPEAVSGYGGNDTVWSVTHS |

**Deide_1p00954** conserved protein of unknown function (86aa)
MIILQGTYQVAPTKRLLTILESHGQGKGLATIDALSGKCAQGGKCD1TVTTQHGMPTGTYEKKPRKLSWQFEGHLSFPQRS

Blastp: homolog only of other new leaderless gene Deide_02488 and DGo_CA2102 (86aa)
Tblastn: not more hits

| Deide_02488 | MIILQGTYQVAPTKRLLTILESHGQGKGLATIDALSGKCAQGGKCD1TVTTQHGMPTGTYEKKPRKLSWQFEGHLSFPQRS |
| Deide_11736 | MIILQGTYQVAPTKRLLTILESHGQGKGLATIDALSGKCAQGGKCD1TVTTQHGMPTGTYEKKPRKLSWQFEGHLSFPQRS |
| DGo_CA2102  | MIILQGTYQVAPTKRLLTILESHGQGKGLATIDALSGKCAQGGKCD1TVTTQHGMPTGTYEKKPRKLSWQFEGHLSFPQRS |

---
Deide_04426 conserved exported protein of unknown function (58aa)
MKRTIPLLIAALLLASCDDGAE TEDTDSTSTTTTTSTDQEDTTDTQSTSTTTTEEEK

Blastp: several hits (lipoprotein signal peptide followed by T-rich region
is found in various proteins), but only DR_1317 and DGo_CA2846 of similar
size.

Blastn: not clearly more hits

Deide_04426
MKRTIPLLIAALLLASCDDGAE TEDTDSTSTTTTTSTDQ
DR_1317
MRCSRSHARSLAADFNGA KRAVLA
DGo_CA2846
MLDTEKR---LPLLAALLLAGCSNQGSG-TSSSTTTTKFDSSG

**:    :* *.. ..  *.:**:**.: ..:

Deide_04426
EDTTDTQSTSTTTTEEEK
DR_1317
QNTGTTTTTTTTTTTNK-
DGo_CA2846
QPAGTSSTTTTESNK---

Probable correct start of DR_1317 and DGo_CA2846 are in green.

The D at +2 after the cysteine indicates attachment of the mature protein to the inner
membrane.

LPAM_[pfam08139], Prokaryotic membrane lipoprotein lipid attachment site; In prokaryotes,
membrane lipoproteins are synthesized with a precursor signal peptide, which is cleaved by a
specific lipoprotein signal peptidase (signal peptidase II). The peptidase recognizes a
conserved sequence and cuts upstream of a cysteine residue to which a glyceride-fatty acid
lipid is attached.

ProtParam of the mature protein starting with C:
Number of amino acids: 42
Molecular weight: 4455.3
Theoretical pI: 3.40
Amino acid composition:

|   | Ala (A) | Arg (R) | Asn (N) | Asp (D) | Cys (C) | Glu (E) | His (H) | Ile (I) | Leu (L) | Lys (K) | Met (M) | Pro (P) | Ser (S) | Thr (T) | Tyr (Y) |
|---|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|
|   | 2.4%    | 0.0%    | 0.0%    | 14.3%   | 2.4%    | 4.8%    | 0.0%    | 0.0%    | 0.0%    | 2.4%    | 0.0%    | 0.0%    | 11.9%   | 45.2%   | 0.0%    |
**Deide_14223** conserved protein of unknown function (56aa)
MRIDDHNLNEQAQMGGATIEQARRMRELLERKPRATEDFTGKWAEAVLLEATR

Blastp: only one good hit of similar size (DR_0413)
Tblastn: not more hits

| Deide_14223 | MRIDDHNLNEAQHMGATIEQARRMRELLERKPRATEDFTGKWAEAVLLEATR |
|-------------|----------------------------------------------------------|
| DR_0413     | MKKREWKEKPERACATGKEKDRMDLQELATQGSGEDTDAAALRDLGLTTQGEREDFTS |

Deide_14223 ---
DR_0413    ---

Possible correct start of DR_0413 in **green**

---

**Deide_07364** conserved protein of unknown function (36aa)
MREFLNDWRLLGKLTAATLAIPVVLWGLLVLLGILR

Blastp: three *Deinococcus* homologs ([^ in Deima_1393 = GTG])
Tblastn: indicates homologs also in other *Deinococcus*, see below

| Deide_07364 | MREFLNDWRLLGKLTAATLAIPVVLWGLLVLLGILR |
|-------------|-------------------------------------|
| Deipr_1862  | MPEWLNDWRLKLTVSLAVFVPLLWALLWNVAGVLK |
| Deima_1393  | MKVPWRTFAARGPCYAEEKRFNDWRLKLTVLSAVPVLVYLLLWNVGLK |
| Deipe_3104  | MREFWYWFRLKTFGVALLVPLLWLLMLAVGLLRL |

**tblastn:**

*Deinococcus gobiensis* I-0, complete genome
Features:
- 87 bp at 5' side: Phosphoglucomutase, alpha-D-glucose phosphate-specific
- 194 bp at 3' side: putative Fructose-bisphosphatase

Query 1
MREFLNDWRLLGKLTAATLAIPVVLWGLLVLLGILR 36
MREFLNDWRL KLTAATLAIPVVLWGLLVLLGILR
Sbjct 2750143 MREFLNDWRLLKTLTVSLAVFVPLLWALLWNVAGVLK 2750250

