Early evolution: prokaryotes, the new kids on the block

Anthony Poole,* Daniel Jeffares, and David Penny

Summary

Prokaryotes are generally assumed to be the oldest existing form of life on earth. This assumption, however, makes it difficult to understand certain aspects of the transition from earlier stages in the origin of life to more complex ones, and it does not account for many apparently ancient features in the eukaryotes. From a model of the RNA world, based on relic RNA species in modern organisms, one can infer that there was an absolute requirement for a high-accuracy RNA replicase even before proteins evolved. In addition, we argue here that the ribosome (together with the RNAs involved in its assembly) is so large that it must have had a prior function before protein synthesis. A model that connects and equates these two requirements (high-accuracy RNA replicase and prior function of the ribosome) can explain many steps in the origin of life while accounting for the observation that eukaryotes have retained more vestiges of the RNA world. The later derivation of prokaryote RNA metabolism and genome structure can be accounted for by the two complementary mechanisms of *r*-selection and thermoreduction. *BioEssays* 21:880–889, 1999.

Introduction

It is widely accepted that an RNA world, with neither DNA nor genetically encoded proteins, was a necessary stage during the origin of life.[1,2] The discovery that RNA acts in both information storage and biological catalysis led Gilbert to coin the term the ''RNA world''[1] for these intermediate stages. RNA is, however, inferior to DNA and protein in information storage and catalysis,[2–5] respectively. If RNA predated protein as a biological catalyst, then during evolution there must have been a transfer of catalysis from RNA-enzymes to ribonucleoproteins (RNPs—catalytic RNA complexed with protein), to proteins: RNA → RNP → protein. The corollary of this is that classes of RNA that are central to metabolism and/or are catalytic (ribozymes) are relics of the RNA world, not recent additions. Although new RNA functions could arise after the advent of genetically encoded protein synthesis, the assumption is that RNA never took over a new function from existing protein enzymes—rather, the direction of change was the reverse. Many RNAs, such as the ribosomal RNAs, were so central to metabolism that these could not have been replaced by proteins—like the genetic code, they can be considered frozen accidents.[6]

The existence of these relics of the RNA world allows a new approach to the study of early evolution. By considering the properties of presumptive RNA relics, we derive a model for the process of early biological evolution, from the origin of encoded protein synthesis to the ''last universal common ancestor'' (LUCA). We present that model here, focusing on the advent of genetically encoded protein synthesis, along with some steps in the evolution of DNA and use these events to predict the nature of the LUCA (Fig. 1). The analysis leads to the conclusion that prokaryotes are not the earliest forms of modern life but rather the specialized derivatives of an earlier form of life that used many more catalytic RNAs, including those now found only in eukaryotes.[7] Throughout this article, ''prokaryotes'' is used in its original sense (absence of a membrane-bound nucleus) and includes eubacteria and archaea, irrespective of their precise evolutionary relationship. (Carl Woese and co-workers[8,9] first recognized that the prokaryotes consist of two domains, the eubacteria and the archaea. Our decision to refer to these as ''prokaryotes'' is due to the limitations of the dataset—RNA relics in

Institute of Molecular BioSciences, Massey University, Palmerston North, New Zealand.
*Correspondence to: Anthony Poole, Department of Molecular Biology, Stockholm University, 106 91 Stockholm, Sweden. E-mail: anthony.poole@molbio.su.se

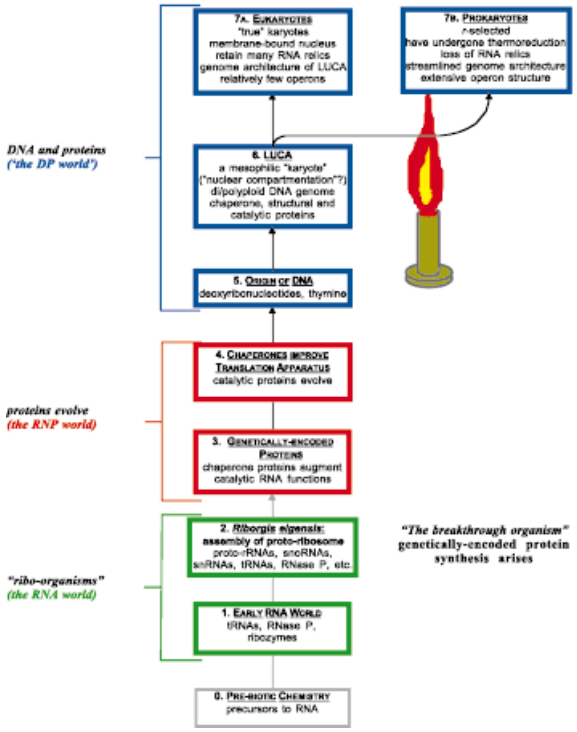

Figure 1. Overview of steps from the prebiotic world to eukaryotes and prokaryotes. A ''timeline'' of events can be deduced, with the RNA world currently being the earliest identifiable form of life. Simpler ribo-organisms must have been present before the relatively complex breakthrough organism, *Riborgis eigensis*. This possessed a relatively complex molecular machinery for processing RNAs, tRNA charging, and ssRNA which gave rise to mRNA. Translation by the ''all-RNA'' prototype ribosome occurred to produce the first proteins—low complexity RNA-chaperones. By the time of LUCA (last universal common ancestor), both catalytic proteins and DNA had arisen. Our model predicts that the LUCA may have been a ''karyote'' with a fragmented, linear genome structure, and RNA metabolism more similar to eukaryotes than to prokaryotes. (This is not the same usage as Rivera and Lake's[23] definition of a karyote based on phylogenetic analysis but implies the likely existence of some form of nuclear and/or nucleolar compartmentation.) The explanation for the origin of prokaryotes is a period of *r*-selection/thermoreduction, which resulted in the loss of a number of heat-labile ssRNAs present in the LUCA. The prokaryotes are thus considered derived rather than ancestral.

modern metabolism and genome architecture only permit the resolution of the eukaryotes from the prokaryotes and cannot address the question of the monophyly of the prokaryotes.) We acknowledge, however, that the phenotypic grouping of ''prokaryotes'' is not necessarily a monophyletic one.

