

Part I

# Evolution of Life

COPYRIGHTED MATERIAL



# Chapter 1

---

## Evolutionary Genomics Leads the Way

*David Penny and Lesley J. Collins*

- 1.1 INTRODUCTION
  - 1.2 EVOLUTION AND THE POWER OF GENOMES
  - 1.3 THE PROBLEM OF DEEP PHYLOGENY AND “THE TREE”
  - 1.4 FRED, THE LAST COMMON ANCESTOR OF MODERN EUKARYOTES
  - 1.5 EUKARYOTE ORIGINS: CONTINUITY FROM THE RNA WORLD?
  - 1.6 MINIMAL GENOMES AND REDUCTIVE EVOLUTION
  - 1.7 EVOLUTIONARY GENOMICS FOR THE FUTURE
- REFERENCES

### 1.1 INTRODUCTION

When the older of our authors was an undergraduate (we won't tell you how long ago, but it was certainly way back in the last millennium), there were considered three “Great Scientific Problems.” All three were questions about origins that might (in principle) have genuine scientific answers, but at that time they were thought to be so complex that we might never find them—the questions might just be too big to ever find a scientific solution. The questions were

1. the origin of humans,
2. the origin of life, and
3. the origin of the universe.

It is brilliant that in a single working life the first is answered and the second is crumbling away. The third (the origin of the universe) we recognized even “way back then” as a question of a different kind in that, in principle, could lead to an infinite regress. That is, solve the question about the origin of our universe (say, hypothesis A) and it immediately

opens up another question, namely, what explains hypothesis A. So the third question is best left to the physicists or philosophers! As we show below, analyzing genomics datasets allows us to address such major questions where solutions were not possible even 5–10 years ago.

## 1.2 EVOLUTION AND THE POWER OF GENOMES

The first step is showing how access to information about complete genomes allowed the first of the above three questions to be answered; so in this section, we will only refer to the comparison of the chimpanzee and human genomes. The practical point is how these two genomes can be used as a test of the question whether microevolutionary processes are *sufficient* for macroevolution (Penny and Phillips, 2004). Can the origin of the human genome be understood *solely* in terms of the normal microevolutionary processes that occur in natural populations? This is a major scientific question—perhaps “the” major question.

The genetic processes in populations (that we know about) include point mutations (SNPs, single nucleotide polymorphisms), small insertions and deletions (indels, from a single nucleotide to larger indels), variations in copy number (of a gene or other fragment of DNA, CNVs; Redon et al., 2006), inversions and translocations of sections of chromosome and also chromosome fusions, and activated retrotransposable elements. These are precisely the differences we see between the human and chimpanzee genomes, and the classes of differences are outlined in Table 1.1 (see Li and Saunders (2005) and Levy and Strausberg (2008)).

Thus, the conclusion is that the human genome arises strictly from the natural processes that occur in plant and animal populations—we can find nothing different or unexpected about the genetic processes leading to humans. The conclusion is extremely powerful and should be reiterated continually by biologists talking to members of the public, especially in more religious countries. In principle, other possibilities exist for the origin of the human genome. Although we make jokes about it, saying that perhaps a Kindly Creator, or a Group of Itinerant Space Travellers (the GIST model), might have inserted into the human genome a whole lot of genes for both wisdom and intelligence. Just think of that, we tell our students, many, many genes in the human genome for wisdom and intelligence. After the appropriate pause, we continue—all we have to do now is to find how to turn those genes on!

All right, so on the surface the story is a joke, but the story has a very serious purpose and carries a much deeper significance. As far as we can tell, the human genome arose from an

**Table 1.1** Natural Microevolutionary Differences Between Chimpanzee and Human Genomes

	Natural
One chromosome fusion	<input checked="" type="checkbox"/>
Many point mutations	<input checked="" type="checkbox"/>
An enzyme lost	<input checked="" type="checkbox"/>
Many differences in copy number	<input checked="" type="checkbox"/>
Many small inversions	<input checked="" type="checkbox"/>
Different transposable elements activated	<input checked="" type="checkbox"/>
Many indels (insertions/deletions)	<input checked="" type="checkbox"/>
Some introns expanded/contracted	<input checked="" type="checkbox"/>

The differences between the human and chimpanzee genomes are all the normal microevolutionary processes seen within populations and sibling species (see Li and Saunders (2005) and Levy and Strausberg (2008)).

ape-like genomic ancestor through 100% normal microevolutionary processes—processes that occur within populations or between sibling species. So yes, in principle, there were other alternatives. If normal microevolutionary processes are *sufficient* to lead to humans, then that is a very powerful conclusion about the sufficiency of microevolutionary processes.