*Deinococcus geothermalis* DSM 11300, complete genome
Features:
- 115 bp at 5' side: phosphoglucomutase, alpha-D-glucose phosphate-specific
- 187 bp at 3' side: fructose-1,6-bisphosphatase, class II

Query 1
MREFLNDWRLLGKLTAATLAIPVVLWGLLVLLGILR 17
MREFLNDWRL GKLTAATLAIPVVLWGLLVLLGILR
Sbjct 1973393 MREFLNDWRLLGKLIA 1973443

*Deinococcus radiodurans* R1 chromosome 1, complete sequence
Features:
- hypothetical protein (= DR_2037 on opposite strand)

Query 1
MREFLNDWRLLGKLTAATLAIPVVLWGLLVLLGILR 36
MREFLNDWRL KLTVSLAVFVPLLWALLWNVAGVLK
Sbjct 2053590 MRELLTLWDRIFKLVAGICGPLLINGALVNVGVLK 2053483

Further analysis reveals the following entire non-annotated proteins:

* Dgob: MREFLNDWRLLGKLTAATLAIPVVLWGLLVLLGILR
* Dgeo: MREFLNDWRLLGKLIAVGLALPLLWGLLVAGVILH
* Drad: MRELLTLWDRIFKLVAGICGPLLINGALVNVGVLK

| Deide_07364 | MREFLNDWRLLGKLTAATLAIPVVLWGLLVLLGILR |
|-------------|-------------------------------------|
| Deipr_1862  | MPEWLNDWRLKLTVSLAVFVPLLWALLWNVAGVLK |
| Deima_1393  | MKVPWRTFAARGPCYAEEKRFNDWRLKLTVLSAVPVLVYLLLWNVGLK |
| Deipe_3104  | MREFWYWFRLKTFGVALLVPLLWLLMLAVGLLRL |
| Dgob        | MREFLNDWRLLGKLIAVGLALPLLWGLLVAGVILH |
| Dgeo        | MRELLTLWDRIFKLVAGICGPLLINGALVNVGVLK |

---

(Figure S6, continued)
Deide_12656 conserved exported protein of unknown function (70aa)
Protein detected by proteomics (Figure S4).

Amino acid composition:

| Amino Acid | Number |
|------------|--------|
| Ala        | 3      |
| Asn        | 5      |
| Cys        | 0      |
| Glu        | 5      |
| His        | 1      |
| Leu        | 6      |
| Met        | 1      |
| Pro        | 0      |
| Thr        | 9      |
| Tyr        | 0      |

Theoretical pI: 5.65

Molecular weight: 5270.8

with some more homologs:

Deipe_0721
Deima_0510
FJSC11DRAFT_0576
Ava_2326
Glo7428_3587
Nos7107_2275
Deipr_2033
DR_1067
Dgeo_1211
Deide_12656

Blastp: many hits (note conserved C-terminal region, which is present in many more homologs of similar size)

Deide_12656
Deipe_0721
Deima_0510
FJSC11DRAFT_0576
Ava_2326
Glo7428_3587
Nos7107_2275
Deipr_2033
DR_1067
Dgeo_1211
Deide_12656

with some more homologs:

Deide_12656
Deipe_0721
Deima_0510
FJSC11DRAFT_0576
Ava_2326
Glo7428_3587
Nos7107_2275
Deipr_2033
DR_1067
Dgeo_1211
Deide_12656

ProtParam of mature protein starting with Q:

Number of amino acids: 48
Molecular weight: 5270.8
Theoretical pI: 5.65

Amino acid composition:

| Amino Acid | Number |
|------------|--------|
| Ala (A)    | 3      |
| Asn (N)    | 5      |
| Cys (C)    | 0      |
| Glu (E)    | 5      |
| His (H)    | 1      |
| Leu (L)    | 6      |
| Met (M)    | 1      |
| Pro (P)    | 0      |
| Thr (T)    | 9      |
| Tyr (Y)    | 0      |

Arg (R) 5 10.4%
Asp (D) 1 2.1%
Gln (G) 1 2.1%
Gly (G) 7 14.6%
Ile (I) 1 2.1%
Lys (K) 0 0.0%
Phe (F) 0 0.0%
Ser (S) 0 2.1%
Val (V) 0 2.4%

SP='YES' Cleavage site between pos. 22 and 23: AGA-QE D=0.848 D-
cutoff=0.510 Networks=SignalP-TM.

(Figure S6, continued)
Deide_14766 conserved protein of unknown function (34aa)
MKGLGEFIEWLREVLKGASQPQPQPVPVRVRQR

Blastp: homology with Deipe_0001 and Deima_1530
Tblastn: also homology with other Deinococcus, see below

Deide_14766  MKGLGEFIEWLREVLKGASQPQPQPVPVRVRQR
Deima_1530   MSAPYNEAKMAGEDFLQKREILRAQT
Deipe_0001   MTLRSEKRLILVMDDLKKLSALRHEERLLQ
syc0626_d    MGLVDQILSRLQQDLARKLEALFQPEAQEPFPVPVPVRQR

Tblastn:
Deinococcus radiodurans R1 chromosome 1, complete sequence
Features:
111 bp at 5' side: GTP pyrophosphokinase
45 bp at 3' side: peptidyl-prolyl cis-trans isomerase, FKBP-type
Query 1   MKGLGEFIEWLREVLKGASQPQPQPV  26
Sbjct 1866952  VKGLSEFLEWRLGELGEPRPQVIPVRTR

