Article

The Tree of Life: Tracing the Genetic Pathway from the Last Universal Common Ancestor to Homo Sapiens (Part I)

Chris King*

ABSTRACT

This series of articles are fully referenced research reviews to overview progress in unraveling the details of the evolutionary Tree of Life, from life's first occurrence in the hypothetical RNA-era, to humanity's own emergence and diversification, through migration and intermarriage, using research diagrams and brief discussion of the current state of the art. The Tree of Life, in biological terms, has come to be identified with the evolutionary tree of biological diversity. It is this tree which represents the climax fruitfulness of the biosphere and the genetic foundation of our existence, embracing not just higher Eucaryotes, plants, animals and fungi, but Protista, Eubacteria and Archaea, the realm, including the extreme heat and salt-loving organisms, which appears to lie almost at the root of life itself. To a certain extent the notion of a tree based on generational evolution has become complicated by a variety of compounding factors. Gene transfer is not just vertical carried down the generations. There is also evidence for promiscuous incidences of horizontal gene transfer, genetic symbiosis, hybridization and even the formation of chimeras. This review will cover all these aspects, from the first life on Earth to Homo sapiens.

Part I of this article includes: 1. Introduction; 2. LUCA: The Universal Common Ancestor; and 3.Two or Three Domains of Life?

Key Words: tree of life, genetic pathway, common ancestor, Homo Sapiens, biological diversity.

1. Introduction

The Tree of Life, in biological terms, has come to be identified with the evolutionary tree of biological diversity. It is this tree which represents the climax fruitfulness of the biosphere and the genetic foundation of our existence, embracing not just higher Eucaryotes, plants, animals and fungi, but Protista, Eubacteria and Archaea, the realm, including the extreme heat and salt-loving organisms, which appears to lie almost at the root of life itself. To a certain extent the notion of a tree based on generational evolution has become complicated by a variety of compounding factors. Gene transfer is not just vertical carried down the generations. There is also evidence for promiscuous incidences of horizontal gene transfer, genetic symbiosis, hybridization and even the formation of chimeras. This review will cover all these aspects, from the first life on Earth to Homo sapiens. This series of articles are fully referenced research reviews to overview progress in unraveling the details of the evolutionary Tree of Life, from life's first occurrence in the hypothetical RNA-era, to humanity's own emergence and diversification, through migration and intermarriage, using research diagrams and brief discussion of the current state of the art.

* Correspondence: Chris King http://www.dhushara.com E-Mail: chris@sexualparadox.org

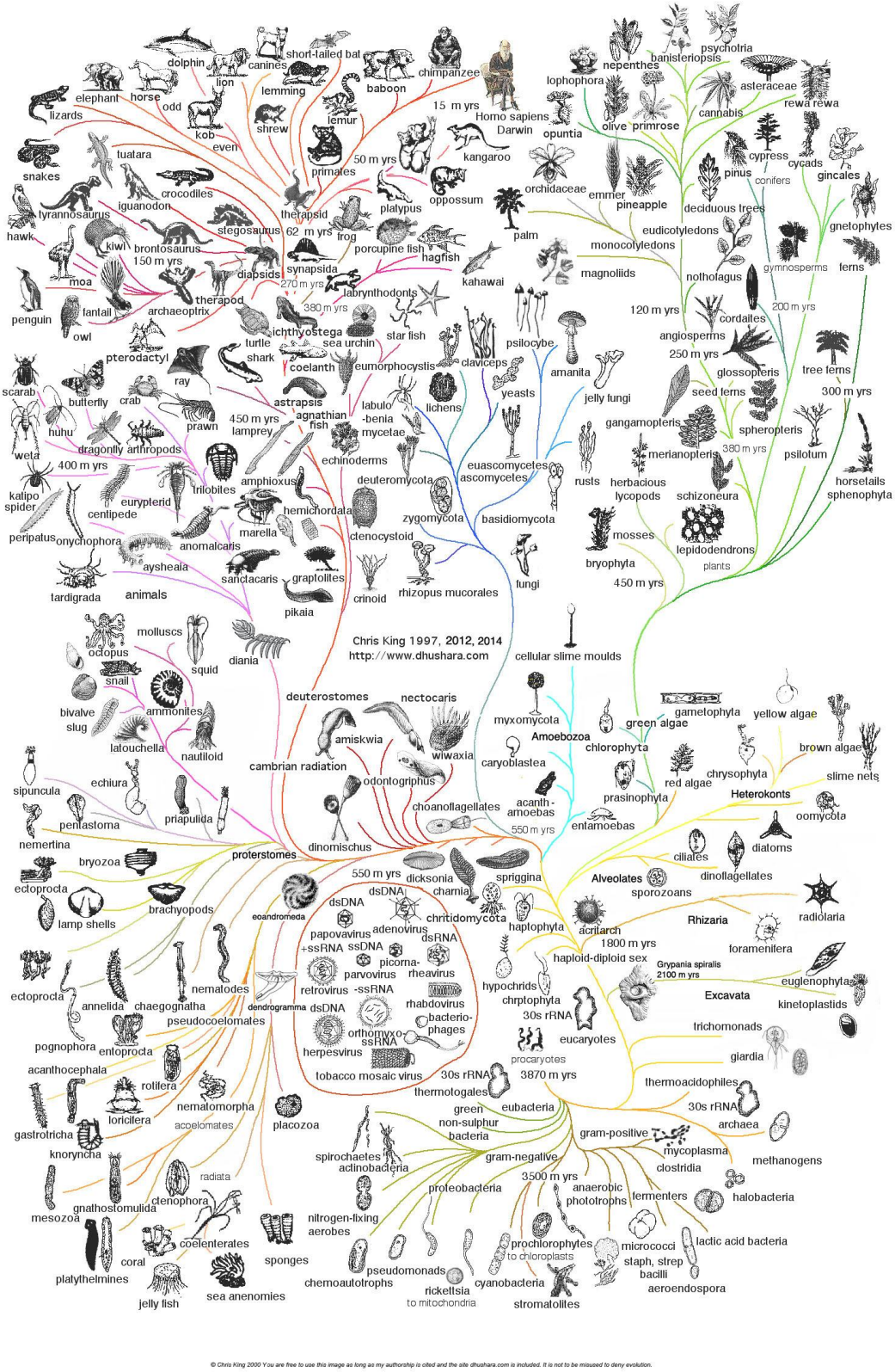

Fig 1: The Tree of Life

2. LUCA: The Last Universal Common Ancestor

Following a phase of biogenesis possibly based on cosmic symmetry-breaking (King 1978, 2004), based on spontaneous prebiotic RNA synthesis (Powner et. al. 2009, 2010) recent research suggests that the last universal common ancestor (LUCA) of all life on the planet may have arisen before the first cells, from a phase interface between alkaline hydrogen-emitting undersea vents and the archaic acidified iron-rich ocean (Martin and Russel 2003) in which differential dynamics in membranous micropores in the vents managed to concentrate polypeptides and polynucleotides to biologically sustainable levels (Baaske et. al. 2007, Budin et. al. 2009), giving rise to the RNA era, while at the same time providing a free energy source based on proton transport across membranous microcellular interfaces resulting from fatty acids also being concentrated above their critical aggregate concentration. The transition to enclosed cells is likely to have been in an active iron-sulphur reaction phase still present in living cells and associated with sodium-proton anti-porters activating ATP (Lane and Martin 2012, Lane 2009b), leading in turn to electron transport and some of the most ancient proteins, such as ferredoxin,

