

Correspondence

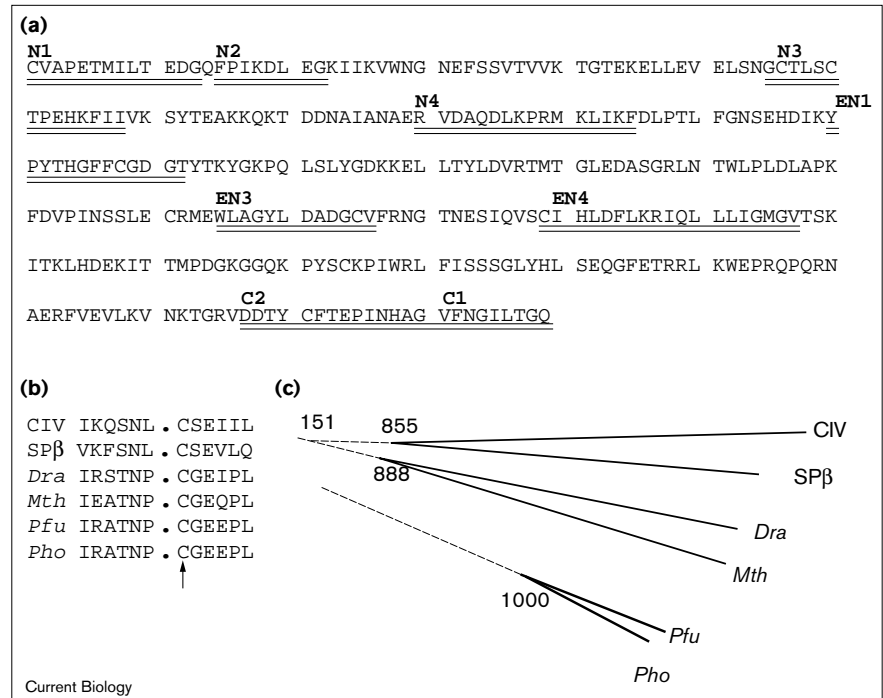
Identification of a virus intein and a possible variation in the protein-splicing reaction Shmuel Pietrokovski

Inteins are protein ‘introns’ found inside other proteins. They remove themselves from their host proteins post-translationally in an autoproteolytic protein-splicing reaction [1]. Besides their protein-splicing activity, many inteins also have endonucleolytic activity believed to mediate the homing of their genes to corresponding unoccupied integration sites. Both functions benefit the inteins: protein splicing averts deleterious effects caused by their insertion into the protein hosts and homing disseminates inteins by horizontal gene transfer. Inteins therefore appear to be selfish genetic elements [2].

Organisms from all three kingdoms of life — eucarya, bacteria and archaea — contain inteins, but their distribution is sporadic. The presence of inteins in evolutionarily distant organisms and their variable appearance in related organisms is generally explained by horizontal transfer [3,4]. The extent of intein horizontal transfer has health implications: *Mycobacterium tuberculosis*, *Mycobacterium leprae* and *Candida tropicalis* are major human pathogens that contain inteins [3]. Here, I describe the first identified insect virus intein. On the basis of the sequence relation of the viral intein and its protein host to other proteins, I propose viruses as vehicles of intein dispersion. A new type of residue in the intein’s carboxy-terminal end suggests a variation of the protein-splicing mechanism.

The ribonucleotide reductase (RNR) large subunit of the *Chilo*

Figure 1



The CIV ribonucleotide reductase (RNR) intein. **(a)** Sequence of the intein found in positions 272–610 of the CIV RNR (NCBI database accession number 2738400; shown in the single-letter amino-acid code). Conserved intein sequence motifs (see [9,10]) are double-underlined and denoted as N (amino-terminal), EN (endonuclease) or C (carboxy-terminal) domains as in [15] (EN1 and EN3 are the LAGLIDADG motifs). The expectant value for finding the multiple conserved intein regions is 9.8×10^{-9} [16]. **(b)** RNR intein integration points in CIV (CIV_RIR1 intein), SPβ (SPB_RIR1), *Dra* (Dra_RIR1 from *Deinococcus radiodurans*), *Mth* (Mth_RIR1, *Methanobacterium thermoautotrophicum*), *Pfu* (Pfu_RIR1-2, *Pyrococcus furiosus*) and *Pho* (Pho_RIR1, *Pyrococcus horikoshii* OT3). Intein names