Deinococcus gobiensis I-0, complete genome
Features:
147 bp at 5' side: Peptidylprolyl isomerase FKBP-type
228 bp at 3' side: ppGpp synthetase I, SpoT/ReiA
Query 1   MKGLGEFIEWLREVLKGASQPQPQPV  29
Sbjct 1110944  MKGLREFIDWLRETLQGAQPQPVPVP

Deinococcus geotherma DSM 11300, complete genome
Features:
242 bp at 5' side: sigma 54 modulation protein/ribosomal protein S30EA
354 bp at 3' side: ABC transporter related protein
Query 1   MKGLGEFIEWLREVLKGASQPQPQPV  26
Sbjct 1466316  MKGLREIFDLKRELKAGSASQPPVFP

Further analysis reveals the entire non-annotated proteins, included in the multiple alignment:

Deima_1530   MSAPYNEAKMAGEDFLQKREILRA-GTTPKPALVPVPVRTRQPR
Drad    MKGLGEFIEWLREVLKGLQVLTG-LGEPROQPVPVPVRTRERR-
Deide_14766  MKGLGEFIEWLREVLKGASQPQPQPO-QPVPVRVRQR
Dgob    MKGLREFIDWLRETLQQAQPQPFP--FPVPVRVRDRE-
Deipe_0001  MTLRSEKRLILVMDLKKLSALRHEERLLQ--APQQPVPVPVPVRRRR---
syc0626_d    MGLVDQIFSLQDLARKLEALFQPEAQEPFPVPVPVRQR---
Dgeo    MKGLREIFDLKRELKAGSASQPQPVPVPVRNRDRE---

(Figure S6, continued)
Deide_15148 conserved protein of unknown function (91aa)
MSGFSGGFSGFHRSGHRGGMVGGLLGSSSHSGRRHGVPQGGRQAQRRRSGGSLGAFLVT
AGLAGAVMGVLVSLIA

Blastp: only two good homologs
Tblastn: also homology in *D. geothermalis*

| ID            | Description                       |
|---------------|-----------------------------------|
| Deide_15148   | MSGFSGGFSGFHRSGHRGGMVGGLLGSSSHSGRRHGVPQGGRQAQRRRSGGSLGAFLVT |
| DGo_CA2041    | MSGFSGGFSGFHRSGHRGGMVGGLLGSSSHSGRRHGVPQGGRQAQRRRSGGSLGAFLVT |
| Deima_0925    | MSGFSGGFSGFHRSGHRGGMVGGLLGSSSHSGRRHGVPQGGRQAQRRRSGGSLGAFLVT |
| Deide_15148   | MSGFSGGFSGFHRSGHRGGMVGGLLGSSSHSGRRHGVPQGGRQAQRRRSGGSLGAFLVT |
| DGo_CA2041    | MSGFSGGFSGFHRSGHRGGMVGGLLGSSSHSGRRHGVPQGGRQAQRRRSGGSLGAFLVT |
| Deima_0925    | MSGFSGGFSGFHRSGHRGGMVGGLLGSSSHSGRRHGVPQGGRQAQRRRSGGSLGAFLVT |

**tblastn:**
Deinococcus geothermalis DSM 11300, complete genome

| Features | Description |
|----------|-------------|
| aminoglycoside phosphotransferase |

| Query | SBjct |
|-------|-------|
| Deide_15148 | MSGFSGGFSGFHRSGHRGGMVGGLLGSSSHSGRRHGVPQGGRQAQRRRSGGSLGAFLVT |
| DGo_CA2041 | MSGFSGGFSGFHRSGHRGGMVGGLLGSSSHSGRRHGVPQGGRQAQRRRSGGSLGAFLVT |
| Deima_0925 | MSGFSGGFSGFHRSGHRGGMVGGLLGSSSHSGRRHGVPQGGRQAQRRRSGGSLGAFLVT |

Further analysis reveals the following non-annotated Dgeo protein:

| Dgeo     | Description |
|----------|-------------|
| MSGFSGGFSGFHRSGHRGGMVGGLLGSSSHSGRRHGVPQGGRQAQRRRSGGSLGAFLVT |

Amino acid composition of the cytoplasmic domain:

For the cytoplasmic domain only:
Number of amino acids: 67
Molecular weight: 7031.6
Theoretical pI: 12.37

Amino acid composition:

| AA     | Count | Percentage |
|--------|-------|------------|
| Ala (A) | 2     | 3.0%       |
| Arg (R) | 10    | 14.9%      |
| Asn (N) | 0     | 0.0%       |
| Asp (D) | 0     | 0.0%       |
| Cys (C) | 0     | 0.0%       |
| Gln (Q) | 2     | 3.0%       |
| Glu (E) | 0     | 0.0%       |
| Gly (G) | 19    | 28.4%      |
| His (H) | 7     | 10.4%      |
| Ile (I) | 0     | 0.0%       |
| Leu (L) | 2     | 3.0%       |
| Lys (K) | 1     | 1.5%       |
| Met (M) | 2     | 3.0%       |
| Phe (F) | 5     | 7.5%       |
| Pro (P) | 0     | 0.0%       |
| Ser (S) | 13    | 19.4%      |
| Thr (T) | 0     | 0.0%       |
| Trp (W) | 0     | 0.0%       |
| Tyr (Y) | 2     | 3.0%       |
**Deide_19985** conserved protein of unknown function (67aa)

