The origins and evolution of prokaryotes and eukaryotes.

A thesis presented in partial fulfilment of the requirements for the degree of

Doctor of Philosophy
in Molecular BioSciences

at Massey University

Anthony Masamu Poole

2001

Contents

Papers and manuscripts included in this thesis:

1. Jeffares DC, **Poole AM** & Penny D. Relics from the RNA world. *J Mol Evol* 46, 18-36 (1998). ***Reprinted with permission from Springer-Verlag New York Inc.***
2. **Poole AM**, Jeffares DC & Penny D. The path from the RNA world. *J Mol Evol* 46, 1-17 (1998). ***Reprinted with permission from Springer-Verlag New York Inc.***
3. RNA evolution: separating the new from the old. (manuscript).
4. **Poole A**, Jeffares D, Penny D. Early evolution: prokaryotes, the new kids on the block. Bioessays 21, 880-889 (1999). ***Reprinted by permission of Wiley-Liss, Inc., a subsidiary of John Wiley & Sons, Inc.***
5. Penny D & **Poole A.** The nature of the Last Universal Common Ancestor. *Curr Opin Genet Dev* 9, 672-677 (1999). ***Reprinted with permission from Elsevier Science.***
6. The origin of the nuclear envelope and the origin of the eukaryote cell. (manuscript).
7. **Poole AM**, Phillips MJ & Penny D. Prokaryote and eukaryote evolvability. *Biosystems* (submitted).
8. **Appendix: Poole A** & Penny D. Does endosymbiosis explain the origin of the nucleus? *Nature Cell Biol* 3, E173. [Letter]

Related papers not included in this thesis:

- Jeffares DC, **Poole AM** & Penny D. Pre-rRNA processing and the path from the RNA world. *Trends Biochem Sci* 20, 298-299 (1995). [Letter]
- **Poole A**, Penny D & Sjöberg B-M. Methyl-RNA: an evolutionary bridge between RNA and DNA? *Chem Biol* 7, R207-R216 (2000).
- **Poole A**, Penny D & Sjoberg B-M. Confounded cytosine! Tinkering and the evolution of DNA. *Nature Reviews Mol Cell Biol* 2, 147-151 (2001).
- **Poole AM**, Logan DT & Sjöberg B-M. The evolution of the ribonucleotide reductases: much ado about oxygen. *J Mol Evol* (accepted).

Acknowledgements.

Dad, thanks for teaching me how to think. If I can learn to be half as inquisitive and half as good a thinker as you, I'll be pretty chuffed. Mum, you taught me how to push myself when it is so easy to content oneself with the bare minimum. (Just as well with David as a supervisor!) I remember your reaction when I got 98.5% in a chemistry test at school. Rather than being happy that I scored so well, you asked me why I didn't get 100%!

Angelik and Emma, you deserve a big mention because you challenged me to become an efficient worker so that I wouldn't forget to have a life outside science. (Nevertheless, I've been in front of the computer and ignoring you for weeks now.) You also made me happy even after the crappiest of days at work and helped me to believe that what I was doing wasn't a complete waste of time. Angelik, tack också för att du visade mig Sverige. Visst är det bra att ha en hemlig kod här i Nya Zeeland också! I love you both more than you can know.

Big thanks also to Alan & Dennis for putting up with me and all the antics that caused you earache! Annette! Thanks for all the letters, phone calls, great discussions and advice over the years, and for showing me Denmark!

David, your fourth year course is to blame for you having to deal with me on and off over the last 7 years! You have always surprised me by giving me what I thought was an unreasonable amount of work, only for me to find that I *could* cope with the extra load (though at times I was swearing under my breath!), which has meant that I've learnt a lot of new skills that I otherwise wouldn't have made the effort to acquaint myself with. Oh, and your infinite store of anecdotes still continues to amaze me!

Dan, the clever one in our twosome. I have to admit that many of the clever bits in this thesis are your fault. Pink Giraffes on the Pink Serengeti! Matt, cheers for your ESNDs (Paper 7) another good idea in this thesis that I can't take credit for! I'm still waiting for the opportunity to strap torches to the backs of your knees to see if I can reset your biological (un)clock. Was the best of times flatting with you both, even if Matt is a filthy b'stard in the kitchen and Dan drank all my whisky!

Thanks Monbusho for the two years I spent in Japan, prior to beginning my PhuD, where I had the opportunity to broaden my mind, and do what I wanted *when* I felt like it! Thanks to Hideo Oka for helping me get into the University of Tokyo and Yoshiki Hotta, under whom I worked at Todai. When I got bored with what I was doing in his lab, Hotta sensei was kind enough to let me instead work on the RNA world problem! Very few people, if any, would give you such an opportunity.

Thanks to Matt Ridley, who thought our ideas were worth telling the world about, and who is responsible for much of the attention that our ideas have been given. It was an incredible thrill to be reading his book the Origins of Virtue at the time he contacted us—indeed an eerie coincidence!

I also wish to thank Britt-Marie Sjöberg, who accepted me into her group when Angelik and I had to move to Sweden. While the work I did with her is not included in this thesis, she has been a fantastic person to work with. Thanks to two years in her group, my knowledge of biochemistry is a bit more respectable than before! The stupid immigration laws in place in New Zealand at the time when I left for Sweden actually did me a favour...

Also, many thanks to Trish, who will be helping me bind this beggar before I get on the plane on Tuesday!

Thesis survival kit: Powerbook G3, Blur 13, Komeda Pop på svenska, Björk Debut & Homogenic, the Cardigans Emmerdale, Ryuichi Sakamoto 1996, Pearl Jam Yield.

Finally finally, thanks Dad for introducing me to The Hitch-Hikers Guide to the Galaxy and Homer's Odyssey all those years ago!

Ant.

Introduction

Introduction.

Candidate's note.

This thesis is a collection of papers, either published, submitted, or in preparation for submission, to international journals. Each chapter is a paper with an introduction, and can be read as a stand-alone paper, the purpose of the thesis introduction is to give an overview of the motivation for the work. It also reviews other approaches being taken ular with respect to establishing the evolutionary relationships between the three domains of life, archaea, bacteria and eukarya.

Problems with the accepted scenario for the origin of life.

For most biologists, the big picture regarding the origin and evolution of prokaryotes and eukaryotes is not at issue, and recent evidence only serves to back up the intuitively obvious: complex eukaryotes evolved from simpler prokaryotic ancestors. In the standard account, prokaryotes predated eukaryotes by at least 800 million years, as evidenced by cyanobacterial microfossils dating back 3.5 billion years [e.g. Schopf & Packer 1987, Walsh 1992]. (The finding of molecular markers of eukaryote metabolism by Brocks et al. [1999] has pushed back the emergence of the earliest eukaryotes from 2.1 billion years to 2.7 billion years.) Establishing the root of the tree of life has shown that prokaryotes in fact consist of two domains, the archaea and bacteria, that the Last Universal Common Ancestor (LUCA) of all extant life lived at extremely high temperatures and that the eukaryotes emerged from the archaea [Woese & Fox 1977, Woese 1987, Woese et al. 1990]. Prior to the emergence of cyanobacteria, life arose from prebiotic conditions on the early earth, and at some stage, possessed an RNA-rich metabolism. This period, dubbed the RNA world [Gilbert 1986, Benner et al. 1989], predated both the emergence of genetically-encoded proteins and of DNA as genetic storage molecule.

The standard picture is therefore that, after the period of heavy bombardment that is suggested to have vapourised the oceans on Earth perhaps as recently as 3.8 billion years ago [reviewed in Nisbet & Sleep 2001], life emerged, went through an RNA world period, a thermophilic prokaryote LUCA, and developed into cyanobacteria in an astonshingly short period of time - perhaps 300 million years [Lazcano & Miller 1994]. Indeed, life may have arisen in an even shorter timeframe than this. Among the oldest rocks are those from the Isua belt of Southwest Greenland, which arguably date back around 3.85 billion years. Enrichment of the ^{13}C isotope of carbon in these rocks have been argued to betray evidence of biological carbon fixation [Mojzsis et al. 1996].

A closer look at any one of these 'established facts', as with any area in science, suggests that none are as clear-cut as various popular science commentaries suggest. For instance, the earliest stromatolites do not contain microfossils, and may have an abiological origin [Lowe 1994, Grotzinger & Rothman 1996], unlike those inhabited by modern cyanobacteria. The dating of the Isua belt is controversial, as is the argument that the enrichment of ^{13}C found in rock samples from the belt is indicative of life [reviewed in Nisbet & Sleep 2001]. Furthermore, a hot earth rules out the possibility of an RNA world, given the instability of both single-stranded RNA [Forterre 1995a], and of the bases which make up RNA, particularly cytosine [Miller & Bada 1988, Levy & Miller 1998, Shapiro 1999]. The suggestions of a faint early sun and a 'snowball earth' [Bada et al. 1994, Nisbet & Sleep 2001] potentially fit better with an RNA-rich period in the origin of life [Moulton et al. 2000], yet perhaps one of the few points on which prebiotic researchers agree is that life could not have begun with RNA [e.g. Joyce & Orgel 1999, Nelson et al. 2000], there must have been earlier phases.

Another issue is the reliability of microfossil classification. Biologists seem to take the finding of 3.5 billion year old cyanobacteria as fact, yet forget that until the early work of Woese & Fox [1977], there were only prokaryotes. The primary domains archaea and bacteria were indistinguishable morphologically and were initially characterised solely on the basis of phylogenetic grouping from sequence motifs. Another concern is that modern cyanobacteria carry out oxygenic photosynthesis, yet the evolution of atmospheric oxygen probably did not occur until around 2.5-2.2 billion years ago [Ohmoto 1996, Summons et al. 1999]. Furthermore, it is not even always possible to distinguish prokaryote from eukaryote on the basis of morphology. Microbial symbionts in the gut of Surgeonfish were first characterised as eukaryotic protists [Fishelson et al. 1985], and it was not until rRNA sequences were obtained that it was possible to establish unequivocally that these large symbionts were in fact prokaryotes [Angert et al. 1993].

Finally, a hyperthermophilic LUCA is also at issue. Early work on reverse gyrase by Forterre [1995a] suggested that hyperthermophiles were not ancestral to mesophiles, and more recently, reconstruction of ancestral GC content suggests the LUCA was mesophilic [Galtier et al. 1999]. While the domains archaea, bacteria and eukarya are now generally accepted, it has become clear that horizontal transfer of genes between these lineages has probably occurred at significant levels, so simple phylogenetic reconstruction from a single gene may not be an accurate reflection of the evolution of the three domains [e.g. Martin 1999]. Moreover, the finding that microsporidia have been incorrectly placed as deep diverging eukaryotes [reviewed by Keeling & McFadden 1998] has served as a reminder that there are fundamental phylogenetic problems that have yet to be resolved in the reconstruction of deep divergences [e.g. Lockhart et al. 1996, Forterre 1997b, Philippe & Laurent 1998]. Indeed, as argued by Forterre [1995a,b, 1997a,b], we should not only be cautious about the claim that the LUCA was a hyperthermophile, but moreover, it has never actually been established that prokaryotes preceded eukaryotes in evolution. The evidence is at best circumstantial, and conclusions are accepted largely on the basis of the widespread assumption that this must be correct.

Right or wrong, the assumption that, owing to their greater complexity, eukaryotes have evolved from prokaryotes, definitely holds sway. Consequently, the approach that many biologists take in approaching the origin of a particular structure is to use the diversity of modern structures to try and build a picture of how the structure gradually became more complex. That is, a succession of forms from the modern prokaryotic apparatus, to the modern eukaryotic apparatus. This is flawed for several reasons. First, the assumption is made that whatever is prokaryotic must be ancient, and second, that there has been negligible change in the prokaryotic form since its advent. That the biolgical community can accept extensive horizontal transfer between prokaryotic organisms and extensive adaptation by prokaryotes to a wide range of dissimilar niches, at the same time as arguing that all prokaryotic structures are effectively living fossils, is amazing! Perhaps the most disturbing consequence of accepting *a priori* that prokaryotes predate eukaryotes is that the evolution of complex biological phenomena is approached as a purely descriptive problem. The direction of evolution is already known—simple to complex, and prokaryote to eukaryote.

However, there is no inherent reason under Darwinian evolution that evolution should proceed from simple to complex [Szathmáry & Maynard Smith 1995] - simplification may equally occur, as is evident in many examples of parasite evolution [e.g. Andersson & Kurland 1998, Grbic 2000, Wren 2000]. More problematically, with the solution implicit in the assumption, selection pressures are usually not given in trying to explain the origins of a structure, rather, the emphasis is on explaining the diversification/complexification of that structure, perhaps with natural selection as an afterthought [Paper 6]. This problem is in some respects parallel to the problem in developmental biology of always applying adaptationist reasoning in describing the evolution of structures, it is widely assumed that every observable trait must have a function, but this is unlikely to be the case [Gould & Lewontin 1979, Gibson 2000, Paper 7].

Given that reductive processes are as much a feature of evolutionary change as is complexification (as exemplified by parasite evolution), I have avoided making the assumption that prokaryotes are ancestral simply because they appear simpler. Instead, I have examined a range of data relevant to extant prokaryotes and eukaryotes to establish the nature of the processes underlying evolution in these groups [Paper 7]. I have also examined how the origin of the three domains fits with the RNA world period in the evolution of life [Papers 1-4]. My conclusion, and the main point of this thesis, is that the prokaryote lineages appear to have undergone reductive evolution, whereas the beginnings of eukaryote complexity may date back to early inefficient metabolic genetic and cellular systems [Papers 2, 4-6]. Thus prokaryotes are simple because they are streamlined, while eukaryotes are perhaps complex by historical accident [Paper 7].

The tree of life & the LUCA.

The tree of life as it currently stands aims to describe the evolutionary relationships between all organisms on Earth, but also to provide, by extrapolation, insights into the likely nature of the Last Universal Common Ancestor (LUCA). The crowning achievement was the tree of life from small subunit rRNA sequences [Woese & Fox 1977], which established the relationships between representatives of a wide spread of organisms. The work resulted in the discovery of the archaebacteria, later renamed archaea [Woese et al. 1990], which as distinct as eubacteria and eukaryotes. This was a major improvement for understanding the relationships between prokaryotes (and also single celled eukaryotes) since many species appeared very similar in terms of morphology and ultrastructure. Subsequently, attempts were made to root the tree of life using paralogous gene sets (gene pairs which had a common origin, and which were expected to have undergone a duplication and divergence from a single original gene prior to the emergence of the three domains) [Gogarten et al. 1989, Iwabe et al. 1989]. The overall aim of this was two-fold: to build a phylogeny describing the relationships of all organisms on the planet, and to determine which of the three domains is most like the Last Universal Common Ancestor (LUCA). While the pursuit of a tree of life has been plagued with difficulties such as the problem of long-branch attraction [Hendy & Penny 1989; Philippe & Laurent 1998, Forterre & Philippe 1999], finding suitable genes for rooting the tree [Lopez et al. 1999], the need to improve on the rates across sites assumption [Lopez et al. 1999, Brinkmann & Philippe 1999] and horizontal transfer [Teichmann & Mitchison 1999, Martin 1999, Doolittle 1999], and weaknesses and conflicts between individual gene data [e.g. Baldauf et al. 2000] there is still confidence that the correct tree can eventually be recovered.

The controversy over the tree of life and difficulties with the dataset and methods used is not an issue that I consider in this thesis. Numerous articles in the literature discuss this issue [e.g. Doolittle 1999, Snel et al. 1999, Brinkmann & Philippe 1999, Teichmann & Mitchison 1999, Stiller & Hall 1999, Forterre & Philippe 1999, Philippe & Forterre 1999, Baldauf et al. 2000, Penny et al. 2001].

Instead, I will consider the problems inherent in using the tree for inferring the nature of the LUCA. Reconstructing the tree of life to is central to understanding evolutionary relationships between all organisms on Earth. Continuing attempts should be made, despite the problems inherent with recovering phylogenetic relationships for such deep divergences [Penny et al. 2001]. I shall suggest however that, even if the correct tree were recovered, it would be largely uninformative for gaining an insight into the nature of the LUCA. It is my aim to describe exactly how the tree could be useful, and what the caveats and limitations of using the tree for evolutionary inference are. Important to that discussion is the issue of how horizontal transfer affects the tree, and whether the effect is so great that the tree becomes unresolvable, as has been suggested by Woese [1998].

Attempts have been made to overlay characters onto the tree (such as thermophily) in order to examine the LUCA. However, there has been little consideration of the compatibility with earlier scenarios for the origins of life, based on physicochemical data. For instance, if the LUCA was thermophilic [Woese 1987], given the thermolability of RNA [Forterre 1995a, Papers 2&4], it is difficult explain how presumed relics from the RNA world have been retained. Indeed, establishing the position of the root cannot provide an answer to the question of the nature of the LUCA—it is virtually uninformative from this viewpoint [Forterre 1997b, Paper 5].

An approach I take in this thesis is to consider the diversity of RNA in modern organisms. Taking the model for the RNA world, a physicochemical approach to understanding the replacement of RNA by protein in evolution is possible. By recognising the properties of RNA, it is possible to identify niches where RNA would be expected to be lost, or at the very least reduced severely in its use. The link with the RNA world, plus the adherence to the properties of RNA enabled me to take a model for the RNA world [Paper 1] and apply it to the problem of the nature of the LUCA [Papers 2&4]. This was done by examining the phylogenetic distribution of putative RNA world relics. Furthermore, the properties of RNA meant it was possible to examine the problem of polyphyletic gene loss for the RNA dataset, which gives a marked improvement over application of simple parsimony [Paper 5].

The minimal genome concept and reconstruction of the LUCA.

Currently, an active area of research has been in trying to derive a minimal genome, that is, the smallest gene set required for a functional cell [Mushegian & Koonin 1996, Mushegian 1999, Hutchison III et al. 1999]. Initially, it was considered that this approach would provide a useful means of examining the likely genomic make-up of the LUCA [Mushegian & Koonin 1996], though it is now being acknowledged that a minimal genome and the LUCA are not one and the same thing [Mushegian 1999; Paper 5].

A minimal genome is defined by the nature of its environment, and hence will differ depending on the genomes compared. In their initial work, Mushegian & Koonin [1996] compared the genomes of *Haemophilus influenzae* and *Mycoplasma genitalium* (at the time, the only two genomes available for analysis). Their reconstruction produced a minimal genome of 256 genes that could be argued to be both necessary and sufficient for the function of a modern cell. This minimal gene set was criticised by Becerra et al. [1997] because it led Mushegian & Koonin [1996] to argue that the LUCA had an RNA genome! Mycoplasmas are parasitic and the alternative explanation for the lack of *de novo* deoxyribonucleotide synthesis is that they obtain these from their host. This is a likely example of loss resulting from intracellular parasitism, and highlights the shortcomings of a minimal gene set as an approximation of the LUCA.

Leipe et al. [1999] contend that the LUCA had a genome consisting of both RNA and DNA, since their genomic analysis suggests that the bacterial DNA replication machinery is unrelated to the archaeal and eukaryal machinery. The coding capacity of RNA is so low that it is unlikely that an organism as complex as the LUCA had an RNA genome. Likewise, the ubiquity and common origin of ribonucleotide reductases argues against this [Poole et al. 2000]. Forterre [1999] has also pointed out that other DNA replication proteins share a common origin, and that anomalies in the others may be a result of non-orthologous gene displacements.

Following on from their systematic construction of a modern day minimal gene set, Mushegian & Koonin [1996] suggested how this gene set could be reduced to a set that would provide a model of a simpler ancestral cell:

i. Examine pathways requiring complex cofactors and eliminate those of them that can be bypassed without the use of the cofactors.
ii. Eliminate the remaining regulatory genes.
iii. Delineate paralogs and replace at least the most highly conserved families with a single, presumably multifunctional ''founder.''
iv. Apply the parsimony principle: those systems and genes that are not found in both bacteria and eukaryotes or both bacteria and archaea are unlikely to come from a primitive cell.

They also suggest: 'It has to be kept in mind that not only reduction but also certain additions to the minimal gene are likely to be required to produce a realistic model of a primitive cell. The most important of such additions may be a simple system for photo- or chemoautotrophy'.

Points i-iii are simplifications for which the only basis is the notion that the direction of evolution was always from simple to complex. There is no inherent requirement that organisms will tend towards greater complexity during evolution [Szathmáry & Maynard Smith 1995]. Indeed it has been argued that prokaryotes arose through a process of reductive evolution, with aspects of eukaryote genome architecture and RNA processing being more indicative of the make-up of the LUCA than those found in prokaryotic organisms [Forterre 1995a, Glansdorff 2000, Papers

Table: Difficulties with using distribution to establish whether a gene was a feature of the LUCA.

	Bacteria	Eukaryotes	Archaea	HT[a]	RNA world relic	In LUCA?
Gene 1	✓	✓	✓	✖	✖	YES Ubiquitous No HT
Gene 2	✖	✓	✖	✖	✓	YES Predates LUCA
Gene 3	✓	✓	✓	✓	✖	UNCERTAIN Unplaceable if extensive HT
Gene 4	✓	✖	✖	✖	✖	UNCERTAIN[b]
Gene 5	✖	✓	✓	✖	✖	UNCERTAIN[c]

[a] HT: Horizontal transfer.

[b] If eukaryotes and archaea are monophyletic, Gene 4 could either be argued to be a feature of the LUCA (with a single loss prior to the archaea-eukaryote divergence), or to have arisen in the bacterial lineage after it split from archaea-eukaryotes. If bacteria and archaea are monophyletic, Gene 4 could be a feature of the LUCA with two independent losses (once from archaea and once from eukaryotes), or may have arisen specifically in the bacterial lineage, after it split from archaea.

[c] If eukaryotes and archaea are monophyletic, it is as likely that Gene 5 arose in the common ancestor of these two groups as it is that it was a feature of the LUCA. If bacteria and archaea are monophyletic, parsimony would suggest the gene was a feature of the LUCA, with loss from bacteria.

2, 4 & 5]. Finally, reductive evolution is a hallmark of the mycoplasmas [Fraser et al. 1995] and such reductive evolution may be a hallmark of the parasitic lifestyle of the organism [Andersson & Kurland 1998, Paper 7]. An example is the different degrees of degradation of the S-adenosylmethionine synthetase gene in 8 species of *Rickettsia* [Andersson & Andersson 1999], which are obligate intracellular parasites. Thus the minimal genome concept may better represent the minimal parasitic/obligate intracellular symbiont genome; further reduction would produce an even more extremely minimal parasitic genome, not an approximation of the LUCA.

Mitigating against points i-iii is their final comment. However, this reduces the worth of the minimal genome approach to understanding the LUCA, since one may add or remove anything, without a specified framework that enables additions or removals to be evaluated. The RNA world model suggests that many RNA processing pathways absent from prokaryotes should be included in any reconstruction of the make-up of the LUCA [Papers 2,4&5].

The likelihood then is that the LUCA was not 'minimal' as mycoplasmas or other obligate intracellular parasites are. Importantly, paralogous genes (point iii) are expected to have been a feature of the LUCA, and these have figured in attempts to root the tree of life [see Forterre & Philippe 1999, Glansdorff 2000, for review]. While paralogous genes have originated from a single "founder", the duplications that gave rise to some paralogues will have occurred prior to the emergence of the three domains of life. More generally, throwing away paralogues may mean that a minimal gene set could be underestimating the level of complexity of the LUCA. The problem with which we are faced is then, given a minimal gene set as a starting point, how to decide what features should be removed, and what should be added?

Finally, point iv is that simple parsimony is a useful tool for reconstructing the LUCA. Given the three domains, archaea, bacteria and eukaryotes, the presence of a trait in two of the three is not in itself strong evidence for the presence of that trait in the LUCA. If agreement on the topology of the tree, and hence the position of the root, can be reached, this may guide the use of parsimony in tracing genes back to the LUCA [Forterre 1997a, Papers 3 & 5]. Rigid application of parsimony however may wrongly exclude genes that can be traced back to the LUCA on other grounds, or exclude genes for which no other evidence of their ancestry is evident.

Building on the minimal genome.

In terms of reconstruction of the LUCA, the minimal genome concept should not be abandoned, but its limitations should be noted. It may help to take the minimal genome concept as a starting point, as it provides a powerful way of sorting through a large number of traits to establish which can possibly be traced back to the LUCA. Certainly, the conceptual difficulty of reconstructing the RNA world [Papers 1,3&4] is similar in this regard, but the nature and size of the dataset makes it easier to distinguish, *ad hoc*, putative RNA world relics from RNAs that have evolved more recently [Paper 3]. Based on Mushegian & Koonin's [1996] original proposal, along with current attempts to reconstruct the LUCA using a model for the RNA world [Papers 2, 4 & 5], I suggest the following amendment, where I remove and replace criteria i-iii, amend iv and effectively expand their final point on additions to include the RNA world data (see table). This provides a tentative method for how to go about reinserting some traits into a minimal gene set to improve the reconstruction of the LUCA:

1. Inclusion of synthetic pathways for pyridine nucleotide cofactors because these are likely RNA world relics, though not necessarily of pathways requiring these cofactors. Rather, it is the *generic* reaction chemistries that should be considered ancestral.
2. Inclusion of putative RNA world relics, even where these are not universal in distribution.
3. Reintroduce paralogues in those cases where these clearly diverged prior to the divergence of the LUCA into the three domains.
4. Apply simple parsimony with caution: under certain circumstances, it is weak or misleading (see table). Current disagreements on the position of the root (and therefore the relationships between the three domains) makes it difficult to use this in examining possible polyphyletic losses or gains.
5. The ability to describe a large number of traits as ancestral or derived on the basis of a single selection pressure should permit reconsideration of some datasets which may not otherwise be included in the minimal genome.

The problem of horizontal transfer.

Much has been made of the question of horizontal transfer in the three lineages. It is still debated how extensive this is - some authors have argued for massive unbridled horizontal transfer events [Woese 1998, Doolittle 1998], some have argued that there are detectable patterns to the process [e.g. Jain et al. 1999, Lan & Reeves 2000, Paper 7], and some have suggested there is very little transfer at all [Snel et al. 1999]. The other issue is whether this transfer is extensive and ongoing [Ochman et al. 2000, Lan & Reeves] or whether it was extensive and has possibly slowed [Woese 1998]. The need for caution is obvious: horizontal transfer of genes will blur the ability to trace a given gene back to the LUCA, meaning that until it is possible to recognise even ancient horizontal transfer events, it will pay to be judicious with the application of parsimony. This may mean in effect that careful studies of the distributions of various genes within the diversity of life will be essential, and furthermore, that it will be crucial to develop ever more sensitive ways of recognising potential cases of transfer. Again, the tree of life will be a useful tool here, as limited distribution of a gene within one domain *may* provide a means of homing in on potential transfer events. Nevertheless, like the simple parsimony approach to the three domains, this will require that we have reconstructed the correct tree if it is to be of any use.

A clear example of how the difficulties of tree topology and possible horizontal transfer weakens the propensity for theory to examine events in early evolution is that of the 'respiration early' hypothesis [Castresana & Saraste 1995, Castresana & Moreira 1999]. Here the authors acknowledge that their argument rests on the assumption that the position of the root is correct, and that horizontal transfer has had no impact on the traits they examine. The hypothesis is inherently testable, but the prerequisite for testing it is that tree topology can be established, and that the impact of horizontal transfer can be evaluated. If one takes the extreme view of Woese [1998, 2000], it is not possible to test any such hypotheses, and the result is a situation whereby competing theories are evaluated on intuition or popularity, not on hypothesis testing.

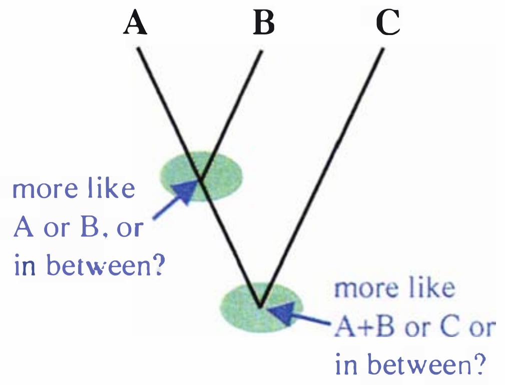
Current evidence argues that while genes involved in metabolic processes may transfer extensively, those involved in informational processes [*sensu* Rivera et al. 1998] tend not to be transferred very frequently, and some may not transfer at all [Jain et al. 1999]. It is thus a crucial goal of genomics to determine how frequent horizontal transfer is, between which types of organisms it tends to occur, and whether it applies to all genes [Martin 1999, Lan & Reeves 2000]. The ultimate goal is to construct a network describing genomic evolution, with those components of the genome that are subject to horizontal transfer overlain on a tree that describes organismal relationships, as determined by vertical transmission [Martin 1999]. Horizontal transfers have been suggested to contribute strongly to speciation events [de la Cruz & Davies 2000, Lawrence 1999], though currently there is no reason to suggest that these are more frequent than speciation by descent, particularly when one considers that there can be large intraspecies genome differences in prokaryotes [Lan & Reeves 2000]. Indeed, as Lan & Reeves [2000] point out, applying the species concept to prokaryotes will require a very different approach to the framework used for sexual organisms. In multicellular eukaryotes, where extensive cell specialisation makes transfers less likely than in single-celled organisms, speciation through horizontal transfer is likely to be rare [Paper 7]. However, in both unicellular and multicellular eukaryotes, there are strong indications that many genes have been transferred from organelles to the nucleus [Martin et al. 1998, McFadden 1999, Berg & Kurland 2000].

A tree of genomes is most likely to be part tree, part network and would indicate organismal relationships in terms of descent by modification, and gene relationships in terms of mode of transition. Some regions of the tree may have limited network structure, some may have extensive network structure, with tree branches being highly unreliable [Martin 1999].

Given known difficulties with phylogenetic analyses for deep divergences [Lockhart et al. 1996, Philippe & Laurent 1998, Lockhart et al. 1998, Philippe & Forterre 1999, Penny et al. 2001] how can cases of transfer be distinguished from problems of phylogenetic reconstruction? There are two aspects. One is to determine the nature and extent of horizontal transfer, and should be approached as a biological problem. What is the evolutionary basis for horizontal transfer between organisms, and what patterns emerge? Does transfer occur non-specifically given proximity between two organisms, or is transfer dependent on selection? Some aspects of horizontal transfer are considered in papers 6 and 7. In paper 7, I consider horizontal transfer from the viewpoint of organismal evolvability, and argue that extensive horizontal transfer has a selective component. The other aspect, which is not considered in any depth in this thesis, is how cryptic transfers can mislead phylogenetic reconstructions [Teichmann & Mitchison 1999, Philippe et al. 1999], and bioinformatic [Lawrence & Ochman 1998, Nelson et al. 1999, Ochman et al. 2000] and experimental [reviewed in Lan & Reeves 2000] approaches for establishing patterns of transfer.

Given the correct tree, some transfer events may in principle be identifiable, and so should traits dating back to the LUCA [Paper 5]. A trait that is found on both sides of the root can be best explained as loss in one of the three domains, and hence the most parsimonious explanation is a strong one. A trait that appears in two of the three domains, but where the two domains containing this trait group together (i.e. are monophyletic), is uninformative, and parsimony is not sufficient. Without further knowledge, it is not clear if the trait is ancestral or derived since the grouping of the two domains means the tree is reduced to a 'V' shape (Figure), with the two domains that form a monophyly being represented by a single branch. Nevertheless, the topology makes the application of parsimony weak, and it is also important to note that independent losses are much more likely than independent gains [Forterre 1997a].

In reconstructing the LUCA, it should be possible to examine whether there are other arguments for the inclusion of a particular gene, even if it has undergone horizontal transfer. Since function is of greater importance than whether there has been horizontal transfer, there may be cases where, say, a metabolic pathway can be included in the LUCA, even though one or more of the genes has been shown to have undergone horizontal transfer. For instance, numerous arguments have been made for an early origin for the TCA cycle [Wächtershäuser 1992, Morowitz et al. 2000], so this may be a good candidate for inclusion on the basis of function as opposed to inclusion on the basis of presence in the minimal genome dataset. In Paper 3 a similar approach is taken in distinguishing betwen the ultimate origin of an RNA, and recent recruitment to new function (proximate origin).


A
B
C
more like A or B, or in between?
more like A+B or C or in between?

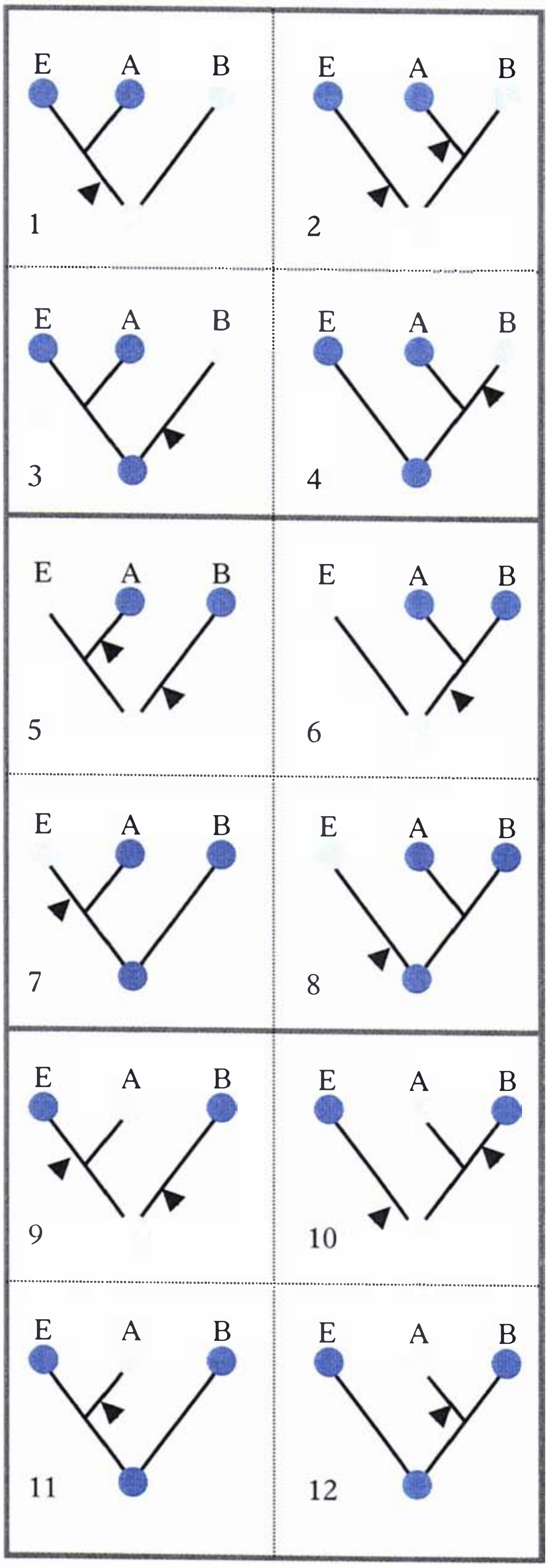


Figure: *The topology of the tree of life is uninformative as to the nature of the organism at the root.* **Above:** If topology alone is considered, it is not possible to establish whether the organism at the root is most like lineage C, or A+B, or in between these. Indeed, the same holds for the A-B monophyly. Overlaying characters on this tree to establish the nature of the root is likewise problematic, especially since horizontal transfer may mislead such analyses. **Right:** A shared character in all possible combinations, overlain on trees either rooted by bacteria or eukaryotes. Blue = presence, Grey = absence. For 2,5,9 & 10, independent gains are unlikely. All other trees are equally parsimonious for each shared character combination. If blue denotes loss, then these trees are still favoured as independent losses are easier to explain than independent gains. E.g. for 5 & 6, if grey is an RNA world relic, one vs two independent losses could only be evaluated by knowing the position of the root [Paper 5]. Trees 9 & 10 could be explained by mitochondrion to nucleus gene transfer. Extended from Forterre [1997a].

Using the tree for reconstructing LUCA

Broadly, the problems faced in reconstructing the tree of life are two-fold: current phylogenetic techniques are not able to recover the correct tree with any certainty, and horizontal transfers may further complicate reconstruction [Paper 5]. If, even with extensive horizontal transfer, the three domains, archaea, bacteria and eukaryotes can be shown to hold, a low-resolution tree of life will be recoverable, and that this can be rooted using various tricks such as using a paralogous gene as an outgroup (building separate unrooted trees from two genes that duplicated before the divergence of the three domains in order to root one tree with the other) [e.g. Gogarten et al. 1989, Iwabe et al. 1989, Brinkmann & Philippe 1999], can we then use the tree to obtain information on the root?

The fundamental problem with the tree as it currently stands (technical difficulties in reconstructing relationships aside) is that, at its lowest resolution, it attempts to describe the relationships between three monophyletic groups: archaea, bacteria and eukaryotes. Wherever the root is placed, it is difficult to infer much about the evolutionary relationships between groups of organisms (even when characters are overlain - see figure), and a rooted three-pronged tree can in principle establish whether two of those groups come together as a monophyletic group. Rooting the tree in the phylogenetic sense is an important means by which to examine the monophyly of the prokaryotes [Brinkmann & Philippe 1999]. What it absolutely cannot do however is to establish the nature of the LUCA. The outgroup is often argued to indicate which lineage is most likely to resemble the organism at the root, but this is incorrect (Figure). The structure of the tree is uninformative, and importantly, phylogenetic trees do not in themselves describe a process of evolutionary change. Their utility comes when, given the correct tree, various characters or traits can be overlaid upon the tree, giving a more complete picture of evolution. A recent example is the use of both fossils and molecular sequence data in reconstruction of the evolution of echolocation in bats [Springer et al. 2001].

The topology problem in the tree of life is fairly straightforward (Figure). The process of inference from phylogenetic trees has been to argue that the deepest-diverging groups in the branch that leads to the root provide insight to the nature of the LUCA. This has led to the widely-accepted proposal that the LUCA was hyperthermophilic and much like modern bacteria [e.g. Woese 1987].

Without considering the phylogenetic arguments for and against this proposal, let us first consider the implication of a split in the tree defining two domains (Figure). If domain A and B are shown to be related in the tree with the exclusion of group C, what can we infer about the common ancestor of A and B? Was it more like A, more like B, or did it have traits characteristic of both, some of which they still share in common? Or was it still like C? Considering the whole tree results in the same problem—it is not possible to decide if organisms that constitute 'outgroup' C in general, and deep-dranching members of group C in particular are more representative of the organism at the root. The branch that leads to the 'monophyletic' grouping of A and B could potentially provide just as much information on the nature of the organism at the root of the tree. If one of these three has maintained most metabolic traits of the common ancestor, it is not clear from the pattern of divergence given by the tree which of these three this is. When the ancestors of A, B and C diverged, it could have been that C underwent a series of reductions, whereby many ancestral traits were lost in the evolution of this domain, so that, even though the other two groups diverged from each other more recently, one or both may have retained more ancestral traits than has C. Alternatively, it could be the opposite!

Rooting of a tree with three groups (Figure) implies that A and B are monophyletic, and hence the tree could be represented in simplified form with two branches, and A and B together constituting one domain. Which group then is most similar to the organism at the root—the AB monophyly or C? No such information can be recovered simply by looking at branching patterns on a tree.

The tree clearly gives us important information on evolutionary splits between major lineages, but it offers no information on which traits can be traced back to the ancestor of all three groups. That said, evolutionary inference based on the tree of life has not relied solely on the topology - the standard interpretation is that thermophily appears in the deepest branches of both archaeal and bacterial domains, leading to the contention that the LUCA was a hyperthermophile [Woese 1987]. Given that rooting the tree supported the grouping together of archaea and eukaryotes to the exclusion of bacteria, this was a correct conclusion, assuming the relationships between the three domains were correctly recovered, and assuming that hyperthermophily evolved only once. If so, then, given hyperthermophily is recovered in both branches of the tree (i.e. it traverses the root), this argues that this is the ancestral state (Tree 7 in figure).

The bacterial rooting is subject to continued scrutiny as phylogenetic methods improve, and the hypothesis that the LUCA was a hyperthermophile is likewise testable. Indeed, there have been several criticisms on both the rooting of the tree, and the conclusion that the LUCA was a hyperthermophile. The competing hypothesis is that the bacterial rooting is a consequence of long branch attraction [Brinkmann & Philippe 1999, Lopez et al. 1999, Forterre & Philippe 1999]. An examination of the phylogenetic distribution of putative RNA world relics [Papers 2 & 4], gyrases and topoisomerases [Forterre 1995a], ancestral GC content [Galtier et al. 1999] and low stability of RNA at high temperature [Moulton et al. 2000] argues that the LUCA was mesophilic. These independent approaches argue that eukaryotes have retained a number of ancestral features that date back to the LUCA, while archaea and bacteria have lost these. Furthermore, the stability of hyperthermophily as a character has also been questioned, with several reports that hyperthermophilic traits common to both bacteria and archaea having undergone horizontal transfer [Nelson et al. 1999, Aravind et al. 1999, Forterre et al. 2000], and other traits, such as the lipid composition of hyperthermophile membranes [reviewed in Daniel & Cowan 2000], suggest hyperthermophily has evolved twice independently [Forterre 1996].

The tree of life displays the evolutionary relationships between extant organisms as patterns of divergence on a tree. All living organisms are thus billions of years removed from the LUCA, such that the deep branches do not necessarily represent 'living fossils', only the pattern of evolutionary divergence. Indeed, indications from current tree building methods are that it is the fastest-evolving lineages that are most likely to take basal positions because most current tree reconstruction methods tend to provide a measure of evolutionary distance which is affected by rate of evolutionary change [Laurent & Philippe 1998, Stiller & Hall 1999, Brinkmann & Philippe 1999]. The pattern of evolutionary divergence is not recovered because it has not been possible to build trees that correctly take into account rate variation between lineages. Brinkmann & Philippe [1999] have been able to demonstrate how Long Branch Attraction [Hendy & Penny 1989] affects the overall topology of the tree, using an implementation [Lopez et al. 1999] of the covarion model [Fitch & Markowitz 1970, Fitch 1971] to separate out fast-evolving and slower-evolving sites. With the fast-evolving sites, which will tend to become saturated, archaea and eukaryotes group together, but taking the slower-evolving sites returns a tree where the root is in the eukaryote branch, and the prokaryotes are monophyletic. If correct, the tree severely weakens the conclusion that the LUCA was a hyperthermophile, as this trait is now found in one branch only: the monophyletic prokaryotes (see trees 6 & 8 in figure).

Phylogenomics.

Nevertheless, the problem remains. Given the alternative trees: Brinkmann & Philippe's [1999] bacteria-archaea monophyly or the eukaryote-archaea monophyly [Woese et al. 1990, Iwabe et al. 1989, Gogarten et al. 1989], which is right? One alternative has been to move away from single genes and attempt to use whole genomes in phylogenetic analyses [e.g. Sicheritz-Pontén & Andersson 2001].

Genomics (unlike conventional phylogenetic analyses of one gene conserved across all organisms in the study) promises to allow us to compare all genes in a group of organisms. This is achieved in two ways. The simplest is counting the number of genes that are shared. Relatedness is based on the number of genes in common with other species in the study [Snel et al. 1999]. The other is carrying out a global phylogenetic analysis of genes that are shared in order to try and build a composite tree using sequence data. A more modest and potentially very powerful approach is a composite tree, where genes which have individually been shown to be informative in reconstructing distant phylogenetic relationships are used to produce a combined dataset. A recent analysis of the phylogeny of eukaryotes is one such example [Baldauf et al. 2000].

Nevertheless, these approaches are not necessarily expected to provide significant improvements to single-gene trees. A consensus tree over all, or for each of, the three domains, where there is general agreement for several different genes, all of which contain sufficient phylogenetic information from which to build a tree is not yet achievable. Protein and RNA trees give conflicting results [Forterre 1997b, Philippe & Forterre 1999]. At worst, large-scale 'phylogenomic' analysis simply amounts to adding more data without attempting to address limitations of models in current tree-building algorithms [Lopez et al. 1999, Penny et al. 2001] which require each site always to evolve at the same rate. Furthermore, it is not clear how genome-level comparisons will be able to deal with the problem of horizontal transfer. Snel et al. [1999] used gene presence and absence in 13 genomes as a phylogenetic character, claiming that their analysis supports the 16S rRNA tree and that horizontal transfer was not extensive. However, such an analysis might miss orthologous gene replacements as well as independent gains and losses through horizontal transfer.

If it is assumed that the problem of horizontal transfer is real, and that those genes which do transfer can be distinguised from those that do not, should the former be eliminated from reconstructions of the LUCA? These cannot be reliably traced back to the LUCA, *unless* independent criteria for their inclusion can be used (see table). From the subset that are primarily transmitted vertically, which are ancestral traits, and which are derived? That is, which were present in the LUCA, and which arose later? The difficulty here is that there is no good methodology for deciding this. One could use parsimony, such that where two of the three have a trait it is ancestral, and where two of the three lack it, it is derived. Parsimony as a rule is fraught with problems, especially where one applies it to three groups, as it could easily lead to artificial groupings of ancestral and derived traits [Forterre 1997a, figure]. Gene loss versus the origin of novel genes cannot be inferred without some evolutionary precedent, and parsimony is insufficient in three-domain problem [discussed in Paper 3 for the origin of snoRNAs]. Nor, as we have seen, does the tree give such precedent (e.g. if it is in C it is ancestral, if it is A and B but not C, it is derived), so this must be established through other lines of inquiry.

Non-phylogenetic approaches.

Using a genomic approach, many traits are simply not amenable to analysis, either because of horizontal transfer, or because traits which are not ubiquitous in distribution cannot always be reliably argued to date back to the LUCA on the basis of parsimony alone (Table). With current methods, those that turn out to have been subject to extensive horizontal transfer may not be reliably examined in the context of the LUCA problem, though cases where transfer turns out to be only very limited might be expected to be.

Since the reconstruction of the LUCA depends most on rebuilding a rough picture of metabolism before the emergence of the three domains, it is not necessary to use phylogenetic-based approaches in justifications for the antiquity of a given trait. While less ambitious than the minimal gene set [Mushegian & Koonin 1996], an alternative is to try to identify ancient metabolic traits, even if they are limited in distribution.

In this thesis, I have attempted to do just that. A means of examining some aspects of extant metabolism is the application of the RNA world theory to the problem, in the first instance to identify RNA species which are likely to be ancient [Papers 1&3], and subsequently, to explain the asymmetric distribution of these in modern species based on known principles. Since the tree gives us very limited information on the likely nature of the LUCA, owing to the rooting problem, an alternative that examines this is essential.

While the notion of an RNA world may or may not represent an intermediate in the evolution of life, currently there is no real alternative for understanding the origins of proteins and DNA. Certainly, it seems highly likely that RNA played a more prominent role in metabolism than it currently does, and not only is there a good physicochemical and biochemical basis for expecting RNA would be replaced over time by proteins and DNA, a number of RNAs, such as rRNA, tRNA, srpRNA and RNase P, are found to be ubiquitous [Papers 1&3]. The biggest problem is trying to identify candidate relics and, although criteria have been put forth that aid in distinguishing between relic RNAs and recent additions to metabolism, the approach is necessarily *ad hoc* [Papers 1,3&4]. Importantly, it is not an absolute requirement for candidate RNA relics to be ubiquitous, and this offers an improvement over parsimony, and abrogates the need for the correct tree in evaluating aspects of the nature of the LUCA.

Expanding LUCA: how easy or hard is identification of ancient metabolic traits?

Some ancient metabolic traits can be identified if they are ubiquitous and have been demonstrated not to have been subject to horizontal transfer. This is in itself likely to pose a difficult technical problem, as horizontal transfer would make it impossible to judge on distribution alone whether or not the trait was ancient.

Those traits that are not ubiquitous represent an equally formidable problem. How can such ancient traits be identified from a tree based on a single gene, or, from a tree based on comparisons of genome content (where presence/absence of a gene is a character), or a composite tree where several ubiquitous genes give the same tree?

Again, one could apply parsimony. However, a tree cannot be used to infer evolutionary pressures that account for changes along a branch, because the branching pattern alone cannot identify such pressures [Forterre 1997a]. It may however point us in the right direction, provided the topology problem is taken into account. For instance, if we are able to unambiguously determine the relationships between the archaea, bacteria and eukaryotes, the monophyly of two, for example archaea and bacteria, can greatly improve the usefulness of the parsimony rule in certain situations. For instance, given the tree in the Figure, if a gene known not to have been subject to horizontal transfer is found in organisms in groups A and C, but not B, and if the grouping (AB)C is correct, we can argue from parsimony that the trait was lost from group B, and that it can be traced back to the LUCA. If the trait were in C only, or in A and B but not C, parsimony cannot be used, so the tree cannot be used to determine whether the trait dates back to the LUCA.

In concluding the introduction, the main point I will be arguing with regard to reconstructing the LUCA is that the framework of the RNA world hypothesis provides one way of establishing some events in early evolution, and with greater certainty than searching for patterns in genomic data. This approach provides hard data on the metabolic make-up of the LUCA, and leads to testable hypotheses (described in the section on future work). However it cannot replace phylogenetic approaches for classifying taxa. It cannot even examine the question of the monophyly of the prokaryotes. Indeed, as described in Paper 5, if eukaryotes and archaea do turn out to be monophyletic, this does not affect the conclusion that the LUCA possessed some eukaryote-like features. Rather, it highlights how uninformative the root is - contrary to the interpretation that many non-phylogeneticists have, the outgroup is not indicative of the LUCA, and the direction of evolutionary change cannot be inferred solely from the topology.

What the approach in this thesis does allow is a hypothesis-driven approach to understanding eukaryote and prokaryote evolution. It provides continuity between the RNA world, the LUCA, and the subsequent divergence of the three domains. Furthermore, it makes a significant shift away from the preconception that prokaryotes predate eukaryotes by establishing important factors that influence evolution in extant prokaryotes and eukaryotes [Paper 7]. This provides an insight into evolutionary processes and establishes how the process of natural selection has operated in the evolution of prokaryotes and eukaryotes. Such insight cannot be established through phylogenetic analyses or comparative genomics alone.

References.

Andersson JO, Andersson SGE: Genome degradation is an ongoing process in *Rickettsia.* Mol Biol Evol 1999, 16, 1178-1191.

Andersson SGE, Kurland CG: Reductive evolution in resident genomes. Trends Microbiol 1998, 6, 263-268.

Angert ER, Clements KD, Pace NR: The largest bacterium. Nature 1993, 362, 239-241.

Aravind L, Tatusov RL, Wolf YI, Walker DR, Koonin EV: Evidence for massive gene exchange between archaeal and bacterial hyperthermophiles. Trends Genet 1998, 14, 442-444.

Bada JL, Bigham C, Miller SL: Impact melting of a frozen ocean on the early earth and the implication for the origin of life. Proc Natl Acad Sci USA 1994, 91, 1248-1250.

Baldauf SL, Roger AJ, Wenk-Siefert I, Doolittle WF: A kingdom-level phylogeny of eukaryotes based on combined protein data. Science 2000, 290, 972-977.

Benner SA, Ellington AD, Tauer A: Modern metabolism as a palimpsest of the RNA world. Proc Natl Acad Sci USA 1989, 86, 7054-7058.

Berg OG, Kurland CG: Why mitochondrial genes are most often found in nuclei. Mol Biol Evol 2000, 17, 951-961.

Becerra A, Islas S, Leguina JI, Silva E, Lazcano A: Polyphyletic gene losses can bias backtrack characterizations of the cenancestor. J Mol Evol 1997, 45, 115-117.

Brinkmann H, Philippe H: Archaea sister-group of Bacteria? Indications from tree reconstruction artifacts in ancient phylogenies. Mol Biol Evol 1999, 16, 817-825.

Brocks JJ, Logan GA, Buick R, Summons RE: Archaean molecular fossils and the early rise of eukaryotes. Science 1999, 285, 1033-1036.

Castresana J, Moreira D: Respiratory chains in the Last Common Ancestor of living organisms. J Mol Evol 1999, 49, 453-460.

Castresana J, Saraste M: Evolution of energetic metabolism: the respiration-early hypothesis. Trends Biochem Sci 1995, 20, 443-448.

Daniel RM, Cowan DA: Biomolecular stability and life at high temperatures. Cell Mol Life Sci 2000, 57, 250-264.

de la Cruz F, Davies J: Horizontal gene transfer and the origin of species: lessons from bacteria. Trends Microbiol 2000, 8, 128-133.

Doolittle WF: You are what you eat: a gene transfer ratchet could account for bacterial genes in eukaryotic nuclear genomes. Trends Genet 1998, 14, 307-311.

Doolittle WF: Phylogenetic classification and the universal tree. Science 1999, 284, 2124-2128.

Fishelson L, Montgomery WL, Myrberg Jr AA: A unique symbiosis in the gut of tropical herbivorous surgeonfish (Acanthuridae: Teleostei) from the Red Sea. Science 1985, 229, 49-51.

Fitch WM: Rate of change of concomitantly variable codons. J Mol Evol 1971, 1, 84-96.

Fitch WM, Markowitz E: An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. Biochem Gen 1970, 4, 579-593.

Forterre P: Thermoreduction, a hypothesis for the origin of prokaryotes. CR Acad Sci III 1995a, 318, 415-422.

Forterre P: Looking for the most "primitive" organism(s) on Earth today: the state of the art. Planet Space Sci 1995b, 43, 167-177.

Forterre P: Archaea: what can we learn from their sequences? Curr Opin Genet Dev 1997a, 7, 764-770.

Forterre P: Protein versus rRNA: problems in rooting the universal tree of life. ASM News 1997b, 63, 89-95.

Forterre P: Displacement of cellular proteins by cellular analogues from plasmids or viruses could explain puzzling phylogenies of many DNA informational proteins. Mol Microbiol 1999, 33, 457-465.

Forterre P, Bouthier De La Tour C, Philippe H, Duguet M: Reverse gyrase from hyperthermophiles: probable transfer of a thermoadaptation trait from archaea to bacteria. Trends Genet 2000, 16,152-154.

Forterre P, Philippe H: Where is the root of the universal tree of life? BioEssays 1999, 21, 871-879.

Fraser CM, Gocayne JD, White O, Adams MD, Clayton RA, Fleischmann RD, Bult CJ, Kerlavage AR, Sutton G, Kelley JM, Fritchman JL, Weidman JF, Small KV, Sandusky M, Fuhrmann J, Nguyen D, Utterback TR, Saudek DM, Phillips CA, Merrick JM, Tomb J-F, Dougherty BA, Bott KF, Hu P-C, Lucier TS, Peterson SN, Smith HO, Hutchison III CA, Venter JC: The minimal gene complement of *Mycoplasma genitalium.* Science 1995, 270, 397-403

Galtier N, Tourasse N, Gouy M: A nonhyperthermophilic common ancestor to extant life forms. Science 1999, 283, 220-221.

Gibson G: Evolution: Hox genes and the cellared wine principle. Curr Biol 2000, 10, R452-R455.

Gilbert W: The RNA world. Nature 1986, 319, 618.

Glansdorff N: About the last common ancestor, the universal life-tree and lateral gene transfer: a reappraisal. Mol Microbiol 2000, 38, 177-185.

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.

Gould SJ, Lewontin RC: The spandrels of San Marco and the Panglossian paradigm: a critique of the adaptationist program. Proc R Soc Lond B 1979, 205, 581-598.

Grbic M: "Alien" wasps and evolution of development. BioEssays 2000, 22, 920-932.

Grotzinger JP, Rothman DH: An abiotic model for stromatolite morphogenesis. Nature 1996, 383, 423-425.

Hendy MD, Penny D: A framework for the quantitative study of evolutionary trees. Syst. Zool. 1989, 38, 297-309.

Hutchison III CA, Peterson SN, Gill SR, Cline RT, White O, Fraser CM, Smith HO, Venter JC: Global transposon mutagenesis and a minimal mycoplasma genome. Science 1999, 286, 2165-2169.

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.

Jain R, Rivera MC, Lake JA: Horizontal transfer among genomes: the complexity hypothesis. Proc Natl Acad Sci USA 1999, 96, 3801-3806.

Joyce GF, Orgel LE: Prospects for understanding the origin of the RNA world. In: Gesteland RF, Cech TR, Atkins JF eds. The RNA World. 2nd ed. Cold Spring Harbor Laboratory Press, New York, 1999, p49-77.

Keeling PJ, McFadden GI: Origins of microsporidia. Trends Microbiol 1998, 6, 19-23.

Lan R, Reeves PR: Intra-species variation in bacterial genomes: the need for a species genome concept. Trends Microbiol 2000, 8, 396-401.

Lawrence J: Selfish operons: the evolutionary impact of gene clustering in prokaryotes and eukaryotes. Curr Opin Genet Dev 1999, 9, 642-648.