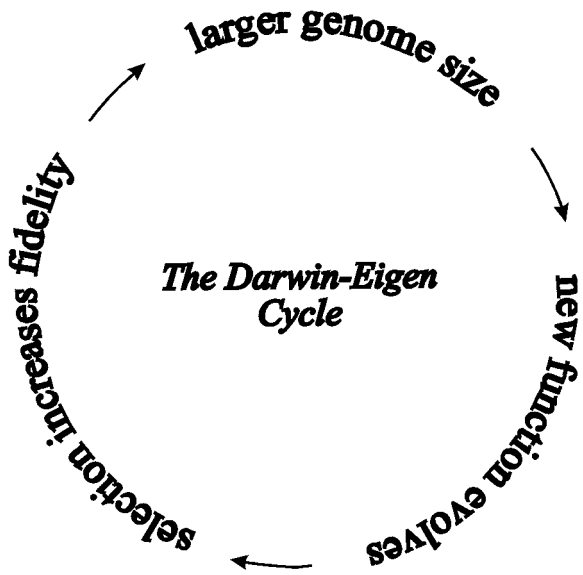

Figure 2. The Darwin-Eigen Cycle is a positive feedback mechanism. Improvements in fidelity of replication increase the upper limit (Eigen limit) on the length of the genome, and allow the evolution of a larger genome size. This allows the evolution of new function, which could further improve replication fidelity, and so on. Thus, successive cycling of this system in early ribo-organisms could, through natural selection of higher fidelity of replication, increase the upper limit (Eigen limit) on the length of the genome. The Darwin-Eigen cycle has been central to the events from the RNA world, to the evolution of the ribosome to proteins (chaperone-like and then catalytic) and the sequential steps in the origin of DNA.

Principles and concepts

It is difficult to infer the nature of events in the remote past, so in order for the reader to follow our reasoning, we specify here the principles and concepts on which the argument is based (see also Jeffares et al.[3]).

Darwin-Eigen cycle and the error cascade

The maximum length (informational content) of a nucleic acid sequence is inversely proportional to the error rate of its replication (the Eigen limit). When the error rate exceeds this limit, new errors accumulate in the system, compounding the original error and resulting in eventual randomization—''mutational meltdown'' or ''error-cascade.''[10] The inevitability of this enforces the maximum sequence length (Eigen limit) for a given fidelity of replication. There is, however, a potential positive feedback mechanism (the Darwin-Eigen cycle, Fig. 2) by which natural selection of traits that reduce the error rate allow, in principle, greater amounts of information to be coded. Selection processes operating on this additional information could reduce the error rate even further, allowing even longer sequences.

Known physicochemical properties of molecules
It is important that any hypothesis be framed in light of our understanding of the physico-chemical properties of molecules (such as RNA, amino acids, peptides, etc.). For instance, the thermolability of both RNA[7,11] and free glutamine[7] are important examples overlooked by those who propose a hot origin of life.

Primary sequence data are unlikely to be reliable for the oldest divergences
Both theory and simulations[12,13] (Penny et al., unpublished) indicate that many sites in a sequence (those that are not kept constant by selection) will be so saturated by mutational events after several hundred million years that they could be unreliable for reconstructing evolutionary trees. One major concern is that trees that assume equal rates of change across sites may fall foul of long branch attraction if one group is evolving faster than the others—it appears more divergent and, hence, appears to be an early diverging group.[13] Furthermore, primary sequence data cannot unambiguously identify a tree and its root.[14] Given the uncertainty of current phylogenetic methods, we started examining RNA biochemistry to see what its properties might suggest about the origins and early stages of life.

RNA does not replace a protein catalyst; the trend RNA → RNP → protein is unidirectional
This conclusion is based on the kinetic properties of catalytic RNA, RNP, and protein.[3] Protein enzymes exhibit shorter reaction times (k_{cat}/K_m) and much greater turnover numbers (K_{cat}) than RNA enzymes.[3] If the RNA world idea is correct, then ribozymes predated protein enzymes, yet there are few ribozymes remaining. The replacement of ribozymes by catalytic protein apparently involved RNP intermediates because even nonspecific RNA-protein associations appear to improve the rate of ribozymes.[15,16] RNA could have evolved a new function de novo (see Appendix) but our major assumption is that RNA will *not* replace a catalytic protein. It is necessary to establish criteria for ascribing particular kinds of RNAs to the RNA world in order to differentiate between ancient and more recent RNAs.[3]

r-K selection and the RNA → RNP → protein transition
The rate of the transition from RNA to RNP to protein would depend on the selection pressure. An organism that is being selected for a shorter life cycle, faster response times to a fluctuating food supply, higher rate of reproduction, smaller size and, consequently, a shorter generation time (that is, *r*-selection) is under strong pressure to minimize rate-limiting factors in gene expression. In contrast, minimizing times for gene expression is not critical for an organism that is *K*-selected (more stable food supply, slower rate of reproduction, larger size and, consequently, longer generation time), and hence is not under strong selection for this feature. What this means in terms of the RNA to RNP to protein transition is that *r*-selected organisms will retain fewer RNAs than those organisms that are *K*-selected, and this is because RNA-dependent processes often (though not always—see Ref. 3) constitute the rate-limiting step, and hence there will be selection to replace RNA or RNP with protein.[3,7] Prokaryotic lineages have undergone *r*-selection to a greater extent than eukaryotes[17] and in general respond to environmental stimuli faster than eukaryotes, an example being gene activation.

Thermoreduction and the origin of prokaryotes
The thermoreduction hypothesis[18,19] is that many features of prokaryotes could originally be adaptations to high temperature, by minimizing thermolabile traits. The inherent difficulty with adaptation to high temperatures is the instability of single-stranded RNA (ssRNA) in particular. Little ssRNA remains in prokaryotes, and this is processed quickly where it occurs (such as in mRNA and rRNA processing). Yet ssRNA would need to have been long-lived in the original RNA world.

Continuity of stages
A goal for any evolutionary theory is to show that any endpoint could be preceded by a continuum of functional intermediates. Thus a ''single continuous theory,'' based on the above principles, is necessary and would link the RNA world through to the tree of life (Fig. 3). This, however, requires a consideration of other questions, such as the evolution of encoded protein synthesis and the transition from RNA to a DNA genome (Fig. 1).

In general, a theory that encompasses a range of phenomena is preferred over several explanations, each covering only a single aspect of a problem. There have been individual explanations for many features, such as the origin of protein synthesis, or the origin of eukaryotes, but little attempt has been made to link these. Because of the necessarily speculative nature of the problem, a theory is only useful if, as well as explaining current data, it leads to further predictions. Considering RNA world relics in extant organisms allows many aspects of early evolution to be integrated into one hypothesis. Using these principles as a guide, we will propose a series of stages from the late RNA world to prokaryotes.

