



**HAL**  
open science

# Viruses are essential agents within the roots and stem of the tree of life

Luis P. Villarreal, Guenther Witzany

► **To cite this version:**

Luis P. Villarreal, Guenther Witzany. Viruses are essential agents within the roots and stem of the tree of life. *Journal of Theoretical Biology*, 2010, 262 (4), pp.698. 10.1016/j.jtbi.2009.10.014. hal-00564083

**HAL Id: hal-00564083**

**<https://hal.science/hal-00564083>**

Submitted on 8 Feb 2011

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

## Author's Accepted Manuscript

Viruses are essential agents within the roots and stem of the tree of life

Luis P. Villarreal, Guenther Witzany

PII: S0022-5193(09)00489-5  
DOI: doi:10.1016/j.jtbi.2009.10.014  
Reference: YJTBI5742

To appear in: *Journal of Theoretical Biology*

Received date: 29 May 2009  
Revised date: 28 September 2009  
Accepted date: 8 October 2009

Cite this article as: Luis P. Villarreal and Guenther Witzany, Viruses are essential agents within the roots and stem of the tree of life, *Journal of Theoretical Biology*, doi:10.1016/j.jtbi.2009.10.014

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting galley proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.



[www.elsevier.com/locate/jtbi](http://www.elsevier.com/locate/jtbi)

# 1 Viruses are essential agents within the roots and stem of the tree of life

2  
3  
4 Luis P. Villarreal<sup>1</sup> and Guenther Witzany<sup>2</sup>

5  
6 <sup>1</sup>Department of Molecular Biology and Biochemistry, University of California, Irvine, California 92697

7 <sup>2</sup>Telos – Philosophische Praxis,

8 correspondence: Email: [witzany@sbg.at](mailto:witzany@sbg.at)

9 Tel./Fax: ++43 6274 6805

10 Vogelsangstraße 18c,

11 A-5111-Buermoos,

12 Austria

## 13 14 15 16 Abstract

17  
18 In contrast with former definitions of life limited to membrane-bound cellular life forms  
19 which feed, grow, metabolise and replicate (i) a role of viruses as genetic symbionts, (ii)  
20 along with peripheral phenomena such as cryptobiosis and (iii) the horizontal nature of  
21 genetic information acquisition and processing broaden our view of the tree of life. Some  
22 researchers insist on the traditional textbook conviction of what is part of the community of  
23 life. In a recent review\* they assemble four main arguments which should exclude viruses  
24 from the tree of life because of their inability to self-sustain and self-replicate, their  
25 polyphyly, the cellular origin of their cell-like genes and the volatility of their genomes. In  
26 this article we will show that these features are not coherent with current knowledge about  
27 viruses but that viral agents play key roles within the roots and stem of the tree of life.

28  
29 **Keywords:** viral interactions, RNA world agents, eukaryotic nucleus, biocommunication

30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  

---

\* Moreira D, Lopez-Garcia P (2009) Ten reasons to exclude viruses from the tree of life. Nat. Rev. Microbiol. 7, 306-311.

## Introduction

Within the Darwin project of the European Space Agency (ESA) four or five spacecraft will search for planets similar to earth around other stars and analyse their atmospheres for the chemical signature of life. Its search for extraterrestrial life raise the question of an appropriate definition of life, because it could be expected to find certain indicators on other planets which differ from terrestrial life forms. Current debate on common characteristics of terrestrial living agents assemble traditional textbook convictions such as that life includes those cellular life forms which feed, grow, metabolise and reproduce, i.e. membrane-bound life forms which divide cytoplasmatic space from outer environments. All of these concepts share the opinion that life emerged by interacting non-living chemical components (Greener, 2008).

A useful definition of life must also coherently explain (i) the peripheral phenomena of life, such as cryptobiosis, where the metabolic activity is barely discernible, (ii) the contrasting concept to the selfish gene hypothesis, i.e. the well documented phenomena of symbiogenesis, (iii) the role of non-lytic but persistent virus life-strategies which serve as main regulatory elements in all cellular life forms such as mobile genetic elements and noncoding RNAs and last but not least (iv) the current knowledge of genetic information processing.

In a recent article, Moreira and Lopez Garcia defend the traditional organism based tree and outline their opinion that there are ten reasons to exclude viruses from this tree of life model. At the end of their article they reduce these to four relevant arguments. 'Taken together their inability to self-sustain and self-replicate, their polyphyly, the cellular origin of their cell-like genes and the volatility of their genomes through time make it impossible to incorporate viruses into the tree of life' (Moreira and Lopez Garcia, 2009: 311).

As we will see, there are several arguments which contradict this opinion. Viruses are not metabolising cells, true enough. But all the functions within living cells such as replication, transcription and repair as well as their fine-tuned regulatory order are now known to also be of viral origin (Tang *et al*, 1997; Villarreal, 2005). Therefore it must be considered that viruses have played and still play crucial evolutionary roles and are essential agents within the tree of life. Are the essential agents within the roots of a tree not part of a tree? What would remain of the tree of life if we would substract all viral properties? We will try to show that at least three of the four arguments of Moreira and Lopez-Garcia in particular indicate viruses to be essential parts of the tree of life.

In a first step we will look at prokaryotic genomes and their gene word order determined by their colonizing viral agents. The second step exemplifies the identification of viruses and viral lineages. The third point will argue that the cell first perspective has to outline as good arguments as the virus first perspective. In a fourth point we will see that there is a lot of virus-virus interaction. A fifth chapter will demonstrate that the predecessors of cells had to be polyphyletic. At the sixth point we will see that the ancient and the current RNA-world is a main force in life. A seventh point will show that more than prokaryotes eukaryotic cells also depend on evolutionary roles of viruses, and this offers solutions to (i) the origin of the eukaryotic nucleus, (ii) the role of mobile genetic elements and (iii) the role of non-coding RNAs, i.e. key features with viral origins. Last but not least there is a eighth argument introducing semiotic principles that viruses are an essential part of the tree of life and provide coherent explanations of genetic information processing by the biocommunicative approach.

### 1) Prokaryotic genomes

1 The arguments of Moreira and Lopez-Garcia to exclude viruses of the tree of life are based on  
2 the comparison of viral features and cellular features. But as “cellular” features they assume  
3 in most cases eukaryotic cells. Prior to them we have to look at the prokaryotic genomes and  
4 their gene word order if we shall have an appropriate basis of the roles of viruses in the  
5 evolution of prokaryotes as hypothesised predecessors of eukaryotes. Interestingly the serial  
6 endosymbiotic theory which identified key components of the eukaryotic cell such as  
7 mitochondria and chloroplasts as former free living bacteria is supported by the suggestion  
8 that the eukaryotic nucleus derived from a large doublestranded DNA virus, which represent  
9 the properties of eukaryotic nucleus not present in any known prokaryote (Villarreal, 2005).  
10 Additionally this is supported by a recent review regarding the role of viral polymerases in the  
11 origin of mitochondria and chloroplasts (Filée and Forterre, 2005; Brussow, 2009).

12 The oceans (and the world) are intensely viral. All life must survive this viral laden  
13 habitat and survivors generally retain prophage (or provirus) or their defectives. This includes  
14 marine bacteria (Krupovic and Bamford, 2007), extremophiles in deep sea hydrothermal vents  
15 (Williamson *et al*, 2008a) as well as the organisms of the Antarctic (Angly *et al*, 2006;  
16 Dinsdale *et al*, 2008). If we imagine that 1ml of seawater contains one million bacteria and  
17 ten times more viral sequences it can be determined that  $10^{31}$  bacteriophages infect  $10^{24}$   
18 bacteria per second (Tettelin *et al*, 2005). Since the beginning of life this has been an ongoing  
19 process. The enormous viral genetic diversity in the ocean seems to have established  
20 pathways for the integration of complete and complex genetic data sets into host genomes,  
21 e.g. acquisition of complex new phenotypes. A prophage can provide the acquisition of more  
22 than 100 new genes in a single genome editing event (Campbell, 2007; Canchaya *et al*,  
23 2003a; Brussow *et al*, 2004; Villarreal, 2009; Ryan, 2009).

24 It is now 13 years since first bacterial genome was sequenced, and comparative  
25 genomics now provides us a very clear picture of prokaryotic evolution; both bacteria and  
26 archaea show dominant force of evolution is mediated by horizontal gene transfer (HTG)  
27 (Koonin and Wolf, 2008). Comparing metabolic pathways of 160 prokaryotic species, shows  
28 acquisition of gene sets not by point mutation but by rapid and massive acquisition of gene  
29 groups (Iwasaki and Takagi, 2009). Clearly, as sex is not common to most bacteria, this is  
30 mainly mediated by phage action as such changing gene clusters (aka phage islands) are  
31 adjacent to tRNA integration sites. This inherently symbiogenic situation was also apparent  
32 with the initial sequencing of *B. subtilis* as the second complete bacterial genome (Sonenshein  
33 *et al*, 2002). Thus, the “Tree of Life” concept has been severely undermined and can not  
34 apply to such large scale lateral gene transfer processes (Baptiste *et al*, 2005; Lopez and  
35 Baptiste, 2009) or explain the role of viruses (Sinkvics, 2001; Brussow, 2009). Yet a tree-  
36 like structure of genetic evolution is observed in all domains of life, including most viruses.  
37 Thus HGT is colonizing an existing tree from non-ancestral (viral) sources. However, ‘Tree-  
38 thinking’ which explains tree growth by ancestral variation and natural selection continues to  
39 be vigorously defended leading many to dismiss the prokaryotes as ‘odd-balls’ that evolve  
40 differently from other life. Evidence now compels us to revise our definition and vision of  
41 the Tree of Life to include viruses. Figure one thus presents a schematic of how an inherently  
42 fuzzy virus community (shown as a cloud) provides the information and process of host  
43 colonization. Reticulate evolution and symbiosis apply to all life and must now be  
44 incorporated into our conceptual framework (Beiko, Doolittle, and Charlebois, 2008). We  
45 also clearly see is recurrent endosymbiosis as a main creative force the evolution of  
46 eukaryotic life (Ryan, 2009).

## 47 48 **2) We can trace viral genes**

49  
50 The opinions of Moreira and Lopez Garcia on viral lineages is that (i) there is no viral  
51 phylogenetic tree, (ii) viral genes in general have been acquired from their host by HGT, (iii)

1 viral lineages cannot have persistence in the presence of high HGT rates, because genetic  
2 contents are not stable but highly volatile and (iv) the cell to virus gene flux is quantitatively  
3 overwhelming in contrast to the opposite. But current knowledge on viral lineages contradicts  
4 this opinion clearly. The presence of viral specific tRNA genes, their distinct G/C  
5 composition and the Restriction-Modification word bias in viral genomes allows us to identify  
6 and differentiate viral genes from host using codon and word bias based methods and apply  
7 them to DNA viruses of both prokaryotes and eukaryotes (e.g., using non-BLAST methods).  
8 This was initially done when a set of T-even phage were first sequenced (Nolan *et al*, 2006),  
9 but has been followed by other viruses (Monier *et al*, 2007; Pride and Schoenfeld, 2008).  
10 Also, it is known that DNA viruses often have an unusually large number of small genes  
11 (Villarreal, 2005; Ogata and Claverie, 2007) which makes them good sources of novel but  
12 coordinated functions. Indeed, a general assessment of viral ORFs of unknown function  
13 indicates a clear trend to code for small genes (Yin and Fischer, 2008) and most ORFs in  
14 prokaryotic genomes apparently came from viruses (Cortez *et al*, 2009). Studies of viruses in  
15 the ocean, are consistent with other virus studies and show their evolution is mosaic involving  
16 reticulated gene exchange with many other viral lineages, but they tend to most conserve their  
17 capsid genes (Rohwer *et al*, 2000; Rohwer 2003). We see lots of phage to phage DNA  
18 transfers and much evidence of intron related endonucleases involvement (Chibani-Chennoufi  
19 *et al*, 2004).

20 What has been missing in our thinking about host survival, however, are the effects of  
21 virus-virus interactions (i.e, lysogeny and lytic virus). In an intensely viral habitat, host  
22 survival depends fully on a dynamic system of virus-virus evolution that promotes the  
23 horizontal transfer of DNA. Such transfers of virus derived information are clearly seen in  
24 the genomes of all life (Villarreal, 2005; Villarreal, 2009; Ryan, 2009).

### 25 26 **3) Cell first vs. virus first perspectives**

27  
28 The authors of the mentioned tree-of-life article share the widely held opinion that viruses are  
29 products of cells and are evolved by cells. But phylogenetic analyses and comparative  
30 genomics contradict this opinion. Viruses are most abundant agents in the oceanic biosphere  
31 and metagenomic screens indicate that 3% of total protein in the oceans correspond with  
32 capsid genes of PSSM4-(T4-like) cyanophage. Viruses provide the largest reservoir of genes  
33 known in the biosphere (Comeau and Krisch, 2008; Filee *et al*, 2005) but were not 'stolen'  
34 from host. Such capsids cannot be of host origin. But these viruses can also have host-like  
35 genes. The cyanophage may have a nearly universal presence of pbsA core photosynthetic  
36 genes (Bench *et al*, 2007; Sullivan *et al*, 2006), clearly resembling genes from host. This  
37 similarity of virus and host genes is uncritically cited by Moreira and Lopez Garcia as  
38 evidence that such genes are host derived based on the common belief that when viral and  
39 host genes cluster together, it is likely that the virus derived the gene from the host. Yet, this  
40 phage psb gene is within a highly conserved gene cluster associated with energy metabolism  
41 that also contains the mazG gene (Bryan *et al*, 2008). But this mazG gene does not cluster  
42 with related host genes thus it cannot be explained by acquisition from host. We can  
43 conclude from these results that co-clustering does not identify the direction of gene transfer.  
44 Thus, a viral origin of the psb-A genes can also be asserted. Indeed metagenomic  
45 measurements of photosynthetic genes in marine habitats indicates that the majority of all  
46 photosynthesis genes are viral derived (Sharon *et al*, 2007) and are undergoing selection  
47 independent of the host. (Lindell *et al*, 2007). As these genes display virus-like codon bias,  
48 this further supports the idea that viral photosynthetic genes evolve independently of hosts  
49 (Zeidner *et al*, 2005) (Weigle *et al*, 2007) This argument for viral origin of psb genes is fully  
50 developed in Villarreal 2009.

1 Cyanophage mediate cyanobacterial evolution. Indeed, comparative genomics of  
2 *Prochlorococcus* cyanobacterial biotypes (sub-species variants) indicates that they differ  
3 mainly by gene sets or ‘phage islands’ (Coleman *et al*, 2006). These surveys: “lend strong  
4 support to the notion that viral-mediated gene acquisition is a common and ongoing  
5 mechanism for generating microbial diversity in the marine environment.” (Williamson *et al*,  
6 2008b). Accordingly, it is most commonly observed that recruitment of highly variable  
7 genome ‘fragments’ is the prevalent evolutionary process in the biosphere and such  
8 hypervariable genomic islands were too variable to assemble into coherent trees (Rusch *et al*,  
9 2007).