Lawrence JG, Ochman H: Molecular archaeology of the *Escherichia coli* genome. Proc Natl Acad Sci USA 1998, 95, 9413-9417.

Lazcano A, Miller SL: How long did it take for life to begin and evolve to cyanobacteria? J Mol Evol 1994, 39, 546-554.

Leipe DD, Aravind L, Koonin EV: Did DNA replication evolve twice independently? Nucleic Acids Res 1999, 27, 3389-3401.

Levy M, Miller SL: The stability of the RNA bases: implications for the origin of life. Proc Natl Acad Sci USA 1998, 95, 7933-7938.

Lockhart PJ, Larkum AWD, Steel MA, Waddell PJ, Penny D: Evolution of chlorophyll and bacteriochlorophyll: the problem of invariant sites in sequence analysis. Proc Natl Acad Sci USA 1996, 93, 1930-1934.

Lockhart PJ, Steel MA, Barbrook AC, Huson DH, Howe CJ: A covariotide model describes the evolution of oxygenic photosynthesis. Mol Biol Evol 1998, 15, 1183-1188.

Lopez P, Forterre P, Philippe H: The root of the tree of life in light of the covarion model. J Mol Evol 1999, 49: 496-508.

Lowe DR: Abiological origin of described stromatolites older than 3.2 Ga. Geology 1994, 22, 387-390.

Martin W: Mosaic bacterial chromosomes: a challenge en route to a tree of genomes. Bioessays 1999, 21, 99-104.

Martin W, Stoebe B, Goremykin V, Hansmann S, Hasegawa M, Kowallik KV: Gene transfer to the nucleus and the evolution of chloroplasts. Nature 1998, 393, 162-165.

McFadden GI: Endosymbiosis and evolution of the plant cell. Curr Opin Plant Biol 1999, 2, 513-519.

Miller SL, Bada JL: Submarine hot springs and the origin of life. Nature 1998, 334, 609-611.

Mojzsis SJ, Arrhenius G, McKeegan KD, Harrison TM, Nutman AP, Friend CR: Evidence for life on Earth 3800 million years ago. Nature 1996, 384, 55-59. [Erratum: Nature 1997, 386, 665]

Morowitz HJ, Kostelnik JD, Yang J, Cody GD: The origin of intermediary metabolism. Proc Natl Acad Sci USA 2000, 97, 7704-7708.

Moulton V, Gardner PP, Pointon RF, Creamer LK, Jameson GB, Penny D: RNA folding argues against a hot-start origin of life. J MolEvol 2000, 51, 416-421.

Mushegian A: The minimal genome concept. Curr. Opin. Genet. Dev. 1999, 9, 709-714.

Mushegian AR, Koonin EV: A minimal gene set for cellular life derived by comparison of complete bacterial genomes. Proc Natl Acad Sci USA 1996, 93, 10268-10273.

Nelson KE, Clayton RA, Gill SR, Gwinn ML, Dodson RJ, Haft DH, Hickey EK, Peterson JD, Nelson WC, Ketchum KA, McDonald L, Utterback TR, Malek JA, Linher KD, Garrett MM, Stewart AM, Cotton MD, Pratt MS, Phillips CA, Richardson D, Heidelberg J, Sutton GG, Fleischmann RD, Eisen JA, Whilte O, Salzberg SL, Smith HO, Venter JC, Fraser CM: Evidence for lateral gene transfer between Archaea and Bacteria from genome sequence of *Thermotoga maritima.* Nature 1999, 399, 323-329.

Nelson KE, Levy M, Miller SL: Peptide nucleic acids rather than RNA may have been the first genetic molecule. Proc Natl Acad Sci USA 2000, 97, 3868-3871.

Nisbet EG, Sleep NH: The habitat and nature of early life. Nature 2001, 409, 1083-1091.

Ochman H, Lawrence JG, Groisman EA: Lateral gene transfer and the nature of bacterial innovation. Nature 2000, 405, 299-304.

Ohmoto H: Evidence in pre - 2.2 Ga paleosols for the early evolution of atmospheric oxygen and terrestrial biota. Geology 1996 24(12) 1135-9

Penny D, Foulds LR, Hendy MD: Testing the theory of evolution by comparing phylogenetic trees constructed from five different protein sequences. Nature 1982, 297, 197-200.

Penny D, McComish BJ, Charleston MA, Hendy MD: Mathematical elegance with biochemical realism: the covarion model of molecular evolution J Mol Evol 2001, (in press).

Philippe H, Forterre P: The rooting of the tree of life is not reliable. J Mol Evol 1999, 49, 509-523.

Philippe H, Laurent J: How good are deep phylogenetic trees? Curr Opin Genet Dev 1998, 8, 616-623.

Philippe H, Budin K, Moreira D: Horizontal transfers confuse the prokaryotic phylogeny based on the HSP70 protein family. Mol Microbiol 1999, 31, 1007-1009.

Poole A, Penny D, Sjöberg B-M: Methyl-RNA: an evolutionary bridge between RNA and DNA? Chem Biol 2000, 7, R207-R216.

Schopf JW, Packer BM: Early Archean (3.3 billion to 3.5 billion year old) microfossils from Warrawoona Group, Australia. Science 1987, 237, 70-73.

Shapiro R: Prebiotic cytosine synthesis: a critical analysis and implications for the origin of life. Proc Natl Acad Sci USA 1999, 96, 4396-4401.

Sicheritz-Pontén T, Andersson SGE: A phylogenomic approach to microbial evolution. Nucleic Acids Res 2001, 29, 545-552.

Snel B, Bork P, Huynen MA: Genome phylogeny based on gene content. Nat Genet 1999, 21, 108-110.

Springer MS, Teeling EC, Madsen O, Stanhope MJ, de Jong WW: Integrated fossil and molecular data reconstruct bat echolocation. Proc Natl Acad Sci USA 2001, 98, 6241-6246.

Stiller JW, Hall BD: Long-branch attraction and the rDNA model of early eukaryotic evolution. Mol Biol Evol 1999, 16, 1270-1279.

Summons RE, Janhke LL, Hope JM, Logan GA: 2-methylhopanoids as biomarkers for cyanobacterial oxygenic photosynthesis. Nature 1999, 400, 554-557.

Szathmáry E, Maynard Smith J: The major evolutionary transitions. Nature 1995, 374, 227-232.

Teichmann SA, Mitchison G: Is there a phylogenetic signal in prokaryote proteins? J Mol Evol 1999, 49, 98-107.

Wächtershäuser G: Groundworks for an evolutionary biochemistry: the iron-sulphur world. Prog Biophys Mol Biol 1992, 58, 85–201.

Walsh MM: Microfossils and possible microfossils from the Early Archean Onverwacht Group, Barberton Mountain Land, South Africa. Precambrian Res 1992, 54, 271-293.

Woese CR: Bacterial evolution. Microbiol Rev 1987, 51, 221-271.

Woese CR: The universal ancestor. Proc Natl Acad Sci USA 1998, 95, 6854-6859.

Woese CR: Interpreting the universal phylogenetic tree. Proc Natl Acad Sci USA 2000, 97, 8392-8396.

Woese CR, Fox GE: Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc Natl Acad Sci USA 1977, 74, 5088-5090.

Woese CR, Kandler O, Wheelis ML: Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eukarya. Proc Natl Acad Sci USA 1990, 87:4576-4579.

Wren BW: Microbial genome analysis: insights into virulence, host adaptation and evolution. Nat Rev Genet 2000, 1, 30-39.

Paper 1

Jeffares DC, **Poole AM** & Penny D.
Relics from the RNA world.
Journal of Molecular Evolution 46, 18-36 (1998).
Reprinted with permission from Springer-Verlag New York Inc.

Paper 2

Poole AM, Jeffares DC & Penny D.
The path from the RNA world.
Journal of Molecular Evolution 46, 1-17 (1998).
Reprinted with permission from Springer-Verlag New York Inc.

Paper 3

RNA evolution: separating the new from the old.
Manuscript.

RNA evolution: separating the new from the old.

Abstract.

The existence of an RNA world, an RNA-rich period in the early evolution of life, is widely accepted, as is the idea that many cellular RNAs can be traced back to this period. However, while some RNAs may derive from the very earliest stages of life, others have arisen comparatively recently in evolution. A further difficulty is that some RNAs may have arisen early in evolution, but may have changed their role during evolution. It is therefore useful to distinguish between the 'ultimate' origin of an RNA and a 'proximate' origin, where it evolved into its present function. A number of RNAs have not been unequivocally placed as 'new' or 'old', including group I & II introns, snRNAs, tmRNA and snoRNAs. In this article, we examine how RNA world 'relics' might be distinguished from RNAs with a more recent origin, why there are problems or controversies in establishing the evolutionary origins of some RNAs, and whether it is possible to resolve these.

Introduction.

In eukaryotes it is well-established that RNA is central to a number of molecular processes, including protein synthesis, mRNA editing and splicing, rRNA and tRNA processing and telomere replication. Some of these RNAs are also found in archaea and eubacteria, though in general it appears that RNA plays a less prominent role in metabolism in these organisms (Wassarman et al., 1999). Indeed, this differential use of RNA is claimed to be a fundamental one, and may be the basis for very different evolutionary mechanisms employed in the diversification of prokaryotes and eukaryotes (Herbert & Rich, 1999a,b).

It is generally accepted that many RNAs are evolutionarily very ancient. The RNA world hypothesis (Gilbert, 1986) is that, prior to the advent of genetically-encoded proteins and DNA, RNA was both genetic material and major biological catalyst. With the advent of protein synthesis, and later, ribonucleotide reduction, RNA is believed to have gradually lost its central role as catalyst and information storage molecule. Those few RNAs that remain in modern metabolism are widely considered to be 'relics' from the RNA world period (Benner et al., 1989; Jeffares et al., 1998). However, with the number of novel RNAs growing, it is clear that many RNAs may have arisen more recently in evolution to fulfill specific functions and do not date back to the RNA world period (Eddy, 1999).

In this article, we briefly review the current state of the RNA world hypothesis insofar as it allows us to distinguish between RNAs that are likely to be ancient in origin and those which are more recent. We define 'ancient' as prior to the emergence of the three domains, archaea, bacteria and eukaryotes, that is pre-Last Universal Common Ancestor (pre-LUCA), and 'recent' as post-LUCA. A broad survey of RNAs that are probably recent innovations suggests that RNA is a potent source of novel function in eukaryotes. In addition, we will focus on those RNAs where the evolutionary origins are a current source of controversy.

Central to this problem is the need to distinguish between 'ultimate' origins and 'proximate' origins, thereby providing a distinction between the origin of a given RNA and the role it currently plays in modern metabolism. This distinction is in effect the same as the same as the use of the terms paralogous and orthologous in descriptions the evolutionary history of gene families. Orthologous genes have arisen through from a single ancestral gene through duplication and divergence and maintained the same function over time. Paralogous genes have also arisen from a single ancestral gene through duplication and divergence but now perform different functions. An example of orthologous RNA genes are the RNase P genes from *E. coli* and yeast. An example of paralogous RNA genes are RNase P and RNase MRP. In this paper, we are particularly interested in the latter, case. Where two related RNAs perform different functions, what is the ultimate origin of this family of RNA?

What is old and what is new?

We have previously suggested several criteria as an aid for drawing the line between relic RNAs and recently-evolved RNAs (Poole et al., 1999). These are:

1. That the RNA is ubiquitous in distribution.
2. That the RNA is central to metabolism.
3. Whether proteins perform the function equally well in other organisms.
4. That the RNA is catalytic[1].

These criteria are helpful, but are not necessarily sufficient to give a reliable indication of the likely status for every RNA. Criterion 1 is the strongest argument for the RNA world ancestry of a given RNA, and one can assign relic status to a number of RNAs, on this criterion alone. Obvious examples are tRNA, rRNA, RNase P and srpRNA (4.5S in bacteria, 7S in eukaryotes & archaea) (Jeffares et al., 1998). In the case of criterion 2, where an RNA is not ubiquitous, one may argue for an RNA world origin on functional grounds. In this manner, Maizels and Weiner (1999) have argued for the antiquity of telomerase function, which is further supported by a strong selection pressure for the circularisation of chromosomes in the prokaryotes being a derived trait, and thus not present in the RNA or RNP (ribonucleoprotein) worlds (Forterre, 1995). In spite of the example of telomerase, arguing just from criterion 2 is difficult, since it is a matter of opinion as to what is central to metabolism.

[1] The term catalytic RNA is used either in a chemical sense or a functional sense. In the chemical sense, a catalytic RNA is one which can catalyse a chemical reaction without the aid of protein, that is, the RNA is necessary and sufficient for catalysis. In a functional sense, an RNA which is necessary but not sufficient for catalysis is still a catalytic RNA. Bacterial RNase P RNA is catalytic in both senses, but human RNase P RNA is only catalytic in the functional sense.

The third criterion is of fundamental importance, and stems primarily from the argument that proteins are in general better catalysts than RNA (Jeffares et al., 1998; Poole et al., 1999). This suggests that, given the general trend is replacement of catalytic RNA with protein during evolution, in cases where in one lineage a protein performs a function identical to that of RNA in another lineage, the RNA is ancestral. However, certain functions may simply be better-suited to RNA (a point to which we shall return), and hence, not all RNAs should be placed automatically in the RNA world (Eddy, 1999). By itself, criterion 3 may be insufficient, but it is an important consideration, particularly where a function is argued to be central to metabolism. We consider several examples where criteria 2 and 3, combined, are important in assigning putative relic status.

Criterion 4 is more complex than it appears, which may be somewhat surprising, given the importance that catalytic RNA studies have played in the development of the RNA world hypothesis. Distinguishing between functional and chemical definitions for catalysis is helpful however. We will argue here that all RNAs defined as functionally catalytic but very few RNAs defined as chemically catalytic are direct descendents from the RNA world (see Table), though the latter are nevertheless important exemplars of RNA world complexity.

RNA as a source of novel function

As the RNA universe expands, it is becoming clear that RNA is more than just a relic from early evolution. 'New' RNAs in many cases can be readily picked out simply because the role they play is highly specialised and their phylogenetic distribution is very limited, indicating recent origins. It seems likely that the growing list of newly discovered RNAs (Table) is but the tip of the iceberg, especially given that current genomic search strategies (e.g. BLAST) do not perform well for RNA families, which in general retain very little primary sequence information (e.g. Ganot et al., 1997a; Lowe & Eddy, 1999; Collins et al., 2000). Likewise, large-scale identification techniques such as those possible with EST databases are biased against detection of noncoding RNAs (Eddy, 1999, though see Hüttenhofer et al., 2001).

Recent reviews (Eddy, 1999; Wassarman et al., 1999; Erdmann et al., 2001) cover much of the developments in RNA identification (for summary and relevant references from the literature, see Table), so we limit ourselves to a number of examples where it might be argued that RNA is inherently better suited to certain roles than protein. Furthermore, we consider briefly how RNA impacts on the evolvability of organisms.

RNA editing in kinetoplastids of trypanosomes.

RNA editing, whereby the sequence of a transcript is changed prior to translation, is widespread, and occurs via widely different mechanisms. The mechanisms appear unrelated and have limited distribution (Smith et al., 1997). RNA editing is particularly prevalent in organelles, and the best explanation for this is that editing is a response to mutational pressures from the operation of Muller's Ratchet in organellar genomes (Börner et al., 1997). Muller's Ratchet is the slow accumulation of slightly deleterious mutations in the absence of recombination (reviewed in Andersson & Kurland, 1998; Blanchard & Lynch, 2000). The largest number of editing events observed in a single organelle is in kinetoplastids of trypanosomes, where uridine insertion and deletion occurs in about 12 of 18 mRNA transcripts, creating start codons, frameshift corrections, and even entire open reading frames (Estévez & Simpson, 1999). As well as being the most extensive form of transcript editing, it is also the only form where RNA guides are involved.

The information for transcript editing is housed on separate minicircles in the form of guide RNA genes. Depending on the organism (see Simpson et al., 2000) there are approximately 50 maxicircles which house the mitochondrial genes, and >1000 guide RNA coding minicircles. Given that editing in general (Börner et al., 1997), the breaking of a single chromosome into several smaller pieces (Reanney, 1986) and mutational buffering through presence of multiple copies, are all expected to slow the loss of genetic information through Muller's Ratchet, and given the limited phylogenetic distribution of guide RNA-mediated uridine insertion/deletion editing (Simpson et al., 2000), this is extremely unlikely to date back to the RNA world.

Covello and Gray (1993) have introduced a three-step model for the evolution of RNA editing in general, and kinetoplastid RNA editing, the latter having been extended by Stoltzfus (1999). In kinetoplastid editing (and editing in general) it is not necessary for there to be a selective advantage for fixation of editing. It may simply arise through suitable preconditions. Stolzfus (1999) points out that recruitment of the editing machinery can be explained by tinkering, since it involves enzymes that are known in other functions. Furthermore, multiple genome copies will slow Muller's Ratchet, and redundancy can result in the accumulation and tolerance of variance between copies. Thus the emergence of a mutation (that can be neutral, slightly deleterious or lethal with only a single copy of the genome) in one copy of a given gene will always be neutral. Likewise, expression of an antisense transcript from another unaltered copy of the gene, which can bind to the mRNA produced from the mutant gene copy, has no fitness effect. Such potential precursors may arise and subsequently disappear through drift, and the same is expected for an interaction that is edited by chance. While the genotypes may differ, the phenotype for edited and unedited versions is identical, and under a neutral or even slightly deleterious model (i.e. Muller's Ratchet), both can become fixed.

As fixation at more sites occurs, while variation in the position of editing will be stochastic (for editing events where the change is neutral), the probability that all revert through back mutation is extremely low. Moreover, at functionally important sites, editing becomes maintained by natural selection (Covello & Gray 1993). This is because some editing events have become essential for production of the protein product. Loss of a key editing enzyme, which would affect all edited sites, would thus be lethal and selected against.

Strong evidence for the continuing role of neutral processes and drift in guide RNA-mediated editing includes the presence of multiple copies of both minicircles and maxicircles and large size variation for both minicircles and maxicircles across a range of organisms, large variability in minicircle copy number within strains over time and between species, presence of guide RNA genes on both minicircles and maxicircles, and existence of variant guide RNAs with mismatches in the guide regions (Simpson et al., 2000).

In summary, the suggestion that the effect of Muller's Ratchet on organellar genomes resulted in the independent evolution of unrelated forms of RNA editing in eukaryotic organelles (Börner et al., 1997), provides a strong precedent for considering uridine insertion/deletion editing to be a recently-evolved trait, and not an RNA world relic. It also underpins the evolutionary utility of RNA—where a class of RNA is limited in phylogenetic distribution and acts as a guide, it may be a recently-evolved trait.

RNA as a 'riboregulator'.

Riboregulators are RNAs that act to regulate gene expression, usually through base-pairing, and, as such, are expected to evolve readily. A number of well-understood examples are known, and a long list of possibles are currently under investigation (Erdmann et al., 2001). A number of these RNAs are included in Table 1, and an exciting finding is that 'riboregulation' is not limited to mRNA binding (as with *lin-4* and *let-7* antisense RNAs from *C. elegans*). It may also occur through other processes, such as RNA-protein interactions, as exemplified by CsrB RNA inhibition of CsrA protein activity in *E. coli* (Romeo, 1998), and meiRNA interaction with mei2 protein in regulation of meiosis in *S. pombe* (Watanabe & Yamamoto, 1994).

Another exciting prospect for the 'modern RNA world' is that unrelated RNAs have appeared in nearly identical functions, where either these functions are known to have evolved more than once, or where the evolutionary origins of the recruited RNAs can be discerned. For instance, BC1 and BC200 are RNAs with similar functions, the former having been identified in rodents (Muslimov et al., 1998), the latter being found in primates (Skryabin et al., 1998). Both appear to have a role in translation regulation in dendrites, and both apparently bind the same protein (Kremerskothen et al., 1998; Brosius, 1999). While convergence of function has yet to be conclusively demonstrated, their evolutionary origins are clear; BC1 appears to have been recruited from tRNA^{Ala}, while BC200 was originally an Alu element, a type of transposable element derived from eukaryotic srpRNA (Brosius, 1999). Given that searches have so far not yielded other such functionally analogous RNAs within mammals, yet the proteins known to make up the BC1/BC200 are conserved (Brosius, 1999), it will be interesting to see if there is evidence for non-orthologous replacement by one/both RNAs. Is RNA is inherently better suited to certain functions, being selected for over and over again for the same class of function? To this question we shall return.

RNAs in dosage compensation.

An even more dramatic example of functional convergence is emerging in studies of dosage compensation. In organisms with sex chromosomes, the number of sex chromosomes is unequal between the sexes. In *Drosophila* and mammals, males are XY, and females are XX. The unequal number of Xs means that gene expression from the X differs between the sexes, and there are mechanisms which compensate for this. In *Drosophila*, dosage is turned up in males, making expression from their single X equivalent to the two X chromosomes in females. In mammals, one X is inactivated in females, so expression is halved, making it equivalent to the single X carried by males. Furthermore, *C. elegans* takes a third strategy; expression from both copies of the X in hermaphrodites is halved relative to males (which are XY). Given multiple solutions to this problem, it is clear that mechanisms for dosage compensation have evolved more than once (Pannuti & Lucchesi, 2000; Marín et al., 2000).

In mammals and flies not only are the mechanisms of dosage compensation unrelated, they both make use of RNA for marking the X for either inactivation or upregulation, respectively (Kelley & Kuroda, 2000). The RNAs (*roX1* & roX2 in *Drosophila*, and *Xist*, which is regulated by an antisense RNA, *Tsix* in human) are unrelated, yet provide an analogous function—in both systems, RNA is thought to facilitate interaction at numerous points along the length of the target X chromosome, and the RNA genes are themselves to be found on the X chromosome. Importantly, the systems must operate via different mechanisms; in mammals, only one female X is inactivated, and it is therefore unsurprising to find that the mode of inactivation is via some mechanism that occurs exclusively *in cis*. In flies, there is no such requirement, as might be expected, given that dosage compensation is through upregulation of the single male X.

While it is still unclear how RNA is involved in these systems, it is intriguing that RNA has apparently been independently recruited to an analogous function on separate occasions. How does dosage compensation in *C. elegans* operate? Does this likewise require RNA, and indeed, in other organisms such as birds and reptiles, where sex chromosomes are different again, is dosage compensation also an RNA-dependent process?

Unclear origins of tmRNA

In bacteria, it is well established that release from ribosomal stalling on damaged mRNA is an RNA-mediated process. tmRNA, so called because of its dual role as tRNA and mRNA, allows a stalled ribosome to be uncoupled from the mRNA upon which it is stalled by virtue of the tRNA moiety of tmRNA, which is charged with alanine. The tRNA moiety accesses the A site of the ribosome and the alanine with which it is charged is then added to the partially-synthesised peptide. Next, the ribosome switches template by virtue of a conformational change in the tmRNA, and the ribosome uses the tmRNA as a template. The tmRNA encodes a string of alanines, of length 10, that labels the damaged peptide for degradation, and the ribosome is released (Keiler et al., 1996).

So far, this process has only been identified in bacteria where, it appears ubiquitous (Keiler et al., 1999). Given the dual role of the tmRNA as both tRNA and mRNA, it might be considered a candidate for the RNA world. Indeed Maizels and Weiner (1999) have speculated that such an RNA could have been the RNA world counterpart of initiator tRNA in contemporary translation. However, it is equally likely that this is a recent innovation (i.e. post LUCA) specific to the bacterial lineage. In eukaryotes, only mRNAs that possess a 5' cap structure and polyA tail pass a prerequisite quality control check before translation (Ibba & Söll, 1999). Damaged mRNAs are degraded via a nonsense-mediated decay pathway (Culbertson, 1999), reducing the production of truncated proteins during translation.

There is clearly selection for release of stalled ribosomes and tagging of damaged peptide for protein degradation in a sophisticated protein synthetic machinery, and a scenario for RNA world origins such as that suggested by Maizels and Weiner (1999) is difficult to test. What will be tractable is extending the search for tmRNA to eukaryotes and archaea. Indeed, even with quality control in eukaryote translation, mRNA may occasionally be damaged during translation, so it is possible that eukaryotes possess tmRNA. A more extensive search will thus aid in establishing whether tmRNA may have been a feature of the LUCA. Certainly, given the ubiquity of the cellular protein degradation apparatus, the proteasome (Baumeister et al., 1998; Bouzat et al., 2000), and the fact that search strategies for tmRNA identification have not yet been fully applied to eukaryotes and archaea, it will be interesting to see if stalled ribosome release occurs via a similar mechanism in these lineages.

Many naturally-occurring catalytic RNAs are not RNA world relics.

As we have already seen, not all criteria need necessarily apply for an RNA to be designated a relic, and for all but the first, the application of the criterion may not in itself provide sufficient information for the status of relic to be assigned. Criterion 4 is whether or not an RNA is catalytic. The RNA world hypothesis states that RNA catalysts pre-dated proteins in the evolution of catalysis, and the idea has been extended to a two-step transition, RNA→RNP→protein, that more accurately explains the process by which an RNA is replaced by a catalytic protein, and identifies catalytic perfection as central to understanding how come there are any ribozymes remaining at all (Jeffares et al., 1998; Poole et al., 1999).

The term catalytic RNA is most often used in a chemical sense, that is, a naked RNA that is capable of catalysis without cognate proteins. This definition excludes the peptidyl transferase activity of large subunit ribosomal RNA, eukaryotic RNase P, and spliceosomal snRNA. All are nevertheless putative RNA world relics, and in all cases, the RNA component is absolutely required for catalysis (Noller et al., 1992; Muth et al. 2000; Nissen et al. 2000; Kirsebom & Altman, 1999; Yean et al., 2000; Nilsen, 2000).

Surprisingly, the sole case where a catalytic RNA (in the chemical sense of being necessary and sufficient to carry out catalysis) can unequivocally be placed in the RNA world is that of RNase P. This has been found in all organisms examined to date, and is universally required for tRNA maturation. Bacterial RNase P has several additional substrates, including srpRNA (4.5S RNA) and tmRNA precursors (Kirsebom & Altman, 1999), and hence can be claimed under criteria 1 and 2 also. The related RNase MRP, which is involved in pre-rRNA processing in eukaryotes, is more limited in distribution, and its evolutionary origins are less clear. In considering a possible RNA world origin for RNase MRP, perhaps the most important piece of evidence is the position at which RNase MRP cleaves pre-rRNA in eukaryotes (Morrissey and Tollervey, 1995; Venema & Tollervey, 2000)—the A_3 site in eukaryotic pre-rRNA is at an equivalent position to a tRNA found in archaeal and bacterial pre-rRNAs, and Morrissey and Tollervey (1995) have argued that the tRNA has been lost from the eukaryote pre-rRNA, while cleavage at this site has been maintained. Furthermore, that RNase P is ubiquitous while RNase MRP has only been found in eukaryotes, suggests that MRP is derived from P by duplication and divergence, and bolsters the claim that the original state was tRNA processing from within pre-rRNA. While MRP may post-date the LUCA, its function in pre-rRNA processing is effectively one in the same as P in prokaryotic pre-rRNA processing.

As far as the additional substrates of bacterial RNase P are concerned, it is currently hard to establish the antiquity of these. While srpRNA is ubiquitous, the eukaryote and archaeal versions srpRNAs (7S RNAs), are not known to be processed by RNase P, and tmRNA is only known in bacteria, and, as described above, its status as an RNA world relic is uncertain. Certainly there is a precedent for post-RNA world functional diversification, as *E. coli* RNase P is also known to process phage RNAs and the polycistronic *his* operon mRNA (Altman & Kirsebom, 1999).

Another example which may clarify the discussion is the finding that there are two spliceosomes in metazoans (Tarn & Steitz, 1997; Burge et al., 1999). Both have the same origin, but the minor variant arguably arose more recently, through duplication and divergence. The function of both is identical (both excise introns from pre-mRNA, though the class of introns recognised is different), but one probably has a more recent origin (Burge et al., 1999) so in the strictest sense is not a relic, even though splicing in general arguably originated in the RNA world (see next section). In the case of RNases P and MRP, a more recent duplication and divergence event for these is possible, assuming RNase P carried out both functions initially (Morrissey & Tollervey, 1995).

These examples serve to point out that in some cases, it may difficult to separate the ultimate origin from the proximate origin. This is similar to the problem of trying to establish the ultimate origin of a family of proteins which carry out a range of functions. Where the function of an RNA has remained essentially unchanged since the RNA world, it is possible to identify the ultimate origin. In the case of MRP, the function it carries out is arguably ancient, but the origin of MRP itself cannot be unequivocally linked with this function, hence, it is unclear whether it should be assigned relic status. Morrissey and Tollervey's (1995) model best fits the data, though other scenarios can be envisaged (Collins et al., 2000).

Other naturally-occurring ribozymes, including the hammerhead, hairpin, hepatitis delta virus and neurospora VS ribozymes (Table, Symons, 1997; Carola & Eckstein, 1999) are examples of recently-evolved catalytic RNAs, since these are used in novel strategies for viral or plasmid (Neurospora VS ribozyme and Salamander hammerhead-like RNA) genome replication. It has been argued recently that all these ribozymes have a common origin (Harris & Elder, 2000), but even if this is the case, this does not require that they originated in the RNA world. That said, these ribozymes demonstrate a potential mechanism for genome replication, as well as contributing to the reconstruction of a putative RNA world. The HDV ribozyme is a particularly salient example, since it has been shown to carry out self-cleavage through general acid-base catalysis (Perrotta et al., 1999; Nakano et al., 2000), as opposed to metal ion catalysis (Westhof, 1999). Likewise, the hairpin ribozyme may also make use of general acid-base catalysis (Rupert & Ferré-D'Amaré, 2001), and excitingly, this is also the case for the peptidyl transferase subunit of the ribosome (Muth et al., 2000). The similarity to the catalytic reaction carried out by peptidyl transferase certainly establishes the relevance of these viral RNAs to catalysis in the RNA world, but also raises the point that ribozymes could have arisen multiple times in evolution with similar chemistry.

mRNA splicing and self-splicing introns.

A less clear case is presented by the group I and II self-splicing introns (Table). Broadly, the phylogenetic distribution of these two ribozymes is bacteria and eukaryotic organelles (see Figure 4 in Lykke-Andersen et al., 1997; Cech & Golden, 1999) Group I introns make use of the 3'-OH of free guanosine as nucleophile in the first step of splicing, while in group II introns, the nucleophile is provided in *cis*, and consequently, this is a 2'-OH group. Splicing in both cases is via a two step transesterification. The spliceosome, a large ribonucleoprotein complex responsible for splicing out of introns from eukaryotic nuclear pre-mRNA, also makes use of an internal 2'-OH for the first transesterification. At the core of the spliceosome are 5 small nuclear snRNAs: U1, U2, U4, U5 and U6.

A common origin of group II introns and the spliceosome has been suggested by numerous authors (e.g. Sharp, 1985, 1991, 1994; Cech, 1986; Copertino & Hallick, 1993; Stoltzfus 1999). This possibility revolves around the idea that a group II intron evolved into a 5-piece RNA complex. This idea is gaining ground, with similarities in chemical mechanism of cleavage, structurally analogous regions and ligation by a two-step transesterification (Sharp, 1985; Cech, 1986; Chanfreau & Jacquier 1994; Sontheimer et al., 1999; Gordon et al. 2000; Boudvillain et al. 2000; Yean et al.

2000). Strikingly, Hetzer et al. (1997) removed the ID3 subdomain of a group II intron, which reduced exon anchoring during ligation, and were able to reconsititute this by supplying U5 snRNA in *trans*. In addition to the direct comparisons between canonical group II and spliceosomal splicing, the feasibility of a common origin has been given support from a number of sources. Formation of group II intron structure from three separate transcripts has been observed in *Chlamydomonas reinhardii* chloroplasts (Goldschmidt-Clermont et al. 1991), demonstrating that trans-splicing can arise from cis-splicing, and that the proposal of fragmentation of a single functional RNA (as envisaged for the evolution of the spliceosome) is not without precedent. Group III introns, degenerate group II introns found as 'twintrons' (an intron within an intron) in *Euglena* chloroplast DNA, lack much of the canonical structure of group II introns, and probably require additional functions in *trans* for splicing (Copertino & Hallick 1993). Again, this has been considered as support for the possibility that the five snRNAs could have arisen from a single precursor. Moreover, Copertino et al. (1994) have described a group III twintron which excises via a lariat intermediate, analogous to the formation of a lariat in the excised spliceosomal introns.

With so much circumstantial evidence, it seems likely that the spliceosomal RNAs and group II introns have a common origin. However, such similarities may either belie a common ancestry or they might be a result of convergence owing to 'chemical determinism' (Weiner, 1993). Given that splicing always begins by nucleophilic attack of the phosphate-sugar backbone by a hydroxyl group on ribose, the different strategies used by group I and II introns (3'-OH of GTP supplied in trans versus 2'-OH of adenosine supplied in cis) might be the only two possible ways of initiating this reaction. That the spliceosome makes use of the same mechanism as group II introns could therefore be a consequence of 'chemical determinism' (and therefore convergence), not common origin (Weiner 1993). Indeed, in all three cases, splicing is carried out through two transesterifications. Chemical similarities and functional parallels provide an inroad into understanding the evolution of splicing, but given Weiner's (1993) point, they are not particularly informative in terms of distinguishing between convergence and divergence. Structural studies may help shed light on this question, in much the same way as this has resolved the question of whether the different classes of ribonucleotide reductase are convergent or divergent (Logan et al., 1999).

If it is nevertheless concluded that the similarities between group II introns and pre-mRNA splicing are sufficient to rule out convergence (that there several examples of alternative cleavage reactions available to RNA (see Westhof, 1999) in addition to those in group I and II introns might suggest this), how is the direction of evolution established? It is as conceivable that group II introns are derived from the snRNAs through fusion and reductive evolution as the possibility that snRNAs evolved from a group II intron.

In examining the evolutionary origins of splicing, there are two major questions:

- Does splicing date back the the RNA world?
- Did group II introns give rise to the snRNAs of the eukarotic spliceosome, or vice versa?

The short answer to first quesion is that an RNA world origin for splicing is likely, but the argument is over whether such splicing was group II-like, spliceosome-like, or both. In addressing the second question, it is assumed that group II and pre-mRNA splicing are related by descent. We begin with an overview of the first question, specifically with respect to the intron-exon structure of eukaryotic nuclear genes, since this has been the source of greatest controversy.

Eukaryotic pre-mRNA splicing has been argued to be an ancient process from which protein diversification by exon shuffling could have subsequently arisen (see Gilbert, 1978; Doolittle, 1978; Blake, 1978). It was argued that through the presence of splicing, discrete protein modules could have been mixed and matched, producing protein diversity from functional building blocks encoded by 'exon shuffling'. Indeed, shuffling is seen to some extent, in the form of processes such as alternative splicing, where an mRNA can be spliced in different ways to yield different products (reviewed by Graveley 2001). The implication of the 'introns-early' hypothesis for the origin of introns is that the eukaryote splicing apparatus and the intron-exon structure of genes arose very early in evolution, and were subsequently lost from prokaryote genomes. This explanation, while potentially explaining a role for splicing in protein diversification through exon shuffling, runs into two problems. First, it does not actually explain intron origins, rather, only a possible role for these in exon shuffling, *after* the advent of an intron-exon gene structure. Exon shuffling as an explanation for the origin of the intron-exon structure of genes implies that introns arose *in order to* shuffle exons. That is, it implies evolutionary forethought (Blake 1978; Doolittle, 1978). A consequence of the origin of introns might be exon shuffling, but that separates the origin of introns from the emergence of exon shuffling.

Second, the specific prediction of exon shuffling is that in at least some cases, the intron-exon structure of a gene should reflect the existence of discrete functional protein modules. Overall, the data are not strong, and even if there are cases of ancient exon shuffling, it may not be possible to detect these if intron sliding (for which there is no support [Stoltzfus et al. 1997]) is permitted (Rzhetsky et al. 1997). Indeed, the data accumulated to date (see Logsdon 1998; Wolf et al. 2000) are most compatible with the alternative theory, 'introns-late', that the 5 snRNAs of the spliceosome arose from group II introns which originated in the bacterial lineage as selfish elements, and that introns represent insertion of selfish genetic elements. Under 'introns-late', group II introns entered the eukaryote genome via the mitochondrion (members of the α-proteobacteria, which, among extant bacteria, share the most recent common ancestor with mitochondria, have been shown to possess group II introns), and this is known as the 'mitochondrial seed' hypothesis (Cavalier-Smith, 1991; Logsdon, 1998).

Importantly, phylogenetic evidence suggests that all extant amitochondrial eukaryotes once possessed mitochondria (or hydrogenosomes, which share a common origin with mitochondria - see Embley & Hirt, 1998; Rotte et al., 2000). This can be taken as evidence to support the scenario described by Logsdon (1998), since all modern eukaryotes arose from an ancestral cell which harboured an endosymbiont. Hence the advent of splicing specifically in eukaryotes could be explained by endosymbiont to host transfer of a group II intron this direction of transfer is well supported by independent evidence [Blanchard & Lynch, 2000]), followed by complexification to form the modern spliceosome.

Introns in are in fact found in all three domains. Archaeal introns are not self-splicing, but are positionally conserved with eukaryotic tRNA introns, and both make use of a conserved LAGLIDADG endoribonuclease in the cleavage and ligation reaction (Lykke-Andersen et al., 1997; Trotta & Abelson, 1999). Group I introns are found in bacteria and both the eukaryote nucleus and organelles (Lykke-Andersen et al., 1997; Cech & Golden, 1999), while group II introns are found in bacteria and eukaryote organelles (mitochondria and chloroplasts) (Logsdon, 1998). However, it is hard to argue for a common origin for the three types of intron (group I, groupII/spliceosomal, tRNA), so on phylogenetics, introns may have arisen more than once, and do not clearly date back to the RNA world. A common origin is not impossible, just not readily testable, given current data.

While many consider the introns early-late debate to be largely over, there are nevertheless shortcomings in the introns-late scenario. Furthermore, alternatives exist to exon-shuffling as an explanation for the origin of introns and the spliceosomal RNAs in the RNA world. While there are continued arguments for the validity of exon shuffling (de Souza et al., 1998), we think the evidence does not favour this scenario (see Logsdon, 1998).

That modern eukaryotes are all likely to have descended from a mitochondrion-bearing ancestor adds weight to the suggestion that the spliceosome arose specifically within that lineage subsequent to transfer of mitochondrial group II introns to the nucleus[2]. However, a serious problem for this account is that, because the model does not involve a selective advantage for the emergence of splicing, it is hard to understand how a group II intron became fragmented into five-pieces, and associated with a large number of conserved proteins. There is nothing at fault with not invoking a selective pressure in the evolution of complex structures. As described above, this has provided valuable insight into the evolution of kinetoplastid editing.

[2] For simplicity, we imply the host was a eukaryote with a nucleus, and the endosymbiont was a mitochondrion. The nature of the endosymbiont and host are currently the subject of intense debate (Andersson & Kurland, 1999; Rotte et al. 2000), but we note that on current data, it is simplest to describe the endosymbiont as mitochondrial, since it is in these organelles that group II introns have been identified (Logsdon, 1998).

An additional problem with this scenario is that it relies on inference. It cannot be directly tested using phylogenetic analyses in the same way as other mitochondrial to nucleus transfers (reviewed in Embley & Hirt, 1998; Philippe et al. 2000). This is because both sequence and structure of group II introns and spliceosomal RNAs are too divergent to be able to use either of these for phylogenetic reconstruction of their histories. Assuming group II and spliceosomal RNAs have a common origin, it is not possible to distinguish between a common origin in LUCA or transfer from mitochondrion to nucleus on the current dataset (Figure 1).

The model advocated by Logsdon (1998) requires transfer of non-fragmented group II introns to the nucleus (no examples of fragmented group II introns in mitochondria have been described) where these then insert into the host DNA, and excise during mRNA expression. Then, over time, the mechanism shifts from *cis* splicing to *trans* splicing by a complex of 5 RNAs. The first point is uncontroversial given that group II intron mobility is known (though no examples of nuclear group II introns are known) to be mediated via an intron-encoded reverse transcriptase (Lambowitz et al., 1999). The second is harder to explain. The fragmentation process was either extremely fast, predating divergence of the major eukaryote lineages, or, there was selection for the modern spliceosome over other versions, or least likely, the modern 5-piece spliceosome was fixed through drift.

No suggestions have been made regarding the second two possibilities, and the third is becoming more problematic since the previous consensus on eukaryote phylogenetics based on rRNA phylogeny (Sogin, 1991) has been challenged by the finding that microsporidia are not deep-diverging eukaryotes as per the rRNA trees, but rather are a sister group of fungi (reviewed in Keeling & McFadden, 1998). The emergence of the modern splicing apparatus must predate the diversification of eukaryotes, but is also constrained by the endosymbiosis event. In the absence of apparent selection for the origins of the spliceosome late (Stoltzfus, 1999), there ought to be spliceosomes intermediate to the 5-piece spliceosome.

A further point is that both chromosome (Backert et al., 1997; Watanabe et al., 1999; Zhang et al., 1999), gene (Estévez & Simpson, 1999) and RNA gene (Keiler et al., 2000) fragmentation is found in mitochondria and chloroplasts. A similar architecture is seen in RNA viruses, and this has been argued to be a means of slowing the accumulation of slightly deleterious mutations arising via Muller's Ratchet (Reanney, 1986). Hence, while fragmentation might be a predicted consequence of an organellar location for group II introns (no fragmented introns have been documented in free-living bacteria), it is not expected for genes located in the nucleus, given that the ratchet does not operate at the same levels as in organellar genomes (Blanchard & Lynch, 2000).

Currently there is limited information on the nature of splicing in protists. Spliceosomal introns and all five snRNAs have been identified in *Euglena gracilis* (Breckenridge et al. 1999, and references therein), *Trypanosoma brucei* and *T. cruzi* (Mair et al. 2000, and references therein). The *Giardia lamblia* genome project

(McArthur et al. 2000) is underway, and it will be interesting to see whether splicing occurs and whether snRNAs are present. Given the *Trypanosoma* and *Euglena* examples, it would be a surprise to find any protists without 5 snRNAs (unless only trans-splicing is present in which case U1 may be expected to be absent - see Breckenridge et al., 1999; Mair et al., 2000). This suggests it is at least feasible that, prior to the endosymbiosis event that gave rise to the mitochondrion, proto-eukaryotes possessed splicing.

Insertion of 'selfish' elements into genomes also deserves consideration. Insertion is not a widespread feature of prokaryotic genomes, while it varies from almost none, to extreme in eukaryotes. In extant bacteria there is good evidence for loss of any sequence that is not under immediate selection, including periodically-selected functions (reviewed in Poole et al., 2001). In bacteria the rate of genome replication is likely to be limited by a single origin of replication, and with fast response times being crucial to proliferation upon detection of an energy source, there is strong selection for sequence loss in the absence of direct selection for the sequence. In general, eukaryotes do not compete via fast reaction times, though this may be more prevalent among 'simple' eukaryotes (see Poole et al. 2001). Without such competition, there is no inherent selective disadvantage to selfish element insertion if the only consequence is an increase in genome size. With these differences, it is clear that bacterial genomes have not simply remained in some 'primitive' status quo with eukaryotes having diversified through complexification. With a precedent for loss in bacteria, it is as likely that group II introns represent the remnants of eukaryotic mRNA splicing (surviving as selfish elements through intron mobility) as the standard view that splicing has complexified in eukaryotes. Equally, if group II introns did enter eukaryote nuclear genes via the mitochondrion, invasion and proliferation is expected.

In examining the case for the spliceosome and mRNA introns in the RNA world, there are two major questions. First, what role might splicing have played in an RNA world, and second, is there any evidence for an RNA world origin? As described above, the exon shuffling theory does not explain the origin of introns, and nor is it well supported in specific and genome-wide analyses. Nevertheless, this does not preclude an RNA world origin for introns. An RNA world origin is not incompatible with the majority of introns being inserted during eukaryote evolution, and it does not require that putatively ancient introns adhere to the exon shuffling theory.

Two roles for splicing in the RNA world have been suggested. First, splicing might have been a mechanism for recombination as a buffer against accumulation of deleterious mutation (Reanney 1984; Darnell & Doolittle, 1986; Jeffares et al., 1998). Again, this role would be separate from the origin of an intron-exon structure. An explanation for the origin of splicing comes from examining the origin of chromosomes (Maynard Smith & Szathmáry, 1993; Szathmáry & Maynard Smith, 1993). At a very early stage in the evolution of the cell, genes would not have been maintained on chromosomes. The advantages of chromosomes are that, upon cell division, both daughter cells are guaranteed to receive a copy of all genes, and the spread of selfish genes that replicate faster than the other genes is limited (Maynard Smith & Szathmáry, 1993).

In the early RNA world, where gene and product were one and the same, the advent of the chromosome would have a step toward the separation of phenotype and genotype. Either transcription would have to become separated from replication (see Maizels & Weiner, 1999), or the whole chromosome would be transcribed and subsequently cut up to produce functional products (that is the chromosome and transcript are not distinguishable, unless all functional RNAs are on the same strand). Both these alternatives are likely, though the latter probably predated the former as a means of expressing RNA genes (Poole et al., 1998; 1999).

The emergence of physical linkage of genes on chromosomes in an RNA world provides a selection for splicing in the RNA world but does not explain the origins of the intron-exon structure of genes, nor whether group II introns predate the spliceosome. The emergence of an intron-exon structure may have simply been a consequence of absence of selection against the emergence of linker regions as a result of low replication fidelity. The presence of additional nucleotides at the 5' and/or 3' end might not have affected function appreciably, though there is no inherent reason for splicing to have been an inaccurate process. If it did cleave at specific sites, insertions between RNA genes resulting from low copying fidelity would not be selectively disadvantageous.

There is however a strong argument that splicing from a transcript/chromosome could not have been carried out by group II introns in the RNA world. Consider a chromosome with 5 RNA genes on it, and with group II introns between the RNA genes. Upon self-splicing of the group II introns out of the transcript copy, the 5 genes would still be unprocessed; only the group II introns will have been released from the transcript. Gilbert and de Souza (1999) have suggested that group II introns interrupted RNA genes, with splicing yielding a functional RNA. They also suggest that, with recombination, this architecture would enable RNA domain shuffling; that is, exon shuffling for RNA instead of proteins. There are examples of RNAs with introns **(e.g. U3 snoRNA, U5?),** but it is not possible to establish whether these date back to the RNA world, or represent recent insertions.

More problematically, the scenario proposed by Gilbert and de Souza (1999) requires a one gene, one chromosome model, with group II introns fulfilling a solely 'selfish' role. 'Selfish' elements are likely to be an emergent feature of any replicative system. However, for chromosomes to evolve, splicing in *trans* is required in order to express functional RNAs from a precursor transcript. Group II introns would not have provided this function, since they self-excise then splice together the two exons! Furthermore, the propensity for self-splicing introns to insert into a sequence is not a property of the RNA, but of the associated proteins (Lambowitz et al., 1999). Without a mechanism for insertion, there would be a tendency for 'selfish' self-splicing introns to be lost, since the processed chromosome would function equally well without these. In fact, without insertion, it is difficult to see how these introns could be parasitic on early RNA genomes. Hence, self-splicing introns, if they date back to the RNA world, would have had insert themselves as well as excise themselves. Given that modern group I and II introns only do the latter, it is as likely that these post-date the RNA world, arising subsequent to DNA endoribonucleases and reverse transcriptases and associated factors requried for insertion (Lambowitz et al., 1999). If tRNA introns date back to the RNA world, they have lost both splicing and insertional functions (Trotta and Abelson, 1999).

For expression of several functional RNAs from a single transcript RNA/chromosome (and assuming that these functional RNAs were not all self-splicing), what is needed is the reverse of modern day splicing (where the junk is cut out and the coding regions are spliced together). That is, in an RNA world, modern-day introns would have been the coding genes, and modern-day exons would have been the junk (Figure 2).

The brief description of the origin of chromosomes given above is not a new one, but the finding of the exact same structure in modern genomes has rekindled the argument that the intron-exon structure of genes dates back to the RNA world (Poole et al. 1998, 1999). Several eukaryotic genes are now known where the introns code for functional RNAs (small nucleolar snoRNAs), the exons being non-coding (Tycowski et al., 1996a; Bortolin & Kiss, 1998; Pelczar & Filipowicz, 1998; Smith & Steitz, 1998). In snoRNA expression in these genes, the snoRNA-containing introns are spliced out and the noncoding exons are spliced together. Gene expression from chromsomes would have been identical in the RNA world (Figure 2).

Excitingly, the production of a junk RNA from a series of non-coding exons could also solve the problem of where mRNA came from (Poole et al., 1999). In a tightly-packed genome of RNA genes, there would have been no raw material for the ribosome to act upon. However, if RNAs were excised from precursor transcripts, with the junk being spliced together, this could have provided the raw material from which protein genes arose (Figure 2). Under this model, there would be no correlation between exons and protein modules, since the proto-exons would have been continuous structures, not modular as per the exon shuffling theory.

A good number of snoRNAs are intron-encoded, with almost all vertebrate snoRNAs being intronic, and moreover, these are found in ribosomal and nucleolar proteins (Weinstein & Steitz 1999). The latter group are of particular interest, since models for the origin of protein synthesis involve a positive feedback loop: proteins stabilise and increase the accuracy of the ribosome, which makes proteins more accurately, and these further enhance the accuracy of the ribosome (see Poole et al. 1999, and references therein).

It has been variously argued that this is an ancient system (Poole et al., 1998; 1999), and that snoRNAs arose by recently by diversification (Morrissey & Tollervey, 1995; Lafontaine & Tollervey, 1998). Many snoRNAs have now been identified, and almost all are involved in rRNA processing, being essential for 2'-*O*-ribose methylations, pseudouridylations or precursor rRNA cleavage (reviewed by Weinstein & Steitz, 1999). Pre-rRNA processing can certainly be argued to be central to metabolism since it is processing of an ubiquitous RNA, as with processing of tRNA by RNase P. Nevertheless, establishing the antiquity of snoRNAs is not straightforward. Both hypotheses have their merits, and are not necessarily incompatible in all respects (Poole et al., 2000). This debate we shall consider further, and try to establish an approach that could resolve this issue.

snoRNAs

SnoRNAs are involved in extensive processing of eukaryotic rRNA (Smith & Steitz, 1997; Weinstein & Steitz, 1999), and some process spliceosomal RNAs (Tycowski et al., 1998; Jady & Kiss, 2001). Two families have been characterised, C/D and H/ACA, on the basis of sequence elements. The C/D family guides 2'-*O*-methylation of ribose, and in yeast 51 of 55 rRNA methylations have been shown to be snoRNA-guided (Lowe & Eddy, 1999). The H/ACA family snoRNAs guide isomerisation of uridine to form pseudouridine. In yeast, based on the number of pseudouridylations of rRNA (Ofengand & Fournier, 1998) the number of H/ACA snoRNAs is predicted to be comparable to C/D snoRNAs. In humans, this number is expected to be near 100 for each family, again on the basis of the number of modifications made to the rRNA (Smith & Steitz, 1997). Members of each class are also involved in cleavage of pre-rRNA during rRNA maturation (reviewed in Smith & Steitz, 1997). Recently, a 'chimeric' snoRNA, which guides both pseudouridylation and methylation on snRNA U5, has been characterised (Jady & Kiss, 2001). However, with the exception of this snoRNA, all other snoRNAs fall neatly into the two families, C/D and H/ACA.

The distribution of snoRNAs varies across the three domains. Eukaryotes contain both C/D and H/ACA family snoRNAs, involved in 2'-*O*-methylation and pseudouridylation, and representatives of both families participate in pre-rRNA cleavage (reviewed in Morrissey & Tollervey, 1995; Smith & Steitz, 1997; Lafontaine & Tollervey 1998; Smith & Steitz, 1999). Bacteria are not expected to possess snoRNA-like RNAs, having a limited number of 2'-*O*-methylations and pseudouridylations, all of which are produced by protein enzymes in bacteria studied to date (Bachellerie & Cavaillé, 1998; Ofengand & Fournier, 1998). Cleavage of pre-rRNA in bacteria is likewise carried out by proteins (Morrissey & Tollervey, 1995).

A more complex picture has emerged in archaea. Both the crenarchaea and euryarchaea possess extensive 2'-*O*-methylation of rRNA, guided by a family of small RNAs homologous to eukaryotic C/D snoRNAs (Gaspin et al., 2000; Omer et al., 2000). However, the number of pseudouridylations in archaeal rRNA is low, as per bacteria (Lafontaine & Tollervey, 1998). No homologues of H/ACA snoRNA-associated proteins have been identified, suggesting that the pseudouridylation apparatus may be protein-mediated like in bacteria (Lafontaine & Tollervey, 1998;

Charette & Gray, 2000). Less is known about the pre-rRNA processing events involving cleavage in archaea. Evidence to date suggest this aspect of pre-rRNA processing does not involve snoRNA-like RNAs, but one or more novel endonucleases (Russell et al., 1999). However, an in-*cis* snoRNA U3-like function (U3 functions in pre-rRNA cleavage in eukaryotes [see Smith & Steitz, 1997]) for sequences within the 5' external transcribed spacer of pre-rRNA has been suggested for both archaea and bacteria (Dennis et al., 1997), and homologues of the snoRNA U3-associated protein IMP4, have been identified in Archaea (Mayer et al., 2001). If snoRNA-mediated cleavage of pre-rRNA is not demonstrated in archaea, the existence of proteins homologous to the eukaryotic snoRNP-based processing system, and the existence of C/D family homologues for pre-rRNA 2'-*O*-methylation, might be best interpreted as loss from archaea, especially given that some of the eukaryotic snoRNAs involved in cleavage are C/D family members. Furthermore, if Dennis et al. (1997) are correct in their suggestion of an in-cis U3-like function for the 5'ETS, this may suggest that the snoRNA system for cleavage is, in some form, ancestral, as suggested by Jeffares et al. (1998). With the paucity of information currently available for archaeal pre-rRNA cleavage events, it is not possible to establish whether it is more like the eukaryote or bacterial pathway, or indeed, whether it is unique to the archaeal domain.

We have previously argued that both families of snoRNAs date back to the RNA world (Jeffares et al. 1998; Poole et al. 1999), while Tollervey and colleagues have argued for more recent origins, with the C/D family arising in the ancestor of eukaryotes and archaea and the H/ACA family perhaps arising in the eukaryotes, after divergence from the two prokaryotic lineages. Which scenario is correct, and how does one establish this? There are several aspects to the snoRNA problem:

- Consideration of the phylogenetic distribution of C/D and H/ACA family snoRNAs, as outlined above.
- Problems with the rooting of the tree of life, and how this may influence conclusions.
- Selection.
- That an RNA world origin for snoRNAs does not preclude recent diversification.

It is necessary to consider all aspects in any theory that attempts to account for the origin, evolution and modern distribution of snoRNAs. We shall review relevant aspects of the tree of life problem, and present a theory for the origin of snoRNAs that accounts for all the data.

Currently the interrelationships between the three domains is still in dispute, with the widely accepted monophyly of archaea and eukaryotes (Figure 3a, Iwabe et al., 1989; Gogarten et al., 1989; Woese et al., 1990) having been challenged in the light of new techniques, which suggest that the bacteria appear more divergent because of 'long-branch attraction' (Brinkmann & Philippe, 1999; Lopez et al., 1999), wherein a faster rate of evolution incorrectly groups the two slower-evolving groups (archaea and eukaryotes). Removing the 'long-branch attraction' artefact places the two prokaryotic groups together, with the root falling on the eukaryote branch (figure 3b). The traditional tree suggests that the snoRNAs arose in the common ancestor of the archaea and eukaryotes, and may or may not have been present in the Last Universal Common Ancestor (LUCA), the latter point depending on whether the bacterial rRNA processing system is ancestral or derived (Figure 3a). The newly-proposed tree places the snoRNAs in the LUCA (assuming the distribution is not a result of horizontal transfer), as they are represented in both major branches of the tree, so parsimony can be applied to argue that bacteria almost certainly lost these (Figure 3b).

Since the position of the root of the tree of life is not known with any certainty, it is difficult to establish the origin of a feature based on its distribution across the three domains. Even if the root is established, it is difficult to use this information to establish the nature of the LUCA. A feature found on both sides of the root can be argued to be present in the LUCA, assuming no horizontal transfer or convergent evolution. A feature which is present in only one lineage, e.g. H/ACA snoRNAs in eukaryotes, must be treated slightly differently however. Multiple losses are far more likely than multiple gains (as exemplified by multiple independent losses of primary synthetic pathways in parasitic and endosymbiotic bacteria [Andersson & Andersson, 1999]). Hence, if H/ACA snoRNAs are not found in archaea or bacteria, this does not rule out the possibility that it was a feature of the LUCA (Forterre, 1997; Penny & Poole, 1999).