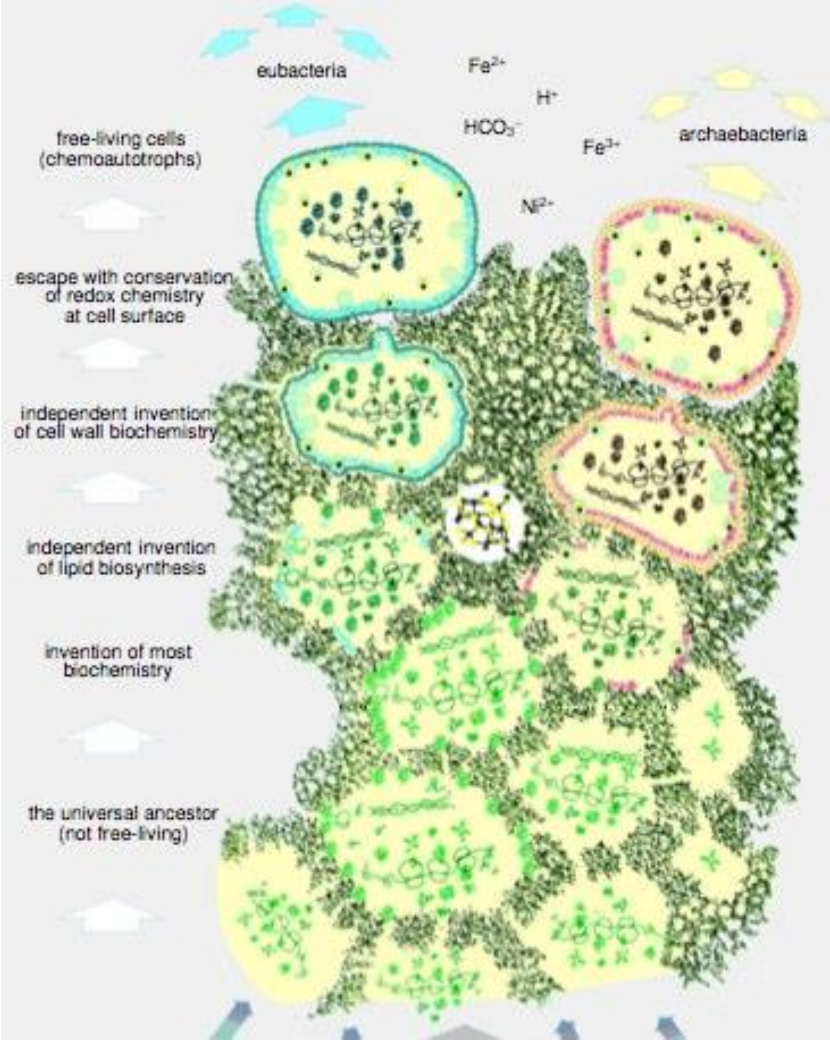

Fig 1a: Proposed scheme for the universal common ancestor (Martin and Russel 2003)

The universal common ancestor of the three domains of life may have thus been a proton-pumping membranous interface from which archaea and bacteria emerged as free-living adaptions. This is suggested by fundamental differences in their cell walls and other details of evolutionary relationships among some of the oldest genes.

Among the archaea, halobacteria still use a form of photosynthesis generating ATP from H^+ gradients generated by a rhodopsin protein and those in hydrothermal vents rely on Na^+-H^+ antiporters to generate ion gradients, and their membrane proteins, such as the ATP synthase, are compatible with gradients of sodium ions or protons (Lane and Martin 2012, Yong 2012).

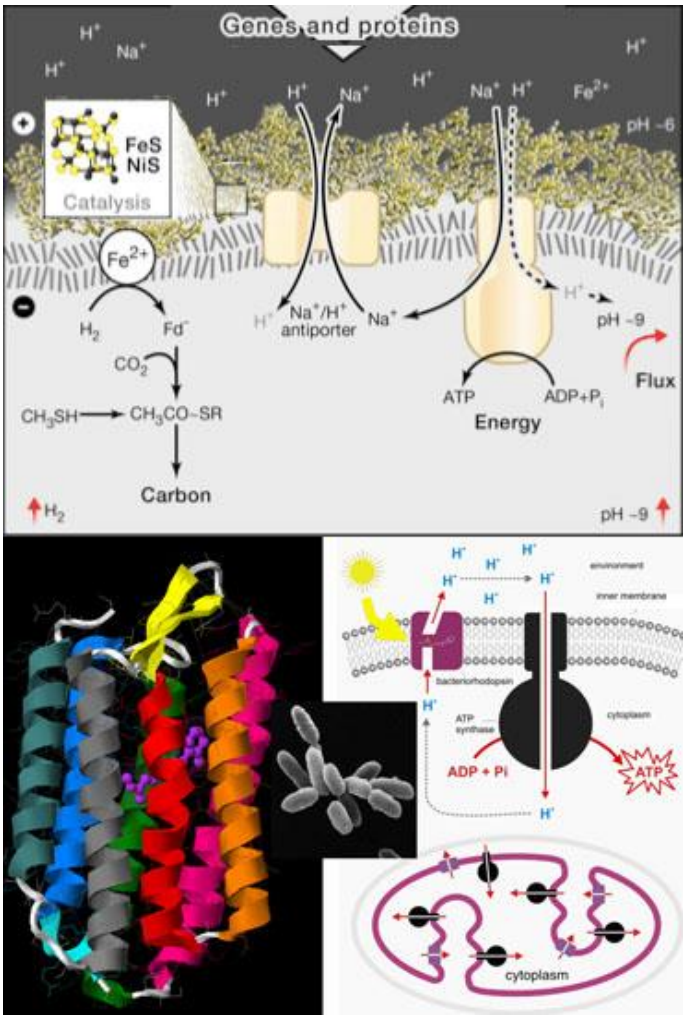

Fig 1b: (Above) founding metabolism based on Na^+-H^+ anti-transported, ATP synthetase and FeSNiS containing vents (Lane and Martin 2012). The extremely ancient origin of the rhodopsin family of heptahelical receptors can be seen from the ultra-primitive archael photosynthesis in Halobacteria, which relies on direct coupling between photo-stimulated chemiosmotic H^+ pumping and H^+ generated ATP formation, based on bacteriorhodopsin, which is heptahelical, uses a form of retinal and whose helices share a distant sequence homology with vertebrate rhodopsin (Ihara et al 1999).

It has also been proposed, on the basis of the highly-conserved commonality of transcription and translation proteins to all life, but the apparently independent emergence of distinct DNA replication enzymes in archaea/eucaryotes and eubacteria, that the last universal common ancestor had a mixed RNA-DNA metabolism based on reverse transcriptase, pinpointing it to the latter phases of the RNA era (Leipe et. al. 1999).

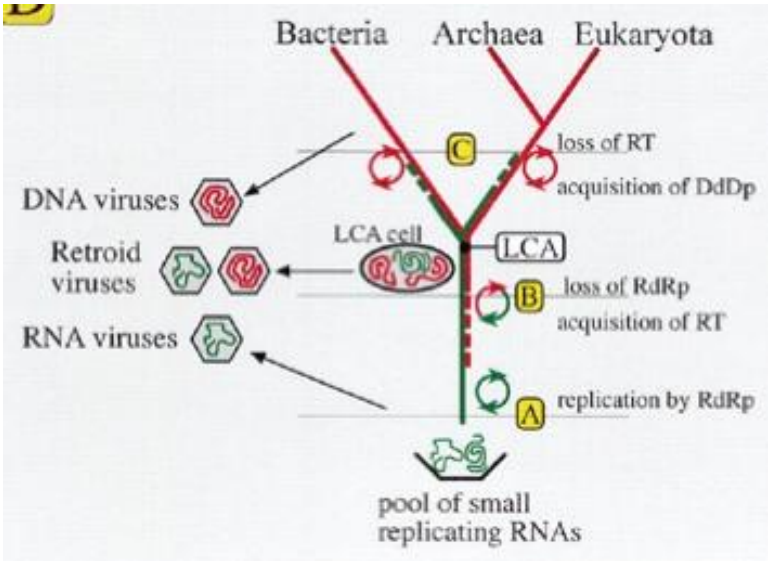

Fig 1c: Hypothetical branching and evolution of RNA and DNA replication machinery (Leipe et. al. 1999) suggests viruses were pivotal in the transition from RNA to DNA (see below)

To get a characterization of LUCA at the point it diversified into the three domains of life Archaea, Eucaryotes and Bacteria, one cannot rely on nucleotide gene sequences because these would have mutated beyond recognition, but amino acid sequences mutate more slowly because neutral mutations leave the amino acid sequence fixed and the tertiary folded structure of a protein is even more strongly conserved.

The validity of the RNA-era concept and the capacity for RNAs to be both replicating informational and active ribo-enzymes is emphasized by the continuing dependence of the ribosome on rRNA rather than the protein components demonstrated by the 3-dimensional realizations of the two subunits in fig 4, which show that the rRNA molecules are still carrying out the central task of protein assembly with only minor modification due to the 'chaperoning' proteins, despite 3.8 billion years of evolution.