10 Additionally it is well accepted by virologists that viruses often contain many complex  
11 genes (including core genes) that cannot be attributed to having been derived from host genes.  
12 Indeed, comparative viral genomics and the study of virus evolution often depends on  
13 precisely such non-host gene conservation (Villarreal, 2005; Domingo *et al*, 2008). Indeed,  
14 there are many indicators that the converse relationship is prevalent as demonstrated by phage  
15 genes that colonize host.

16 Comparative genomics of *E. coli* dramatically demonstrate the importance of large  
17 scale genetic variation even within one bacterial ‘species’. It is established that *E. coli*  
18 genome variation is from 4.6 to 5.5 Mbp and much of this variation appears to be of phage  
19 origin (Binnewies *et al*, 2006). The best bacterial example are the highly sequenced genomes  
20 of pathogenic *E. coli*, such as 0157(E2348/69) (Iguchi *et al*, 2009). Indeed, it has recently  
21 become clear that chromosomal integration hotspots are occurring adjacent to *lueX* tRNA  
22 indicating a heavy phage involvement in *coli* adaptation (Lescat *et al*, 2009) (Kirsch *et al*,  
23 2004). However, non-pathogenic commensal *E. coli* has similar phage mediated genomic  
24 plasticity (Oshima *et al*, 2008). And it is clear that highly conserved prophage elements are  
25 mediating the extraordinary adjacent genomic instability in both pathogenic and commensal  
26 isolates (Bielaszewska *et al*, 2007) (Yang *et al*, 2009). From the comparative genomics of  
27 several pathogenic *E. coli* strains, it is also clear that independent infections with similar but  
28 distinct bacteriophages were deeply involved in the evolution of these *E. coli* (Ogura *et al*,  
29 2007). Furthermore, it appears these *Ler* and *Pch* phage can orchestrate the coordination of  
30 the scattered transcription of various genes involved pathogenicity (Abe *et al*, 2008). Since a  
31 similar ‘horizontal’ tRNA(*leuX*) adjacent genetic pattern applies to the differences between  
32 related bacterial but distinct bacterial ‘species’ such as *E. coli* and *Salmonella enterica*  
33 (Bishop *et al*, 2005) or *Yersinia* (Rakin *et al*, 2001), the *E. coli* example appears represent a  
34 generalized and prominent mechanism of prokaryotic evolution. Indeed, the initial sequencing  
35 of a second bacterial species (*B. subtilis*) also identified tRNA adjacent gene sets as the main  
36 difference with *E. coli* (Sonenshein *et al*, 2002).

#### 37 38 **4) There is much evidence for virus-virus interactions**

39  
40 In prokaryotes, viruses appear to exist into two broad but interacting relationships with host;  
41 lytic and persistent (prophage) (Villarreal, 2005) with two distinct evolutionary dynamics  
42 (Gelfand and Koonin, 1997). T4 are strictly lytic phage, and are the iconic example of the T  
43 even phage. Although they frequently recombine with each other, T-even phage don’t  
44 exchange very much DNA with *E. coli* (Nolan *et al*, 2006). These lytic phage are not  
45 ‘moronic’ (prophage derived small inserts into ORFs, next to promoter) like the persisting  
46 lambdoid phage (Hendrix *et al*, 2003). Their genomes are partitioned and mosaic, via  
47 extensive exchange with mostly other phage, but retain expression strategy and morphology  
48 (head and tail most conserved). Since phage replication is often recombination dependent,  
49 they have notoriously reticulated patterns of evolution. Yet even the most intensively studied  
50 phage of all (T4), retains many poorly characterized virus-specific genes (i.e. of the 300

1 genes, nearly 130 remain uncharacterized and unrelated to other genes). But some of these  
2 genes are known to affect virus-virus interactions.

3 Host fitness is not usually considered from a viro-centric perspective. Yet we can assert  
4 with confidence that numerically viruses rule the world. Consider the well studied T4 SegB  
5 gene which is a homing endonuclease, is highly conserved in T4 (but not other phage) and is  
6 needed for preferred inheritance of T4 tRNA gene region (Brok-Volchanskaya *et al*, 2008).  
7 Homing endonucleases are found in both host and virus, but those in T4 are often used to  
8 preclude non-T4-like genomes. The T4 SegG appears to provide T2 exclusion (Liu *et al*,  
9 2003) as does SegF. T4 has many such genes and similar use of endonucleases for genetic  
10 exclusion may apply to other T4 like phage, although the specific endonucleases are different  
11 (Sandegren *et al*, 2005) (see Brussow *et al*, 2004; Chibani-Chennoufi *et al*, 2004). Indeed  
12 group I introns are found in about half of all Streptococcus, Lactococcus, and Lactobacillus  
13 phage, often interrupting the lysine gene (Foley *et al*, 2000). This suggests wide spread virus-  
14 virus based selection which should have big effects on host survival.

15 The prophage - lytic phage relationship is directly relevant to this issue and has been well  
16 examined experimentally with lactic acid bacteria (Desiere *et al*, 2001; Desiere *et al*, 2002;  
17 Canchaya *et al*, 2003b). Here, it is clear that prophages are major mediators of lateral gene  
18 transfer as seen via tRNA adjacent integration (Brussow and Desiere, 2001; Canchaya *et al*,  
19 2004; Cheetham and Katz, 1995) (Brussow, 2007). Lysogenic conversion is also well  
20 established (Canchaya *et al*, 2003a). This type phage-host evolutionary process also appears  
21 to apply to the well studied bacterial pathogens (Brussow *et al*, 2004). And although extensive  
22 exchange of genes with host can be inferred, 80% of the viral orthologous groups have no  
23 host counterparts (Liu *et al*, 2006). Even cryptic prophages have been well established to  
24 exclude lytic T4 (Mehta *et al*, 2004; Toothman and Herskowitz, 1980a; Toothman and  
25 Herskowitz, 1980b), thus, even viral 'junk' matters greatly to host survival.

## 26 **5) The predecessors of the three domains of life had to be polyphyletic**

27  
28  
29 There may be several non-cellular origins of viruses consistent with their polyphyletic nature.  
30 A pre-cellular RNA world with self-replicating and self-cleaving ribozymes contains the  
31 essential host and parts for ancient RNA viruses. But beneath that we can find ancient but  
32 related DNA viruses infecting all three domains of life which completely differ from RNA  
33 viruses as well as retroviruses (a kind of RNA virus but with reverse reading direction).

34 If we look at the three domains of life we can identify three different DNA  
35 polymerases of the family B extension polymerases of the DNA replication complex which  
36 are clearly polyphyletic (Forterre, 2001, 2002, 2005, 2006a) similar to the DNA  
37 topoisomerases which also seem to be polyphyletic (Forterre and Gadelle 2009). This means  
38 at the roots of the tree of life there must be at least three polyphyletic predecessors. Only this  
39 could explain the three remaining polymerases. Since RNA genomes most probably existed  
40 before the appearance of DNA genomes (Villarreal, 2005, Forterre, 2006 b), DNA can be  
41 considered to be a modification of RNA, requiring ribonucleotide reductase, followed by only  
42 two thymidylate synthases.

43 Prior to DNA, there were only two informational components RNA molecules and  
44 proteins. Hypothesised proto-cells initially were colonized by only a few individual DNA  
45 components that still operated through RNA. In turn, these DNA agents facilitated the  
46 scission into two separate lineages – DNA viruses and RNA viruses – whereby a DNA virus  
47 was capable of infecting and persisting as stable DNA in an RNA virus host. Eventually,  
48 even more stable colonizing genomes developed dsDNA which were recombinogenic and  
49 able to capture host genes and establish permanent persistence (symbiosis). This model  
50 would at least account for the existence of virally-encoded DNA transaction proteins for

1 which no cellular counterpart exists. In addition, this model would explain the existence of  
2 two dissimilar DNA replication systems (Villarreal, 2005).

3 Even though the DNA cells gained distinct selective advantages, the RNA parasites  
4 still had an astonishingly powerful genomic creativity (Ryan, 2006, Ryan, 2009); this  
5 imparted distinct survival advantages if environmental conditions changed considerably.  
6 Today we know that the DNA world, on its own, would not have brought forth such an  
7 incredible diversity by natural selection (Gabora, 2006), let alone established the necessary  
8 genetic precondition to create such a high degree of complexity. Overall, the genomic  
9 innovations of the RNA world complements that of the more conservative and stable DNA  
10 world. Thus the consortial (commonly shared) volatility and “lack of structural continuity”  
11 (Moreira and Lopez-Garcia, 2009) of viral genomes should be an argument for the inclusion  
12 of viruses in the tree of life and provides the precondition for the evolution of the complexity  
13 of the tree of life as well as the emergence of a commonly used code for life. Even Darwin did  
14 not exclude a multiple origin of life: ‘There is grandeur in this view of life, with its several  
15 powers, having been originally breathed by the creator into a few forms or into one; and that,  
16 whilst this planet has gone cycling on according to the fixed law of gravity, from so simple a  
17 beginning endless forms most beautiful and most wonderful have been, and are being,  
18 evolved’ (Charles Darwin, *The Origin of Species*).

## 19 20 **6) The ancient and the current RNA world**

21  
22 In some sections the authors seem to suggest that the RNA world hypothesis remained a  
23 hypothesis and not well proven. However, the RNA-World research indicates a contrasting  
24 perspective. From the early RNA-world perspective the whole diversity of processes within  
25 and between cells (intracellular and intercellular cell communication) depends on various  
26 RNAs. Therefore the nature of modern RNA suggests a precellular RNA world which must  
27 have been dominated by quasi-species consortia-based evolution just like current RNA  
28 viruses (Domingo *et al*, 2008). A variety of RNAs can be identified at the roots of the tree of  
29 life such as

- 30 • genetic polymers inside membrane vesicles of a hypothesised protocell (Chen *et al*,  
31 2006), riboswitches (Breaker, 2006),
- 32 • a variety of catalytic strategies of self-cleaving ribozymes (Ke and Doudna, 2006),
- 33 • the structure and function of group I introns (Houglund *et al*, 2006),
- 34 • the roles of RNA in the synthesis of proteins (Moore and Steitz, 2006).

35  
36 Important features like the role of ribosomes in the translation from RNA into protein (Noller,  
37 2006) and great diversity of action potential of RNAs in the modern DNA world like

- 38 • the role of ribonucleoproteins (RNPs) (Cech *et al*, 2006),
- 39 • the list of functions necessary for each single cellular process such as small nuclear  
40 RNPs (Tycowski *et al*, 2006),
- 41 • small nucleolar RNPs (Matera *et al*, 2007),
- 42 • the assemblies which build spliceosomes with viral origins,
- 43 • the insertion/deletion competence for site-specific modifications of RNA molecules  
44 (Simpson, 2006),
- 45 • the unique feature of all retroelements, i.e. reverse transcriptase and other telomerases  
46 (Blackburn, 2006).

47  
48 Last but not least we can find crucial parts of the present RNA world with clear viral origin  
49 such as

- 50 • group II introns with splicing competences (Pyle and Lambowitz, 2006),
- 51 • the important roles of SINES and LINEs (Weiner, 2006),

- 1 • the whole range of non-coding RNAs (Witzany, 2009a)
- 2 • also the RNA ligase as found in T4 is found in all three domains of life (Ho and
- 3 Shuman, 2002).

4 All of these agents can be identified as descendants of an early RNA (quasi-species  
5 dominated) world which evolved prior to cellular life and are predecessors of cellular life-  
6 functions present since the last universal common ancestor of all cellular life (LUCA) in all  
7 three domains of life. On the basis of these RNAs we are able to reconstruct the emergence of  
8 the tree of life from its roots to its top.

9 Viruses can parasitize almost any replication system (i.e., other viruses including  
10 themselves) - even prebiotic ones - and probably emerged well before the appearance of  
11 cellular life forms. RNA viruses store crucial and dynamic information (in RNA-“clouds”)  
12 that not only pertains to (a) replication proteins but also to (b) morphology and (c) phenotypic  
13 diversity and retains a history of past selections. Based on this and the results of phylogenetic  
14 analyses and comparative genomics, it is possible to establish viral lines of ancestral origin.  
15 These lines of origin can also be non-linear because different parts of viruses contain different  
16 evolutionary histories (Villarreal, 2005, Domingo *et al*, 2008).

17 Since viruses with RNA genomes are the only living agents that use RNA as a storage  
18 medium, they are considered to be witnesses of an earlier RNA world, of a time when DNA  
19 did not exist yet (Forterre 2001, 2002, 2005, 2006b; Koonin *et al*, 2006). Negatively stranded  
20 RNA viruses have genome structures and replication patterns that are dissimilar to all known  
21 cell types. As demonstrated by phylogenetic analyses, cellular replicases are related to each  
22 other; however, there is no similarity between RNA-viral replicases and those of any known  
23 cell types (Zanotto *et al*, 1996). This suggests the existence of negatively stranded viral RNA-  
24 replicases even before cellular life came into being (Villarreal, 2005). DNA viruses, too, do  
25 not give any reference to a cellular origin. Phylogenetic analyses point to an older timescale,  
26 as DNA-repairing proteins of DNA viruses do not have any counterparts in cellular biota.

## 27 28 **7) The Virophere is not restricted to prokaryotes**

29  
30 Eukaryotes also show that viruses mediated the most dynamic part of their genomes as well.  
31 But here, retrovirus related agents are much more active than the DNA prophage or DNA  
32 episomes of prokaryotes (see figure 1). Also, eukaryotic retroviruses more often seem to  
33 manipulate (reprogram) gene regulation and function, not invent new genes. There are also  
34 lots of eukaryotic large DNA viruses in the sea as reported by metagenomic screens (via DNA  
35 pol, Mimivirus and algal viruses) (Monier *et al*, 2007, 2008). And we can see major  
36 consequences to host by these viruses. For example, *Ostrecoccus tauri*, the smallest marine  
37 photosynthetic eukaryote, is host for large DNA virus. This virus lacks  
38 restriction/modification enzymes seen in other viruses of microalgae (Derelle *et al*, 2008).  
39 Yet, like cyanobacteria, it too is a highly gene dense chromosome that seems to evolve mostly  
40 via the action of horizontal transfer (Palenik *et al*, 2007). But this eukaryotic host also has  
41 two chromosomes that are structurally distinct, have biased G/C content and codon usage and  
42 one of which contains the majority of transposable elements (Derelle *et al*, 2006). Thus even  
43 in representatives of early eukaryotes, we can see a large expansion in the kind and quantity  
44 of (former parasitic) virus derived DNA. Thus they too have been molded by symbiotic viral  
45 mediated events (horizontal, reticulated and symbiotic acquisitions). But what really differs is  
46 the nature of virus involved in eukaryotes and how they control their host. Eukaryotes are not  
47 gene dense, do not support DNA prophage, but they have acquired regulatory genetic  
48 complexity via the action of mostly retroviral (and viral defective) colonization. We can  
49 clearly demonstrate this assertion by comparative genomics. For example, a comparison of  
50 human and chimpanzee DNA establishes the prevalent and recent role of endogenous

1 retroviruses (Ryan 2009). Viruses contributed clearly our recent ancestors. So why would we  
2 choose to deny this incontrovertible fact and propose to exclude virus from the tree of life?