As the tree describes the relationships between three monophyletic lineages, any argument from parsimony should be treated with caution. More importantly, even with horizontal transfer excluded (as far as we are aware, there is no evidence for horizontal transfer of snoRNAs or associated proteins), the uncertainty of the topology of the tree of life makes it uninformative (Forterre, 1997; Penny & Poole, 1999).

The problems of using the tree in establishing the evolution of the snoRNAs calls into question the robustness of Tollervey and colleagues' conclusions (Morrissey & Tollervey, 1995; Lafontaine & Tollervey, 1998) because their scenario for the origin of the snoRNAs is based on two assumptions: that the bacterial rooting of the tree of life is correct; and that the corollary of the placement of the bacterial lineage as the outgroup is that bacterial features are ancestral and those shared by archaea and eukaryotes are derived. It is currently unclear whether the bacterial rooting is the correct one, but in placement of the root in the bacterial lineage does not imply that bacterial traits are ancestral, or that shared archaeal-eukaryote traits arose post-LUCA (Forterre, 1997). This latter point does not in itself invalidate the evolutionary scheme described Tollervey and colleagues' papers, but it does cast doubt on it.

The case for snoRNAs as RNA relics.

As the tree of life cannot be used to establish the antiquity of snoRNAs, it is necessary to establish an alternative approach to examining the origin of snoRNAs. One way to do this is to establish whether there is a role for methylation and pseudouridylation in the RNA world. Both types of modification are ubiquitous, so can be argued to date back to the RNA world (Martínez Giménez et al., 1998; Cermakian & Cedergren, 1998). This suggestion is relatively uncontroversial since it is based on the ubiquity of these modifications, and on arguments for their utility prior to the emergence of protein synthesis. Pseudouridylation might have originally been selected for the increased H-bonding that is possible compared with uridine (see Ofengand & Fournier, 1998; Charette & Gray, 2000). It might therefore be important in the specification of tertiary structure, or a folding pathway. 2'-*O*-methylation alters the 2'-OH moiety of ribose, and this could have two roles. First, this modification eliminates the reactivity of the 2'-OH, so 2'-*O*-methylated ribose cannot be involved in catalytic reactions. Moreover, the addition of a methyl group will restrict the potential for hydrogen bonding at that position. Hence, 2'-*O*-methylation would prevent cross-reactivity or unwanted self-cleavage, and furthermore, influencing hydrogen bonding might specify or favour a particular folding pathway (Bachellerie & Cavaillé, 1998; Poole et al., 2000). 2'-*O*-methylation is expected to be possible without protein, consistent with a possible RNA world origin for this modification (Poole et al., 2000), though it is less clear whether pseudouridylation could be catalysed by RNA. In both cases, this could be established through *in vitro* selection experiments. A final point is that cleavage reactions analogous to those in pre-rRNA processing are known for RNA, an example being that carried out by RNases P and MRP.

The theory proposed by Tollervey and colleagues (Morrissey & Tollervey, 1995; Lafontaine & Tollervey, 1998) would require that these modifications were present in the RNA world in limited numbers (or perhaps even absent altogether), with the snoRNA apparatus only arising post-LUCA. If this argument is accepted, an explanation must be given for the very limited use of these functional groups in the RNA world and the LUCA, with emergence of high levels of rRNA methylation in archaea, and both methylation and pseudouridylation in eukaryotes. It also must explain the utility of such rRNA modifications specifically in these two groups, and not bacteria. The alternative is that modification of rRNA dates back to the RNA world, and that it was snoRNA mediated (Poole et al., 1998,1999). Protein-RNA interactions subsequently replaced the role of such modifications in folding, and in silencing sites of potential catalytic activity (Poole et al., 2000). Detailed structural information of the bacterial ribosome is now available (Muth et al., 2000; Nilssen et al., 2000, Yusupov et al., 2001), and eventually it may become possible, through comparative structures, to establish whether eukaryotic modifications serve an equivalent function to RNA-protein interactions.

If it is assumed that pseudouridylation and 2'-*O*-methylation date back to the RNA world, was relatively extensive, and that modification was either mediated or catalysed by snoRNA, an explanation for the complete absence of snoRNAs from bacteria, and of H/ACA snoRNAs from archaea must also be given.

The bacterial rooting of the tree of life, and the position of thermophiles at the base of both the archaeal and bacterial domains has been taken as evidence to support a thermophilic LUCA (Woese, 1987). However, single-stranded RNA is unstable at high temperatures, and a strong counter argument for the reduction in RNA processing, and putative RNA relics, in prokaryotes is that either the ancestor of prokaryotes was a thermophile, or, that thermophily arose twice (Forterre, 1995; Poole et al., 1998, 1999). In both scenarios, eukaryotes would never have undergone a period of adaptation to high temperatures, and the LUCA would have been a mesophile (Forterre, 1995; Poole et al., 1998, 1999). In addition to the expectation that RNA processing would be reduced during adaptation to high temperatures, circular chromosomes may also be an adaptation to high temperature, solving the problem of 'frayed ends' (Marguet & Forterre, 1994; Poole et al., 1999) and also supporting the argument that linear chromosomes and telomerase RNA is the ancestral state (Maizels & Weiner, 1999; Poole et al., 1999). Independent evidence that the LUCA was mesophilic comes from reconstruction of the ancestral GC content by comparing archaeal, bacterial and eukaryote genomes (Galtier et al., 1999). Even when mesophiles were removed from the dataset the conclusion reached was the same (Galtier et al., 1999). Finally, three independent reports have now suggested that traits contributing to hyperthermophily may have been subject to horizontal transfer (Aravind et al., 1998; Nelson et al., 1999; Forterre et al., 2000).

Neither scenario can readily explain the snoRNA data however. In addition to the roles for 2'-*O*-methylation described above, it has also been shown that this type of modification serves to stabilise RNA, and that the extent of modification is positively correlated with growth temperature in thermophilic archaea (Noon et al., 1998). If the LUCA were a thermophile, there ought to have been selection for extensive methylation in all groups, yet single-stranded RNA should not be favoured since it is thermolabile (Forterre, 1995). SnoRNA-mediated 2'-*O*-methylation is found in archaea and eukaryotes, but not in bacteria, whereas a thermophilic common origin for all three domains would predict that all three would have extensive methylation, and, if anything, eukaryotes would be the strongest candidates to have lost these. Likewise, a thermophilic ancestor for prokaryotes does not readily explain the presence of extensive methylation in archaea, and near absence in bacteria. However it can potentially explain the loss of pseudouridylation in both lineages, since there is no obvious role for this type of modification in RNA thermostability. Nevertheless, given the inconsistency with the 2'-*O*-methylation data, this is too simplistic an explanation.

As opposed to the scenario given by Lafontaine & Tollervey (1998), where C/D family snoRNAs emerged in the archaeal-eukaryote lineage, and H/ACA snoRNAs emerged in eukaryotes after divergence from archaea, we favour the following possibility.

Given a likely RNA world role for both pseudouridylation and 2'-*O*-methylation, the bacterial site-specific protein system for modification is most likely to be derived. The simplest explanation for snoRNAs is therefore that they date back to the RNA world, and hence that these were a feature of the LUCA (Poole et al., 1998, 1999). The presence of C/D family snoRNA-like sRNAs in archaea (Gaspin et al., 2000; Omer et al., 2000) and their absence in bacteria, and absence of H/ACA snoRNAs from both can be explained by the loss of snoRNAs from the bacterial lineage prior to thermoadaptation, while snoRNAs were present in the ancestors of archaea prior to thermoadaptation. In adaptation to high temperatures in general, there will be the tendency to minimise use of single-stranded RNA, owing to its instability at high temperatures, and hence RNA processing is expected to have been reduced in lineages which underwent a period of thermoadaptation. For RNA to nevertheless be maintained, there must be counter-selection for RNA protection.

We suggest that in the archaea, H/ACA snoRNAs were lost since there was selection for reduction of RNA processing, with one consequence being that extensive pseudouridylation was replaced by protein-RNA interactions. In the case of C/D snoRNAs, there was still selection for reduction of RNA processing, but 2'-*O*-methylation was selectively advantageous since it imparted greater stability on the modified RNAs. Consequently, this pathway of RNA processing was retained, though there was selection for reduction in size of C/D family snoRNAs, regularity in structure, and for maximal modification from minimal numbers of RNAs (see Omer et al., 2000), so those which performed two modifications were selected over those that directed just one modification.

In the case of bacteria, we suggest that snoRNA-mediated modifications had been lost prior to thermoadaptation, and that these had been replaced by RNA-protein interactions. The selection we have proposed for loss of RNA processing is response time in organisms competing for limited resources that fluctuate in availability (Poole et al., 1998; Poole et al., 1999). In bacteria, a fast response time is required in order to act upon detection of a nutrient source. Action requires gene expression and subsequent utilisation of that source, and the faster this is achieved, the more progeny that are produced (Carlile, 1982). Fast gene expression requires fast protein synthesis, and it is notable that in bacteria, translation begins before transcription is complete, and that ribosome assembly requires fewer steps than in eukaryotes, since there is relatively little processing of the rRNA. In eukaryotes, ribosome assembly takes much longer, and gene expression requires many processing steps, as well as export from the nucleus (see Poole et al., 1998). We therefore suggest that competition drove the streamlining of the RNA processing apparatus in the ancestors of bacteria, prior to thermoadaptation. Consequently, when bacterial lineages colonised high temperature environments, RNA-protein interactions in the ribosome provided thermostability.

In eukaryotes we favour the scenario put forth by Lafontaine and Tollervey (1998), who argue that duplication & divergence conceivably resulted in expansion of the modification snoRNAs in this lineage. Duplication and divergence is more likely to lead to new function in eukaryotes than in archaea or bacteria since in the latter two groups, the rate of genome replication is under selection. Successful individuals are not only those that respond to a new nutrient, but those that can divide the fastest (see Poole et al., 2001). Duplication events in eukaryotes are not in themselves selectively disadvantageous, and could lead to the emergence of two snoRNAs from a single ancestral snoRNA which carried out two modifications. Once this had occurred, there would be a low probability that reversion could have restored the original state. While a few eukaryote snoRNAs can mediate two modifications, the majority carry out just a single modification (Kiss-László et al., 1996; Tycowski et al., 1996b; Ni et al., 1997; Ganot et al., 1997b; Lowe & Eddy, 1999).

Duplication and divergence would also have resulted in potential for expansion of the role of snoRNAs. As has been recently documented (Cavaillé et al., 2000), some snoRNAs in mouse and human are expressed specifically in the brain, and are targeted to mRNA, possibly playing a role in the regulation of editing which produces alternative gene products. These brain-specific snoRNAs (Cavaillé et al., 2000) provide a clear example of RNAs with different proximate and ultimate origins. Even if snoRNAs are a recent development (i.e. post-LUCA), it is possible to establish the ultimate (original) function as being in rRNA processing, as this is conserved between archaea and eukaryotes.

Given that the ancestral state would be two modifications per snoRNA, this would have been maintained, or selected for in the C/D box s(no)RNAs of archaea, owing to the thermolability of RNA, whereas loss of this organisation might be an expected outcome of duplication and divergence. As for an explanation for the ancestral state being two modifications and not one, this is unclear, and indeed one evolutionary explanation may simply be that this is what emerged. An alternative possibility is that in the RNA world, two modifications (as is presumably the ancestral state for both C/D and H/ACA snoRNAs) may have represented the optimal number of modifications by a single RNA, given low coding capacity.

Conclusions.

The evidence we review here argues that new RNAs do evolve *de novo*, that this process is ongoing, and central to evolution of new cellular functions. Likewise, new RNA functions can arise through duplication and divergence. Nevertheless, it is still possible to distinguish between RNAs which arose very early in evolution and those which have a relatively recent origin. This distinction is not necessarily on the basis of function alone, and the necessarily *ad hoc* nature of this classification results in some RNAs being harder to place. However, on current evidence, and consistent with the RNA world theory (Jeffares et al., 1998), we conclude that newly-evolved RNAs do not appear to displace proteins, whereas proteins have probably replaced RNAs on many occasions during evolution.

A question of central evolutionary importance is whether, as argued by Eddy (1999), RNA may be inherently better suited to certain roles than are proteins. RNA can readily form complementary base pairs, making it effective in regulation of gene expression, guide-mediated site-specific modification, and, moreover, such functions may arise readily, for instance, through duplication and expression of an antisense RNA from the duplication. While a reasonable suggestion; proteins families have also evolved diverse specific RNA binding function. A good example is the large number of restriction-modification systems, where pairs of evolutionarily unrelated endonucleases and methylases recognise the same sequence. A common origin for a range of restriction endonucelases (Jeltsch et al. 1995; Bujnicki, 2000) demonstrates that extensive diversification is possible from a single protein.

Indeed, arguing that RNA is *inherently better* than protein runs counter to the process by which new functions evolve. There is no requirement that the molecule that becomes selected for that function is the 'best' possible for that role, and this is exactly the point of Jacob's (1977) analogy of evolution as a tinkerer, not an engineer—selection merely requires that a function confers an advantage. It does not require that only the best possible molecule is the only molecule that can come under selection.

It is not clear that RNA is inherently better than protein, even if this apparently makes intuitive sense. RNA may be more readily recruited into functions where base recognition is required, perhaps suggesting that potential antisense molecules are readily generated in cells. Proteins are able to recognise specific sequences of considerable length, and regulate gene expression through nucleic acid binding. Hence, there is not the same clear picture as for the evolution of catalysis (Jeffares et al. 1998). Notably, even with the evolution of catalysis, it is possible that some RNAs may never be replaced by proteins if the only criterion is catalytic efficiency, since it is possible for ribozymes to reach catalytic perfection, selection for a faster chemical step in catalysis will only occur when substrate diffusion is not the rate limiting step in the reaction; the larger the substrate, the slower it diffuses (Jeffares et al. 1998).

Arguments such as Eddy's (1999) lump the propensity for recruitment together with the propensity for function. In a hypothetical situation where only protein was available, no amount of tinkering would result in an RNA being selected for a given function (even though it might be better than protein) simply because there is no RNA for selection to act on.

We therefore suggest that the recruitment of either RNA or protein into new function depends on what is available, not what is best. For catalysis, where there is selection for evolution towards catalytic perfection, protein may replace RNA if an RNA cannot reach rates of catalysis where diffusion becomes the rate-limiting step, but not for a ribozyme where substrate diffusion is rate limiting (Jeffares et al. 1998). For site-specific recognition, we suggest that recruitment of RNA or protein has more to do with what is available, and that there is no evidence supporting the possibility that RNA is inherently better than protein in this role. In general, the propensity for RNA to be selected over protein in a sequence-recognition role will depend on the initial 'environment' not the inherent properties of the molecule. Where this may break down is at high temperature, where RNA will be selected against.

The snoRNAs constitute the only case where it is argued that RNA could have displaced proteins (Lafontaine & Tollervey, 1998), and, at least in respect to their role as guides for post-transcriptional modification, this is not unreasonable. The alternative scenario, that snoRNAs pre-date protein-enzymes is also feasible (Poole et al., 1999). For a resolution of this issue, two questions must be addressed. First, what is the biological function of 2'-*O*-methylated ribose and pseudouridine, the products of snoRNA-mediated modification? Second, in the context of the two theories, what selection pressures could account for the diversification of these in eukaryotes (and archaea) or the reduction of these in bacteria? Elsewhere, we have offered a selection pressure for the loss of modifications in bacteria (Poole et al., 1999). In contrast, an argument for the diversification of snoRNA-mediated modifications in eukaryotes based on selection has yet to be proposed.

Several exciting developments with respect to the evolutionary origins of snoRNAs and snRNAs are coming from the examination of the protein constituents of the RNPs. For example, the C/D family snoRNPs and U4 snRNP possess a common core protein that binds to an equivalent motif in both C/D family snoRNAs and U4 snRNA (Watkins et al., 2000; Peculis, 2000). As per the problems with establishing the evolutionary relationships between snRNAs and group II introns, it is not possible to tell whether this similarity is due to convergence or divergence from a common ancestor. Likewise, the common H/ACA motifs shared by telomerase RNA and H/ACA box snoRNAs could be divergent or convergent (Mitchell et al., 1999), as could the demonstration that both associate with the same set of core proteins (Pogacic et al. 2000; Dez et al., 2001). With respect to snRNA origins, it is interesting to note that Sm proteins have now been detected in archaea (Salgado-Garrido et al., 1999). Sm proteins are part of the spliceosome, but have recently shown to be involved in mRNA degradation (Bouveret et al., 2000). The function of Sm proteins in archaea is unknown (Salgado-Garrido et al., 1999), as is the pathway of RNA degradation in this domain.

In conclusion, information on phylogenetic distribution, together with metabolic context may provide an important test for resolving problematic data sets, such as the snoRNA data set. This is essential primarily because there is no clear way of objectively evaluating the two theories as they currently stand. A major hurdle that needs to be overcome before this approach can be reliably applied is for phylogenetics to unambiguously establish the relationships of the three domains archaea, bacteria and eukaryotes. Finally, it will also be important to test the evolutionary relationship of C/D family snoRNAs in eukaryotes and sRNAs from archaea. It is difficult to predict whether it will be possible to establish if these are related by descent, or are convergent. However the task ought to be simpler than demonstrating relationships between functionally unrelated RNAs such as H/ACA snoRNAs and telomerase RNA, U4 snRNA and C/D snoRNAs, or group II introns and the spliceosomal RNAs.

References.

Altman S, Kirsebom L. 1999. Ribonuclease P. In: Gesteland R, Cech T, Atkins J, eds. The RNA World, 2nd Ed. New York: Cold Spring Harbor Laboratory Press. pp 351-380.

Altuvia S, Zhang A, Argaman L, Tiwari A, Storz G. 1998. The Escherichia coli OxyS regulatory RNA represses fhlA translation by blocking ribosome binding. *EMBO J 17:* 6069-6075.

Andersson JO, Andersson SGE. 1999. Insights into the evolutionary process of genome degradation. *Curr Opin Genet Dev 9:* 664-671.

Andersson SGE, Kurland CG. 1998. Reductive evolution of resident genomes. *Trends Genet 6:* 263-268.

Andersson SGE, Kurland CG. 1999. Origins of mitochondria and hydrogenosomes. *Curr Opin Microbiol 2:* 535-541.

Aravind L, Tatusov RL, Wolf YI, Walker DR, Koonin EV. 1998. Evidence for massive gene exchange between archaeal and bacterial hyperthermophiles. *Trends Genet 14:* 442-444.

Bachellerie J-P, Cavaillé J. 1998. Small nucleolar RNAs guide the ribose methylations of eukaryotic rRNAs. In: Grosjean H, Benne R, eds. Modification and Editing of RNA. Washington, DC: ASM Press. pp 255-272.

Backert S, Nielsen BL, Börner T. 1997. The mystery of the rings: structure and replication of mitochondrial genomes from higher plants. *Trends Plant Sci 2:* 477-483.

Baumeister W, Walz J, Zühl F, Seemüller E. 1998. The proteasome: Paradigm of a self-compartimentalizing protease. *Cell 92:* 367-380.

Been MD, Wickham GS. 1997. Self-cleaving ribozymes of hepatitis delta virus RNA. *Eur J Biochem 247:* 741-753.

Benner SA, Ellington AD, Tauer A. 1989. Modern metabolism as a palimpsest of the RNA world. *Proc Natl Acad Sci USA 86:* 7054-7058.

Blake CCF. 1978. Do genes-in-pieces imply proteins-in-pieces? *Nature 273:* 267.

Blanchard JL, Lynch M. 2000. Organellar genes: why do they end up in the nucleus? *Trends Genet. 16:* 315-320.

Börner GV, Yokobori S-I, Mörl M, Dörner M, Pääbo S. 1997. RNA editing in metazoan mitochondria: staying fit without sex. *FEBS Lett. 409:* 320-324.

Bortolin ML, Kiss T. 1998. Human U19 intron-encoded snoRNA is processed from a long primary transcript that possesses little potential for protein coding. *RNA 4:* 445-454.

Boudvillain M, de Lencastre A, Pyle AM. 2000. A tertiary interaction that links active-site domains to the 5' splice site of a group II intron. *Nature 406,* 315-318.

Bouveret E, Rigaut G, Shevchenko A, Wilm M, Séraphin B. 2000. A Sm-like protein complex that participates in mRNA degradation. *EMBO J 19:* 1661-1671.

Bouzat JL, McNeil LK, Robertson HM, Solter LF, Nixon JE, Beever JE, Gaskins HR, Olsen G, Subramaniam S, Sogin ML, Lewin HA. 2000. Phylogenomic Analysis of the α Proteasome Gene Family from Early-Diverging Eukaryotes. *J Mol Evol 51:* 532–543.

Breckenridge DG, Watanabe Y, Greenwood SJ, Gray MW, Schnare MN. 1999. U1 small nuclear RNA and spliceosomal introns in *Euglena gracilis. Proc Natl Acad Sci USA 96:* 852-856.

Brinkmann H, Philippe H. 1999. Archaea sister-group of Bacteria? Indications from tree reconstruction artifacts in ancient phylogenies. *Mol Biol Evol 16:* 817-825.

Brosius J. 1999. RNAs from all categories generate retrosequences that may be exapted as novel genes or regulatory elements. *Gene 238:* 115–134.

Brown JW, Haas ES, Pace NR. 1993. Characterization of ribonuclease P RNAs from thermophilic bacteria. *Nucleic Acids Res 21:*671-679.

Bujnicki JM. 2000. Phylogeny of the restriction endonuclease-like superfamily inferred from comparison of protein structures. *J Mol Evol 50:* 39-44.

Cavalier-Smith T. 1991. Intron phylogeny: a new hypothesis. *Trends Genet 7:* 145-148.

Cech TR. 1986. The generality of self-splicing RNA: Relationship to nuclear RNA splicing. *Cell 44:* 207-210.

Cermakian N, Cedergren R. 1998. Modified nucleotides always were: an evolutionary model. In: Grosjean H, Benne R, eds. Modification and Editing of RNA. Washington, DC: ASM Press. pp. 535-541.

Chanfreau G, Jacquier A. 1994. Catalytic site components common to both splicing steps of a group II intron. *Science 266:* 1383-1387.

Charette M, Gray MW. 2000. Pseudouridine in RNA: What, Where, How, and Why. *IUBMB Life 49:* 341-351.

Collins LJ, Moulton V, Penny D. 2000. Use of RNA secondary structure for studying the evolution of RNase P and RNase MRP. *J Mol Evol 51:* 194-204.

Copertino DW, Hall ET, Van Hook FW, Jenkins KP, Hallick RB. 1994. A group III twintron encoding a maturase-like gene excises through lariat intermediates. *Nucleic Acids Res 22:* 1029-1036.

Copertino DW, Hallick RB. 1993. Group II and group III introns of twintrons: potential relationships with nuclear pre-mRNA introns. *Trends Biochem Sci 18:* 467-471.

Covello PS, Gray MW. 1993. On the evolution of RNA editing. *Trends Genet 9:* 265-268.

Culbertson MR. 1999. RNA surveillance: unforseen consequences for gene expression, inherited genetic disorders and cancer. *Trends Genet. 15:* 74-80.

Darnell JE, Doolittle WF. 1986. Speculations on the early course of evolution. *Proc Natl Acad Sci USA 83:* 1271-1275.

de Souza SJ, Long M, Klein RJ, Roy S, Lin S, Gilbert W. 1998. Toward a resolution of the introns early/late debate: Only phase zero introns are correlated with the structure of ancient proteins. *Proc Natl Acad Sci USA 95:* 5094–5099

Delihas N. 1995. Regulation of gene expression by trans-encoded antisense RNAs. *Mol Microbiol 15:* 411-414.

Dennis PP, Russell AG, Moniz De Sa M. 1997. Formation of the 5' end pseudoknot in small subunit ribosomal RNA: involvement of U3-like sequences. *RNA 3:* 337-343.

Dez C, Henras A, Faucon B, Lafontaine D, Caizergues-Ferrer M, Henry Y. 2001. Stable expression in yeast of the mature form of human telomerase RNA depends on its association with the box H/ACA small nucleolar RNP proteins Cbf5p, Nhp2p and Nop10p. *Nucleic Acids Res 29:* 598-603.

Doolittle WF. 1978. Genes in pieces: were they ever together? *Nature 272:* 581-582.

Eddy SR. 1999. Non coding RNA genes. *Curr Opin Genet Dev 9:* 695-699.

Embley TM, Hirt RP. 1998. Early branching eukaryotes? *Curr Opin Genet Dev 8:* 624-629.

Erdmann VA, Barciszewska MZ, Szymanski M, Hochberg A, de Groot N, Barciszewski J. 2001. The non-coding RNAs as riboregulators. *Nucleic Acids Res 29:* 189-193.

Estévez AM, Simpson L. 1999. Uridine insertion/deletion editing in trypanosome mitochondria—a review. Gene *240:* 247-260.

Forterre P. 1995. Thermoreduction, a hypothesis for the origin of prokaryotes. *CR Acad Sci Paris III 318:* 415-422.

Forterre P. 1996. A hot topic: the origin of hyperthermophiles. *Cell 85:* 789-792.

Forterre P. 1997. Archaea: what can we learn from their sequences?. *Curr. Opin. Genet. Dev. 7:* 764-770.

Forterre P, Bouthier De La Tour C, Philippe H, Duguet M. 2000. Reverse gyrase from hyperthermophiles: probable transfer of a thermoadaptation trait from archaea to bacteria. *Trends Genet 16:* 152-154.

Franke A, Baker BS. 1999. The rox1 and rox2 RNAs are essential components of the compensasome, which mediates dosage compensation in Drosophila. *Mol Cell 4:* 117-122.

Fung PA, Gaertig J, Gorovsky MA, Hallberg RL. 1995. Requirement of a small cytoplasmic RNA for the establishment of thermotolerance. *Science 268:* 1036–1039.

Galtier N, Tourasse N, Gouy M. 1999. A nonhyperthermophilic common ancestor to extant life forms. *Science 283:* 220-221.

Ganot P, Caizergues-Ferrer M, Kiss T. 1997a. The family of box ACA small nucleolar RNAs is defined by an evolutionarily conserved secondary structure and ubiquitous sequence elements essential for RNA accumulation. *Genes Dev 11:* 941-956.

Garrett TA, Pabon-Pena LM, Gokaldas N, Epstein LM. 1996. Novel requirements in peripheral structures of the extended satellite 2 hammerhead. *RNA 2:* 699–706.

Gaspin C, Cavaillé J, Erauso G, Bacherllerie J-P. 2000. Archaeal homologs of eukaryotic methylation guide small nucleolar RNAs: lessons from the Pyrococcus genomes. *J Mol Biol 297:* 895-906. [Erratum in *J Mol Biol 300:* 1017-1018.]

Gilbert W. 1978. Why genes in pieces? *Nature 271:* 501.

Gilbert W. 1986. The RNA world. *Nature 319:* 618.

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. 1989. Evolution of the vacuolar H + -ATPase: implications for the origin of eukaryotes. *Proc Natl Acad Sci USA 86:* 6661-6665.

Goldschmidt-Clermont M, Choquet Y, Girard-Bascou J, Michel F, Schirmer-Rahire M, Rochaix JD. 1991. A small chloroplast RNA may be required for trans-splicing in Chlamydomonas reinhardtii. *Cell 65:* 135–143.

Gordon PM, Sontheimer EJ, Picirilli JA. 2000. Metal ion catalysis during the exon-ligation step of nuclear pre-mRNA splicing: Extending the parallels between the spliceosome and group II introns *RNA 6:* 199-205.

Graveley BR. 2001. Alternative splicing: increasing diversity in the proteomic world. *Trends Genet 17:* 100-107.

Harris RJ, Elder D. 2000. Ribozyme relationships: the hammerhead, hepatitis delta, and hairpin ribozymes have a common origin. *J Mol Evol 51:* 182-4.

Herbert A, Rich A. 1999a. RNA processing in evolution. The logic of soft-wired genomes. *Ann N Y Acad Sci 870:* 119-132.

Herbert A, Rich A. 1999b. RNA processing and the evolution of eukaryotes. *Nat Genet 3:* 265-269.

Hetzer M, Wurzer G, Schweyen RJ, Mueller MW. 1997. Trans-activation of group II intron splicing by nuclear U5 snRNA. *Nature 386:* 417-420.

Hüttenhofer A, Kiefmann M, Meier-Ewert S, O'Brien J, Lehrach H, Bachellerie J-P, Brosius J. 2001. RNomics: an experimental approach that identifies 201 candidates for novel, small, non-messenger RNAs in mouse. *EMBO J 20:* 2943-2953.

Iwabe N, Kuma K-I, Hasegawa M, Osawa S, Miyata T. 1989. Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. *Proc Natl Acad Sci USA 86:* 9355-9359.

Jeffares DC, Poole AM, Penny D. 1995. Pre-rRNA processing and the RNA world. *Trends Biochem Sci 20:* 298-299.

Jeffares DC, Poole AM, Penny D. 1998. Relics from the RNA world. *J Mol Evol 46:* 18-36.

Jeltsch A, Kroger M, Pingoud A. 1995. Evidence for an evolutionary relationship among type-ii restriction endonucleases. *Gene 160:* 7-16.

Keeling PJ, McFadden GI. 1998. Origins of microsporidia. *Trends Microbiol 6:* 19-23.

Keiler K, Waller P, Sauer R. 1996. Role of a peptide tagging system in degradation of proteins synthesized from damaged messenger RNA. *Science 271:* 990-993.

Keiler KC, Shapiro L, Williams KP. 2000. tmRNAs that encode proteolysis-inducing tags are found in all known bacterial genomes: A two-piece tmRNA functions in *Caulobacter. Proc Natl Acad Sci U S A 97:* 7778-7783.

Kelley RL, Kuroda MI. 2000. The role of chromosomal RNAs in marking the X for dosage compensation. *Curr Opin Genet Dev 10:* 555-61.

Kiss-Lászlő Z, Henry Y, Bachellerie J-P, Caizergues-Ferrer M, Kiss T 1996. Site-specific ribose methylation of preribosomal RNA; a novel function for small nucleolar RNAs. *Cell 85:* 1077-1088.

Kremerskothen J, Nettermann M, op de Bekke A, Bachmann M, Brosius J. 1998. Identification of human autoantigen La/SS-B as BC1/BC200 RNA-binding protein. DNA *Cell Biol 17:* 751-759.

Lafontaine DLJ, Tollervey D. 1998. Birth of the snoRNPs: the evolution of the modification-guide snoRNAs. *Trends Biochem Sci* 23: 383-388.

Lambowitz AM, Caprara MG, Zimmerly S, Perlman PS. 1999. Group I and group II ribozymes as RNPs: clues to the past and guides to the future. In: Gesteland R, Cech T, Atkins J, eds. The RNA World, 2nd Ed. New York: Cold Spring Harbor Laboratory Press. pp 451-485.

Lease R, Belfort M 2000. Riboregulation by DsrA RNA: *trans*-actions for global economy. *Mol Microbiol 38:* 667-672.

Lee JT, Davidow LS, Warshawsky D 1999. Tsix, a gene antisense to Xist at the X-inactivation centre. *Nat Genet 21:* 400-404.

Lee JT, Jaenisch R. 1997. The (epi)genetic control of mammalian X-chromosome inactivation. *Curr Opin Genet Dev 7:* 274-280.

Lee RC, Feinbaum RL, Ambros V. 1993. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell 75:* 843-854.

Logan DT, Andersson J, Sjöberg B-M, Nordlund P. 1999. A glycyl radical site in the crystal structure of a class III ribonucleotide reductase. *Science 283:* 1499-1504.

Logsdon JM Jr. 1998. The recent origin of spliceosomal introns revisited. *Curr Opin Genet Dev 8,* 637-648.

Lopez P, Forterre P, Philippe H. 1999. The root of the tree of life in light of the covarion model. *J Mol Evol 49:* 496-508.

Lowe TM, Eddy SR. 1999. A computational screen for methylation guide snoRNAs in yeast. *Science 283:* 1168-1171.

Lykke-Andersen J, Aagaard C, Semionenkov M, Garrett RA. 1997. Archaeal introns: splicing, intercellular mobility and evolution. *Trends Biochem Sci 22:* 326-331.

Mair G, Shi H, Li H, Djikeng A, Aviles HO, Bishop JR, Falcone FH, Gavrilescu C, Montgomery JL, Santori MI, Stern LS, Wang Z, Ullu E, Tschudi C. 2000. A new twist in trypanosome metabolism:*cis*-splicing of pre-mRNA. *RNA 6:* 163-169.

Maizels N, Weiner AM. 1999. The genomic tag hypothesis: what molecular fossils tell us about the evolution of tRNA. In: Gesteland R, Cech T, Atkins J, eds. The RNA World, 2nd Ed. New York: Cold Spring Harbor Laboratory Press. pp 79-111.

Marguet E, Forterre P. 1994. DNA stability at temperatures typical for thermophiles. *Nucleic Acids Res 22:* 1681-1686.

Marín I, Siegal ML, Baker BS. 2000. The evolution of dosage compensation mechanisms. *BioEssays 22:* 1106-1114.

Martínez Giménez JA, Sáez GT, Seisdedos RT. 1998. On the function of modified nucleosides in the RNA world. *J Theor Biol 194:* 485-490.

Mayer C, Suck D, Poch O. 2001. The archaeal homolog of the Imp4 protein, a eukaryotic U3 snoRNP component. *Trends Biochem Sci 26:* 143-144.

McArthur AG, Morrison HG, Nixon JE, Passamaneck NQ, Kim U, Hinkle G, Crocker MK, Holder ME, Farr R, Reich CI, Olsen GE, Aley SB, Adam RD, Gillin FD, Sogin ML. 2000. The *Giardia* genome project database. *FEMS Microbiol Lett 189:* 271-273.

Mitchell JR, Cheng J, Collins K. 1999. A box H/ACA small nucleolar RNA-like domain at the human telomerase RNA 3' end. *Mol Cell Biol 19:* 567-576.

Morrissey JP, Tollervey D. 1995. Birth of the snoRNPs: the evolution of RNase MRP and the eukaryotic pre-rRNA-processing system. *Trends Biochem Sci 20:* 78-82.

Moss E, Lee R, Ambros V. 1997. The cold shock domain protein LIN-28 controls developmental timing in C. elegans and is regulated by the lin-4 RNA. *Cell 88:* 637-646.

Muller B, Schümperli D. 1997. The U7 snRNP and the hairpin binding protein: key players in histone mRNA metabolism. *Semin Cell Dev Biol 8:* 567-576.

Muslimov IA, Banker G, Brosius J, Tiedge H. 1998. Activity-dependent regulation of dendritic BC1 RNA in hippocampal neurons in culture. *J Cell Biol 141:* 1601-1611.

Muth GW, Ortoleva-Donnelly L, Strobel SA. 2000. A single adenosine with a neutral pKa in the ribosomal peptidyl transferase center. *Science 289:* 947-950.

Nakano S-I, Chadalavada DM, Bevilacqua PC. 2000. General acid-base catalysis in the mechanism of a hepatitis delta virus ribozyme. *Science 287:* 1493-1497.

Nelson KE, Clayton RA, Gill SR, Gwinn ML, Dodson RJ, Haft DH, Hickey EK, Peterson JD, Nelson WC, Ketchum KA, McDonald L, Utterback TR, Malek JA, Linher KD, Garrett MM, Stewart AM, Cotton MD, Pratt MS, Phillips CA, Richardson D, Heidelberg J, Sutton GG, Fleischmann RD, Eisen JA, Whilte O, Salzberg SL, Smith HO, Venter JC, Fraser CM. 1999. Evidence for lateral gene transfer between Archaea and Bacteria from genome sequence of *Thermotoga maritima. Nature 399:* 323-329.

Nilsen TW. 2000. RNA splicing: The case for an RNA enzyme. *Nature 408:* 782-783.

Nissen P, Hansen J, Ban N, Moore PB, Steitz TA. 2000. The structural basis of ribosome activity in peptide bond synthesis. *Science 289:* 920-930.

Noller HF, Hoffarth V, Zimniak L. 1992. Unusual resistance of peptidyl transferase to protein extraction procedures. *Science 256:* 1416-1419.

Noon KE, Bruenger E, McCloskey JA. 1998. Post-transcriptional modifications in 16S and 23S rRNAs of the archaeal hyperthermophile *Sulfolobus solfataricus. J Bacteriol 180:* 2883-2888.

Ofengand J, Fournier MJ. 1998. The pseudouridine residues of rRNA: number, location, biosynthesis, and function. In: Grosjean H, Benne R, eds. Modification and Editing of RNA. Washington, DC: ASM Press. pp. 229-253.

Omer AD, Lowe TM, Russell AG, Ebhardt H, Eddy SR, Dennis PP. 2000. Homologs of small nucleolar RNAs in Archaea. *Science 288:* 517-522.

Pannuti A, Lucchesi JC. 2000. Recycling to remodel: evolution of dosage compensation complexes. *Curr Opin Dev Genet 10:* 644-650.

Pasquinelli AE, Reinhart BJ, Slack F, Martindale MQ, Kuroda MI, Maller B, Hayward DC, Ball EE, Degnan B, Müller P, Spring J, Srinivasan A, Fishman M, Finnerty J, Corbo J, Levine M, Leahy P, Davidson E, Ruvkun G. 2000. Conservation of the sequence and temporal expression of *let-7* heterochronic regulatory RNA. *Nature 408:* 86-89.

Pelczar P, Filipowicz W. 1998. The host gene for intronic U17 small nucleolar RNAs in mammals has no protein-coding potential and is a member of the 5'-terminal oligopyrimidine gene family. *Mol Cell Biol 18:* 4509-4518.

Penny D, Poole A. 1999. The nature of the Last Universal Common Ancestor. *Curr Opin Genet Dev 9:* 672-677.

Perrotta AT, Shih I-H, Been MD. 1999. Imidazole rescue of a cytosine mutation in a self-cleaving ribozyme. *Science 286:* 123-126.

Philippe H, Germot A, Moreira D. 2000. The new phylogeny of eukaryotes. *Curr Opin Genet Dev 10:* 596-601.

Pogacic V, Dragon F, Filipowicz W. 2000. Human H/ACA small nucleolar RNPs and telomerase share evolutionarily conserved proteins NHP2 and NOP10. *Mol Cell Biol. 20:* 9028-9040.

Poole A, Jeffares D, Penny D. 1999. Early evolution: prokaryotes, the new kids on the block. *Bioessays 21:* 880-889.

Poole A, Penny D, Sjöberg B-M. 2000. Methyl-RNA: an evolutionary bridge between RNA and DNA? *Chem. Biol. 7:* R207-R216.

Poole AM, Jeffares DC, Penny D. 1998. The path from the RNA world. *J Mol Evol 46:* 1-17.

Poole AM, Phillips MJ, Penny D. 2001. Prokaryote and eukaryote evolvability. *Biosystems*, submitted.

Rastogi T, Beattie TL, Olive JE, Collins RA. 1996. A long-range pseudoknot is required for activity of the Neurospora VS ribozyme. *EMBO J. 15:* 2820–2825.

Reanney DC. 1984. RNA splicing as an error-screening mechanism. *J. Theor. Biol.* 110: 315–321.

Reanney DC. 1986. Genetic error and genome design. *Trends Genet. 2:* 41-46.

Romeo T. 1998. Global regulation by the small RNA-binding protein CsrA and the non-coding RNA molecule CsrB. *Mol Microbiol 29:* 1321-1330.

Rotte C, Henze K, Müller M, Martin W. 2000. Origins of hydrogenosomes and mitochondria. *Curr Opin Microbiol 3:* 481-486.

Rupert PB, Ferré-D'Amaré AR. 2001. Crystal structure of a hairpin ribozyme-inhibitor complex with implications for catalysis. *Nature 410:* 780-786.

Rzhetsky A, Ayala FJ, Hsu LC, Chang C, Yoshida A. 1997. Exon/intron structure of aldehyde dehydrogenase genes supports the 'introns-late' theory. *Proc Natl Acad Sci USA 94:* 6820-6825.

Salgado-Garrido J, Bragado-Nilsson E, Kandels-Lewis S, Séraphin B. 1999. Sm and Sm-like proteins assemble in two related complexes of deep evolutionary origin. *EMBO J 18:* 3451-3462.

Saville BJ, Collins RA. 1991. RNA-Mediated Ligation of Self-Cleavage Products of a Neurospora Mitochondrial Plasmid Transcript. *Proc Natl Acad Sci USA 88:* 8826–8830.

Sharp PA. 1985. On the origin of RNA splicing and introns. *Cell 42:* 397-400.

Sharp PA. 1991. "Five easy pieces". *Science 254:* 663.

Sharp PA. 1994. Split genes and RNA splicing. *Cell 77:* 805-815.

Simpson L, Thiemann OH, Savill NJ, Alfonzo JD, Maslov DA. 2000. Evolution of RNA editing in trypanosome mitochondria. *Proc Natl Acad Sci USA 97:* 6986-6993.

Skryabin BV, Kremerskothen J, Vassilacopoulou D, Disotell TR, Kapitonov VV, Jurka J, Brosius J. 1998. The BC200 RNA gene and its neural expression are conserved in Anthropoidea (Primates). *J Mol Evol 47:* 677-685.

Smit AFA. 1999. Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Curr Opin Genet Dev 9:* 657-663.

Smith HC, Gott JM, Hanson MR. 1997. *RNA 3:* 1105-1123.

Smith CM, Steitz JA. 1997. Sno storm in the nucleolus: new roles for myriad small RNPs. *Cell 89:* 669-672.

Smith CM, Steitz JA. 1998. Classification of gas5 as a multi-small-nucleolar RNA (snoRNA) host gene and a member of the 5'-terminal oligopyrimidine gene family reveals common features of snoRNA host genes. *Mol Cell Biol 18:* 6897-6909.

Sontheimer, EJ, Gordon PM, Piccirilli JA. 1999. Metal ion catalysis during group II intron self-splicing: parallels with the spliceosome. *Genes Dev 13:* 1729-1741.

Stoltzfus A. 1999. On the possibility of constructive neutral evolution. *J Mol Evol 49:* 169-181.

Stoltzfus A, Logsdon JM Jr, Palmer JD, Doolittle WF. 1997. Intron 'sliding' and the diversity of intron positions. *Proc Natl Acad Sci USA 94:* 10739-10744.

Symons RH. 1997. Plant pathogenic RNAs and RNA catalysis. *Nucleic Acids Res 25:* 2683-2689.

Tam W-Y, Steitz JA. 1997. Pre-mRNA splicing: the discovery of a new spliceosome doubles the challenge. *Trends Biochem Sci 22:* 132-137.

Trotta CR, Abelson J. 1999. tRNA splicing: an RNA world add-on or an ancient reaction? In: Gesteland R, Cech T, Atkins J, eds. The RNA World, 2nd Ed. New York: Cold Spring Harbor Laboratory Press. pp 561-584.

Tycowski KT, Shu MD, Steitz JA. 1996a. A mammalian gene with introns instead of exons generating stable RNA products. *Nature 379:* 464-466.

Tycowski KT, Smith CM, Shu M-D, Steitz JA. 1996b. A small nucleolar RNA requirement for site-specific ribose methylation of rRNA in *Xenopus. Proc Natl Acad Sci USA 93:* 14480-14485.

Tycowski KT, You Z-H, Graham PJ, Steitz JA. 1998. Modification of U6 spliceosomal RNA is guided by other small RNAs. *Mol. Cell 2:* 629-638.

Wassarman KM, Storz G. 2000. 6S RNA regulates E. coli RNA polymerase activity. *Cell 101:* 613-623.

Wassarman KM, Zhang A, Storz G. 1999. Small RNAs in Escherichia coli. *Trends Microbiol 7:* 37-45.

Watanabe KI, Bessho Y, Kawasaki, M, Hori H. 1999. Mitochondrial genes are found on minicircle DNA molecules in the mesozoan animal *Dicyema J Mol Biol 286:* 645-650.

Watanabe Y, Yamamoto M. 1994. S. pombe mei2+ encodes and RNA-binding protein essential for premeiotic DNA synthesis and meiosis I, which cooperates with a novel RNA species meiRNA. *Cell 78:* 487-498.

Watkins NJ, Segault V, Charpentier B, Nottrott S, Fabrizio P, Bachi A, Wilm M, Rosbash M, Branlant C, Lührmann R. 2000. A common core RNP structure shared between the small nucleoar box C/D RNPs and the spliceosomal U4 snRNP. *Cell 103:* 457-466.

Weiner AM. 1993. mRNA splicing and autocatalytic introns: distant cousins or the products of chemical determinism? *Cell 72:* 161–164

Weinstein L, Steitz JA. 1999. Guided tours: from precursor snoRNA to functional snoRNP. *Curr Opin Cell Biol 11:* 378-384.

Westhof E. 1999. Chemical diversity in RNA cleavage. *Science 286:* 61-62.

Wightman B, Ha I, Ruvkun G. 1993. Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans. *Cell 75:* 855-862.

Woese CR. 1987. Bacterial evolution. *Microbiol Rev 51:* 221-271.

Woese CR, Kandler O, Wheelis ML. 1990. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eukarya. *Proc Natl Acad Sci USA 87:* 4576-4579.

Wolf YI, Kondrashov FA, Koonin EV. 2000. No footprints of primordial introns in a eukaryotic genome. *Trends Genet 16:* 333-334.

Yean S-L, Wuenschell G, Termini J, Lin R-J. 2000. Metal-ion coordination by U6 small nuclear RNA contributes to catalysis in the spliceosome. *Nature 408:* 881-884.

Yusupov MM, Yusupova GZ, Baucom A, Lieberman K, Earnest TN, Cate JH, Noller HF. 2001. Crystal structure of the ribosome at 5.5 Å resolution. *Science 292:* 883-896.

Zhang A, Altuvia S, Tiwari A, Argaman L, Hengge-Aronis R, Storz G. 1998a. The OxyS regulatory RNA represses rpoS translation and binds the Hfq (Hf-I) protein. *EMBO J 17:* 6061-6068.

Zhang F, Lemieux S, Wu X, St.-Arnaud D, McMurray C, Major F, Anderson D. 1998b. Function of hexameric RNA in packaging of bacteriophage ϕ29 DNA in vitro. *Mol Cell 2:* 141-147.

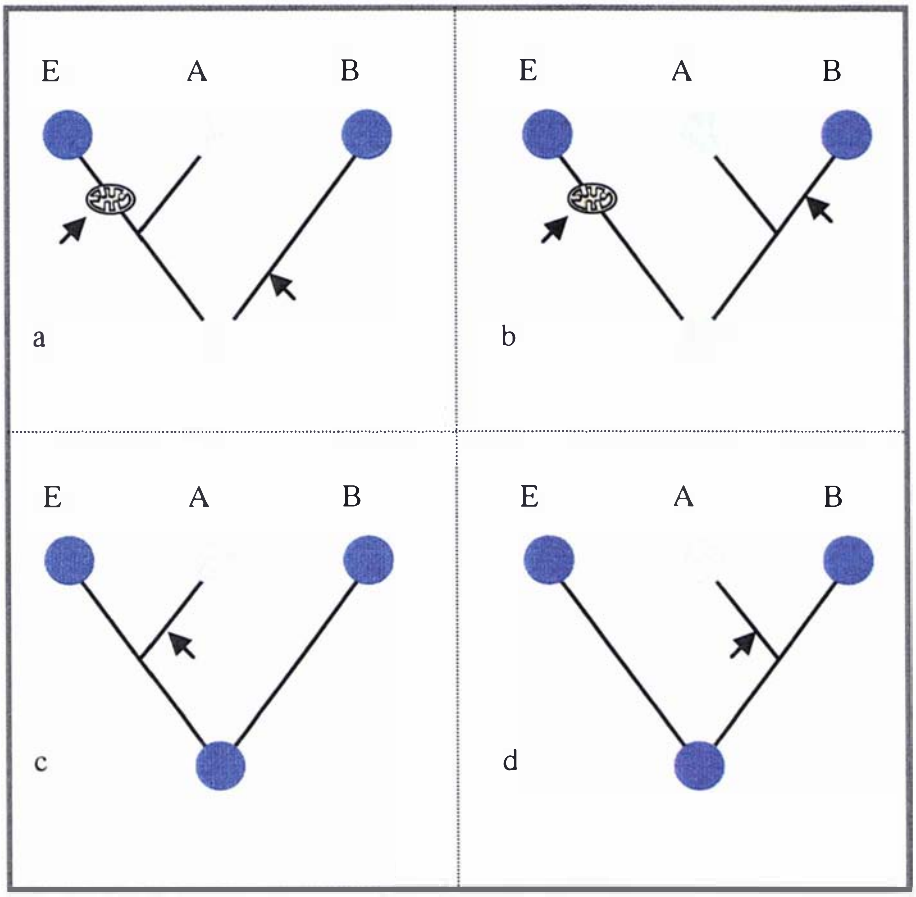
Zhang Z, Green BR, Cavalier-Smith T. 1999. Single gene circles in dinoflagellate chloroplast genomes. *Nature 400:* 155-159.

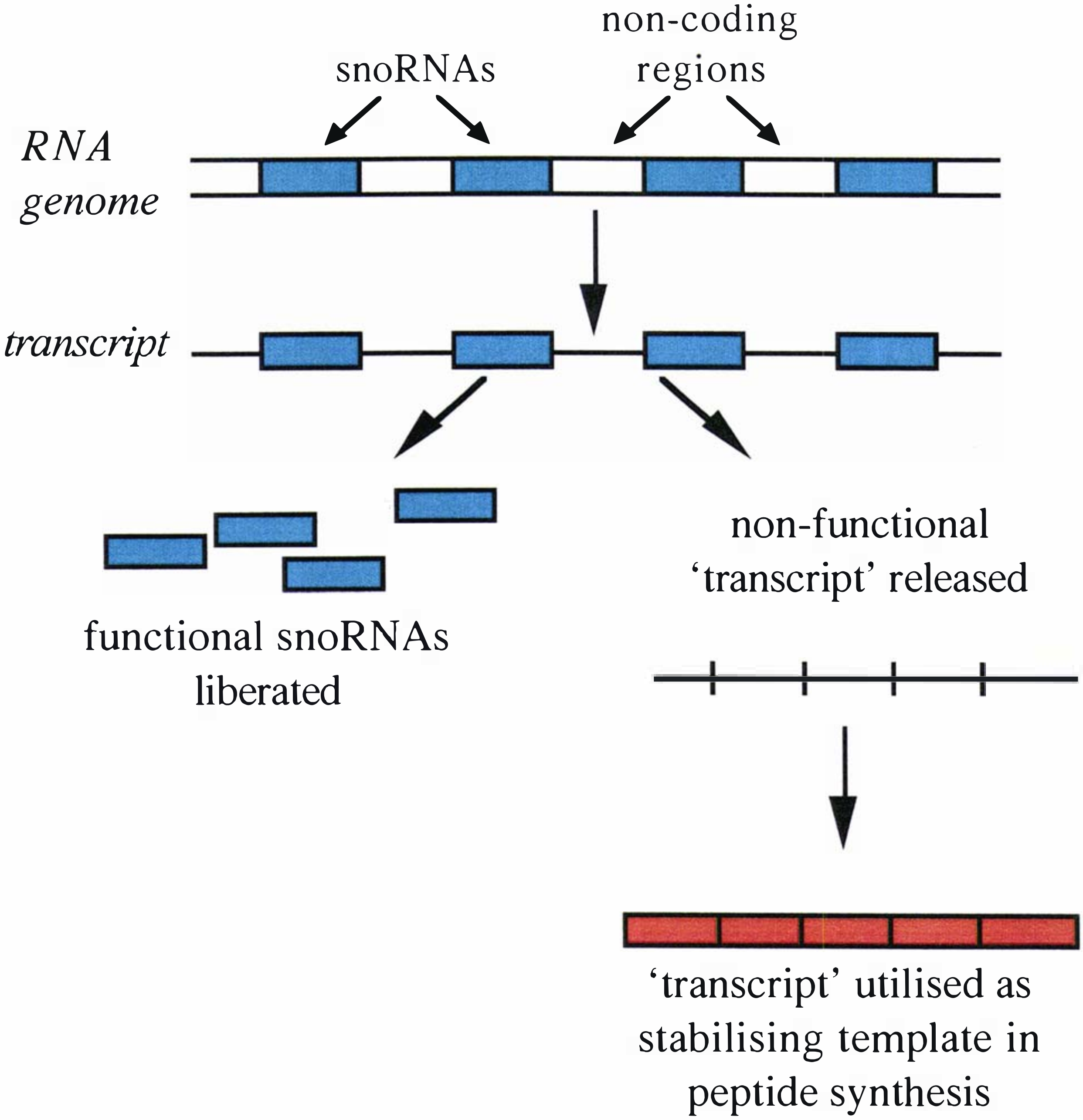
Figure legends.

Figure 1. Problems for inferring ancestry of group II introns and spliceosomal RNAs from the tree of life.
In a and c, the bacterial rooting is shown, in b and d, the eukaryote rooting is shown. Blue dots represent group II introns and spliceosomal RNAs, grey dots denote absence of these. The position of the root does not allow evaluation of the different trees with simple parsimony since trees a and b show origin of spliceosomal RNAs through 'seeding' from the mitochondrion. Consequently all four trees are equally likely. Testing the 'seed' hypothesis by examining the bacterial distribution of group II introns will be inconclusive for two reasons. First, group II introns are mobile, and second, limited distribution can equally be explained by polyphyletic losses. Likewise, ubiquity of group II introns in bacteria cannot be taken as support for a common ancestor of group II introns and spliceosomal RNAs in the LUCA, since the former are mobile elements. Finding group II introns in archaea can also be ambiguously interpreted.

Figure 2. Introns first hypothesis.
The final step in the origin of genetically-encoded protein-synthesis is presumed to be the origin of mRNA. We propose that the non-coding 'transcripts', produced as a by-product in the processing of precursor transcripts containing functional RNAs (such as snoRNAs), were the source of the first genetically-encoded proteins. These were utilised by the proto-ribosome to stabilise the interaction between two charged tRNAs, during non-genetically-encoded peptide synthesis. As primary sequence structure appears unimportant for non-specific RNA-binding, we propose that the first proteins produced in this manner were not catalytic, and could retain function despite a high mutation rate in the genomic sequence. Hence, we postulate that it was by virtue of the coupling of cleavage and ligation (a transesterification) in the proto-spliceosome that the first genetically-encoded proteins arose.

Figure 3. SnoRNAs in the LUCA?
The suggestion that snoRNAs date back to the RNA world may be independently examined depending on the placing of the root of tree of life. Currently the position of the root is unresolved, with bacterial and eukaryote rootings being considered as possibilities. A. If the bacterial rooting is correct, it is not possible to establish from the tree alone if the LUCA possessed snoRNAs. B. If the eukaryote rooting is correct, the most parsimonious explanation is that the LUCA contained snoRNAs, since these are then found on both sides of the root. The position of the root is in dispute, and since the rooting drastically affects the utility of the tree, it is difficult to use phylogenetic distribution to resolve the debate. Until a consensus is reached, biochemical arguments have to be relied upon (see text).