Relics from the RNA world

Both catalytic and non-catalytic RNAs are abundant in modern organisms but which of them are likely to be relics of the RNA world? Each must be examined with respect to its biological context and position in metabolism (see Appendix). Current models for the most complex stage of the RNA world (the last ribo-organism, which we nicknamed *Riborgis eigensis*) describe the RNA world at the point where genetically encoded protein synthesis arose.[3,4,7] Apart from identifying putative relics (see Appendix) we can also estimate aspects

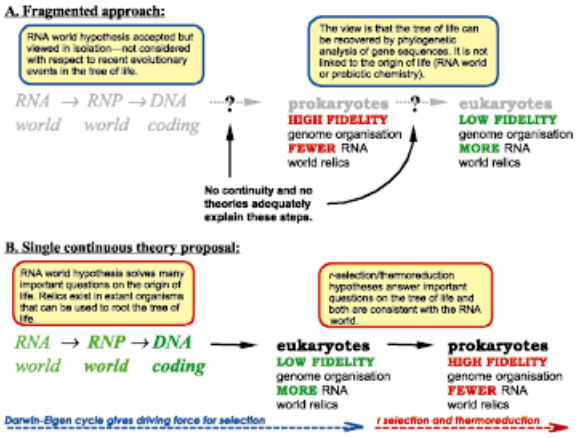

Figure 3. Integrating the RNA world with the tree of life. **A:** Fragmented approach: Each stage of the evolution of life is considered independently. Earlier events are unlinked to the tree of life, and hence the question of the RNA world and other periods in early evolution are considered using a different dataset to that used in building the tree of life. **B:** Single continuous theory: Stages from the origin of protein synthesis in the RNA world to LUCA are considered. A valid solution to one problem (say, rooting the tree of life) is favored if it allows a solution to other parts of the process (such as the origin of protein synthesis). RNA relics are currently the most informative data (see Appendix), and coupled with the principles described in the text, allow many events to be identified in terms of selection pressures and known mechanisms of evolution.

of the size and organization of the genome of *R. eigensis*.[3] RNA relics which remain today are an *integral* part of current metabolism and appear unchanged from their likely function in the RNA / early RNP world. This approach can partially reconstruct the metabolism of the LUCA. The RNA dataset, however, does not allow us to resolve the relationships between archaea and eubacteria, which is why we can only refer to these in terms of their phenotype. Thus, the dataset gives us an idea of the RNA metabolism of the LUCA, and allows a different method to root the tree of life[7] independent of traditional methods using RNA or protein gene sequences.[20–24]

The distribution of RNA world relics shows that all known relics occur in the eukaryote lineage—only a subset of these occur in the two prokaryote lineages.[3] From the relics of the RNA world and its genome organization, we have built a model of LUCA, and its genome organization more closely resembles that of eukaryotes than that of prokaryotes. (We do *not* consider cytological organization, such as intracellular compartmentation, or whether it had a membrane-bound nucleus, though this is now becoming possible.[25]) If it appears that the LUCA in its *genome organization* was in some sense eukaryote-like, we must answer the question: what events led to the derivation of the prokaryotes? An understanding of the properties of RNA and the RNA-world genome organization is also crucial to addressing this question.

The origin of ribosomal RNA and protein synthesis

The evolutionary origin of the ribosome, the biggest RNA complex within the cell, presents a major conundrum. If one accepts the idea of an RNA world, then a precursor for the ribosome must have existed in that world,[3,4,7,26] since a complex structure cannot appear de novo. The catalytic core of the ribosome is the RNA with protein providing, among other things, a scaffold which supports this core[27–29] (Table 1). The assembly of the ribosome in eukaryotes is even more complicated, with over 50 small nucleolar (sno) RNAs involved in its maturation.[30] Formation of all sites of 2′OH ribose methylation and of pseudouridine modification so far examined require complementary snoRNAs.[31,32] Also, a number of snoRNAs are essential for catalytic cleavage of pre-rRNAs (reviewed in Ref. 33).

RNA processing reveals a cascade of events in the production of mature rRNA: the spliceosome is required for the liberation of at least some snoRNAs from their positions in the introns of protein genes.[34–36] In turn, snoRNAs are required to produce mature rRNA, as described above. Furthermore, snRNPs, which are assembled in the cytoplasm, move to the nucleolus where they may be methylated and pseudouridylated by snoRNPs before becoming active in nuclear splicing of mRNA.[37]

$$\underset{\text{(spliceosome)}}{\text{SnRNA}} \rightarrow \underset{\text{(processosome)}}{\text{mature snoRNA}} \rightarrow \underset{\text{(ribosome)}}{\text{mature rRNA}}$$

If protein synthesis arose from the RNA world, this cascade is exactly what is expected, since protein processors did not exist during this period in the evolution of life.

The size of the rRNAs, their complex maturation, and the cascade of processing steps with RNA acting on RNA demonstrates that the ancient proto-ribosome was *central* to the metabolism of *R. eigensis.* Ribosomal RNA, and presumably some of its complex processing, must have predated translation and must have had a function in the RNA world. Whatever this precise function was, it must have been essential.

Before considering a potential solution, it is necessary to raise a related question. We have deduced that *Riborgis eigensis* was complex, possessing a large RNA genome of at least 15 kb, but may well have reached twice this size.[3] This exceeds the usual size of ssRNA viruses, and approaches the size of dsRNA viruses,[38] even though the RNA world would have lacked the more accurate protein RNA-polymerases. Given the Eigen limit of fidelity on genome size, *R. eigensis* has an absolute requirement for extremely accurate

TABLE 1. Comparison of Features Required in a Protoribosome With Those of the Modern Ribosome

Feature	RNA world replicase ('triplicase')	Genetically encoded protein synthesis
rRNA	16/18S and 23S; binds ssRNA (template) and tRNA, catalyses synthesis (new RNA)	16/18S and 23S; binds ssRNA (messenger) and tRNA, catalyses synthesis (peptide)
ssRNA	Template for replication/transcription. ssRNA fed through replicase by ratchet	Messenger for translation mRNA fed through ribosome by ratchet
'tRNA'	tRNA-like molecule source of triplet for elongation Anticodon added to growing RNA	tRNA has anticodon and amino acid acceptor Amino acid added, anticodon for recognition
tRNA-ssRNA binding	tRNA-like substrates bind to template RNA for elongation in replication	tRNA binds to mRNA for elongation in protein synthesis
Anticodon	Specific binding to template and added to growing RNA	Specific binding to codon only
Three-nucleotide ratchet	Needed for triplet addition during RNA elongation*	Needed for tRNA–mRNA binding/release cycle during peptide elongation
tRNA–rRNA binding	tRNA is substrate, rRNA is replicase	tRNA carries substrate, rRNA is ribosome core
Decoding	Template-triplet recognition	Codon–anticodon recognition
Transesterification	For adding triplets	Lost as no longer carries out RNA polymerization
Genomic-tag (tRNA)	Transcription/replication toggle	tRNA only used when charged with amino acid
Charging tRNA with amino acid	Increased accuracy of triplet recognition, code established (in part) before protein synthesis	Required for translation of genetic code to protein
The triplet code	Initially became important as amino acid tags, an aid to anticodon–codon recognition Many-to-one relationship between codons and tags as tags were probably only positively charged amino acids	The universal code Many-to-one relationship between codons and amino acids (genetic code) becomes established
Dipeptide formation	Only occurred on replicase because charged tRNAs bring amino acids into close proximity Peptidyl transferase carried out dipeptide formation	Peptidyl transferase reaction constitutes core of the protein synthesis reaction