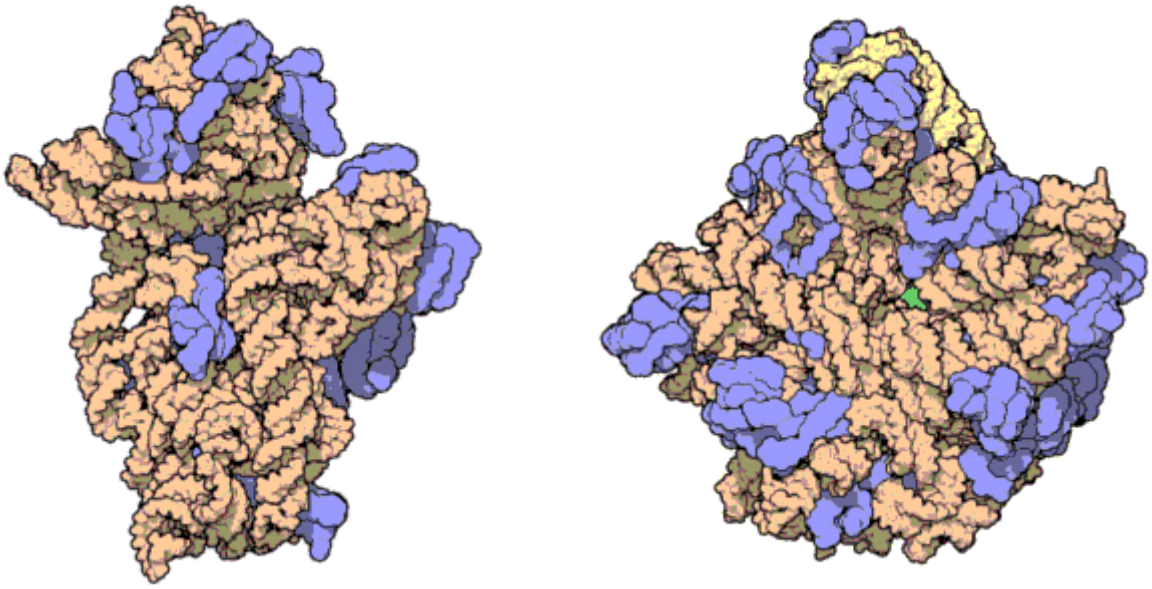

Fig 4: Small and large rRNA subunits of the eubacteia Thermus thermophilus and the archaeon Haloarcula marismortui. RNA orange and yellow, protein blue and active site green. (Wikipedia Ribosome).

Brooks et al. (2002) have found that the amino acids used in sections of genes common to life which are believed to originate with LUCA show amino acid distributions reflecting the relative abundance of such amino acids in primitive synthesis, indicating that the first translational genes used the amino acids which were spontaneously available.

One intriguing indication of the state of genetic translation in LUCA is the incorporation of selenocysteine into the genetic code. Selenoenzymes which contain selenocysteine as a genetically translated amino acid are essential to the three domains of life and source back to LUCA, despite the fact that the 21st coded amino acid selenocysteine could not be fitted into the genetic code. An ingenious piece of genetic software engineering evolved in which the amber stop codon UAG is overridden if the m-RNA possessses a motif called SECIS (selenocysteine insertion sequence) and selenocysteine is then inserted instead of termination and translation continues.

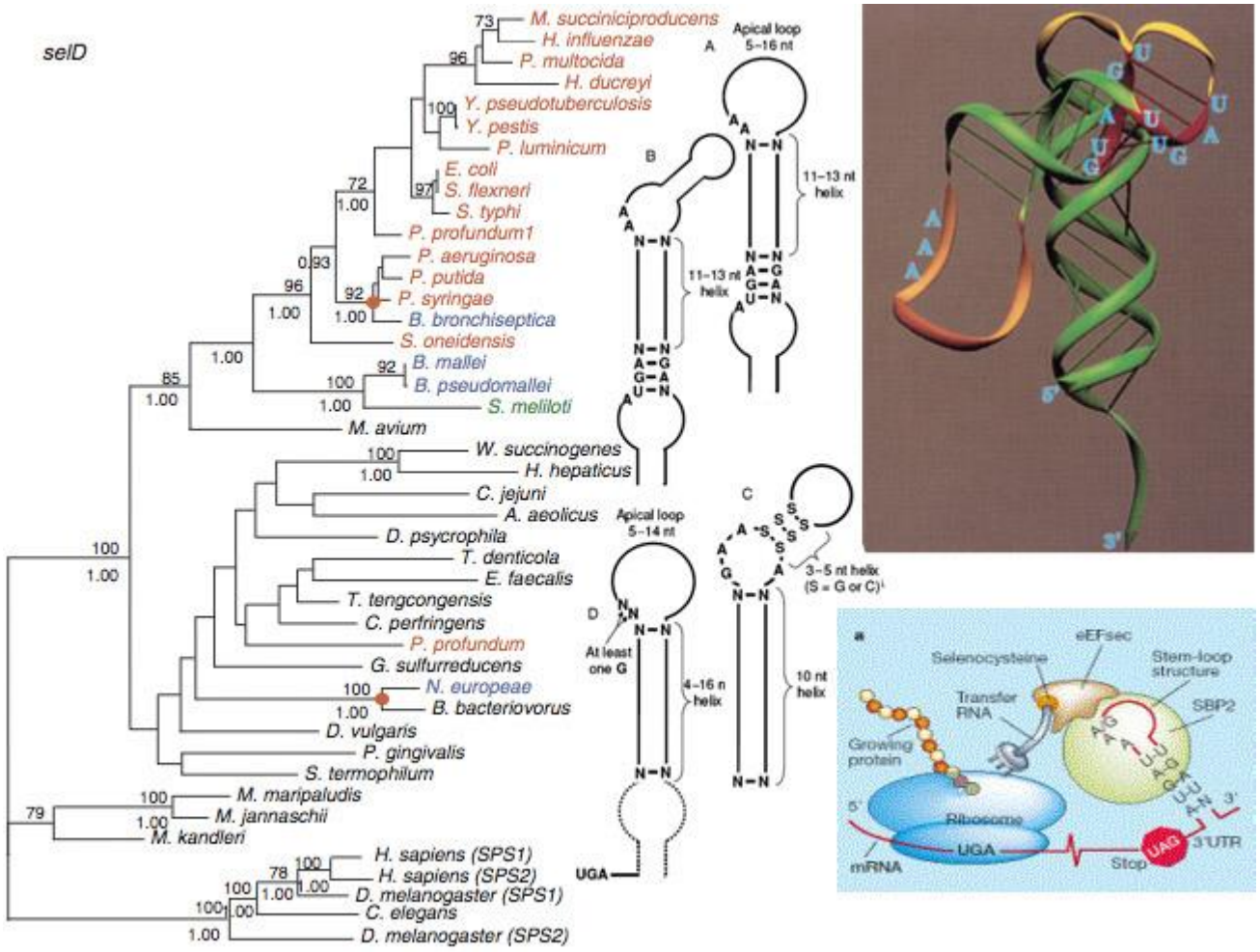

Fig 1c2: Left: Evolutionary tree of selenophosphate synthetase (Romero et al. 2005) spans the three domains of life. Centre: SECIS hairpins of archaea (A), bacteria (B) and corresponding eukaryote variants (C, D) (Moldave ed 2006). Top right: Tertiary structure of SECIS showing highly conserved regions (hot) (Walczak et al. 1996). Lower right: SECIS acts as an RNA-enzyme to attach the selenocysteine t-RNA to the nascent protein.

SECIS is an unusual hairpin loop structure which has varying forms in archaea and prokaryotes with both forms appearing in eucaryotes, but they have a common feature of a highly conserved hairpin loop forming an RNA translational catalyst, which literally takes over some of the ribosomal RNA function, binding to the selenocysteine t-RNA and coupling selenocysteine to the nascent protein chain, as shown in the above figure. It is clear that this unique piece of genetic software engineering evolved in LUCA because the wobble positions of three other essential amino acid t-RNAs, lysine, glutamine and glutamic acid (those with two wobble positions XAA-XAG, the fourth set being amber and ochre stop codons), all depend on a modified 2-seleno-uridine base to function and this has to be generated from selenophosphate, which in turn is generated by selenophosphate synthetase. As shown above left, this enzyme has an evolutionary tree extending back to LUCA confirming the obvious - that the genetic code cannot exist without the 21st software engineered amino acid selenocysteine!

To reconstruct the set of proteins LUCA could make, Kim and Caetano-Anollés (direct link) searched a database of proteins from 420 modern organisms, looking for structures that were common to all. Of the structures he found, just 5 to 11 per cent were universal, meaning they were conserved enough to have originated in LUCA. By looking at their function, they conclude that LUCA had enzymes to break down and extract energy from nutrients, and some protein-making equipment, but it lacked the enzymes for making and reading DNA molecules.