E
A
B
a
E
A
B
b
E
A
B
c
E
A
B
d


snoRNAs
non-coding regions
RNA genome
transcript
functional snoRNAs liberated
non-functional 'transcript' released
'transcript' utilised as stabilising template in peptide synthesis

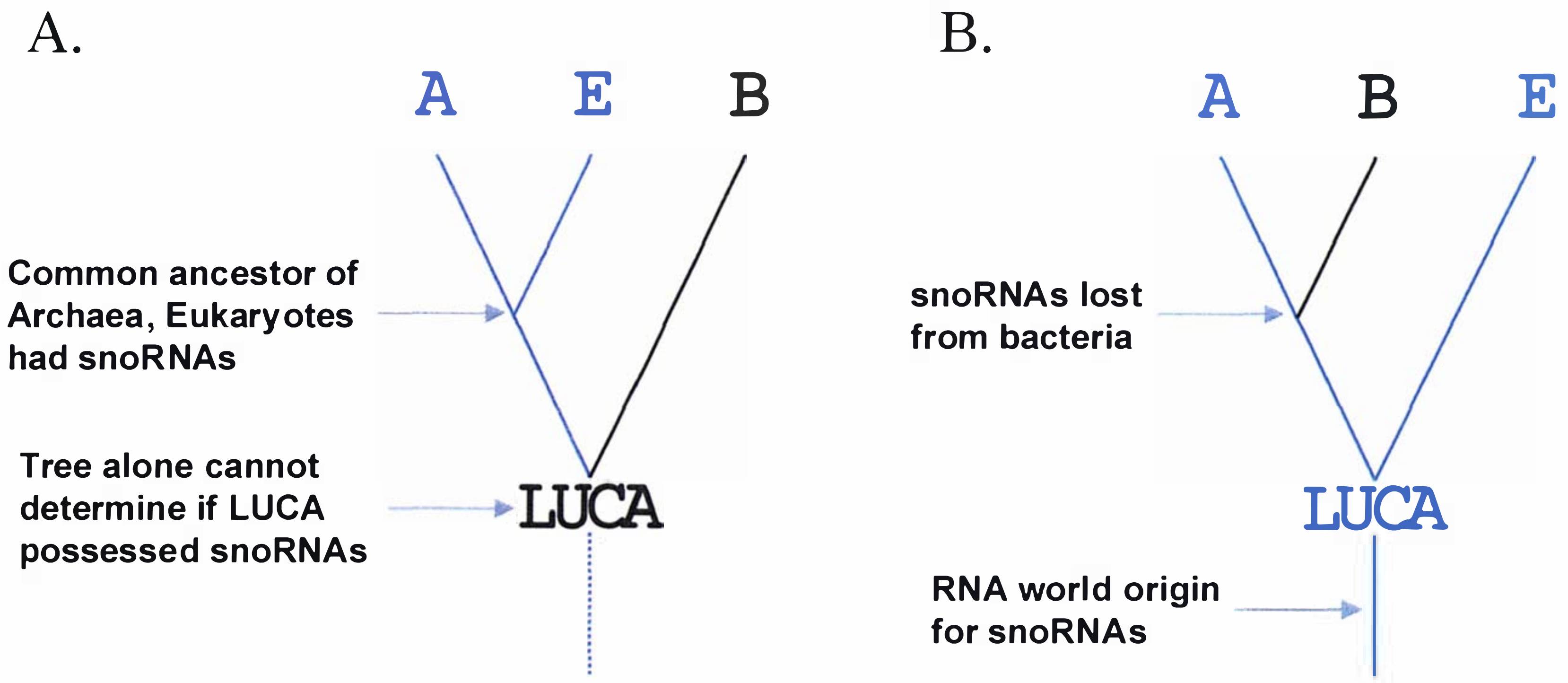

A.
A
E
B
Common ancestor of Archaea, Eukaryotes had snoRNAs
Tree alone cannot determine if LUCA possessed snoRNAs
LUCA
B.
A
B
E
snoRNAs lost from bacteria
LUCA
RNA world origin for snoRNAs

Table 1. Candidate post-RNA world RNAs.

RNA	Distribution	Function	Comments	References
roX1 & roX2	*D. melanogaster*	Dosage compensation	*roX1* & *roX2* are unrelated, and neither are related to *Xist* or *Tsix*. *Tsix* is an antisense regulator of *Xist*.	Franke & Baker, 1999.
Xist & Tsix	Mammals			Lee et al., 1999; Kelley & Kuroda, 2000.
BC 200	Primates	Translation regulation in dendrites	BC1 and BC200 are unrelated, but may be serve analogous roles. Both bind a protein homologous between Primates and Rodents.	Skryabin et al., 1998.
BC 1	Rodents			Muslimov et al., 1998; Kremerskothen et al., 1998.
lin-4	*C. elegans, C. briggsae*	Antisense regulator of *lin-14* and *lin-28*.		Lee et al., 1993; Wightman et al., 1993; Moss et al., 1997.
let-7	Bilaterian animals	Antisense regulator of *lin-41* probably in late temporal transitions in development.		Pasquinelli et al., 2000.
OxyS RNA	*E. coli*	Oxidative stress-indiced antisense global inhibitor of translation initiation.		Altuvia et al., 1998; Zhang et al., 1998a.
DsrA RNA	*E. coli*	Antisense regulator of translation initiation of global transcription regulators H-NS	Inhibits H-NS translation, but stimulated RpoS translation, acting through RNA-RNA	Lease & Belfort, 2000

		and RpoS.	interactions.	
MicF RNA	Gram-negative bacteria	Activator of translation initiation of OmpF		Delihas, 1995.
DicF RNA	*E. coli*	Antisense regulator in cell division.		Delihas, 1995.
meiRNA	*Schizosaccharomyces pombe*	Regulation of meiosis		Watanabe & Yamamoto, 1994; Ohno & Mattaj, 1999.
tmRNA	Bacteria	Ribosome/mRNA/protein release		Keiler et al., 1996; Keiler et al., 2000.
CsrB	*E. coli, Erwinia carotovora*	Binds and inhibits CsrA global regulatory protein		Romeo, 1998.
G8 RNA	*Tetrahymena thermophila*	Establishment of themotolerance.		Fung et al., 1995.
6S RNA	*E. coli*	Modulation of RNA polymerase activity		Wassarman & Storz, 2000.
gRNAs	Kinetoplastids of trypanosomes	Editing of mRNA transcripts	RNA editing by guide RNA argued to be ancient, but is most probably an adaptation to Muller's ratchet (see text).	Estévez & Simpson, 1999; Simpson et al. 2000.

Bacteriophage φ29 RNA	Bacteriophage φ29	RNA hexamer required for DNA packaging		Zhang et al., 1998b.
Hammerhead ribozymes	Plant pathogenic RNAs Salamander nuclear DNA	Genome replication Transcript processing		Symons, 1997; Garrett et al., 1996.
Hairpin ribozyme	Plant pathogenic RNAs	Genome replication		Symons, 1997.
Hepatitis delta virus ribozyme	Hepatitis delta virus	Viral genome replication		Been & Wickham, 1997.
Neurospora VS ribozyme	*Neurospora*	Transcript processing in mitochondrial DNA plasmid		Saville et al., 1991; Rastogi et al., 1996.
U7 snRNA	Metazoa	Histone pre-mRNA processing	While histones are found in Archaea, the limited distribution of U7 suggests it arose in eukaryotes, though more data are needed.	Muller & Schümperli, 1997.
Group I introns	Eukaryotic organelles & nucleus, Phage, Bacteria	Mobile selfish elements	Catalysis is via 3'OH of guanosine, supplied *in trans*, a mechanism distinct from group II/spliceosomal catalysis.	Cech & Golden, 1999; Lykke-Andersen et al., 1997.
Group II introns	Phage, Eukaryote organelles, Bacteria	Mobile selfish elements	Argued to be either evolutionarily related to the spliceosome or evolved	Logsdon, 1998 Cech & Golden, 1999.

			recently *de novo* (see text).	
Diversity of C/D & H/ACA snoRNAs	Eukaryote nucleolus	Cleavage, methylation & pseudouridylation of rRNA, and probably other RNAs.	C/D box family are found in Archaea also. Opinion is divided on whether these are RNA world relics (see text).	Weinstein & Steitz, 1999; Omer et al., 2000.
		Some C/D snoRNAs appear to be involved in regulation of brain-specific gene expression.	These snoRNAs are most probably recent innovations.	Cavaillé et al., 2000.

Paper 4

Poole A, Jeffares D, Penny D.

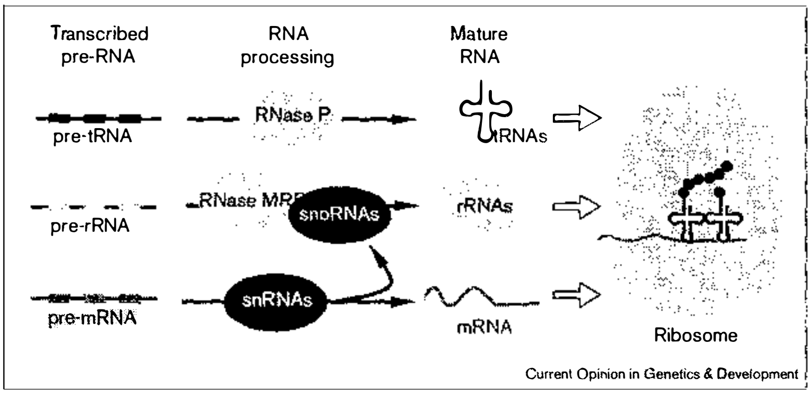
Early evolution: prokaryotes, the new kids on the block.

Bioessays 21, 880-889 (1999).

Paper 5

Penny D & **Poole A.**
The nature of the Last Universal Common Ancestor.
Current Opinion in Genetics & Development 9, 672-677 (1999).
Reprinted with permission from Elsevier Science.

Figure 3



The RNA processing pattern in eukaryotes reflects that of the LUCA. An examination of RNAs involved in translation reveals a striking pattern. Precursor RNAs are processed by RNPs (ribonucleoproteins–RNA plus cognate protein) to yield mature RNAs. Furthermore, RNPs process other RNPs – snoRNAs are released by sn RNAs, the RNA component of the splicing machinery, which in turn are crucial for rRNA processing. In prokaryotes, some of these RNAs have been lost (shaded region), and indeed, in the case of pre-mRNA, the processing step has been lost completely. Eukaryotes have retained a more complete record of the supposed RNA-world processing pathway than have prokaryotes.

of life between eukarya and archaea-bacteriais consistent with the conclusion that the genome architecture of the LUCA more closely resembled that of eukarya.

Thermoreduction and prokaryote origins

In postulating the nature of the LUCA, it is essential to consider the selective forces that would give rise to either prokaryotes or eukaryotes. Two selective forces that reinforce each other have been proposed by which prokaryotes could have evolved from an ancestor containing a eukaryote-like genome: thermoreduction and *r*-selection, [20••,21••,24]. *r*-selected organisms are fast-growing, competing for nutrient sources which fluctuate greatly in abundance. Yeast is *r*-selected when compared to an oak tree, which grows slowly, has a slow generation time and a fairly constant nutrient source (and is thus K-selected), and prokaryotes are *r*-selected relative to eukaryotes. *r* selection generally results in extremely fast and efficient use of resources, because limited availability produces strong competition for these. At the molecular level, the result is that enzymes that affect metabolite utilisation and organismal growth rate will be driven toward perfection at a faster rate than in organisms not under *r* selection. Thus, *r* selection may at least account partially for the observed replacement of RNA enzymes by protein in the prokaryote lineages [20••,21••].

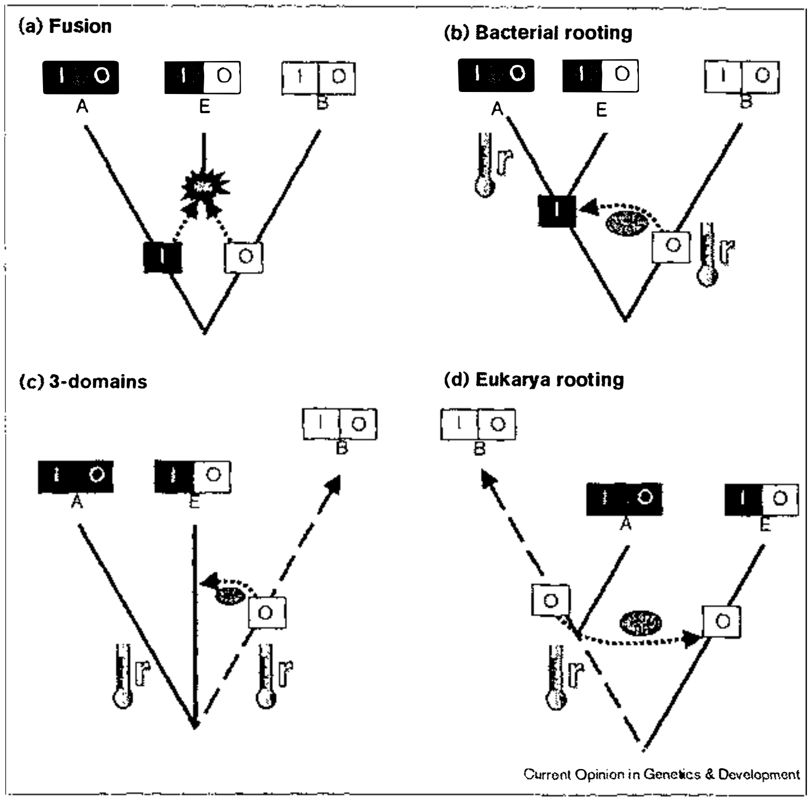
The thermoreduction hypothesis [24] is that prokaryotes arose from mesophiles by adaptation, via the loss of thermolabile traits, to high-temperature environments. This explains the loss of the ssRNA processing pathways (Figure 3) dating back to the RNA-world. Single-stranded RNA is heat labile, and would have been the Achilles' heel of early thermophiles. Accelerating ssRNA processing (mRNA, tRNA and rRNA) from hours (eukaryotes) to minutes (prokaryotes) would increase the viability of an organism at high temperatures. This loss of pre-mRNA processing, as well as the replacement of snoRNA-mediated rRNA processing with a protein enzyme system, would have been important steps in the evolution of thermophily. Unlike RNA, proteins are capable of extreme thermostability [25]. Furthermore, circular chromosomes are more thermostable than linear [26].

Other important molecules, such as glutamine [27] and carbamoyl phosphate [28], are also thermolabile. Glutamine is a protein amino.acid and major nitrogen donor whereas carbamoyl phosphate is a crucial intermediate in the formation of pyrimidines and arginine. Pathways where carbamoyl phosphate and/or glutamine are used may have been affected by thermoreduction. For instance, in the hyperthermophilic archaeon *Pyrococcus furiosus*, carbamoyl phosphate is used immediately after synthesis by metabolite channelling, and has ammonia rather than glutamine as amino donor [28]. A second example of metabolite channelling is mischarging of glutaminyl–tRNA with glutamate, thereby making glutamine synthesis the final step before incorporation into protein; this is widespread within the prokaryotes but absent from eukaryotes [20••]. Although the area requires more investigation, the distribution of these traits in archaea and bacteria is predicted by the thermoreduction hypothesis.

Another dataset consistent with the LUCA being mesophilic comes from reconstructions of ancestral GC content. Galtier *et al.* [29••] have estimated its GC content and find it much lower than that characteristic of thermophiles. Moreover, a comparable result was obtained using only the thermophiles in their dataset. All work involving ancient sequence comparisons needs to be rigorously scrutinised but, in light of all the above data, the result is compelling nonetheless. In addition, that nucleotides themselves are unstable at high temperatures [30•] is consistent with a more mesophilic origin of life.

Overall, the thermoreduction hypothesis predicts a mesophilic LUCA with a genome and RNA-processing system more characteristic of eukarya. The power of the thermoreduction hypothesis is that it predicts a range of phenomena, rather than relying on *ad hoc* explanations of individual phenomena. Fossil dates do not contradict this picture because rocks from 2700

Figure 4



Fitting the data to the trees. Given our current understanding, several alternative trees could fit the data without altering either main conclusion. These are that the eukarya retain the greatest amount of biochemical similarity to the LUCA and that the prokaryotes have been through a period of reductive evolution, mainly through evolving to life at high temperatures. Some possible trees are as follows (episodes of thermoreduction and the origin of mitochondria are indicated). **(a)** The origin of eukarya (E) by fusion of a bacterium (B) and an archeon (A) fits the informational (I) and operational (O) gene distribution but is hard to fit all the data. It does not explain the origin of the nuclear membrane, however, which is assembled and disassembled during cell division, quite unlike organellar membranes (see [21••]). **(b)** Rooting the tree in the bacterial branch fits the data provided the biochemistry of the LUCA is understood to be more closely similar to that of modern eukaryotes than that of eubacteria. A bacterial rooting would require that the archaea and bacteria arose independently via r-selection and thermoreduction. **(c)** The classic 3-domain tree can also fit the data, provided the greater divergence of bacterial informational genes can be ascribed to higher rates of evolution. There would be transfer of operational genes back into eukarya through endosymbiosis. **(d)** The tree where the root is on the eukarya branch is perhaps the simplest with respect to the biochemical data. It is consistent with all the other data, provided (as for [c]) that the bacterial informational genes are indeed evolving at a faster rate.

Mya appear to have organic molecules characteristic of both prokaryotes and eukaryotes retained [31••].

Integrating data from genomes

Although data gleaned from biochemical approaches allows tentative reconstruction of the 'bare bones' LUCA, whole genomes will ultimately uncover much more information. Genomics allows metabolic traits to be compared through the presence or absence of genes, and by sequence comparisons. However, simple comparison of the presence or absence of homologous genes does not take into account the problems of gene loss or acquisition by horizontal transfer. Initial reconstruction of the 'minimal gene set' [32] highlights this caveat: being criticised because it resulted in exclusion of *de novo* pathways for deoxyribonucleotide synthesis, leading the authors to conclude that the LUCA had an RNA genome [33].

There is a difference between reconstructing the minimal gene set for cellular life, and the set of genes which the LUCA had. Greater caution is required when examining all three domains, as eukaryotes received prokaryotic genes subsequent to the endosymbioses of mitochondria and chloroplasts [34••,35••]. Replacement of unrelated, distantly related, or paralogous genes by functional counterparts is 'non-orthologous displacement' [36] and 'may' be central to understanding how the existing distribution of genes has arisen.

If we expect a eukaryote-like genome for LUCA as a starting point, how does this then fit with the data on operational and informational genes (Figure 1)? It is necessary to identify the direction of transfer. The complexity hypothesis [3••] places limits on gene transfer, such that we expect the transfer of mostly the operational genes in explaining the apparent chimerism. It has been suggested that acquisition of prokaryotic operational genes by eukaryotes results from their diet [37•]. There is no apparent selective advantage to such uptake, however, even though the mechanism might contribute to gene acquisition.

Another possibility is that the eukarya received the largest number of bacterial operational genes from the mitochondrion [38••]. Two established evolutionary mechanisms together favour this and are compatible with a eukaryal root: the increased rate of evolution toward catalytic perfection under r-selection [19••,21••], and Müller's ratchet.

Müller's ratchet is the term given to the continual accumulation of slightly deleterious mutations in lineages lacking recombination. It has been shown that Müller's ratchet is active in organelles [39] and that it drives the gene loss there (and also from obligate intracellular parasites) [34••,35••]. Most importantly, relocation of organellar genes to the nucleus benefits both host and symbiont. If the action of Müller's ratchet on the organelle drives gene loss, this can compromise the host-endosymbiont relationship and thus there is selection to relocate useful genes to the nucleus, where mutation rate is lower. The majority of endosymbiont genes were not expected to fit this category, however, and it was assumed these were lost over time, since equivalent functions already resided in the nucleus; but the simplest explanation of the evidence is that many were transferred [38••].

Figure 4 illustrates that the bioinformatic data, the RNA relic data, plus the evolutionary mechanisms that gave rise to the three domains can *still* fit several trees. Thus even with the nature of the LUCA, the branching order of the universal tree is not yet sufficiently informative to resolve all the issues. This is because each domain is a monophyletic group, so the basal branches of the tree (dividing the domains) can only take on a very limited number of trees. Hence, the metabolic data set cannot be used as an unambiguous outgroup for rooting the tree.

Conclusions

An interesting picture of the LUCA is emerging. It was a fully DNA and protein-based organism with extensive processing of RNA transcripts by RNPs (Figure 3). It had an extensive set of proteins for DNA, RNA and protein synthesis, DNA repair, recombination, control systems for regulation of genes and cell division, chaperone proteins, and probably lacked operons. Biochemistry favours a mesophilic LUCA with eukaryote-like RNA processing, though it is still possible to fit the data to several different trees (Figure 4). A eukaryote-like LUCA is not a new idea and can be traced back to Reanney [40].

Details of energy source(s) are unclear, partly because operational genes apparently undergo frequent horizontal transfers. Comparative genomics promises a clearer picture but apparent intermingling of lineages via horizontal transfer is a major obstacle [38••]. Increasingly, models need to fit our understanding of evolutionary theory and population genetics -it is essential to have plausible mechanisms and selective forces. The extent and direction of horizontal gene transfer needs accurate estimates before concluding the theory of descent does not hold for the earliest divergences [8,42,43]. Nevertheless, it is unclear whether the LUCA was a single 'species' or whether there was extensive horizontal transfer between divergent life forms. An outstanding issue is the origin of nuclear/cytoplasmic compartmentation as the concentration of RNA relics within the nucleus suggests this organelle is more ancient than previously supposed.

Acknowledgements

We thank H Philippe, P Forterre, H Brinkmann and P Lopez for sending manuscripts prior to publication.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. •• Doolittle WF: **Phylogenetic classification and the universal tree.** *Science* 1999, **284**:2124-2128.

A recent review of developments in the fields of phylogenetics and bioinformatics as applied to the question of the root of the tree of life. An important aspect is the discussion of horizontal transfer, how this could affect the search for the root, and the issue of whether informational genes could potentially transfer between lineages as readily as operational genes are suggested to.

2. Woese CR, Kandler O, Wheelis ML: **Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eukarya.** *Proc Natl Acad Sci USA* 1990, **87**:4576-4579.

3. •• Jain R, Rivera MC, Lake JA: **Horizontal transfer among genomes: the complexity hypothesis.** *Proc Natl Acad Sci USA* 1999, **96**:3801-3806.

The authors argue for extensive gene transfer between prokaryotes during evolution and that it is genes of the operational class that are transferred most frequently. They suggest that transfer of informational genes is hindered by the many intermolecular interactions in which these macromolecules are involved. Informational genes include those for transcription, translation, replication, and GTPases. Operational genes are those for nearly all of metabolism, including regulation.

4. • Nelson KE, Clayton RA, Gill SR, Gwinn ML, Dodson RJ, Haft DH, Hickey EK, Peterson JD, Nelson WC, Ketchum KA *et al.*: **Evidence for lateral gene transfer between Archaea and Bacteria from genome sequence of *Thermotoga maritima.*** *Nature* 1999, **399**:323-329.

Another whole microbial genome sequence from TIGR. This paper considers especially the issue of horizontal gene transfer, concluding on the basis of conserved gene order that some horizontal transfer occurs between eubacteria and archaea.

5. • Aravind L, Tatusov RL, Wolf YI, Walker DR, Koonin EV: **Evidence for massive gene exchange between archaeal and bacterial hyperthermophiles.** *Trends Genet* 1998, **14**:442-444.

Describes evidence that horizontal transfer from hyperthermophilic archaea to hyperthermophilic bacteria occurs more readily than to mesophilic bacteria. The authors conclude that this transfer may have been the defining event in the origin of hyperthermophilic bacteria.

6. •• Rivera MC, Jain R, Moore JE, Lake JA: **Genomic evidence for two functionally distinct gene classes.** *Proc Natl Acad Sci USA* 1998, **95**:6239-6244.

Using whole-genome data, the authors class genes as either operational or informational on the basis of function and demonstrate that the operational gene sets of bacteria and eukaryotes are more closely related than that of the archaea, whereas the archaea–eukaryote grouping holds for the informational gene set.

7. •• Snel B, Bork P, Huynen MA: **Genome phylogeny based on gene content.** *Nat Genet* 1999, **21**:108-110.

The authors use the gene content of 13 completely sequenced genomes for reconstructing the tree of life and rooting it. Unlike sequence-based phylogenies, the tree is built by examining similarities and differences in gene content, so that the presence or absence of a gene is counted as a character. The authors conclude that massive horizontal transfer events between distant groups is not supported by their results, and that their data largely support the 16S rRNA tree topology for the 13 genomes.

8. Woese CR: **The last universal common ancestor.** *Proc Natl Acad Sci USA* 1998, **95**:6854-6859.

9. 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.

10. Gogarten JP, Kibak H, Dittrich P, Taiz L, Bowman EJ, Bowman BJ, Manolson MF, Poole RJ, Date T, Oshima T *et al.*: **Evolution of the vacuolar H+-ATPase: implications for the origin of eukaryotes.** *Proc Natl Acad Sci USA* 1989, **86**:6661-6665.

11. •• Brinkmann H, Philippe H: **Archaea sister-group of Bacteria? Indications from tree reconstruction artifacts in ancient phylogenies.** *Mol Biol Evol* 1999, **16**:817-825.

Uses a new method which applies the 'covarion' model to building a tree of life from signal recognition particle proteins. The authors conclude that the root is in the eukaryote branch and that earlier trees built with these sequences placing eubacteria as basal were as a result of long-branch attraction.

12. Lopez P, Forterre P, Philippe H: **The root of the tree of life in light of •• the covarion model.** *J Mol Evol* 1999, 49:496-508.
A description of a new method for applying the 'covarion' model to tree building that allows sites to alter their rate of evolution as secondary and tertiary structure evolves. Applied to the rooting of the tree of life, and with elongation factors, the authors conclude that the eubacteria are evolving at a higher rate than either archaea or eukaryotes, accounting for their basal position in earlier trees.

13. Penny D, Foulds LR, Hendy MD: **Testing the theory of evolution by comparing phylogenetic trees constructed from five different protein sequences.** *Nature* 1982 **297**:197-200.

14. Philippe H, Laurent J: **How good are deep phylogenetic trees?** *Curr •• Opin Genet Dev* 1998, 8:616-623.
A brief appraisal of the problems associated with building phylogenetic trees for deep divergences. Problems that are often ignored such as differences in evolutionary rate, sequence saturation, and fast-evolving lineages are addressed, and limitations of existing methods as well as possible solutions are described.

15. Teichmann SA, Mitchison G: **Is there a phylogenetic signal in •• prokaryote proteins?** *J Mol Evol* 1999, 49:98-107.
Using a data set of 32 proteins, it is shown that one gene which has undergone horizontal transfer can heavily influence the construction of a phylogenetic tree, even for a data set of 32 proteins. Upon removal of the offending gene, the remainder of the data set contained little information.

16. Forterre P, Philippe H: **Where is the root of the universal tree of •• life?** *Bioessays* 1999, **21**:871-879.
An overview of problems associated with rooting the tree of life, along with arguments favouring prokaryotes being derived from a mesophilic ancestor that was eukaryote-like in many respects. The paper also reviews details of how eukaryotes and prokaryotes have arisen from such an ancestor.

17. Lockhart PJ, Steel MA, Barbrook AC, Huson DH, Howe CJ: • **A covariotide model describes the evolution of oxygenic photosynthesis.** *Mol Biol Evol* 1998 **15**:1183-1188.
A mathematical test is introduced that can detect some cases where sites in a sequence are evolving under different constraints in different parts of the tree. It is well known in structural biology that two- and three-dimensional structrures of macromoles evolve over time (as predicted under W Fitch's covarion model); however, standard tree-building methods assume a site is always under the same contrstraints.

18. Benner SA, Ellington AD, Tauer A: **Modern metabolism as a palimpsest of the RNA world.** *Proc Natl Acad Sci USA* 1989, **86**:7054-7058.

19. Jeffares DC, Poole AM, Penny D: **Relics from the RNA world.** *J Mol •• Evol* 1998, **46**:18-36.
An updated model of the RNA-world, describing arguments, both old and new, for the placing of various RNAs in the RNA-world, what gaps there are in our present understanding of this period, plus a review of ancient genome architecture from the viewpoint of information theory. The paper also describes a novel way of viewing the evolutionary transition from RNA to protein catalysts and explains why some RNAs have persisted while others have not.

20. Poole AM, Jeffares DC, Penny D: **The path from the RNA world.** •• *J Mol Evol* 1998, **46**:1-17.
Here we attempted to establish what was known about the evolutionary transitions in going from an RNA-world to the emergence of the three domains of life. Included is a discussion of the origins of protein synthesis, the first proteins, messenger RNA, as well as aspects of the origins of DNA. Notably, we put forth a new hypothesis on the origin of introns, which we call 'introns-first'. Also discussed is the validity of using RNA-world relics for 'rooting' the tree of life. We conclude that the data we assemble are incompatible with a prokaryote-like Last Universal Common Ancestor.

21. Poole AM, Jeffares DC, Penny D: **Early evolution: prokaryotes, the •• new kids on the block.** *Bioessays* 1999, **21**:880-889.
We argue for re-evaluating the nature of the Last Universal Common Ancestor. Emphasises the importance of continuity of function in evolution, and suggests that our understanding of the RNA-world and the Last Universal Common Ancestor should be mututally compatible. It proposes a feedback process (the Darwin-Eigen cycle) where improved accuracy of replication permits a larger genome size, which permits coding for more features, which permit more accurate replication.

22. Reanney DC: **Genetic error and genome design.** *Trends Genet* 1986 **2**:41-46.

23. Reanney DC: **Genetic error and genome design.** *Cold Spring Harb Symp Quant Biol* 1987, **52**:751-757.

24. Forterre P: **Thermoreduction, a hypothesis for the origin of prokaryotes.** *CR Acad Sci Paris III* 1995, **318**:415-422.

25. Hiller R, Zhou ZH, Adams MW, Englander SW: **Stability and dynamics in a hyperthermophilic protein with melting temperature close to 200 degrees C.** *Proc Natl Acad Sci USA* 1997, **94**:11329-11332.

26. Marguet E, Forterre P: **DNA stability at temperatures typical for thermophiles.** *Nucleic Acids Res* 1994, **22**:1681-1686

27. Greenstein JP, Winitz M: **Glutamic acid and glutamine.** *In Chemistry of the Amino Acids.* New York: John Wiley and Sons; 1961:1929-1954.

28. Legrain C, Demarez M, Glansdorff N, Piérard A: **Ammonia-dependent synthesis and metabolic channelling of carbamoyl phosphate in the hyperthermophilic archaeon *Pyrococcus furiosis.*** *Microbiology* 1995, **141**:1093-1099.

29. Galtier N, Tourasse N, Gouy M: **A nonhyperthermophilic common •• ancestor to extant life forms.** Science 1999, **283**:220-221.
The authors compare the GC content of modern organisms in order to understand more on the nature of the LUCA and concludes that the ancestral GC content was too low for it to have been hyperthermophilic. A similar ancestral GC content was found using only the thermophilic organisms in the dataset.

30. Levy M, Miller SL: **The stability of the RNA bases: implications for • the origin of life.** *Proc Natl Acad Sci USA* 1998, **95**:7933-7938.
The half-lives of RNA bases at temperatures characteristic of hyperthermohiles is shown to be too rapid for bases to accumulate in a prebiotic world. The authors conclude that life must originate at low temperatures or that theories for the high temperature origin of life must exclude the four bases in RNA.

31. Brocks JJ, Logan GA, Buick R, Summons RE: **Archean molecular •• fossils and the early rise of eukaryotes.** *Science* 1999 285:1033-1036.
Identification of molecular biomarkers in 2700-million-year-old Archaean shales in Australia argues for the presence of photosynthetic organisms hundreds of millions of years before the atmosphere became oxidising. Perhaps more strikingly, the research also points to the presence of eukaryotes at this time, pushing back the earliest identification of these organisms by 600 million years.

32. Mushegian AR, Koonin EV: **A minimal gene set for cellular life derived by comparison of complete bacterial genomes.** *Proc Natl Acad Sci USA* 1996, **93**:10268-10273.

33. Becerra A, Islas S, Leguina JI, Silva E, Lazcano A: **Polyphyletic gene losses can bias backtrack characterizations of the cenancestor.** *J Mol Evol* 1997, **45**:115-117 [Mushegian AR, Koonin EV: **Response.** *J Mol Evol* 1997, **45**:117-118.]

34. Andersson SGE, Zomorodipour A, Andersson JO, Sicheritz-Pontén T, •• Alsmark UCM, Podowski RM, Näslund AK, Eriksson A-S, Winkler HH, Kurland CG: **The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria.** *Nature* 1998, **396**:133-140.
The first complete genome for an α proteobacterium and thereby among the closest ancestors to mitochondria. The work demonstrates the effects of Müller's ratchet (accumulation of deleterious alleles in the absence of recombination) on the evolution of intracellular obligate microbes.

35. Martin W, Stoebe B, Goremykin V, Hansmann S, Hasegawa M, •• Kowallik KV: **Gene transfer to the nucleus and the evolution of chloroplasts.** *Nature* 1998, **393**:162-165
Using whole genomes, mostly of chloroplasts, the paper demonstrates patterns of gene loss and of gene transfer to the nucleus. They find independent gene losses in multiple lineages and identify a large set (44) of chloroplast genes which had transferred from chloroplast to nucleus. It is concluded that gene loss and transfer in organelles is best explained in terms of Müller's ratchet.

36. Koonin EV, Mushegian AR, Bork P: **Non-orthologous gene replacement.** *Trends Genet* 1996, **12**:334-336.

37. Doolittle WF: **You are what you eat: a gene transfer ratchet could • account for bacterial genes in eukaryotic nuclear genomes.** *Trends Genet* 1998 **14**:307-311.
The author proposes a mechanism for gene transfer from bacteria to the eukaryote nucleus. The model attempts to account for the apparent chimeric makeup of the nuclear genome and is a plausible mechanism by which eukaryotes could acquire genetic information from the diet.

38. Martin W: **Mosaic bacterial chromosomes: a challenge en route to •• a tree of genomes.** *Bioessays* 1999, **21**:99-104.
A broad review of current knowledge on horizontal transfer. The author describes some interesting consequences of horizontal transfer with respect to phylogenetic analyses. Additionally, the likely differences between prokaryote–prokaryote transfer and prokaryote–eukaryote transfer are discussed.

39. Lynch M: **Mutation accumulation in nuclear, organelle, and prokaryotic transfer RNA genes.** *Mol Biol Evol* 1997, **14**:914-925.

40. Reanney DC: **On the origin of prokaryotes.** *J Theor Biol* 1974, 48:243-251.

41. Pennisi E: **Genome data shake the tree of life.** *Science* 1998, 280:672-673.

42. Pennisi E: **Is it time to uproot the tree of life?** *Science* 1999, 284:1305-1307.

Paper 6

The origin of the nuclear envelope and the origin of the eukaryote cell.
Manuscript.

The origin of the nuclear envelope and the origin of the eukaryote cell.

Summary.

Establishing the origin of the nucleus is central to understanding the evolution of the eukaryotic cell. One feature of virtually all discussions of nuclear origins to date is the lack of discussion of the nuclear envelope. Here I attempt to ask how such a unique membrane structure could have arisen in evolution, when this occurred, what the selection pressure might have been, and why it is not found in prokaryotes.

Introduction.

Ever since the first descriptions of prokaryote and eukaryote cell structure, researchers have sought an explanation for the origins of the nucleus. While progress in understanding the molecular details of nuclear function is moving at a fast pace (Olson, et al. 2000; Wente, 2000; Lewis and Tollervey, 2000), progress on nuclear origins is slow. Scenarios for the evolution of the nucleus include an endosymbiotic origin (e.g. Lake and Rivera, 1994), autogenous origins in eukaryotes (e.g. Cavalier-Smith, 1988), and emergence subsequent to a fusion between an eubacterium and an archaeon (e.g. Gupta and Golding, 1996; Martin and Müller, 1998; Moreira and López-García, 1998; Margulis, et al., 2000).

With the explosion of new comparative data and, in particular, the finding that a number of nuclear features (e.g. histones & small nucleolar RNAs) are also present in representatives of the archaea, the relationship between the three domains is again becoming clouded. Forterre (1997) has made the important point that, given the lack of consensus on the relationships between the three domains, it is problematic to assume the direction of evolution on the basis of shared archaeal-eukaryote characters. These could be readily explained either as dating back to the Last Universal Common Ancestor (LUCA) (having been lost from bacteria), or by emergence in the common ancestor of archaea and eukaryotes, after their divergence from bacteria.

Indeed, the rapidly growing data from whole genomes has established that eukaryote genomes contain genes whose closest counterparts are in the eubacteria, and genes whose closest counterparts are archaeal (Ribeiro and Golding, 1998; Rivera et al., 1998; Horiike et al., 2001). Horiike et al. (2001) provide the most intuitive description of the pattern: eukaryotic genes which appear most closely related to bacterial genes function in the cytoplasm, while those apparently related to archaeal sequences generally function in the nucleus. However, as per Forterre's caveat that single traits can fit more than one scenario when the relationships between the three domains is not known (Forterre, 1997), the same applies for the genome data. This pattern could be interpreted in several ways with respect to the relationships between archaea, eubacteria and eukaryotes, with interpretation being further complicated by differing accounts of the degree of horizontal transfer between the three domains (Martin, 1999a; Penny and Poole, 1999, for review).

Inherent in discussion of the problem of nuclear origins, is the assumption that (chloroplasts, hydrogenosomes and mitochondria aside) the greater complexity of the eukaryote cell has evolved from a simpler prokaryotic cell ultrastructure. This has seemed reasonable, and indeed is implicit in almost all discussions of the evolution of prokaryotes and eukaryotes (see Forterre and Philippe, 1999, for a critique), so the debate has largely centred on how the nucleus arose in eukaryotes after their divergence from prokaryotes. While it seems intuitive that complex eukaryotic organisms, and with them, the nucleus, must have evolved from simpler prokaryotic organisms, evolution does not necessarily result in complexification with time. Reductive evolution is generally accepted in endosymbiosis and parasitology (Fraser et al., 1995; Razin et al., 1998; Andersson and Andersson, 1999a), but the idea that the prokaryote lineages have as a whole undergone a process of reductive evolution (Reanney, 1974; Forterre, 1995a; Poole et al., 1998) has been less popular. However, a number of groups are finding that a broad range of biochemical, biophysical, phylogenetic and genetic data are more compatible with this scenario than with the traditionally-accepted prokaryote to eukaryote transition (Forterre and Philippe, 1999; Poole et al., 1999; Penny and Poole, 1999; Glansdorff, 2000).

Indeed, a previously unconsidered dataset is the concentration of putative RNA world relics in the eukaryote nucleus (Poole et al., 1998, 1999). The identification of snoRNAs in archaea (Omer et al., 2000; Gaspin et al., 2000) means there are more such RNA world traits in the archaea than in eubacteria. Both archaea and bacteria nevertheless appear to have undergone reductive evolution, losing a number of these traits (Poole et al., 1999; Penny and Poole, 1999). An archaeal origin of the nucleus, or the host of the mitochondrion/hydrogenosome (Gupta and Golding, 1996; Martin and Müller, 1998; Moreira and López-García, 1998; López-García and Moreira, 1999), does not explain the likely RNA world origin of a number of traits (though see Sogin et al., 1996, for a more inclusive model) (Table 1). Selection pressures for the loss of putative RNA relics in the bacteria and archaea (either once or twice) have been described (Forterre, 1995a; Poole et al., 1999). Since most fusion scenarios cannot explain the observation that the greatest diversity of RNA relics are in eukaryotes, they are problematic (Sogin et al., 1996; Penny & Poole, 1999).

In this paper, I concentrate on what I consider to be the three most problematic issues surrounding the origins of the nucleus:

1. The selection pressure that drove the evolution of the nucleus.
2. The nature of the organism in which this developed.
3. Whether or not the nucleus arose prior to organellar endosymbioses.

Current theories fall far short of explaining the entire range of genetic, structural and biochemical data, and in my mind, this is symptomatic of many discussions of early evolution. The prokaryote dogma is a large part of the problem (see Forterre, 1995b; Forterre & Philippe 1999, for detailed discussion). This is simply that eukaryotes evolved from prokaryote-like ancestors, and is underpinned by the identification of 3.5 billion-year old microfossils classified as cyanobacteria (Schopf & Packer 1987).

The existence of stromatolites as far back as this has likewise been taken to suggest the existence of cyanobacteria (Walsh, 1992), since modern stromatolites are formed by cyanobacteria. However, the earliest stromatolites lack microfossils and can abiotic processes can explain their formation (Lowe, 1994; Grotzinger & Rothman, 1996). Furthermore, in recent phylogenetic reconstructions, modern cyanobacteria appear as a derived group (Lockhart et al., 1998).

There are also difficulties in establishing taxonomy from ultrastructure. The best known example is the identification of the domain archaea which required sequence comparisons (Woese and Fox, 1977), but there have also been difficulties in distinguishing some protists and bacteria on morphology alone. *Epulopiscium fishelsoni*, a symbiont living in the gut of surgeonfishes, was originally thought to be a protist (Fishelson et al., 1985; Montgomery & Pollak, 1988). Electron microscopy suggested that, despite massive cell size, these symbionts might in fact be prokaryotes (Clements and Bullivant, 1991), but it was only with phylogenetic analysis that this could be confirmed (Angert et al., 1993).

The problem of the nucleus is, without exception, framed within the assumption of prokaryote ancestry. Hence, when it is asked, 'What is the origin of the nucleus?' the question really is, 'Given we know that eukaryotes are derived, how did the nucleus arise specifically in this lineage after their split from bacteria and archaea?' It is worth pointing out that many theories do not even get this far, providing nothing more than a description of which bits could have evolved after which other bits to give the modern nucleus! All biochemical data that show any relationship between archaea and eukaryotes tend to considered in this light, archaeal histones (Pereira & Reeve, 1998) and snoRNA homologues (Omer et al. 2000), being two such examples.

The prokaryote dogma may be correct, or it may be incorrect (its validity has been challenged, but is hardly debated) but the problem lies with the application of the assumption in general. By making this assumption, the question is answered before the data are even looked at. Thus only one scenario can ever be considered and, while details may differ, there is only one possible conclusion!

There are strong grounds on which to challenge the prokaryote dogma, and that, without questioning its validity as a central tenet of early evolution, it is impossible to make progress in understanding cellular evolution. Thus, in this paper I aim to reexamine the question of the origin of the nucleus without first requiring that, as a corollary of the prokaryote dogma, the nucleus must have arisen in the eukaryote lineage and is a derived trait. Indeed, because all discussions on the origin of the nucleus that I am aware of assume that it arose specifically in the eukaryote lineage, I will take the other end of the spectrum: that the nuclear envelope predates the LUCA and arose concurrent with the first cells. It may eventually be possible to reject this extreme position, but in the meantime it is interesting to see where the argument leads.

The proposal is based in part on a previous finding that the largest collection of putative RNA world relics is found in the eukaryote nucleus, suggesting it is perhaps a more ancient structure than previously supposed (Poole et al., 1998, 1999; Penny and Poole, 1999). I further suggest that a double membrane structure (i.e. nuclear and cytoplasmic) could have served as a buffer to osmotic pressure in the cell prior to the advent of sophisticated gates, channels and pumps for osmoregulation.

Moreover, I argue that the unique structure of the nuclear envelope may hold the key to how the presumed transition from surface chemistry to the first cells occurred, as simple pores of either protein or RNA are possible without the requirement that these traverse the lipid bilayer. I subsequently argue that, in the absence of selection pressure to remove the nuclear membrane, this was never lost from eukaryotes, while in prokaryotes it was selectively advantageous to have coupled transcription and translation.

The ideas discussed are speculative, and may well be incorrect. However, if the paper serves to drive debate on early evolutionary events away from discussion of narrow datasets and preconceptions, it will have served an important purpose. Work to date tends to focus on phylogenetic patterns and gene distributions, or the study of candidate 'living fossil' organisms or groups. There is a paucity of discussion on the structure of the nuclear envelope as compared with other cellular membrane structures and almost no discussion on selection for its origin and evolution exclusively in eukaryotes (Martin 1999b; Poole and Penny, 2001). Ignoring the unique structure of the nuclear envelope in proposing a theory is unforgivable, yet a number of authors do this when proposing that the nucleus was an endosymbiont (Rivera and Lake 1994; Gupta and Golding, 1996; Horiike et al., 2001). These scenarios explain only a single dataset: that the eukaryote genome appears to be chimeric, individual genes being either most closely related to bacterial genes or archaeal ones. In itself, this is a salient observation based initially on confusing gene relationships (Gupta and Singh, 1994) and later, from larger genomic analyses (e.g. Rivera et al., 1998; Ribeiro & Golding, 1998; Horiike et al., 2001). However, in concluding from these data that the nucleus was an endosymbiont, crucial differences between nuclear structure and function and organelles of clear endosymbiont origin (mitochondria, hydrogenosomes, chloroplasts) are either overlooked or ignored.

Likewise, the gap between prebiotic chemists, who are largely in favour of surface chemistry as a crucial step the origin of life, and molecular biologists, who expect that the earliest life forms were cellular, is large. A cell with an almost impermeable lipid membrane is not a likely intermediate between these two presumed stages. How the first cells might have regulated their intracellular environment relative to the external environment is largely unexplored.

The problem with purely descriptive explanations for the origin of the nucleus.

Most current theories on the origin of the nucleus attempt to address the growing evidence (Gupta and Golding 1996; Ribeiro and Golding, 1998; Rivera et al., 1998; Horiike et al. 2001) that eukaryote genomes represent a mixture of archaeal- and eubacterial-like genes. However, all lack a crucial component: no clear selection pressures are given for the origin of this structure. This is in stark contrast to research into the origins of mitochondria, hydrogenosomes and chloroplasts, where there is general agreement, and good experimental support (Andersson et al., 1998; Martin et al., 1998; Gray et al., 1999; McFadden, 1999) to show that these organelles arose by endosymbiosis from free living bacteria. Current theories in that field attempt to identify a selection pressure for the initial symbiosis, as well as the process of gene loss from organelles and gene transfer to the nucleus (e.g. Martin and Müller, 1998; Martin et al., 1998; Moreira and López-García, 1998; Andersson and Kurland, 1999; Andersson and Andersson, 1999a,b; Blanchard and Lynch, 2000).

What connects the question of the origin of endosymbiotic organelles and the origin of the nucleus has been the difficulty in establishing whether, prior to the origin of the mitochondrion, amitochondriate eukaryotes existed at all (Sogin, 1997; Embley & Hirt, 1998; Martin and Müller, 1998; Philippe et al., 2000a). As pointed out by Martin, "for organelles to take up residence in a cytoplasm, there had to be a host" (Martin, 1999b). If all the DNA-containing organelles of the cell arose through endosymbiosis, and the nucleus was the first to arise this way, what happened to the genetic material of the original host? Vellai et al. (1998) have noted that for an endosymbiotic event to occur, there must have been some mechanism of phagocytosing the endosymbiont, and, so far, only eukaryota have been demonstrated to be capable of this. Indeed, if it is contended that a prokaryotic organism was the original host, it seems odd that phagocytosis is no longer a feature of extant prokaryote lineages!

Two broad variant theories for the origin of the nucleus through endosymbiosis have been suggested, those where the endosymbiont is an archaeon, and those where the host is an archaeon. The first, that the nucleus was an endosymbiont archaeon that took over the host cell (Rivera & Lake, 1994; Horiike et al. 2001), not only fails to explain how an archeal cell membrane could have become the nuclear envelope, it also requires that the endosymbiont gained genes from the host (Poole and Penny, 2001). In addition, it requires that the endosymbiont changed its lipid composition, from ether-linked lipids to the phospholipids found in the nuclear envelope. It also requires a change in structure from a simple lipid bilayer to a structure where inner and outer nuclear membranes are continuous. The outer membrane is also continuous with the endoplasmic reticulum, forming a continuous lumen. Furthermore the nuclear pores do not traverse the lipid bilayer as such, but are instead formed at regions where the inner and outer nuclear membranes meet.

What was the selection pressure that drove this event? How can the lack of similarity of nuclear membrane structure (an envelope with pores) and nuclear chromosomes (with those of prokaryotes, chloroplasts and mitochondria) be accounted for? How is it possible to account for the disappearance, and later reformation, of the nuclear envelope at cell division (meiosis and mitosis) in some eukaryotic groups? Other organelles of endosymbiotic origin, regardless of placement on the eukaryote tree, do not undergo this process. While 'closed' mitosis, where the nuclear envelope remains intact throughout, is known in various protists, algae and fungi, it is not clear whether this is ancestral or derived.

Furthermore, the structure of the nuclear envelope bears no resemblance to *any* biological membranes in archaea and bacteria. The nuclear envelope is unlike the membrane structure of any prokaryote, consisting of a flattened continuous lipid bilayer with nuclear pores allowing free diffusion of molecules 20-40kDa in size across the envelope (Wente, 2000; Allen et al., 2000). Engulfment to form chloroplasts and mitochondria has not produced such structures, and both the membrane and most porins are of Gram-negative bacterial 'origin' (Cavalier Smith 2000; Flügge 2000; Soll et al. 2000). Nor has a structure equivalent to the nuclear envelope appeared in cases of secondary endosymbioses where one eukaryotic cell engulfs another. An exception is the nucleomorph, which is clearly a relic of the nucleus of the eukaryotic endosymbiont (Gilson et al., 1997; Cavalier-Smith, 2000; Douglas et al. 2001).

If the nucleus is archaeal in origin (Lake, 1994; Gupta and Golding, 1996; Moreira and López-García, 1998; Horiike et al. 2001), then these issues are unexplained. The worst oversight here is that it requires that the endosymbiont *gained* genes from the host, with the latter presumably losing *all* its genes, including a significant proportion to the endosymbiont (Poole & Penny, 2001). This is inconsistent with all documented cases of endosymbiosis and intracellular parasitism by prokaryotes, and eukaryotes (Andersson et al. 1998; Moran & Baumann 2000; Wren 2001; Keeling & McFadden 1998; Douglas et al., 2001). Upon entering a symbiotic or parasitic relationship with the host, the endosymbiont, by utilising host metabolites, over time loses the capacity to synthesise these metabolites. This pattern has been clearly established through numerous whole genome studies (Fraser et al. 1995, 1997, 1998; Himmelreich et al. 1996; Andersson & Andersson 1999a,b; Kalman et al. 1999; Cole et al. 2001). In time, this irreversible process presumably results in host dependence, with the endosymbiont becoming obligate.

Mitochondria, chloroplasts and hydrogenosomes, which are of endosymbiotic origin, have suffered this fate (Blanchard & Lynch 2000; Martin et al. 1998), with hydrogenosomes having completely lost their genome in all but a few cases (Akhmanova et al. 1998). The intracellular lifestyle places endosymbionts and parasites under mutational pressure, particularly in an obligate intracellular lifestyle. This is best explained as being due to Muller's ratchet, the gradual accumulation of slightly deleterious mutations in asexual organisms with small population size. Muller's ratchet has been shown to affect free-living bacteria (Andersson & Hughes 1996), endosymbionts such as *Buchnera* (Moran 1996; Moran & Baumann 2000), intracellular parasites such as the *Rickettsiae* and *Chlamydiae* (Andersson & Andersson 1999b; Kalman et al. 1999) as well as organellar genomes (Berg & Kurland 2000; Blanchard & Lynch 2000).

The hydrogen hypothesis (Martin and Müller, 1998) does not fall foul of the above criticisms as it instead suggests the host was an archaeon, the endosymbiont was the forerunner to mitochondria and hydrogenosomes, and that the nucleus evolved subsequent to endosymbiosis (Martin and Müller, 1998; Martin, 1999b). Other chimeric theories where the host is an archaeon exist, but most are largely descriptive and do not attempt to establish selection pressures for the origin of the nucleus. For reviews of the various chimeric hypotheses, see Gupta and Golding (1996), Katz (1998), López-García and Moreira (1999) and Margulis et al. (2000).

The hydrogen hypothesis (Martin and Müller, 1998) and the related but independently conceived syntrophy hypothesis (Moreira and López-García, 1998) are perhaps the most interesting. Both provide a detailed and feasible scenario for the origin of the eukaryote cell, ultimately by fusion between an archaeon and a bacterium (two bacteria in the case of syntrophy). They do not fall foul of any of the criticisms levelled at competing theories. Similarities and differences between the hydrogen and syntrophy hypotheses have been discussed elsewhere (López-García and Moreira, 1999) and I do not cover these in depth here. Suffice it to say both represent plausible scenarios for the metabolic basis for the establishment of symbiosis, as opposed to simply suggesting that the symbiont gave away ATP. However, the question I consider is the nature of the host, as opposed to the nature of the initial interaction. For simplicity, I shall consider the simpler of the two scenarios, where there is only a single symbiont (the hydrogen hypothesis).

In the hydrogen hypothesis endosymbiosis occurs, though not as an initial step. In this scenario, the endosymbiont is the ancestor of both hydrogenosomes and mitochondria. However, it does not explicitly describe the origins of the nucleus, other than to say that the possession of numerous traits common to both archaea and eukaryotes makes it feasible to suggest the nucleus arose after the endosymbiosis that spawned hydrogenosomes and mitochondria. In a separate paper, Martin (1999b) does however discuss nuclear origins under the hydrogen hypothesis. To this I shall return.

A third class of theory for the origin of eukaryotes avoids the problem by invoking a proto-eukaryotic host (possibly with a nucleus), thereby explaining the chimeric origin of nuclear genes. This is the traditional formulation of the endosymbiont hypothesis (as revived by Margulis, 1970), and which has been most extensively developed by Cavalier-Smith. He proposed that extant amitochondriate protists, which he named the Archaezoa (Cavalier-Smith, 1983, 1987, 1988), were the ancestors of mitochondriate eukaryotes, predating endosymbiosis in the eukaryote lineage. While the member composition of the Archaezoa has been variable (see Table 2 in Patterson, 1999), it is now widely thought that the Archaezoa may all be secondarily amitochondriate (reviewed by Keeling, 1998; Embley and Hirt, 1998). However, that these extant protists are not the 'missing link' in the evolution of the eukaryote cell does not necessarily mean that the host could not have been a proto-eukaryote.

The wealth of data on mitochondria and hydrogenosomes suggests endosymbiosis of an ancient facultative α-proteobacterium best accounts for a single origin for these organelles (Rotte et al. 2000). While the details are strongly debated (Andersson & Kurland 1999; Rotte et al. 2000), the issue of the origins of the nucleus tend to take a back seat (though see Martin, 1999b). Indeed, as has been debated recently (Biagini & Bernard 1999; Martin 1999c) there is much difficulty in establishing the nature of the host. Was it an archaeon, with the nucleus arising only after the initial endosymbiosis, or an amitochondrial proto-eukaryote with a nucleus?

Did the nucleus arise after mitochondria/hydrogenosomes?

Aside from genomic data suggesting that the eukaryote nucleus has a chimeric gene composition, little has been said about the origins of the nucleus. There has been no systematic attempt to establish whether the host was 'eukaryotic' with a nucleus, or whether it was an archaeon (with the nucleus being a late development). Nor has there been much attempt to suggest plausible selection pressures for its origin under either one of these scenarios.

In the current context, the debate between proponents of the 'ox-tox' hypothesis (that the original interaction between proto-eukaryote host and ancestors of mitochondria was based on oxygen detoxification of the host by the symbiont [Andersson and Kurland, 1999]) and the hydrogen hypothesis is interesting. While the details differ, the common feature is that all agree on a common origin for mitochondria and hydrogenosomes (Andersson and Kurland, 1999; Rotte et al., 2000). However, neither theory addresses the origin of the nuclear envelope. The 'ox-tox' hypothesis envisages a proto-eukaryotic host that may or may not possess a nucleus (Andersson and Kurland, 1999), while the hydrogen hypothesis argues for an archaeal host, so requires the nucleus to have arisen after the endosymbiosis that gave rise to mitochondria and hydrogenosomes (Martin, 1999b).

The greater potential for oxidative damage in the mitochondrion (and chloroplast) is probably one pressure for many (though not all) genes to be relocated to the nucleus (Allen and Raven, 1996; Race et al., 1999). 'Ox-tox' is potentially compatible with this, requiring that strictly anaerobic eukaryotes arose from aerobic ancestors. A theory put forth by Vellai et al. (1998) is intermediate in that it proposes an archaeal host, but an aerobic basis for endosymbiosis.

One argument for the origin of the nucleus is that it served to protect host DNA from oxidative damage resulting from leakage of reactive oxygen species from the mitochondrion (see Li, 1999). This theory is interesting, being based on observed differences in oxidative damage in the nucleus and mitochondria (Richter et al., 1988; Ljungman and Hanawalt, 1992). However, the nuclear envelope allows free diffusion of small molecules up to ~40kDa, so is unlikely to represent a barrier to oxygen radicals. Furthermore, reactive oxygen species are dealt with by superoxide dismutases, catalases and glutathione peroxidases, not compartmentation (McCord, 2000). Nor does oxidative damage explain the absence of a nucleus-like structure in aerobic prokaryotes, or suggest how a nuclear envelope might protect an anaerobe against oxygen, and reactive oxygen species. Finally, if the ancestral endosymbiosis was based on an anaerobic symbiosis (Martin and Müller, 1998; Moreira and López-García, 1998) this theory cannot readily account for the origin of the nucleus in anaerobic eukaryotes, though in all current theories, the endosymbiont is considered to be facultatively aerobic.