*See text for discussion of triplet addition by replicase.

replication for an RNA-based system. Recombination, copy number (ploidy), multiple origins of replication, and a genome fragmented into chromosomes provide ways of reducing error catastrophe for the maintenance of a genome of this size.[3,39,40] Nevertheless, our current understanding still requires a highly accurate RNA polymerase in *R. eigensis*. The absence of an obvious relic of this polymerase constitutes a huge gap in our reconstruction of the RNA world.

Could the ribosome have evolved from an ancient RNA replicase?

Since there is no obvious relic of an RNA replicase—and no obvious function for the proto-ribosome—one possible solution is that the proto-ribosome was itself involved in RNA replication.[7,26] This solution is attractive because it is economical: it gives an RNA world replicase/polymerase, and an origin for the ribosome and its RNA processing cascade. We favor a model with an RNA polymerase that adds trinucleotides to the growing RNA chain.[7,26] Important parallels can be drawn between RNA replication and translation: both involve a ratchet to move an enzyme along an RNA template, and the base pairing of complementary RNA (nascent RNA strand, or the tRNA anticodon) to this sequence (Table 1).

An advantage of trinucleotide addition over single nucleotides is that trinucleotides H-bond longer to the RNA template, giving the replicase sufficient time to carry out polymerization. Protein RNA polymerases can accurately incorporate a single nucleotide but it is expected that the slower turnover times of ribozymes[3] could result in dissociation of nucleotide and template before the reaction was complete. Increasing the length of an oligonucleotide increases the stability of base pairing of the structure and the duration of binding exponentially. The number of possible oligonucleotides, however, would also increase exponentially; there are four times as many potential substrates for each additional nucleotide. Oligonucleotides other than triplets are possible, but only triplets are consistent with the origin of encoded protein synthesis.

The evolution of the ribosome is perhaps best understood as a *sequential* process. It probably came together through the interaction of a number of modules and there is thus no need to invoke massive new complexes of interacting catalytic RNA arising in one step (Table 1). Crucial to our model is to explain the coupling of the different RNAs (rRNA and mRNA) in a single complex. The precursor of the contemporary 16S rRNA may have been the earliest component, since it is involved in decoding the information on the mRNA. The 23S rRNA (the peptidyl transferase)[27] would most likely have been a later addition to the function of the replicase, since a model with amino acid tags, particularly positively charged ones, would have improved fidelity, perhaps by aiding in recognition or by stabilizing the interaction between the

negatively charged RNAs.[41] Stabilizing the two charged ''tRNAs'' with RNA complementary to the anticodons would greatly facilitate peptidyl transferase efficiency. The crucial event comes when the replicase, the peptidyl transferase, and the stabilizing RNA molecule all interact; if the stabilizing RNA (the prototype mRNA) is genetically encoded, the first genetically encoded protein is, by default, produced.[7] The use of amino acid tags bound to the tRNA for trinucleotides (now anticodons) being added to the growing RNA chain means that a relationship between codon and amino acid is already forged in the RNA world.[42] The full genetic code need not have arisen at this point but a triplet code is at least partially established before protein synthesis.

The origins of mRNA and genetically encoded proteins. Were introns "first"?

Can we predict the origin of messenger RNA, and the likely nature of the first genetically encoded proteins? The first proteins are expected be chaperone-like RNA binding proteins which increase the accuracy and the speed of ribozymes. Because the *precise* sequence of the protein is not essential for such a function,[15,16] the first proteins could have arisen from sequences in the genome that were initially noncoding. Thus, the benefits of genetically encoded protein synthesis are realized by virtue of the stabilizing role of the first proteins. An increase in catalytic efficiency is therefore possible without the need for accurately coded catalytic proteins.

A possible clue to the origin of mRNA is that a large proportion of the snoRNAs in eukaryotes are found within the introns of genes that code for ribosomal proteins.[33] Because snoRNAs are crucial to the maturation of the ribosomal RNAs, we argue that they *must predate* genetically encoded proteins. If so, *the introns (coding snoRNAs) were there first*; the current exons that surround snoRNAs are more recent. Not all snoRNAs are encoded in introns of ribosomal proteins—but other genes that contain snoRNAs in their introns would be considered ancient proteins on our model. They are not catalytic proteins, but rather act as multifunctional RNA binding proteins or ''RNA chaperones''[7] and heat shock proteins. Ribosomal proteins are obvious examples of ancient multifunctional RNA-binding proteins that stabilize catalytic RNA.[43,44]

The Darwin-Eigen cycle (Fig. 2; see above) also applies to the evolution of the ribosome; the first nonspecific RNA binding proteins would allow improvement in accuracy of translation, resulting in an increased capacity to translate protein more accurately. Ad infinitum. Thus the improvement in the catalytic efficiency of RNA-enzymes by their association with nonspecific protein feeds back on the fidelity of translation.[10] With time, this would allow the evolution of more complex catalytic proteins, where primary sequence is more critical to function.