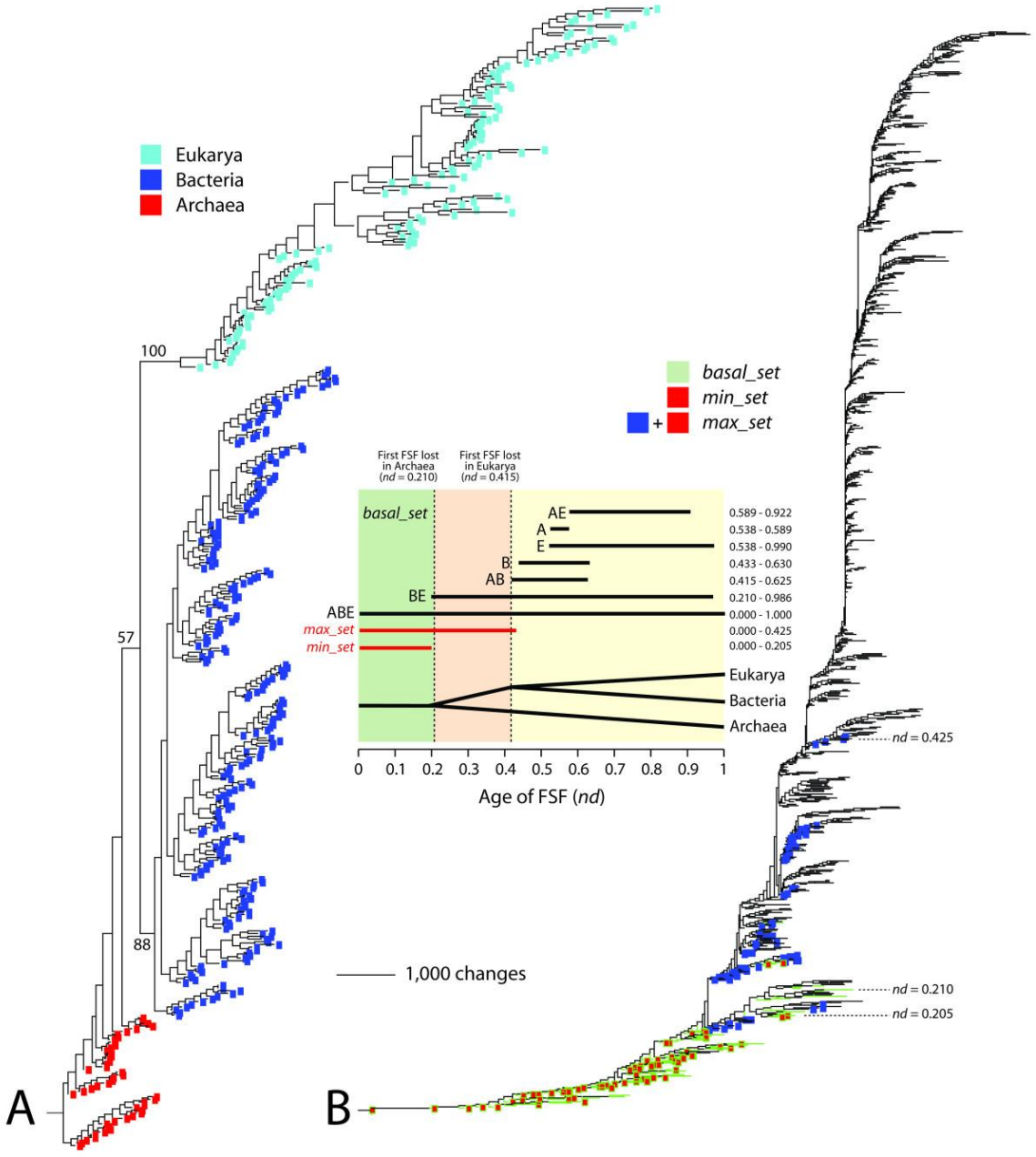

Fig 1d: Phylogenomic tree of proteomes describing the evolution of 420 FL organisms. phylogenomic study of protein domain structure in the proteomes of 420 free-living fully sequenced organisms. Domains were defined at the highly conserved fold superfamily (FSF) level of structural classification (Kim and Caetano-Anollés).

Organelles were thought to be the preserve of eukaryotes, but in 2003 researchers found an organelle called the acidocalcisome also occurred in bacteria. Caetano-Anollés' team has now found that tiny granules in some archaea are also acidocalcisomes, or at least their precursors. That means acidocalcisomes are found in all three domains of life, and date back to LUCA (Seufferheld et al. - direct link).

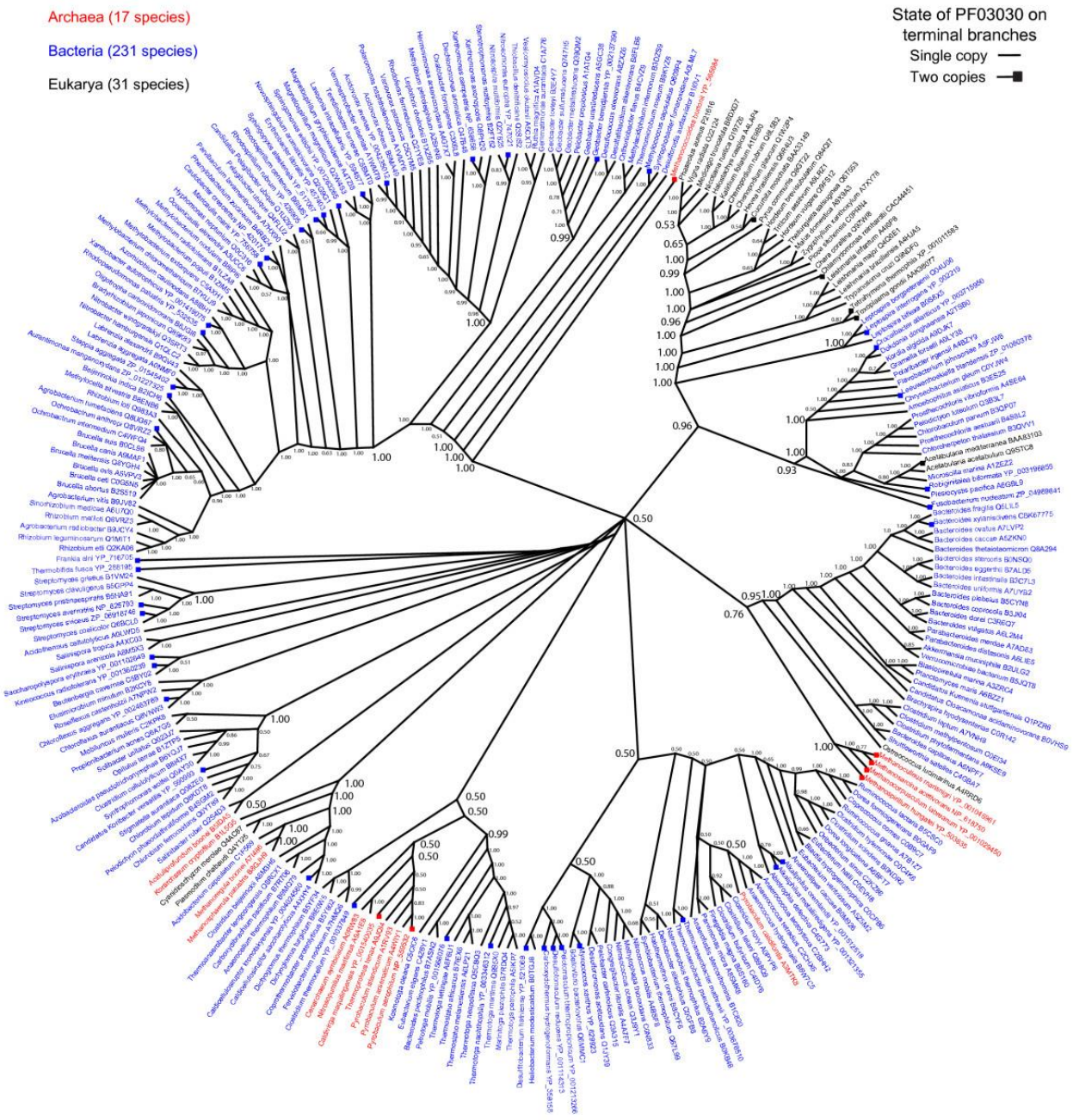

Fig 1e: Tangled web linking acidocalcisomes in existent archaea, bacteria and eucaryote species (Seufferheld et al.), overlaying electron micrographs of acidocalcisomes in Agrobacterium tumefaciens(a, b) and Methanosarcina acetivorans (c, d).

Acidocalcisomes were originally discovered in Trypanosomes (sleeping sickness and Chagas disease) but have since been found in *Toxoplasma gondii* (toxoplasmosis), *Plasmodium* (malaria), *Chlamydomonas reinhardtii* (a green alga), *Dictyostelium discoideum* (a slime mould), bacteria and human platelets. Their membranes contain a number of protein pumps and antiporters, including aquaporins, ATPases and Ca^{2+}/H^{+} and Na^{+}/H^{+} antiporters. Acidocalcisomes have been implied in osmoregulation. They were detected in vicinity of the contractile vacuole in Trypanosoma cruzi and were shown to fuse with the vacuole when the cells were exposed to osmotic stress. Presumably the acidocalcisomes empty their ion contents into the contractile vacuole, thereby increasing the vacuole's osmolarity. This then causes water from the cytoplasm to enter the vacuole, until the latter gathers a certain amount of water and expels it out of the cell.