MDLDSWTPDNNARRLTLATAVGVTFTVALWLGLASHALLGLVLGAVGLVGVVVFIARRLVSFWRR

**Blastp:** only several homologs (35-50%) in *Deinococcus*

- **Deide_19985**
- **Deipr_0048**
- **DGo_PC0138**
- **Deide_3p01320**
- **DR_1299**
- **Deipe_2825**

- **Deide_20865** protein of unknown function, partial (72aa)

VPPATPGEMPSSHREYARLARELGYYRPWTRVLGSGDPELTFDFQKAQCHRKWTEVGVVKTAPombokQHLI

**Blastp:** hits with larger proteins

- **Deide_20865**
- **DR_0468**
- **DGo_CA2682**
- **Deima_1024**
- **B14911_25565**

---

(Figure S6, continued)
Deide_23068 conserved protein of unknown function (61aa)
MTDDQKKPQGHDPAEQSPAEGQSHAIPDAAQGKNPGVDPAAKGAPAEGGRDEVEGSSTPGQ

Blastp: several “good” (50%) homologs in Deinococcus only
Tblastn: more hits, see below

Deide_23068  ---------------------------------------------MTDDQKKP---------------------------------------------
Dgeo_2289   MKHPMNSMTPTSTRKAPLGSHPKSGPGRGGEVLSHSSSGLQDRPQVRSESPRSDDL
Deima_0533  ---------------------------------------------MTDETKQQ---------------------------------------------
Deipe_1333  . . . . . . .

Deide_23068  ---------------------------------------------QGHDPAEQSPAEG---------------------------------------------
Dgeo_2289   LPLNKPRESSKLQGRACRQRAN[EDDSKRFYDPANTAPAEGQSHIPPQQQNPAYTDPA
Deima_0533  ---------------------------------------------NLQPDADQEQAE---------------------------------------------
Deipe_1333  : : : : *: :*

Deide_23068  -----------------------------QSHAIPDAAQGKNPGVDPAAKGAPAEGGRDEVEGSSTPGQ--------------------
Dgeo_2289   LPLNKPRESSKLQGRACRQRAN[EDDSKRFYDPANTAPAEGQSHIPPQQQNPAYTDPA
Deima_0533  ---------------------------------------------NLQPDADQEQAE---------------------------------------------
Deipe_1333  : : : : *: :*

** tblastn:**

Deinococcus gobiensis I-0, complete genome
Features: hypothetical protein
Query 1      MTDDQKKPQGHDPAEQSPAEGQSHAIPDAAQGKNPGVDPAAKGAPAEGGRDEVEGSSTPGQ  60
M+DD KK  G+DPA QSPAEGQSHAIP +GK+P +DPAK PAEGGRDEVEGSSTPGQ
Sbjct 43089  MSDDPKK--GYDPANSTPASQRPPEEEDRKGAPNADPDPAAKDEPAEGGRDEVEGSSTPGQ  43262

Deinococcus radiodurans R1 chromosome 1, complete sequence
Features: 21 bp at 5’ side: endonuclease III
84 bp at 3’ side: conserved hypothetical protein
Query 1      MTDDQKKPQGHDPAEQSPAEGQSHAIPDAAQGKNPGVDPAAKGAPAEGGRDEVE  54
M+DD KK  G+DPA QSPAEGQSHAIP +GK+P +DPAK PAEGGRDEVEGSSTPGQ
Sbjct 2439859 MSDDPKK--GYDPANSTPASQRPPEEEDRKGAPNADPDPAAKDEPAEGGRDEVEGSSTPGQ  2439704

Further analysis reveals the following entire non-annotated proteins:
Dgob MSDDPKKGYDPANSTPASQRPPEEEDRKGAPNADPDPAAKDEPAEGGRDEVEGSSTPGQ
Drad MSDDPKKGYDPANSTPASQRPPEEEDRKGAPNADPDPAAKDEPAEGGRDEVEGSSTPGQ

(Figure S6, continued)
(Figure S6, continued)

**Deide_2p00483** conserved protein of unknown function (49aa)

MTKKKTGTTSPRVAKASELLSNPKSAAAVKVSAASALANADAPKQKK

Upstream of Deide_2p00480 (integrase).
Blastp: several homologs of similar size.
Blastn: no more hits.
Looks bit like N-terminus of HU (also histones among blastp hits).
In some others, gene adjacent to phage-associated genes.

Deide_2p00483

| MTKKKTGTTSPRVAKASELLSNPKSAAAVKVSAASALANADAPKQKK |

Blastp: several homologs of similar size.
Blastn: no more hits.
Looks bit like N-terminus of HU (also histones among blastp hits).
In some others, gene adjacent to phage-associated genes.