### 3 4 **7.1 A viral-driven and symbiotic origin of eukaryotes**

5  
6 Serial Endosymbiotic Theory (SET) suggested that the eukaryotic cell did not result from  
7 random mutations, but from the coordinated union of former free-living prokaryotes  
8 (Margulis, 1996, 1999, 2004; Margulis *et al.* 2000; Margulis and Sagan, 2002; Witzany, 2006  
9 a). So far it is clear that mitochondria and other organelles descended from these micro-  
10 organisms (Odintsova and Yurina, 2000, 2005) and it has also been asserted that the  
11 eukaryotic nucleus is of archaeal or bacterial descent.

12 In the meantime, however, there are good reasons to support the idea that eukaryotic  
13 nuclei originated before the symbiogenetic integration with mitochondria and chloroplasts,  
14 (see Villarreal, 2005, Villarreal, 2009). In fact, the nucleus has basic properties that are  
15 otherwise absent in prokaryotic cells (Bell, 2001, 2006).

16 Prokaryotes do have circular chromosomes with uniform standardised origins of  
17 replication. Their chromosomes are only loosely attached to chromatin proteins and have  
18 different control regions that coordinate and terminate DNA replication. All eukaryotic  
19 proteins involved in DNA replication differ from those found in prokaryotes. Hence, nuclear  
20 properties of eukaryotes are completely different from those of prokaryotes (Villarreal, 2004).  
21 These differences include, for example, use of linear chromosomes, with elaborately  
22 controlled and multiple origins of replication, repetitive termination points, transcription and  
23 translation which are separated via nuclear membrane and the existence of complex nuclear  
24 pore structures that actively mediate RNA translocation.

25 All these properties represent complex phenotypes, which require complex co-  
26 ordination of numerous protein functions. None of these functions can be found in  
27 prokaryotes even though they are considered to be the predecessors of the eukaryotic nucleus.

28 The eukaryotic nucleus contains three kinds of DNA-dependent RNA polymerases that  
29 differ significantly from RNA polymerases of prokaryotes (Villarreal, 2005). Even the three  
30 kinds of splicing group I-introns (DNA transposase, reverse transcriptase and micro-RNAs)  
31 are largely nonexistent in prokaryotes, but they are present in viruses of prokaryotes. In  
32 addition, no single prokaryotic process is known to account for the tasks of membrane  
33 disintegration and restoration as observed in eukaryotes, but this too can be seen in DNA  
34 viruses (Villarreal, 2005).

35 Viral genes are directly involved in tasks of transposition. Viruses generally mark their  
36 genome, their RNA and their proteins with various virus-specific enzymes, such as  
37 methylases; e.g. via enzymatic reactions known as base methylation (Villarreal, 2005).

38 Moreira and Lopez-Garcia suggest that viruses are gene-“robbers” of cellular life  
39 (Moreira and Lopez Garcia, 2009: 309). However, as asserted above, it is a well-known fact  
40 that viruses create new genes as a result of their evolutionary line of descent, via tremendous  
41 rates of recombination (often replication dependent) and a high tolerance for errors. Just by  
42 looking at the baculoviruses, for example, with reference to GenBank database investigations,  
43 we can find that 80% of their genes are unique to this group and found nowhere else (Herniou  
44 and Jehle, 2007; Herniou *et al.*, 2001, Villarreal, 2005). Gene losses have been documented in  
45 baculoviruses but the twelve losses documented therein are countered by a staggering  
46 acquisition of 255 new genes. Similar observations of gene novelty apply to other families of  
47 large DNA viruses found in both prokaryotes (Hendrix, 2009) and eukaryotes (Domingo *et al.*,  
48 2008, Villarreal, 2009).

49 Between 1950 and 1980 scientists realised that the T4 phage-polymerase proteins are  
50 much more similar to the eukaryotic DNA polymerase protein than to any prokaryotic  
51 polymerase. Today we know that the eukaryotic DNA polymerase and the T4 DNA

1 polymerase do have common origins. Indeed, the T4-like viruses (tailed icosahedral phage,  
2 Caudavirales) represent a huge family of viruses that is capable of infecting both bacteria and  
3 archaea (Villarreal, 2004). Hence, it is not surprising that T4-DNA polymerases (and capsids)  
4 are found in all three domains of life: archaea, bacteria and eukarya - although it must be  
5 noted that the phage HK97 capsid protein shared by Caudavirales doesn't indicate ancestral  
6 relationships.

7 Algae were among the first more complex eukaryotic organisms that had to deal with  
8 viruses. Thus, viruses that infected microalgae must have had a large adaptive potential that  
9 accompanied the evolutionary pathway and which must have included the characteristic of a  
10 protonucleus. Here in particular we can think of the phycodnavirus and therefore one has to  
11 examine the entire GenBank database for sequences that may be similar to the DNA  
12 polymerase of this particular virus (Villarreal, 2004). Such sequences must include replication  
13 polymerases of all higher eukaryotes as well as of all larger eukaryotic DNA viruses, primer  
14 polymerases of eukaryotes, and repair polymerases of both archaea and bacteria.

15 The DNA polymerase of the CSV1 virus can be found at the phylogenetic origin of  
16 any of the eukaryotic replication DNA polymerases and can be considered as a precursor of  
17 all polymerases which are involved in replication of the eukaryotic genome. So far no other  
18 viral or prokaryotic DNA polymerase that shares these features is known (Villarreal 1999,  
19 DeFilippis and Villarreal 2001, Villarreal, 2004).

20 The membrane-bound separation of transcription and translation is a characteristic of  
21 the pox viruses; more concretely, of the vaccinia and other large DNA viruses (Mimivirus)  
22 (Villarreal 2005). Moreover, these viruses have a very simple pore structure that has actively  
23 been incorporated from the membrane-bound RNA into the cytoplasm of the host. A similar  
24 situation can be documented with the small chromatin proteins and the linear chromosomes  
25 along with their repetitive telomer tails that are so characteristic of various cytoplasmid DNA  
26 viruses, TTV1 and phycodna viruses. Even the highly complex function of tubulin as an  
27 important coordinating element during chromosomal separation of duplicated strands can be  
28 found in DNA viruses with exactly the same set of functions (Villarreal, 2004).

29 It became increasingly obvious that all properties of the eukaryotic nucleus are  
30 compatible with having been derived from a large, stable and persistent DNA virus with  
31 linear chromosomes. The precursor of the eukaryotic nucleus indeed appears to have been a  
32 huge membrane-covered DNA virus (similar to Mimivirus or poxvirus) that persistently  
33 colonised a prokaryotic host (Villarreal, 2004, 2009; Bell, 2001, 2006). Therefore, the hosting  
34 cell must have lost its cell wall with the virus incorporating the prokaryotic genes into its pre-  
35 nuclear genome: particularly in cases of encoding for metabolism and translation. This virus  
36 was probably non-lytic, as it coordinated both its own replication and its transcription genes,  
37 and it had a double-layered membrane (reminiscent of that seen in herpesviruses) and a  
38 tubulin system in order to wrap chromosomes. Its persistence and its reactivation would imply  
39 that (a) the process of cell division (nuclear envelope dispersion and reformation), (b) mitotic  
40 duplication (doubling of the chromosomes and allocating them to the progeny cells) and (c)  
41 the viral DNA correspond to the sexual reproductive cycle of the host organism. Such an  
42 infectious origin for the nucleus would also be compatible for the prevalence of infectious  
43 nuclei, seen in many species of parasitic red algae that represent a basal eukaryote) (Goff *et*  
44 *al.*, 1997). Red algae are clearly the oldest eukaryote that can be found in the fossil record.  
45 Comparative genomics is consistent with this fossil evidence (Cole and Sheath, 1990).  
46 Interestingly all these properties can be found in various prokaryotic viruses such as  
47 cyanophage, archaeal phage, mycobacterial phage and eubacterial phage (Villarreal, 2005,  
48 2009).

49 Not only can the eukaryotic nucleus, however, to be considered of viral origin, but as  
50 referenced above, RNA polymerase, DNA polymerase and DNA helicase which transcribe

1 and replicate DNA in modern mitochondria may also be of viral origin (Filee and Forterre,  
2 2005).

## 3 4 **7.2 Ongoing roles of retroviruses and mobile genetic elements in Eukaryotes**

5  
6 Although DNA prophage no longer modify the eukaryotic genome as seen in prokaryotes,  
7 endogenous retroviruses (proviruses) do. The human genome has many more elements  
8 derived from retroviruses (LTR containing) than it has genes (ORFs) (Hughes and Coffin,  
9 2001) (Kim *et al*, 2004). And these endogenous retroviruses (ERVs) distinguish human DNA  
10 from all other primates (Andersson *et al*, 2002) (Mayer and Meese, 2005) (Polavarapu *et al*,  
11 2006). Such ERVs have colonized Eukaryotic genomes from exogenous viral (non-ancestral)  
12 sources. All domains of eukaryotic life have their own peculiar pattern of colonization by  
13 such viruses (Villarreal, 2005, 2009). And it has been experimentally verified that various  
14 crucial and complex functions of eukaryotes (vertebrates) are directly mediated by these  
15 ERVs (see Ryan 2009). This includes the emergence of the adaptive immune system  
16 (Villarreal, 2009) as well as vivipary in mammals (Mi *et al*, 2000; Dupressoir *et al*, 2005;  
17 Dunlap *et al*, 2006). The most numerous mobile elements in the eukaryotic genome integrate  
18 into a host genome via an RNA intermediate, reverse transcriptase.

19 Copying from RNA into DNA generally involves reverse transcriptases. Mobile  
20 elements are important for genotype processing, with far-reaching consequences for  
21 phenotype expression during its various developmental stages (Jurka *et al*, 2007). Recent  
22 research has demonstrated that overlapping epigenetic marking in eukaryotic cells is an  
23 important evolutionary feature to silence the expression of mobility of these mobile elements  
24 (Slotkin and Martienssen, 2007). Mobile elements can silence single genes as well as larger  
25 chromosomal regions and therefore, play an important role in the evolution of diversity. They  
26 share their competence to recombine, rearrange and insert into genomic content with other  
27 retroelements (Coffin *et al*, 1997). They influence neighbouring genes through alternative  
28 splicing and are active agents as enhancers and promoters or act by polyadenylation patterns  
29 (Slotkin and Martienssen, 2007). Indeed, related elements appear to account for the origin of a  
30 substantial part of the regulatory sequences in the human genome (Jordan *et al*, 2003).

31 ERV related retroposons have direct repeats at its ends (LTR), others transposons do  
32 not (non-LTRs). Interestingly, the number of retroposons increases with every transposition  
33 (transposition duplication) so that they can expand host genomes: LINE-1 is 20 % of the  
34 human genome (Maita *et al*, 2004). Like some ERVs, full LINEs contain a code for the  
35 transposase protein, which have been proposed to be evolutionary related to retroviral and  
36 ERV integrase (Capy *et al*, 1997). This enzyme identifies the terminal inverted repeats which  
37 flank mobile elements, excises them and integrates itself in place of them. The gap at the  
38 donor site is repaired in a cut-and-paste transposition or filled up with a copy of the  
39 transposon by a gap repair technique (Slotkin and Martienssen, 2007). Although many think  
40 LINEs are independent of ERVs, deep genome database comparison to early ERVs  
41 (Chromoviruses, DIRs) suggests that most transposons and retroposons originally descended  
42 from recombinant viruses which persistently integrated into host genomes and became  
43 defective (for references see Villarreal, 2004, 2005, 2009, Domingo, *et al*, 2008. Weiss,  
44 2006).

## 45 46 **7.3 Non-lytic but persistent viruses as non-coding regulatory agents for cellular needs**

47  
48 Like persisting prophage of prokaryotes, the persistence of retrovirus derived elements in  
49 Eukaryotes have major consequences. Principally, they provide a diffuse but coordinated  
50 system to edit and control the genome, but also affect virus susceptibility. Some thousands of  
51 endogenous retroviral sequences have been integrated into the human genome, and until now

1 there are 22 independent retroviral families identified (Bannert and Kurth, 2004; Bromham,  
2 2002; Buzdin *et al*, 2002; Hughes and Coffin, 2001; Khodosevich *et al*, 2002; Sverdlov, 2000,  
3 Villarreal, 2004, 2005; Ryan, 2006; Griffith and Voisset, 2008). A quantity of remaining  
4 former viral gene embedding repetitive elements embracing an enormous genetic diversity  
5 originally accompanied the protein coding sequences as control- and/or identification  
6 segments. Most endogenous retroviruses have been degraded into formerly connected  
7 domains, but they can still be recognised by their three genes gag, pol and env (Gao *et al*,  
8 2003; Ryan, 2004).

9 It became obvious that mobile sequences such as transposons and retroposons (Volff,  
10 2006) and non-coding repetitive elements such as LTRs, SINEs, LINEs (long terminal  
11 repeats, short interspersed elements, long interspersed elements) make far-reaching eukaryotic  
12 DNA rearrangement and reorganisation possible (Shapiro, 2002; Sternberg, 2002; Shapiro  
13 and Sternberg, 2005). Together, they play a decisive role in the evolution of new genomic  
14 structures (Shabalina and Spiridonov, 2004; Shapiro, 2004; Sternberg and Shapiro, 2005).  
15 Being dependent on the state of development, the varying chromatin markers are thus  
16 capable, through different methylation patterns, histone modifications and alternative splicing,  
17 of coming up with a set of multiple protein meanings, from one and the same genetic dataset  
18 (Turner, 2000, 2002; Jenuwein and Allis, 2001; Brett *et al*, 2002; Xu *et al*, 2002; Jaenisch and  
19 Bird, 2003; True *et al*, 2004; Wang *et al*, 2004). This marks even the rise of epigenetics and  
20 its research object that phenotypic variations, which are even heritable, must not depend on  
21 genetic alterations (Jablonka and Lamb, 1989, 2002, 2006; Van De Vivjer *et al*, 2002; Van  
22 Speybroeck *et al*, 2002). The question arises as to how and why the evolution of higher  
23 genetic complexity is connected to non-coding virus derived DNA, that was formerly called  
24 'junk' DNA.

25 Although it has been known for several decades that the unbelievable diversity of  
26 enzyme proteins is a practical tool for DNA editing processes, it was unclear by which rules  
27 or higher-order regulations they are governed (Witzany, 1995, 2000, 2005). Later on it  
28 became obvious that higher-order regulations such as co-suppression, suppression of  
29 transposition, position effect variegation, several start- and stop-signals, RNA interference,  
30 imprinting, chromosomal methylation, transvection and transcriptional and post-  
31 transcriptional gene silencing are processed by non-protein-coding RNAs, especially micro-  
32 RNAs (Mattick, 2001, 2003, 2005; Mattick and Gagen, 2001). New research indicates that  
33 these repetitive non-coding sequences originated primarily from retroviral RNA (Villarreal,  
34 2004, 2005).