LUCA may have used RNA rather than DNA, as there is no evidence LUCA possessed ribonucleotide reductases, which create the deoxy versions of ribonucleotides the building blocks of DNA (Lundin et al - direct link). Rather it appears these functions have been transferred from bacteria back to archaea by horizontal transfer on at least two separate occasions (arrows in fig 1e). Eucaryotes (mid green) would also have received theirs after LUCA diversification.

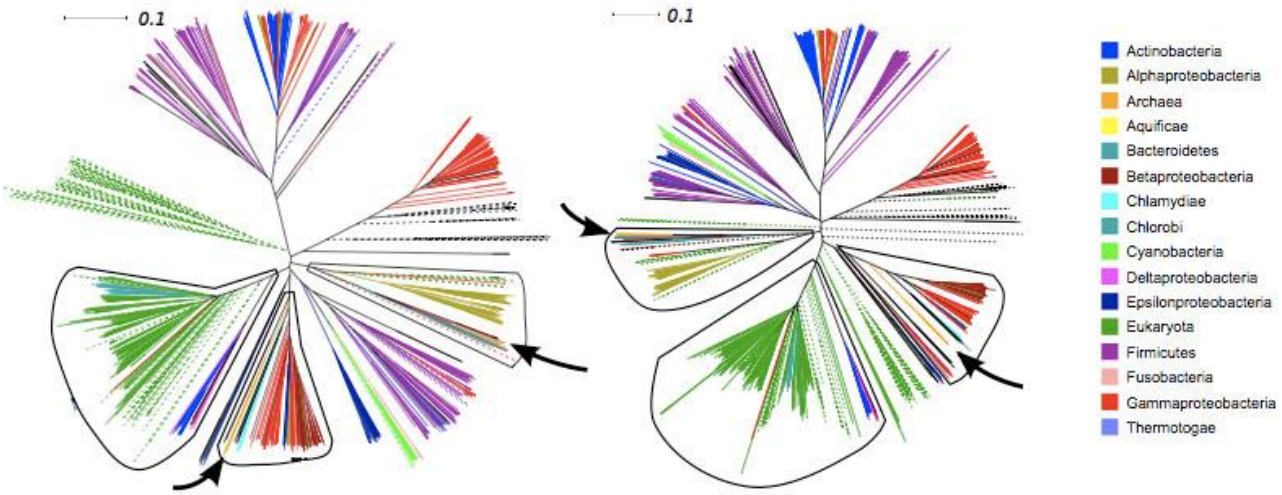

Fig 1f: Ribonucleotide reductase trees showing bacterial, eucaryote and archaeal branches, with evidence of two events of horizontal transfer from bacteria to archaea (arrows) after the diversification of LUCA (Lundin et al).

LUCA was a "progenote". Progenotes can make proteins using genes as a template, but the process is so error-prone that the proteins can be quite unlike what the gene specified. Both Di Giulio and Caetano-Anollés have found evidence that systems that make protein synthesis accurate appear long after LUCA. In order to cope, the early cells must have shared their genes and proteins with each other. Caetano-Anollés says the free exchange and lack of competition mean this living primordial ocean essentially functioned as a single mega-organism.

3. Two or Three Domains of Life?

Life today is informationally based on the sequences of the four bases A, G, T and C in DNA, with messenger copies of the genetic sequence in mRNA (with U replacing T) forming intermediates in the assembly of proteins, as the cell's primary active chemical and structural agents. This is achieved through a process of translation at the ribosome - a supra-molecular complex composed of some 50 chaperoning proteins surrounding a core composed of three rRNA units, fed by amino-acid coupled tRNAs. The RNAs carry out the essential function, supporting the idea that translation was at first a purely RNA-based process of protein construction. In line with this and other RNA fossils found particularly in Eukaryotes, it is widely believed that life began based on RNA, which shares both the capacity for complementary replication of DNA and the formation of 3-dimensional chemically reactive conformations, similar to proteins, after which the ribosome evolved, transferring the reactive burden on to proteins sequenced through the genetic code. Some time later, the informational genome was consolidated into more stable DNA.

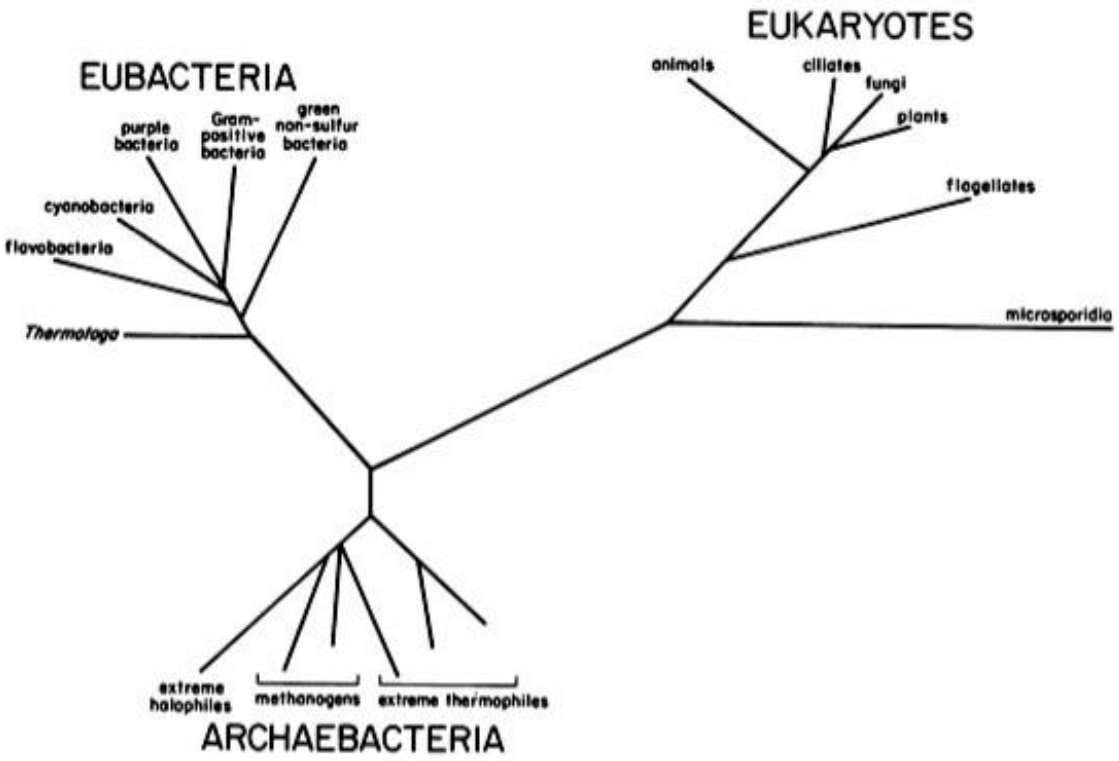

Fig 2: The initial tree of rRNAs shows three distinct founding domains

Originally the Bacteria and Archaea were thought to be one large diverse family of prokaryotes until Carl Woese (1977, 1978, 1987, 1990) and others investigated the evolutionary tree of ribosomal RNAs and found that there were three distinct founding evolutionary domains, then named eubacteria, archaebacteria along with the eukaryotes.

This gave the Eukaryotes a closer founding status as well, by contrast with the idea that the procaryotic bacteria came first and then, somehow the higher Eukaryote organisms with their complex cellular

structures, including among others - the endoplasmic reticulum, along with the nuclear envelope and Golgi apparatus - all parts of a common complex of internal membranous partitions - and the architecture of microtubules, including centrioles, and the Eukaryote flagellum, as well as the Eukaryotes endosymbiont mitochondria and chloroplasts.