are from the Intein Database [9] and their accession numbers are given in the Supplementary material. Intein positions are marked by bullets and the active site thyl radical by an arrow. **(c)** Relation between RNR inteins. Part of an intein dendrogram showing the same RNR inteins as in (b). Numbers are bootstrap values from 1000 trials. Sequence motifs of protein-splicing and endonuclease domains – see (a) – were used for calculating the dendrogram. Gaps were used for the endonuclease motifs of the *Mth* RIR1 intein, which does not have this domain. The dendrogram was constructed with the neighbor-joining and bootstrapping procedures of the CLUSTAL W program [17]. Similar results were obtained with the PROTDIST and SEQBOOT programs of the PHYLIP package [18].

iridescent virus (CIV) contains an intein. CIV is an insect-infecting iridovirus [5]; members of this double-stranded DNA virus family are known to infect invertebrates, amphibia and fish [6]. All the expected protein-splicing motifs are found in the intein together with the typical dodecapeptide LAGLIDADG homing-endonuclease domain (Figure 1a). The glutamine at the carboxy-terminal intein end, where only

asparagine has been observed before, indicates a variation on the previously described protein-splicing mechanism (see below).

RNRs catalyze the essential reaction producing deoxyribonucleotides from ribonucleotides. Three classes of RNRs have been described, distinguished by their sequence, subunit composition and cofactors. Nevertheless, enzymatic mechanism, allosteric control and subtle sequence

similarities suggest that all RNR classes are derived from one common ancestor [7]. The CIV RNR is a class I enzyme and the intein is integrated directly amino-terminal to the thiol radical active site (Figure 1b) [8]. Remarkably, other distantly related RNRs also have inteins in this position: the class II RNRs of the archaea *Pyrococcus furiosus*, *Pyrococcus horikoshii* OT3 and *Methanobacterium thermoautotrophicum* and of the bacterium *Deinococcus radiodurans* and the class I RNR of the bacteriophage SP β (Figure 1b; see [9,10]).

Viruses are potential horizontal transfer vectors of inteins across species. Analyses of class I and II RNR enzymes and of RNR intein sequences support this hypothesis. Most RNR enzyme sequences cluster according to RNR class type and taxonomy. Intein-containing RNR sequences are not particularly similar to each other, except for the two *Pyrococcus* sequences (see Supplementary material published with this paper on the internet). SP β and CIV RNR inteins are probably homologous as they are significantly related to each other — clustering together within all other known intein sequences (Figure 1c). This incongruence between the sequence relation among the viral RNR inteins and among their hosts is consistent with horizontal transfer of these inteins. The same situation is also found for the *M. thermoautotrophicum* and *D. radiodurans* RNR inteins. The details of the actual physical encounter necessary for these horizontal transfer events are unknown. *Chlorella* viruses have been suggested to mediate the horizontal transfer of group I introns between algae and protists [11] and various viruses are known to integrate foreign DNA from their host and other co-infecting viruses [12,13].

Glutamine is the carboxy-terminal residue of the CIV intein and of the *P. horikoshii* polC intein. Only asparagine has previously been described at this position [3]. This asparagine plays a key role in the

protein-splicing reaction: it undergoes cyclization, releasing the amino-terminally cleaved intein from the ligated host flanks in the penultimate step of the reaction [1]. I suggest that these two inteins, integrated in highly conserved protein motifs, are active and splice using a variation of the previously described protein-splicing reaction in which their carboxy-terminal glutamines undergo cyclization analogous to that of the intein carboxy-terminal asparagine forming a glutarimide ring. Such a cyclization is chemically plausible and has been predicted [14]. These two inteins probably evolved their peculiar carboxy-terminal ends independently, as they are not particularly similar by sequence and are also different in the highly conserved penultimate residue (see Supplementary material).

Eucaryotic inteins have been found only in single-cell organisms and chloroplast genomes [3]. Conspicuously, these invasive parasitic elements are missing from the nuclear genomes of multicellular organisms. Nevertheless, the presence of inteins in animal pathogens, particularly ones that are intracellular, such as the virus reported here, and mycobacteria, provides an opportunity for inteins to invade metazoan genomes. To date, this has not yet been observed.