More importantly, the 'Ox-tox' hypothesis, while less developed than the hydrogen hypothesis, permits the origin of the nucleus to be either prior to endosymbiosis, or to post-date it. The hydrogen hypothesis, in arguing for an archaeal host, requires that the nucleus (and a number of other eukaryote-specific traits) arose after hydrogenosomes and mitochondria. It does not suggest what selection pressures might account for the origin of the nuclear envelope, endomembrane system, and other features that separate archaea from eukaryotes.

Martin (1999b) has argued for a fortuitous emergence of the eukaryote endomembrane system using the symbiosis described by the hydrogen hypothesis as a starting point. I will argue that this hypothesis, in requiring an archaeal host, does not explain many aspects of modern eukaryote cells. However this does not mean that I think the hydrogen hypothesis should be rejected outright. In terms of establishing a biochemical basis for the symbiotic interaction that gave rise to mitochondria and hydrogenosomes, it is not only plausible, but provides in many respects a substantial improvement over previous theories. At issue here is the nature of the host, not the nature of the symbiosis that gave rise to mitochondria and hydrogenosomes.

One argument that has been made in favour of the possibility that the host was an archaeon is that the amitochondriate group of eukaryotes, the Archaezoa, are probably all secondarily amitochondriate, suggesting all extant eukaryote lineages once harboured mitochondria (Keeling, 1998; Embley and Hirt, 1998). This has led to the suggestion that the origin of mitochondria & hydrogenosomes is concurrent with the origin of the eukaryote cell (Martin & Müller, 1998; Martin, 1999b). This argument is as problematic as the former assumption that the ancestral state for eukaryotes was nucleate but amitochondrial (Cavalier-Smith, 1983, 1987), yet is presently being strongly argued for because of the absence of any evidence for the Archaezoa being genuinely amitochondriate as opposed to secondarily so (e.g. Martin, 1999b).

In the same way as there may be no *bona fide* Archaezoa, there are no anucleate eukaryotes/archaea which harbour mitochondria/hyderogenosomes or endosymbionts. The hydrogen hypothesis (Martin & Müller, 1998) points to modern-day examples of symbioses between archaea and bacteria much like those argued in that hypothesis to provide the basis for the interaction that ultimately led to the α-proteobacterial symbiont becoming an intracellular organelle (mitochondria/hydrogenosomes). But what selection pressures might have led to all these subsequent eukaryote-specific traits?

No intermediates between the modern examples of archaeal-bacterial symbioses and modern eukaryotes with hydrogenosomes/mitochondria have been identified. Obvious examples would be phagocytic archaea, archaea with linear chromosomes maintained by telomerase with multiple origins of replication, or 'eukaryotes' or archaea with intracellularly located hydrogenosomes/mitochondria but no nucleus. Both Archaezoan and hydrogen hypotheses demand that ancestral forms went extinct, presumably through competition. Hence, arguing that the absence of one presumed ancestral form supports the alternative hypothesis is not only incorrect it is moot!

In an important sense, arguing for an eukaryotic nuclear host is easier than arguing for an archaeal host. One has to accept that no modern examples exist, but it permits the host to be endophagocytic, and does not require that a range of eukaryote-specific features (linear chromosomes with telomeres and multiple origins of replication, the nuclear envelope and nuclear pore complex, endoplasmic reticulum, golgi) all evolved subsequent to endosymbiosis.

If current speculations of a eukaryote 'big bang' (Philippe et al., 2000a,b) are supported, this could be argued to account for the extinction of earlier forms, and could in principle fit with either a proto-eukaryote or archaeal host. On current evidence of formerly deep diverging amitochondrial eukaryotes being derived (Keeling, 1998), it could either be argued that the endosymbiosis event resulted in extinction of all proto-eukaryotic lineages, or that the advent of the nucleus and other eukaryote-specific features, subsequent to endosymbiosis, resulted in the extinction of intermediate forms. The general agreement that hydrogenosomes and mitochondria are of a common endosymbiotic origin, as well as organelle to nucleus transfer of genes not strictly required for the function and maintenance of the endosymbiont (e.g. glycolysis) tentatively suggests the former.

Although the 'big bang' hypothesis is far from accepted, it does lend some credibility to the fact that both nucleus-first and endosymbiont-first theories require extinction of intermediate forms, and this point bears further inspection. Since the endosymbiont-first theory has been detailed elsewhere (Martin & Müller, 1998; Martin, 1999b), I will limit discussion to two issues. The first is whether there is an evolutionary precedent for the extinction of intermediate forms, and the second is whether assuming the derivation of eukaryotic nuclear traits from archaeal traits is reasonable.

Extinction of intermediate forms?

Currently, there are no known intermediate forms between archaea and modern eukaryotes that might favour the idea that eukaryote features arose subsequent to the endosymbiosis event. Nor are there any *bona fide* Archaezoa to support the idea that the host was nucleate. Two possibilities are immediately obvious: 1. That the limited sampling of eukaryote and archaeal diversity is such that intermediate forms have simply not been found (Embley & Hirt, 1998; Keeling, 1998)

2. That intermediate forms have been outcompeted by the ancestors of the extant lineage, which would potentially account for the eukaryote 'big-bang' suggested from phylogeny (Phillipe et al. 2000a,b). Until intermediate forms are identified, neither theory fares better than the other and the debate cannot be readily resolved on this point alone.

Nevertheless, one can speculate on the feasibility of an across the board extinction of intermediate forms. If all Archaezoa turn out to be secondarily amitochondriate, a revised Archaezoan hypothesis would require that ancestrally nucleate forms were outcompeted across all environments by nucleate eukaryotes carrying a facultatively aerobic endosymbiont. The hydrogen hypothesis is slightly trickier. The initial formulation (Martin & Müller, 1998) does not address the origin of eukaryote-specific traits in detail, and does not involve a symbiont in an intracellular location. Instead, a symbiosis event similar to modern symbioses between archaea and bacteria is argued as the initial state. Nevertheless, given the intracellular location of mitochondria and hydrogenosomes in extant eukaryotes, endosymbiosis must have ultimately ensued. In a subsequent paper by Martin (1999b), the origin of the endomembrane system is argued to be a consequence of the relocation of symbiont genes for lipid synthesis to the host chromosomes. The wording is ambiguous with respect to whether the symbiont was intracellular by this time. However, the statement that, 'Gene transfer from the symbiont's genome to the cytosolic chromosomes of the host could have genetically cemented two prokaryotes into a single, biochemically compartmented, but nucleus-lacking common ancestor of eukaryotes' suggests this. Gene decay and symbiont to host gene transfer are features of endosymbionts and obligate intracellular parasites (Andersson & Andersson, 1999a,b; Moran & Baumann, 2000). Given such a location for hydrogenosomes and mitochondria, it makes most sense that, by this time, the symbiont in Martin's scenario is intracellularly located.

Considering the hydrogen hypothesis first, if biological competition was responsible for extinction of intermediate forms, extinction must occur as a consequence of the evolution of eukaryote-specific features subsequent to endosymbiosis. These eukaryote-specific features would need to be selectively advantageous in all ancestral eukaryote environments, displacing existing forms, as well as being maintained in the subsequent colonisation of aerobic environments. The theory must explain the ubiquity of eukaryote features such as linear chromosomes with telomeres and multiple origins of replication, an endomembrane system consisting of nuclear envelope, endoplasmic reticulum and golgi, and a cytoskeleton (Table 1). Either the final feature to appear was so superior as to outcompete ancestral forms (with the other features being fixed), or these cell structural features in combination were. There is difficulty even coming up with a selection pressure for the emergence of such features, let alone establishing how such features could come to define the eukaryote cell architecture. What is it about the endomembrane system that makes it so superior to an anuclear host with an endosymbiont?

The other possibility is a modified version of the traditional argument, that the endosymbiont took up residence in a proto-eukaryotic host which, other than the lack of hydrogenosomes or mitochondria, was structurally similar to modern eukaryotic cells. The host would have already been separated from the archaeal lineage, and was phagocytic. The ancestral cell would have endophagocytosed the ancestral α-proteobacterium, and the nature of the interaction could still have been initially anaerobic, as per the hydrogen hypothesis, with the endosymbiont being facultatively anerobic and the proto-eukaryotic host being an anaerobe. I will discuss the points in favour of a proto-eukaryotic host in the next section.

In the absence of intermediate forms, this scenario, as with the hydrogen hypothesis, would require extinction of intermediate forms, though in this case, there is only one form, proto-eukaryote lineages without an endosymbiont. Thus, the presence of an anaerobic endosymbiont in one lineage would have to be argued to be sufficient to outcompete all other proto-eukaryote lineages in all environments, and account for the colonisation of aerobic environments by its descendents.

In order to consider this in depth, I shall introduce the concept of Evolutionarily-Stable Niche-Discontinuity (ESND) (Poole et al., 2001; M.J. Phillips, in prep.). Put simply, the ESND concept describes limits on potential evolvability as a result of within species competition between individuals, and the existence of a valley of low fitness between two niches. An individual that displays a trait which shifts it away from its (original) niche toward a second, occupied niche will be selected against within its own niche. It will still be too far away from the second niche to be able to compete successfully within the latter. The dual requirement of gradual changes across multiple traits, coupled with specialisation within a niche thus results in a discontinuity, and inhabitants of one niche cannot reach another (occupied) niche. An example given in Poole et al. (2001) is that of cats, which are fast-burst strike predators, and dogs, which are indurance predators.

ESNDs are predicted to exist between eukaryotes and prokaryotes, the latter in general being r-selected relative to the former (Poole et al., 2001). Two important aspects of prokaryotes and eukaryotes (when viewed not as phylogenetic groups, but evolutionary strategies) are evident. First, prokaryotes are able to respond quickly to the presence of a new nutrient by virtue of transcription and translation being coupled. Thus, before the transcript has been completely synthesised, translation of the protein it encodes has begun. In eukaryotes, the transcript is synthesised, capped, polyadenylated, spliced and then exported to the nucleus before it is synthesised. Secondly, prokaryote genome size is at a premium. There are limits to the rate at which a circular chromosome with a single origin of replication can be copied, and is be the rate-limiting step during exponential growth in *E. coli* (Poole et al., 2001, and references therein). With such strong selection on genome size in prokaryotes, only a single origin of replication per chromosome[1], and selection for fast response times, it is likely that there is an ESND between r-selected eukaryotes and prokaryotes. The former group possesses multiple origins per chromosome, and response time is limited by physical separation of transcription and translation provided by the nucleus. The number of changes required for eukaryotes to become established in niches currently inhabited by prokaryotes (or vice versa) are too great, given intermediate low fitness. ESNDs may break down when organisms that inhabit similar niches and have never been in contact (e.g. because of geographical isolation) are brought into contact, or in organisms where horizontal gene transfer is possible (Poole et al., 2001).

With regard to the possibility of replacement of ancestral nucleate eukaryotes by the lineage that possessed an endosymbiont, the ESND concept provides a useful way of looking at how across the board displacements could have occurred. First of all, one of the consequences of selection for fast response times and subsequent exponential growth in prokaryotes is that enzymes will tend to evolve towards catalytic perfection at a faster rate than in eukaryotes (Jeffares et al., 1998; Poole et al., 1999, 2001). Catalytic perfection is achieved when the rate-limiting step in a reaction is the diffusion of substrate to the active site (Albery and Knowles, 1976).

This may account for the observation that more than just endosymbiont-specific genes have been transferred to the eukaryote nucleus (Berg & Kurland, 2000; Blanchard & Lynch, 2000). Notably, genes for glycolysis have been argued to be of endosymbiotic origin (e.g. Martin et al., 1993; Keeling & Doolittle, 1997; Henze et al., 1998; Liaud et al., 2000), perhaps consistent with the possibility that ancestral prokaryotic metabolic genes were superior in terms of catalytic efficiency to those of the host. This might likewise account for the chimeric genome of eukaryotes, where most genes of probable bacterial origin are 'cytoplasmic' (i.e. involved in metabolism in the eukaryote cytoplasm, *sensu* Horiike et al., 2001, see also Rivera et al., 1998). Selection for relocation of beneficial endosymbiont genes to the host can be argued on the basis of Muller's Ratchet (Blanchard and Lynch, 2000; Berg and Kurland, 2000), and furthermore, replacement of eukaryote genes by endosymbiont orthologues (non-orthologous gene replacement, *sensu* Koonin et al., 1996) might be argued given the predicted catalytic superiority of endosymbiont metabolic enzymes. One could argue for other sources for the bacterial genes, but the simplest, most parsimonious, and most obvious source of the bulk of bacterial genes is the endosymbiont.

Returning to the question of how a biologically driven 'mass extinction' of nucleate eukaryotes by endosymbiont-harbouring relatives might have occurred, endosymbiosis would have provided two selectively advantageous and immediately

[1] Putative origins of replication have been identified in *Pyrococcus abyssi* (Myllykallio et al., 2000), *Pyrococcus horikoshii, Methanobacterium thermoautotrophicum* (Lopez et al., 1999) and *Thermotoga maritima* (Lopez et al., 2000). This work suggests archaeal replication is analogous to bacterial replication, being bidirectional, with a single origin per chromosome.

realised traits that might result in ESND breakdown and therefore extinctions. First, the established endosymbiont provided the host with ATP[2], and second, I argue that it had a large number of orthologous enzymes that were catalytically superior to the host complement. Third, the symbiosis presumably allowed previously anaerobic cells to diversify into aerobic and facultatively aerobic niches, which were inaccessible to their ancestors.

I will skip the establishment of symbiosis, since this has been covered by other authors (Martin and Müller, 1998; Andersson and Kurland, 1999; Rotte et al., 2000). Instead, I will concentrate on loss of redundant genes versus transfer to of genes to the nucleus. Selection within the endosymbiont-containing eukaryotes for fittest variants may have possibly resulted in a significant proportion of metabolic pathway orthologues being transferred to the nucleus, though some are expected to be lost owing to redundancy, as is seen in contemporary endosymbionts.

There may be an important difference between modern examples and the initial endosymbiosis however, and it might be predicted that the outcome would have been different depending upon whether the host is presumed to be archaeal or proto-eukaryotic. As described above, relative to extant eukaryotes, extant prokaryotes are r-selected. I predict that catalytic efficiency of archaeal and bacterial proteins carrying out identical reactions will be comparable. Assuming that this is true, genome fusion ought to reveal a chimeric origin for metabolic pathways. This is not the case however, with evidence to date (Ribeiro & Golding, 1998; Rivera et al., 1998; Horiike et al., 2001) suggesting host ancestral metabolic pathways have been replaced by endosymbiont pathways.

Modern eukaryotes are K-selected relative to bacteria, and if this is this niche discontinuity is an ancestral one, there would have been strongest selection in the latter for evolution of catalysis towards catalytic perfection. In metabolic pathways, where substrates, intermediates and products are usually similar, or can have a similar outcome (e.g. generation of ATP), the pathway is not as important as the outcome, since it is the products that are utilised. I therefore suggest that with the redundancy of orthologous metabolic pathways in the initial endosymbiosis, the endosymbiont pathways would have prevailed, being faster and more efficient. Furthermore, equivalent (analogous) pathways would be displaced from the host repertoire in favour of the endosymbiont pathways.

So, even though there would be a significant degree of loss through redundancy (as is seen in contemporary endosymbionts), those genes that conferred an advantage under endosymbiosis would tend to be maintained in the population. If

[2] I am not describing how the initial endosymbiosis was established, but rather how, subsequent to the development of the contemporary situation, the endosymbiont provided the host with energy. Both hydrogen and ox-tox hypotheses point out that the initial symbiosis was probably based on different interactions (Martin and Müller, 1998; Moreira and López-García, 1998; Andersson and Kurland, 1999)

these are orthologues or analogues of nuclear genes, the end result is selection for these over the nuclear genes, the latter being lost, and the former being ultimately transferred to the nucleus, as a result of the operation of Muller's Ratchet, and perhaps also oxidative DNA damage (Allen and Raven, 1996). However, moving to the nucleus in has the disadvantage that gene expression is slowed, meaning slower response times. Some products must be targetted to the endosymbiont, but transfer to the nucleus would alleviate the mutational pressure of being located in the endosymbiont.

The limitations of eukaryotic gene expression[3] would have meant that the organism could never have competed with prokaryote ancestors of the endosymbiont, but could however have resulted in extinction of proto-eukaryotes without an endosymbiont, within-population selection favouring those individuals that made use of the endosymbiont genes. I suggest that endosymbiosis caused a breakdown of ESNDs in eukaryotes effectively because of horizontal gene transfer.

This model is speculative, but in the next section I argue that there is a strong case for a proto-eukaryote host, based on the unique characteristics of eukaryote genome architecture, and the presence of putative RNA world relics in the nucleus, many of these having been lost from prokaryotes.

RNA relics in the nucleus.

If the nucleus is argued to be present in the host that endosymbiosed the ancient α-proteobacterium that gave rise to hydrogenosomes and mitochondria, then when did the nucleus arise? The standard argument is that it evolved in the eukaryotic branch subsequent to the split from archaea. In both pre- and post-endosymbiotic scenarios for the evolution of the nucleus, a key point is the presence in archaea of genes that contribute to eukaryote-specific traits (e.g. Martin & Müller, 1998; Moreira and López-García, 1998). This is used to suggest the evolutionary building blocks for the emergence of eukaryote-specific features evolved in the archaeal-like common ancestor of archaea and eukaryotes. It is equally feasible to argue that the presence of these genes in archaea, while suggesting a more recent common ancestry between archaea and eukaryotes than either with bacteria, is evidence for a eukaryote-like common ancestor, and loss of specific structures in archaea through reductive evolution!

One feature of the nucleus which might favour the latter possibility is that the nucleus is the site of the RNA processing events that produce mature functional RNAs and mRNAs (Lewis & Tollervey 2000). Most of these processing events

[3] It has been argued that eukaryote individual 'informational' genes (*sensu* Rivera et al., 1998) would not have been so readily replaced because of their involvement in large multimeric complexes with many interactions, as per the ribosome (Jain et al., 1999). Another explanation would be that if the host had a different genome architecture, i.e., much like that of modern eukaryotes, replacement could not occur without a fundamental change in architecture.

require functional RNAs that have been argued to be of RNA world origin, and a number of these are present only in eukaryotes (Poole et al. 1999). Any theory for the origin of the nucleus must consider the concentration of relic RNAs within this eukaryotic organelle, and the smaller numbers of RNA relics in prokaryote lineages (Penny & Poole 1999). I shall briefly review the distribution of RNA world relics before examining a possible scenario for the origin of the nuclear envelope prior to the emergence of eukaryotes, archaea and bacteria.

The concept of an RNA world is now well established, and enjoys a prominent position in origin of life studies, being pursued both through the identification of putative relics (Jeffares et al., 1998; Poole et al., 1999) and through *in vitro* selection studies (Yarus, 1999). While it is not clear whether an RNA-only world existed *sensu stricto*, it is certainly clear that there was an earlier period in the evolution of life where RNA played a more prominent role in cellular processes than now. At present, the main difficulty is that work on this problem has become separated from later periods in early evolution (Poole et al., 1999), with the question of the nature of the last universal common ancestor (LUCA) now being largely the domain of phylogenetics (Doolittle, 1999).

The RNA world model leads to the finding that the greatest diversity of putative RNA relics, as well as probable ancestral genome architecture, are concentrated in the nucleus of modern eukaryotes (Poole et al., 1998, 1999). Since this is the subject of other recent articles, I refer the reader to these for detail (Poole et al., 1998, 1999; Penny and Poole, 1999), and limit discussion here to a brief overview of the main points.

Several lines of argument suggest that features of the prokaryote lineages are derived, and that processes that have been considered ancient owing to their apparent simplicity may have evolved from a more complex (inefficient) precursor through reductive evolution (Reanney, 1974; Darnell and Doolittle, 1986; Forterre, 1995a; Poole et al., 1999). In extant eukaryotic cells, there exists a general processing pattern, where pre-tRNA, pre-rRNA and pre-mRNA are transcribed, processed to produce mature rRNA, tRNA and mRNA, exported from the nucleus, and then become involved in translation in the cytoplasm (Poole et al., 1999; Penny and Poole, 1999). The processing of all three occurs via ribonucleoprotein complexes, with tRNA being processed by the ubiquitous RNase P[4], which is a strong RNA world candidate with the RNA alone being sufficient for catalysis in some organisms (Altman and Kirsebom, 1999; Pannucci et al. 1999). Both spliceosomal snRNAs (Darnell and Doolittle, 1986, Gilbert & de Souza, 1999), and the snoRNAs involved in rRNA

[4] Bacterial RNase P is also involved in processing of rRNA, 4.5S RNA (srpRNA) and tmRNA (Altman & Kirsebom 1999). Eukaryotic RNase MRP, often considered a snoRNA, is specific for rRNA processing, carrying out the equivalent cleavage to that of bacterial RNase P on rRNA. On function (Venema & Tollervey 1999) and phylogeny (Collins et al. 2000), both RNase P and MRP appear to have a common origin.

processing (Poole et al., 1998, 1999, 2000) have been argued to date back to the RNA world. The origins of tRNA (Maizels and Weiner, 1999), rRNA (Noller, 1999; Poole et al., 1999) and mRNA (the 'introns first' hypothesis - see Poole et al., 1998, 1999) prior to the evolution of protein synthesis, is generally accepted. Strong arguments have also been made for the origin of telomerase & telomeres in the RNA world (Maizels and Weiner, 1999). Loss of a number of RNA traits, and reduction in RNA processing in prokaryotes, is considered more consistent with the RNA world hypothesis (Forterre, 1995a; Poole et al., 1999; Penny and Poole, 1999). As yet, no examples of RNA replacing protein have been found (Poole et al., 1999; A. Poole and D. Penny, in preparation).

Likewise, an argument for major features of the eukaryotic genome architecture (introns, multiple origins of replication, redundancy, linear chromosomes) being ancestral is well-developed (Poole et al., 1999), and supported by theoretical studies on the evolution of early genetic systems (Eigen and Schuster, 1979; Koch, 1984; Scheuring, 2000). Furthermore, the hypothesis that prokaryotes (but not eukaryotes) underwent a period of thermoadaptation from a mesophilic ancestor (Forterre, 1995a; Galtier et al., 1999) is consistent with circular genomes being found only in these lineages. (The only apparent selection pressure for circular genome architecture is its greater thermostability when compared with linear DNA [Marguet and Forterre, 1994]). This is consistent with the argument that eukaryotic telomerase RNA has its roots in the RNA world (Poole et al., 1999).

Several independent datasets thus point to the prokaryotes as being derived from a LUCA that had a number of features now found only in modern eukaryotes (Penny and Poole, 1999; Glansdorff, 2000). The concentration of putative RNA world relics in the eukaryote nucleus means that assumptions as to its origins should be reevaluated. The evolution of the nucleus needs to be considered in selective terms, and the assumption that it arose in the eukaryotes after they split from the prokaryote lineages must be relaxed. The absence of the nucleus in prokaryote lineages might equally be as a result of adaptive processes (through reduction), so selection scenarios for both gain and loss of such a structure need to be considered if progress is to be made on this problem. In the following sections, I will outline possible selection pressures for the origin of the nuclear envelope, and for its later loss in the lineages that ultimately gave rise to modern prokaryotes.

Nuclear Envelope-like structure for the first cells?

The RNA world theory represents the most ancient period in the evolution of life that can be reached using the 'top-down' approach, that is, working from extant biochemistry back towards the origin of life. This period is still far removed from the first steps toward life which have been established via the 'bottom-up' approach taken by prebiotic chemists. As Joyce and Orgel (1999) have pointed out, the 'Molecular Biologist's dream' is the 'Prebiotic Chemist's nightmare', with the latter group favouring one or more alternative genetic systems as intermediates between the origin of life, and the emergence of RNA (Joyce & Orgel, 1999; Maurel & Décout, 1999; Shapiro, 1999; Nelson et al., 2000).

In addition to the problems facing prebiotic chemists trying to understand the origin of life and later emergence of an RNA world is that of metabolism. RNA relics shed almost no light on essential biosyntheses, or possible energy sources for the first living entities (Jeffares et al., 1998). It is currently considered most likely that such prebiotic processes were carried out on two-dimensional surfaces than in a prebiotic soup. Surface chemistry avoids the problems of low concentration of precursors and hyrolysis by water that are expected in a prebiotic soup (Wächtersäuser, 1990, 1992; Maurel & Décout, 1999).

Whether or not life began on surfaces, it at some point became cellular, and this is a major problem for origin of life scenarios. It is not clear whether, by the time RNA arose, life had become cellular—there is no evidence for or against this possibility. However, a major problem with cellularisation is that a simple lipid bilayer closed in on itself is largely impermeable, and does not seem to be a likely intermediate in the evolution of modern cells. The advantage of cellular compartmentation is not only the concentration of substrates, products, and a genetic apparatus; a cell must also be 'leaky'. That is, it must allow waste out, and nutrients in.

Leakiness in the broadest sense has the disadvantage of making the cell completely at the mercy of the surrounding environment, so that a change in osmolarity can potentially pop the cell. Modern cells have a sophisticated system of pumps, gates and channels for regulating the concentration of protons and various ions within the cell, provide an effective way by which to buffer the cell from changes in the external environment, and allow nutrients in, and waste out.

A potential link between surface metabolism and cells is the semicell (Wächtershäuser, 1992; Maynard Smith and Szathmáry, 1995), helping in bridging the gap from surface metabolism to cells. A semicell would allow its contents to interact with its environment (the surface) without the requirement of a leaky membrane. However, this model must assume that the nutrients at the surface are replenished through diffusion, as it is not clear how the semi-cell could have divided, and without movement, it would simply use up all the resources at a given site.

This issue aside, in moving from hypothetical semicell (or some equivalent structure) to cell there is nevertheless the requirement for a 'leaky cell' stage in the origin of the cell, that is, a cell that could interact with the external environment. A cell with such pores may have been a necessary precursor to the modern cell.

Membrane pores are formed from proteins that traverse the lipid bilayer, requiring a hydrophobic region that traverses the bilayer, and a hydrophilic centre, through which ions and small molecules can pass, as well as hydrophilic extremities, on either side of the bilayer. Gram-negative bacteria have two membranes, separated by a periplasmic space. The inner membrane contains most of the well-known pumps and transporters, while the outer membrane is best described as leaky, owing to the presence of porins, such as OmpF. As homotrimers, these form pores which allow hydrophilic molecules up to 600Da to diffuse freely between the extracellular environment and the periplasmic space, though there are also those which facilitate uptake of specific metabolites (Koebnik et al., 2000). The minimum requirement for a transmembrane protein is an outer hydrophobic surface and an inner hydrophilic surface, and for a pore to be opened up. OmpF from *E. coli,* just such a protein, is comprised of 16 β-strands, producing a hydrophilic channel through the outer membrane (Cowan et al., 1992). However, the nuclear pore complex suggests there is a simpler alternative to a transmembrane pore.

The nuclear envelope is unique among biological membranes. It is a double membrane structure, as is found in Gram-negative bacteria and organelles, but the difference is that the inner and outer membranes are continuous. Nuclear pores do not traverse the lipid bilayer in the conventional sense (Goldberg and Allen, 1995). The nuclear pore complex is made up of around 30 different nucleoporins in yeast and around 50 in vertebrates, has a complex stoichiometry, and allows free diffusion of molecules up to ~40kDa (Kerminer and Peters, 1999; Allen et al., 2000; Rout et al., 2000; Shulga et al., 2000). It is anchored to the surrounding nuclear envelope by a small number of transmembrane proteins—three different types in yeast (Rout et al., 2000)—but the pore itself does not punch through the lipid bilayer (Rout et al., 2000; Wente 2000).

My suggestion is that a lipid bilayer arrangement like that seen in the extant nuclear envelope may be a better candidate for the first cell membrane, since, in principle, it permits very simple pores. This is because the interaction between protein forming the pore and the membrane does not require hydrophobic and hydrophilic regions. While the integral membrane proteins of the nuclear pore complex are central to the modern structure, and no doubt produce a much more stable pore-membrane interaction, a pore could in principle be constructed without such proteins, relying only on hydrophilic interactions with the polar head groups at the lipid-solvent interface. Thus, with such membrane architecture, a pore is possible through much simpler chemical interactions, than with transmembrane proteins.

Osmotic buffering.

The structure of the nuclear envelope and the general architecture of the nuclear pore complex provide insights into the possible structure of the first pore-containing leaky cells, without the requirement for transmembrane proteins. It does not however explain the early origin of the nucleus, only the possible utility of a simplified version of the nuclear envelope architecture.

A huge improvement on the leaky pore-containing cell, that would not require the advent of complex ion transporters or other complex proteins for osmoregulation, would be to have a second leaky membrane outside the first. This would provide a buffer region between the cell core and the environment, thereby serving to reduce the effect of small changes in osmotic pressure. For a roughly spherical cell, as the radius increases linearly, the volume increases by the cube of the radius. Thus, the 'cytoplasm' has the potential occupy a large volume, acting as a buffer region to the cell core, even though some of the volume of this buffer region is taken up by the nucleus.

Leakiness could likewise be regulated by the level of pore protein or the amount of lipid produced. The fewer pores and the larger the outer membrane, the less susceptible the cell is to osmotic pressure. Such regulation requires a mechanism for sensing pore density within the membrane, and I do not consider this likely in the earliest cells.

The buffered region resulting from a second, outer, pore-containing membrane might also provide the cell core with a nutrient-containing region, but these nutrients could only be utilised through diffusion into the core, and it is not obvious that all available nutrients would diffuse into the core. The presence of this proximal source of nutrients might therefore drive the evolution of protein or RNA transport out of the core, thereby allowing nutrients in the buffer region to be metabolised. One point that is worth raising is that, in addition to the effect of concentration gradients on determining the direction of diffusion, the rate of diffusion is inversely proportional to the square root of the molecular weight of the molecule (Graham's Law of diffusion). Thus, with large molecules and a shallow gradient, diffusion will be much slower than for small molecules and a steep gradient.

With use of metabolites such as ATP (in energy storage and nucleic acid synthesis) in the cell core, production of further ATP in the outer region would presumably result in diffusion into the core, where concentration is lower due to use. Furthermore, following Graham's Law, breakdown of large-sized nutrients with storage of the energy in a small molecule would result in faster diffusion because of size. That said, a higher concentration in the buffer, with lower concentrations in both the core and the external milieu would also result in metabolite loss from the leaky cell into the surrounding milieu. The development of better regulation of diffusion at the outer membrane would therefore be selectively advantageous not only because of improved buffering to fluctuating osmotic pressure, but because of nutrient/energy loss.

Transport of enzymes from the core to the buffer would presumably require transport of a diverse range of enzymes of varying sizes. While some would be effectively transported, others may not be transported at all. However, under general selection for transport of proteins and RNA into the buffer, if the components of the translation apparatus became transported, this would create a situation equivalent to the transport of all proteins to the site of metabolism. It would however require mRNA transport, and also create the opposite problem, in that proteins required in the core would need to be transported from the site of synthesis back into the core.

Such transport would probably not be 100% efficient, such that the translation apparatus would be present in both compartments, as would proteins and RNAs, some of which would be produced in the compartment where they were utilised, and others, which were superfluous to the functioning of the compartment. Such doubling up, resulting from inefficient mRNA and protein transport, would set the stage for selection of successively directional and specific RNA and protein transport pathways, as are seen in modern eukaryotic cells (Nakielny & Dreyfuss, 1999). Selection for translation in what is now the cytoplasm was presumably stronger than selection for nuclear translation.

The above discussion is speculative, but important in that it suggests a selection pressure for the localisation of translation in the cytoplasm. Viewing the cytoplasm as a buffer also suggests an important relationship between diffusion and active transport of proteins and RNA, the latter being selected for in that metabolism in the cytoplasm can potentially produce an artificial gradient, directing nutrients to the nucleus. It also implicitly involves proteins, and the possible relationship of such a scenario to the RNA world has not been examined. Before I do so, I shall first discuss the difference between the cytoplasmic and nuclear membranes.

The cytoplasmic membrane.

In arguing for the possibility that the nuclear envelope represents a relic of an ancient strategy for cellularisation, and that cells may have evolved a nucleus-cytoplasm form of compartmentation very early in the evolution of life, it is also necessary to address the nature of the extant cytoplasmic membrane. If I am to argue that the structure of the nuclear envelope was common to both the ancestral nuclear and cytoplasmic membranes, the theory must also explain how the cytoplasmic membrane of modern eukaryotes arose, which, structurally, is a conventional lipid bilayer, and not an envelope.

As protein synthesis became more accurate, hydrophobic transmembrane proteins would have become possible, and pores would be selected against in the outer cytoplasmic envelope, since these are leaky. Furthermore, I assume that an envelope is more prone to disruption, since the interactions between pore and the polar groups at the membrane surface can be disrupted by competing interactions with other molecules. In contrast, hydrophobic interactions are not easily disrupted in aqueous solution, so a single lipid bilayer with hydrophobic transmembrane proteins is expected to be a more robust architecture for the outer cell membrane. Indeed, the presence of anchoring proteins in the modern nuclear pore complex suggests this provides a stronger interaction between pore and membrane, reducing the possibility of structural disintegration at the interface between pore and bilayer. Such an arrangement could potentially have arisen early, and from this, the transmembrane proteins of the cytoplasmic membrane, though there is no evidence to support such conjecture.

What is harder to explain is the persistence of an envelope structure in the nucleus. And indeed, this is as problematic whether one argues for a late origin, as per the standard model, or for an early origin, as is being considered here. The continuity of the nuclear envelope with the endoplasmic reticulum is one possible issue to examine. Is the endoplasmic reticulum only possible because of the nuclear envelope, or vice versa? Gupta & Golding (1996) have drawn a schematic that suggests the two membrane systems arose from the invaginations of the host, with the endosymbiont eventually losing its membrane. Their imaginative picture goes some way toward describing how the two membranes could form from an endosymbiotic event, but does not explain the origin of the nuclear pore complex, nor the function of the endoplasmic reticulum.

The origin of the endoplasmic reticulum is as problematic as the origin of the nuclear envelope, and I do not address this question here. With regard to the persistence of a nuclear envelope structure in eukaryotes, this is likewise difficult to establish. The above suggestion that there was selection for a robust outer (cytoplasmic) membrane, due to disruption, would not necessarily apply for the nucleus. Under the model described in this paper, the persistence of the nuclear envelope in the eukaryotic cell is not clear. In light of this, the best option is to apply the neutral theory (Stoltzfus, 1999), and argue that there was simply no selection to remove this structure, and over time, it would have become essential simply because other functions revolved around its presence.

An RNA cell?

While the minimal requirements for the formation of a cell are hard to ascertain (Szostak et al., 2001), one major issue is whether a pre-protein cell is at all feasible. It is not likely that RNA could form pores that traverse the lipid bilayer, since this requires both hydrophobic and hydrophilic moieties. Highly modified nucleosides have nevertheless been shown to have pore-forming capabilities via a G-quartet structure (Forman et al., 2000), but there is no evidence such modifications were part of the RNA world repertoire, and while these studies do not make use of RNA derivatives, this could probably be achieved.

More tantalisingly, Khvorova et al. (1999) carried out in vitro selection experiments to screen for RNA that bound phospholipid membranes, and subsequently examined their ability to alter phospholipid membrane permeability. Their experiments revealed that RNAs can increase the ion permeability of phospholipid bilayers. While RNA is predicted to readily interact with the polar head group and glycerol phosphate moieties of phospholipids, it is less obvious that RNA could traverse the bilayer. However, interactions between RNA and hydrocarbons, in the form of the side chains of valine and isoleucine, have previously been demonstrated (Majerfeld & Yarus, 1994; 1998). These interactions were mediated by specific hydrophobic pockets within the RNAs, thereby adding hydrophobic chemistry to the list of RNA chemistries.

It is known that 2'-*O*-ribose methylation is found in all three domains of life, and likely dates back to the RNA world (Poole et al. 2000). Complete 2'-*O*-ribose methylation of double-stranded RNA produces a hydrophobic cushion in the deep groove of the helix (Popenda et al. 1997; Adamiak et al. 1997). Ribose methylation could therefore be a potential means of producing RNAs with hydrophobic moieties, and might be one direction to take in building upon Majerfeld & Yarus's work on RNA hydrophobicity. Nevertheless, it remains unclear whether such interactions could be sufficient to form a transmembrane pore, and I favour the possibility of a non-hydrophobic, non-transmembrane RNA pore.

Given that a membrane structure like the nuclear envelope in principle permits pores without the requirement for these to traverse the lipid bilayer, the naturally-occurring G-quartet structure taken on by RNAs such as telomerase RNA (Williamson et al., 1989) may be the strongest link to a cellular RNA world. When stacked, G-quartets produce a pore-like structure that has even been shown to permit ion diffusion (Gilbert and Feigon, 1999; Hud et al., 1999). The pore does not need to be formed solely from G residues, and can include Us. The structure is simple, can self-assemble, and should be capable of interacting with the surface of a lipid bilayer. Importantly, the plausibility of a nuclear envelope-like membrane with RNA pores can be tested *in vitro*.

Loss of the Nucleus from prokaryotes?

In arguing for the antiquity of the overall arrangement of the eukaryote cell, that is, a nuclear envelope and a cytoplasmic membrane, the absence of the nuclear envelope from prokaryotes must also be accounted for. In eukaryotes, transcription and RNA processing occur in the nucleus (Lewis and Tollervey, 2000), while translation occurs in the cytoplasm. If this division of processes is ancestral, an argument for the loss of this structure is possible under Forterre's thermoreduction hypothesis (Forterre, 1995a) and/or under r selection.

The general argument for loss of eukaryote structures or processes has been developed extensively elsewhere (Forterre, 1995a; Poole et al., 1998, 1999, 2001), so I will only provide a brief treatment here. Relative to eukaryotes, prokaryotes can be considered r-selected (Carlile, 1982; Poole et al., 1999, 2001). In short, prokaryotes display a fast response to the presence of a nutrient, this response involving activation of gene expression for metabolising the nutrient, and subsequent entry into exponential growth. Nutrient availability fluctuates in the environment, so there is selection for fast metabolism upon detection, and fast doubling times—those organisms that proliferate fastest will tend to win out over slower competitors. One consequence of such competition is that increases in the rate of gene expression are expected to be selectively advantageous (Poole et al., 2001). If the ancestor of modern prokaryotes expressed genes in a similar way to modern eukaryotes, the expectation would be that there would be strong selection for loss of the nuclear envelope. In modern eukaryotes, a transcript is synthesised, spliced, capped and polyadenylated, undergoes a quality control check for damage (Ibba and Söll, 1999), and translation occurs after transport across the nuclear membrane. In bacteria, translation begins while the transcript is still being synthesised, processing events are minimal, and quality control is skipped altogether, with damaged mRNAs being translated anyway. mRNA damage causes a ribosome to stall, and this is released via a specific mechanism involving tmRNA (Ibba and Söll, 1999). Under r selection, there would be selection for faster rates of gene expression (Poole et al., 1998, 1999). If the nucleus were ancestral, loss would have been advantageous in that it would have removed a step in the gene expression pathway, speeding the response time, and therefore enabling faster protein production. This would have been important not only for speeding response times, but would also have permitted faster cell division, assuming protein synthesis was a rate-limiting step in this process.

The thermoreduction hypothesis (Forterre, 1995a) is that prokaryote lineages underwent a period of adaptation to high temperature environments which resulted in reduction in thermolabile traits. Single-stranded RNA is known to be thermolabile (see Forterre, 1995a), so the loss of an ancestral nuclear envelope can also be argued from the viewpoint that this would shorten the time between mRNA synthesis and protein synthesis, and thus reduce the chance of transcript thermodegradation. Thermoreduction and r selection can both account for the reduction of RNA world relics from prokaryotes, and are indistinguishable, both on reduction of relic RNAs, and on the possible loss of the nuclear envelope, from prokaryote lineages (through reductive evolution). Again, what is important here is not whether this scenario is correct but that the extreme position can be argued, and that this is consistent with other explanations for the origin and evolution of eukaryotes and prokaryotes.

Conclusions & outstanding problems.

In this paper, I have examined a number of issues regarding the origin of the eukaryote nucleus. In addressing the areas that are currently problematic, I have advanced an extreme viewpoint. This can be summarised as four separate conclusions:

1. That the RNA world dataset suggests proto-eukaryotes were the host of the endosymbiont that gave rise to hydrogenosomes and mitochondria.
2. That, given evidence that the prokaryote lineages have arisen through reductive evolution, the LUCA may have had a nucleus.
3. That an argument can be made that the origin of the nucleus is concurrent with origin of the first cells, eliminating the problem of low permeability of lipid bilayers, and at the same time, minimising the problems for an early leaky cell.
4. That pores similar to those formed by the modern nuclear pore complex could have predated transmembrane pores, and that this architecture might even be feasible for an RNA world.

That the conclusions can be treated separately is an important point. For instance, it may be accepted that the host for the forerunner to mitochondria and hydrogenosomes was nucleate, as per conclusion 1, without requiring that the nucleus was a feature of the LUCA. Similarly, conclusions 3 and 4 may be of interest to understanding the origin of the first cells without requiring that conclusions 1 and 2 are correct.

The hypothesis I have advanced is necessarily speculative and, if previous attempts at this question are anything to go by, probably wrong. However, the value of taking an evolutionary approach to understanding modern cell structure and function is that it ties together a wide range of experimental data and observation. Evolutionary hypotheses allow an explanation of unrelated phenomena within the context of a theory. An evolutionary approach also provides a framework for asking new questions that would otherwise not get asked. While evolutionary hypotheses are often not amenable to simple tests to prove or disprove them (there is no 'killer experiment' as Maizels and Weiner (1999) have put it), they can nevertheless advance knowledge by providing a framework for understanding existing and subsequent experimental results, and may even provide a novel way of choosing subsequent experiments. The following points serve to illustrate that the same data often fit more than one hypothesis, so it is worth being cautious.

Reliance on the presence or absence of a trait in order to determine 'modern' and 'ancestral', and therefore how these 'ancestral' creatures became 'modern' is nonsense if the theory has already assumed the direction of evolution. A good evolutionary theory should aim to identify potential selection pressures that can account for change, and hence might then help us establish in *which direction* evolution went.

The hypothesis I present is worthwhile because it identifies a selection pressure for the origin of the nucleus-cytoplasm organisation of eukaryotic cells. This represents a significant departure from most treatments of the problem, which are largely descriptive in nature. Furthermore, having argued that the nucleus predates the divergence of eukaryotes, archaea and bacteria, I identify a clear selection for the loss of nuclear structure in the latter two groups. The theory also provides an explanation of the differences between eukaryotic and prokaryotic ultrastructure that is not at odds with the identification of the eukaryotic nucleus as the source of greatest diversity in RNA world relics. The 'nucleus first' hypothesis is consistent with both the thermoreduction hypothesis as an explanation for the origin of circular genomes in prokaryotes as an adaptation to high temperature, and the loss of RNA world relics from prokaryotes as a derived trait (Forterre, 1995a; Poole et al., 1999). The similarity of the genome organisation of modern eukaryotes (linear with multiple chromosomes, each with multiple origins of replication, and not haploid) with what is predicted from theory to be a low-fidelity genome architecture able to withstand high error rates also argues that this organisation never arose from a prokaryotic ancestor with a circular genome (Poole et al., 1998; 1999). This, together with the RNA relic dataset, strongly suggests that the nucleus was not an endosymbiont and cannot be readily understood as having an archaeal origin. Rather, the nucleus is currently best viewed as a candidate for the most ancient 'living fossil' of early evolution. The prokaryotic lineages are considered to have lost the nucleus in response to selection to reduce the time it takes to synthesise a protein, as well as to reduce the risk of mRNA degradation at elevated temperatures (Poole et al., 1998).

What constitutes a useful theory is the ability to continue to explain the existing data as well as new data that come to hand, where other theories fail to. In the current case, I have argued that this theory best explains the available data, *without* justifying the direction of evolution based on a transition from simple to complex, but rather in terms of selection. The theory may be replaced in time, but it succeeds over current theories because it is better able to explain the available data in an evolutionary context. The possibility of a leaky nuclear envelope with G-quartet structures acting as pores is experimentally testable. Crucially, the hypothesis describes a selection pressure for the origin of a nucleus-like organisation early in the evolution of the cell, and is the first serious attempt to address the structure of the nuclear envelope in the context of the origins of this structure. Whether pores were provided by G-quartets or not, I suggest that the continuous double membrane system is a likely structure for an early cell membrane since it does not require amphipathic membrane-spanning proteins for pore production. Channel-shaped or leaky basic proteins are all that is required in this role as pore-forming proteins only ever come into contact with the polar head groups of the membrane surface.

A final question that should be addressed is the function of additional membranes in organisms such as members of the *Pirellula* (Fuerst and Webb, 1991; Lindsay et al., 1997) and the Gram-negative bacteria (Gupta, 1998). In the case of the *Pirellula*, it will be important to establish what the role of the nucleus-like compartmentation is, and whether there are any similarities with the eukaryote nucleus over and above the two-membrane structure seen in *P. marina* and *P. stalyei* (Lindsay et al., 1997). This may open up the possibility for detailed understanding of the structure, function and evolution of such ultrastructure, and progress with these organisms may also shed new light on the question of the selection pressure that gave rise to the nucleus.

Postscript.

This work is very much a work in progress. The argument that the origin of the nuclear envelope is concurrent with the first cells is particularly interesting in light of Blobel's (1980) concept of an inside-out cell, which has been extended by Cavalier-Smith (1987). The latter author has argued that the first cells were Gram-negative bacteria, with inside-out cells containing cell wall material in their lumen, allowing them to bend. Eventually, these would have bent round on themselves, to form double membrane-bound cells (Cavalier-Smith, 1987). This looks good on paper, but there are two difficulties. First, what selection is there for bending in the first place? Second, and of greater interest, how would a cell that closed in on itself interacted with its environment? A double membrane with a cell wall between the two membranes would have been a particularly impermeable structure!

Blobel's idea is nevertheless interesting in light of the recent suggestions that many problems for the RNA world could in principle be solved by a 'lipid world' (Luisi et al., 1999; Segré and Lancet, 2000; Segré et al., 2001). Excitingly, the biosynthesis of phospholipids involves activated precursors consisting of a nucleotide moiety and a hydrophobic moiety (e.g. CDP-diacylglycerol in phosphatidyl serine and phosphatidyl inositol synthesis, CDP-choline in phosphatidyl choline synthesis, CDP-ethanolamine in phosphatidyl ethanolamine synthesis, and addition of sugar residues such as UDP-glucose, UDP-galactose). Not only are these molecules potentially synthesisable under prebiotic conditions (Rao et al., 1982, 1987; Mar et al., 1987), they provide a possible link to the RNA world since nucleotide cofactors are arguably relics of the RNA world (White, 1976). The possibility that genetic information was initially encoded by heterogeneous lipid vesicles (Segré and Lancet, 2000) is in itself interesting, with progress on autocatalytic self-replicating micelles and vesicles, central to the feasibility of a lipid world (Bachmann et al., 1992; Veronese and Luisi, 1998).

What is exciting for the origin of the RNA world and the link between surface and cell metabolism is that hypothesised genetic take-over by RNA does not require invoking intermediate steps for which there are no identifiable relics (e.g. clay surfaces or PNA). Phosphate chemistry is common to both phospholipids and RNA and phosphate chemistry is considered to have played a central role in the emergence of life (Westheimer, 1987; Baltscheffsky, 1997). Moreover, not only can phosphatidylnucleosides self-assemble to form vesicles, they can in principle permit Watson-Crick base pairing via the lipid head groups possessing nucleosides (Berti et al., 1998). Not only do the head groups permit a link between lipid and RNA worlds, lipids are able to carry out catalyses (see Segré et al., 2001), and the surface of a lipid bilayer would provide a two-dimensional surface for sequestering molecules, as per other surface scenarios.

The inside-out cell might thus have initially provided a surface on which to carry out catalysis. Cooperation between such cells would lead to aggregations of inside-out cells, and these would form the basis of a 'pseudocell', where the 'cytoplasm' represents an inner compartment which resembles the nucleus (minus nuclear pores) in structure. This can explain the transition from surface to cell without invoking a semicell, and would also account for the unique structure of the nuclear envelope. A leaky membrane would not initially be at issue, as surface chemistry would initially dominate, and pores could form without the requirement for spanning a membrane. I plan to develop this idea (and those described in the latter sections of the paper) more thoroughly, having represented only the broad concepts here.

References.

Adamiak, D.A., Milecki, J., Popenda, M., Adamiak, R.W., Dauter, Z. and Rypniewski, W.R. (1997) Crystal structure of 2'-O-Me(CGCGCG)2, an RNA duplex at 1.30 A resolution. Hydration pattern of 2'-O-methylated RNA. *Nucleic Acids Res* **25**, 4599-4607.

Akhmanova, A., Voncken, F., van Alen, T., van Hoek, A., Boxma, B., Vogels, G., Veenhuiss, M. and Hackstein, J.H.P. (1998) A hydrogenosome with a genome. *Nature*, **396**, 527-528.

Albery, W.J. and Knowles, J.R. (1976) Evolution of enzyme function and the development of enzyme efficiency. *Biochemistry* **15,** 5631-5640.

Allen, J.F. and Raven, J.A. (1996) Free-radical-induced mutation vs redox regulation: costs and benefits of genes in organelles. *J. Mol. Evol.* **42,** 482-492.

Allen, T.D., Cronshaw, J.M., Bagley, S., Kiseleva, E. and Goldberg, M.W. (2000) The nuclear pore complex: mediator of translocation between nucleus and cytoplasm. *J. Cell. Sci.*, **113**, 1651-1659.

Altman, S. and Kirsebom, L.A. (1999) Ribonuclease P. In: The RNA World, 2nd Edn. Gesteland, R.F., Cech, T.R. and Atkins, J.F., eds. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York. pp. 351-380.

Andersson, D.I. and Hughes, D. (1996) Muller's ratchet decreases fitness of a DNA-based microbe. *Proc. Natl. Acad. Sci. USA*, **93**, 906-907.

Andersson, J.O. and Andersson, S.G.E. (1999a) Insights into the evolutionary process of genome degradation. *Curr. Opin. Genet. Dev.*, **9**, 664-671.

Andersson, J.O. and Andersson, S.G.E. (1999b) Genome degradation is an ongoing process in *Rickettsia. Mol. Biol. Evol.,* **16**, 1178-1191.

Andersson, S.G.E. and Kurland, C.G. (1999) Origins of mitochondria and hydrogenosomes. *Curr. Opin. Microbiol.,* **2**, 535-541.

Andersson, S.G.E., Zomorodipour, A., Andersson, J.O., Sicheritz-Pontén, T., Alsmark, U.C.M., Podowski, R.M., Näslund, A.K., Eriksson, A.-S.,Winkler, H.H. and Kurland, C.G. (1998) The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. *Nature*, **396**, 133-140.

Angert, E.R., Clements, K.D. and Pace, N.R. (1993) The largest bacterium. *Nature*, **362**, 239-241.

Bachmann, P.A., Luisi, P.L and Lang, J. (1992) Autocatalytic self-replicating micelles as models for prebiotic structures. *Nature*, **357**, 57-59.

Baltscheffsky, H. (1997) Major "anastrophes" in the origin and early evolution of biological energy conversion. *J. Theor. Biol.*, **187**, 495-501.

Berg, O.G. and Kurland, C.G. (2000) Why mitochondrial genes are most often found in nuclei. *Mol. Biol. Evol.*, **17**, 951-961.

Berti, D., Baglioni, P., Bonaccio, S., Barsacchi-Bo, G. and Luisi, P.L. (1998) Base complementarity and nucleoside recognition in phosphatidylnucleoside vesicles. *J. Phys. Chem. B*, **102**, 303-308.

Biagini, G.A. and Bernard, C. (1999) Primitive anaerobic protozoa: a false concept? *Mol. Microbiol.,* **146**, 1019-1020.

Blanchard, J.L. and Lynch, M. (2000) Organellar genes: why do they end up in the nucleus? *Trends Genet.*, **16**, 315-320.

Blobel, G. (1980) Intracellular protein topogenesis. *Proc. Natl. Acad. Sci. USA*, **77**, 1496-1500.

Carlile, M.J. (1982) Prokaryotes and eukaryotes: strategies and successes. *Trends Biochem. Sci.*, **7**, 128–130.

Cavalier-Smith, T. (1983) A six-kingdom classification and a unified phylogeny. *Endocytobiol.*, **2**, 1027-1034.

Cavalier-Smith, T. (1987) The origin of cells: a symbiosis between genes, catalysts, and membranes. *Cold Spring Harb. Symp. Quant. Biol.*, **52**, 805-824.

Cavalier-Smith, T. (1988) Origin of the cell nucleus. *BioEssays* **9,** 72-78.

Cavalier-Smith, T. (2000) Membrane heredity and early chloroplast evolution. *Trends Plant Sci.,* **5**, 174-182.

Clements, K.D. and Bullivant, S. (1991) An unusual symbiont from the gut of surgeonfishes may be the largest known prokaryote. *J. Bact.*, **173**, 5359-5362.

Cole, S.T., Eiglmeier, K., Parkhill, J., James, K.D., Thomson, N.R., Wheeler, P.R., Honoré, N., Garnier, T., Churcher, C., Harris, D., Mungall, K., Basham, D., Brown, D., Chillingworth, T., Connor, R., Davies, R.M., Devlin, K., Duthoy, S., Feltwell, T., Fraser, A., Hamlin, N., Holroyd, S., Hornsby, T., Jagels, K., Lacroix, C., Maclean, J., Moule, S., Murphy, L., Oliver, K., Quail, M.A., Rajandream, M.-A., Rutherford, K.M., Rutter, S., Seeger, K., Simon, S., Simmonds, M., Skelton, J., Squares, R., Squares, S., Stevens, K., Taylor, K., Whitehead, S., Woodward, J.R., Barrell, B.G., 2001. Massive gene decay in the leprosy Bacillus. Nature 409, 1007-1011.

Collins, L.J., Moulton, V. and Penny, D. (2000) Use of RNA secondary structure for studying the evolution of RNase P and RNase MRP. *J. Mol. Evol.*, **51**, 194-204.

Cowan, S.W., Schirmer, T., Rummel, G., Steiert, M., Ghosh, R., Pauptit, R.A. et al. (1992) Crystal structures explain functional properties of two *E. coli* porins. *Nature,* **358**, 727-733.

Darnell, J.E. and Doolittle, W.F. (1986) Speculations on the early course of evolution. *Proc. Natl. Acad. Sci. USA*, **83**, 1271-1275.

Doolittle, W.F. (1999) Phylogenetic classification and the universal tree. *Science*, **284**, 2124-2128.

Douglas, S., Zauner, S., Fraunholz, M., Beaton, M., Penny, S., Deng, L.-T., Wu, X., Reith, M., Cavalier-Smith, T. and Maier, U.-G. (2001) The highly reduced genome of an enslaved algal nucleus. *Nature* **410**, 1091-1096.

Eigen, M. and Schuster, P. (1979) The hypercycle: a principle of natural self-organization. Springer-Verlag, Berlin.

Embley, T.M. and Hirt, R.P. (1998) Early branching eukaryotes? *Curr. Opin. Genet. Dev.* 1998, **8**, 624-629.

Fishelson, L., Montgomery, W.L. and Myrberg, Jr., A.A. (1985) A unique symbiosis in the gut of tropical herbivorous surgeonfish (Acanthuridae: Teleostei) from the Red Sea. *Science*, **229**, 49-51.

Flügge, U.-I. (2000) Transport in and out of plastids: does the outer envelope membrane control the flow? *Trends Plant Sci.,* **5**, 135-137.

Forman, S.L., Fettinger, J.C., Pieraccini S., Gottarelli G., Davis, J.T. (2000) Toward artificial ion channels: a lipophilic G-quadruplex. *J. Am. Chem. Soc.* **122**, 4060-4067.

Forterre, P. (1995a) Thermoreduction, a hypothesis for the origin of prokaryotes. *C.R. Acad. Sci. Paris III,* **318**, 415-422.

Forterre, P. (1995b) Looking for the most "primitive" organism(s) on Earth today: the state of the art. *Planet. Space Sci.*, **43**, 167-177.

Forterre, P. (1997) Archaea: what can we learn from their sequences?. *Curr. Opin. Genet. Dev.*, **7**, 764-770.

Forterre, P. and Philippe, H. (1999) Where is the root of the universal tree of life? *Bioessays,* **21**, 871-879.