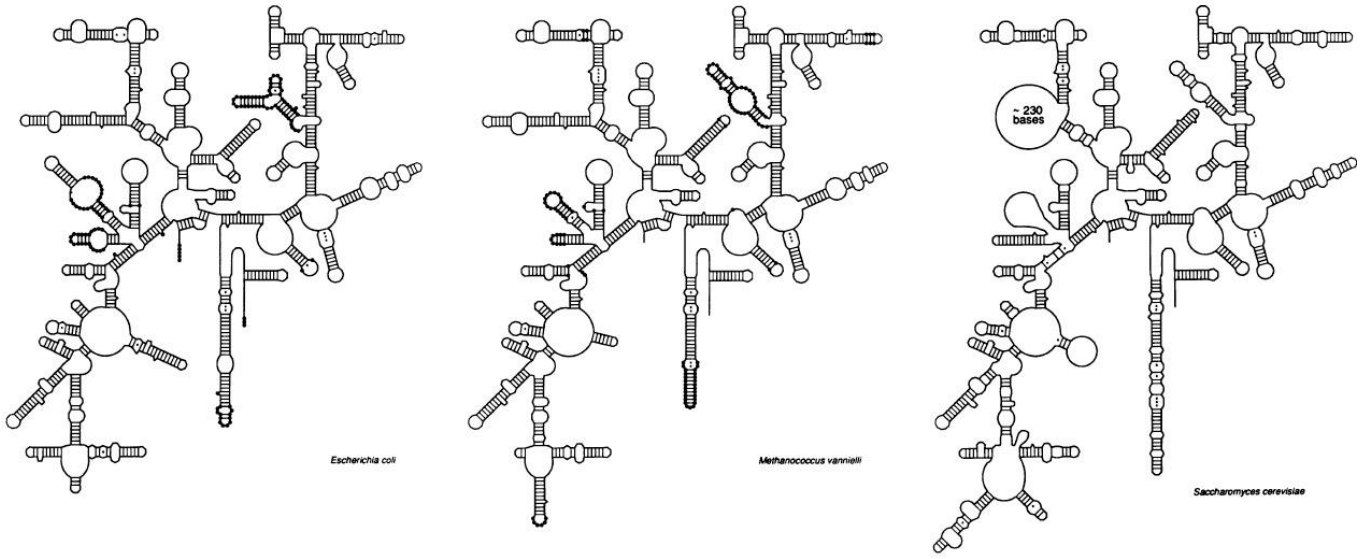

Fig 3: Key structural differences separating the larger rRNA units of the three domains (Woese 1987)

In addition to their evolutionary sequence divergence, the smaller 30s ribosomal RNAs of each domain, show distinct structural features characteristic of their own domain, but also emphasizing structural links between Bacteria and Archaea on the one hand and Archaea and Eukaryotes on the other, qualitatively confirming the central place of the Archaea in the divergence.

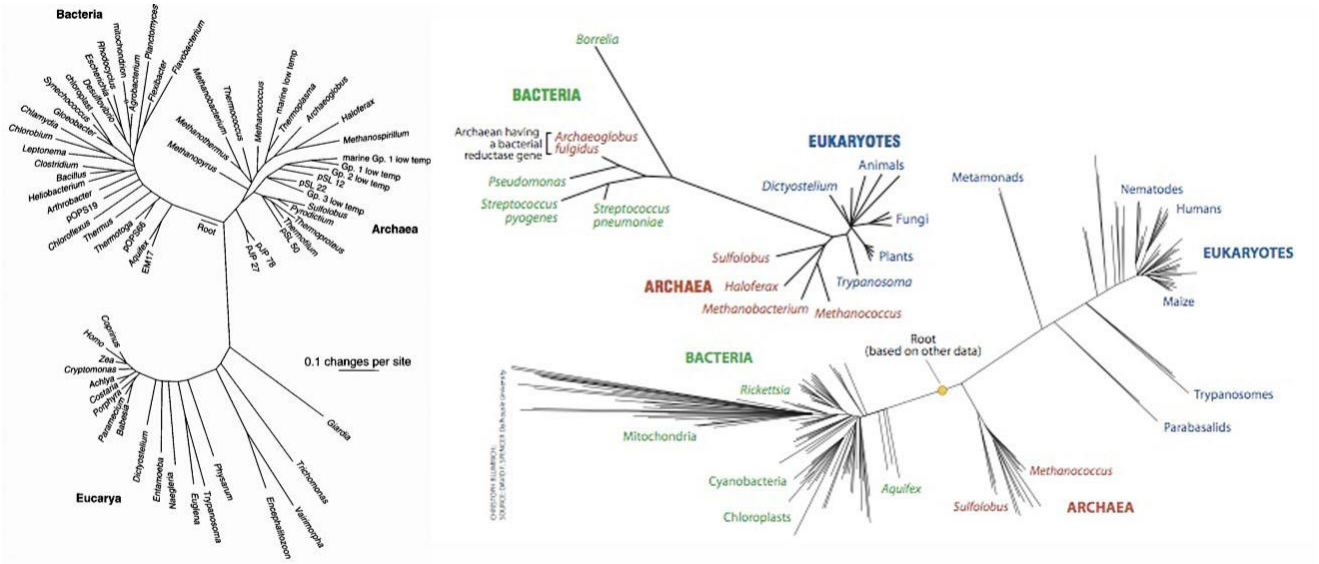

Fig 5: (a) Further elaboration of the rRNA tree (Pace 1997) (b) A third rRNA tree which suggests Archaea lie very close to the root is contrasted with that for the enzyme HMGCoA reductase (c), which also shows evidence of horizontal transfer to an Archaean (ex Doolittle 2000).

Norman Pace subsequently enlarged the scope and accuracy of the rRNA tree, including a greater diversity of organisms. This tree has become the basis of several other studies (see e.g. fig 11).

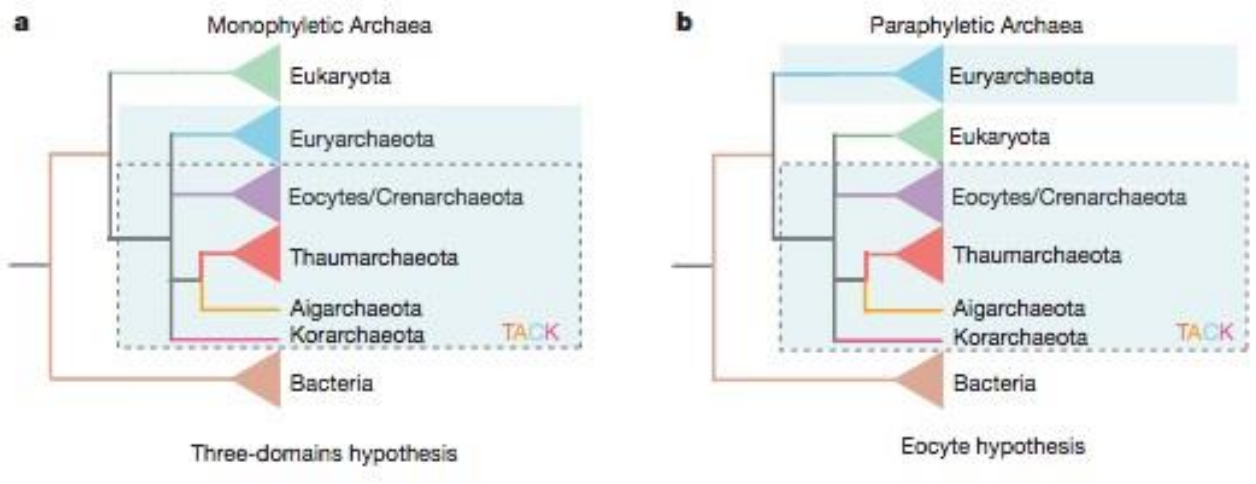

Fig 5b: Three domains (a) is contrasted with a recent version of the "eocyte" hypothesis (b) showing the eucaryotes emerging from the wider crenarcheota grouping (TACK) after divergence from euryarcheota, implying the amoeboid ancestor of the eucaryotes was an "eocyte" (Williams et al. 2013).

However James Lake (1988) had already challenged the notion of three domains, with an analysis claiming that the eucaryotes instead branched off form only one line of the archaea, the eocytes or chrenarcheota. This view has been confirmed by accumulating genetic studies (Williams & Embley 2014, Williams et al.2013, Foster, Cox & Embley 2009, Cox et al. 2008).

The Copernican principle asserts that the Earth is a typical rocky planet in a typical planetary system, located in an unexceptional region of a common barred-spiral galaxy, hence it is probable that the universe teems with complex life. This is supported to a reasonable extent by the discovery of an increasing number of planets including some putative "Goldilocks" zone planets where water would be liquid and life as we know it could potentially exist. Set against this, the "rare earth" hypothesis argues that the emergence of complex life requires a host of fortuitous circumstances, including a galactic habitable zone, a central star and planetary system having the requisite character, the circumstellar habitable zone, the size of the planet, the advantage of a large satellite, conditions needed to assure the planet has a magnetosphere and plate tectonics, the chemistry of the lithosphere, atmosphere, and oceans, the role of "evolutionary pumps" such as massive glaciation and rare bolide impacts, and whatever led to the still mysterious Cambrian explosion of animal phyla. This might mean that planets able to support a bacterial level of life are not so uncommon, but those supporting complex multicellular life might be.