Molecular weight: 5066.9
Theoretical pI: 10.69
Amino acid composition:

| Amino Acid | Count | Percentage |
|------------|-------|------------|
| Ala (A)    | 10    | 20.4%      |
| Arg (R)    | 1     | 2.0%       |
| Asn (N)    | 2     | 4.1%       |
| Asp (D)    | 1     | 2.0%       |
| Cys (C)    | 0     | 0.0%       |
| Glu (E)    | 1     | 2.0%       |
| Gly (G)    | 1     | 2.0%       |
| His (H)    | 0     | 0.0%       |
| Ile (I)    | 0     | 0.0%       |
| Leu (L)    | 3     | 6.1%       |
| Lys (K)    | 12    | 24.5%      |
| Met (M)    | 1     | 2.0%       |
| Phe (F)    | 0     | 0.0%       |
| Pro (P)    | 3     | 6.1%       |
| Ser (S)    | 6     | 12.2%      |
| Thr (T)    | 4     | 8.2%       |
| Trp (W)    | 0     | 0.0%       |
| Tyr (Y)    | 0     | 0.0%       |
| Val (V)    | 3     | 6.1%       |

**Deide_2p01755** conserved membrane protein of unknown function (92aa)

MNPVREWNNWKSAGAWLLGALLLUVLVLVYQLSGTHLEAYQVELSLISMILMVLYATDRTFVLRGTYRMALGNAFFCTVALMLQARSLLMMVRS

Blastp: only one homolog
Blastn: no other hits

Deide_2p01755

| MNPVREWNNWKSAGAWLLGALLLUVLVLVYQLSGTHLEAYQVELSLISMILMVLYATDRTFVLRGTYRMALGNAFFCTVALMLQARSLLMMVRS |

Deide_15680

| MNPVQEDKNGAAGWALLLAVVLVYQSEAYLEYQVFISVTSTILLAMHRVTFLW |

Deide_2p01755

| RQGDYRMLGNAFFCTVALMLQARSLLMMVRS |

Deide_15680

| RQGDHRMALISAGIMAVVLLRAFSSLVMMRY |

Number of amino acids: 49
**Figure S7. New proteins detected by proteomics.** New gene labels are mentioned. Tryptic or chymotryptic peptides are indicated. Transcription start site, when found, is mentioned. Results of BLASTP and TBLASTN (if any) and alignments are shown.

**Deide_05864** conserved protein of unknown function (76aa)
MLTVKMHLAGGDIALNMTPSQKRNLSTKINTQAALPLPFTANVDGVDEVIEIPWRISYISSYPQVSSPVLREAAM

Detected peptides: TINQAQLPTLPFTANVDGVDEVIEIPWR, MHLAGGDIALNMTPSQK, TANVDGVDEVIEIP.
Transcription start at -6 of start codon.
Blastp: only one homolog, Dgeo_2254
Tblastn: two additional homologs in *D. deserti*: new predicted Deide_05654 (no identified TSS), unpredicted Deide_11206 (TSS at -6 as for Deide_05864).

Deide_05864 MHLTVKMHLAGGDIALNMTPSQKRNLSTKINTQAALPLPFTANVDGVDEVIEIPWRISYISSYPQVSSPVLREAAM
Deide_11206 MLTVKMHLAGGDIALNMTPSQKRNLSTKINTQAALPLPFTANVDGVDEVIEIPWRISYISSYPQVSSPVLREAAM
Deide_05654 MHLTVKMHLAGGDIALNMTPSQKRNLSTKINTQAALPLPFTANVDGVDEVIEIPWRISYISSYPQVSSPVLREAAM

**Deide_13059** protein of unknown function (90aa)
MPILVVTISKGGHRNYGQEVQGYIVDQYIQFSSGGPNILEFNNSDPNQPSQDISMNITSMVHPDDLQSVP EPITDAVHDNAMDKK

Detected peptides: SSSGPNILEF, RGSQEVLQUERY, RGSQEVLQUERY.
Transcription start at 1st nt of start codon.
Blastp: no hits
Tblastn: no hits

**Deide_1p00482** protein of unknown function (69aa)
MDHGVQYPYVEALQERGCVTOHPDGRYSVTLPDGETIEPGAPSVHPQTPWSLIEACRSLNVTVPFG

Detected peptides: LNVTPFGE, SVTLPDGETIEPGAPSVHPQTPW, VEALQERGCVT.
Transcription start at 1st nt of start codon.
Blastp & tblastn: nothing

**Deide_2p01542** protein of unknown function (113aa)
MERPDEVNGLTFNAQRDHGFRHVEAGFMPMLPVPFALLQADRSAPEETTRSSLLKQLRLKTTAPEAFSKET SGFLTTATFMTNVSEDEYFRNRLTFLVEAYRKHATSD

Detected peptides: MTNVSEDEYFRNRL, ALLQADRSAPEETTRSSLL
Transcription start at 1st nt of start codon.
Blastp & tblastn: nothing
Deide_15253 conserved exported protein of unknown function (158aa)
MKTLVLALLASAPVAYQDTTDTTPTETATETTDDVTGETTGTDTGTDTGTDAGTDATDTMGTDTETDADATGTTDTDADATATETETEATDATDAATTTSSGTIVETENRSGFWGLGGLGLAGLACGNRTHAHTTTTQTDTRRP

Detected peptides: SGFWGLGGLGLAGLACGR, MKTL

Blastp: hits with DR_2344 (78/167 = 46%) and Deima_0987 (81/186 = 43%)
(DR_2344 probably too long: start at MKK.. gives predicted signal peptide)

Tblastn: nothing

Deide_15253

Deima_0987

Detected peptides:

Deide_12656 conserved exported protein of unknown function (70aa)
MTKLKLASSAVLAIPNVAGAQETNTTTIIMENNERTDGMGLGLGGLAGLACGRHVETSTVRR

Detected peptide: GTDMGWGLAGLGLAGLACGR

Transcription start at 1st nt of start codon.