Fraser, C.M., Gocayne, J.D., White, O., Adams, M.D., Clayton, R.A., Fleischmann, R.D., Bult, C.J., Kerlavage, A.R., Sutton, G., Kelley, J.M., Fritchman, J.L., Weidman, J.F., Small, K.V., Sandusky, M., Fuhrmann, J., Nguyen, D., Utterback, T.R., Saudek, D.M., Phillips, C.A., Merrick, J.M., Tomb, J.-F., Dougherty, B.A., Bott, K.F., Hu, P.-C., Lucier, T.S., Peterson, S.N., Smith, H.O., Hutchison III, C.A. and Venter, J.C. (1995) The minimal gene complement of Mycoplasma genitalium. *Science*, **270**, 397-403.

Fraser, C.M., et al., 1997. Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi.* Nature 390, 580-586.

Fraser, C.M., et al., 1998. Complete genome sequence of *Treponema pallidum*, the syphilis spirochaete. Science 281, 375-388.

Fuerst, J.A. and Webb, R.I. (1991) Membrane-bounded nucleoid in the eubacterium *Gemmata obscuriglobus. Proc. Natl. Acad. Sci. USA*, **88**, 8184-8188.

Galtier, N., Tourasse, N. and Gouy, M. (1999) A nonhyperthermophilic common ancestor to extant life forms. *Science* **283**, 220-221.

Gaspin, C., Cavaillé, J., Erauso, G. and Bachellerie, J.P. (2000) Archaeal homologs of eukaryotic methylation guide small nucleolar RNAs: lessons from the Pyrococcus genomes. *J. Mol. Biol.*, **297**, 895-906.

Gilbert, W. and de Souza, S.J. (1999) Introns and the RNA world. In: The RNA World, 2nd Edn. Gesteland, R.F., Cech, T.R. and Atkins, J.F., eds. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York. pp. 221-231.

Gilbert, D.E. and Feigon, J. (1999) Multistranded DNA structures. *Curr. Opin. Struct. Biol.*, **9**, 305-314.

Gilson, P.R., Maier, U.-G. and McFadden, G.I. (1997) Size isn't everything: lessons in genetic miniturisation from nucleomorphs. *Curr. Opin. Genet. Dev.*, **7**, 800-806.

Glansdorff, N. (2000) About the last common ancestor, the universal life-tree and lateral gene transfer: a reappraisal. *Mol. Microbiol.*, **38**, 177-185.

Goldberg, M.W. and Allen, T.D. (1995) Structural and functional organization of the nuclear envelope. *Curr. Opin. Cell Biol.,* **7**, 301-309.

Gray, M.W., Burger, G. and Lang, B.F. (1999) Mitochondrial evolution. *Science*, **283**, 1476-1481.

Grotzinger, J.P. and Rothman, D.H. (1996) An abiotic model for stromatolite morphogenesis. *Nature,* **383**, 423-425.

Gupta, R.S. (1998) What are archaebacteria: life's third domain or monoderm prokaryotes related to Gram-positive bacteria? A new proposal for the classification of prokaryotic organisms. *Mol. Microbiol.*, **29**, 695-707.

Gupta, R.S. and Golding, G.B. (1996) The origin of the eukaryotic cell. *Trends Biochem Sci.* **21**, 166-171.

Gupta, R.S. and Singh, B. (1994) Phylogenetic analysis of 70kD heat shock protein sequences suggests a chimaeric origin for the eukaryotic nucleus. *Curr. Biol.*, **4**, 1104–1114.

Henze, K., Morrison, H.G., Sogin, M.L. and Müller, M. (1998) Sequence and phylogenetic position of a class II aldolase gene in the amitochondriate protist, *Giardia lamblia. Gene*, **222**, 163-168.

Himmelreich, R., et al., 1996. Complete genome sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae.* Nucleic Acids Res. 3, 109-136.

Horiike, T., Hamada, K., Kanaya, S. and Shinozawa, T. (2001) Origin of eukaryotic cell nuclei by symbiosis of Archaea and Bacteria is revealed by homology-hit analysis. *Nat. Cell Biol.* **3**, 210-214.

Hud, N.V., Schultze, P., Sklenár, V. and Feigon, J. (1999) Binding sites and dynamics of ammonium ions in a telomere repeat DNA quadruplex. *J. Mol. Biol.*, **285**, 233-243.

Ibba, M. and Söll, D. (199) Quality control mechanisms during translation. *Science*, **286**, 1893-1897.

Jain, R., Rivera, M.C. and Lake, J.A. (1999) Horizontal transfer among genomes: the complexity hypothesis. *Proc. Natl. Acad. Sci. USA*, **96**, 3801-3806.

Jeffares, D.C., Poole, A.M. and Penny, D. (1998) Relics from the RNA world. *J. Mol. Evol.*, **46**,18-36.

Joyce, G.F. and Orgel, L.E. (1999) Prospects for understanding the origin of the RNA world. In: The RNA World, 2nd Edn. Gesteland, R.F., Cech, T.R. and Atkins, J.F., eds. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York. pp. 49-77.

Kalman, S., Mitchell, W., Marathe, R., Lammel, C., Fan, J., Hyman, R.W., Olinger, L., Grimwood, J., Davis, R.W., Stephens, R.S., 1999. Comparative genomes of *Chlamydia pneumoniae* and *C. trachomatis. Nat. Genet.,* **21**, 385-389.

Katz, L.A. (1998) Changing perspectives on the origin of eukaryotes. *Trends Ecol. Evol.*, **13**, 493-497.

Keeling, P.J. (1998) A kingdom's progress: archaezoa and the origin of eukaryotes. *Bioessays*, **20**, 87-95.

Keeling, P.J. and Doolittle, W.F. (1997) Evidence that eukaryotic triosephosphate isomerase is of alpha-proteobacterial origin. *Proc. Natl. Acad. Sci. USA*, **94**, 1270-1275.

Keeling, P.J. and McFadden, G.I. (1998) Origins of microsporidia. *Trends Microbiol.*, **6**, 19-23.

Kerminer, O. and Peters, R. (1999) Permeability of single nuclear pores. *Biophysical J.*, **77**, 217-228.

Khvorova, A., Kwak, Y.-G., Tamkun, M., Majerfeld, I. and Yarus, M. (1999) RNAs that bind and change the permeability of phospholipid membranes. *Proc. Natl. Acad. Sci. USA*, **96**, 10649-10654.

Koebnik, R., Locher, K.P. and Van Gelder, P. (2000) Structure and function of bacterial outer membrane proteins: barrels in a nutshell. *Mol. Microbiol.*, **37**, 239-253.

Koch, A.L. (1984) Evolution vs the Number of Gene Copies Per Primitive Cell. *J. Mol. Evol.* **20**, 71-76.

Koonin, E.V., Mushegian, A.R. and Bork, P. (1996) Non-orthologous gene replacement. *Trends Genet.*, **12**, 334-336.

Lake, J.A. and Rivera, M.C. (1994) Was the nucleus the first endosymbiont? *Proc. Natl. Acad. Sci. USA*, **91**, 2880-2881.

Lewis, J.D. and Tollervey, D. (2000) Like attracts like: getting RNA processing together in the nucleus. *Science*, **288**, 1385-1389.

Li, Y.-L. (1999) The primitive nucleus model and the origin of the cell nucleus. *Endocyt. Cell Res.*, **13**, 1-86.

Liaud, M.-F., Lichtlé, C. Apt, K., Martin, W. and Cerff, R. (2000) Compartment-specific isoforms of TPI and GAPDH are imported into diatom mitochondria as a fusion protein: evidence in favor of a mitochondrial origin of the eukaryotic glycolytic pathway. *Mol. Biol. Evol.*, **17**, 213-223.

Lindsay, M.R., Webb, R.I. and Fuerst, J.A. (1997) Pirellulosomes: a new type of membrane-bounded cell compartment in planctomycete bacteria of the genus *Pirellula. Microbiology*, **143**, 739-748.

Ljungman, M. and Hanawalt, P.C. (1992) Efficient protection against oxidative DNA damage in chromatin. *Mol. Carcinog.*, **5**, 264-269.

Lockhart, P.J., Steel, M.A., Barbrook, A.C., Huson, D.H., and Howe, C.J. (1998) A covariotide model describes the evolution of oxygenic photosynthesis. *Mol. Biol. Evol.* **15**, 1183-1188.

Lopez, P., Forterre, P., le Guyader, H. and Philippe, H. (2000) Origin of replication of *Thermotoga maritima. Trends Genet.*, **16,** 59-60.

Lopez, P., Philippe, H., Myllykallio, H. and Forterre, P. (1999) Identification of putative chromosomal origins of replication in Archaea. *Mol. Microbiol.*, **32,** 883-886.

López-García, P. and Moreira, D. (1999) Metabolic symbiosis at the origin of eukaryotes. *Trends Biochem. Sci.*, **24**, 88-93.

Lowe, D.R. (1994) Abiological origin of described stromatolites older than 3.2 Ga. *Geology* **22**, 387-390.

Luisi, P.L., Walde, P. and Oberholzer, T. (1999) Lipid Vesicles as Possible Intermediates in the Origin of Life. *Curr. Opin. Colloid Interface Sci.*, **4**, 33-39.

Maizels, N. and Weiner, A.M. (1999) The genomic tag hypothesis: what molecular fossils tell us about the evolution of tRNA. In: The RNA World, 2nd Edn. Gesteland, R.F., Cech, T.R. and Atkins, J.F., eds. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, p79-111.

Majerfeld, I. and Yarus, M. (1994) An RNA pocket for an aliphatic hydrophobe. *Nat. Struct. Biol.*, **1**, 287-292.

Majerfeld, I. and Yarus, M. (1998) Isoleucine:RNA sites with associated coding sequences. *RNA*, **4**, 471-478.

Marguet, E. and Forterre, P. (1994) DNA stability at temperatures typical for thermophiles. *Nucleic Acids Res.*, **22**, 1681-1686.

Margulis, L. (1970) Origin of eukaryotic cells. Yale University Press, New Haven.

Margulis, L., Dolan, M.F. and Guerrero, R. (2000) The chimeric eukaryote: Origin of the nucleus from the karyomastigont in amitochondriate protists. *Proc. Natl. Acad. Sci. USA*, **97**, 6954-6959.

Martin, W., Brinkmann, H., Savona, C. and Cerff, R. (1993) Evidence for a chimeric nature of nuclear genomes: eubacterial origin of eukaryotic glyceraldehyde-3-phosphate dehydrogenase genes. *Proc. Natl. Acad. Sci. USA*, **90**, 8692-8696.

Martin, W. and Müller, M. (1998) The hydrogen hypothesis for the first eukaryote. *Nature*, **392**, 37-41.

Martin, W. (1999a) Mosaic bacterial chromosomes: a challenge en route to a tree of genomes. *Bioessays*, **21**, 99-104.

Martin, W. (1999b) A briefly argued case that mitochondria and plastids are descendents of endosymbionts, but that the nuclear compartment is not. *Proc. R. Soc. Lond. B*, **266**, 1387-1395.

Martin, W. (1999c) Primitive anaerobic protozoa: the wrong host for mitochondria and hydrogenosomes? *Mol. Microbiol.* **146**, 1021-1022.

Martin, W., Stoebe, B., Goremykin, V., Hansmann, S., Hasegawa, M. and Kowallik, K.V. (1998) Gene transfer to the nucleus and the evolution of chloroplasts. *Nature*, **393**, 162-165.

Mar, A., Dworkin, J. and Oró, J. (1987) Non-enzymatic synthesis of the coenzymes, uridine diphosphate glucose and cytidine diphosphate choline, and other phosphorylated metabolic intermediates. *Origins Life Evol. Biosph.*, **17**, 307-319.

Maurel, M.-C. and Décout, J.-L. (1999) Origins of life: molecular foundations and new approaches. *Tetrahedron,* **55**, 3141-3182.

Maynard Smith, J. and Szathmáry, E. (1995) The major transitions in evolution. W.H. Freeman, Oxford.

McCord, J. (2000) The evolution of free radicals and oxidative stress. *Am. J. Med.*, **108**, 652-659.

McFadden, G.I. (1999) Endosymbiosis and evolution of the plant cell. *Curr. Opin. Plant Biol.*, **2**, 513-519.

Montgomery, W.L. and Pollak, P.E. (1988) *Epulopiscium fishelsoni* N. G., N. Sp., a protist of uncertain taxonomic affinities from the gut of an herbivorous reef fish. *J. Protozool.*, **35**, 565-569.

Moran, N. and Baumann, P. (2000) Bacterial endosymbionts in animals. *Curr. Opin. Microbiol.*, **3**, 270-275.

Moran, N.A. (1996) Accelerated evolution and Muller's Ratchet in endosymbiotic bacteria. *Proc. Natl. Acad. Sci. USA*, **93**, 2873-2878.

Moreira, D. and López-García, P. (1998) Symbiosis between methanogenic archaea and delta-proteobacteria as the origin of eukaryotes: the syntrophic hypothesis. *J. Mol. Evol.*, **47**, 517-530.

Myllykallio, H., Lopez, P., López-García, P., Heilig, R., Saurin, W., Zivanovic, Y., Philippe, H. and Forterre, P. (2000) Bacterial mode of replication with eukaryotic-like machinery in a hyperthermophilic archaeon. *Science*, 288, 2212-2215.

Nakielny, S. and Dreyfuss, G. (1999) Transport of proteins and RNAs in and out of the nucleus. *Cell*, **99**, 677-690.

Nelson, K.E., Levy, M. and Miller, S.L. (2000) Peptide nucleic acids rather than RNA may have been the first genetic molecule. *Proc. Natl. Acad. Sci. USA,* **97**, 3868-3871.

Noller, H.F. (1999) On the origin of the ribosome: coevolution of subdomains of tRNA and rRNA. In: The RNA World, 2nd Edn. Gesteland, R.F., Cech, T.R. and Atkins, J.F., eds. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York. pp. 351-380.

Olson, M.O.J., Dundr, M. and Szebeni A. (2000) The nucleolus: an old factory with unexpected capabilities. *Trends Cell Biol.*, **10**, 189-196.

Omer, A.D., Lowe, T.M., Russell, A.G., Ebhardt, H., Eddy, S.R. and Dennis, P.P. (2000) Homologs of small nucleolar RNAs in Archaea. *Science*, **288**, 517-522.

Pannucci, J.A., Haas, E.S., Hall, T.A., Harris, J.K. and Brown, J.W. (1999) RNase P RNAs from some Archaea are catalytically active. *Proc. Natl. Acad. Sci. USA,* **96**, 7803-7808.

Patterson, D.J. (1999) The diversity of eukaryotes. *Am. Nat.*, **154**, S96-S124.

Penny, D. and Poole, A. (1999) The nature of the last universal common ancestor. *Curr. Opin. Genet. Dev.*, **9**, 672-677.

Pereira S.L. and Reeve J.N. (1998) Histones and nucleosomes in Archaea and Eukarya: a comparative analysis. *Extremophiles,* **2**, 141-148.

Philippe, H., Germot, A. and Moreira, D. (2000a) The new phylogeny of eukaryotes. *Curr. Opin. Genet. Dev.* **10**, 596-601.

Philippe, H., Lopez, P., Brinkmann, H., Budin, K., Germot, A., Laurent, J., Moreira, D., Müller, M., Le Guyader, H. (2000b) Early branching or fast evolving eukaryotes? An answer based on slowly evolving positions. *Philos. Trans. R. Soc. Lond. B* **267,** 1213-1221.

Poole, A., Jeffares, D. and Penny, D. (1999) Early evolution: prokaryotes, the new kids on the block. *Bioessays*, **21**, 880-889.

Poole, A. and Penny, D. (2001) Does endosymbiosis explain the origin of the nucleus? *Nat. Cell Biol.* 3, E173.

Poole, A., Penny, D. and Sjöberg, B.-M. (2000) Methyl-RNA: an evolutionary bridge between RNA and DNA? *Chem. Biol.* **7**, R207-R216.

Poole, A.M., Jeffares, D.C. and Penny, D. (1998) The path from the RNA world. *J. Mol. Evol.*, **46**, 1-17.

Poole, A.M., Phillips, M.J. and Penny, D. (2001) Prokaryote and eukaryote evolvability. *Biosystems*, submitted.

Popenda, M., Biala, E., Milecki, J. and Adamiak, R.W. (1997). Solution structure of RNA duplexes containing alternating CG base pairs: NMR study of r(CGCGCG)2 and 2'-O-Me(CGCGCG)2 under low salt conditions. *Nucleic Acids Res.* **25**, 4589-4598.

Race, H.L., Herrmann, R.G. and Martin, W. (1999) Why have organelles retained genomes? *Trends Genet.* **15**, 364-370.

Rao, M., Eichberg, J. and Orό, J. (1982) Synthesis of phosphatidylcholine under possible primitive earth conditions. *J. Mol. Evol.*, **18**, 196-202.

Rao, M., Eichberg, J. and Orό, J. (1987) Synthesis of phosphatidylethanolamine under possible primitive earth conditions. *J. Mol. Evol.*, **25**, 1-6.

Razin, S., Yogev, D. and Naot, Y. (1998) Molecular biology and pathogenicity of mycoplasmas. *Microbiol. Mol. Biol. Rev.*, **62**, 1094-1156.

Reanney, D.C. (1974) On the origin of prokaryotes. *J. Theor. Biol.*, **48**, 243-251.

Ribeiro, S. and Golding, G.B. (1998) The mosaic nature of the eukaryotic nucleus. *Mol. Biol. Evol.*, **15**, 779-788.

Richter, C., Park, J.W. and Ames, B.N. (1988) Normal oxidative damage to mitochondrial and nuclear DNA is extensive. *Proc. Natl. Acad. Sci. USA*, **85**, 6465-6467.

Rivera, M.C., Jain, R., Moore, J.E. and Lake, J.A. (1998) Genomic evidence for two functionally distinct gene classes. *Proc. Natl. Acad. Sci. USA*, **95**, 6239-6244.

Rotte, C., Henze, K., Müller, M. and Martin, W. (2000) Origins of hydrogenosomes and mitochondria. *Curr. Opin. Microbiol.*, **3**, 481-486.

Rout, M.P., Aitchison, J.D., Suprapto, A., Hjertaas, K., Zhao, Y. and Chait, B.T. (2000) The yeast nuclear pore complex: composition, architecture, and transport mechanism. *J. Cell Biol.,* **148**, 635-651.

Scheuring, I. (2000) Avoiding Catch-22 of early evolution by stepwise increase in copying fidelity. *Selection* **1**, 135-145.

Schopf, J.W., Packer, B.M. (1987) Early Archean (3.3 billion to 3.5 billion year old) microfossils from Warrawoona Group, Australia. *Science,* **237**, 70-73.

Segré, D., Ben-Eli, D., Deamer, D.W., Lancet, D. (2001) The lipid world. *Origins Life Evol. Biosph.* **31**, 119-145.

Segré, D. and Lancet, D. (2000) Composing life. *EMBO Rep.*, **1**, 217-222.

Shapiro, R. (1999) Prebiotic cytosine synthesis: a critical analysis and implications for the origin of life. *Proc. Natl. Acad. Sci. USA,* **96**, 4396-4401.

Shulga, N., Mosammaparast, N., Wozniak, R. and Goldfarb, D.S. (2000) Yeast nucleoporins involved in passive nuclear envelope permeability. *J. Cell Biol.,* **149**, 1027-1038.

Sogin, M.L. (1997) History assignment: when was the mitochondrion founded? *Curr. Opin. Genet. Dev.*, **7**, 792-799.

Sogin, M.L., Silberman, J.D., Hinkle, G. and Morrison, H.G. (1996) Problems with molecular diversity in the eukarya. In: Evolution of Microbial life. Roberts, D.M., Sharp, P., Alderson, G. and Collins, M., eds. Cambridge University Press, Cambridge. pp. 167-184.

Soll, J., Bölter, B., Wagner, R. and Hinne, S.C. (2000) ...response: The chloroplast outer envelope: a molecular sieve? *Trends Plant Sci.,* **5**, 137-138.

Stoltzfus, A., (1999) On the possibility of constructive neutral evolution. *J. Mol. Evol.*, **49**, 169-181.

Szostak, J.W., Bartel, D.P. and Luisi, P.L. (2001) Synthesizing life. *Nature*, **409**, 387-390.

Vellai, T., Takács, K. and Vida, G. (1998) A new aspect to the origin and evolution of eukaryotes. *J. Mol. Evol.*, **46**, 499-507.

Venema, J. and Tollervey, D. (1999) Ribosome synthesis in *Saccharomyces cerevisiae*. *Annu. Rev. Genet.*, **33**, 261-311.

Veronese, A. and Luisi, P.L. (1998) An autocatalytic reaction leading to spontaneously assembled phosphatidyl nucleoside giant vesicles. *J. Am. Chem. Soc.*, **120**, 2662-2663.

Wächtershäuser, G. (1990) Evolution of the first metabolic cycles. *Proc. Natl. Acad. Sci. USA,* **87**, 200-204.

Wächtershäuser, G. (1992) Groundworks for an evolutionary biochemistry: the Iron-Sulphur World. *Prog. Biophys. Mol. Biol.*, **58,** 85-201.

Walsh, M.M. (1992) Microfossils and possible microfossils from the Early Archean Onverwacht Group, Barberton Mountain Land, South Africa. *Precambrian Res.*, **54**, 271-293.

Wente, S.R. (2000) Gatekeepers of the nucleus. *Science*, **288**, 1374-1377.

Westheimer, F.H. (1987) Why nature chose phosphates. *Science*, **235**, 1173-1178.

White, H.B. (1976) Coenzymes as fossils of an earlier metabolic state. *J. Mol. Evol.*, **7**, 101-104.

Williamson, J.R., Raghuraman, M.K. and Cech, T.R. (1989) Monovalent cation-induced structure of telomeric DNA: the G-quartet model. *Cell* **59**, 871-880.

Woese, C.R. and Fox, G.E. (1977) Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc. Natl. Acad. Sci. USA*, **74**, 5088-5090.

Wren, B.W. (2001) Microbial genome analysis: insights into virulence, host adaptation and evolution. *Nat. Rev. Genet.* **1**, 30-39.

Yarus, M. (1999) Boundaries for an RNA world. *Curr. Opin. Chem. Biol.* **3**, 260-267.

Table 1. Features that need to be explained, and issues that must be addressed in examining the origins of the eukaryote nucleus.

1. Chimeric nature of eukaryotic genome.
2. Absence of phagocytosis in bacteria and archaea
3. The structure of the nuclear envelope
4. The nuclear pore complex
5. Nuclear export and import processes.
6. Disappearance of the nuclear membrane, but not of other organellar membranes, during cell division in some eukaryotes.
7. The origin of meiosis/mitosis.
8. Eukaryotic linear chromosomes with multiple origins of replication and telomeres
9. Preservation of RNA world relics in eukaryotes, and reduction in prokaryotes.
10, Coupled transcription and translation in prokaryotes compared with mRNA splicing and processing in eukaryotes.
11. Any theory for the origins of the nucleus must also explain the absence of this structure in prokaryotes.

Paper 7

Poole AM, Phillips MJ & Penny D.
Prokaryote and eukaryote evolvability.
Biosystems (submitted).

Prokaryote and Eukaryote Evolvability.

Anthony M Poole*, Matthew J Phillips & David Penny

Institute of Molecular BioSciences

Massey University

Private Bag 11222

Palmerston North

New Zealand

*Corresponding author.

Email: a.m.poole@massey.ac.nz

Fax: +64 6 350 5688

Abbreviations:

ESND: Evolutionarily-Stable Niche-Discontinuity

PSF: Periodically-selected function

LUCA: Last Universal Common Ancestor

Abstract:

The concept of evolvability covers a broad spectrum of, often contradictory, ideas. At one end of the spectrum it is equivalent to the statement that evolution is possible, at the other end are untestable *post hoc* explanations, such as the suggestion that current evolutionary theory cannot explain the evolution of evolvability. We examine similarities and differences in eukaryote and prokaryote evolvability, and look for explanations that are compatible with a wide range of observations. Differences in genome organisation between eukaryotes and prokaryotes meets this criterion. The single origin of replication in prokaryote chromosomes (versus multiple origins in eukaryotes) accounts for many differences because the time to replicate a prokaryote genome limits its size (and the accumulation of junk DNA). Both prokaryotes and eukaryotes appear to switch from genetic stability to genetic change in response to stress. We examine a range of stress responses, and discuss how these impact on evolvability, particularly in unicellular organisms versus complex multicellular ones. Evolvability is also limited by environmental interactions (including competition) and we describe a model that places limits on potential evolvability. Examples are given of its application to predator competition and limits to lateral gene transfer. We suggest that unicellular organisms evolve largely through a process of metabolic change, resulting in biochemical diversity. Multicellular organisms evolve largely through morphological changes, not through extensive changes to cellular biochemistry.

Keywords:

evolvability, evolutionarily-stable niche-discontinuity, eukaryote evolution, genome evolution, prokaryote evolution.

Introduction.

Evolvability is a central concept in evolution but is easily misconstrued, hence its use must be defined carefully. At a basic level, evolvability is the fundamental concept of evolution. From the late 17th to mid-19th centuries it was generally assumed that species had an unchangeable 'essence'. This Platonic concept was introduced in the late 17th century when it became increasingly clear that continuing spontaneous generation of larger life forms did not occur (see Farley 1977). If species had an unchangeable essence then, by definition, there could be no evolution, even if individual organisms deviated from the 'ideal type'. 'Evolvability', by denying species have an unchangeable essence, is central to evolution. Since all evolutionists agree, this definition is not that interesting.

Burch and Chao (2000) offer a more limited definition, "the ability to generate adaptive mutations". We consider the two aspects of this definition: 'adaptive mutations' and 'ability to generate'. That adaptive mutations occur is the evolvability concept from the previous paragraph, but in modern terminology: some mutations are advantageous. In the early 19th century many accepted selection, but only in elimination of deleterious variants. Selection, by eliminating such variants, tended to preserve the unchanging essence of the species. In contrast, the existence of adaptive variants and positive selection allows evolution through time and is an essential part of evolvability.

The 'ability to generate' adaptive mutations is more problematic, and is mirrored in Kirschner & Gerhart's (1998) definition: 'the *capacity to generate* [our emphasis] heritable, selectable phenotypic variation'. If it is simply the observation that advantageous mutations occur, then, again, the usage is uncontroversial, though uninteresting. If it implies that advantageous mutations can be generated 'on demand' (e.g. Cairns et al. 1988) then it is a specialised (and controversial) usage. Some discussions on evolvability appear to give the impression of 'the more change the better' - yet most major change is highly deleterious. For instance, Radman et al. (1999) point out that selection for increased fidelity of DNA synthesis has been achieved in the lab (Fijalkowska et al. 1993), and that this demonstrates 'there was no durable selective pressure in nature for maximal fidelity'.

However, the majority of discussions on evolvability (e.g. Wagner & Altenberg 1996; Wagner 1996; Kirschner & Gerhart 1998; Partridge & Barton 2000), acknowledge directed mutation is not required to understand evolvability. Nevertheless, confusion arises easily, as shown by reactions to work from Lindquist's group (Rutherford & Lindquist 1998; True & Lindquist 2000). Other workers concluded (Dickinson and Seger 1999; Partridge & Barton 2000) that these authors favoured the idea that certain traits have been selected for their utility to contribute to organismal evolvability, *and nothing else.* While Lindquist points out that this was never her interpretation (Lindquist 2000), the subsequent correspondence generated by this work (Dickinson and Seger 1999; Partridge & Barton 2000; Dover 2000) illustrates how problematic this concept can be. There is no agreed definition for evolvability that explicitly avoids the problem of evolutionary forethought. Indeed, whenever the phrase 'the evolution of evolvability' is used, there is the possibility of it being misconstrued. This is not because evolvability cannot evolve through accepted processes of evolution. Rather, under known processes of Darwinian evolution, evolvability cannot evolve *in itself* because the origin and maintenance of a trait would have to precede selection for the trait.

Evolvability can be a by-product of selection however. For example, activation of a transposable element might lead to a mutation that is selected, thereby inadvertently leading to additional mutations (through additional element insertions) in the future. Such future mutations may be deleterious or advantageous; the increased mutation rate is a by-product of the transposable element hitchhiking with the selected mutation.

Still at issue is the evolutionary origin of traits that contribute to evolvability and adaptive mutations. Examination of the origins of such traits is an important step in alleviating controversy surrounding this area. This is particularly so with evolvability in multicellular organisms, where one gets the impression that we should be in awe of the exciting molecular and genetic mechanisms that contribute to eukaryote evolvability (Kirschner & Gerhart 1998; Herbert & Rich 1999). Other reviews on the evolution of evolvability (e.g. Partridge and Barton 2000; Kirschner and Gerhart 1998; Moxon and Thaler 1997) identify mechanisms by which genome architecture can influence this (see also Box 3). We focus here on the genome organisation of prokaryotes and eukaryotes, and interactions between these and the environment.

We review recent work on the evolutionary origins of the differing genome organisation prokaryotes (archaea and bacteria) and eukaryotes, and how this impacts on our understanding of the 'evolution of evolvability'. Our previous work, and that of others (particularly on parasites), suggests that many of the differences can be explained through constraints (or lack thereof) on genome size and architecture. Another significant area is stress responses. Experimental data, both with bacterial and metazoan models, point towards a general response to stress as being important in understanding how traits contributing to evolvability may have hitchhiked on survival of individuals. Horizontal gene transfer, stationary phase hypermutation, switching between sexual and asexual cycles, and the role of HSP 90 in *Drosophila* are considered.

Finally, we discuss how the physical and biotic environment limits potential evolvability, allowing a distinction to be drawn between this and realised evolvability (Fig. 2). Our model, which we call Evolutionarily-Stable Niche-Discontinuity (ESND), describes how competition allows colonisation of a fitness peak, and subsequently, how intraspecific competition limits movement away from that peak (Fig. 1). Examples of interspecies competition and predator-prey coevolution are considered, and are aimed at understanding evolvability in eukaryotes.

Assumptions versus hypotheses.

It is almost universally assumed that eukaryotes evolved from ancestral prokaryote forms, an assumption that seems intuitively correct. However, it is just that - an intuitive bias that simple evolves to complex - and is taken as given by a large majority of researchers (see Forterre & Philippe 1999 for critique). An extensive body of literature and ongoing research challenges this notion (Reanney 1974; Darnell & Doolittle 1986; Forterre 1995; Poole et al. 1998, 1999; Forterre & Philippe 1999; Penny & Poole 1999; Glansdorff 2000). What is important for evolvability studies is that the assumption of a prokaryote to eukaryote transition effectively removes selection from discussions on the evolution of prokaryotes - they are by definition the ancestral state. Since the direction of change is assumed to be from simple prokaryote cells to complex eukaryote cells, the question becomes, by default, what drove eukaryote genomes to become so complex? We will argue that factors affecting the origin of prokaryotic genome organisation are equally important.

There are strong parallels in the evolution of complexity and the evolution of evolvability. Neither complexity nor evolvability can be directly selected for; both impact future evolution, and hence are in violation of evolution as tinkering. Szathmáry and Maynard Smith (1995) point out that 'There is no theoretical reason to expect evolutionary lineages to increase in complexity with time, and no empirical evidence that they do so'. Unlike with evolvability however, there is little apparent controversy here. It is accepted that complexity is sometimes a consequence of evolution, but not a predictable outcome of evolution. Reductive evolution in parasites and eukaryotic organelles are important examples (see below).

How can we account for traits that contribute to complexity which are conserved in most eukaryotes when we know that, as with evolvability, complexity is not directly selectable? It is not sufficient to claim that a trait conserved across a broad range of species is evidence for selection. A recent example is that junk DNA has a function because a survey of genome size shows that it correlates with cell size in cryptomonads (Beaton & Cavalier-Smith 1999). The argument seems to be that selection for increased cell size has led to the expansion of junk regions because these take up space, and therefore the amount of DNA can 'specify' cell size. Correlation is ambiguous, and in this case it is unclear which is cause and which effect. Junk DNA may persist because it has not been selected against.

A theory that explains a range of phenomena (explanatory power) and leads to new tests (predictive power) is certainly preferable to *post hoc* explanations. These one-off explanations are proposed *after* a discovery has been made, hence *post hoc* – 'after the event'. When explaining to students the lack of scientific rigour in *post hoc* explanations, we use the story of Darryl (Box 1). The humour is incidental to the main point, that scientific statements are best made as predictions, not thought up after the event. An example is an old natural theology explanation of why the earth changes its tilt on its axis as it rotates around the sun. The change in tilt is *for* generating the seasons. A delightful *post hoc* explanation!

This criticism of *post hoc* explanations is similar to Gould and Lewontin's (1979) critique of the 'adaptionist program', that everything about an organism can be explained as aiding some aspect of its life cycle. *Post hoc* explanations may nevertheless be correct (equally, 'good' theories can be incorrect). The aim should be to reformulate them into testable hypotheses, and to look for explanations that account for a range of phenomena (not just the original observation that led to the hypothesis). Gibson (2000) points out that the tendency for researchers to give *post hoc* adaptationist explanations is still alive and well in developmental biology. He writes that, 'selection should only be invoked when the null hypothesis of neutrality cannot explain the data'. In molecular evolution the importance of neutral evolution is often taken into account, and extremely complex traits such as the spliceosome, mRNA editing in trypanosomes, and the scrambled genes of ciliates have been argued to be neutral (Stolzfus 1999). While it is not certain if any of these traits originated through neutral evolution, the idea is an important one, since it shifts theorising away from *post hoc* explanations, and frames the problems in the manner advocated by Gibson (2000).

Returning to the evolution of prokaryotes and eukaryotes, we shall argue that many complexities of the eukaryote genome can be explained by the null hypothesis of neutralism, while the prokaryote genome cannot. This is an important point, since it changes our view of the evolution of genomic features contributing to evolvability.

Origins of prokaryote and eukaryote genome architecture.

Key aspects of eukaryotic genome architecture appear to be conserved from a very early period in evolution, pre-dating the Last Universal Common Ancestor (LUCA). In contrast, prokaryote genome architecture results from one or more periods of reductive evolution (Poole et al. 1999; Penny & Poole 1999). Others (Forterre 1995, Forterre & Philippe 1999, Galtier et al. 1999) have developed similar views from different data. Our argument is based on extant genome architectures and the observation that the greatest diversity of RNA world relics (RNAs that appear to predate the origins of proteins and DNA) are found in eukaryotes. For prokaryotes, both the loss of ancient RNA genes and their genome architecture can be explained in terms of reductive evolution.

Some of our reasoning is given below, but it is not necessary to accept all our conclusions to accept our general argument on eukaryote and prokaryote evolvability. Our conclusions are consistent with Kirschner and Gerhart's (1998) description of prokaryote and eukaryote modes of evolvability. Prokaryotes, 'have undergone limited morphological change but instead have achieved extensive biochemical diversification'. Similarly, multicellularity in eukaryotes, specifically metazoa, 'achieved extensive control over the milieu of internal cells and generated many physiologically sensitive micro-environments in that milieu'. In this latter multicellular group, biochemical evolution is limited, and cells receive a more constant level of nutrition with little or no variation in the *type* of nutrients available. If evolution is biochemically conservative in metazoa and biochemically innovative in prokaryotes, it is perhaps no surprise to find ancient biochemical traits conserved in eukaryotic cells, while these have been lost from prokaryotes.

Broad differences between eukaryote and prokaryote lifestyle have been described in terms of r and K selection (Carlile 1982), terms derived from the equation for the rate of population growth (Box 2). Relative to prokaryotes, eukaryotes are K-selected, where K-selected organisms are broadly defined as having a relatively slow rate of reproduction and longer generation time, a stable (though limiting) nutrient supply, relatively stable populations, and are larger in size. In contrast, prokaryotes are relatively more r-selected, with faster reproduction and short generation times, small size, fast response times to a fluctuating nutrient supply, and with large fluctuations in population size. There is a spectrum of values with perhaps *E. coli* and yeast near the r-selection end, and elephants and oak trees near the K-selection end of the spectrum.

Prokaryote genomes.

Prokaryote genomes possess only one origin of replication per chromosome. Consequently, size places limits on the rate of chromosome replication. As the fidelity is affected by replication rate, so rate will be constrained by the need to faithfully copy and maintain the genome.

Transient global hypermutation occurs in stationary phase (Table 1), whereas selection for fast replication operates during periods of exponential growth. There is no precedent for assuming that higher mutation rates will be selected for during exponential growth where proliferation of a successful strategy is required. Rather, a quick response to nutrient availability, followed by clonal proliferation, is advantageous. r-selection revolves around competition (during exponential growth) for resources that fluctuate in availability, and this places the reproductive rate under selection (Box 3).

That replication is rate-limiting during exponential growth has been documented for *E. coli*, where genome doubling takes one hour, and cell doubling occurs every 20 minutes (Alberts et al. 1994). The effect on the genome is straightforward - anything that can be lost will eventually be lost. Selection does not distinguish between junk, and what may be advantageous later (e.g. on a new nutrient source), so even essential functions required only periodically may be lost from the genome. It is therefore of little surprise to find that, in both *E. coli* and *Salmonella enterica*, genome size varies within species by around 20%. Similar variability is found in *Helicobacter pylori* and *Neisseria meningitidis*, and is interpreted as different genes being maintained in different isolates, which often inhabit different niches (Lan & Reeves 2000).

Periodically-selected functions (PSFs) are regularly lost from individuals, but are maintained in bacterial populations through lateral gene transfer. PSFs are essential in the long term, given that environmental fluctuation is normal and that organisms must continually cope with such fluctuations. In a completely clonal population where replication time is rate limiting, PSFs would be irreversibly lost. Constant selection of PSFs within a population, coupled with lateral transfer is likely central to prokaryote genome architecture, permitting maintenance of PSFs crucial to long term survival under conditions where these are frequently lost.

Plasmids are a complete transferable unit that can be immediately expressed, but do not increase the replication time of the genome. While a genomic copy of a PSF must be lost through gene decay (mutations and deletions) and reestablishment requires reinsertion, a plasmid can be lost without gene decay (this would be advantageous during exponential growth), is readily reacquired, can be replicated in parallel with the genome. Supernumerary chromosomes in fungi have been likened to plasmids, as they are not permanent and in several cases have been found to carry genes for pathogenicity, detoxification of host antimicrobials, and antibiotic resistance (Covert 1998).

An obvious solution to the prokaryote dilemma is to distribute genes across several chromosomes and with multiple origins of replication, thereby permitting a larger genome without slowing replication. A number of prokaryote genomes are spread across multiple chromosomes, and some may possess more genes than yeast (Bendich & Drlica 2000). Circular chromosomes with single origins of replication nevertheless place limits on individual chromosome size.

That circular chromosomes are only found in prokaryotes may be historical accident. Forterre (1995) has argued that the prokaryote lineages arose through adaptation to high temperatures (the thermoreduction hypothesis). Currently his is the best explanation for the presence of circular chromosomes in prokaryotes; circular DNA is more thermostable than linear (Marguet & Forterre 1994). Other data are also consistent with thermodreduction (Poole et al. 1999, Penny & Poole 1999), and while some prokaryotes possess linear genomes (Bendich & Drlica 2000), this state appears derived (Poole et al. 1998, 1999).

Eukaryotes.

K-selected organisms have a steadier rate of reproduction, with relatively smaller population fluctuation, particularly in multicellular eukaryotes. Eukaryote chromosomes possess multiple origins of replication, and accumulation of repetitive elements largely accounts for the 80,000-fold genome size variation in this domain (Hartl 2000). In many cases, increases in size are probably not a result of selection (Hartl 2000), and consequently, some eukaryote genome sizes are probably only limited by the fidelity of replication (see Table 4 in Drake 1999).

With few apparent constraints on genome size, gene duplication followed by divergence is an effective means for the evolution of new functions. Neither duplication, nor the presence of pseudogenes, is inherently deleterious in eukaryotes (in contrast to prokaryotes). Gene duplication and divergence has resulted in major expansions of developmental gene families, e.g. the homeobox family (Ruddle et al. 1999). Genome duplication is also considered a feature of eukaryote genome evolution (Wolfe et al. 1997; Ruddle et al. 1999), a good example being polyploidy in plants.

Lack of constraint on genome size has enabled large numbers of 'selfish' elements to co-exist in eukaryotic genomes (Smit 1999; Brosius 1999). Such elements can occasionally be recruited into the cellular repertoire. Examples include dendrite-specific RNAs, rodent BC1 and primate BC200. BC1 has been recruited from $tRNA^{Ala}$ and BC200 from an Alu element (Brosius 1999). V(D)J recombination in the vertebrate immune system is another example. Proteins RAG1 & RAG2 mediate V(D)J recombination, forming a site-specific recombinase which recognises and cleaves DNA at conserved recombination signal sequences (Agrawal et al. 1998; Hiom et al. 1998). Similarities in gene organisation, signal sequences, mechanism of action, and the presence of a transposase DDE motif in RAG1 (Landree et al. 1999) suggests this system originated through a germline transposition event into a receptor gene in the ancestor of jawed vertebrates (Agrawal et al. 1998; Plasterk 1998). An unforeseen consequence of the recruitment that gave rise to V(D)J joining is that it also appears to participate in at least some chromosomal translocation events, though probably at low frequency (Melek & Gellert 2000).

Aspects of placental development in eutherian mammals appear similar to viral infection (Larsson and Andersson 1998; Harris 1998). Cell fusion, forming the placental syncytium, is also a feature of endogenous retroviruses (providing an efficient means of infecting new cells). In human placental development, an envelope protein from the endogenous retrovirus ERV-3 is responsible for cell fusion and other differentiation events during formation of the syncytium (Lin et al. 1999). Production of endogenous retroviral particles early in placental development increases the chance of germline insertion, but also provides immunosuppression, thereby preventing the maternal immune system from rejecting the foetus. Indeed, retroviral envelope protein expression suppresses the immune response (Mangeney & Heidmann 1998).

These examples highlight the centrality of the tinkering concept in evolution (Jacob 1977). In all cases, the evolution of complex structures appears to have arisen from selfish elements. Occasional recruitment of such elements into new function appears a consequence of the lack of selection against genome size, making the genomes of higher eukaryotes more vulnerable to intragenomic parasites. Overall, the neutrality of non-coding sequences in chromosomes with multiple centres of replication explains many aspects of eukaryote evolvability.

Transcript processing.

Extensive transcript processing is a feature of eukaryotes, and includes mRNA splicing (Sharp 1994), editing (Smith et al. 1997), and snoRNA-mediated cleavage, methylation and pseudouridylation of RNA (Weinstein & Steitz 1999). Splicing and editing are absent from prokaryotes, and snoRNA-mediated modifications are absent in bacteria (though methylation is present in archaea). Though disputed (Lafontaine & Tollervey 1998; Sontheimer et al. 1999), splicing and snoRNA-mediated modifications probably predate the LUCA (Poole et al. 1998, 1999).

Under r-selection and a single origin of replication, spliceosomal introns and snoRNA-mediated modifications are expected to be reduced or lost. mRNA processing delays the expression of proteins, the transcript being processed largely by RNA-mediated reactions. Methylation and pseudouridylation of RNA is ubiquitous, though heavily reduced in bacteria. In archaea, methylation is extensive, and requires snoRNA-like sRNAs (Omer et al. 2000), smaller than in eukaryotes. Each sRNA guides two methylations (the majority of eukaryotic snoRNAs guides just one). Pseudouridylation is minimal in archaea, with numbers comparable to those for bacteria (Charette & Gray 2000). Modifications in bacteria are limited to highly conserved regions of the rRNA, which may explain their maintenance, while methylation may be important in archaeal rRNA for stability at high temperature (Omer et al. 2000).

In scenarios of the evolution of snoRNAs post-LUCA, the argument has largely been *post hoc*, with the emphasis being on how these RNAs could have diversified in eukaryotes (Morrissey & Tollervey 1995; Lafontaine & Tollervey 1998). The finding of sRNAs in archaea requires a revision of that theory. The alternative, loss under r-selection in prokaryotes, is the best explanation for the current data.

Some snoRNAs are paternally imprinted in rodent and human brain, and do not direct methylation of rRNA or other functional RNAs (Cavaillé et al. 2000). One of these may regulate A-to-I editing and/or alternative splicing of the serotonin 5-HT$_{2C}$ receptor mRNA through methylation (Cavaillé et al. 2000; Filipowicz 2000). Indeed, splicing and A-to-I editing, perhaps also modification by methylation and pseudouridylation, are central to the generation of multiple products from one mRNA (Herbert & Rich 1999). It is unclear how A-to-I editing of nuclear mRNAs arose in evolution, but the targets have largely been found in signalling in the nervous system of both invertebrates and vertebrates (Reenan 2001). The role of splicing in generating alternative protein products, and in regulating developmental fate (Graveley 2001), is possibly a consequence of its maintenance in the absence of selection to remove this apparatus long after its hypothesised role in early genomes would have become redundant. RNA processing pathways can be co-opted and contribute to evolvability, but clearly had other origins.

Cytosine methylation, a double-edged sword.

Cytosine methylation is widespread in eukaryotes, and is considered to provide a mechanism for gene silencing, and parental imprinting. Cytosine is an unstable base, readily deaminating to uracil, which, if unrepaired will result in a C•G to T•A mutation in one of two daughter copies. Methylation of cytosine produces 5-methylcytosine (5-meC) which deaminates more rapidly than unmethylated cytosine, yielding thymine (Poole et al. 2001). Cytosine methylation, while apparently providing a means of epigenetic control, also produces mutational hotspots, and this can potentially be beneficial or deleterious, depending on context.

Gene silencing has been considered to represent the main function of cytosine methylation, but Yoder et al. (1997) point out that evidence is limited. The majority of 5-meC residues are found in transposable elements, not promoters. They suggest that methylation is primarily a mechanism for silencing transposons, with the corollary that 5-meC to T deamination is largely beneficial because it results in faster inactivation of these elements through mutation. That this cannot be the only function of cytosine methylation is supported by the existence of at least two repair mechanisms (Schärer &Jiricny 2001; Poole et al. 2001). If both gene regulation and transposon inactivation are mediated by cytosine methylation, there is a trade-off because in the former 5-meC to T deaminations are potentially deleterious, whereas in the latter they are potentially beneficial. The presence of deamination repair mechanisms would therefore be important for repairing damaged genes, but weaken the potential for transposon inactivation (Poole et al. 2001).

The picture is further complicated, because methylation of transposable elements may contribute to epigenetic effects on adjacent genes (Whitelaw & Martin 2001). Patterns of methylation are known to be inherited, and to have a phenotypic effect. An example is agouti locus in mice, where coat colour is inherited epigenetically through the female line in the absence of genetic variation (Morgan et al. 1999). Whitelaw & Martin (2001) coined the term epigenotype for the effect that epigenetic inheritance has on phenotype, and excitingly, this may provide a means of exploring phenotypic space. However, work on agouti demonstrated that, even with selection for a given epigenotype, the original proportions of epigenotypes may reappear (Morgan et al. 1999, Whitelaw & Martin 2001), making it hard to see how parental imprinting mechanisms could lead to genetic fixation of a phenotypic trait. However, Monk (1995) has proposed that 5-meC deamination may contribute to fixation, since this would make permanent the silencing effect at a given site. In this way, the epigenotype could permit exploration of alternative phenotypes that could then become 'hard-wired' in the genome.

Again, this mechanism impacts on evolvability, but did not evolve for evolvability's sake. Prerequisites for such complex regulation may instead have been the invasion of eukaryote genomes by transposable elements, and selection to silence these, given the apparent inability to prevent their insertion. The conflicting need to eliminate these and the recruitment of methylation into gene regulation, perhaps through adjacent transposons may have set up the requirement to repair 5-meC to T deaminations. Imperfect repair of these (Holliday & Grigg 1993) may be the cost associated with the conflicting roles of methylation in the genome. However, it may provide a mechanism where 5-meC to T deamination gives rise to a heritable phenotypic trait from an epigenetic trait with limited heritability. Again, it is difficult to establish which came first, transposon inactivation or gene regulation, but the example serves to make the point that it is necessary to examine the origins of a process when considering the evolution of evolvability.

Another example is somatic hypermutation at the V(D)J locus in formation of the antibody variable region by C to U editing (Muramatsu et al. 2000; Revy et al. 2000). This is effectively enzyme-catalysed cytosine deamination at hotspots (contingency loci). The function is opposite to the uracil-DNA glycosylases, which are involved in repair of cytosine deaminations (Schärer &Jiricny 2001), and is also seen in apolipoprotein B transcript editing (Herbert & Rich 1999).

Parasites: evolvability or reductive evolution?

Parasites are interesting in regard to evolvability because they represent a strategy common to eukaryotes, prokaryotes, viruses, and selfish elements. Parasites are often fast-evolving, and have often moved from a non-parasitic to a parasitic lifestyle. We consider the following questions:

- Were ancestral groups from which parasites arose inherently more ‘evolvable’?
- If there is fast evolution in parasites, are they inherently more evolvable?
- Is the concept of evolvability useful here?

Parasitism is widespread - in plants, fungi, insects, worms, protists, bacteria, etc. Conspicuously absent are parasitic mammals, birds, amphibians and reptiles (tetrapods). Is this due to limited evolvability or an ecological limitation? We think the latter. The dependence of the eutherian embryo on the mother for nutrients is much like the dependence of endoparasitic larvae on the host for nutrition (Grbic 2000). Suckling in mammals, and nutritional dependence of juvenile birds and mammals on parents might to a lesser extent be seen in this light. Indeed, juvenile parasitic stages in early development serve the same role as parentally-supplied nutrition, and it is worth noting that egg-yolk mass has become reduced in endoparasitic wasps (Grbic 2000). Clearly, this *modus operandi* of early development has been made use of in mammals, and absence of true parasitism in tetrapods may simply reflect an absence of niches, though a few examples, such as brood parasitism exist.

What distinguishes lineages that have become obligate parasites from those that are free-living? The discussion above suggests it is the presence of an available niche, not limits on evolvability. However, there must be adaptation in order to secure nutrients from the host, fine-tune development to coincide with host life cycle, and not kill the host before the parasite has matured or moved to the next host. Studies of unicellular parasite genomes suggest that the loss of traits no longer required in the parasitic lifestyle accounts for most change. For example, in the *Rickettsiae*, adenosylmethionine synthetase is in the process of being lost from the genomes of this genus (Andersson & Andersson 1999). Likewise, cases of loss from parasitic genomes of primary biosynthetic pathways, such as amino acid synthesis and *de novo* pathways for deoxyribonucleotide synthesis (Fraser et al. 1998; Andersson et al. 1998b) are consequent to the evolution of mechanisms for extracting these nutrients from the host.

Genome reduction and higher rates of evolution appear to be general features of parasitic genomes, being reported in leprosy bacillus (Cole et al. 2001), the obligate intracellular parasites *Chlamydia* (Kalman et al. 1999) and *Buchnera*, and other endosymbiont bacteria (Moran & Baumann 2000). Genome reduction is likely a consequence of redundancy, while the higher rates of evolution seen in parasites are attributable to Muller's Ratchet, the fixation of slightly deleterious mutants within small asexual populations (Moran 1996).

Genome reduction is extreme in chloroplasts (McFadden 1999), mitochondria (Gray et al. 1999), and nucleomorphs, the remains of nuclei in secondary endosymbionts (Douglas et al. 2001). There is a difficulty in separating selection for evolvability *per se* from other potential selective pressures. Reductive evolution, and increased rates of evolution are *consequences* of parasitism or endosymbiosis, so while the process of adaptation can be extensively studied, the initial conditions cannot. What can be said is that to the parasite or endosymbiont, the host is a resource, so general models of evolvability are likely to be useful in understanding parasitism. In the following section, we consider this problem in greater depth.

The stress response and evolvability.

In this section, we consider how stress responses promote organismal survival. Hypermutation (adaptive evolution), horizontal transfer, sex in organisms with an asexual cycle, recombination, cell-cell interactions, and cell specialisation can all be understood as stress adaptations (Table 1). That they contribute to evolvability in prokaryotes and unicellular organisms is consequential - these traits have not been selected for their propensity to promote evolvability. and the evolutionary origins of these phenomena need not be in the adaptation to stress. Rather, what is important is that they currently contribute to adaptation to stress in a range of organisms, and that this has an impact on evolvability.

We suggest that these mechanisms are important for understanding periods of genetic stability versus genetic change within the lifecycle of a range of organisms. Respectively, these might be described as 'if it ain't broke, don't fix it' and 'adapt or die' strategies. Switching between strategies is expected to be more effective in prokaryotes and unicellular eukaryotes than in multicellular eukaryotes since, as described below, mechanisms for alleviating lethal stresses exist in the first two groups, but not the third.

A range of starvation responses, which can be described as 'adapt or die' strategies, are seen in prokaryotes (Table 1). In *Bacillus subtilis*, sporulation and genetic competence (to take up DNA from the external milieu) are both controlled by an extracellular peptide, CSF (competence and sporulation factor). At low concentrations, CSF stimulates competence, and this occurs 2-3 generations prior to entry into stationary phase. At high concentrations, which arise shortly after entry into stationary phase, CSF inhibits competence, and stimulates sporulation (Lazazzera et al. 1999). Importantly, the SOS response and competence are coinduced and DNA uptake may provide a template for repair of endogenous DNA (Tortosa & Dubnau 1999). Alternatively, formation of double-strand breakages may permit integration of foreign DNA concurrent with uptake. Perhaps favouring the first possibility is the observation that these 'quorum sensing' mechanisms are often strain-specific, which may favour uptake from closely-related strains.

Concurrent with competence (and controlled by the same pathway), degradative enzymes are expressed and these may act to increase the availability of extracellular nutrients (Tortosa & Dubnau 1999). The same situation is seen in sexual sporulation in the fungus *Aspergillus nidulans*, where the α-(1,3)-glucan, which makes up the vegetative hyphal wall, is degraded to glucose (Champe et al. 1994).

A parallel to meiosis and sexual sporulation in fungi is evident here. Meiosis and competence precede sporulation, and DNA uptake in some bacteria may be most favoured between closely related strains, thereby approximating sex. The response to starvation is to change from a mode of development where genetic change is minimised, to one where there is active change, before dispersal to a new environment.

In *Aspergillus*, hyphae are sent out into the medium in a radial pattern away from the centre of the colony. Closer to the centre, asexual spores develop, which allow dispersal to new nutrient sources. This strategy is analogous to exponential growth in bacteria. Sexual sporulation occurs later in the lifecycle of the fungus; sexual spores are formed, at the centre of the original colony, where nutrients will have been most exhausted (Champe et al. 1994), and this is equivalent to the stationary phase events of genetic competence and sporulation in *Bacillus.*

DNA uptake by prokaryotes is apparently not always an approximation of eukaryote sex. Distant transfers between archaea and bacteria have been documented (Nelson et al. 1999; Forterre et al. 2000), and both *Neisseria* and *Haemophilus* are apparently competent all the time (Solomon & Grossman 1996), each containing well over a thousand copies of a DNA uptake signal sequence (Smith et al. 1999).

It seems unlikely that horizontal transfer is unbridled and without patterns, despite the vigour with which many in the phylogenetics community have taken on this idea as a *post hoc* explanation for current difficulties in explaining conflicting datasets (Woese 1998; Doolittle 1998). DNA loss, due to constraints on replication rate during exponential growth, suggests that any sequences taken up will only be fixed if they confer a selective advantage to the organism. Greater promiscuity permits greater sampling of environmental DNA, potentially bestowing a greater propensity to adapt to environmental change (greater evolvability). Greater promiscuity may also equate to greater parasite susceptibility, which might explain the existence of strain-specific competence factors.

There is now overwhelming evidence for transient hypermutation, induced by the SOS response to starvation (Torkelson et al. 1997; Foster 1999; McKenzie et al. 2000). Metzgar and Wills (2000) argue that it may simply be a spandrel, that is, a by-product, not a directly-selected adaptation. The DNA polymerases involved in the response have been selected to copy highly damaged DNA, which constitutive polymerases (with higher replication fidelity) are unable to copy. The lower-fidelity polymerases repair damaged DNA, but the lower specificity of polymerisation required to bypass lesions also results in a transient increase in mutation rate.

In the lab, global mutators have been successfully selected for, and tend to outcompete nonmutators (Sniegowski et al. 1997). Mutators can arise by chance, and, it has been argued that they could be maintained in asexual populations through genetic hitch-hiking on an advantageous allele created as a result of mutation. While it is thought that complete fixation of mutators would be rare, there seems to be a correlation between elevated mutation rate and virulence in pathogens (see Metzgar & Wills 2000 for discussion). Perhaps this is not surprising, given that their hosts make use of somatic hypermutation in antibody formation, setting up a Red Queen race. However, the side effects for bacterial mutators are potentially worse; mutational meltdown due to the accumulation of deleterious mutations.