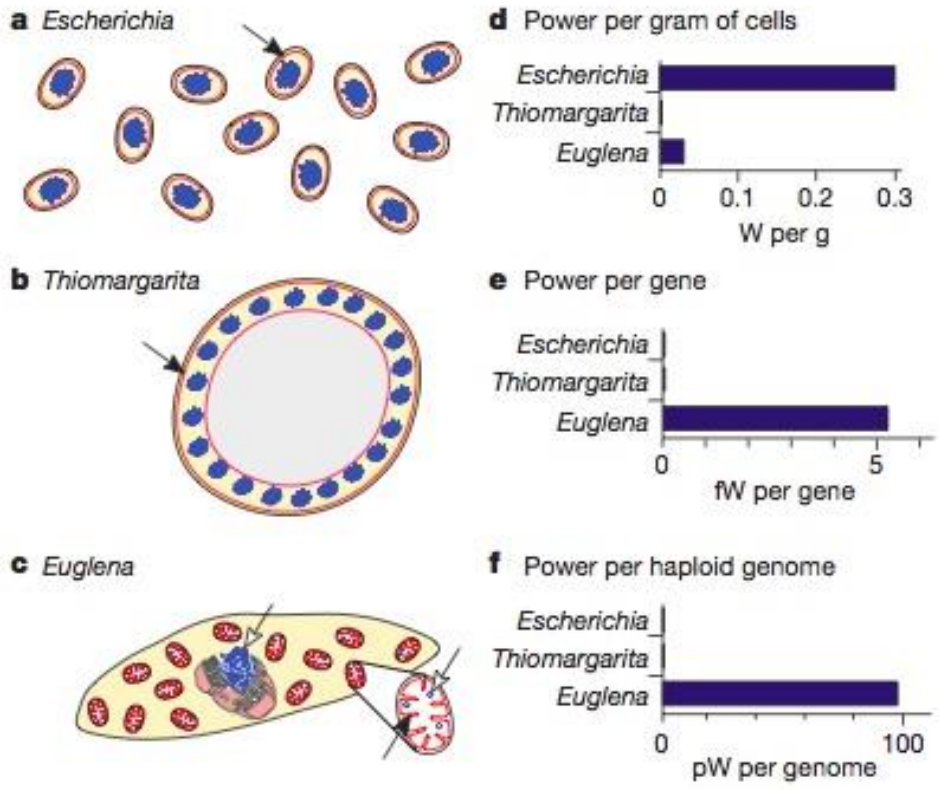

Fig 6: Metabolic power of eucarote cells per haploid genome and hence the capacity for genomic complexity depends on the rspiraqtory power of mitochondria (Lane and Martin).

Bringing this question to a pivotal crux in our context, the emergence of mitochondria as endosymbionts has been proposed to be a critical bottleneck which allowed complex life to evolve only once on Earth, because, only in this effectively fractal cellular architecture, can the membrane surface areas necessary to support the chemical reactions enabling the vastly larger number of genes in a complex organism's genome to maintain metabolic stability (Lane and Martin). Lane and Martin note "The cornerstone of eukaryotic complexity is a vastly expanded repertoire of novel protein folds, protein interactions and regulatory cascades. The eukaryote common ancestor increased its genetic repertoire by some 3,000 novel gene families. The invention of new protein folds in the eukaryotes was the most intense phase of gene invention since the origin of life. Eukaryotes invented five times as many protein folds as eubacteria, and ten times as many as archaea. Even median protein length is 30% greater in eukaryotes than in prokaryotes". Whether such endo-symbiosis is rare. or a common extreme of parasitic or predatory relationships would then determine how likely or unlikely complex life might be.

This massive increase in complexity remains obscure in the genetic and fossil records and requires some ingenious model construction to envisage how mitosis, meiosis, sexuality, the nuclear envelope, endoplasmic reticulum, cytoskeleton, and all the complexities of eucaryote regulation evolved. For a seminal work on this see (Cavalier-Smith 2010).

Regardless of this, Lane and Martin's metabolic approach explains neatly why there is little sign of any of these structures in any existing prokaryote. In effect endo-synbiosis created a completely new

energetic regime, in which the only niche players were the newly formed endo-symbiotic chimeras themselves, who then underwent a massive adaptive radiation to form ever more complex forms of cellular machinery and ultimately LECA and the diversity of eucaryotes as we now know them. There are echoes in this metabolic shangri-la of the conditions in lost city vents that we are coming to understand may have likewise given rise much earier to LUCA.

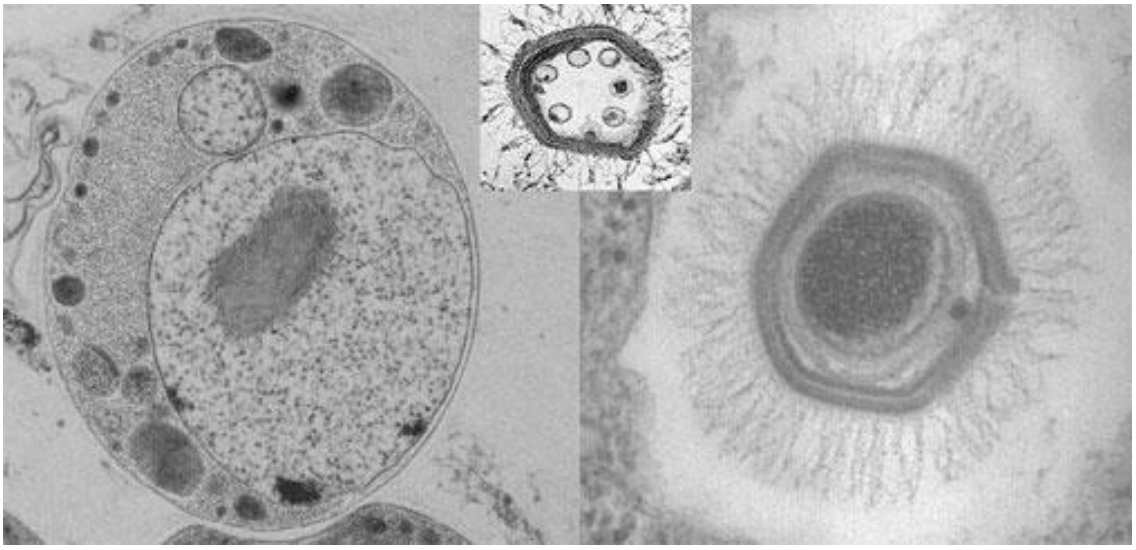

Fig 6b: Left: Bacterium Gemmata obscuriglobus with internal nuclear envelope and vaccuoles (Rachel Melwig & Christine Panagiotidis / EMBL). Right: Ultrathin EM section of a mimivirus in amoeba (Jean-Michel Claverie) Inset: Mamavirus infected by sputnik phage.

Offset against both the uniqueness of the mitochondrial endo-symbiosis and the closely linked, but independent question of the origin of the nucleus and nuclear envelope, has been the discovery of mimi-, mama-, mega- and pandora-viruses infecting amoeba (Raoult et, al., Philippe et al) and related very large aquatic viruses such as CroV infecting single celled plankton species (Fisher et. al.), which despite their recent discovery, appear from ocean gene analyses to be potentially ubiquitous and widespread in the oceans and possibly playing a crucial role in regulating the atmospheric-oceanic pathways, such as carbon sequestration. These form an intermediate genetic position between viruses and cells, having the largest genomes, with extensive cellular machinery and larger than the smallest completely autonomous bacterial and archaeal genomes.

Megavirus chilensis, for example is 10 to 20 times wider than the average virus. The particle measures about 0.7 micrometres (thousandths of a millimetre) in diameter. It just beats the previous record holder, Mimivirus, which was found in a water cooling tower in the UK in 1992. A study of the megavirus's DNA shows it to have more than a thousand genes. The mimivirus genome is a linear, double-stranded molecule of DNA with 1.18 Mbp in length. Megavirus has 1.25 Mbp. Like Mimivirus, Megavirus has hair-like structures, or fibrils, on the exterior of its shell, or capsid, that probably attract unsuspecting amoebas looking to prey on bacteria displaying similar features. These viruses show many characteristics at the boundary of living and non-living. They are as large as several bacterial species, such as Rickettsia conorii and Tropheryma whipplei, possess a genome of comparable size to several bacteria, including those above, and code for products previously not thought to be encoded by viruses. Mimivirus has genes coding for nucleotide and amino acid synthesis, which even some small obligate

intracellular bacteria lack. However, it lacks genes for ribosomal proteins, making it dependent on a host cell for protein translation and energy metabolism.