Supplementary material

An illustration of the possible variation in the protein-splicing reaction and a dendrogram of RNR enzymes are available as Supplementary material published with this article on the internet.

Acknowledgements

I thank E.A. Greene, P. Talbert, C. McCallum, K. Ahmad, L. Comai and S. Henikoff for commenting on the manuscript. The author is a Howard Hughes Medical Institute Fellow of the Life Sciences Research Foundation. This work was supported by an NIH grant (GM29009).

References

- Perler FB: **Protein splicing of inteins and hedgehog autoproteolysis: structure, function, and evolution.** *Cell* 1998, **92**:1-4.
- Belfort M, Reaban ME, Coetzee T, Dalgaard JZ: **Prokaryotic introns and inteins: a panoply of form and function.** *J Bacteriol* 1995, **177**:3897-3903.
- Perler FB, Olsen GJ, Adam E: **Compilation and analysis of intein sequences.** *Nucleic Acids Res* 1997, **25**:1087-1093.
- Derbyshire V, Belfort M: **Lightning strikes twice — intron-intein coincidence.** *Proc Natl Acad Sci USA* 1998, **95**:1356-1357.
- Bahr U, Tidona CA, Darai G: **The DNA sequence of Chilo iridescent virus between the genome coordinates 0.101 and 0.391; similarities in coding strategy between insect and vertebrate iridoviruses.** *Virus Genes* 1997, **15**:235-245.
- Williams T: **The iridoviruses.** *Adv Virus Res* 1996, **46**:345-412.
- Reichard P: **From RNA to DNA, why so many ribonucleotide reductases?** *Science* 1993, **260**:1773-1777.
- Stubbe J: **Ribonucleotide reductases in the twenty-first century.** *Proc Natl Acad Sci USA* 1998, **95**:2723-2724.
- The New England Biolabs Intein Database. http://www.neb.com/neb/inteins/intein_intro.html
- Shmuel Pietrokovski: **Inteins — protein introns.** <http://www.blocks.fhcrc.org/~pietro/inteins/>
- Yamada T, Tamura K, Aimi T, Songsri P: **Self-splicing group I introns in eukaryotic viruses.** *Nucleic Acids Res* 1994, **22**:2532-2537.
- Isfort RJ, Witter R, Kung HJ: **Retrovirus insertion into herpesviruses.** *Trends Microbiol* 1994, **2**:174-177.
- Jehle JA, Nickel A, Vlak JM, Backhaus H: **Horizontal escape of the novel Tc1-like lepidopteran transposon TcP3.2 into cydia pomonella granulovirus.** *J Mol Evol* 1998, **46**:215-224.
- Geiger T, Clarke S: **Deamidation, isomerization, and racemization at asparaginyl and aspartyl residues in peptides. Succinimide-linked reactions that contribute to protein degradation.** *J Biol Chem* 1987, **262**:785-794.
- Pietrokovski S: **Modular organization of inteins and C-terminal autocatalytic domains.** *Protein Sci* 1998, **7**:64-71.
- Henikoff S, Henikoff JG: **Protein family classification based on searching a database of blocks.** *Genomics* 1994, **19**:97-107.
- Thompson JD, Higgins DG, Gibson TJ: **CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.** *Nucleic Acids Res* 1994, **22**:4673-4680.
- Felsenstein J: **Phylogenies from molecular sequences: inference and reliability.** *Annu Rev Genet* 1988, **22**:521-565.

Address: Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue, North Seattle, WA 98109, USA.