Horizontal transfer and genome copy number may be crucial in the maintenance of elevated global mutation rates resulting from the appearance of heritable global mutators. Tenaillon et al. (2000) point out that horizontal transfer provides a potential mechanism for the spread of selectively advantageous mutations (such as those rare beneficial mutations arising during hypermutation) within a population. This might result in the advantageous allele being selected for while the mutator is selected against (due to an increase in deleterious mutations) and thus lost. The ability to segregate the beneficial mutation from the mutator phenotype may serve to provide a mechanism for the elimination of mutator alleles from a population in the long term.

Prokaryotes with multiple copies of the genome are widespread (Bendich & Drlica 2000), perhaps even the rule. For instance, *E. coli* is polyploid throughout its cell cycle (Åkerlund et al. 1995). Multiple genomic copies will serve as a buffer to deleterious mutation, minimising the detrimental effects of hypermutation, and at the same time, permitting new alleles to arise and be selected for (Koch 1984). *Azotobacter vinlandii* maintains over 100 genomic copies in stationary phase (Maldonado et al. 1994), making it a potentially very interesting model organism for mutation studies.

Another mechanism contributing to adaptive evolution is transient gene amplification of the *lac* operons of *Salmonella typhimurium* (Andersson et al. 1998a) and *E. coli* (Hastings et al. 2000). Multiple copies of a mutant locus with residual activity produces an unstable 'wild type' revertant. At the same time, presence of multiple copies increases the likelihood of a true reversion event. This last point is important, since, in effect, multiple copies provide mutation with a bigger 'target' without deleterious changes being lethal. This mechanism (Andersson et al. 1998a) may be important in rescuing periodically-selected functions (PSFs) from loss during selection to reduce genome size. While Hastings et al. (2000) did not find such revertants in their studies on *E. coli*, this does not necessarily imply that this cannot occur.

An additional link between stress response and evolvability is reported in *Drosophila.* Rutherford and Lindquist (1998) mutated the *hsp83* locus (encoding HSP90), finding mutations of unrelated morphological traits in heterozygotes. The morphological mutations are stable even after subsequent crosses restore progeny to wild type. They argue that such a situation might arise in nature due to titration of HSP 90 during heat shock, or other stresses where heat shock proteins are expressed.

In contrast to the previous examples where change is immediate, the stress, and the release from HSP 90 buffering, would presumably have to be sustained across generations for an alternate phenotype to be expressed and for selection to act upon this. Developmental processes (formation of adult structures, for instance) must run before phenotype is expressed. The comparison highlights the difference in the nature of adaptation between unicellular and multicellular organisms. A relaxation of buffering in response to stress could promote survival through expression of new variants, but the stress must be sustained and non-lethal. A lethal stress such as application of an antibiotic can however be dealt with in unicellular organisms, where beneficial mutations or genes received through horizontal transfer confer instant alleviation of the stress.

A parallel system exists in yeast, where, under conditions of heat shock, the PSI protein, which has a role in translation termination, undergoes a conformational change, becoming a prion (True & Lindquist 2000). This conformational switch impairs translation termination, and there is extensive readthrough, producing alternative protein products. Reversion to the non-prion form is possible, and the process can result in heritable changes. As Metzgar and Wills (2000) point out, it is not possible to establish whether these examples are best described as spandrels, or whether there was selection for the buffering of variability in the absence of stress, and release from buffering during stress. The latter scenario is not incompatible with current evolutionary theory, as demonstrated by the above discussion of stress response in unicellular prokaryotes and eukaryotes, but given Rutherford & Lindquist's (1998) titration model, we favour the first possibility.

In Table 1 sporulation, and cell-cell interaction are also listed as environmentally regulated and promoting survival during stress. Sporulation or cell-cell aggregation to form fruiting bodies, biofilms and other transient multicellular structures in response to environmental stress is not controversial. The difference between these, and the more controversial mechanisms is that the controversial mechanisms require mutation. If such responses can be selected for under lethal conditions, such as starvation, then so can the latter. However, that transient hypermutation and horizontal transfer are selected is best explained as occurring through hitch-hiking, not direct selection. The twist is that the fixation and subsequent maintenance of adaptive evolutionary traits through hitch-hiking may be on different loci at each round of selection.

To conclude this section, while the evolutionary origins of many of the stress responses in Table 1 are still obscure, it is nevertheless possible to identify selection pressures which result in their maintenance and heritability. These are all 'adapt or die' strategies with a short term survival advantage, consistent with standard evolution. As pointed out by Metzgar & Wills (2000) and Hastings et al. (2000) there is no requirement for evolutionary forethought. If the ultimate consequence of starvation (or other environmental stresses) is death, then individuals in which elevated mutation rates, genetic competence or locus specific amplification are induced may survive. There are therefore two aspects: the ability to induce the mechanism to generate variability, and advent of a new function which may alleviate the stress.

An ecological perspective: Evolutionarily-stable niche-discontinuity (ESND).

Between groups of (complex multicellular) taxa, there often appear to be long-term stable niche boundaries. In a fitness landscape these boundaries limit access to a single peak, or sub-set of peaks, and thus limit evolutionary potential. For example, the vertebrate flying insectivore niche has been occupied by birds at day and bats at night for over 55 million years

(Novacek, 1985) with little crossover between nocturnal and diurnal niches. Dinosaurs and mammals may have provided niche boundaries for each other for over 150 million years until many of the great Mesozoic reptiles became extinct around the Cretaceous-Tertiary boundary (Bromham *et al.*, 1999; Sereno, 1999).

Typically, niche restrictions are explained as dominance resulting from specialisation of the incumbent (Rosenzweig and McCord, 1991). This is basically inter-specific competition, with the species occupying the niche having had time for many optimisations compared with a potential competitor. We introduce the concept of evolutionarily-stable niche-discontinuity (ESND) to explain the maintenance of niche boundaries; in addition to interspecies competition, it attributes a major role to intraspecific competition within the competitor. A shift in an individual competitor (toward an alternative niche) typically involves a deleterious trade-off between interspecific and intraspecific competition. That is, a small heritable shift away from the fitness peak of the competitor's own gene pool will result in a greater fitness reduction (due to intraspecific competition) than the fitness increase from increased resources via interspecific competition.

Figure 1 depicts a possible ESND for two taxa (1 & 2) that specialise on different food resources, with each taxon located near its own peak of fitness. The black and grey curves show the relative fitness derived from resources A and B respectively. The contributions sum (dashed line) to give the relative fitness for a hypothetical character. Models of resource partitioning among mammals (Phillips, in prep.) suggest that an ESND between two taxa can be maintained where potentially competing taxa specialise respectively on either side of an environmental discontinuity that may be physical (night vs. day) or biological (e.g. different prey species).

Niche partitioning among large cursorial carnivores illustrates ESND maintained by specialisation in several characters, and coevolution with resources. Throughout Eurasia, Africa and America, the cat and dog groups of carnivores fill niches for fast-burst and endurance predators respectively. As predators, cats and dogs have many differences (Jones and Stoddart, 1998). As fast-burst, first-strike predators, cats have a high proportion of fast twitch glycolytic muscle (like olympic sprinters), powerful jaws and crushing canines, as well as having forelimbs as part of the killing mechanism. Conversely, as well as behavioural differences, large dogs, as endurance predators, have a low proportion of fast twitch glycolytic muscle (like marathon runners), slashing jaws and canines, and forelimbs (specialised for long-distance running) are not included in the killing mechanism.

Thus multiple specialisations reinforce the ESND. Consider an individual dog (or cat) with a heritable shift in one of these characters towards the optimum phenotype of the other, but without concurrent shifts in the others. This change will reduce fitness in its own niche, but will still be of little benefit in accessing the other niche. Additionally, coevolution between predators and prey can strengthen the ESND. A dog with a slightly higher ratio of glycolytic to oxidative muscle is unlikely to benefit as a fast-burst predator because potential prey has coevolved with the faster burst-predators (cats). Yet other dogs will leave this mutant dog behind before they reach their endurance limit – intraspecific competition is strong. A consequence of ESND development for coevolution with prey resources is that evolvability may be more affected by ESNDs among taxa that prey on live organisms, than taxa that are autotrophs or detritavores.

Given the prevalence in nature of physical and biological discontinuities, in the absence of extrinsic extinction and immigration of foreign (non-coevolved) competitors ESNDs should develop between coevolved taxa that compete for resources. As such, it is not surprising that catastrophic physical events have so often been suggested to catalyse evolvability (Jablonski, 1986; Roy 1996). Although such events may not directly affect molecular and developmental mechanisms, they free lineages from ESND-restricted evolutionary trajectories.

The establishment of ESNDs may differ between eukaryotes and prokaryotes in that horizontal transfer may break such barriers down in prokaryotes. For example, pathogenic Shigella strains of *E. coli* appear to have multiple independent origins within *E. coli*, probably concurrent with receipt of a plasmid carrying pathogenesis genes, and subsequent convergent gene losses (Pupo et al. 2000). Operons in both prokaryotes (Lawrence 1999) and fungi

(Walton 2000) are also interesting in this regard, since, like plasmids, they represent a distinct, potentially transferable unit, such as an entire biosynthetic pathway, complete with regulatory sequences.

Horizontal transfer of genes that allow an organism to compete in a new niche may have a number of outcomes. 1, the incumbent is better adapted and the invader cannot colonise the niche. 2, the invader is better adapted (will depend on genetic background of the trait under selection in the niche). 3, both have similar fitness, which may result in further competition, extinction of one or the other, or specialisation leading to two new niches. In the context of evolvability it is not sufficient just to consider interspecific competition between a potential invader and the incumbent species. Evolvability depends also on intraspecific competition within the invader, and coevolution between different levels of the food chain.

Functional interactions between organisms and their environment necessarily invoke evolutionary constraints. Flowers which interact with pollinators are subject to greater evolutionary constraints than are parts such as leaves and bark, which are not required to interact specifically with other organisms (Raven et al., 1986). Evolutionary stability conferred on plant reproductive structures has made them more useful than (for example) bark or leaves in determining phylogenetic relationships.

Evolutionary constraint can also result when environmental interactions change during development. Many amphibian and reptile taxa experience dramatic shifts in their environment through development, essentially having to function in different niches. For instance, the komodo dragon (*Varanus komodoensis*) begins life as an arboreal predator of small insects, progressively moves onto larger insects, small vertebrates and eggs, then larger vertebrates and eventually fills a terrestrial large predator/scavenger niche. Mutations providing a potential fitness advantage at any point along this continuum may be deleterious somewhere else during growth. This effect is less in mammals and birds because they typically feed their young until they can occupy the adult niche.

Compared with other vertebrates, mammals and birds are also notable for an increased emphasis on homeostasis, particularly endothermy (Ruben, 1995), so stabilising internal biochemical and physiological conditions. Both effects, reducing the range of niches during development and stabilising internal conditions, should enhance morphological evolvability. Indeed, while mammals and birds have diversified into widely different niches and morphologies from their ancestors that shared the planet with dinosaurs 65 million years ago, amphibians, turtles, lepidosaurs (snakes and lizards) and crocodilians typically have not (Benton, 1993).

Plasticity, Learning and Evolvability

Population genetics typically considers just the genetic contribution to the phenotype on the grounds that the genetic component is selectable. Phenotypic plasticity, such as the specific branching pattern of a tree that has grown into a gap of light in the forest, is not genetically determined - yet has an important bearing on evolvability. One suggestion, often called the Baldwin effect (Baldwin (1896), though also proposed by others), is that useful non-genetically acquired phenotypes will eventually tend to be determined genetically. Schmalhausen (1949) and Simpson (1953) explained the Baldwin effect genetically, without the inheritance of acquired characters. These explanations however assumed that the plasticity was eventually lost as the optimal phenotype became the only developmental possibility, and therefore heritable. However, this approach does not seem useful; a tree in the forest still needs to be able to grow into a new gap where there is light—plasticity needs to be retained.

Baldwin (1896) also proposed that learning tends to hasten the rate of evolution. Traditionally (e.g. Wright 1931, Grant 1991) learning, or any non-genetic component of phenotypic variability, was thought to slow the rate of evolution by diluting the genetic component, thereby reducing the efficiency of natural selection in sorting genetic variance. However, quantitative genetic models (Anderson, 1995) suggest that after an environmental change, populations of individuals able to ‘search phenotype space’ and those that can learn, will tend to find fitness peaks faster. Using neural networks, Hinton and Nowlan (1987) showed that non-genetically acquired phenotypes could allow an organism to find a fitness peak faster than networks that only had genetically determined variability. In terms of fitness landscapes, it is straightforward to produce models where a combination of phenotypic flexibility and genetic variants will find a new optimum faster than the same model with only the genetic component. Testing this hypothesis may be challenging, though we note the parallels with the earlier discussion on epigenotypes.

Wyles et al. (1983) reported that land vertebrates had an increasing rate of morphological evolution with increasing brain size to body size (encephalisation). How could larger brain size lead, on average, to a faster rate of morphological evolution? Their suggestion was that the more flexible behaviour of larger-brained animals allows them to broaden, for example, their use of food sources. Because the behaviour of the species is more flexible, it is possible that a new morphological variant would be advantageous in using the new food source. In this suggestion there is no direct linkage between relative brain size and morphological evolution. Mutations leading to improved learning ability could be selected for if behaviour was more flexible, and quite independently this could allow a different mutation to be selected that modified some aspect of morphology. To follow the idea further, the plasticity of flowering plants in varying their growth form in response to their local environment is considered the plant equivalent of flexible behaviour. For example, the phytochrome pigment system by detecting the level of shade, produces etiolation in plants (Smith 1974).

An important conclusion of these last two sections is that the potential to evolve is dependent on other organisms in the environment, with both intra- and inter-group competition being important. Potential evolvability is thus greater than realised evolvability.

Conclusions.

In this paper, we have examined a wide range of biological phenomena relevant to the concept of evolvability. In agreement with most authors, we conclude that there is no need to explain evolvability as having evolved in itself; the evolution of phenomena contributing to evolvability can be explained by current evolutionary theory. It is important to base models for evolvability on a range of data, rather than establishing *post hoc* explanations for a single dataset. To this end, we have examined how genome architecture affects evolvability in prokaryotes and eukaryotes.

In prokaryotes, an r-selected lifestyle is characterised by exponential growth in response to an energy source, with competition driving shorter doubling times. That prokaryotes possess a single replication origin places pressure on chromosome size, since replication is the rate-limiting step in cell doubling under exponential phase. Consequently, there is selection for elimination of superfluous DNA, including periodically-selected functions (PSFs). PSFs can be maintained by horizontal transfer, permitting more or less continual selection within a population or wider unit. Numerous prokaryotes maintain multiple genomic copies which may buffer against gene loss, provide a means of sidestepping the rate-limiting effect of replication by genome copy stockpiling, and may also permit the emergence of biochemical novelty through divergent evolution at identical copies of a given locus. This latter point, given the potential for additional catalytic activities in numerous enzymes (O'Brien & Herschlag 1999), may explain how prokaryotes have become so biochemically diverse and colonised so many environments (Rothschild & Mancinelli 2001), even with ongoing sequence elimination.

In general, eukaryotes are K-selected relative to prokaryotes (Carlile 1982). They possess multiple origins of replication per chromosome, and, with relatively stable nutrient sources, doubling times are not the major component to competition. Genome size is therefore not limited by replication rate, but by replication fidelity. Consequently, the accumulation of junk DNA is not in itself selected against. In eukaryotes, neutral evolution appears to be central to understanding complexity and evolvability. Accumulation of junk DNA is neutral, and conducive to occasional co-option of junk or duplicated DNA into a new function.

Both prokaryotic and eukaryotic parasites and endosymbionts have repeatedly undergone reductive evolution, losing massive amounts of genetic material. This is a convergent feature resulting from redundancy subsequent to the evolution of mechanisms for nutrient import. There may be less pressure for loss of superfluous sequences compared to free-living prokaryotes, as suggested by the 24% non-coding content of the *Rickettsia* genome, compared with around 10% for other bacterial genomes (Andersson et al. 1998b).

No hard boundary delineates r and K lifestyles which are best considered as a spectrum, helpful in understanding general patterns, but problematic if used to compare specific taxa. The utility of describing an r-K spectrum can be seen when comparing unicellular and simple eukaryotes to prokaryotes and complex multicellular eukaryotes. Unicellular eukaryotes appear to make use of horizontal transfer and tend to lose and gain PSFs, as supernumerary chromosomes in fungi (Covert 1998) demonstrate, but the eukaryote translation apparatus makes for response times on the order of an hour in yeast compared with minutes in *E. coli.*

Where prokaryotes and, to a lesser extent, unicellular eukaryotes have diversified through biochemical adaptation to a wide range of environments, multicellular eukaryotes have tended to colonise niches very similar to the initial niche. These can be reached by virtue of changes in structures, rather than the underlying biochemistry (e.g., the beaks of Darwin's finches (Lawrence 1999)). The emergence of an internal biochemical environment that can be regulated in response to starvation (e.g. by release of large reserves of stored energy) may have been a prerequisite to the emergence of morphological evolution in multicellular organisms, permitting the colonisation of new niches, but precluding access to ancestral niches.

Mechanisms for dealing with environmental stresses are also different between eukaryotes and prokaryotes On the whole, changes in environment which are lethal to the organism will result in extinction in specialised multicellular eukaryotes whereas adaptation to non-lethal, sustained changes in environment may be possible. The process of heritable adaptation cannot happen within-generation because developmental programs cannot be re-run to produce new, slightly modified structures in an adult. In prokaryotes, unicellular eukaryotes, and to some extent plants (which produce multiple centres of reproduction from vegetative tissue), there is the possibility of within-generation adaptation through immediate expression of a beneficial mutation or acquired gene. Viewed in these terms, prokaryote 'adapt or die' strategies make them more evolvable in response to environmental stress, while mechanisms to stabilise the internal environment in complex multicellular eukaryotes serve as a buffer to the external environment. Unicellular and simple multicellular eukaryotes are perhaps somewhere in the middle.

An important consequence of this is that the extensive biochemical change seen in prokaryotes and unicellular eukaryotes, together with reductive evolution, may explain the observation that r-selected organisms appear to have lost more early biochemical relics than multicellular eukaryotes (Poole et al. 1998, 1999). Much more of multicellular biochemistry may in fact be a frozen accident, though many processes would have been lost because of the diminished requirement for interaction with fluctuating environments. The relevance of organisms in extreme environments as models for the earliest organisms (Nisbet & Sleep 2001) must be reconsidered within this framework.

The effects of stress have been very important in experimental studies relevant to evolvability (particularly in prokaryotes), but we emphasise that we have still not covered all aspects of evolvability. Questions such as redundancy and modularity need more consideration, and other aspects of the system will affect potential evolvability in more ways than those described in our treatment of genome architecture and environmental interactions. A formal treatment of time scale, from within generations, to millions or billions of years, is also required.

Finally, our evolutionarily-stable niche-discontinuity (ESND) model emphasises the difference between potential and realised evolvability, the latter including limits placed on organisms from constraints in their environment. Lateral transfer in prokaryotes may break down some ESNDs in a way that is similar to the niche competition when organisms adapted to previously isolated niches are able to interact (e.g. geological changes allowing interaction of isolated biota, or the introduction of exotic species into an environment). Likewise, ESNDs can break down in some cases where complex behaviour is a trait in one organism, humans being the prime example. The emergence of plasticity, including complex behaviour, further separates organism from environmental changes because this allows a wider range of responses for a given genotype. The effect of organisms restricting the potential evolvability of others needs more consideration, as does plasticity (including learning).

Evolvability has been a loosely defined concept and it is important to avoid *post hoc* usages of it. As a final comment, evolvability, in one sense, never needed to evolve because information transfer is always error prone - early biological systems were of much lower fidelity, and therefore inherently ‘evolvable’.

Acknowledgements.
We thank David Martin for helpful discussions regarding the epigenotype concept. This work was supported by the New Zealand Marsden Fund.

References.

Adams, T.H., Wieser, J.K., Yu, J.-H., 1998. Asexual Sporulation in *Aspergillus nidulans.* Microbiol. Mol. Biol. Rev. 62, 35-54.

Agrawal, A., Eastman, Q.M., Schatz, D.G., 1998. Implications of transposition mediated by V(D)J-recombination proteins RAG1 and RAG2 for origins of antigen-specific immunity. Nature 394, 744-751.

Åkerlund, T., Nordstr•m, K., Bernander, R., 1995. Analysis of cell size and DNA content in exponentially growing and stationary-phase batch cultures of *Escherichia coli.* J. Bacteriol. 177, 6791-6797.

Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K., Watson, J.D., 1994. Molecular Biology of the Cell, 3rd. Ed. Garland Publishing, NY.

Anderson, R.W., 1995. Learning and evolution: a quantitative genetics approach. J. Theor. Biol. 175, 89-101.

Andersson, D.I., Slechta, E.S., Roth, J.R., 1998a. Evidence that gene amplification underlies adaptive mutability of the bacterial lac operon. Science 282, 1133-1135.

Andersson, J.O., Andersson, S.G.E., 1999. Genome degradation is an ongoing process in *Rickettsia.* Mol. Biol. Evol. 16, 1178-1191.

Andersson, S.G.E., et al., 1998b. The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. Nature 396, 133-140.

Baldwin, J.M., 1896. A new factor in evolution. Am. Nat. 30, 441-451.

Banuett, F., 1998. Signalling in the Yeasts: An Informational Cascade with Links to the Filamentous Fungi. Microbiol. Mol. Biol. Rev. 62, 249-274.

Beaton, M.J., Cavalier-Smith, T., 1999. Eukaryotic non-coding DNA is functional: evidence from the differential scaling of cryptomonad genomes. Proc. Roy. Soc. Lond. B 266, 2053-2059.

Bendich, A.J., Drlica, K., 2000. Prokaryotic and eukaryotic chromosomes: what's the difference? Bioessays 22, 481-486.

Benton, M.J., 1993. The Fossil Record 2. Chapman and Hall, London.

Bromham, L., Phillips, M .J., Penny, D., 1999. Growing up with dinosaurs: molecular dates and the mammalian radiation. Trends Ecol. Evol. 14, 113-118.

Brosius, J., 1999. RNAs from all categories generate retrosequences that may be exapted as novel genes or regulatory elements. Gene 238, 115-134.

Burch, C.L., Chao, L., 2000. Evolvability of an RNA virus is determined by its mutational neighbourhood. Nature 406, 625-628.

Cairns, J., Overbaugh, J., Miller, S., 1988. The origin of mutants. Nature 335, 142-145.

Carlile, M.J., 1982. Prokaryotes and eukaryotes: strategies and successes. Trends Biochem. Sci. 7, 128–130.

Cavaillé, J., et al., 2000. Identification of brain-specific and imprinted small nucleolar RNA genes exhibiting an unusual genomic organization. Proc. Natl. Acad. Sci. USA 97, 14311-14316.

Champe, S.P., Nagle, D.L., Yager, L.N., 1994. Sexual sporulation. Prog. Ind. Microbiol. 29, 429-454.

Charette, M., Gray, M.W., 2000. Pseudouridine in RNA: What, Where, How, and Why. IUBMB Life 49, 341-351.

Cole, S.T., et al., 2001. Massive gene decay in the leprosy Bacillus. Nature 409, 1007-1011.

Covert, S.F., 1998. Supernumerary chromosomes in filamentous fungi. Curr. Genet. 33, 311-319.

Crespi, B.J., 2001. The evolution of social behaviour in microorganisms. Trends Ecol. Evol. 16, 178-183.

Darnell, J.E., Doolittle, W.F., 1986. Speculations on the early course of evolution. Proc. Natl. Acad. Sci. USA 83, 1271-1275.

Dickinson, W.J., Seger, J., 1999. Cause and effect in evolution. Nature 399, 30.

Doolittle, W.F., 1998. You are what you eat: a gene transfer ratchet could account for bacterial genes in eukaryotic nuclear genomes. Trends Genet. 14, 307-311.

Douglas, S., et al., 2001. The highly reduced genome of an enslaved algal nucleus. Nature 410, 1091-1096.

Dover, G., 2000. Results may not fit well with current theories... Nature 408, 17.

Drake, J.W., 1999. The distribution of rates of spontaneous mutation over viruses, prokaryotes, and eukaryotes. Ann. NY Acad. Sci. 870, 100-107.

Farley, J., 1977. The spontaneous generation controversy from Descartes to Oparin. Johns Hopkins University Press, Baltimore MD.

Fijalkowska, I.J., Dunn, R.L., Schaaper, R.M., 1993. Mutants of *Escherichia coli* with increased fidelity of DNA replication. Genetics 134, 1023-1030.

Filipowicz, W., 2000. Imprinted expression of small nucleolar RNAs in brain: Time for RNomics. Proc. Natl. Acad. Sci. USA 97, 14035-14037.

Finkel, S.E., Kolter, R., 1999. Evolution of microbial diversity during prolonged starvation. Proc. Natl. Acad. Sci. USA 96, 4023-4027.

Forterre, P., 1995. Thermoreduction, a hypothesis for the origin of prokaryotes. CR Acad. Sci. Paris III 318, 415-422.

Forterre, P., Philippe, H., 1999. Where is the root of the universal tree of life? Bioessays 21, 871-879.

Forterre, P., Bouthier de la Tour, C., Philippe, H., Duguet, M., 2000. Reverse gyrase from hyperthermophiles: probable transfer of a thermoadaptation trait from Archaea to Bacteria. Trends Genet. 16, 152-154.

Foster, P.L., 1999. Mechanisms of stationary phase mutation: a decade of adaptive mutation. Annu. Rev. Genet. 33, 57-88.

Fraser, C.M., et al., 1998. Complete genome sequence of *Treponema pallidum*, the syphilis spirochaete. Science 281, 375-388.

Galtier, N., Tourasse, N., Gouy, M., 1999. A nonhyperthermophilic common ancestor to extant life forms. Science. 283, 220-221.

Gibson, G., 2000. Evolution: Hox genes and the cellared wine principle. Curr. Biol. 10, R452-R455.

Glansdorff, N., 2000. About the last common ancestor, the universal life-tree and lateral gene transfer: a reappraisal. Mol. Microbiol. 38, 177-185.

Gould, S. J., Lewontin, R. C., 1979. The spandrels of San Marco and the Panglossian paradigm: A critique of the adaptationist program. Proc. R. Soc. Lond. B 205, 581-598.

Grant, V., 1991. The Evolutionary Process. Columbia Univ. Press, New York.Gray, M.W., Burger, G., Lang, B.F., 1999. Mitochondrial evolution. Nature 283, 1476-1481.

Graveley, B.R., 2001. Alternative splicing: increasing diversity in the proteomic world. Trends Genet. 17, 100-107.

Gray, M.W., Burger, G., Lang, B.F., 1999. Mitochondrial evolution. Science. 283, 1476-1481.

Grbic, M., 2000. "Alien" wasps and evolution of development. BioEssays 22, 920-932.

Harris, J.R., 1998. Placental endogenous retrovirus (ERV): structural, functional, and evolutionary significance. BioEssays 20, 307-316.

Hartl, D.L., 2000. Molecular melodies in high and low C. Nat. Rev. Genet. 1, 145-149.

Hastings, P.J., Bull, H.J., Klump, J.R., Rosenberg, S.M., 2000. Adaptive amplification: an inducible chromosomal instability mechanism. Cell 103, 723-731.

Herbert, A., Rich, A., 1999. RNA processing in evolution. The logic of soft-wired genomes. Ann. N.Y. Acad. Sci. 870, 119-132.

Hinton, G.E., Nowlan S.J., 1987. How learning can guide evolution. Complex systems 1, 495-502.

Hiom, K., Mele, M., Gellert, M., 1998. DNA transposition by the RAG1 and RAG2 proteins: a possible source of oncogenic translocations. Cell 94, 463-470.

Holliday, R., Grigg, G.W., 1993. DNA methylation and mutation. Mutat. Res. 285, 61-67.

Hood, D.W., et al., 1996. DNA repeats identify novel virulence genes in *Haemophilus influenzae*. Proc. Natl. Acad. Sci. USA 93, 11121-11125.

Jablonski, D., 1986. Background and mass extinctions: the alteration of macroevolutionary regimes. Science 231, 129-133.

Jacob, F., 1977. Evolution and Tinkering. Science 196, 1161-1166.

Jacobs, H., Bross, L., 2001. Towards an understanding of somatic hypermutation. Curr. Opin. Immunol. 13, 208-218.

Jones, M.E., Stoddart, D.M., 1998. Reconstruction of the predatory behaviour of the extinct marsupial thylacine (*Thylacinus cynocephalus*). J. Zool. Soc. Lond. 246, 239-246.

Kalman, S., et al. 1999. Comparative genomes of *Chlamydia pneumoniae* and *C. trachomatis.* Nat. Genet. 21, 385-389.

Kasak, L., Horãk, R., Kivisaar, M., 1997 Promoter-creating mutations in *Pseudomonas putida*: A model system for the study of mutation in starving bacteria. Proc. Natl. Acad. Sci. USA 94, 3134-3139.

Kirschner, M., Gerhart, J., 1998. Evolvability. Proc. Natl. Acad. Sci. USA 95, 8420-8427.

Koch, A.L., 1984. Evolution vs the number of gene copies per primitive cell. J. Mol. Evol. 20, 71-76.

Lafontaine, D.L.J., Tollervey, D., 1998. Birth of the snoRNPs: the evolution of the modification-guide snoRNAs. Trends Biochem. Sci. 23, 383-388.

Lan, R., Reeves, P.R., 2000. Intra-species variation in bacterial genomes: the need for a species genome concept. Trends Microbiol. 8, 396-401.

Landree, M.A., Wibbenmeyer, J.A., Roth, D.B., 1999. Mutational analysis of RAG1 and RAG2 identifies three catalytic amino acids in RAG1 critical for both cleavage steps of V(D)J recombination. Genes Dev. 13, 3059-3069.

Larsson, E., Andersson, G., 1998. Beneficial role of Human Endogenous Retroviruses: Facts and Hypotheses. Scand. J. Immunol. 48, 329-338.

Lawrence, J., 1999. Selfish operons: the evolutionary impact of gene clustering in prokaryotes and eukaryotes Curr. Opin. Genet. Dev. 9, 642-648.

Lazazzera, B.A.,Kurtser, I.G., McQuade, R.S., Grossman, A.D., 1999. An autoregulatory circuit affecting peptide signalling in *Bacillus subtilis.* J. Bact. 181, 5193-5200.

Levin, P.A., Grossman, A.D., 1998. Cell cycle and sporulation in *Bacillus subtilis.* Curr. Opin. Microbiol. 1, 630-635.

Lin, L., Xu, B., Rote, N.S., 1999. Expression of Endogenous Retrovirus ERV-3 Induces Differentiation in BeWo, a Choriocarcinoma Model of Human Placental Trophoblast. Placenta 20, 109-118.

Lindquist, S., 2000. ...but yeast prion offers clues about evolution. Nature 408, 17-18.

Maldonado, R., Jimenez, J., Casadesus, J., 1994. Changes of ploidy during the *Azotobacter vinelandii* growth cycle. J. Bacteriol. 176, 3911-3919.

Mangeney, M., Heidmann, T., 1998. Tumor cells expressing a retroviral envelope escape immune rejection *in vivo.* Proc. Natl. Acad. Sci. USA 95:14920-14925.

Marguet, E., Forterre, P., 1994. DNA stability at temperatures typical for thermophiles. Nucleic Acids Res. 22, 1681–1686.

McFadden, G.I., 1999. Endosymbiosis and evolution of the plant cell. Curr. Opin. Plant Biol. 2,513-519.

McKenzie, G.J., Harris, R.S., Lee, P.L., Rosenberg, S.M., 2000. The SOS response regulates adaptive mutation. Proc. Natl. Acad. Sci. USA 97, 6646-6651.

Melek, M., Gellert, M., 2000. RAG1/2-mediated resolution of transposition intermediates: two pathways and possible consequences. Cell. 101, 625-633.

Metzgar, D., Wills, C., 2000. Evidence for the adaptive evolution of mutation rates. Cell 101, 581-584.

Monk, M., 1995. Epigenetic programming of differential gene expression in development and evolution. Dev. Genet. 17, 188-197.

Moran, N.A., 1996. Accelerated evolution and Muller's Ratchet in endosymbiotic bacteria. Proc. Natl. Acad. Sci. USA 93, 2873-2878.

Moran, N., Baumann, P., 2000. Bacterial endosymbionts in animals. Curr. Opin. Microbiol. 3, 270-275.

Morgan, H.D., Sutherland, H.G.E., Martin, D.I.K., Whitelaw, E., 1999. Epigenetic inheritance at the agouti locus in the mouse. Nat. Genet. 23, 314-318.

Morrissey, J.P., Tollervey, D., 1995. Birth of the snoRNPs: the evolution of RNase MRP and the eukaryotic pre-rRNA-processing system. Trends Biochem. Sci. 20, 78-82.

Moxon, E.R., Rainey, P.B., Nowak, M.A., Lenski, R.E., 1994. Adaptive evolution of highly mutable loci in pathogenic bacteria. Curr. Biol. 4, 24-33.

Moxon, E.R., Thaler, D.S., 1997. The Tinkerer's evolving toolbox. Nature 387, 659-662.

Muramatsu, M., Kinoshita, K., Fagarasan, S., Yamada, S., Shinkai, Y., Honjo, T., 2000. Class switch recombination and hypermutation require activation-induced cytidine deaminase (AID), a potential RNA editing enzyme. Cell 102, 553-563.

Nelson, K.E., et al., 1999. Evidence for lateral gene transfer between Archaea and Bacteria from genome sequence of Thermotoga maritima. Nature 399, 323-329.

Nisbet, E.G., Sleep, N.H., 2001. The habitat and nature of early life. Nature 409, 1083-1091.

Novacek, M.J., 1985. Evidence for echolocation in the oldest known bats. Nature 306, 683-684.

O'Brien, P.J., Herschlag, D., 1999. Catalytic promiscuity and the evolution of new enzymatic activities. Chem. Biol. 6, R91-R105.

Omer, A.D., Lowe, T.M., Russell, A.G., Ebhardt, H., Eddy, S.R., Dennis, P.P., 2000. Homologs of small nucleolar RNAs in Archaea. Science 288, 517-522.

Partridge, L., Barton, N.H., 2000. Evolving evolvability. Nature 407, 457-458.

Penny, D., Poole, A., 1999. The nature of the Last Universal Common Ancestor. Curr. Opin. Genet. Dev. 9, 672-677.

Plaga, W., Schairer, H.U., 1999. Intercellular signalling in *Stigmatella aurantiaca.* Curr. Opin. Microbiol. 2, 593-597.

Plasterk, R., 1998. V(D)J recombination. Ragtime jumping. Nature 394, 718-719.

Poole, A.M., Jeffares, D.C., Penny, D., 1998. The path from the RNA world. J. Mol. Evol. 46, 1-17.

Poole, A., Jeffares, D., Penny, D., 1999. Early evolution: prokaryotes, the new kids on the block. Bioessays 21, 880-889.

Poole, A., Penny, D., Sjöberg, B.-M., 2001. Confounded cytosine! Tinkering and the evolution of DNA. Nat. Rev. Mol. Cell. Biol. 2, 147-151.

Powell, S.C., Wartell, R.M., 2001. Different characteristics distinguish early versus late arising adaptive mutations in *Escherichia coli* FC40. Mutat. Res. 473, 219-228.

Pupo, G.M., Lan, R., Reeves, P.R., 2000. Multiple independent origins of Shigella clones of *Escherichia coli* and convergent evolution of many of their characteristics. Proc. Natl. Acad. Sci. U.S.A. 97, 10567-10572.

Radman, M., Matic, I., Taddei, F., 1999. Evolution of evolvability. Ann. N.Y. Acad. Sci. 870, 146-155.

Raven, P.H., Evert, R.E., Eichhorn, S.E., 1986. Biology of Plants. Worth Publishers, New York.

Reanney, D.C., 1974. On the origin of prokaryotes. J. Theor. Biol. 48, 243-251.

Reenan, R.A., 2001. The RNA world meets behavior: A-I pre-mRNA editing in animals. Trends Genet. 17, 53-56.

Revy, P., et al. 2000. Activation-induced cytidine deaminase (AID) deficiency causes the autosomal recessive form of the Hyper-IgM syndrome (HIGM2). Cell 102, 565-575.

Rosenzweig, M.L., McCord, R.D., 1991. Incumbent Replacement: evidence for long-term evolutionary progress. Paleobiology 17, 202-213.

Rothschild, L.J., Mancinelli, R.L., 2001. Life in extreme environments. Nature 409, 1092-1101.

Roy, K., 1996. The roles of mass extinction and biotic interaction in large-scale replacements: a reexamination using the fossil record of stromboidean gastropods. Paleobiology 22, 436-452.

Ruben, J., 1995. The evolution of endothermy in mammals and birds: from physiology to fossils. Ann. Rev. Physiol. 57, 69-95.

Ruddle, F.H., et al., 1999. Evolution of chordate Hox gene clusters. Ann. N.Y. Acad. Sci. 870, 238-248.

Rutherford, S.L., Lindquist, S.L., 1998. Hsp 90 as a capacitor for morphological evolution. Nature 406, 336-342.

Schärer, O.D., Jiricny, J., 2001. Recent progress in the biology, chemistry and structural biology of DNA glycosylases. Bioessays 23, 270-281.

Schmalhausen, I.I., 1949. Factors of Evolution. University of Chicago Press, Chicago.

Sereno, P., 1999. The evolution of dinosaurs. Science 284, 2137-2147.

Sharp, P.A., 1994. Split genes and RNA splicing. Cell 77, 805–815

Simpson, G.G., 1953. The Baldwin effect. Evolution 7, 110-117.

Smit, A.F.A., 1999. Interspersed repeats and other mementos of transposable elements in mammalian genomes. Curr. Opin. Genet. Dev. 9, 657-663.

Smith, H., 1974. Phytochrome and photomorphogenesis; an introduction to the photocontrol of plant development. McGraw-Hill, London.

Smith, H.O., Gwinn, M.L., Salzberg, S.L., 1999. DNA uptake signal sequences in naturally transformable bacteria. Res. Microbiol. 150, 603-616.

Smith, H.C., Gott, J.M., Hanson, M.R., 1997. A guide to RNA editing. RNA 3, 1105-1123.

Sniegowski, P.D., Gerrish, P.J., Lenski, R.E., 1997. Evolution of high mutation rates in experimental populations of *E. coli.* Nature 387, 703-705.

Solomon, J.M., Grossman, A.D., 1996. Who's competent and when: regulation of natural genetic competence in bacteria. Trends Genet. 12, 150-155.

Sontheimer, E.J., Gordon, P.M., Piccirilli, J.A., 1999. Metal ion catalysis during group II intron self-splicing: parallels with the spliceosome Genes Dev. 13, 1729-1741.

Stoltzfus, A., 1999. On the possibility of constructive neutral evolution. 49, 169-181.

Szathmáry, E., Maynard Smith, J., 1995. The major evolutionary transitions. Nature 374, 227-232.

Torkelson, J. Harris, R.S., Lombardo, M-J., Nagendran, J., Thulin, C., Rosenberg, S.M., 1997. Genome-wide hypermutation in a subpopulation of stationary-phase cells underlies recombination-dependent adaptive mutation. EMBO J. 16, 3303–3311

Tortosa, P., Dubnau, D., 1999. Competence for transformation: a matter of taste. Curr. Opin. Microbiol. 2, 588-592.

True, H.L., Lindquist, S.L., 2000. A yeast prion provides a mechanism for genetic variation and phenotypic variability. Nature 407, 477-483.

Varon, M., Choder, M., 2000. Organization and cell-cell interaction in starved *Saccharomyces cerevisiae* colonies. J. Bacteriol. 182, 3877-3880.

Wagner, A., 1996. Does evolutionary plasticity evolve? Evolution 50, 1008-1023.

Wagner, G.P., Altenberg, L., 1996. Complex adaptations and the evolution of evolvability. Evolution 50, 967-976.

Walton, J.D., 2000. Horizontal gene transfer and the evolution of secondary metabolite gene clusters in fungi: an hypothesis. Fungal Genet. Biol. 30, 167-171.

Ward, M.J., Zusman, D.R., 1999. Motility in Myxococcus xanthus and its role in developmental aggregation. Curr. Opin. Microbiol. 2, 624-629.

Weinstein, L.B., Steitz, J.A., 1999. Guided tours: from precursor snoRNA to functional snoRNP. Curr. Opin. Cell. Biol. 11, 378-384.

Whitelaw, E., Martin, D.I.K., 2001. Retrotransposons as epigenetic mediators of phenotypic variation in mammals. Nat. Genet. 27, 361-365.

Woese, C.R., 1998. The universal ancestor. Proc. Natl. Acad. Sci. USA 95, 6854-6859.

Wolfe, K.H., Shields, D.C., 1997. Molecular evidence for an ancient duplication of the entire yeast genome. Nature 387, 708-713.

Wright, S.K., 1931. Evolution in Mendelian populations. Genetics 16, 97-159.

Wyles, J.S., Kunkel, J.G., Wilson, A.C., 1983. Birds, behavior, and anatomical evolution. Proc. Natl. Acad. Sci. USA 80, 4394-4397.

Yoder, J.A., Walsh, C.P., Bestor, T.H., 1997. Cytosine methylation and the ecology of intragenomic parasites. Trends Genet. 13, 335–340.

Table 1. Examples of stress response which mayaffect evolvability.

Prokaryotes

Mechanism	**Activating stress**	**Organism(s)**	**Notes**	**References**
Global hypermutation	Occurs in stationary phase, thus likely to be a starvation response.	*E.coli* *Pseudomonas putida?*	Hypermutation is transient, recombination-dependent and in stationary phase.	Torkelson et al.1997 McKenzie et al. 2000 Kasak et al. 1997
Local hypermutation (contingency loci)	Recurrent selection, such as in host-parasite coevolution.	*H. influenzae* *E. coli* *S. typhimurium*	E.g. phenotypic switching of surface antigens, hypermutable virulence factors. cf V(D)J hypervariability.	Moxon et al. 1994 Hood et al. 1996
Gene amplification	Occurs in stationary phase, thus likely to be a starvation response.	*S. typhimurium* *E. coli*	Requires residual activity at amplified locus. In late arising colonies.	Andersson et al. 1998a Powell & Wartell 2001 Hastings et al. 2000
Genetic competence (DNA uptake)	Occurs in stationary phase.	*B. subtilis* *Streptococcus pneumoniae* *H. influenzae*	Extracellular signalling molecules indicate a cell density 'quorum' which establishes competence.	Solomon & Grossman 1996 Tortosa & Dubnau 1999
Sporulation		*B. subtilis*	Sporulation controlled by the same pathway as competence.	Levin & Grossman 1998
Cell-cell interaction	Starvation	*Stigmatella auantiaca* *Myxococcus xanthus*	Sporulation occurs in response to starvation in these myxobacteria	Ward & Zusman 1999 Plaga & Schairer 1999

Eukaryotes

Mechanism	Activating stress	Organism	Notes	References
Sexual sporulation	Starvation	*S. cerevisiae* *A. nidulans*	*Saccharomyces* enters meiosis upon nitrogen starvation. *Aspergillus* sporulates sexually at low glucose concentrations,. At high glucose it switches to asexual sporulation (dispersal)	Banuett 1998 Adams et al. 1998
Supernumerary chromosomes		Fungi	Not usually stably maintained in the genome. cf plasmids.	Covert 1998
Cell-cell interaction	Starvation	*S. cerevisiae* *D. discoideum*	In yeast, connecting filaments form between cells. Starvation promotes fruiting body and spore formation.	Varon & Choder 2000 Crespi 2001
PSI-dependent translation readthrough.	Heat shock protein-mediated	*S. cerevisiae*	PSI normally translation terminator. Change in protein conformation occurs.	True & Lindquist 2000
Hsp 90-mediated phenotype exploration.	Heat stress, other stresses involving Hsp 90.	*D. melanogaster*	Hypothesised that Hsp 90 titration during heat stress lifts buffering, resulting in hidden phenotypes being tested.	Rutherford & Lindquist 1998
Local hypermutation	Host-parasite interactions	Mammals	Somatic hypermutation of V(D)J genes in antibody formation.	Jacobs & Bross 2001

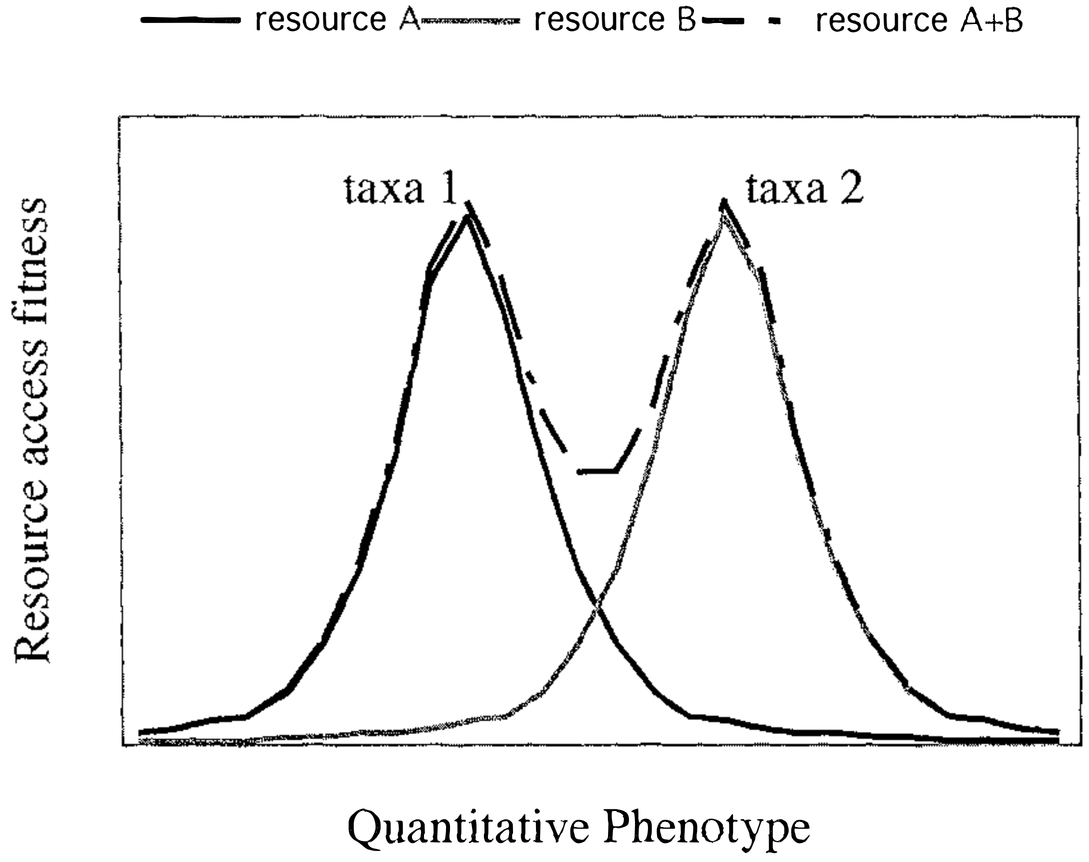


Figure 1. Evolutionarily stable niche discontinuity between two taxa. The curves represent relative fitness contribution derived by an organism from access to resources A (black line) and B (grey line), as dependent on a quantitative phenotype. The sum of the curves for resources A+B (dashed line) represents the overall relative fitness of organism's with respect to a quantitative phenotype. The signature of an ESND is a direction of selection pattern creating a valley of low fitness. This is expected to occur where there is a deleterious phenotype shift trade-off between interspecific and intraspecific competition.

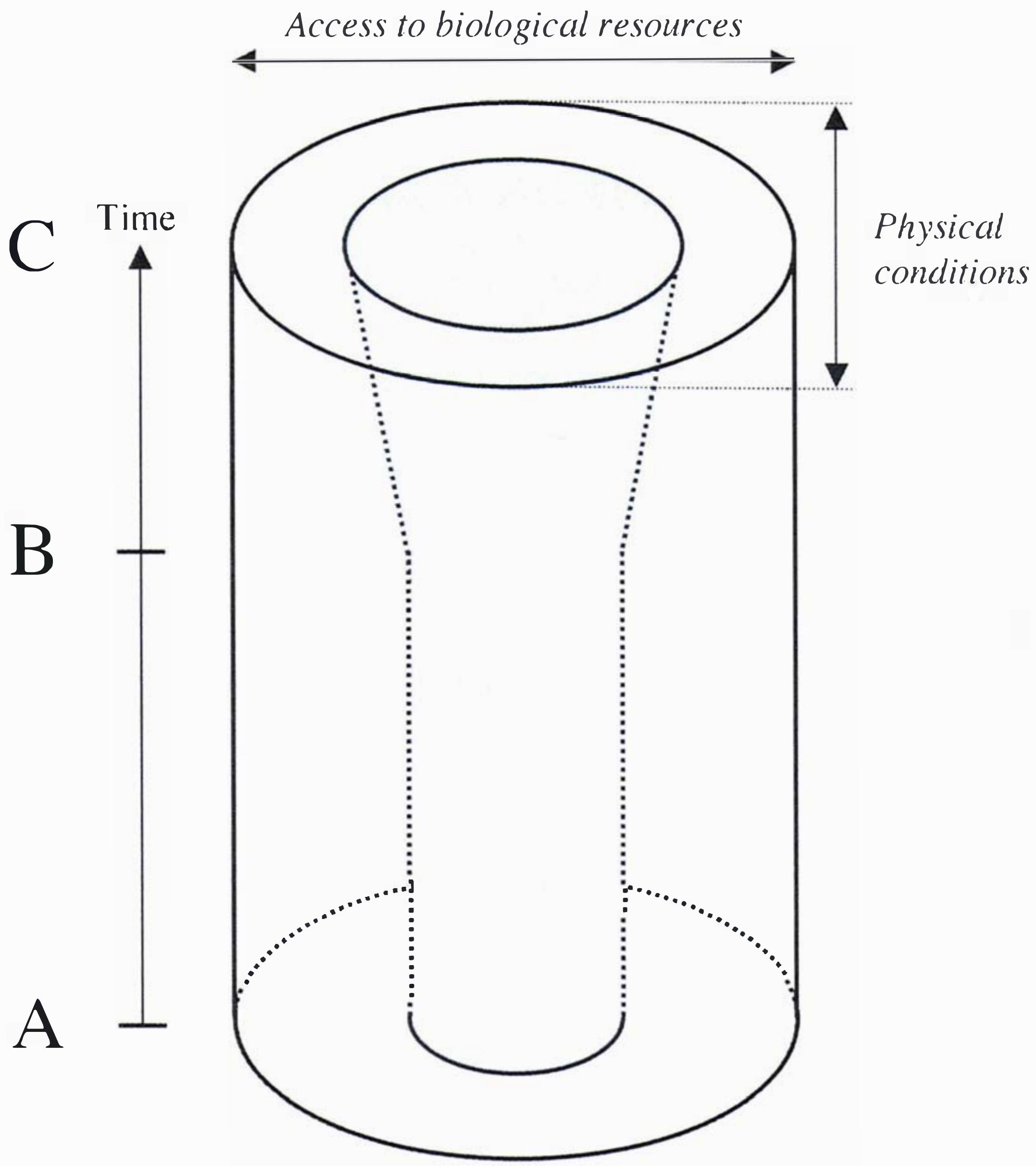


Figure 2 traces the relationship between the potential (transparent) and realised (shaded) niche through time for a hypothetical organism. The potential niche includes the full range of physical (axis 1) and biotic (axis 2) conditions for which the organism can survive and reproduce. The effect of competition and predation on fitness contracts this range, leaving the realised niche (shaded) that naturally occurs. Extinction of a predator at time B allows the expansion of the realised niche (within the bounds of the potential niche). Changes to the potential niche may follow due to alteration of the fitness landscape owing to the expansion of the realised niche.

Box 1 – Post hoc explanations

The Story of Darryl.

Darryl lived in a small farmhouse on the edge of an isolated village. Perhaps as a result of generations of inbreeding, he was slow, but very gentle and wouldn't even harm a fly. Darryl had one ability that really endeared him to the locals. He was a fantastic shot.

The wall of Darryl's barn was covered with small round circles, each with a small hole right in the centre where a bullet had hit. An intrigued journalist from the neighboring town arranged an interview for a feature story.

"Tell me Darryl", she said, "how is it that you are such a good shot with a rifle?" Darryl replied,

"It's veeery simple, --- I taaakes my rifle, --- aaaims it at the wall, --- puuulls the trigger, --- fiiinds where it hits, --- and draaaws a circle around it."

Box 2. r and K selection.

Rate of population growth, R, is given by the equation:

$$R = dN/dt = rN(1-N/K)$$

Where:

r = maximum intrinsic rate of increase for a population

N = number of organisms

K = carrying capacity (of the environment)

r-selected organisms:	*K-selected organisms:*
• small	• large
• high reproductive rates	• lower, more constant, reproductive rate
• short life cycles	• longer life cycles
• live in unpredictable environments	• live in more stable environments
• fluctuation in resource availability and type requires fast response times	• resources in more constant supply (though limited in amount)
• population size varies hugely	• population size relatively stable

r- and K- selection is a relative measure. While specific application of this concept is problematic (organism A may be r-selected relative to organism B, but K-selected relative to organism C), it is no more problematic than fitness, which is also a relative measure. The concept is useful in general discussions such as this since it aims to explain many aspects of prokaryotes and eukaryotes, rather than invoke special explanations for each feature.

Box 3. Selection pressures on genome organisation and consequences for evolvability.

Prokaryotes.

Fast reproductive rate during exponential growth is a consequence of r selection.

- Under r selection, fast replication is selectively advantageous.
- Under fast replication, a single origin of replication per chromosome limits genome size.
- Consequently there is selection against multiple copies of genes, 'junk DNA', and genes that are only rarely required (periodically-selected').
- Horizontal transfer is advantageous for recovering periodically-selected' genes.

Copy number.

- Retaining multiple copies of the genome appears widespread (Bendich & Drlica 2000).
- This redundancy provides a buffer to deleterious mutation, and is expected to promote survival during hypermutation in the stationary phase (Finkel & Kolter 1999).
- Redundancy may favour diversification of new functions similar to duplication and divergence in eukaryote genomes.
- May maintain faster cell division through genome stockpiling – overcomes a problem if replication takes an hour, but cells can double in 20 minutes during exponential growth.

Plasmids.

- Maintain periodically-selected functions in r-selected populations; a gene on a plasmid can be retained within a population though lost from individuals.

Operons.

- Transferable units of metabolism. The origin of the operon organisation is debated, but once formed, an operon may be spread through horizontal transfer (Lawrence 1999).

Response times.

- Ability to respond quickly to changes in environment, e.g. presence of a new substrate, is a feature of r-selected organisms.
- Beginning translation before transcription is finished allows fast response. mRNA being extensively processed, and then exported from the nucleus, makes response times much slower even in r-selected eukaryotes such as yeast. Response times are in minutes in prokaryotes, and of the order of an hour in yeast (Alberts et al. 1994).
- Loss of extensive transcript processing will be selected for.

Environmental interactions.

- Regulation of developmental pathways are strongly linked to environmental cues. Examples are fruiting body formation (asexual sporulation), genetic competence, biofilm formation, regulation of virulence (see Table 1 in Crespi 2001).

Eukaryotes.

In K-selected organisms reproductive rate is slower.

- Given many centres of replication (and replicons) there are few constraints on genome size, and accumulation of junk DNA is not inherently disadvantageous. Thus expansion of genome size through transposable elements, retroviral incorporation, duplication of genes or genomic regions can occur frequently.
- Occasional recruitment of new function from this pool is possible.
- Similarly, duplication and divergence of genes is a major source of evolutionary novelty.

Extensive transcript processing.

- K-selected organisms tend to occur where nutrient supply is more stable.
- Fast gene expression is therefore not strongly selected, so extensive transcript processing is not strongly disadvantageous.
- Any potential benefits of processing, such as alternative splicing and RNA editing can therefore be realised, and lead to many RNA intermediates from one gene, resulting in a more complex genotype-phenotype relationship (the ribotype concept of Herbert & Rich 1999).

Constitutive multicellularity in eukaryotes.

- Increased propensity for division of labour among 'obligate cooperators' results in cell specialisation. (This occurs transiently in other eukaryotes and in prokaryotes).
- Specialisation also results in different, irreversible, developmental fates of cells, tissues and organs, larval and adult stages in metazoa, polyphenic insects.
- Specialisation can lead to efficient mechanisms for large-scale nutrient storage (e.g. adipose tissue, glycogen, and starch), further stabilising the control of nutrients.
- Specialisation permits heavy investment in specific structures such as organs and mechanical tools for nutrient acquisition, defence or competition.
- Regulation of developmental pathways is less dependent on environmental cues, with greater internal control.