35 The ERV related and repetitive sequences are highly species-specific and more  
36 suitable for the determination of species than the corresponding coding sequences (Villarreal,  
37 2005, 2009). Each taxon organises and formats its genome architecture differently, i.e.  
38 regulation of expression, transcription, replication, and translation are species-specific. Within  
39 each taxon, these processes along with the associated gene architecture must run in a highly  
40 coordinated manner so that they do not disturb each other. Only a co-ordination of the  
41 individual steps ensures that these different actions are performed and maintained  
42 successfully; i.e. DNA sequences must be read at the right site and at the right time. The  
43 precise spatio-temporal coordinations are essential in order to sustain vital processes;  
44 nonetheless, and in principle, these coordinations can also fail, either by sequence-damage  
45 (i.e. mutations) or by organism-induced translational, transductional, repairing or other  
46 rearrangement disturbances.

47 Experience has shown that (1) excision, (2) insertion and (3) combination of the  
48 genomic texts are the keys in DNA editing and the basis of evolutionary processes. In  
49 contemporary terms, we could call it 'natural genetic engineering' (Shapiro, 2004).

## 50 8. The tree of life from a biocommunicative perspective

1  
2 Finally, we should look at an additional argument for integration of viral consortia into the  
3 tree of life which is not part of cell biological and molecular biological but of bioinformatic  
4 perspectives. Although bioinformatics serves as important tool in detection and comparing  
5 sequences of genetic texts we must look at complementary perspectives on genetic  
6 information processing such as biolinguistics, biosemiotics or the biocommunicative  
7 approach. But isn't it a commonly shared and widely accepted assumption that molecules of  
8 nucleic acid sequences are the result of randomly derived mixtures? How should we think  
9 about evolution editing biological codes? What would be needed? Agents like viral or virus-  
10 like consortia can coordinate their behavioural strategies via an inherent capacity to identify  
11 sequence specific content arrangements. Can this allow them to integrate a defense of an  
12 attack, and invent (i) *de novo* (elongate) code or (ii) edit code through rearranged new  
13 sequence combination (recombination), (iii) and/or provide alternative regulations? Can  
14 viruses thus provide new coordinated code sequences?

15 Nevertheless it seems to be reasonable to reflect on this feature because recent  
16 knowledge indicates the fundamental role of information processing in both viral life  
17 strategies and key regulatory novelty in cell biology.

18 If we look at life from an encoding and editing perspective in a recent article Patrick  
19 Forterre suggests defining life as both 'ribosome encoding organisms and capsid-encoding  
20 organisms and their ancestors' (Forterre, 2009). Yet both encode (using the same molecular  
21 alphabet) albeit encoding different products. They share a competence, i.e. to code, in contrast  
22 with entities which do not share the competence to encode. But what does 'encode' mean? It  
23 has something to do with a code, i.e. the genetic code, which as DNA serves as an  
24 information storage medium, as RNAs serves as information-based editing agents. But  
25 'information' is something which seems to be very useful the first time, but not the second  
26 time. There are approximately 60 different definitions of 'information', many of them  
27 incompatible with each other.

28 Several scientific approaches focus on the linguistic-like structure and function of the  
29 genetic code such as biolinguistics (Popov *et al*, 1996; Ji, 1997, 1999; Searls, 2002; Chomsky,  
30 2004; Zhang, 2006), biohermeneutics (Chebanov, 1994; Markos, 2002), biosemiotics  
31 (Florkin, 1974; Sebeok and Umiker-Sebeok, 1992; Hoffmeyer, 1996; Barbieri, 2001, 2007),  
32 protein-linguistics (Gimona, 2006) and biocommunicative approach (Witzany, 1995, 2000,  
33 2009b). Their common focus: a code is a language-like structure. This at first glance seems to  
34 be difficult because investigations on 'language' and its concrete use in 'communication' are  
35 not core competences of natural sciences. Any trials of a coherent and sufficient definition of  
36 'language' and 'communication' by information theoretical, cybernetic, systems theoretical,  
37 mathematical, statistical, and mechanistic methods failed (Witzany, 2000), because formal or  
38 even comparative analysis of the combinatorial patterns of sign sequences, i.e. the syntactical  
39 level, cannot identify the pragmatic *interactional* contexts in which living agents are  
40 interwoven that determine meaning/functions of sign sequences.

41 For example biolinguistics interprets and investigates genetic structures in the light of  
42 linguistic categories (Popov *et al*, 1996; Ji, 1997, 1999; Searls, 2002; Chomsky, 2004; Zhang,  
43 2006). Similarly to bioinformatics they use statistical methods and algorithms to identify  
44 sequence orders for measurements of sequence-length and content homologies. Biolinguistics  
45 follows bioinformatics and its model of language as a quantifiable set of signs and is still  
46 convinced that it would be possible to extract semantic contents by analysis of the 'universal  
47 syntax'. In a limited sense this is possible, e.g. in genetic sequence comparison, i.e.  
48 comparative genomics. But an unambiguous determination of genetic semantics through  
49 analysis of the molecular syntax of genetic code is not possible in principle, because analysis  
50 of the molecular syntax does not tell us anything about the context in which the genetic  
51 content bearer of the genetic information is interwoven in real life. This context plays an

1 important role in epigenetic imprinting and therefore in the construction of different  
2 methylation patterns which then are the determinants for alternative splicing pathways of the  
3 same genetic datasets. This crucial role of pragmatic contexts is not part of the methods of  
4 biolinguistics and bioinformatics.

5 One result of these deficiencies is that invention of new and even complex genetic data  
6 sets or, as they may be called, gene blocks and the coherent integration of new genes or gene  
7 blocks in pre-existent genetic content arrangements by competent agents is not part of  
8 bioinformatics or biolinguistics, because innovative generation of new genetic content which  
9 is not randomly assembled cannot be deduced out of a mathematic model of language, i.e.  
10 formalisable procedures such as algorithms.

11 As in any language there are certain characteristics of the genetic code which are  
12 fundamental and cannot be reduced to each other. This means they have complementary  
13 functions: a language/code needs an alphabet, i.e. signs which can be combined in certain  
14 ways with the result of different modes of combination. The rules (not laws) of correct  
15 combination we term syntax. Manfred Eigen spoke of the ‘molecular syntax of nucleotide  
16 sequences’ (Eigen and Winkler, 1975). Eigen also introduced the term quasispecies (in the  
17 oversimplified meaning of a ‘master template’), but we now are able to understand this to be a  
18 consortial and dynamic system of virus information processing, adaptation and evolution with  
19 no master code, (Domingo *et al.*, 2008) but with coordinated capacity (i.e. competence) to  
20 edit. In terms of philosophy of science Eigen followed the most prominent “linguistic turn”  
21 i.e. syntax analyses should offer understanding of semantic contents of code-sequences  
22 exclusively, and dismissed primacy of pragmatics, i.e. real-life situations in which sequence  
23 generating consortia are interwoven (Witzany, 1995).

24 A code or a language, however, does not code itself nor does a language speak  
25 (Witzany, 2000, 2006 b). A code or a language (‘nucleic acid language’) is generated and  
26 used by some living agents in real life-worlds to coordinate and organise via communication  
27 processes (Witzany, 2007), i.e. they are interwoven in constantly changing environmental  
28 processes. To balance life *interactionally* in real-life habitats is a kind of behaviour, in  
29 contrast with non-behavioural cause-effect reactions, underlying natural laws strictly. By  
30 using codes or languages, living agents can modulate the information that determines their  
31 behaviour. To generate different modes of behaviour in a coordinated way living agents  
32 follow not syntactic rules but pragmatic rules. According to different kinds of pragmatic  
33 situational-contexts living agents are able to use a limited number of signs (alphabet) to  
34 generate different messages, i.e. varying sign-sequence arrangements. This means that an  
35 identical sign sequence can be used to transport different messages according to the different  
36 needs of the living agent as a result of the different situational contexts in which it is  
37 interwoven. Therefore – this is really important – pragmatics determines the meaning/function  
38 of code-sequences or language-sequences (we term “sentences”).

39 Within the three semiotic (sign-theoretical) levels of rules (syntax, pragmatics,  
40 semantics) which are essential characteristics of any real code or language it must be noted  
41 that no single living agent is capable of generating or using a code or a language. Ludwig  
42 Wittgenstein once noted that is not possible for only one person to follow a rule only once  
43 (Wittgenstein, 1972). The capability to follow rules depends on a historically grown ‘culture’  
44 (customs) of social interactors, i.e. group behaviour. The emergence of codes or languages  
45 needs, as a fundamental precondition, a consortium of living agents in principle. Therefore it  
46 seems likely that living agents being competent to encode coherent content sequences in the  
47 nucleic acid language emerged in parallel and in great number. In this, it is thus most  
48 interesting that RNA virus evolution and adaptation is fundamentally consortial (Domingo *et al.*  
49 *et al.*, 2008). And because they most likely evolved prior to DNA based cellular life forms also  
50 the assumption of a single last universal common ancestor (LUCA) as suggested by the  
51 authors of the current review is less plausible. If nucleic acid coding agents invented

1 membrane bound cellular life which implies a commonly shared translational code it must be  
2 assumed that they emerged at once or in parallel in great numbers so that we must speak in  
3 the plural of LUCAs (Witzany, 2008) as being quasispecies based and reticulated.

4 Should we term these coding agents as viruses or subviral RNA species? Viruses  
5 themselves do not differentiate this issue very much. The multiplicity reactivation and/or  
6 complementation of defective individuals described in several articles (Bailey et al., 2004;  
7 Villarreal, 2005; Zeidner et al., 2005), is common and occurs at a critical level of viral parts  
8 such that these viral parts are able to recreate a complete virus after being damaged by e.g.  
9 UV radiation. This indicates that competent code editing could also occur on a subviral  
10 ribozymatic level by an otherwise defective consortia.

11 Not even one known natural code or language is the result of a random-like mixture of  
12 signs/parts of an alphabet. Any structure which functions as code or language has to be  
13 generated by a consortium of living agents which share a common competence to use signs  
14 according to basic semiotic (syntactic, i.e. combinatorial, pragmatic, i.e. context-sensitive,  
15 semantic, i.e. content-specific) rules. These rules in most cases are very conservative but in  
16 contrast with natural laws they are changeable in principle, which is an advantage for  
17 adaptational purposes as well as for the generation of new sequences *de novo* which never  
18 existed before. Viruses inherently provide this essential capacity.

19 This view about 'What is life?' avoids suggesting that the term 'organism' is crucial to  
20 the tree of life, because it may be paradox to describe an ancient hammerhead ribozyme with  
21 self-cleaving competence as an organism. 'Organism' from this point of view seems to be  
22 appropriate for cellular life forms. Although viruses and subviral agents are both driving  
23 forces in the evolution of organisms and represent key functions of all cellular organisms they  
24 cannot be considered to be organisms even if sometimes able to re-animate organisms. From  
25 the biocommunicative perspective current knowledge about viruses and their varying life  
26 strategies indicates their driving force and crucial role in the evolution of all cellular life  
27 forms and key processes as well as key regulations of cells. Therefore the 'origin of life' at  
28 the roots of the tree of life from this perspective started with the origin of agents with  
29 encoding competences, a feature which is absent in non-animated nature. There are no  
30 (semiotic) rule-following agents if water freezes to ice (Witzany, 2009b).

31 From the biocommunicative aspect life starts with the competent interaction between  
32 swarms of ribozymes, i.e. based on identity/difference of encoded (molecular) syntactic order  
33 of nucleotides. This does not include the evolution of cellular metabolism at this stage even if  
34 some compartmentalisation is needed such as a different molecular syntactic structures. This  
35 means that we must decide whether ribozymatic swarms on extraterrestrial planets we term  
36 presence of life or presence of proto-life. This leads into a paradigmatic discussion with  
37 cellular textbook conviction on the one side and bioinformatic innovation at the other. At the  
38 current stage the position in our article is rather conservative but doesn't exclude further  
39 developments.

## 40 41 **Conclusion**

42  
43 Phylogenetic analyses as well as GenBank database and genome comparison show that most  
44 adaptations are of external (horizontal) and mostly viral origin. Thus natural genome-editing  
45 competences are not of cellular origin but represent original skills of viruses. Viruses have  
46 two completely different life strategies, which are clearly reflected in their genomes. In  
47 comparison, acute viruses that exhibit lytic action induce disease and even death, whereas the  
48 life strategy of persistent viruses implies compatible interactions with the host, either by being  
49 integrated into the hosting genome or within the cell plasma, and act non-destructively during  
50 most life stages of the host. These two viral life strategies often oppose one another in a  
51 population dependent way.

1 The persistent lifestyle allows the virus to transmit complex viral phenotypes to the  
2 hosting organism. This enables the host to broaden evolutive potentials that may well lead to  
3 the formation of new species. This, along with providing resistance to lytic virus, is an  
4 advantage for host survival also for the persisting agent.

5 The natural genome-editing competences of viruses are most complex in prokaryotes,  
6 in which the complete nucleotide word order is largely determined, combined, and  
7 recombined by viruses. Hence, the main genomic novelties are found in the prokaryotic  
8 domain from where they originally evolved into the higher life forms. Probably all basic  
9 enzymatic variations originated therein. Massive viral colonisation occurred from the very  
10 beginning of life starting with the evolution of bacteria and archaea, and later that of  
11 protoctista and multicellular eukaryota. The formation of all kingdoms, their families, genera,  
12 and species relies on the effects of viral colonisation and results in diversified lineages and  
13 ultimately in the evolution of new species.

14 Increasing complexity and diversity are caused by genetic innovations, new  
15 combinatorial patterns of genetic content, non-coding regulatory networks and modifications  
16 of the genomic architecture. Interestingly, increasing eukaryotic complexity correlates with  
17 expansion of non-coding DNA (Taft and Mattick, 2004). Combinatorial and rearranging  
18 processes in evolutionary and developmental processes occur non-randomly. They need to  
19 successfully and coherently abide by the rules of molecular syntax (Eigen and Winkler 1975;  
20 Witzany, 1995). Genetic content arrangements within the genomic matrix depends on  
21 situational contexts in which living organisms are involved in vivo (e.g. context of growth,  
22 mating, virulence, stress, etc.) and is therefore able to produce multiple protein meanings, i.e.  
23 different semantic contents of the same genetic dataset. Undoubtedly this promotes the  
24 development of regulatory complexity. This is the prerequisite for epigenetically-induced  
25 evolutionary and developmental processes. These rule-following processes may even fail,  
26 with fatal consequences for the organism (Witzany, 2006 b). If evolutionary processes are  
27 intertwined between different species complexity is even more evident (Villarreal, 2005;  
28 Zhang 2006). This indicates an important role of symbiogenetic processes in enhancing  
29 genetic, genomic and phenotypic complexity and diversity (Ryan, 2009).