E-mail: pietro@sparky.fhcrc.org

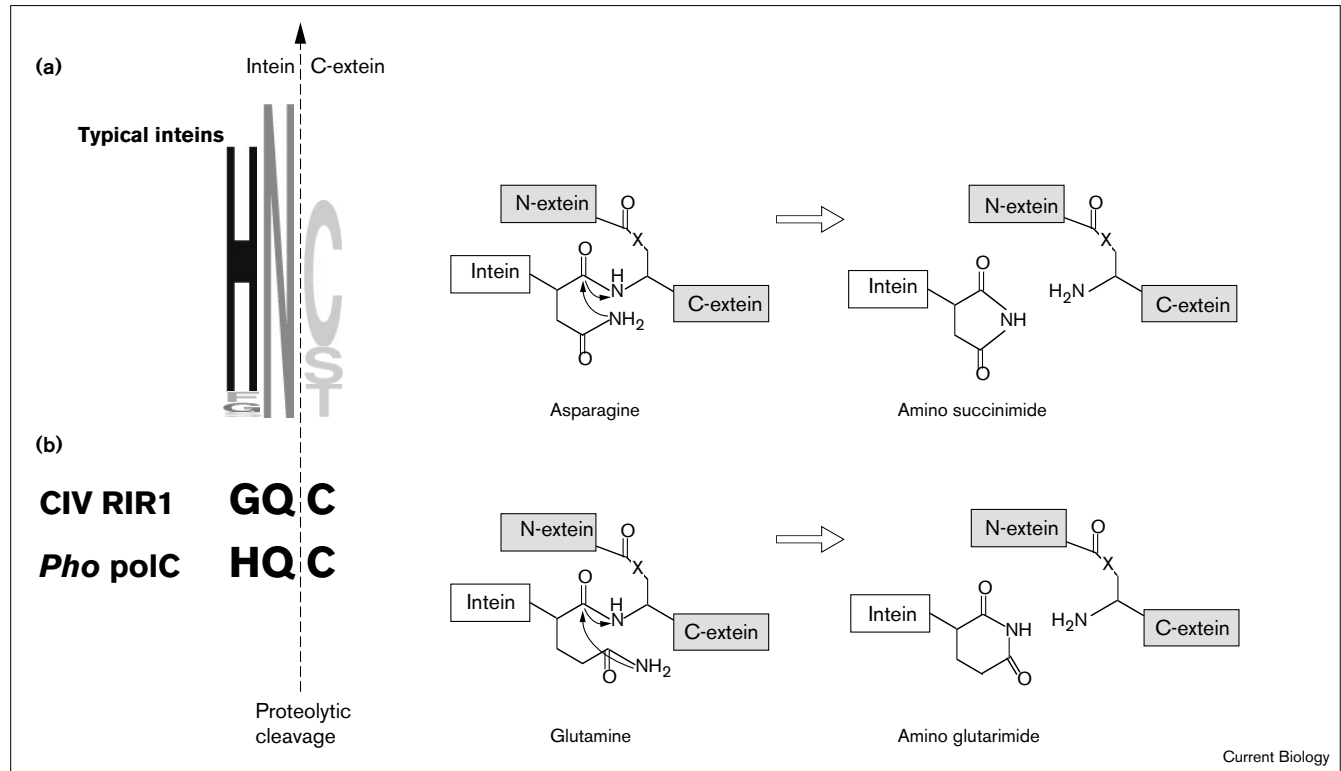
The editors of *Current Biology* welcome correspondence on any article in the journal, but reserve the right to reduce the length of any letter to be published. All Correspondence containing data or scientific argument will be refereed.

Identification of a virus intein and a possible variation in the protein-splicing reaction

Shmuel Pietrokovski

Current Biology 10 September 1998, 8:R634–R635

Figure S1



The carboxy-terminal end of inteins and its cleavage from the ligated host protein. **(a)** The sequences found in the carboxy-terminal end of a typical intein (left) and a scheme for the release of the semi-cleaved intein from the ligated host flanks (exteins) by asparagine cyclization (right) [1]. The sequences are represented by a logo, where the height of each residue is related to its frequency. In the scheme, X denotes

the sidechain of the residue carboxy-terminal to the asparagine: a sulfur atom (residue is cysteine) or an oxygen atom (residue is serine or threonine). **(b)** The carboxy-terminal ends of the two known inteins with glutamine at the carboxyl terminus and the mechanism proposed for their release from the ligated host flanks.