All the above are generalisations to which there must be exceptions. Describing an organism as r- or K-selected is relative, and focuses on the extremes (prokaryotes and multicellular eukaryotes). The differences are on a continuum. For instance, unicellular eukaryotes are r-selected relative to their multicellular relatives, and many of the points listed under prokaryotes apply to this group. Transcript processing and junk accumulation is less extensive in unicellular eukaryotes, operons and periodically-selected functions are a feature of their genomes, and developmental regulation is tightly linked to environmental cues. Constitutive multicellularity makes horizontal transfer unlikely, but unicellular eukaryotes may acquire new functions through DNA uptake.

Future work

Future work

Testing the thermoreduction hypothesis.

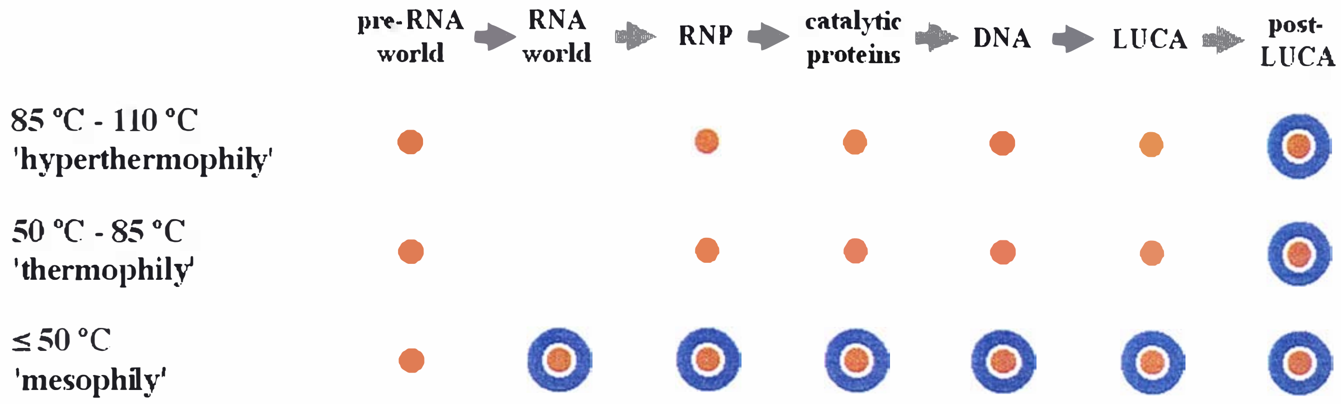
In this thesis, I have examined a wide range of issues with respect to the origin of prokaryotes and eukaryotes. My overall conclusion is that prokaryotes represent derived lineages, not ancestral ones, as has generally been thought. By using a well-developed model for the RNA world as the outgroup, it has been possible to establish that a number of expected features of the LUCA have been maintained in eukaryotes and lost in prokaryotes. Eukaryotes have been more conservative in terms of biochemical evolution, and and I have presented a detailed discussion on prokaryote and eukaryote evolvability to support this claim. I have questioned the dogma that all eukaryote-specific features are recent evolutionary innovations, and have presented a challenge to this dogma in the form of a critique of the problems surrounding the question of the origin of the nucleus.

In each chapter, specific conclusions are presented, and it would be redundant to describe these again. This thesis has concentrated both on the use of RNA relics as a marker for establishing the direction of evolution at the root of the tree of life, and on ecological aspects of prokaryote and eukaryote lifestyle and how this could account for my findings. An independent test is however available, and has been briefly suggested in several of the chapters, but not described in detail.

If Forterre's thermoreduction hypothesis is correct, evidence of past thermophily should be identifiable in prokaryote lineages, but not eukaryotes. If the LUCA was a thermophile however, such evidence will be found in all three lineages. A range of traits which contribute to thermostability at high temperatures might be relevant in testing the thermoreduction hypothesis [see Forterre 1996; Daniel & Cowan 2000] and three specific studies are outlined below.

Thermoreduction can be invoked in understanding the phylogenetic distribution of RNA relics, and r-selection reinforces this. Perhaps the most obvious prokaryotic feature that the latter cannot account for however is the emergence of circular genomes. If the prokaryote lineages did evolve through thermoreduction, this can be tested by looking for signatures of past thermophily in mesophilic prokaryotes and in eukaryotes. The thermoreduction hypothesis predicts that such signatures will be present in mesophilic prokaryotes, but absent from eukaryotes. The 'thermophilic LUCA' hypothesis predicts that evidence of past thermophily will be present in all three domains, though this has never been tested.

Both hypotheses actually cover a range of possibilities, which in a simple form can be considered as 'cold start, hot LUCA' or 'hot start, hot LUCA' for a thermophilic



	pre-RNA world	RNA world	RNP	catalytic proteins	DNA	LUCA	post-LUCA
85 °C - 110 °C 'hyperthermophily'	●		●	●	●	●	◎
50 °C - 85 °C 'thermophily'	●		●	●	●	●	◎
≤ 50 °C 'mesophily'	●	◎	◎	◎	◎	◎	◎

Figure 1. Relationship between temperature and the origin and early evolution of life.

This figure is a formalisation of the relationship between temperature and the origin of life, based on Figure 3 of Moulton et al. [2000]. Red dots indicate whether each stage can, in principle, exist within the temperature ranges indicated to the right. In the case of periods predating the RNA world, it is not clear whether life began at high or low temperatures, and the limits are not well established, because the processes and requisites are not established. For the RNA world period, the upper limit on stability of tertiary structure of naked RNA [Brion & Westhof 1997] limits this period to the lowest temperature range shown. The permissibility of later periods is established by whether modern organisms living at various temperatures possess any of these traits shown. For instance, for the RNP stage, the ribosome is known to be ubiquitous, so RNPs can clearly function at over 100°C. While any combination of stages is possible in principle, the blue rings indicate the hypothesis that best fits with the RNA world data described in this thesis. The data cannot be used to examine earlier periods in the origin of life, as has been pointed out elsewhere [Miller & Bada 1998].

LUCA, and 'cold start, cold LUCA' or 'hot start, cold LUCA' for thermoreduction. This is discussed by Miller & Bada [1988], and by Moulton et al. [2000]. Moulton et al. provide a more detailed set of scenarios, though because they only consider conditions in the RNA world period and later, they do not consider the possibility of a 'hot start' prior to the emergence of RNA. Figure 1, extends the work of Moulton et al. to cover all aspects of the origin of life, in line with the general consensus that the RNA world would have been mesophilic, but earlier prebiotic periods could have involved either low or high temperatures.

The RNA world data presented here supports all scenarios where the RNA world and LUCA existed at moderate temperatures. Specific estimates of the upper temperature limit for both the RNA world and the LUCA are possible, given the conclusions presented here. For the RNA world, the upper temperature limit is dictated by the stability of RNA tertiary structure, which is lost under 50°C (reviewed in Brion & Westhof [1997]). The upper limit might conceivably be increased through stabilisation of RNA by Mg^{2+} [Brion & Westhof 1997], or through methylation [Kowalak et al. 1994; Noon et al. 1998]. The LUCA is more difficult to estimate in that the potential for stabilisation of thermolabile traits is available. Nevertheless, given that the RNA world data establish the eukaryotic lineage as having been more conservative in terms of RNA replacement, it is reasonable to assume that higher temperature tolerance in prokaryotes is derived (and concurrent with replacement of ancestral RNA biochemistry), so the LUCA most probably existed at those temperatures inhabited by modern day eukaryotes. While some putatively 'thermophilic' eukaryotes have been identified (such as desert ants & bees, polychaetes worms from hydrothermal vents [McMullin et al. 2000] and *Tetrahymena thermophila* [Hallberg et al. 1985]) in all cases, none have been shown to stand sustained internal temperatures above 50°C. The Australian ant *Melophorus begoti* is capable of surviving at 54°C for one hour, with a critical thermal maximum of 56.7°C. On phylogeny, these have evolved from more mesophilic organisms, and while little is known of their usage of RNA, the maximum may be predicted to be set by RNA tertiary structure, as suggested by the close correlation between internal body temperature [see McMullin et al. 2000] and upper limits on RNA tertiary structure [see Brion & Westhof 1997]. In the case of the hydrothermal polychaete worms, proteins such as haemoglobin and collagen have been shown to be unstable at temperatures approaching 50°C [reviewed in McMullin et al. 2000].

Extremes of pressure might be relevant to increased stability, but recent studies have suggested that pressure results in unfolding as a result of water penetration into the protein matrix [Silva et al. 2001]. Nevertheless, prokaryotes have clearly surpassed these limits [Rothschild & Mancinelli 2001], and likewise, proteins have been identified that are stable well above the growth temperature of hyperthermophiles [Hiller et al. 1997]. The point that is interesting in light of the RNA world hypothesis is that, in the absence of mechanisms of stability (such as protein-mediated stabilisation), the expectation is that the upper limit for life in this period would not have exceeded 50°C, and the data in this thesis are most consistent with this constraint having been present in the LUCA and in the lineage leading to modern eukaryotes, while the prokaryote lineages developed mechanisms which enabled colonisation of higher temperatures [Figure 1].

A final point is that the interpretation of early evolution described here is not synonymous with the thermoreduction hypothesis, it is merely consistent with it. For any thesis suggesting that a high temperature lifestyle is ancestral, it is still necessary to explain what selection pressure led to the replacement of protein by RNA in ancient processes subsequent to the emergence of eukaryotes from prokaryote-like ancestors. The direction of evolution from RNA to RNP to protein described in this thesis was based not on temperature considerations, but on the evolution of catalytic efficiency. Where thermoreduction is perhaps important is that RNA instability at high temperatures may result in replacement of RNP with protein, even where the RNP had reached catalytic perfection. No detailed argument for protein replacement by RNP has been provided by those who favour the various thermophilic LUCA hypotheses.

In table 1, I have described all patterns in the data that might be observed (including those which have not been observed, and are not predicted by either thermophilic or mesophilic LUCA hypotheses. In addition, I have provided an interpretive framework for the patterns, in the form of the two proposed rootings of the tree of life (bacterial and eukaryotic). Since the archaeal rooting is not seriously considered, this is omitted, but the interpretations would overlap with those for the bacterial rooting.

Importantly, the formal interpretations given in table 1 are expected to be very limited in terms of hypothesis testing, since these consider only a single trait, whereas for thermophily, many traits contribute to this. Using the RNA world as an outgroup for the mesophilic LUCA hypothesis greatly aids interpretation, but there are several possible sources of potential conflict. The simplest would be that a trait contributing to thermophily in archaea and bacteria was also found in eukaryotes, and no evidence of horizontal transfer was detected (scenario 2a). For this data to overturn the conclusion that the LUCA was mesophilic would require that the RNA world dataset, the absence of circular genomes in eukaryotes and Forterre's reverse gyrase data [Forterre 1995] can also be explained within this new context. Indeed, given that the observation in scenario 2a relies on a lack of evidence for horizontal transfer, the simplest interpretation of the data would be that detection of such an ancient horizontal transfer is beyond the limits of current methods. Another is that the origin of the trait was mesophilic, and that it was simply coopted during adaptation to elevated temperatures.

Another complication is demonstrated by scenario 12, where, taking only the trait described, support for a mesophilic or thermophilic LUCA is root-dependent.

The data from whole genome comparisons strongly suggests that eukaryote genomes are chimeric, with operational genes (*sensu* Rivera et al. [1998]) being of bacterial origin [Ribeiro & Golding 1998; Rivera et al. 1998; Horiike et al. 2001]. On the model described in this thesis, the distributional data are best interpreted as the result of transfer of endosymbiont genes to the nucleus, with subsequent widespread replacement of proto-eukaryotic orthologues (scenarios 2b & 3b in table 1). The implication is as follows. If the reductive evolution model described in this thesis is correct, on the order of 50% of genes in eukaryotes are bacterial, and therefore had a hot history under thermoreduction. These genes fit largely into the operational class. With a hot LUCA and an archaeal-bacterial fusion origin for the eukaryote lineage, 100% of eukaryotic genes would be prokaryotic in origin, and would all retain evidence of a hot history. Thus, testing the thermoreduction hypothesis would require looking at the approximately 50% of genes which are argued to be most closely related to archaeal genes, that is, informational genes (*sensu* Rivera et al. [1998]).

Table 1. Interpreting trait distributions within the framework of a thermophilic or mesophilic LUCA, under bacterial or eukaryote rootings of the tree of life.

	Traits[a]			*HT*[b]	*Bacterial rooting*		*Eukaryote rooting*	
	B	***A***	***E***		***Thermophilic (T)LUCA***	***Thermoreduction (M)LUCA***	***Thermophilic (T)LUCA***	***Mesophilic (M)LUCA***
Scenario 0	1	2	-	-	Uninformative on simple parsimony alone - B-A convergence favours rejection.	Uninformative on parsimony alone - but RNA world as outgroup supports (M)LUCA.	B-A convergence precludes interpretation. Ancestral thermophily not supported.	Uninformative on simple parsimony alone - but RNA world as outgroup supports (M)LUCA.
Scenario 1a (Not observed)	1	2	2	✗	Falsified, only A-E MRCA a thermophile, Traits 1&2 suggest B&A thermophily convergent.	Falsified, thermophilic origin for eukaryotes.	Trait 2 suggests (T)LUCA, but trait 1 must be explained by NOR[c] (not testable)	Falsified, thermophilic origin for eukaryotes.
Scenario 1b (Not observed)	1	2	2	✓	Not informative, for A-E MRCA, HT obscures ancestral state. Reconciling convergence of traits 1&2 requires NOR - untestable.	If HT is A→E, not inconsistent with thermoreduction/ (M)LUCA. E→A not consistent with either.	Not informative for LUCA, HT obscures ancestral state. Reconciling convergence of traits 1&2 requires NOR - untestable.	A→E, not inconsistent with thermoreduction/ (M)LUCA. E→A not consistent with either.
Scenario 2a (Not observed)	1	1	1	✗	(T)LUCA supported	(M)LUCA falsified	(T)LUCA supported	(M)LUCA falsified

	Traits[a]			***HT***[b]	***Bacterial rooting***		***Eukaryote rooting***	
	B	***A***	***E***		***Thermophilic (T)LUCA***	***Thermoreduction (M)LUCA***	***Thermophilic (T)LUCA***	***Mesophilic (M)LUCA***
Scenario 2b (Organellar glu mischarging, NH_4-dep. CP synthesis)	1	1	1	✓	Not informative, HT obscures ancestral state.	If B/A→E, not inconsistent with thermoreduction. N.B. For organellar functions, cf 3b	Not informative, HT obscures ancestral state.	If B/A→E, not inconsistent with thermoreduction. N.B. For organellar functions, cf 3b
Scenario 3a (Not observed)	1	2	1	✖	B-E MRCA (LUCA) a thermophile, A-B convergence requires NOR.	Falsified, thermophilic origin for eukaryotes.	B-E MRCA (LUCA) a thermophile, A-B convergence requires NOR.	Falsified, thermophilic origin for eukaryotes.
Scenario 3b (cf 2b)	1	2	1	✓	Not informative, B-E HT obscures ancestral state. Convergence of thermophily favours rejection.	B→E, not inconsistent with (M)LUCA; endosymbiont hypothesis gives additional test. cf Scenario 0.	Not informative, HT obscures ancestral state. Convergence favours rejection.	B→E, not inconsistent with thermoreduction; endosymbiont hypothesis gives additional test. cf Scenario 0.
Scenario 4a	1	1	-	✖	(T)LUCA supported on simple parsimony. N.B. Conclusion is dependent on correct rooting.	(M)LUCA rejected on simple parsimony. N.B. Rejection is dependent on correct rooting.	Uninformative - thermophilic trait only on one side of the root, cannot establish if it is ancestral.	Uninformative on simple parsimony, alone - but RNA world as outgroup supports (M)LUCA.

	Traits[a]			*HT*[b]	***Bacterial rooting***		***Eukaryote rooting***	
	B	***A***	***E***		***Thermophilic (T)LUCA***	***Thermoreduction (M)LUCA***	***Thermophilic (T)LUCA***	***Mesophilic (M)LUCA***
Scenario 4b (Reverse gyrase, *T. maritima* genome)	1	1	-	✓	Uninformative - HT obscures ancestral state.	HT between A&B not inconsistent with (M)LUCA rooted with RNA world dataset.	Uninformative - HT obscures ancestral state. But thermophily only on one side of root - T(LUCA) not supported.	HT between A&B not inconsistent with (M)LUCA rooted with RNA world dataset.
Scenario 5	-	-	3	-	Uninformative under all scenarios with only simple parsimony - not possible to establish whether the trait arose specifically in eukaryotes after divergence from the prokaryote lineages. RNA world dataset provides an exception as traits predate LUCA.			
Scenario 6a (Not observed)	3	-	3	✖	Reject	Consistent with (M)LUCA, but not thermoreduction	Reject	Consistent with (M)LUCA, but not thermoreduction
Scenario 6b (Not observed)	3	-	3	✓	Not supported.	Not inconsistent	Not supported.	Not inconsistent
Scenario 7a (Not observed)	-	3	3	✖	Reject. A-E MRCA mesophilic.	On simple parsimony, cannot establish ancestral state. Consistent with (M)LUCA, but not thermoreduction, using RNA world dataset.	Reject	Consistent with (M)LUCA, but not thermoreduction

	Traits[a]			***HT***[b]	***Bacterial rooting***		***Eukaryote rooting***	
	B	***A***	***E***		***Thermophilic (T)LUCA***	***Thermoreduction (M)LUCA***	***Thermophilic (T)LUCA***	***Mesophilic (M)LUCA***
Scenario 7b (Not observed)	-	3	3	✓	Not supported.	On simple parsimony, cannot establish ancestral state. Consistent with (M)LUCA, but not thermoreduction, using RNA world dataset.	Not supported	HT obscures ancestral state. Using RNA world as outgroup, is consistent with (M)LUCA, but not thermoreduction
Scenario 8 (Not observed)	1	1	3	✗	Supported on simple parsimony. Conclusion is root-dependent.	Rejected on simple parsimony. Conclusion is root-dependent.	Prokaryotic MRCA thermophilic, uninformative for (T)LUCA	(M)LUCA supported using RNA world as outgroup. Prokaryotes monophyletic, thermoreduction supported.
Scenario 9	1	1	3	✓	Uninformative - HT of trait 1 obscures ancestral state.	Not inconsistent with (M)LUCA, using RNA world as outgroup.	Uninformative - HT of trait 1 obscures ancestral state.	(M)LUCA supported using RNA world as outgroup.

	Traits[a]			*HT*[b]	*Bacterial rooting*		*Eukaryote rooting*	
	B	*A*	*E*		*Thermophilic (T)LUCA*	*Thermoreduction (M)LUCA*	*Thermophilic (T)LUCA*	*Mesophilic (M)LUCA*
Scenario 10 (Convergent glu mischarging in archaea-bacteria, direct charging in eukaryotes)	1	2	3	-	Reject on simple parsimony - thermophily convergent, not ancestral. NOR possible, but not demonstrable.	Not inconsistent with (M)LUCA, thermophily as derived.	Reject on simple parsimony - thermophily convergent, not ancestral. Even considering NOR, cannot show thermophily as ancestral.	(M)LUCA supported using RNA world as outgroup, thermophily derived.
Scenario 11 (Not observed)	3	1	3	✖	Reject on simple parsimony.	(M)LUCA supported on simple parsimony, thermoreduction only in archaea.	Reject on simple parsimony.	(M)LUCA supported on simple parsimony, thermoreduction only in archaea.
Scenario 12 (Direct gln charging in G+ bacteria)	3	1	3	✓	Uninformative - HT of trait 3 obscures ancestral state. For gln charging, ancestral state in B is 1, on simple parsimony (& without HT of 1) (T)LUCA supported. Conclusion is root-dependent.	HT from E→B, consistent with endosymbiont hypothesis, but ancestral state obscured. For gln charging, ancestral state in B is 1, on simple parsimony (& without HT of 1) (M)LUCA rejected. Conclusion is root-dependent.	Uninformative - HT of trait 3 obscures ancestral state. For gln charging, ancestral state in B is 1, simple parsimony uninformative.	HT from E→B, consistent with endosymbiont hypothesis, but ancestral state obscured. For gln charging, ancestral state in B is 1, using RNA world dataset, (M)LUCA supported.

	Traits[a]			HT[b]	Bacterial rooting		Eukaryote rooting	
	B	A	E		Thermophilic (T)LUCA	Thermoreduction (M)LUCA	Thermophilic (T)LUCA	Mesophilic (M)LUCA
Scenario 13 (Not observed)	1	3	3	✖	Uninformative, cannot establish ancestral state on simple parsimony.	RNA world outgroup supports (M)LUCA, but not thermoreduction for Archaea.	Simple parsimony, RNA world outgroup support (M)LUCA, but not thermoreduction for Archaea.	RNA world outgroup supports (M)LUCA, but not thermoreduction for Archaea.
Scenario 14	1	3	3	✓	Uninformative, HT obscures ancestral state in A/E.	RNA world outgroup supports (M)LUCA, but not thermoreduction for Archaea, unless NOR - untestable.	HT obscures ancestral state.	HT obscures ancestral state.

[a] Each number represents an independent trait, not related to any of the others by common descent. Red numbers: traits contributing to thermophily. Blue numbers: mesophilic traits. B - Bacteria, A - Archaea, E - Eukaryotes.

[b] HT is short for horizontal transfer.

[c] NOR: Non-orthologous replacement.

Glutamine usage.

As a free amino acid, glutamine is relatively more thermolabile than when incorporated into a peptide chain [reviewed in Greenstein & Winitz 1961]. Early studies described deamidation of free glutamine at higher levels than asparagine on boiling with magnesia [see Chibnall & Westall 1932; Greenstein & Winitz 1961 and references therein], and heating of glutamine at 100°C for 2-3 hours at a range of pH values resulted in extensive deamidation [Chibnall & Westall 1932; Vickery et al. 1935]. Gilbert et al. [1949], measured non-enzymatic deamidation of glutamine in the presence of a range of anions at various concentrations, pH and temperature. Non-enzymatic deamidation of glutamine is extensive in the presence of phosphate (at pH8, and 37°C). Near complete deamidation can be seen within 48 hours at 47°C in the presence of phosphate. Glutamine does not appear to possess an optimal pH for stability, but at extremes of pH, deamidation is greater. However, added phosphate results in increased deamidation at increasing pH, and decreasing deamidation at decreasing pH.

To measure the effect of temperature, Gilbert et al. [1949] made digests with 0.1M glutamine and 0.8M phosphate in buffer at pH8, and incubated these at either 47.4°C or 37°C. After 1.5 hours, 10.6μM and 4.8μM ammonia as liberated at these respective temperatures, and after 3.25 hours, 21.0 and 9.2μM ammonia was liberated respectively (of a total of 90μM for complete deamidation). The temperature coefficient for glutamine in the presence of phosphate is approximately 2 for a difference of 10°C.

These data suggest that glutamine instability should present a significant problem for even moderately thermophilic organisms (i.e. living above 50°C), especially given the greater instability of this amino acid in the presence of phosphate. Indeed there are indications that this may be the case, and that the examination of glutamine usage will shed light on the competing hypotheses of thermoreduction and a thermophilic LUCA.

Glutamine is a major nitrogen donor in eukaryote metabolism, but predicted to be at such low concentrations in thermophiles that ammonia is expected to be used in its place [Papers 2&4]. One example from the hyperthemophilic archaeon *Pyrococcus furiosus* is carbamoyl phosphate synthesis via an ammonia-dependent pathway, as opposed to the standard glutamine-dependent pathway [Legrain et al. 1995]. Another example is glutamate mischarging [Ibba et al. 1997], where glutamate is charged to glutaminyl-tRNA then amidated to form glutamine. This has the effect of making glutamine synthesis the final step before incorporation into protein, suggesting that this is an adaptation to a high temperature environment [Poole et al. 1998]. In support of thermoreduction, mischarging is found in eubacteria, archaea, and eukaryote organelles, but not the cytoplasm [Ibba et al. 1997, Ibba & Söll 2001]. While mischarging of glutamate to glutaminyl-tRNA can be argued to be a thermoadaptation, in those organisms examined to date, the nitrogen donor for amidation of mischarged glutamate is glutamine [Ibba & Söll 2001]! That glutamine is the nitrogen donor in such cases is consistent with glutamine being favoured over ammonia under 'permissive' conditions (no hyperthermophilic pathways have been examined as yet). It also serves to exemplify that looking for relics of thermoreduction will not be a trivial exercise.

Glutamine-dependent metabolism would be broadly classed as falling within the operational class of genes, so might be expected to have been replaced by genes from the endosymbiont. However, while I have argued in this thesis for a general replacement of proto-eukaryote operational othologues by endosymbiont genes, there will be exceptions. I suggest that glutamine-dependent metabolism would be one example, since, if the endosymbiont pathways are ammonia-dependent, they will not be able to supplant the original proto-eukaryotic genes because they cannot utilise glutamine. Indeed, a clear case is in the different pathways for synthesis of carbamoyl phosphate in the cytosol and in the mitochondrion of eukaryotic cells. The cytosolic class of enzyme is glutamine-dependent, while the mitochondrial class is ammonia dependent [Legrain et al. 1995].

Comparative metabolic databases such as WITS [http://wit.mcs.anl.gov/WIT2/] and KEGG [http://www.genome.ad.jp/kegg/], genome data, and the biochemical literature can be searched for all pathways where glutamine and/or ammonia act as nitrogen donors, to look for evidence of past thermophily. Furthermore, pathways involving other thermolabile metabolites such as carbamoyl phosphate [Van de Casteele et al. 1997] will also be examined. It is worth emphasising that while the example of carbamoyl phosphate synthesis represents a clear-cut case, this is not always to be expected. In thermophilic organisms, one should find mechanisms of adaptation to metabolite thermolability. However, in mesophilic prokaryotes, it is signatures of past thermophily that are important, and these will not necessarily be as clear as expected for comparisons between extant hyperthermophiles and eukaryotes. An example is that Gram negative bacteria have a direct pathway for glutaminyl-tRNA charging. In this case however, it is has been argued by several authors that this is as a result of horizontal transfer from a eukaryote source [Lamour et al. 1994, Handy & Doolittle 1999]. Upon readaptation to mesophilic temperatures, free glutamine can become available intracellularly, so glutamine-dependent pathways could potentially replace ammonia-dependent pathways.

One problem with thermolability studies is that these are generally carried out *in vitro*. This helps in establishing the physicochemical properties of a molecule, but this alone is not necessarily informative in all cases, as exemplified by the use of carbamoyl phosphate in hyperthermophiles. That this metabolite is used in an organism such as *Pyrococcus furiosus* might be considered anomalous if it were not known that metabolite channelling protects carbamoyl phosphate from being degraded [Van de Casteele 1997]. In the case of glutamine, it is therefore of great interest to establish its stability intracellularly when present as a free amino acid, and moreover, to establish the fates of the ammonia and glutamate moieties in thermophilic and hyperthermophilic organisms.

Nuclear Magnetic Resonance (NMR) studies provides the sort of resolution required, and have already been used in this way to examine intracellular free amino acid dynamics in archaea (e.g. Robertson et al. 1992). The advantage for the study of intracellular glutamine is that it would be possible to label both the nitrogen (^{15}N), which is released subsequent to deamidation, as well as ^{13}C-label the glutamine [see Lundberg et al. 1990, for review]. In this way, the fates of both moieties could be examined, in particular, making use of the fact that enzymatic deamidation of glutamine yields glutamate, while non-enzymatic deamidation yields pyrrolidonecarboxylic acid [Chibnall & Westall 1932; Vickery et al. 1935; Greenstein & Winitz 1961]. It should also be possible to establish whether glutamine is directly incorporated into protein. Likewise, the fate of free ammonia and glutamate and could be examined to see whether these are coincorporated into protein.

Some data are available on intracellular glutamine concentration in the archaeon *Methanobacterium thermoautotrophicum*, based on NMR studies of nitrogen assimilation [Choi et al. 1986, Choi & Roberts 1987]. *M. thermoautotrophicum* can utilise glutamine, urea or ammonia as sole nitrogen source. Choi & Roberts [1987] report that when cells were grown on [δ-^{15}N] glutamine as sole nitrogen source, intracellular concentrations of this amino acid were too low to be detectable, yet glutamine was reported to be stable for several days in the presence of an anaerobic cell extract (though the incubation temperature was not described). The authors conclude that their data best support the existence of an efficient glutamine permease for uptake, coupled with glutamate synthase. They rule out the presence of a glutaminase on the basis of the stability of glutamine when incubated with cell extract. This explanation requires the glutamate synthase to be located at the membrane, coupled to the permease. It has alternately been suggested that non-enzymatic degradation in the cell medium prior to uptake is also a possibility that has been suggested [Friedman & Thauer 1987].

The genome sequence of *M. thermoautotrophicum* [Smith et al. 1997] sheds some light on these conflicting positions. No glutaminase was detected in the published annotation, consistent with Choi & Roberts' conclusion. An ABC transporter for glutamine is present, and enzymes such as glutamine synthetase and glutamate synthase are also detected, again consistent with Choi & Roberts [1987]. Nevertheless, ammonium transporters are also present, so, under some conditions, ammonia liberated from glutamine would be taken up. Indeed, given that *M. thermoautotrophicum* has been documented to grow at temperatures ranging from 40-70°C [Smith et al. 1997, and references therein], it would be interesting to examine the fate of extracellular glutamine at various temperatures, labelling with both ^{13}C and ^{15}N.

While such studies were outside the scope of this thesis, the use of NMR spectroscopy for examining intracellular glutamine concentrations and fate are technically feasible, and together with genome data, will provide a rich source of data of relevance to studies of the nature of the LUCA. This would serve to establish substrate usage and/or preference where ammonia and glutamine may be interchangeably utilised, and under a range of growth temperatures.

Other thermolabile metabolites can also be examined. Forterre [1996] points out that ATP is thermolabile, and that its use is avoided in hyperthermophilic archaea, which make use of ADP or PP_i for energy storage in glycolysis. Nevertheless, given that ATP stores more energy than these other two energy cofactors, it would not necessarily be an argument against thermoreduction to find that these are now used in non-hyperthermophilic archaea - it may simply reflect selection for ATP over the other two cofactors at lower temperatures. There are a number of key metabolites and coenzymes which are thermolabile at hyperthermophilic growth temperatures, such as NAD, pyridoxal phosphate and acetyl phosphate [see table 1 in Daniel & Cowan 2000], yet these are nevertheless present in hyperthermophiles, suggesting mechanisms for prevention of thermodegradation are present [Daniel & Cowan 2000].

Another thermolabile metabolite is carbamoyl phosphate, which is a ubiquitous intermediate in the synthesis of arginine and pyrimidines. Interestingly, it is subject to metabolite channelling (where a metabolite is moves through a channel in a multienzyme complex - the intermediates are not released but instead move along the channel to the next active site), which has the effect of stabilising it at high temperatures [Van de Casteele et al. 1997]. If thermoreduction has occurred, channelling ought to be found in mesophilic prokaryotes, but not necessarily in eukaryotes, assuming that orthologous gene replacement has not occurred.

RNA thermoadaptation.

A pivotal study on RNase P RNA in bacteria established evidence for past thermophily in *E. coli*, where the optimum temperature for operation of this RNA is 50°C, well above the growth temperature of *E. coli* [Brown et al. 1993]. While this was interpreted as evidence for a thermophilic LUCA, eukaryote RNase P RNAs were not compared. This comparison is necessarily difficult since eukaryote RNase P RNA is not catalytically active in the absence of its cognate proteins [Kirsebom & Altman 1999]. A broader study of RNase P, as well as other RNAs can be approached by examining frequency of mismatches and non-canonical base-pairs in helices, percent pairing, percent G-C pairing, and other parameters that were shown to impact on RNA thermostability. In their original analysis, Brown et al. [1993] established that these parameters were more central to thermostability than GC content.

Secondary structure melting profiles for RNA are reasonably accurately predicted by theoretical approaches [Moulton et al. 2000], allowing a wide range of

RNAs to be tested. It is noteworthy that thermostability of methylated RNA, as found in hyperthermophiles, will be underpredicted by such analyses [Kowalak et al. 1994; Noon et al. 1998].

There is little likelihood that horizontal transfers will confound such an analysis. In the case of RNase P, the differences between eukaryotic and prokaryotic RNase P RNPs is significant, and notably, the protein composition is completely different [Altman & Kirsebom 1999]. The same is expected for other ubiquitous RNAs that could be compared, such as the signal recognition particle (srp) RNA [Stroud & Walter 1999; Eichler & Moll 2001].

Protein thermoadaptation.

Several studies have compared proteins between thermophilic and mesophilic prokaryotes [McDonald et al. 1999; Haney et al. 1999; McDonald 2001], finding that serine, asparagine, glutamine, threonine and methionine tend to be reduced in the proteins of thermophiles, while isoleucine, arginine, glutamate, lysine and proline residues tend to be increased. No composition asymmetry analyses have yet been carried out across all three domains. Testing thermoreduction by examining protein thermoadaptation requires such an analysis. Concern has been raised that the effects of thermophily on protein composition would be obscured by effects such as G+C content and environment-specific effects [McDonald 2001]. However, studies to date have focused on quite narrow datasets, and the expectation is that temperature effects should be identifiable from other effects by examining a broad range of proteins from a broad range of organisms and looking for trends common to the entire dataset. Furthermore, by distinguishing physicochemical properties from the outset, it may be possible to carry out a more specific analysis than previous analyses which have concentrated on composition asymmetries at all sites [Haney et al. 1999, McDonald et al. 1999]. For instance, glutamine is more thermolabile when incorporated into a peptide chain, whilst the opposite is true for asparagine [Greenstein & Winitz 1961]. Das & Gerstein [2001], who carried out a comparison of 12 genomes across all three domains reported, among other things, that thermophilic proteins tend to have reduced amounts of glutamine and asparagine compared to mesophilic proteins.

A large scale analysis is likely to be necessary in order to be able to distinguish between fluctuations in individual proteins and a consistent signal. The work would need to be carried out separately for informational and operational genes, and this would be interesting in itself, since, if a signature has been maintained for this length of time, it ought to be seen for eukaryote operational genes, but not for informational genes. The data reported by Das & Gerstein [2000] suggests that it will be possible to examine amino acid composition to see if a signature of past thermophily is detectable in mesophilic prokaryotes, though the signal may be weak. There are a number of other physicochemical factors that impact on protein thermostability, such as prevalence of salt bridges [Das & Gerstein 2000], increased hydrogen bonding, shortening of loops and helix dipole stabilisation [Jaenicke & Bohm 1998]. However, the emerging consensus is that it is the combination of a number of factors which contributes to thermostability, and rather than observing clear common differences between proteins from mesophilic and thermophilic organisms, there appears to be multiple routes to protein stability. Individual proteins may differ substantially in terms of properties which contribute to thermostability [Jaenicke & Bohm 1998].

Thermoreduction: once or twice?

One question which this proposed work might be able to answer is whether thermoreduction has occurred once, implying that the prokaryotes are a monophyletic group, or whether it has happened twice, once for archaea and once for bacteria. This question could not be approached using the data described in this thesis, hence the use of the terms prokaryote and eukaryote, even though it has been accepted throughout that this may not reflect a phylogenetic grouping.

If thermoreduction has happened twice, convergent thermoadaptations should be observed between archaea and bacteria. Interestingly, circular genomes may be such an example, as the origins of replication, and the associated replication proteins in archaea and bacteria are best explained as having evolved independently [Myllykallio et al. 2000]. Another example is the presence of a lipid monolayer in thermophilic archaea as opposed to the bilayer found in bacteria and eukaryotes. Likewise, the latter two groups possess ester-linked lipids, while archaea possess more stable ether-linked lipids [reviewed in Daniel & Cowan 2000]. Ether-linked lipids are however found in some thermophilic bacteria, making for a more complex picture. Forterre [1996] has nevertheless argued that, especially given the presence of lipid monolayers only in thermophilic archaea, mechanisms of membrane stability have evolved independently in thermophilic archaea and bacteria.

Because thermophily is not a single trait, but a descriptive term for lifestyle that comprises many traits, it is imperative to systematically look at a large number of traits to establish whether thermophily arose once or twice. The thermophilic LUCA hypothesis requires that thermophily arose only once, irrespective of whether the bacterial or eukaryotic rooting is correct (though in both cases, it also requires evidence for past thermophily in eukaryotes). Finding that archaea and bacteria have independently adaptated to thermophily would be a falsification of the thermophilic LUCA hypothesis. Thermoreduction is consistent with thermophily evolving once, if the prokaryotes are monophyletic, whereas it is only consistent with independent adaptation to thermophily by archaea and bacteria under the bacterial rooting.

Horizontal transfer would blur these distinctions (see table 1), so not only is it necessary to establish the nature of thermophily in archaea and bacteria, but whether

1) common traits have been subject to horizontal transfer, and 2) whether the entire domain possesses such traits, or whether these are restricted to only some regions of the tree. In addition, it will be necessary to establish 3) whether traits unique to members of a domain are ubiquitous, and 4) whether there have been within-domain transfers. In this respect, correctly rooting the tree is not going to be sufficient to establish which of the two hypotheses are correct.

I fully expect that horizontal transfer from endosymbiont to nucleus (but also other instances of horizontal transfer such as between archaeal and bacterial hyperthermophiles - [Aravind et al. 1999; Kyrpides & Olsen 1999; Nelson et al. 1999; Forterre et al. 2000]) will complicate testing of the thermoreduction and thermophilic LUCA hypotheses, but on current expectations, it ought to be possible to establish such cases, and exclude them from the analysis [see Nara et al. 2000, for a discussion of this with respect to pyrimidine biosynthesis, a carbamoyl phosphate-dependent pathway]. Furthermore, I expect the examination of glutamine-dependent and ammonia-dependent pathways, and pathways involving other thermolabile metabolites will not be straightforward to interpret. Additionally, while the presence of alternate pathways solve the problem of thermolabile metabolites (e.g. NAD(P), acetyl phosphate, ATP - see Daniel & Cowan [2000]), and should be detectable by comparative genome analyses, other mechanisms of thermoadaptation (such as the high catalytic efficiency of phosphoribosyl anthranilate isomerase as a means of abrogating phosphoribosyl anthranilate thermolability [Sterner et al. 1996]) will not be amenable to bioinformatic analyses.

The implicit assumption in the proposed work is that, at temperatures typical of mesophiles, thermoadaptations are not selectively disadvantageous, and hence may persist as relics. Certainly it is not expected that this should be the case for all such adaptations, such that it may be impossible to establish whether traits found in extant thermophiles date back to the common ancestor of one or both prokaryotic lineages. Nevertheless, if a general trend emerges from several unrelated datasets (metabolism, RNA and protein) it might be possible to use these data to test the thermoreduction hypothesis, and might enable an examination of the question of the monophyly of the prokaryotes independent of phylogenetic analyses.

In spite of the potential pitfalls of such analyses, and consistent with thermoreduction twice rather than once (and thus thermoreduction over the thermophilic LUCA hypothesis) is the recent demonstration that the pathways of glutamate mischarging in archaea and bacteria have arisen by independent recruitment of enzymes involved in amino acid metabolism, as opposed to being related by descent [Tumbula et al. 2000]. At the time of writing papers 4&5, these data were not available, but are important for two reasons. First, they are consistent with other data (see above) that suggest archaeal and bacterial thermophiles are convergent rather than divergent, and second, they overturn the other major interpretation of mischarging, that it is a relic from the evolution of the genetic code [Di Giulio 2000].

In concluding this thesis, I wish to underscore the inherent difficulties with phylogenetic approaches to understanding the nature of the LUCA, and the evolution of prokaryotes and eukaryotes. A tree with three major groupings holds too little information to be able to establish the nature of the organism at the root. The tree is useful for systematic analyses, and should in principle be able to establish whether the prokaryotes are monophyletic, assuming that a reliable phylogenetic signal can be recovered. Nevertheless, even with the correct topology, it is not possible to infer the nature of the ancestor from the topology alone. As I have shown in this thesis, the traditional bacterial rooting could be correct and yet the LUCA could still be eukaryote-like. The 3-domain tree simply does not hold enough information to establish the direction of evolutionary change, that is, to determine ancestral and derived states. These have simply been assumed because the notion that prokaryotes predate eukaryotes has been taken as given. This thesis has provided an alternative to phylogenetic approaches which reveals their weakness in terms of inferring the nature of the LUCA, and which has challenged the prokaryote dogma.

A final note: the origin of DNA.

In the sections dealing with the RNA world, I have considered the question of the origin of protein synthesis in depth but the question of the origin of DNA is only briefly mentioned. I have examined this question in depth [Poole et al. 2000, 2001a,b], but this work is not included in this thesis. Since I discuss this question in Poole et al. 1998 and 1999, I shall briefly state my major conclusions for completeness.

Most significantly, I conclude that the RNA to DNA transition had to occur subsequent to the advent of genetically-encoded protein synthesis, and that the low coding capacity of RNA as a genetic material presents a major problem in understanding how ribonucleotide reduction arose. This is counter to earlier suggestions, notably by Benner et al. [1989], who argue for an RNA world with a DNA genome, with protein synthesis arising later. In their account, Benner et al. [1989] argue for deoxyribonucleotide synthesis from glyceraldehyde-3-phosphate and acetaldehyde as opposed to ribonucleotide reduction.

Ribonucleotide reduction is the only known pathway for *de novo* synthesis of deoxyribonucleotides, and requires protein radical chemistry. In all three classes of ribonucleotide reductase, a radical is generated and subsequently transferred to a cysteine residue, forming a thiyl radical. For this reduction to take place, a mechanism for radical generation, storage and specific control and transfer to the substrate is required. Other than radical generation, these roles could not be carried out by RNA, which is non-specifically cleaved by radicals, so either catalytic proteins predate DNA [Poole et al. 2000], or an alternative, chemically simpler but unobserved pathway existed [Benner et al. 1989]. The latter is chemically feasible, given the presence of the degradative pathway in deoxyribonucleotide salvage. Indeed it was considered the most likely route for deoxyribonucleotide synthesis, prior to the denonstration that the sole route was ribonucleotide reduction [reviewed in Reichard 1989]. However, the pathway suggested by Benner et al. [1989] may simply not have been 'discovered' by evolution [Poole et al. 2001b], given that evolution is analogous to tinkering, not engineering [Jacob 1977]. Indeed, the evolution of ribonucleotide reduction as opposed to a simpler reaction may have been contingent on the presence of an established pathway for ribonucleotide synthesis and thus availability of ribonucleotides, with acetaldehyde and glyceraldehyde-3-phosphate perhaps not being available in large enough quantities for deoxyribonucleotide synthesis via this route [Reichard 1989, Poole et al. 2001b].

If ribonucleotide reduction was prerequisite for the advent of DNA synthesis, then this causes problems for the RNA world theory, since, in the absence of proofreading and repair, RNA is not expected to be capable to maintaining sufficiently large amounts of genetic information for proteins of the complexity of ribonucleotide reductase to emerge (see the Darwin-Eigen cycle in Poole et al. [1999]). I have proposed a possible solution to this problem [Poole et al. 2000].

In brief, I have argued that post-replicative 2'-*O*-methylation of RNA could have provided a more stable genetic material than RNA, having some, but not all of the features that makes DNA a more stable information storage molecule than RNA. Post-replicative 2'-*O*-methylation would eliminate the tendency for RNA to self-cleave because the modification renders the reactive 2'-hydroxyl group of the ribose inactive. Consequently, 2'-*O*-methyl RNA would potentially be a more stable genetic material than unmodified RNA. Incomplete ribose modification, post-replicative versus pre-replicative modification (deoxyribonucleotides are synthesised prior to DNA synthesis) and deep groove hydrophobicity resulting from extensive methylation make 2'-*O*-methyl RNA inferior to DNA, and hence provide selection for replacement. 2'-*O*-methylation of RNA is found in all three domains of life and has been argued to be a feature of the RNA world, and the theory describes a plausible scenario for the recruitment of 2'-*O*-methylation from functional RNAs to genomic RNA [Poole et al. 2000].

I have also examined the second stage in the RNA to DNA transition, where uracil was replaced by thymine [Poole et al. 2001a]. The substrates for ribonucleotide reduction are ATP, CTP, GTP and UTP (some ribonucleotide reductases make use of diphosphate substrates), forming dATP, dCTP, dGTP and dUTP. dUTP is subsequently converted to dTTP, and this indirect pathway suggests that the U to T transition occurred subsequent to the replacement of ribose by deoxyribose. However the standard argument for the replacement of uracil by thymine in the evolution of DNA is flawed. It suggests that replacing uracil with thymine eliminated the problem of cytosine deamination to uracil, permitting deaminations to be identified since uracil was no longer native to DNA. This requires evolutionary forethought, since thymine only provides a means of recognising deaminations, not of repairing them. If repair evolved before thymine replaced uracil, this is also problematic, as there would then be no selection for thymine to replace uracil. This problem, and the question of what selection pressure might account for the U to T transition is discussed in Poole et al. [2001a].

Concluding remark.

In this thesis, a case has been made for continuity from the RNA world through to the emergence of the three domains, eukaryotes, bacteria and archaea. The major conclusion is that using the RNA world as outgroup to root the tree of life suggests that eukaryotes have retained more ancestral features than prokaryotes. From this conclusion it is then possible to examine the differing modes of evolution in prokaryotes and eukaryotes, and a biological basis for these differences is described. The work is based on the established principles of the error threshold (Eigen limit), the relationship between rate of diffusion and catalytic efficiency, the physicochemical properties of RNA, r- and K-selection and standard evolutionary theory. I believe it provides a significant improvement over previous studies on early evolution in that a wide range of phenomena can be explained consistently, as opposed to being treated as unrelated problems. Importantly, while the model described may not be correct, it is testable, as described above.

References.

Altman S, Kirsebom L: Ribonuclease P. In: Gesteland R, Cech T, Atkins J, eds. The RNA World, 2nd Ed. New York: Cold Spring Harbor Laboratory Press 1999, pp 351-380.

Aravind L, Tatusov RL, Wolf YI, Walker DR, Koonin EV: Evidence for massive gene exchange between archaeal and bacterial hyperthermophiles. Trends Genet 1998, 14, 442-444.

Benner SA, Ellington AD, Tauer A: Modern metabolism as a palimpsest of the RNA world. Proc Natl Acad Sci USA 1989, 86, 7054-7058.

Brion P, Westhof E: Hierarchy and dynamics of RNA folding. Annu Rev Biophys Biomol Struct 1997, 26, 113-137.

Brown JW, Haas ES, Pace NR: Characterization of ribonuclease P RNAs from thermophilic bacteria. Nucleic Acids Res 1993, 21, 671-679.

Chibnall AC, Westall RG: The estimation of glutamine in the presence of asparagine. Biochem J 1932, 26, 122-132.

Choi B-S, Roberts JE, Evans JNS, Roberts MF: Nitrogen metabolism in *Methanobacterium thermoautotrophicum*: a solution and solid-state ^{15}N NMR study. Biochem 1986, 25, 2243-2248.

Choi B-S, Roberts MF: ^{15}N-NMR studies of *Methanobacterium thermoautotrophicum*: comparison of assimilation of different nitrogen sources. Biochim Biophys Acta 1987, 928, 259-265.

Daniel RM, Cowan DA: Biomolecular stability and life at high temperatures. Cell Mol Life Sci 2000, 57, 250-264.

Das R, Gerstein M: The stability of thermophilic proteins: a study based on comprehensive genome comparison. Funct Integr Genomics 2000, 1, 76-88.

Di Giulio M: The RNA world, the genetic code and the tRNA molecule. Trends Genet 2000, 16, 17-19.

Eichler J, Moll R: The signal recognition particle of Archaea. Trends Microbiol 2001, 9, 130-136.

Friedman HC, Thauer RK: FEMS Microbiol Lett 1987, 40, 179-181.

Forterre P: A hot topic: the origin of hyperthermophiles. Cell 1996, 85, 789-792.

Forterre P, Bouthier De La Tour C, Philippe H, Duguet M: Reverse gyrase from hyperthermophiles: probable transfer of a thermoadaptation trait from archaea to bacteria. Trends Genet 2000, 16,152-154.

Galtier N, Tourasse N, Gouy M: A nonhyperthermophilic common ancestor to extant life forms. Science 1999, 283:220-221.

Gilbert JB, Price VE, Greenstein JP: Effect of anions on the non-enzymatic desamidation of glutamine. J Biol Chem 1949, 180, 209-218.

Greenstein JP, Winitz M: Glutamic acid and glutamine. In Chemistry of the Amino Acids. John Wiley and Sons, NY 1961, pp1929-1954.

Hallberg RL, Kraus KW, Hallberg EM: Induction of acquired thermotolerance in *Tetrahymena thermophila*: effects of protein synthesis inhibitors. Mol Cell Biol 1985, 5, 2061-2069.

Handy J, Doolittle RF: An attempt to pinpoint the phylogenetic introduction of glutaminyl-tRNA synthetase among bacteria. J Mol Evol 1999, 49, 709-715.

Haney PJ, et al.: Thermal adaptation analysed by comparison of protein sequences from mesophilic and extremely thermophilic *Methanococcus* species. Proc Natl Acad Sci USA 1999, 96:3578-3583.

Hiller R, Zhou ZH, Adams MW, Englander SW: Stability and dynamics in a hyperthermophilic protein with melting temperature close to 200 degrees C. Proc Natl Acad Sci USA 1997, 94, 11329-11332.

Horiike T, Hamada K, Kanaya S, Shinozawa T: Origin of eukaryotic cell nuclei by symbiosis of Archaea and Bacteria is revealed by homology-hit analysis. Nat Cell Biol 2001, 3, 210-214.

Ibba M, Curnow AW, Söll D: Aninoacyl-tRNA synthesis: divergent routes to a common goal. Trends Biochem Sci 1997, 22:39-42.

Ibba M, Söll D: The renaissance of aminoacyl-tRNA synthesis. EMBO Rep 2001, 2, 382-387.

Jacob F: Evolution and Tinkering. Science 1977, 196, 1161-1166.

Jaenicke R, Bohm G: The stability of proteins in extreme environments. Curr Opin Struct Biol 1998, 8, 738-748.

Kowalak JA et al.: The role of posttranscriptional modification in stabilization of transfer RNA from hyperthermophiles. Biochemistry 1994, 33:7869–7876.

Kyrpides NC, Olsen GJ: Archaeal and bacterial hyperthermophiles: horizontal gene exchange or common ancestry?/Aravind et al.: Reply. Trends Genet 1999, 15, 298-300.

Lamour V, Quevillon S, Diriong S, N'Guyen VC, Lipinski M, Mirande M: Evolution of the Glx-tRNA synthetase family: the glutaminyl enzyme as a case of horizontal gene transfer. Proc Natl Acad Sci USA 1994, 91, 8670-8674.

Legrain C, Demarez M, Glansdorff N, Piérard A: Ammonia-dependent synthesis and metabolic channelling of carbamoyl phosphate in the hyperthermophilic archaeon *Pyrococcus furiosus*. Microbiol 1995, 141, 1093-1099.

Lundberg P, Harmsen E, Ho C, Vogel HJ: Nuclear Magnetic Resonance Studies of Cellular Metabolism. Anal Biochem 1990, 191, 193-222.

McDonald JH, Grasso AM, Rejto LK: Patterns of temperature adaptation in proteins from *Methanococcus* and *Bacillus*. Mol Biol Evol 1999, 16:785-790.

McDonald JH: Patterns of temperature adaptation in proteins from the bacteria *Deinococcus radiodurans* and *Themus thermophilus*. Mol Biol Evol 2001, 18:741-749.

McMullin ER, Bergquist DC, Fisher CR: Metazoans in Extreme Environments: Adaptations of Hydrothermal Vent and Hydrocarbon Seep Fauna. Gravit Space Biol Bull 2000, 13, 13-23.

Miller SL, Bada JL: Submarine hot springs and the origin of life. Nature 1988, 334, 609-611.

Moulton V et al.: RNA folding argues against a hot-start origin of life. J MolEvol 2000, 51:416-421.

Myllykallio H, et al.: Bacterial mode of replication with eukaryotic-like machinery in a hyperthermophilic archaeon.Science 2000, 288, 2212-2215.

Nara T, Hashimoto T, Aoki T: Evolutionary implications of the mosaic pyrimidine-biosynthetic pathway in eukaryotes. Gene 2000, 257, 209-222.

Nelson KE, et al.: Evidence for lateral gene transfer between Archaea and Bacteria from the genome sequence of *Thermotoga maritima*. Nature 1999, 399, 323-329.

Noon KE, Bruenger E, McCloskey JA: Post-transcriptional modifications in 16S and 23S rRNAs of the archaeal hyperthermophile *Sulfolobus solfataricus*. J Bacteriol 1998, 180, 2883-2888.

Poole AM, Jeffares DC, Penny D: The path from the RNA world. J Mol Evol 1998, 46, 1-17

Poole A, Jeffares D, Penny D: Early evolution: prokaryotes, the new kids on the block. BioEssays 1999, 21, 880-9

Poole A, Penny D, Sjöberg B-M: Methyl-RNA: an evolutionary bridge between RNA and DNA? Chem Biol 2000, 7, R207-R216.

Poole A, Penny D, Sjöberg B-M: Confounded cytosine! Tinkering and the evolution of DNA. Nat Rev Mol Cell Biol 2001, 2, 147-151.

Poole AM, Logan DT, Sjöberg B-M: The evolution of the ribonucleotide reductases: much ado about oxygen. J Mol Evol 2001b, accepted.

Reichard P: Commentary on 'Formation of deoxycytidine 5'-phosphate from cytidine 5'-phosphate with enzymes from *Escherichia coli*' by P Reichard & L Rutberg. Biochim Biophys Acta 1989, 1000, 49-50.

Ribeiro S, Golding, GB: The mosaic nature of the eukaryotic nucleus. Mol Biol Evol 1998, 15, 779-788.

Rivera MC, Jain R, Moore JE, Lake JA: Genomic evidence for two functionally distinct gene classes. Proc Natl Acad Sci USA, 1998, 95, 6239-6244.

Robertson DE, Noll D, Roberts MF: Free amino acid dynamics in marine microorganisms. J Biol Chem,1992, 267, 14893-14901.

Rothschild LJ, Mancinelli RL: Life in extreme environments. Nature 2001, 409, 1092-1101.

Silva JL, Foguel D, Royer CA: Pressure provides new insights into protein folding, dynamics and structure. Trends Biochem Sci 2001, 26, 612-618.

Smith DR, Doucette-Stamm LA, Deloughery C, Lee H, Dubois J, Aldredge T, Bashirzadeh R, Blakely D, Cook R, Gilbert K, Harrison D, Hoang L, Keagle P, Lumm W, Pothier B, Qiu D, Spadafora R, Vicaire R, Wang Y, Wierzbowski J, Gibson R, Jiwani N, Caruso A, Bush D, Safer H, Patwell D, Prabhakar S, Mcdougall S, Shimer G, Goyal A, Pietrokovski S, Church GM, Daniels CJ, Mao J-I, Rice P, Nölling J, Reeve JN: Complete Genome Sequence of *Methanobacterium thermoautotrophicum* ΔH: Functional Analysis and Comparative Genomics. J Bact 1997, 179, 7135-7155.

Sterner R, Kleeman GR, Szadkowski H, Lustig A, Hennig M, Kirschner K: Phosphoribosyl anthranilate isomerase from *Thermotoga maritima* is an extremely stable and active homodimer. Protein Sci 1996, 5, 2000-2008.

Stroud RM, Walter P: Signal sequence recognition and protein targeting. Curr Opin Struct Biol, 1999, 9, 754-759.

Tumbula DL, Becker HD, Chang W-Z, Söll D: Domain-specific recruitment of amide amino acids for protein synthesis. Nature 2000, 407, 106-110.

Van de Casteele M et al.: Molecular physiology of carbamoylation under extreme conditions: what can we learn from extreme thermophilic microorganisms? Comp Biochem Physiol A 1997, 118, 463-473.

Vickery HB, Pucher GW, Clark HE, Chibnall AC, Westall RG: The determination of glutamine in the presence of asparagine. Biochem J 1935, 29, 2710-2720.

Appendix

Poole A & Penny D.

Does endosymbiosis explain the origin of the nucleus?

Nature Cell Biology 3, E173 (2001).

Letter in response to:

Horiike T, Hamada K, Kanaya S & Shinozawa T.

Origin of eukaryotic cell nuclei by symbiosis of Archaea and Bacteria is revealed by homology-hit analysis.

Nature Cell Biology 3, 210-214 (2001).