30 Obviously, evolutionary history emerges from a totally different circumstance than  
31 previously thought. It must not be understood as an aggregation of chance mutations of the  
32 genetic text and its associated selection, but more as permanent and competent processing of  
33 genetic sequences for the purpose of acquisition of previously unknown abilities as an  
34 advantageous productivity to ward off competing parasites from host organisms via genomic  
35 innovation.

36 The very genetic volatility of viruses, used as argument to exclude them from the tree  
37 of life must now be considered as an essential *precondition* for life. Inventions of new  
38 genetic sequences and integration in host organisms change their genetic identity and are an  
39 advantage against competing genetic parasites which are not incorporated in host genomic  
40 contents. But inventions of new genetic sequences must occur through these competitors also,  
41 because otherwise they cannot colonize ever-changing host genomes. Genetic inventions by  
42 competent genetic content editors therefore must be considered as the driving force of  
43 evolution in both viruses and cellular host genomes. The changes of DNA-stored data by  
44 random mutations solely would never result in the great variety of new genetic content  
45 arrangements. Only the creative ('error-prone') consortial RNA gene pool (quasispecies) with  
46 its fast-changing genetic content arrangements and its selection (being stored in the DNA  
47 'evolutionary protocols'. see Vetsigian *et al*, 2006) can serve as an appropriate explanatory  
48 concept. Thus we can start to understand why our own genome is composed of such a large  
49 and varied population of "defective" virus. which now act in their novel roles as "effective"  
50 regulatory tools in host genomes.

1 A solution to the problem of cryptobiosis as mentioned in the introduction can be  
2 suggested now also. It is considered to be a contrasting phenomenon to actual life processes,  
3 because e.g. larvae of the African fly *Polypedilum vanderplanki* survive desiccation for up to  
4 17 years and temperatures ranging from  $-270\text{ }^{\circ}\text{C}$  (liquid helium) to  $106\text{ }^{\circ}\text{C}$  (Watanabe et al,  
5 2002) or some bacterial endospores are able to withstand almost any environment. Isolated  
6 endospores of thermophiles from cold lake sediments could be revived from samples some  
7 100,000 years old (Nicholson et al, 2000). Also virions (in oceans) clearly represent a  
8 conserved possibility of living agents ready for actualisation under the appropriate  
9 environmental conditions such as virions of the ATV (Acidianus-Tailed Virus). These virions  
10 can go for months without any change of their morphology at normal temperatures. When  
11 they are exposed to higher temperatures ( $70\text{ }^{\circ}\text{C}$ ) they dramatically change their structure by  
12 forming two tails at the end of their central body (Häring *et al*, 2005) which may be a  
13 precondition for biologically required changes in morphology and structure (Prangishvili *et*  
14 *al*, 2006). If these environmental conditions are lacking cryptobiotic agents remain  
15 inactivated, a complete and long lasting absence of replication or any metabolizing activity  
16 will occur. Would we assume them as not being part of the tree of life? Under appropriate  
17 environmental conditions they will start as typical for living agents. Their DNA data set is  
18 activated according the inherited regulatory pathways as no inactivation ever has occurred. At  
19 this point we should remember that the gene word order of the DNA datasets as well as the  
20 regulatory ratio is the result of persistent viral colonisation. Should we assume them not being  
21 part of the tree of life?

22 Like any tree, the tree of life is not constituted by cells of the stem, branches and  
23 leaves alone. As we know today, the stem of actual trees is in continuous communication with  
24 the rhizosphere to coordinate growth and development. Without a balanced rhizosphere  
25 ecology trees will have serious problems of survival. This rhizosphere ecology is a particular  
26 example of non-selfish interactional (symbiotic) patterns and co-evolution (Witzany, 2006 c).  
27 It crucially depends on communication processes which function in parallel between three  
28 different types of plant root cells, mychorizal fungi and rhizobia bacteria, all of them being  
29 capable of self/non-self differentiation to coordinate and organise group behaviour. Surely,  
30 this fundamental capability to coordinate group behaviour by communication processes in  
31 bacteria, fungi and plants must inherently be of viral origin. Persistent viral agents which  
32 colonised all of these organisms which are typical of rhizosphere ecology generated a great  
33 variety of host features with which these hosts are able to identify and communicate with  
34 group members. If the tree of life metaphor is to be useful in future, we have to remember that  
35 viruses have been and still are essential agents within the roots and stem of the tree of life.

## 44 References

- 45
- 46 1) Abe, H., Miyahara, A., Oshima, T., Tashiro, K., Ogura, Y., Kuhara, S., Ogasawara, N.,  
47 Hayashi, T., Tobe, T., 2008. Global regulation by horizontally transferred regulators  
48 establishes the pathogenicity of *Escherichia coli*. *DNA Res.* 15(1), R25-R38.
  - 49 2) Andersson, A.C., Venables, P.J., Tonjes, R.R., Scherer, J., Eriksson, L., Larsson, E.,  
50 2002. Developmental expression of HERV-R (ERV3) and HERV-K in human tissue.  
51 *Virology* 297, R220-R225.

- 1 3) Angly, F. E., Felts, B., Breitbart, M., Salamon, P., Edwards, R. A., Carlson, C., Chan,  
2 A. M., Haynes, M., Kelley, S., Liu, H., Mahaffy, J. M., Mueller, J. E., Nulton, J., Olson,  
3 R., Parsons, R., Rayhawk, S., Suttle, C. A., Rohwer, F., 2006. The marine viromes of  
4 four oceanic regions. *PLoS Biol.* 4(11), e368.
- 5 4) Bailey, S., Clokie, M.R., Millard, A., Mann, N.H., 2004. Cyanophage infection and  
6 photoinhibition in marine cyanobacteria. *Res. Microbiol.* 155, R720-R725.
- 7 5) Bannert, N., Kurth, R., 2004. Retroelements and the human genome: new perspectives  
8 on an old relation. *Proc. Natl. Acad. Sci. U S A* 101 Suppl 2, R14572-R14579.
- 9 6) Bapteste, E., Susko, E., Leigh, J., MacLeod, D., Charlebois, R. L., Doolittle, W. F.  
10 2005. Do orthologous gene phylogenies really support tree-thinking? *BMC Evol. Biol.*  
11 5, R33.
- 12 7) Barbieri, M., 2001. *The Organic Codes. The birth of semantic biology.* PeQuod.  
13 Ancona.
- 14 8) Barbieri, M., 2007. (Ed.) *Introduction to Biosemiotics.* Springer, Dordrecht.
- 15 9) Beiko, R. G., Doolittle, W. F., Charlebois, R. L., 2008. The impact of reticulate  
16 evolution on genome phylogeny. *Syst. Biol.* 57, R844-R856.
- 17 10) Bell, P.J.L., 2001. Viral Eukaryogenesis: Was the ancestor of the nucleus a complex  
18 DNAVirus? *J. Mol. Evol.* 53, R251-R256.
- 19 11) Bell, P.J.L., 2006. Sex and the eukaryotic cell cycle is consistent with a viral ancestry  
20 for the eukaryotic nucleus. *J. Theor. Biol.* 243, 54-63.
- 21 12) Bench, S. R., Hanson, T. E., Williamson, K. E., Ghosh, D., Radosovich, M., Wang, K.,  
22 Wommack, K. E., 2007. Metagenomic characterization of Chesapeake Bay  
23 viroplankton. *Appl. Environ. Microbiol.* 73(23), R7629-R7641.
- 24 13) Bielaszewska, M., Dobrindt, U., Gartner, J., Gallitz, I., Hacker, J., Karch, H., Muller,  
25 D., Schubert, S., Alexander Schmidt, M., Sorsa, L. J., Zdziarski, J., 2007. Aspects of  
26 genome plasticity in pathogenic *Escherichia coli*. *Int. J. Med. Microbiol.* 297(7-8),  
27 R625-R639.
- 28 14) Binnewies, T. T., Motro, Y., Hallin, P. F., Lund, O., Dunn, D., La, T., Hampson, D. J.,  
29 Bellgard, M., Wassenaar, T. M., Ussery, D. W., 2006. Ten years of bacterial genome  
30 sequencing: comparative-genomics-based discoveries. *Funct. Integr. Genomics* 6(3),  
31 R165-R185.
- 32 15) Bishop, A. L., Baker, S., Jenks, S., Fookes, M., Gaora, P. O., Pickard, D., Anjum, M.,  
33 Farrar, J., Hien, T. T., Ivens, A., Dougan, G., 2005. Analysis of the hypervariable region  
34 of the *Salmonella enterica* genome associated with tRNA(leuX). *J. Bacteriol.* 187(7),  
35 R2469-R2482.
- 36 16) Blackburn, E.H., 2006. Telomerase RNA. In: Gesteland, R.F., Cech, T.R., Atkins, J.F.  
37 (Eds.), *The RNA World.* 3rd Edition. Cold Spring Harbor Laboratory Press, New York,  
38 pp. 419-436.
- 39 17) Breaker, R.R., 2006. Riboswitches and the RNA World. In: Gesteland, R.F., Cech, T.R.,  
40 Atkins, J.F. (Eds.), *The RNA World.* 3rd Edition. Cold Spring Harbor Laboratory Press,  
41 New York, pp. 89-107.
- 42 18) Brett, D., Pospisil, H., Valcarcel, J., Reich, J., Bork, P., 2002. Alternative splicing and  
43 genome complexity. *Nat. Gen.* 30, R29-R30.
- 44 19) Brok-Volchanskaya, V. S., Kadyrov, F. A., Sivogrivov, D. E., Kolosov, P. M., Sokolov,  
45 A. S., Shlyapnikov, M. G., Kryukov, V. M., Granovsky, I. E., 2008. Phage T4 SegB  
46 protein is a homing endonuclease required for the preferred inheritance of T4 tRNA  
47 gene region occurring in co-infection with a related phage. *Nucleic Acids Res.* 36(6),  
48 R2094-R2105.
- 49 20) Bromham, L., 2002. The human zoo: Endogenous retroviruses in the human genome.  
50 *Trends Ecol. Evol.* 17, R91-R97.

- 1 21) Brussow, H., 2007. Bacteria between protists and phages: from antipredation strategies  
2 to the evolution of pathogenicity. *Mol. Microbiol.* 65, R583-R589.
- 3 22) Brussow, H., 2009. The not so universal tree of life or the place of viruses in the living  
4 world. *Trans. R. Soc. B* 364, R2263–R2274.
- 5 23) Brussow, H., Desiere, F., 2001. Comparative phage genomics and the evolution of  
6 Siphoviridae: insights from dairy phages. *Mol. Microbiol.* 39, R213-R222.
- 7 24) Brussow, H., Canchaya, C., Hardt, W.D., 2004. Phages and the evolution of bacterial  
8 pathogens: from genomic rearrangements to lysogenic conversion. *Microbiol. Mol.*  
9 *Biol. Rev.* 68, R560-R602.
- 10 25) Bryan, M. J., Burroughs, N. J., Spence, E. M., Clokie, M. R., Mann, N. H., Bryan, S. J.,  
11 2008. Evidence for the intense exchange of MazG in marine cyanophages by horizontal  
12 gene transfer. *PLoS ONE* 3(4), e2048.
- 13 26) Buzdin, A., Khodosevich, K., Mamedov, I., Vinogradova, T., Lebedev, Y., Hunsmann,  
14 G., Sverdlov, E., 2002. A technique for genome-wide identification of differences in the  
15 interspersed repeats integrations between closely related genomes and its application to  
16 detection of human-specific integrations of HERV-K LTRs. *Genomics* 79, R413-R422.
- 17 27) Campbell, A., 2007. Phage Integration and Chromosome Structure. A Personal History.  
18 *Annu. Rev. Genet.* 41, R1-R11.
- 19 28) Canchaya, C., Fournous, G., Brussow, H., 2004. The impact of prophages on bacterial  
20 chromosomes. *Mol. Microbiol.* 53(1), R9-R18.
- 21 29) Canchaya, C., Fournous, G., Chibani-Chennoufi, S., Dillmann, M. L., Brussow, H.,  
22 2003a. Phage as agents of lateral gene transfer. *Curr. Opin. Microbiol.* 6(4), R417-  
23 R424.
- 24 30) Canchaya, C., Proux, C., Fournous, G., Bruttin, A., Brussow, H., 2003b. Prophage  
25 genomics. *Microbiol. Mol. Biol. Rev.* 67(2), R238-R276, table of contents.
- 26 31) Capy, P., Langin, T., Higuier, D., Maurer, P., Bazin, C. 1997. Do the integrases of LTR-  
27 retrotransposons and class II element transposases have a common ancestor? *Genetica*  
28 100, R63-R72.
- 29 32) Cech, T.R., Moras, D., Nagai, K., Williamson, J.R., 2006. The RNP World. In:  
30 Gesteland, R.F., Cech, T.R., Atkins, J.F. (Eds.), *The RNA World*. 3rd Edition. Cold  
31 Spring Harbor Laboratory Press, New York, pp. 309-326.
- 32 33) Chebanov, S., 1994. Man as Participant to Natural Creation. *Enlogue and Ideas of*  
33 *Hermeneutics in Biology*. *Rivista di Biologia* 87, R39-R55.
- 34 34) Cheetham, B. F., Katz, M. E., 1995. A role for bacteriophages in the evolution and  
35 transfer of bacterial virulence determinants. *Mol. Microbiol.* 18(2), R201-R208.
- 36 35) Chen, I.A., Hanczyc, M.M., Sazani, P.L., Szostak, J.W., 2006. Protocells: Genetic  
37 Polymers Inside Membrane Vesicles. In: Gesteland, R.F., Cech, T.R., Atkins, J.F.  
38 (Eds.), *The RNA World*. 3rd Edition. Cold Spring Harbor Laboratory Press, New York,  
39 pp. 58-88.
- 40 36) Chibani-Chennoufi, S., Canchaya, C., Bruttin, A., Brussow, H., 2004. Comparative  
41 genomics of the T4-Like *Escherichia coli* phage JS98: implications for the evolution of  
42 T4 phages. *J. Bacteriol.* 186(24), R8276-R8286.
- 43 37) Cole, K.M., Sheath, R.G. 1990. *Biology of the red algae*. Cambridge University Press,  
44 New York.
- 45 38) Coleman, M.L., Sullivan, M.B., Martiny, A.C., Steglich, C., Barry, K., Delong, E.F.,  
46 Chisholm, S.W., 2006. Genomic islands and the ecology and evolution of  
47 *Prochlorococcus*. *Science* 311, R1768-R1770.
- 48 39) Comeau, A. M., Krisch, H. M., 2008. The capsid of the T4 phage superfamily: the  
49 evolution, diversity, and structure of some of the most prevalent proteins in the  
50 biosphere. *Mol. Biol. Evol.* 25(7), R1321-R1332.