As described above, a striking observation is that these putative early ribosomal proteins contain in their introns snoRNAs, which date back to the RNA world.[3,33] Further observations support the idea that the intronic snoRNAs predate the ribosomal proteins, and once existed without host ''genes.'' Several genes have now been found where the *exons* are noncoding, and the introns house one or more functional snoRNAs.[45,46] In one case, eight snoRNAs are contained in the introns of a single pre-mRNA.[46] These snoRNAs are liberated from the precursor mRNA by the spliceosome and the mature message does not get translated. Thus, there is a nonfunctional ''message'' which liberates functional snoRNAs, with the ''exons'' being spliced together by virtue of the coupling of cleavage and ligation functions in the spliceosome. This, in our opinion, is identical to the raw genetic material for the evolution of the first genetically encoded proteins. Notably, this model allows protein evolution and an intron–exon structure without exon-shuffling.[47]

The evolution of DNA is sequential and also adheres to the Darwin-Eigen cycle

As with the ribosome, the DNA replication apparatus can be viewed as having evolved in a sequential manner, with each step having a selective advantage. Steps include: synthesis of deoxyribonucleotides from ribonucleotides; replacement of uracil with thymine; a DNA polymerase (including a mechanism for continuous and lagging strand synthesis); repair mechanisms for damaged DNA; DNA recombination; and a mechanism for transcription/gene expression from the DNA (see also Ref. 48). The main point to stress here is that each additional step towards DNA allows an increase in the Eigen limit, and hence the coding capacity of the genome.

The elimination of the reactive 2′-OH of the ribose by ribonucleotide reductase uses free radical chemistry. This would probably not have arisen until protein catalysts were well established; ribozymes are likely to be too susceptible to backbone cleavage to be able to cope with free radical intermediates. Although the three classes of ribonucleotide reductase have a very low degree of sequence similarity, they are predicted to have a number of structural similarities. It has been proposed, on the basis of cofactors, radical generation mechanisms, similarity in allosteric regulation, and three-dimensional structure that the three classes had a common ancestry.[49,50]

The recruitment of a ribonucleotide reductase is unlikely to have been the final step in the evolution of DNA. The replacement of uracil with thymine probably occurred later because synthesis of dTTP occurs via dUMP (dATP, dCTP, dGTP, and dUTP, are all synthesized from the corresponding ribonucleotide diphosphate). The recruitment of thymine ultimately solved the problem of cytosine deamination to uracil, once uracil was replaced by thymine, a C to U deamination can be repaired with 100% accuracy, allowing increased fidelity—another turn of the Eigen-Darwin cycle.

Replication of both RNA and DNA is very similar in that both involve the addition of a single nucleotide to a growing polymer. Recent studies on the RNA and DNA polymerases, and on reverse transcriptases have shown that changing a single amino acid in the active site can yield a relaxation in specificity for rNTPs and dNTPs.[51–53] Sequence, structural and mechanistic similarities[54] support a common origin. It is likely, then, that dNTPs were first used by a preexisting RNA polymerase.

A possible clue is to be found by a scan of proteins that interact with nucleic acids. Other proteins may have a common origin, as is the case for polymerases; RNA and DNA helicases are one such group.[55] Structural similarities exist between several ribosomal proteins and some DNA binding proteins, including a DNA gyrase.[44] A possible implication is that many of the functions may have been present prior to the advent of DNA, possibly acting on a dsRNA genome. Further advances in our understanding of such is crucial for a model of how the RNA genome was replicated and maintained, and how gene expression occurred from this genome in the RNP world prior to the origin of DNA. From an evolutionary viewpoint, transition to a DNA genome would have many intermediate steps, each increasing fidelity of replication, allowing another turn of the Darwin-Eigen cycle.

Can the RNA world model be used to root the Tree of Life?

In eukaryotes, as described earlier, there is a common pattern in RNA processing where RNA transcripts are processed by RNA (RNP ribozymes) to give functional RNA. This same pattern occurs for tRNA, rRNA, and mRNA, and is largely centered around the ribosome. The pattern appears an ancestral one. Prokaryotic lineages lack a number of these processor RNAs (Table 2)—but to derive eukaryotes from prokaryotes implies that RNA took over catalysis from proteins, violating much evolutionary and biochemical knowledge.[3,7] On current understanding, it appears unreasonable to propose that a ribosome evolved without these RNAs, particularly snoRNAs. Thus, the prokaryotic lineages appear to be derived, the eukaryotic lineage having retained more relics of the RNA world. A number of observations that point to this conclusion are summarized in Table 2 and all are considered with respect to the principles and concepts outlined above. The conclusion that the prokaryotes are derived is independent of the suggestion that the origin of the ribosome was in RNA replication; we are using the same data to consider different questions.

Is there a selective advantage for the reduction of RNA processing in prokaryotes? In fact, there are probably two. The first is *r*-selection, which favors faster response times and thus speeds the replacement of RNA by protein catalysts, as described in ''*r-K* selection and the RNA → RNP → protein transition'' (above). The second is the heat instability of RNA, especially ssRNA at high temperatures and unprotected by proteins. This is Forterre's thermoreduction hypothesis,[18,19] which suggests that during the origin of prokaryotes many thermolabile RNA-catalyzed processing steps had to be eliminated before growth was possible at high temperature. The hypothesis[18] explains many of the differences between prokaryotes and eukaryotes (see Table 2), including the latter's elegantly simplified rRNA and mRNA processing and the lack of nucleolus. It was only after protein catalysts had evolved that thermoreduction could occur—proteins then replaced many steps processed by RNA (Fig. 1).

Additional evidence for prokaryotes being derived is that reverse gyrase, an essential enzyme for life at high temperature, was formed by the fusion of two enzymes (a helicase and a topoisomerase) that exist in all organisms.[19] Life presumably existed at moderate temperatures before fusion of a helicase and topoisomerase formed reverse gyrase, enabling life at high temperatures. Similarly, mischarging of the tRNA for glutamine ($tRNA^{Gln}$) with glutamate, and subsequently converting the bound glutamate to glutamine, is easily seen as an adaptation to bypassing the low intracellular concentration of glutamine at very high temperatures; free glutamine not being stable under such conditions.[7] If the process of *r*-selection in the prokaryote lineages resulted in replacement of RNP with catalytic proteins, this may well have "primed" these lineages for adaptation to higher temperatures.[7]

One of the consequences of thermoreduction and *r*-selection is the loss of introns, as shorter RNAs with fewer

TABLE 2. Derivation of the Prokaryotes

	Riborgis eigensis	LUCA	Eukaryotes	Prokaryotes
DNA genome	−	+	+	+
Diploid or polyploid	+	+	+	−
Telomerase RNA	+	+	+	−
Linear Genome	+	+	+	−
rRNA processing by snoRNA	+	+	+	−
mRNA processing by snRNA[a]	+	+	+	−
tRNA processing by RNase P	+	+	+	+
Coupled rRNA transcription & processing	−	−	−	+
Coupled transcription & translation	−	−	−	+
Genome-encoded CCA tail[b]	−	−	−	+

Table 2 illustrates the similarities between *Riborgis eigensis,* LUCA, and eukaryotes, and contrasts these with the prokaryotes. Given a model for the RNA world and a knowledge of extant prokaryotes and eukaryotes, it is possible to make inferences about the LUCA, based on RNA relics found in extant eukaryotes and prokaryotes. The numerous differences between eukaryotes and prokaryotes can be attributed to the process of thermoreduction and *r*-selection in the derivation of the prokaryote lineages from a mesophilic, eukaryote-like ancestor.