Blastp: several hits (note conserved Cter corresponding to detected peptide, which is present in many more homologs of similar size)
**Deide_3p02615** (Deide_23165) conserved protein of unknown function (78aa)

MREFNSVTAGFDGLAVPGRIELEGGRGLMRVSLNGAPDISEGAEAILMHDGVRFRVAYTERLDDTNEVRMKLLARS

Detected peptides: **VAVTERLDDTNEVR**, MREFNSVTAGFDGLAVPGRIELEGGR, FGDLAVPGRIEAL, FGDLAVPGRIELEGGRGLMRVSLNGAPDISEGAEAILMHDGVRFRVAV

Blastp: two homologs, Deima_0424 & Deipe_2626

Tblastn: one almost identical protein (76/78 = 97%) in *D. deserti* itself (= new Deide_23165); RNAseq indicates better expression of **Deide_23165** than **Deide_3p02615**.

---

**Deide_11207** protein of unknown function (70aa)

MDISQVVRATAHHLFKLYWAMFANIENPEEALASAGQA VVLLYLDCCGMPAQEAAMLRDEIMLSIPTRKM

Detected peptide: **ANIENPEEALASAGQAVLL**

Transcription start at 1st nt of start codon.

Blastp & tblastn: no good hits

---

**Deide_14224** conserved protein of unknown function (90aa)

MTAGLTPQARVVALDGALDGALIMHRTERGAFYTLGRRLSVTLLKDLERLRYVSRASAGARTAVAYELTPGGS AALAQWEGSNPASRG

Detected peptide: **TAVAYELTPGSSAALAQWEGSNPASRG**

Blastp: two homologs, Deima_2324 & Deipe_1712

Tblastn: no other good hits

---

**Deide_3p02814** protein of unknown function (142aa)

LTRVPEKTLRRTLRLGRGHRHPFRKLQ1LIIIQQGSALSAQLQTLAIRALTANLEQRVVLQAISSDNLSVQIKANLS DLKVVAAVQLGLLAKQKQIFPSSQHGRHRSCTPENKRIEHALDEFRVRDGLVLTLQSKATPLGN

Detected peptide: **IQQGSALSAQLQTL**

Blastp & tblastn: no good hits
In figure below, Deide_04940 to Deide_04960 are on reverse strand, ORF C_571026_3 is on the forward strand. Deide_04940 has a potential internal TSS (iTSS) and two potential antisense TSS (aTSS) (one with only few reads). The three peptides in C_571026_3 are downstream of these two aTSS. However, antisense reads mainly found with +TEX, without TEX hardly any reads.

Detected peptides: AGAKVTGSEVVGL, GSRVAGHHEAEVL, QVSQSLSDLAL
(Figure S7, continued)
Figure S8. Examples of several highly induced genes encoding small proteins. Multiple alignments with homologs are shown. For 5 of these *D. deserti* proteins, potential non-annotated homologs in *D. geothermalis* and/or *D. radiodurans* are also indicated. All proteins are of unknown function. Conserved cysteine residues in some of the proteins are indicated.

**Deide_04721** conserved protein of unknown function (74aa)

**Blastp & tblastn: no hits with D. radiodurans & D. gobiensis.**

**Deide_04721**

**Dgeo_0305**

**Deima_3010**

**Deipe_1056**

**Deipr_1808**

with **SSDG_06207** = Predicted protein (100aa) of *Streptomyces pristinaespiralis*.

**Deide_05260** conserved protein of unknown function (62aa)

**Tblastn: additional homolog in *D. geothermalis* (see below)**

**Deide_05260**

**Dgeo_CAI723**

**Deipe_3565**

**Deima_2451**

**Deipr_2498**

**DUF1540 (pfam07561)** (Cd Length: 40 Bit Score: 35:27 E-value: 7.15e-05)

**Tblastn on *D. geothermalis***

123 bp at 5’ side: hypothetical protein (Dgeo_0866 = Deide_05250 homolog)

105 bp at 3’ side: cation diffusion facilitator family transporter (Dgeo_0867)

Query 1

**MNDTEHQSMVGRCDATNCRFNDMECTAQIIEVQMSQMAQCITYPTDGMGESY**

**Sbjct**

925654

**MEDRSHQOQQASIVGRCDATRHNENQEBAQIEVLSQMAQCITYPTQEGM**

925827
Among the homologs are new *D. deserti* genes identified in this study: Deide_3p00225 and Deide_1p00514.

Tblastn: additional homolog in *D. geothermalis* (see below).

With 3 best blastp hits only:

**Deide_19965** conserved protein of unknown function (63aa)

MQELACTNWPGTDLIRVLRKVGSSTIELTSTLRARIFGPQALNLDLYLKGRAVVKADARQVAML.

Among the homologs are new *D. deserti* genes identified in this study: Deide_3p00225 and Deide_1p00514.

Tblastn: additional homolog in *D. geothermalis* (see below).