Of course, we do not know which combinations of mutations (of the many kinds of mutations that can occur) led to which changes in human morphology, behavior, and mental and social abilities. And similarly, we do know which mutations led to which changes on the chimpanzee lineage. Fortunately, we tell our students again, there is certainly a huge amount to be learnt in the future, and there is a major role for the next generation. But what we can say is that at the level of the human genome, there is nothing unusual about humans. That is a great improvement over what could have been said even a decade ago. It illustrates the tremendous importance of having genomes of related organisms. If humans, and our achievements, can arise by natural mechanisms, then complete genome analysis has led to the major conclusion that microevolutionary processes are sufficient for a major macroevolutionary change, that is, the origin of humans, with all our creative (and destructive) powers.

Before leaving humans, one more example of how molecular genetic data is generally so important; in this case, even for interpreting fossils. Again, it comes from the experience of the older author as a graduate student when Louis Leakey toured North America lecturing about their new fossils of early humans in Africa. At a reception afterward, the paleontologists were quite unimpressed, “we know humans evolved in Asia” was their conclusion (for several decades they had scathingly ignored the early human fossils found by Raymond Dart, Robert Bloom, and others in Southern Africa). No way, said the molecular biologists—Morris Goodman (Goodman et al., 1962) is finding a very close molecular relationship between humans, chimpanzees, and gorillas. The latter two are exclusively African, so we should be looking in Africa for early human fossils. “Look,” said the molecular biologists, “Louis Leakey is doing just that and he is finding the predicted fossils in Africa.” Thus, we clearly see that molecular data is critical for many other areas of biology.

So although the human/chimp case is just one example, the evolution of whole genomes is a rapidly developing field. For example, 12 new genomes of fruit fly (*Drosophila*) species were published in a single publication (Clark et al., 2007), more genomes of individual humans are becoming available (e.g., Wang et al., 2008), and now there are proposals for 1000 human genomes, and as soon as possible. We have given one example of how genomes can answer major scientific questions, but in many cases we need to know the evolutionary relationships among the taxa involved; as the next section shows that is not as easy as it sounds for deep divergences.

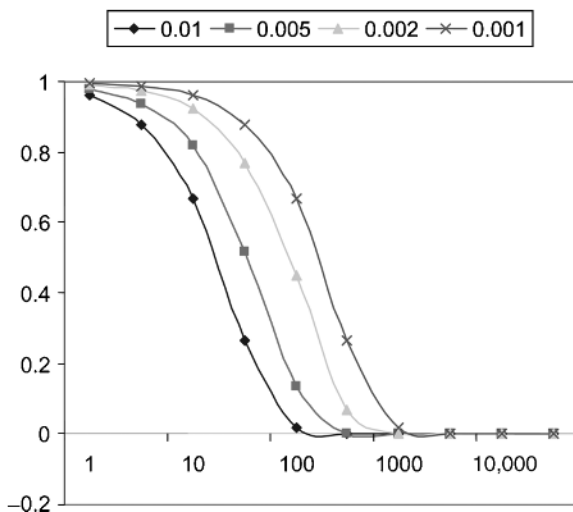
### 1.3 THE PROBLEM OF DEEP PHYLOGENY AND “THE TREE”

We need to know the basic divisions both within and between the archaea, bacteria, and eukaryotes. However, there is a real problem in resolving ancient divergences; almost certainly the deep branching orders cannot be known from aligned DNA and protein sequences. So although genomics data is necessary, we don’t yet have the theory for determining the most powerful approach for comparing genomes.