- 1 40) Chomsky N. (2004) Biolinguistics and the Human Capacity. Delivered at MTA,  
2 Budapest, May 17, 2004.
- 3 41) Coffin, J.M., Hughes, A.H., Varmus, H.E., 1997. Retroviruses. Cold Spring Harbor  
4 Laboratory Press, New York.
- 5 42) Cortez, D., Forterre, P., Gribaldo, S., 2009. A hidden reservoir of integrative elements is  
6 the major source of recently acquired foreign genes and ORFans in archaeal and  
7 bacterial genomes. *Genome Biol.* doi:10.1186/gb-2009-10-6-r65
- 8 43) DeFilippis, V., Villarreal, L.P., Salzberg, S.L., Eisen, J.A., 2001. Lateral Gene Transfer  
9 or Viral Colonization? *Science* 293, 1048a.
- 10 44) DeFilippis, V.R., Villarreal, L.P., 2001. Virus Evolution. In *Fundamental virology*,  
11 Knipe DM, Howley PM, Griffin DE (Eds.), 4th Edition, Lippincott Williams &  
12 Wilkins, Philadelphia, pp. 353-370.
- 13 45) Derelle, E., Ferraz, C., Escande, M. L., Eychenie, S., Cooke, R., Piganeau, G.,  
14 Desdevises, Y., Bellec, L., Moreau, H., Grimsley, N., 2008. Life-cycle and genome of  
15 OtV5, a large DNA virus of the pelagic marine unicellular green alga *Ostreococcus*  
16 *tauri*. *PLoS ONE* 3(5), e2250.
- 17 46) Derelle, E., Ferraz, C., Rombauts, S., Rouze, P., Worden, A. Z., Robbens, S., Partensky,  
18 F., Degroeve, S., Echeynie, S., Cooke, R., Saeys, Y., Wuyts, J., Jabbari, K., Bowler, C.,  
19 Panaud, O., Piegu, B., Ball, S. G., Ral, J. P., Bouget, F. Y., Piganeau, G., De Baets, B.,  
20 Picard, A., Delseny, M., Demaille, J., Van de Peer, Y., Moreau, H., 2006. Genome  
21 analysis of the smallest free-living eukaryote *Ostreococcus tauri* unveils many unique  
22 features. *Proc. Natl. Acad. Sci. U S A* 103(31), R11647-R11652.
- 23 47) Desiere, F., Lucchini, S., Canchaya, C., Ventura, M., Brussow, H., 2002. Comparative  
24 genomics of phages and prophages in lactic acid bacteria. *Antonie Van Leeuwenhoek*  
25 82(1-4), R73-R91.
- 26 48) Desiere, F., McShan, W. M., van Sinderen, D., Ferretti, J. J., Brussow, H., 2001.  
27 Comparative genomics reveals close genetic relationships between phages from dairy  
28 bacteria and pathogenic Streptococci: evolutionary implications for prophage-host  
29 interactions. *Virology* 288(2), R325-R341.
- 30 49) Dinsdale, E. A., Edwards, R. A., Hall, D., Angly, F., Breitbart, M., Brulc, J. M., Furlan,  
31 M., Desnues, C., Haynes, M., Li, L., McDaniel, L., Moran, M. A., Nelson, K. E.,  
32 Nilsson, C., Olson, R., Paul, J., Brito, B. R., Ruan, Y., Swan, B. K., Stevens, R.,  
33 Valentine, D. L., Thurber, R. V., Wegley, L., White, B. A., Rohwer, F., 2008.  
34 Functional metagenomic profiling of nine biomes. *Nature* 452(7187), R629-R632.
- 35 50) Domingo, E., Parrish, C.R., Holland, J.J. 2008. Origin and Evolution of Viruses, 2nd  
36 Edition, Academic Press, San Diego.
- 37 51) Dunlap, K.A., Palmarini, M., Varela, M., Burghardt, R.C., Hayashi, K., Farmer, J.L.,  
38 Spencer, T.E., 2006. Endogenous retroviruses regulate periimplantation placental  
39 growth and differentiation. *Proc. Natl. Acad. Sci. U S A* 103, R14390-R14395.
- 40 52) Dupressoir, A., Marceau, G., Vernochet, C., Benit, L., Kanellopoulos, C., Sapin, V.,  
41 Heidmann, T., 2005. Syncytin-A and syncytin-B, two fusogenic placenta-specific  
42 murine envelope genes of retroviral origin conserved in Muridae. *Proc. Natl. Acad. Sci.*  
43 *U S A* 102, R725-R730.
- 44 53) Eigen, M., Winkler, R., 1975. *Das Spiel – Naturgesetze steuern den Zufall*. Piper,  
45 München.
- 46 54) Filée, J., Forterre, P., 2005. Viral proteins functioning in organelles: a cryptic origin  
47 *Trends Microbiol.* 13, R510-R513.
- 48 55) Filee, J., Tetart, F., Suttle, C. A., Krisch, H. M., 2005. Marine T4-type bacteriophages, a  
49 ubiquitous component of the dark matter of the biosphere. *Proc. Natl. Acad. Sci. U S A*  
50 102(35), R12471-R12476.

- 1 56) Florkin, M., 1974. Concepts of molecular biosemiotics and of molecular evolution. In:  
2 Florkin, M., Stotz, E.H., (Eds.), *Comparative Biochemistry*, Elsevier, Amsterdam.
- 3 57) Foley, S., Bruttin, A., Brussow, H., 2000. Widespread distribution of a group I intron  
4 and its three deletion derivatives in the lysin gene of *Streptococcus thermophilus*  
5 bacteriophages. *J. Virol.* 74(2), R611-R618.
- 6 58) Forterre, P., 2001. Genomics and early cellular evolution. The origin of the DNA world.  
7 *CR. Acad. Sci.* III 324, R1067-R1076.
- 8 59) Forterre, P., 2002. The origin of DNA genomes and DNA replication proteins. *Cur.*  
9 *Opin. Microbiol.* 5, R525-R532.
- 10 60) Forterre, P., 2005. The two ages of the RNA world, and the transition to the DNA  
11 world: a story of viruses and cells. *Biochimie* 87, R793-R803.
- 12 61) Forterre, P., 2006 a. The origin of viruses and their possible roles in major evolutionary  
13 transitions. *Virus Res.* 117, R5-R16.
- 14 62) Forterre, P., 2006 b. Three RNA cells for ribosomal lineages and three DNA viruses to  
15 replicate their genomes: a hypothesis for the origin of cellular domain. *Proc. Natl. Acad.*  
16 *Sci. USA* 103, R3669-R3674.
- 17 63) Forterre, P., Gadelle, D., 2009. Phylogenomics of DNA topoisomerases: their origin and  
18 putative roles in the emergence of modern organisms. *Nucleic Acids Res.* 37, R679-  
19 R692.
- 20 64) Forterre, P. 2009. Defining life: the virus viewpoint. *Orig Life Evol Biosph* (in press)
- 21 65) Frost, L.S., Laplae, R., Summers, A.O., Toussaint, A., 2005. Mobile Genetic Elements:  
22 The Agents of Open Source Evolution. *Nat. Rev. Microbiol.* 3, R722-R732.
- 23 66) Gabora, L., 2006. Self-other organization: Why early life did not evolve through natural  
24 selection. *J. Theor. Biol.* 241, R443-R450.
- 25 67) Gao, X., Havecker, E.R., Baranov, P.V., Atkins, J.F., Voytas, D.F. 2003. Translational  
26 recoding signals between gag and pol in diverse LTR retrotransposons. *RNA* 9, R1422-  
27 R1430.
- 28 68) Gelfand, M.S., Koonin, E.V., 1997. Avoidance of palindromic words in bacterial and  
29 archaeal genomes: a close connection with restriction enzymes. *Nucleic Acids Res.* 25,  
30 R2430-R2439.
- 31 69) Gimona, M., 2006. Protein linguistics - a grammar for modular protein assembly? *Nat.*  
32 *Rev. Mol. Cell Biol.* 7, R68-R73.
- 33 70) Goff, L.J., Ashen, J., Moon, D., 1997. The evolution of parasites from their hosts: A  
34 case study in the parasitic red algae. *Evolution* 51, R1068-R1078.
- 35 71) Greener, M., 2008. It's life but just as we know it. *EMBO Rep* 9, R1067-R1069.
- 36 72) Griffith, D., Voisset, C., 2008. Human endogenous retroviruses: from ancestral  
37 pathogens to bona fide genes. *Microbiol. Tod.* 2, R32-R33.
- 38 73) Häring, M., Vestergaard, G., Rachel, R., Chen, L., Garrett, R.A., Prangishvili, D., 2005.  
39 *Virology: independent virus development outside the host.* *Nature* 436, R1101-R1102.
- 40 74) Hendrix, R.W., 2009. Jumbo bacteriophages. *Curr. Top. Microbiol. Immunol.* 328,  
41 R229-R240.
- 42 75) Hendrix, R.W., Hatfull, G.F., Smith, M.C., 2003. Bacteriophages with tails: chasing  
43 their origins and evolution. *Res. Microbiol.* 154, R253-R257.
- 44 76) Herniou, E.A., Jehle, J.A., 2007. Baculovirus phylogeny and evolution. *Curr. drug targ.*  
45 8, R1043-R1050.
- 46 77) Herniou, E.A., Luque, T., Chen, X., Vlak, J.M., Winstanley, D., Cory, J.S., O'Reilly,  
47 D.R., 2001. Use of whole genome sequence data to infer baculovirus phylogeny. *J.*  
48 *Virol.* 75, R8117-R8126.
- 49 78) Ho, C.K., Shuman, S. 2002. Bacteriophage T4 RNA ligase 2 (gp24.1) exemplifies a  
50 family of RNA ligases found in all phylogenetic domains. *Proc. Natl. Acad. Sci. U S A*  
51 99, R12709-R12714.

- 1 79) Hoffmeyer, J. 1996. Signs of Meaning in the Universe. Indiana University Press,  
2 Bloomington.
- 3 80) Hougland, J.L, Piccirilli, J.A., Forconi, M., Lee, J., Herschlag, D., 2006. How the Group  
4 I Introns Works: A Case Study of RNA Structure and Function. In: Gesteland, R.F.,  
5 Cech, T.R., Atkins, J.F. (Eds.), The RNA World. 3rd Edition. Cold Spring Harbor  
6 Laboratory Press, New York, pp. 133-205.
- 7 81) Hughes, J.F., Coffin, J.M., 2001. Evidence for genomic rearrangements mediated by  
8 human endogenous retroviruses during primate evolution. *Nat. Genet.* 29, R487-R489.
- 9 82) Iguchi, A., Thomson, N. R., Ogura, Y., Saunders, D., Ooka, T., Henderson, I. R., Harris,  
10 D., Asadulghani, M., Kurokawa, K., Dean, P., Kenny, B., Quail, M. A., Thurston, S.,  
11 Dougan, G., Hayashi, T., Parkhill, J., Frankel, G., 2009. Complete genome sequence  
12 and comparative genome analysis of enteropathogenic *Escherichia coli* O127:H6 strain  
13 E2348/69. *J. Bacteriol.* 191(1), R347-R354.
- 14 83) Iwasaki, W., Takagi, T., 2009. Rapid pathway evolution facilitated by horizontal gene  
15 transfers across prokaryotic lineages. *PLoS. Genet.* 5(3), e1000402.
- 16 84) Jablonka, E., Lamb, M.J., 1989. The inheritance of acquired epigenetic variations. *J.*  
17 *Theor. Biol.* 139, R69-R83.
- 18 85) Jablonka, E., Lamb, M.J., 2002. The changing concept of Epigenetics. In: Van  
19 Speybroeck L., Van deVijver, G., Waele, D.D. (Eds.), From Epigenesis to Epigenetics.  
20 The Genome in Context. *Ann. NYAS.* 981, R82-R96.
- 21 86) Jablonka, E., Lamb, M.J., 2006. The evolution of information in the major transitions. *J.*  
22 *Theor. Biol.* 239, R236-R246.
- 23 87) Jaenisch, R., Bird, A., 2003. Epigenetic regulation of gene expression: how the genome  
24 integrates intrinsic and environmental signals. *Nat. Gen. Suppl.* 33, R245-R254.
- 25 88) Jenuwein, T., Allis, C.D., 2001. Translating the histone code. *Science* 293, R1074-  
26 R1080.
- 27 89) Ji, S., 1997. Isomorphism between cell and human languages: molecular biological,  
28 bioinformatic and linguistic implications. *Biosystems* 44, R17-R39.
- 29 90) Ji, S., 1999. The Linguistics of DNA: Words, Sentences, Grammar, Phonetics and  
30 Semantics. *Ann. NYAS.* 870, R411-R417.
- 31 91) Jordan, I.K., Rogozin, I.B., Glazko, G.V., Koonin, E.V. 2003. Origin of a substantial  
32 fraction of human regulatory sequences from transposable elements. *Trends Genet.* 19,  
33 R68-R72.
- 34 92) Jurka, J., Kapitonov, V.V., Kohany, O., Jurka, M.V., 2007. Repetitive Sequences in  
35 Complex Genomes: Structure and Evolution. *Annu. Rev. Genomics Hum. Genet.* 8,  
36 R241-R259.
- 37 93) Kadyrov, F. A., Shlyapnikov, M. G., Kryukov, V. M., 1997. A phage T4 site-specific  
38 endonuclease, SegE, is responsible for a non-reciprocal genetic exchange between T-  
39 even-related phages. *FEBS. Lett.* 415(1), R75-R80.
- 40 94) Ke, A., Doudna, J.A., 2006. Catalytic Strategies of Self-Cleaving Ribozymes: Relics of  
41 an RNA World? In: Gesteland, R.F., Cech, T.R., Atkins, J.F. (Eds.), The RNA World.  
42 3rd Edition. Cold Spring Harbor Laboratory Press, New York, pp. 109-131.
- 43 95) Khodosevich, K., Lebedev, Y., Sverdlov, E., 2002. Endogenous retroviruses and human  
44 evolution. *Comp. Func. Gen.* 3, R494-R498.
- 45 96) Kim, T.H., Jeon, Y.J., Yi, J.M., Kim, D.S., Huh, J.W., Hur, C.G., Kim, H.S., 2004. The  
46 distribution and expression of HERV families in the human genome. *Mol. Cells* 18,  
47 R87-R93.
- 48 97) Kirsch, P., Jores, J., Wieler, L.H., 2004. Plasticity of bacterial genomes: pathogenicity  
49 islands and the locus of enterocyte effacement (LEE). *Berliner und Muenchener*  
50 *tierarztliche Wochenschrift* 117, R116-R129.