[a]By definition, mRNA did not exist in the RNA world since there were no genes coding for proteins. The splicing function of snRNA however is considered a relic—it is the target substrate that has changed.

[b]Not all prokaryotes have this feature but its existence in prokaryotes is considered a derived feature (see text).

processing steps are advantageous. This process is envisaged as the formation of the prokaryote genome by reverse transcription of mRNA back into DNA of an early plasmid-like structure[7] (leading simultaneously to a circular genome). Interestingly, there is evidence that the catalytic subunit of telomerase is a protein that resembles reverse transcriptase.[56] Reverse transcription of mRNAs, and subsequent incorporation into a circular genome, can be seen as advantageous at high temperatures. The following points are relevant here (see Table 2):

1) Loss of introns (with the subsequent loss of the spliceosome) is easier to explain than the de novo origin of a very large RNA enzyme/protein complex, *after* the evolution of efficient protein catalysts.
2) About half the CCA-tails of tRNAs from *B. subtilis* and archaea are genomically encoded, and all are in *E. coli* and *M. genitalium*—in eukaryotes they are rarely so. (The CCA-adding enzyme has been identified in all three domains but is involved in ''repair'' in the case of *E. coli*.[57]) We interpret genomically encoded CCA-tails of prokaryotes as the result of reverse transcription of tRNAs into a circular genome during their early evolution.
3) Circular dsDNA is more thermostable than linear dsDNA.[58–60] Almost all prokaryotes, and all known thermophiles, have circular dsDNA as their genome. Those that do not are predicted to be more recently derived mesophiles.

Our approach does not lead to the conclusion, favored by others, of a hybrid origin of eukaryotes. In this theory, eukaryotes are the result of an ancient fusion between two prokaryotes, putatively an archaeon and a Gram-positive eubacterium.[61] Although this theory could possibly explain the origin of the nuclear membrane, it does not explain all the relic RNAs in the genomes of eukaryotes and their absence in archaea and eubacteria. For example, snoRNAs are not present in any prokaryote lineage, so they cannot be the source of the dozens of snoRNAs involved in ribosomal RNA processing and maturation in eukaryotes. snoRNAs are central to rRNA metabolism in eukaryotes, not late additions; similarly, they are encoded in exons of the oldest proteins. Furthermore, it is doubtful if the endosymbiont theory even explains the origin of the nuclear membrane. It's structure and function differs considerably from an external cell membrane[62] and the nuclear membrane is under cellular control in that it is assembled and disassembled during the cell cycle.

Our theory also offers a functional explanation for the loss of RNA world relics from prokaryotes during a period of *r*-selection and adaptation to high temperatures during early prokaryote evolution. Consider the converse. What selection pressure is there for de novo origin of more than 50 snoRNAs, and sites for methylation and pseudouridylation in eukaryotic rRNA? Prokaryotes already function efficiently without snoRNAs (and with much faster rRNA processing).

Finally, the problem of resolving the tree of life using phylogenetic methods is not at odds with our approach—the two should ideally give the same result, and a good test of the validity of a conclusion is that it can be reached using independent lines of inquiry. Very recent work by Brinkmann and Philippe[63] with signal recognition particle proteins utilized an improved method for retrieving ancient phylogenetic signal by examining only the slowest-evolving sites and assuming a covarion model of evolution. The work demonstrated that, when the slowest-evolving sites are taken for building the tree, the prokaryotes are monophyletic, and the root is located in the eukaryote branch. Those sites that evolve faster mask phylogenetic signal as the sites are saturated. Building a tree with these results in the fastest-evolving lineage, eubacteria, come out at the root as a result of long branch attraction.

Conclusions

All RNA world relics remaining in modern metabolism are found in eukaryotes, but only a subset are found in the prokaryotes. This, plus the apparently simple series RNA → RNP → protein for the evolution of catalysis, leads to the conclusion that LUCA was a ''karyote.'' (See Fig. 1: ''eu'' = true; ''karyote'' = nucleus-containing organism—we do not consider LUCA to necessarily have had a nucleus, though some form of compartmentation (e.g. nucleolus) would have been advantageous for a number of reasons not discussed here.) Together with the Darwin-Eigen cycle, this leads to a number of exciting possibilities. If the protoribosome was an RNA polymerase (adding triplets), there is a relatively straightforward origin for protein synthesis through a series of small incremental steps. There is a simple ordering of the emergence of proteins (with chaperone-like RNA binding proteins first) and then DNA. It gives a new interpretation of the evolution of life since LUCA. Forterre's thermoreduction hypothesis leads to testable predictions. One is that metabolic pathways where glutamine is used as an amide donor will instead use NH_3 in thermophiles. In addition, no hyperthermophiles will be found with linear genomes, and de novo RNA function will not have arisen in thermophiles (because of RNA instability).

Certainly, we feel that the application of basic evolutionary principles to biochemical data is currently the best available for examining the nature of the root of the tree of life. Macromolecule functions and their interactions are conserved even with extensive changes in the encoding gene sequence. The method has limitations; we are not able to date any of the events—only to order them. We cannot resolve the prokaryotic branch of the tree of life with the current dataset so cannot say if the prokaryotic ''phenotype'' arose once or twice–once for archaea and once for eubacteria.[7,64]

The theory is consistent with current models of early evolution, and has considerable explanatory power. It explains the abundance of RNA in eukaryotes, and its scarcity in prokaryotes, and the differences in genome organization between these groups. Furthermore, it offers both plausible selective pressures and a mechanism to explain these differences. Prokaryotes are not just primitive organisms that failed to become eukaryotes. They are very sophisticated in their RNA processing and have an elegant simplicity that is ideal for their lifestyle.

Acknowledgments

We thank Patrick Forterre, James Lake, and two anonymous reviewers for comments on earlier versions of the manuscript, and Hervé Philippe and co-workers for sharing results prior to publication. We are also indebted to the editor of *BioEssays*, Adam Wilkins, for his careful editing, patience, and many helpful comments.