**Deide_19965**

**DGo_CA2814**

**Deide_3p00225**

**Mrub_0267**

**Delma_1438**

**Delma_1206**

**Deipe_2935**

**Deipe_3286**

**Deipe_1p00514**

**Deipe_0580**

**Deipe_1840**

**DGo_CA2814**

**Deide_3p00225**

**Mrub_0267**

**Deima_1438**

**Deima_1206**

**Deipe_2935**

**Deipe_3286**

**Deide_1p00514**

**Deipe_0580**

**Deipe_1840**

**Deide_1p00514**

**Deipe_0580**

**Deipe_1840**

**With 3 best blastp hits only:**

**Deide_19965**

**DGo_CA2814**

**Mrub_0267**

**Delma_1438**

**Delma_1206**

**Deipe_2935**

**Deipe_3286**

**Deipe_1p00514**

**Deipe_0580**

**Deipe_1840**

**Tblastn on *D. geothermalis***:

179 bp at 5’ side: a/b hydrolase superfamily protease and regulatory beta pr... = Dgeo_0365

87 bp at 3’ side: peptidase M20 = Dgeo_0366 (Syntheny with *D. deserti*)

**Query 1**

**Sbjct 372429**

**Query 61**

**Sbjct 372609**
Deide_2p00980 conserved protein of unknown function (64aa)
MTNERGSGSGAGGRDPNGDDKTNNGLGDGRDPGSNHGSPDDREGDGRNGSESGGGKSQTKD

Tblastn indicates additional homolog in D. geothermalis (see below)

Deide_2p00980
DR_A0234
DR_A0234
DR_A0234
DR_A0234

with possible homolog YciG of E. coli:

with GsiB of Bacillus (Glucose starvation-inducible protein B, General stress protein B)
(Induction: Glucose or phosphate starvation, and addition of decoyinine. Also by heat shock, salt stress and oxidative stress):

for E. coli YciG and B. subtilis GsiB:

Family: KGG (PF10685)
Stress-induced bacterial acidophilic repeat motif.
This repeat is found in proteins which are expressed under conditions of stress in bacteria. The repeat contains a highly conserved, characteristic sequence motif, KGG, that is also recognised by plants and lower eukaryotes and repeated in their LEA (late embryogenesis abundant) family of proteins, thereby rendering those proteins bacteriostatic. An example of such an LEA family is LEA-5, PF00477. Further downstream from this motif is a Walker A, nucleotide binding, motif GXXXXGK(S,T), that in YciG of E coli, eg Q8X7B4 is QSGGNKSGKS [URL]. YciG is expressed as part of a three-gene operon, yciGFE, and this operon is induced by stress and is regulated by RpoS, which controls the general stress-response in E coli. YciG was shown to be important for stationary-phase resistance to thermal stress and in particular to acid stress.

tblastn on D. geothermalis:
88 bp at 5' side: UspA
247 bp at 3' side: Rhodanese-like protein
Query  21

D+ N G D
D+ GDGRNGSESGGGKS
3160923
Deide_09148 protein of unknown function (30aa)
MTRPTARQLAMATULLLTLLGGALGRLL

Putative CDS directly downstream of ddrA.
Potential, non-annotated homologs are present downstream of ddrA in D. radiodurans and D. geothermalis.

Deide_09148 ----MTRPTAR-----------------------------QLQLAMATULLLTLLGGALGRLL
Drrd MTRSLTSAELRGGAAPSTDPVMPARVSPARLPDTHLWGAMVNLGLLTLLGGALGRLL
Dgeo ----MSPATPR-----------------------------QLAAVNLGVLTLTLLGGALAKL

# WEBSEQUENCE Length: 30
# WEBSEQUENCE Number of predicted TMHs: 1
# WEBSEQUENCE Exp number of AAs in TMHs: 18.12642
# WEBSEQUENCE Exp number, first 60 AAs: 18.12642
# WEBSEQUENCE Total prob of N-in: 0.65291
# WEBSEQUENCE POSSIBLE N-term signal sequence
WEBSEQUENCE TMHMM2.0 outside 1 9
WEBSEQUENCE TMHMM2.0 TMhelix 10 29
WEBSEQUENCE TMHMM2.0 inside 30 30

Deide_11446 protein of unknown function (57aa)
MAAAQRQDTSMHFIELAQQRQRDLQEAQQRRLIREAQAQRKRRKFRFPSLLGHRLA

Putative CDS, or an ncRNA? Located downstream of, and in opposite direction (convergent) of uvrC.
Weak homology with DGo_CA1576 (55aa) (similarly present downstream and opposite strand of uvrC): Identities = 20/41 (48%), Positives = 25/41 (60%)

Deide_11446 VAAARQDTSMHFIELAQQRQRDLQEAQQRRLIREAQAQRKRRKFRFPSLLGHRLA
DGo_CA1576 --------MNPTFHEFELTFAAQAADHRQEAARDQVARAAQPTFLRTPTRWFTRTRLHPKA
DGo_CA1576 --------MNPTFHEFELTFAAQAADHRQEAARDQVARAAQPTFLRTPTRWFTRTRLHPKA
DGo_CA1576 --------MNPTFHEFELTFAAQAADHRQEAARDQVARAAQPTFLRTPTRWFTRTRLHPKA
ACPL_1607 --------MYP--EINLSLARQGEQLRQEAAYRRARA--AGGRPRBRRSSVRKPVSPR--
Mesil_2337 --------MPHFMEMAKLARERAREAIQOEANQRRLLQEAIGELRPPFHRLAFWLAQLAQRL
ACPL_1607 and Mesil_2337 are not next to uvrC

Deide_11446
DGo_CA1576
ACPL_1607
Mesil_2337
DAEVNLKLIKPSR

(Figure S8, continued)
Deide_20580  conserved hypothetical protein (83aa)
MRA0TIAESRIGFVHTTVNNTLVQEVENGLLAVRRIERQLHLGTVNLDHRDHPLTQTWLASAARAYLITQAERRQAV

Located upstream and divergent of ddrQ (these two genes share the palindromic RDRM). Similar gene pair in others, except D. radiodurans.