- 1 98) Klein, J., 2004. Did viruses play a part in the origin of the adaptive immune system?  
2 *Folia Biol. (Praha)* 50(3-4), R87-R92.
- 3 99) Koonin, E.V., Senkevich, T.G., Dolja, V.V., 2006. The ancient virus world and  
4 evolution of cells. *Biol. Dir.* 1, 29.
- 5 100) Koonin, E. V., Wolf, Y. I., 2008. Genomics of bacteria and archaea: the emerging  
6 dynamic view of the prokaryotic world. *Nucleic Acids Res.* 36(21), R6688-R6719.
- 7 101) Krupovic, M., Bamford, D. H., 2007. Putative prophages related to lytic tailless marine  
8 dsDNA phage PM2 are widespread in the genomes of aquatic bacteria. *BMC Gen.* 8,  
9 R236.
- 10 102) Lescat, M., Calteau, A., Hoede, C., Barbe, V., Touchon, M., Rocha, E., Tenailon, O.,  
11 Medigue, C., Johnson, J. R., Denamur, E., 2009. A module located at a chromosomal  
12 integration hotspot is responsible for the multiple-resistance of a reference strain from  
13 *Escherichia coli* clonal group A. *Antimicrob. Agents Chemother.*  
14 doi:10.1128/AAC.00123-09
- 15 103) Lindell, D., Jaffe, J. D., Coleman, M. L., Futschik, M. E., Axmann, I. M., Rector, T.,  
16 Kettler, G., Sullivan, M. B., Steen, R., Hess, W. R., Church, G. M., Chisholm, S. W.,  
17 2007. Genome-wide expression dynamics of a marine virus and host reveal features of  
18 co-evolution. *Nature* 449, R83-R86.
- 19 104) Liu, J., Glazko, G., Mushegian, A., 2006. Protein repertoire of double-stranded DNA  
20 bacteriophages. *Virus Res* 117, R68-R80.
- 21 105) Liu, Q., Belle, A., Shub, D.A., Belfort, M., Edgell, D.R., 2003. SegG endonuclease  
22 promotes marker exclusion and mediates co-conversion from a distant cleavage site. *J.*  
23 *Mol. Biol.* 334, R13-R23.
- 24 106) Lopez, P., Baptiste, E., 2009. Molecular phylogeny: reconstructing the forest. *C. R.*  
25 *Biol.* 332(2-3), R171-R182.
- 26 107) Maita, N., Anzai, T., Aoyagi, H., Mizuno, H., Fujiwara, H., 2004. Crystal Structure of  
27 the Endonuclease Domain Encoded by the Telomere-specific Long Interspersed Nuclear  
28 Element, TRAS1. *J. Biol. Chem.* 279, R41067-R41076.
- 29 108) Margulis, L., 1996. Archaeal-eubacterial mergers in the origin of Eukarya: Phylogenetic  
30 classification of life. *Proc. Natl. Acad. Sci. USA* 93, R1071-R1076.
- 31 109) Margulis, L. 1999. *Die andere Evolution*. Spektrum Akademischer Verlag, Heidelberg
- 32 110) Margulis, L., 2004. Serial endosymbiotic theory (SET) and composite individuality.  
33 Transition from bacterial to eukaryotic genomes. *Microbiol. Tod.* 31, R173-R174.
- 34 111) Margulis, L., Dolan, M.F., Guerrero, R., 2000. The chimeric eukaryote: origin of the  
35 nucleus from the karyomastigont in amitochondriate protists. *Proc. Natl. Acad. Sci.*  
36 *USA* 97, R6954-R6959.
- 37 112) Margulis, L., Sagan, S., 2002. *Acquiring Genomes. A Theory of the Origin of Species.*  
38 Basic Books, New York
- 39 113) Markos, A., 2002. *Readers of the Book of Life*. Oxford University Press, Oxford, New  
40 York.
- 41 114) Matera, A.G., Terns, R.M., Terns, M.P., 2007. Non-coding RNAs: Lessons from the  
42 small nuclear and small nucleolar RNAs. *Nat. Rev. Mol. Cell Biol.* 8, R209-R220.
- 43 115) Mattick, J.S., Gagen, M.J., 2001. The Evolution of Controlled Multitasked Gene  
44 Networks: The Role of Introns and Other Noncoding RNAs in the Development of  
45 Complex Organisms. *Mol. Biol. Evol.* 18, R1611-R1630.
- 46 116) Mattick, J.S., 2001. Non-coding RNAs: the architects of eukaryotic complexity. *EMBO*  
47 *Rep.* 2, R986-R991.
- 48 117) Mattick, J.S., 2003. Challenging the dogma: the hidden layer of noncoding RNAs in  
49 complex organisms. *BioEssays* 25, R930-R939.
- 50 118) Mattick, J.S., 2005. Das verkannte Genom-Programm. *Spektrum der Wissenschaft* 3,  
51 R62-R69.

- 1 119) Mayer, J., Meese, E., 2005. Human endogenous retroviruses in the primate lineage and  
2 their influence on host genomes. *Cytogenet Genome Res.* 110, R448-R456.
- 3 120) Mehta, P., Casjens, S., Krishnaswamy, S., 2004. Analysis of the lambdoid prophage  
4 element  $\epsilon 14$  in the *E. coli* K-12 genome. *BMC microbiology* 4, R4.
- 5 121) Mi, S., Lee, X., Li, X., Veldman, G.M., Finnerty, H., Racie, L., LaVallie, E., Tang,  
6 X.Y., Edouard, P., Howes, S., Keith, J.C. Jr., McCoy, J.M., 2000. Syncytin is a captive  
7 retroviral envelope protein involved in human placental morphogenesis. *Nature* 403,  
8 R785-R789.
- 9 122) Monier, A., Claverie, J.M., Ogata, H., 2007. Horizontal gene transfer and nucleotide  
10 compositional anomaly in large DNA viruses. *BMC. Gen.* 8, R456.
- 11 123) Monier, A., Claverie, J.M., Ogata, H., 2008. Taxonomic distribution of large DNA  
12 viruses in the sea. *Genome Biol* 9(7), R106.
- 13 124) Moore, P.B., Steitz, T.A., 2006. The Roles of RNA in the Synthesis of Protein. In:  
14 Gesteland, R.F., Cech, T.R., Atkins, J.F. (Eds.), *The RNA World*. 3rd Edition. Cold  
15 Spring Harbor Laboratory Press, New York, pp. 257-285.
- 16 125) Moreira, D., Lopez-Garcia, P., 2009. Ten reasons to exclude viruses from the tree of  
17 life. *Nat. Rev. Microbiol.* 7, R306-R311.
- 18 126) Nicholson, W.L., Munakata, N., Horneck G., Melosh, H.J., Setlow, P., 2000.  
19 Resistance of *Bacillus* endospores to extreme terrestrial and extraterrestrial  
20 environments. *Microbiol. Mol. Biol.Rev.* 64, R548-R572.
- 21 127) Nolan, J.M., Petrov, V., Bertrand, C., Krisch, H.M., Karam, J.D., 2006. Genetic  
22 diversity among five T4-like bacteriophages. *Viol. J.* 3, R30.
- 23 128) Noller, H.F., 2006. Evolution of Ribosomes and Translation. In: Gesteland, R.F., Cech,  
24 T.R., Atkins, J.F. (Eds.), *The RNA World*. 3rd Edition. Cold Spring Harbor Laboratory  
25 Press, New York, pp. 287-307.
- 26 129) Odintsova, M.S., Yurina, N.P., 2000. RNA Editing in Plant Chloroplasts and  
27 Mitochondria. *Fisiol. Rast* 37, R307-R320.
- 28 130) Odintsova, M.S., Yurina, N.P., 2005. Genomics and Evolution of Cellular Organelles.  
29 *Russian J. Genet.* 41, R957-R967.
- 30 131) Ogata, H., Claverie, J. M., 2007. Unique genes in giant viruses: regular substitution  
31 pattern and anomalously short size. *Genome Res* 17(9), R1353-R1361.
- 32 132) Ogura, Y., Ooka, T., Asadulghani, Terajima, J., Nougayrede, J. P., Kurokawa, K.,  
33 Tashiro, K., Tobe, T., Nakayama, K., Kuhara, S., Oswald, E., Watanabe, H., Hayashi,  
34 T., 2007. Extensive genomic diversity and selective conservation of virulence-  
35 determinants in enterohemorrhagic *Escherichia coli* strains of O157 and non-O157  
36 serotypes. *Genome Biol.* 8(7), R138.
- 37 133) Oshima, K., Toh, H., Ogura, Y., Sasamoto, H., Morita, H., Park, S. H., Ooka, T., Iyoda,  
38 S., Taylor, T. D., Hayashi, T., Itoh, K., Hattori, M. 2008. Complete genome sequence  
39 and comparative analysis of the wild-type commensal *Escherichia coli* strain SE11  
40 isolated from a healthy adult. *DNA Res.* 15(6), R375-R386.
- 41 134) Palenik, B., Grimwood, J., Aerts, A., Rouze, P., Salamov, A., Putnam, N., Dupont, C.,  
42 Jorgensen, R., Derelle, E., Rombauts, S., Zhou, K., Otilar, R., Merchant, S. S., Podell,  
43 S., Gaasterland, T., Napoli, C., Gendler, K., Manuell, A., Tai, V., Vallon, O., Piganeau,  
44 G., Jancek, S., Heijde, M., Jabbari, K., Bowler, C., Lohr, M., Robbens, S., Werner, G.,  
45 Dubchak, I., Pazour, G. J., Ren, Q., Paulsen, I., Delwiche, C., Schmutz, J., Rokhsar, D.,  
46 Van de Peer, Y., Moreau, H., Grigoriev, I. V., 2007. The tiny eukaryote *Ostreococcus*  
47 provides genomic insights into the paradox of plankton speciation. *Proc. Natl. Acad.*  
48 *Sci. U S A* 104(18), R7705-R7710.
- 49 135) Paul, A.V., van Boom, J.H., Filippov, D., Wimmer, E., 1998. Protein-primed RNA  
50 synthesis by purified poliovirus RNA polymerase. *Nature* 393, R280-R284.

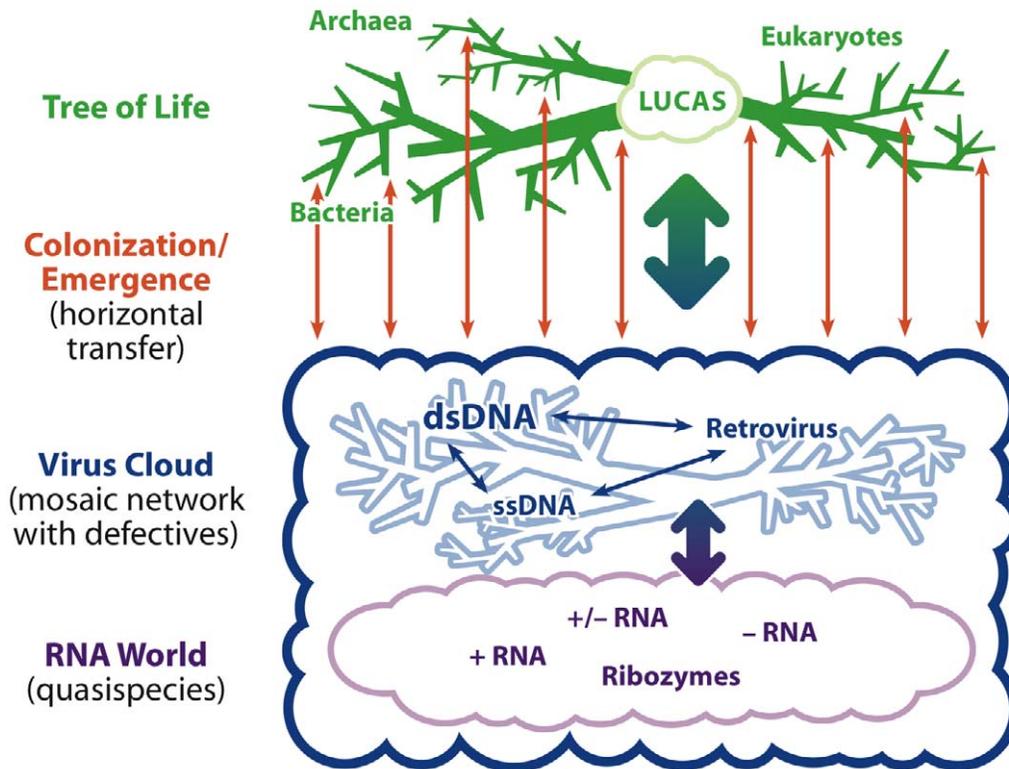
- 1 136) Polavarapu, N., Bowen, N.J., McDonald, J.F., 2006. Identification, characterization and  
2 comparative genomics of chimpanzee endogenous retroviruses. *Genome Biol* 7, R51.
- 3 137) Popov, O., Degal, D.M., Trifonov, E.N., 1996. Linguistic complexity of protein  
4 sequences as compared to texts of human languages. *Biosystems* 38, R65-R74.
- 5 138) Prangishvili, D., Vestergaard, G., Haring, M., Aramayo, R., Basta, T., Rachel, R.,  
6 Garrett, R.A., 2006. Structural and genomic properties of the hyperthermophilic  
7 archaeal virus ATV with an extracellular stage of the reproductive cycle. *J. Mol.Biol.*  
8 359, R1203-R1216
- 9 139) Pride, D.T., Schoenfeld, T., 2008. Genome signature analysis of thermal virus  
10 metagenomes reveals Archaea and thermophilic signatures. *BMC Genomics* 9, R420.
- 11 140) Pyle, A.M., Lambowitz, A.M., 2006. Group II Introns: Ribozymes That Splice RNA  
12 and Invade DNA. In: Gesteland RF, Cech TR, Atkins JF (Eds) *The RNA World*. 3rd  
13 Edition. Cold Spring Harbor Laboratory Press, New York, pp. 469-506.
- 14 141) Rakin, A., Noelting, C., Schropp, P., Heesemann, J., 2001. Integrative module of the  
15 high-pathogenicity island of *Yersinia*. *Mol. Microbiol.* 39(2), R407-R415.
- 16 142) Rohwer, F., 2003. Global phage diversity. *Cell* 113, R141.
- 17 143) Rohwer, F., Segall, A., Steward, G., Seguritan, V., Breitbart, M., Wolven, F., Azam, F.,  
18 2000. The complete genomic sequence of the marine phage Roseophage SIO1 shares  
19 homology with nonmarine phages. *Limnol. Oceanograph.* 45, R408-R418.
- 20 144) Rusch, D. B., Halpern, A. L., Sutton, G., Heidelberg, K. B., Williamson, S., Yooseph,  
21 S., Wu, D., Eisen, J. A., Hoffman, J. M., Remington, K., Beeson, K., Tran, B., Smith,  
22 H., Baden-Tillson, H., Stewart, C., Thorpe, J., Freeman, J., Andrews-Pfannkoch, C.,  
23 Venter, J. E., Li, K., Kravitz, S., Heidelberg, J. F., Utterback, T., Rogers, Y. H., Falcon,  
24 L. I., Souza, V., Bonilla-Rosso, G., Eguiarte, L. E., Karl, D. M., Sathyendranath, S.,  
25 Platt, T., Bermingham, E., Gallardo, V., Tamayo-Castillo, G., Ferrari, M. R.,  
26 Strausberg, R. L., Neilson, K., Friedman, R., Frazier, M., Venter, J. C., 2007. The  
27 Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern  
28 tropical Pacific. *PLoS Biol* 5(3), e77.
- 29 145) Ryan, F.P., 2004. Human endogenous retroviruses in health and disease: a symbiotic  
30 perspective. *J. R. Soc. Med.* 97, R560-R565.
- 31 146) Ryan, F.P., 2006. Genomic creativity and natural selection. A modern synthesis. *Biol. J.*  
32 *Linn. Soc.* 88, R655-R672.
- 33 147) Ryan, F.P. 2009. *Virovolution*. Harper Collins Publishers, London.
- 34 148) Sandegren, L., Nord, D., Sjoberg, B. M., (2005). SegH and Hef: two novel homing  
35 endonucleases whose genes replace the mobC and mobE genes in several T4-related  
36 phages. *Nucleic Acids Res.* 33(19), R6203-R6213.
- 37 149) Searls, D.B., (2002). The language of genes. *Nature* 420, R211-R217.
- 38 150) Sebeok, T., Umiker-Sebeok, J., 1992. *Biosemiotics: the Semiotic Web 1991*. Mouton de  
39 Gruyter, Berlin.
- 40 151) Shabalina, S.A., Spiridonov, N.A., 2004. The mammalian transcriptome and the  
41 function of non-coding DNA sequences. *Gen. Biol.* 5, 105e.
- 42 152) Sharon, I., Tzahor, S., Williamson, S., Shmoish, M., Man-Aharonovich, D., Rusch, D.  
43 B., Yooseph, S., Zeidner, G., Golden, S. S., Mackey, S. R., Adir, N., Weingart, U.,  
44 Horn, D., Venter, J. C., Mandel-Gutfreund, Y., Beja, O., (2007) Viral photosynthetic  
45 reaction center genes and transcripts in the marine environment. *Isme J.* 1(6), R492-  
46 R501.
- 47 153) Shapiro, J.A., 2002. *Genome Organization and Reorganization in Evolution: Formatting*  
48 *for Computation and Function*. In: v. Speybroeck, L., VandeVijver, G., Waele, DD.  
49 (Eds). *From Epigenesis to Epigenetics. The Genome in Context*. Ann NYAS 981,  
50 R111-R134.