As of mid-2013, an even larger virus with a 2.5 Mb genome without morphological or genomic resemblance to any previously defined virus families has been discovered by the same researchers that found mimivirus, in both the same ocean sample off Peru and in a freshwater pond in Australia. Named pandoravirus - reflecting their lack of similarity with previously described microorganisms and the surprises expected from their future study. The researchers suspect that giant viruses evolved from cells. They think that at some point, the dynasty on Earth was much bigger than the three domains of bacteria, archaea and eukaryotes. Some cells gave rise to modern life, and others survived by parasitizing them and evolving into viruses. Pandora might thus provide a complementary relic of the genomes of this wider founding group (Philippe et al). Using the Global Ocean Sampling (GOS) Expedition data to explore variants of recA (the universal DNA repair enzyme) and rpoB (the beta subunit of bacterial RNA polymerase) a team associated with Craig Venter have discovered branches which may also point to a fourth domain (Wu et al).

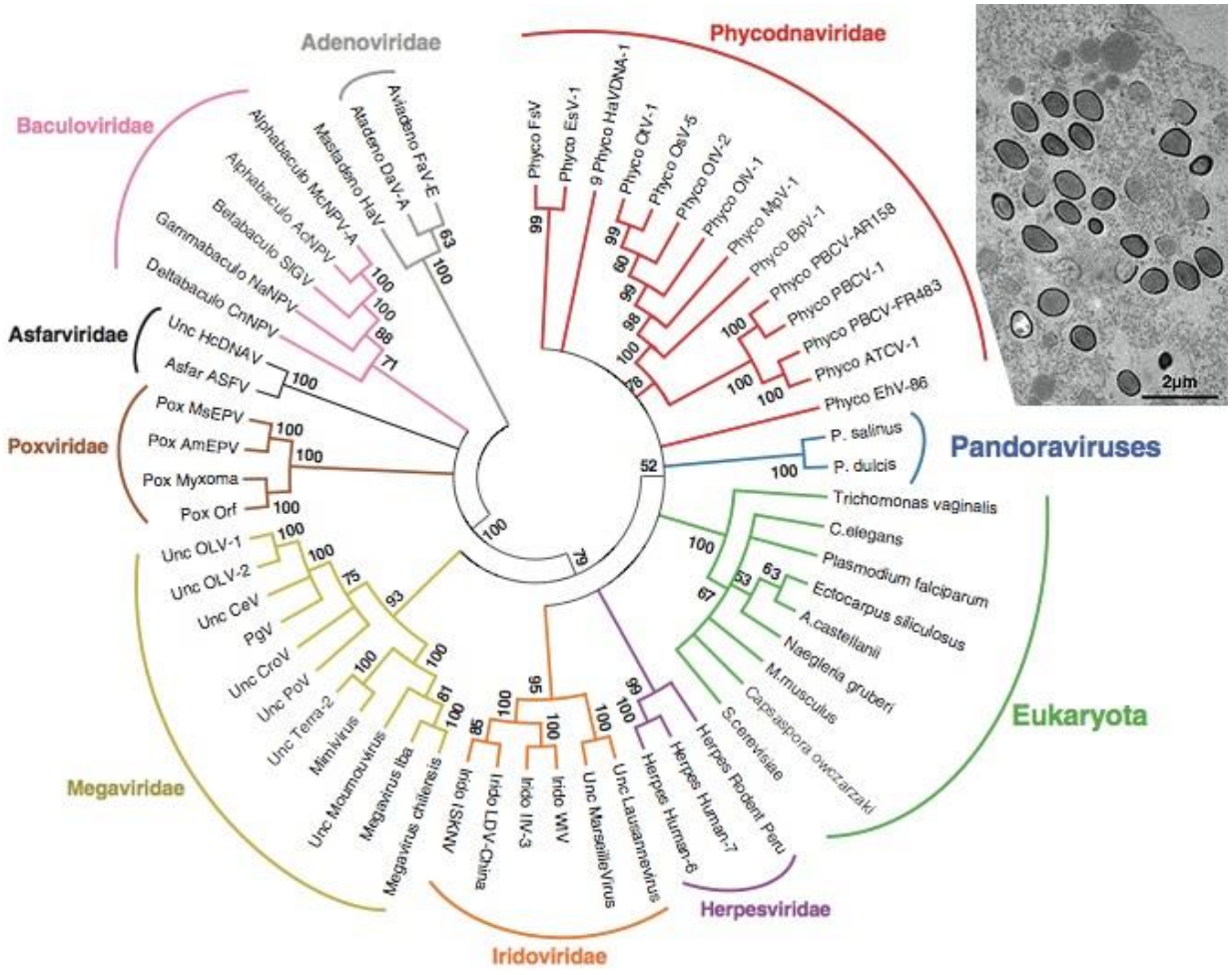

Fig 6c: Evolutionary tree of B-family DNA polymerase showing relationship of pandoravirus to other viruses and eucaryotes. Inset is shown pandoraviruses invading acanthamoeba (Philippe et al).

As an illustration of genes in mimivirus normally appearing only in cellular genomes, the mimivirus has genes for central protein-translation components, including four amino-acyl transfer RNA synthetases, peptide release factor 1, translation elongation factor EF-TU, and translation initiation factor 1. The genome also exhibits six tRNAs. Other notable features include the presence of both type I and type II topoisomerases, components of all DNA repair pathways, although the topoisomerase 1B has a different header structure from the eucaryote form (Brochier-Armanet, Gribaldo & Forterre 2008), many polysaccharide synthesis enzymes, and one intein-containing gene. Inteins are protein-splicing domains encoded by mobile intervening sequences (IVSs). They self-catalyze their excision from the host protein, ligating their former flanks by a peptide bond. They have been found in all domains of life (Eukaria, Archaea, and Eubacteria), but their distribution is highly sporadic. Only a few instances of viral inteins have been described. Self-splicing type I introns are a different type of mobile IVS, self-excising at the mRNA level. They are rare in viruses. Mimivirus exhibits four instances of self-excising intron, all in RNA polymerase genes.

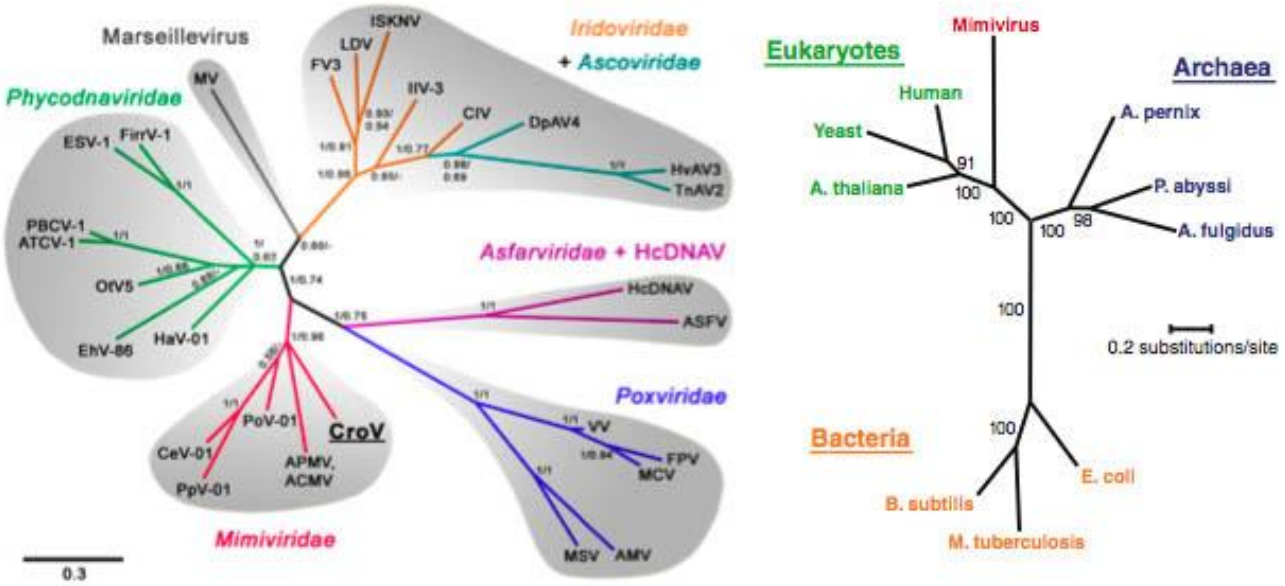

Fig 6d: Evolutionary diversification of Mimiviruses from nucleocytoplasmic large DNA viruses (Fisher et. al.) and in relation to the three domains of cellular life based on the concatenated sequences of seven universally conserved protein sequences (Raoult et. al.)

Mamaviruses also host parasitic virophages, affectionately named sputnik (Pearson 2008) as viral satellites, which piggy back on the metabolism of the large viral factories set up by these giant viral genomes causing the mimiviruses to sicken, and these virophages also contains genes that are linked to viruses infecting each of the three domains of life Eukarya, Archaea and Bacteria (La Scola et. al.). It has thus been suggested that they have a primary role in the establishment of cellular life and that they may have been instrumental in the emergence of the nuclear envelope.

(Continued on Part II)