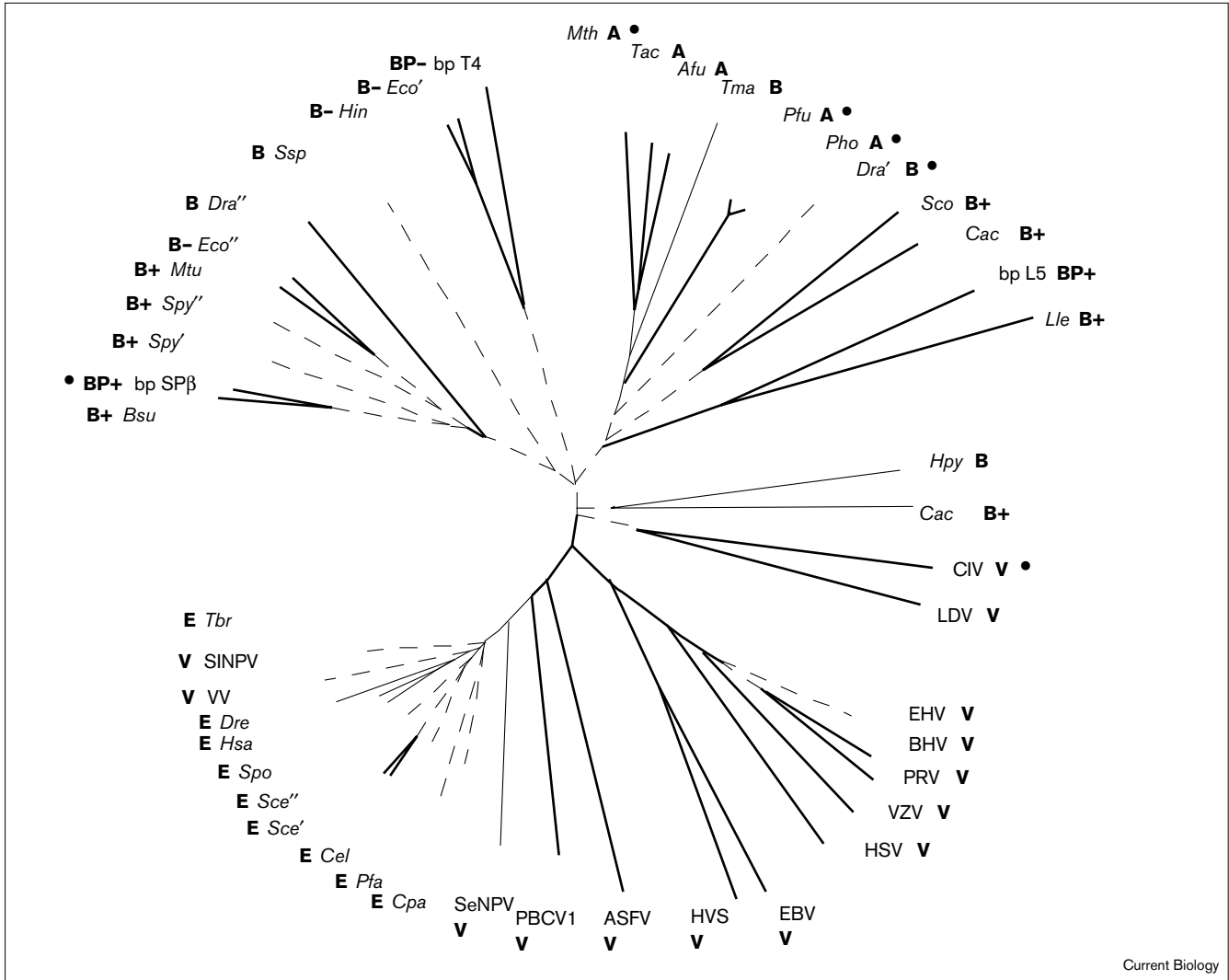


Figure S2

Relation between RNR proteins. A dendrogram of class I and II RNR enzymes. Species names and database accession codes are detailed below. When more than one RNR sequence was found in the same species, one sequence is marked with a single prime (') and the other with a double prime (''). In bold are general taxonomic abbreviations: A, archaea; B, bacteria; BP, bacteriophage; +, gram positive; -, gram negative; V, eucaryotic viruses; E, eucarya. Bullets mark intein-containing proteins. Eight sequence motifs totaling 208 amino acids and found in every sequence were used for calculating this dendrogram. Branches in bold have bootstrap values > 900/1000, solid branches have bootstrap values 800/1000 to 900/1000, and dashed branches have bootstrap values < 800/1000. The dendrogram was constructed with the neighbor-joining and bootstrapping procedures of the CLUSTAL W program [17]. Similar results were obtained with the PROTDIST and SEQBOOT programs of the PHYLIP package [18]. Sequence accession numbers are from the NCBI non-redundant database (gi) [NCBI *Entrez* Protein Sequence Search: <http://www.ncbi.nlm.nih.gov/Entrez/protein.html>], TIGR [The Institute for Genomic Research Microbial Database: <http://www.tigr.org/tdb/mdb/mdb.html>], GTC [Genome Therapeutics Corporation Gene Sequences: <http://web.cric.com/genesequences/clostridium/clospage.html>], and OUACGT [University of Oklahoma Advanced Center for Genome Technology: <http://www.genome.ou.edu/strep.html>]. Species names and sequence database accession numbers are given by order of appearance in the dendrogram clockwise from top – for species with two sequences, the first accession number is for the single prime (') names: *Mth*, *Methanobacterium thermoautotrophicum* (gi|2621734); *Tac*, *Thermoplasma acidophilum* (gi|1657993); *Afu*, *Archaeoglobus fulgidus* (gi|2648891); *Tma*, *Thermotoga