References

1. Gilbert W. The RNA world. Nature 319:618.
2. Gesteland RF, Atkins JF. editors. 1993. The RNA world. Plainview, NY: Cold Spring Harbor Laboratory Press; 1986.
3. Jeffares DC, Poole AM, Penny D. Relics from the RNA world. J Mol Evol 1998;46:18–36.
4. Benner SA, Ellington AD, Tauer A. Modern metabolism as a palimpsest of the RNA world. Proc Natl Acad Sci USA 1989;86:7054–7058.
5. Eigen M, Schuster P. The hypercycle: a principle of natural self-organization. Berlin: Springer-Verlag; 1979.
6. Crick FHC. The origin of the genetic code. J Mol Biol 1968;38:367–379.
7. Poole AM, Jeffares DC, Penny D. The path from the RNA world. J Mol Evol 1998;46:1–17.
8. Woese CR, Fox, GE. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc Natl Acad Sci USA 1977;74:5088–5090.
9. Woese CR, Kandler O, Wheelis ML. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proc Natl Acad Sci USA 1990;87:4576–4579.
10. Hasegawa M, Yano T, Miyata T. Evolutionary implications of error amplification in the self-replicating and protein-synthesizing machinery. J Mol Evol 1984;20:77–85.
11. Lazcano A, Miller SL. The origin and early evolution of life: prebiotic chemistry, the pre-RNA world, and time. Cell 1996;85:793–798.
12. Charleston MA, Hendy MD, Penny D. The effects of sequence length, tree topology and number of taxa on the performance of phylogenetic methods. J Comput Biol 1994;1:133–151.
13. Philippe H, Laurent J. How good are phylogenetic trees? Curr Opin Genet Dev 1998;8:616–623.
14. Forterre P. Protein versus rRNA: problems in rooting the universal tree of life. ASM News 1997;63:89–95.
15. Coetzee T, Herschlag D, Belfort M. Escherichia coli proteins, including ribosomal protein S12, facilitate in vitro splicing of phage T4 introns by acting as RNA chaperones. Genes Dev 1994;8:1575–1588.
16. Herschlag D, Khosla M, Tsuchihashi Z, Karpel RL. An RNA chaperone activity of non-specific RNA binding proteins in hammerhead ribozyme catalysis. EMBO J 1994;13:2913–2924.
17. Carlile M. Prokaryotes and eukaryotes: strategies and successes. Trends Biochem Sci 1992;7:128–130.
18. Forterre P. Thermoreduction, a hypothesis for the origin of prokaryotes. CR Acad Sci Paris Life Sci 1995;318:415–422.
19. Forterre P, Confalonieri F, Charbonnier F, Duguet M. 1995. Speculations on the origin of life and thermophily: review of available information on reverse gyrase suggests that hyperthermophilic procaryotes are not so primitive. Orig Life Evol Biosph 19;25:235–249.
20. Woese CR. Bacterial evolution. Microbiol Rev 1987;51:221–271.
21. Iwabe N, Kuma K-I, Hasegawa M, Osawa S, Miyata T. Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. Proc Natl Acad Sci USA 1989;86:9355–9359.
22. Gogarten JP, Kibak H, Dittrich P, Taiz L, Bowman EJ, Bowman BJ, Manolson MF, Poole RJ, Date T, Oshima T, Konishi J, Denda K, Yoshida M. Evolution of the vacuolar H+-ATPase: implications for the origin of eukaryotes. Proc Natl Acad Sci USA 1989;86:6661–6665.
23. Rivera MC, Lake JA. Evidence that eukaryotes and eocyte prokaryotes are immediate relatives. Science 1992;257:74–76.
24. Gupta RS, Singh B. Phylogenetic analysis of 70kD heat shock protein sequences suggests a chimaeric origin for the eukaryotic nucleus. Curr Biol 1994;4:1104–1114.
25. Faguy DM, Doolittle WF. Cytoskeletal proteins: the evolution of cell division. Curr Biol 1998;8:R338–R341.
26. Gordon KHJ. Were RNA replication and translation directly coupled in the RNA (+protein?) world? J Theor Biol 1995;173:179–193.
27. Noller HF. Peptidyl transferase: protein, ribonucleoprotein, or RNA? J Bacteriol 1993;175:5297–5300.
28. Samaha RR, Green R, Noller HF. A base pair between tRNA and 23S rRNA in the peptidyl transferase centre of the ribosome. Nature 1995;377:309–314.
29. von Ahsen U, Noller HF. Identification of bases in 16S rRNA essential for tRNA binding at the 30S ribosomal P site. Science 1995;267:234–237.
30. Smith CM, Steitz JA. Sno storm in the nucleolus: new roles for myriad small RNPs. Cell 1997;89:669–672.
31. Bachellerie J-P, Cavaillé J. Small nucleolar RNAs guide the ribose methylations of eukaryotic rRNAs. In: Grosjean H, Benne R, editors. Modification and editing of RNA. Washington, DC: ASM Press; 1998. p 255–272.
32. Ofengand JA, Fournier MJ. The pseudouridine residues of rRNA: number, location, biosynthesis, and function. In: Grosjean H, Benne R, editors. Modification and editing of RNA. Washington, DC: ASM Press; 1998. p 229–253.
33. Maxwell ES, Fournier MJ. The small nucleolar RNAs. Annu Rev Biochem 1995;35:897–934.
34. Kiss T, Filipowicz W. Exonucleolytic processing of small nucleolar RNAs from pre-mRNA introns. Genes Dev 1995;9:1411–1424.
35. Fragapane P, Prislei S, Michienzi A, Caffarelli E, Bozzoni I. A novel small nucleolar RNA (U16) is encoded inside a ribosomal protein intron and originates by processing of the pre-mRNA. EMBO J 1993;12:2921–2928.
36. Tycowski KT, Shu M-D, Steitz JA. A small nucleolar RNA is processed from an intron of the human gene encoding ribosomal protein S3. Genes Dev 1993;7:1176–1190.
37. Lamond AI, Earnshaw WC. Structure and function in the nucleus. Science 1998;280:547–553.
38. Atkins JF. Contemporary RNA genomes. In: Gesteland RF, Atkins JF, editors. The RNA world. Plainview, NY: Cold Spring Harbor Laboratory Press; 1993. p 535–556.
39. Koch AL. Evolution vs. the number of gene copies per cell. J Mol Evol 1984;20:71–76.
40. Reanney DC. Genetic error and genome design. Cold Spring Harbor Symp Quant Biol 1987;52:751–757.
41. Maizels N, Weiner AM. Phylogeny from function: evidence from the molecular fossil record that tRNA originated in replication, not translation. Proc Natl Acad Sci USA 1994;91:6729–6734.
42. Szathmáry E. Coding coenzyme handles: a hypothesis for the origin of the genetic. code Proc Natl Acad Sci USA 1993;90:9916–9920.
43. Wool IG. Extraribosomal functions of ribosomal proteins. Trends Biochem Sci 1996;21:164–165.
44. Ramakrishnan V, White SW. Ribosomal protein structures: insights into the architecture, machinery and evolution of the ribosome. Trends Biochem Sci 1998;23:208–212.
45. Bortolin M-L, Kiss T. Human U19 intron-encoded snoRNA is processed from a long primary transcript that possesses little potential for protein coding. RNA 1998;4:445–454.
46. Tycowski KT, Shu M-D, Steitz JA. A mammalian gene with introns instead of exons generating stable RNA products. Nature 1996;379:464–466.
47. Gilbert W, Glynias M. On the ancient nature of introns. Gene 1993;135:137–144.
48. Lazcano A, Guerrero R, Margulis L, Oró J. The evolutionary transition from RNA to DNA in early cells. J Mol Evol 1988;27:283–290.