- 1 154) Shapiro, J.A., 2004. A 21st century view of evolution: genome system architecture,  
2 repetitive DNA, and natural genetic engineering. *Gene* 345, R91-R100.
- 3 155) Shapiro, J.A. Sternberg, R., 2005. Why repetitive DNA is essential to genome function.  
4 *Biol. Rev.* 80, R1-R24.
- 5 156) Simpson, L., 2006. Uridine Insertion/Deletion RNA Editing as a Paradigm for Site-  
6 specific Modifications of RNA Molecules. In: Gesteland RF, Cech TR, Atkins JF (Eds)  
7 *The RNA World*. 3rd Edition. Cold Spring Harbor Laboratory Press, New York, pp.  
8 401-417.
- 9 157) Sinkvics, J.G., 2001. The Place of Viruses in the "Tree of Life" , *Acta Microbiol.*  
10 *Immunol. Hung.* 48 (1), R115-R127.
- 11 158) Slotkin, R.K., Martienssen, R., 2007. Transposable elements and the epigenetic regu-  
12 lation of the genome. *Nat. Rev. Gen.* 8, R272-R285.
- 13 159) Sonenshein, A.L., Hoch, J.A., Losick, R., 2002. *Bacillus subtilis* and its closest  
14 relatives: from genes to cells. ASM Press, Washington.
- 15 160) Sternberg, R., 2002. On the Roles of Repetitive DNA Elements in the Context of a  
16 Unified Genomic-Epigenetic System. In: v. Speybroeck, L., VandeVijver, G., Waele,  
17 DD. (Eds). *From Epigenesis to Epigenetics. The Genome in Context. Ann NYAS* 981,  
18 R154-R188.
- 19 161) Sternberg, R., Shapiro, J.A., 2005. How repeated retroelements format genome  
20 function. *Cytogen. Gen. Res.* 110, R108-R116.
- 21 162) Sullivan, M.B., Lindell, D., Lee, J.A., Thompson, L.R., Bielawski, J.P., Chisholm,  
22 S.W., 2006. Prevalence and Evolution of Core Photosystem II Genes in Marine  
23 Cyanobacterial Viruses and Their Hosts. *PLoS Biol.* 4, e234.
- 24 163) Sverdlov, E.D., 2000. Retroviruses and primate evolution. *Bioessays* 22, R161-R171.
- 25 164) Taft, R.J., Mattick, J.S., 2004. Increasing biological complexity is positively correlated  
26 with the relative genome-wide expansion of non-protein-coding DNA sequences. *Gen.*  
27 *Biol.* 5, R1.
- 28 165) Tang, R.S., Barton, D.J., Flanagan, J.B., Kirkegaard, K., 1997. Poliovirus RNA  
29 recombination in cell-free extracts. *RNA* 3, R624-R633.
- 30 166) Tettelin, H., Massignani, V., Cieslewicz, M.J., Donati, C., Medini, D., Ward, N.L.,  
31 Angiuoli, S.V., Crabtree, J., Jones, A.L., Durkin, A.S., DeBoy, R.T., Davidsen, T.M.,  
32 Mora, M., Scarselli, M., Ros, I.M., Peterson, J.D., Hauser, C.R., Sundaram, J.P.,  
33 Nelson, W.C., Madupu, R., Brinkac, L.M., Dodson, R.J., Rosovitz, M.J., Sullivan, S.A.,  
34 Daugherty, S.C., Haft, D.H., Selengut, J., Gwinn, M.L., Zhou, L., Zafar, N., Khouri, H.,  
35 Radune, D., Dimitrov, G., Watkins, K., O'Connor, K.J.B., Smith, S., Utterback, T.R.,  
36 White, O., Rubens, C.E., Grandi, G., Madoff, L.C., Kasper, D.L., Telford, J.L., Wessels,  
37 M.R., Rappuoli, R., Fraser, C.M., 2005. Genome analysis of multiple pathogenic  
38 isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". *Proc.*  
39 *Natl. Acad. Sci.* 102, R13950-R13955.
- 40 167) Toothman, P., Herskowitz, I., 1980a. Rex-dependent exclusion of lambdoid phages. I.  
41 Prophage requirements for exclusion. *Virology* 102, R133-R146.
- 42 168) Toothman, P., Herskowitz, I., 1980b. Rex-dependent exclusion of lambdoid phages. II.  
43 Determinants of sensitivity to exclusion. *Virology* 102, R147-R160.
- 44 169) True, H., Berlin, I., Lindquist, S.L., 2004. Epigenetic regulation of translation reveals  
45 hidden genetic variation to produce complex traits. *Nature* 431, R184-R187.
- 46 170) Turner, B.M., 2000. Histone acetylation and an epigenetic code. *BioEssays* 22, R836-  
47 R845.
- 48 171) Turner, B.M., 2002. Cellular Memory and the Histone Code. *Cell* 111, R285-R291.
- 49 172) Tycowski, K.T., Kolev, N.G., Conrad, N.K., Fok, V., Steitz, J.A., 2006. The Ever-  
50 Growing World of Small Nuclear Ribonucleoproteins. In: Gesteland RF, Cech TR,

- 1 Atkins JF (Eds) *The RNA World*. 3rd Edition. Cold Spring Harbor Laboratory Press,  
 2 New York, pp. 327-368.
- 3 173) Van De Vijver, G., Van Speybroeck, L., De Waele, D., 2002. Epigenetics: A Challenge  
 4 for Genetics, Evolution and Development. In: v. Speybroeck, L., Van De Vijver, G.,  
 5 DeWaele, D., (Eds.), *From Epigenesis to Epigenetics. The Genome in Context*. Ann  
 6 NYAS 981, R1-R6.
- 7 174) Van Speybroeck, L., Van de Vijver, G., DeWaele, D. 2002. Preface. In v. Speybroeck,  
 8 L., Van DeVijver, G., De Waele, D. (Eds.), *From Epigenesis to Epigenetics. The*  
 9 *Genome in Context*. Ann NYAS 981, vii.
- 10 175) Vetsigian, K., Woese, C., Goldenfeld, N., 2006. Collective evolution and the genetic  
 11 code. Ann NYAS 103, R10696-R10701.
- 12 176) Villarreal, L.P., 1999. DNA virus contribution to host evolution. In: Domingo, E.,  
 13 Webster, R.G., Holland, J.J. (Eds), *Origin and evolution of viruses*, Academic Press,  
 14 San Diego, pp 391-420.
- 15 177) Villarreal, L.P., 2004. Can Viruses Make Us Humans? Proc. Am. Phil. Soc. 148, R296-  
 16 R323
- 17 178) Villarreal, L.P. 2005. *Viruses and the evolution of life*. American Society for  
 18 Microbiology Press, Washington.
- 19 179) Villarreal, L.P., 2007. Virus-host symbiosis mediated by persistence. *Symbiosis* 44, R1-  
 20 R9.
- 21 180) Villarreal, L.P. 2009. *Origin of Group Identity: Viruses, Addiction and Cooperation*.  
 22 Springer, New York.
- 23 181) Volff, J.N., 2006. Turning junk into gold: domestication of transposable elements and  
 24 the creation of new genes in eukaryotes. *BioEssays* 28, R913-R922.
- 25 182) Wang, Y., Fischle, W., Cheung, W., Jacobs, S., Khorasanizadeh, S., Allis, C.D. 2004.  
 26 *Beyond the Double Helix: Writing and Reading the Histone Code*. In: Bock, G., Goode,  
 27 J. (Eds.), *Reversible Protein Acetylation*. Novartis Foundation 2004, 259, pp. 3-17.
- 28 183) Watanabe, M., Kikawada, T., Minagawa, N., Yukuhiro, F., Okuda, T., 2002.  
 29 Mechanism allowing an insect to survive complete dehydration and extreme  
 30 temperatures. *J. Exp. Biol.* 205, R2799-R2802.
- 31 184) Weigele, P.R., Pope, W.H., Pedulla, M.L., Houtz, J.M., Smith, A.L., Conway, J.F.,  
 32 King, J., Hatfull, G.F., Lawrence, J.G., Hendrix, R.W. 2007. Genomic and structural  
 33 analysis of Syn9, a cyanophage infecting marine Prochlorococcus and Synechococcus.  
 34 *Environ. Microbiol.* 9, R1675-R1695.
- 35 185) Weiner, A.M. 2006. SINEs and LINEs: Troublemakers, Saboteurs, Benefactors,  
 36 Ancestors. In: Gesteland RF, Cech TR, Atkins JF (Eds) *The RNA World*. 3rd Edition.  
 37 Cold Spring Harbor Laboratory Press, New York, pp. 507-534.
- 38 186) Weiss, R.A., 2006. The discovery of endogenous retroviruses. *Retrovirology* 3, R67.
- 39 187) Williamson, S.J., Cary, S.C., Williamson, K.E., Helton, R.R., Bench, S.R., Winget, D.,  
 40 Wommack, K.E., 2008a. Lysogenic virus-host interactions predominate at deep-sea  
 41 diffuse-flow hydrothermal vents. *Isme J.* 2(11), R1112-R1121.
- 42 188) Williamson, S.J., Rusch, D.B., Yooseph, S., Halpern, A.L., Heidelberg, K.B., Glass,  
 43 J.I., Andrews-Pfannkoch, C., Fadrosh, D., Miller, C.S., Sutton, G., Frazier, M., Venter,  
 44 J.C., 2008b. The Sorcerer II Global Ocean Sampling Expedition: metagenomic  
 45 characterization of viruses within aquatic microbial samples. *PLoS ONE.* 3, e1456.
- 46 189) Wittgenstein, L. 1972. *Philosophical Investigations*. Basil & Blackwell, Oxford.
- 47 190) Witzany, G. 1995. From the 'Logic of the Molecular Syntax' to Molecular Pragmatism.  
 48 Explanatory deficits in Manfred Eigen's concept of language and communication. *Evol*  
 49 *Cog* 1, R148-R168.
- 50 191) Witzany, G. 2000. *Life: The Communicative Structure. A New Philosophy of Biology*.  
 51 Libri Books on Demand, Norderstedt.

- 1 192) Witzany, G. 2005. Natural history of life: History of communication logics and  
2 dynamics. *S.E.E.D. J.* 5, R27-R55.
- 3 193) Witzany, G., 2006a. Serial Endosymbiotic Theory (SET): The Biosemiotic Update.  
4 *Acta Biotheor.* 54, R103-R117.
- 5 194) Witzany, G. 2006b. The Logos of the Bios 1. Contributions to the Foundation of a  
6 three-leveled Biosemiotics. Umweb, Helsinki
- 7 195) Witzany, G. 2006c. Plant Communication from Biosemiotic Perspective. *Plant Signal*  
8 *Behav* 1(4), R169-R178.
- 9 196) Witzany, G. 2007. The Logos of the Bios 2. Bio-Communication. Umweb, Helsinki.
- 10 197) Witzany, G., 2008. Bio-Communication of Bacteria and their Evolutionary Roots in  
11 Natural Genome Editing Competences of Viruses. *Open Evol. J.* 2, R44-R54.
- 12 198) Witzany, G., 2009a. Non-coding RNAs – persistent viral agents as modular tools for  
13 cellular needs. In: Witzany, G. (Ed). *Natural Genetic Engineering and Natural Genome*  
14 *Editing.* Ann NYAS (in press).
- 15 199) Witzany, G., 2009b. *Biocommunication and Natural Genome Editing.* Springer,  
16 Dordrecht.
- 17 200) Xu, Q., Modrek, B., Lee, C., 2002. Genome-wide detection of tissue-specific alternative  
18 splicing in the human transcriptome. *Nuc. Acid Res.* 30, R3754-R3766.
- 19 201) Yang, Z., Kim, J., Zhang, C., Zhang M., Nietfeldt, J., Southward, C.M., Surette, M.G.,  
20 Kachman, S.D., Benson, A.K., 2009. Genomic instability in regions adjacent to highly  
21 conserved pch prophage in *E. coli* O157:H7 generates diversity in expression patterns of  
22 the LEE pathogenicity island. *J. Bacteriol.* doi:10.1128/JB.01738-08.
- 23 202) Yin, Y., Fischer, D., 2008. Identification and investigation of ORFans in the viral world.  
24 *BMC Gen.* 9, R24.
- 25 203) Zanotto, P.M., Gibbs, M.J., Gould, E.A., Holmes, E.C., 1996. A reevaluation of the  
26 higher taxonomy of viruses based on RNA polymerases. *J. Virol.* 70, R6083-R6096.
- 27 204) Zeidner, G., Bielawski, J.P., Shmoish, M., Scanlan, D.J., Sabehi, G., Beja, O., 2005.  
28 Potential photosynthesis gene recombination between *Prochlorococcus* and  
29 *Synechococcus* via viral intermediates. *Environ. Microbiol.* 7, R1505-R1513.
- 30 205) Zhang, H.Y., 2006. The evolution of genomes and language. *EMBO Rep.* 7, R248-  
31 R249.
- 32  
33  
34  
35

1 Fig 1  
2



3  
4

Accepted