Deide_20580  --------------MRALDTIAESRIGFVHTTVNNTLVQEVENGLLAVRRIERQLHLGTVNLDHRDHPLTQTWLASAARAYLITQAERQAV
Dgeo_0335  --------------MRALDIASIRGVYHTTVNNTLVQEVENGLLAVRRIERQLHLGTVNLDHRDHPLTQTWLASAARAYLITQAERQAV
Deipe_1112  --------------MRALTIANTIKGVYHTTVNNTLVQEVENGLLAVRRIERQLHLGTVNLDHRDHPLTQTWLASAARAYLITQAERQAV
DGo_CA0309  --------------MRALTIASIRGVYHTTVNNTLVQEVENGLLAVRRIERQLHLGTVNLDHRDHPLTQTWLASAARAYLITQAERQAV
Deima_0559  --------------MHTLQRTLIEAETIRQAHQTVHTTVNNTLVQEVENGLLAVRRIERQLHLGTVNLDHRDHPLTQTWLASAARAYLITQAERQAV
Deipr_0092  --------------MQALTRVQTLARTQAHQTVHTTVNNTLVQEVENGLLAVRRIERQLHLGTVNLDHRDHPLTQTWLASAARAYLITQAERQAV
Marky_0724  --------------MMNLRIAHRLKHGRVHPNEALNLIEFDNRAGLHELLEALEATLHRQR
Mrub_2304  --------------MNVKIELIRSVQGAAAPSEVSNALIELDRQKQQLNALSEELRHR
Mesil_2926  --------------MRRNPWERSLGRGQHPQSSAAPSEVSNALIELDRQKQQLNALSEELRHRPAA
Ocepr_1741  --------------MRRQALFRTQ----QALPAAMLLNLIEVDNRAGLHELLEAELAARLQAG

| Deide_20580 | 100% |
| Degeo_0335 | 78% |
| Deipe_1112  | 63% |
| DGo_CA0309 | 63% |
| Deima_0559  | 56% |
| Deipr_0092  | 48% |
| Marky_0724  | 38% |
| Mrub_2304   | 33% |
| Mesil_2926  | 32% |

Deide_07900  conserved protein of unknown function (63aa)
MSNDKNQPAQSADAPQGKDQGEGIKQVQDGQMEEGKRQVDQTPQDVTPQTELDGGAQP

Only homologs in two Deinococcus.
Tblastn indicates homolog in D. radiodurans (see below)

Deide_07900  MSNDKNQPAQSADAPQGKDQGEGIKQVQDGQMEEGKRQVDQTPQDVTPQTELDGGAQP
Deipr_1284  MTQDISAPQGKDQGEGIKQVQDGQMEEGKRQVDQTPQDVTPQTELDGGAQP
DGo_CA1816  MNDKNQPAQSADAPQGKDQGEGIKQVQDGQMEEGKRQVDQTPQDVTPQTELDGGAQP

| Deide_07900 | ** | ** |
| Deipr_1284  | ** | ** |
| DGo_CA1816   | ** | ** |

Tblastn gives also hit with D. radiodurans, but no others

Features:
76 bp at 5’ side: conserved hypothetical protein
68 bp at 3’ side: hypothetical protein

Query 1
MSNDKNQPAQSADAPQGKDQGEGIKQVQDGQMEEGKRQVDQTPQDVTPQTELDGGAQP 51
MS+D K  P +PGKDQGEGIKQVQDGQMEEGKRQVDQTPQDVTPQTELDGGAQP

Sbjct 1917919  MSDKAARMPAARSPEGSDDTLDLSQGEGIKQVQDGQMEEGKRQVDQTPQDVTPQTELDGGAQP 1917764
Deide_13590 conserved protein of unknown function (77aa)
MSDKSTAENMLDAAAAKVNETADRAREAGHNVAHAVTGDAHKKAELEDRGKAELHNRANAEFHEGKHEATDGDGH

Deinococcus-specific

Deide_13590
Dgeo_1167
DGo_Ca1692
DR_1539
Deipr_0475
Deipe_3116
Deide_13821

without Deipe_3116 and Deide_13821:

Deide_13590
Dgeo_1167
DGo_Ca1692
DR_1539
Deipr_0475
Deide_13590
Dgeo_1167
DGo_Ca1692
DR_1539
Deipr_0475
Figure S9. TSS positions relative to RDRM (radiation-desiccation response motif). For the different radiation-induced genes, the arrows indicate the TSS position relative to the 17-bp RDRM. Either gene names or gene numbers (without “Deide_”) are shown.

* = 2x RDRM (ddrA)  
** = 2x TSS (gyrA)

in red: leaderless mRNA

RDRM

20 bp  17 bp  50 bp