49. Reichard P. The evolution of ribonucleotide reduction. Trends Biochem Sci 1997;22:81–85.
50. Logan DT, Andersson J, Sjöberg B-M, Nordlund P. A glycyl radical site in the crystal structure of a class III ribonucleotide reductase. Science 1999;283:1499–1504.
51. Sousa R, Padilla R. A mutant T7 RNA polymerase as a DNA polymerase. EMBO J 1995;14:4609–4621.
52. Gao G, Orlova M, Georgiadis MM, Hendrickson WA, Goff SP. Conferring RNA polymerase activity to a DNA polymerase: a single residue in reverse transcriptase controls substrate selection. Proc Natl Acad Sci USA 1997;94:407–411.
53. Astakte M, Ng K, Grindley NGF, Joyce CM. A single side chain prevents *Escherichia coli* DNA polymerase I (Klenow fragment) from incorporating ribonucleotides. Proc Natl Acad Sci USA 1998;95:3402–3407.
54. Sousa R. Structural and mechanistic relationships between nucleic acid polymerases. Trends Biochem Sci 1996;21:186–190.
55. Gorbalenya AE, Koonin EV. Helicases: amino acid sequence comparisons and structure-function relationships. Curr Opin Struct Biol 1993;3:419–429.
56. Counter CM, Meyerson M, Eaton EN, Weinberg RA. The catalytic subunit of yeast telomerase. Proc Natl Acad Sci USA 1997;94:9202–9207.
57. Yue D, Maizels N, Weiner AM. CCA-adding enzymes and poly(A) polymerases are all members of the same nucleotidyltransferase superfamily: characterisation of the CCA-adding enzyme from the archaeal hyperthermophile *Solfolobus shibatae*. RNA 1996;2:895–908.
58. Vinograd J, Lebowitz J, Watson R. Early and late helix-coil transitions in closed circular DNA. The number of superhelical turns in polysome DNA. J Mol Biol 1968;33:173–197.
59. Marguet E, Forterre P. DNA stability at temperatures typical for thermophiles. Nucleic Acids Res 1994;22:1681–1686.
60. Sen S, Lahiri A, Majumdar R. Melting characteristics of highly supercoiled DNA. Biophys Chem 1992;42:229–234.
61. Gupta RS, Golding GB. The origin of the eukaryotic cell. Trends Biochem Sci 1996;21:166–171.
62. Pennisi E. The nucleus's revolving door. Science 1998;279:1129–1131.
63. Brinkmann H, Philippe H. Archaea sister-group of Bacteria? Indications from tree reconstruction artefacts in ancient phylogenies. Mol Biol Evol 1999; in press.
64. Forterre P. A hot topic: The origin of hyperthermophiles. Cell 1996;85:789–792.
65. Lee RC, Feinbaum RL, Ambros V. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell 1993;75:843–54.
66. Sloof P, Benne R. RNA editing in kinetoplastid parasites: what to do with U. Trends Microbiol 1997;5:189–195.

Appendix

Relic RNAs — the dataset

Distinguishing a Relic RNA from a New RNA

We use four criteria to help estimate whether a given RNA is a relic[3] from the RNA world. It is not necessary for all to be met—some relics are easily placed in the RNA world, some RNAs appear recent additions to metabolism, and a few sit on the fence.

Criteria

1. Central to metabolism.
2. Ubiquitous.
3. Catalytic.
4. The role is equally well carried out by protein in other organisms.

Examples of relic RNAs

1. Ribosomal RNA is central to metabolism, ubiquitous, and catalytic (see text).
2. tRNAs — universal and central to metabolism.
3. snoRNAs — some are essential for the maturation of rRNA, all are required for post-transcriptional modification of rRNA. Are now only present in eukaryotes but their absence from the prokaryotes is readily explained.[3,7]
4. snRNAs — are the RNA constituent of the spliceosome, responsible for mRNA processing in eukaryotes. Recent evolution of this molecule due to recent acquisition of introns is unlikely for two reasons. There is a strong selection pressure for mRNA-intron and spliceosome loss from prokaryotes in response to *r*-selection/thermoreduction (see text). Conversely, there is no satisfactory explanation for the de novo evolution of such a complex RNA–protein complex in eukaryotes. Second, some snoRNAs are excised by the action of the spliceosome on the introns in which these are housed; snRNAs may thus be a relic of an ancient ribozyme/RNA expression system.
5. RNase P — is a ubiquitous, catalytic RNA, and cleaves pre-tRNA to produce active tRNAs.
6. Telomerase RNA. The telomerase RNP enzyme is found in most eukaryotes. The internal template (guide) is supplied by the RNA moiety so that the extreme ends of linear chromosomes are replicated. Telomerase RNA itself does not provide ample evidence for an RNA world origin, but there are several arguments for an ancient linear genome that became circularized in prokaryotes.[3,7,22,26,38] Telomere replication would thus have been necessary in this ancient genome.

Examples of RNAs that evolved after efficient catalytic proteins arose

1. Antisense RNAs, e.g., lin-4, involved in developmental patterning in *C. elegans*.[65]
2. Some mRNAs from any gene that encodes a protein of demonstrably recent origin.
3. Some guide RNAs, e.g., guides in RNA editing in *Trypanosome* kinetoplastids.[66] Editing of pre-mRNA to produce functional mRNA is a solution to the problem of Muller's ratchet — build up of errors is higher than in genomes with recombination.

All these actual *functions* could perhaps be considered relics, but the *specific examples* clearly do not have their origins in the RNA world.