maritima* (gi|2440216); *Pfu*, *Pyrococcus furiosus* (gi|1688292); *Pho*, *Pyrococcus horikoshii* OT3 (gi|3130612); *Dra*, *Deinococcus radiodurans* (TIGR|gdr_149, TIGR|gdr_9); *Sco*, *Streptomyces coelicolor* (gi|2995313); *Cac*, *Clostridium acetobutylicum* (GTC|Contig_477, 4977262_C2_24, GTC|Contig_479, 32616510_F3_12); bp L5, mycobacteriophage L5 (gi|465265); *Lle*, *Lactobacillus leichmannii* (gi|308869); *Hpy*, *Helicobacter pylori* (gi|2500207); CIV, *Chilo* iridescent virus (gi|2738400); LDV, fish lymphocystis disease virus (gi|2276414); EHV, equine herpesvirus type 4 (gi|1710389); BHV, bovine herpesvirus type 1 (gi|1710388); PRV, pseudorabies virus, strain Kaplan (gi|1710393); VZV, Varicella-zoster virus, strain Dumas (gi|132612); HSV, herpes simplex virus, type 1 (gi|132605); EBV, Epstein-Barr virus, strain B95-8 (gi|132602); HVS, herpesvirus saimiri, strain 11 (gi|266914); ASFV, African swine fever virus, strain BA71V (gi|1172937); PBCV1, *Paramecium bursaria Chlorella* virus 1 (gi|2447101); SeNPV, *Spodoptera exigua* nuclear polyhedrosis virus (gi|2052233); *Cpa*, *Cryptosporidium parvum* (gi|2828691); *Pfa*, *Plasmodium falciparum*, isolate FCR-3 (gi|1710392); *Cel*, *Caenorhabditis elegans* (gi|417657); *Sce*, *Saccharomyces cerevisiae* (gi|730514, gi|730516); *Spo*, *Schizosaccharomyces pombe* (gi|1350600); *Hsa*, *Homo sapiens* (gi|132608); *Dre*, *Danio rerio* (gi|1695829); VV, vaccinia virus, strain WR (gi|132611); SINPV, *Spodoptera littoralis* nuclear polyhedrosis virus (gi|2244677); *Tbr*, *Trypanosoma brucei* (gi|2411477); *Bsu*, *Bacillus subtilis* (gi|1710384); bp SPβ, bacteriophage SPβ (gi|2634398); *Spy*, *Streptococcus pyogenes* (OUACGT|Contig286, OUACGT|Contig248); *Mtu*, *Mycobacterium tuberculosis* (gi|1710390); *Eco*, *Escherichia coli* (gi|2507304, gi|2507305); *Ssp*, *Synechocystis species* strain PCC6803 (gi|1653420); *Hin*, *Haemophilus influenzae* (gi|1172938); bp T4, bacteriophage T4 (gi|417656).