The primary approach in phylogeny uses aligned DNA or protein sequences (see Figure 1.1a). Both theory and simulation show that this approach is excellent over a range of times (with some lack of signal at shorter times). However, there is a major loss of information at longer times (say a billion years or more). As in many areas of science, the

Alignment original sequence order	Reordered Alignment shuffled/reordered
AIIFLNSALGSPSELFPPIILATKVL	ASAGPSPPATPLLLIIILLFFNEKV
AIMFLNSALGPPTELFPVILATKVL	ASAGPPTPATPLLLIMVILLFFNEKV
SIMFLNHTLNPTPELFPPIILATETL	SHTNPPTPATPLLLIMIILLFFNEET
TILFLNSSLGLQPEVPTVTLATKTL	TSSGLQPPATPLLLILTIVLTFNEKT
TLLFLNSMLKPPSELFPPIILATKTL	TSMKPPSPATPLLLIIILLFFNEKT
ALLFLNSTLNPPTELFPPIILATKTL	ASTNPPTPATPLLLIIILLFFNEKT
AILFLNSFLNPPKEFFPIILATKIL	ASFNPPKPATPLLLIIILLFFNEKI

**Figure 1.1** Current phylogenetic methods use only a small part of the information in a sequence and are expected to saturate for deep divergences. (a) Aligned amino acid sequences for seven taxa, with amino acids color coded for chemical properties. (b) The same sequence with the columns (sites) reordered; any of the  $c!$  reorderings of the columns always gives the same parsimony or likelihood value, or distance matrix. (c) Probability of recovering the character state at the root for different mutation rates (the  $x$ -axis is time in millions of years). Figure from Penny and Steel, unpublished manuscript. (See insert for color representation of this figure.)



standard mechanism that we use is a Markov model. This is interesting from an evolutionary viewpoint in that Markov models assume “continuity” of process—there were a continuous series of generations of DNA molecules between the starting and end point of the process. Great, that is good evolutionary theory, but the model does not use all the information in the data, for example, the order of the columns in an alignment is not used. Figure 1.1b shows one reshuffling of the columns in Figure 1.1a, and this will give precisely the same tree and parameters as the data from Figure 1.1a. If the number of columns in an aligned data set is  $c$ , then there are  $c!$  ways of shuffling the alignment. (There are  $c$  ways of selecting the first column,  $c - 1$  ways the second column,  $c - 2$  the third, and so on.) There is clearly more information than this in the sequence data; for example, if we shuffled the order of amino acids in a protein, we could get very different 2D and 3D structures.

Markov models are well studied mathematically, and in the case of trees, it is known that at longer times all information about the tree that generated the data is lost (see Mossel and Steel, 2004). These authors show that there is “phase transition,” the probability of either recovering the tree correctly or (given the tree) inferring the ancestral character state, is initially high but decreases markedly in the longer term. The rate of information loss depends on the mutation rate; information loss is obviously faster at higher mutation rates (Figure 1.1c). Thus, there must be a real limit to using aligned sequence data for ancient divergences, and the loss of information will be even faster if there are any systematic errors (Phillips et al., 2004), such as differences in nucleotide composition or in the function or 3D structure of the protein. The calculations of Mossel and Steel (2004) assume the “best possible” case; reality will usually ensure that in practice the loss of information will be even faster. Yes, maybe we can do better by downweighting the fast evolving sites (Jeffroy et al., 2006), but we do not expect this to eliminate all problems.

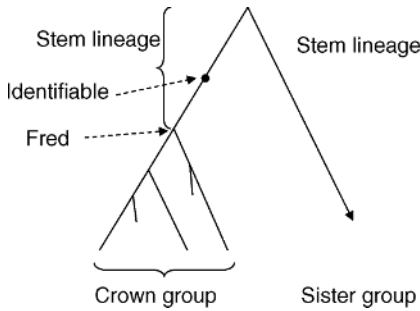
At this point, our earlier comment about our Markov models not using all the information is highly relevant. There is additional information in the data from gene order (Henz et al., 2005), unique structural changes to the genome (such as insertions and deletions, Boore, 2006), and 2D and 3D structures (Caetano-Anollés et al., 2007). So there is plenty of room for progress. It is just that the purists want to check whether these classes of additional information do really lead to similar results, and we also want a better theoretical understanding of the expected changes in information from gene order and 3D structure, especially when there is a change in function of the gene (Lockhart et al., 1996). Despite these qualifications, using the additional information is a top priority for understanding deep divergences.

The rise of RNA has also seen a shift in thinking in phylogenetics. In addition to examining relationships using protein genes (e.g., the ~2% of the human genome), we now have to handle genes that are non-protein coding. In addition, we have to take the structure of a molecule into account as well as the sequence. The idea of using RNA secondary structure has been around for a while, but using it in a practical manner in phylogenetics is still under development despite the volume of software available (Freyhult et al., 2007). With RNA becoming flavor of the month, we expect to see a large jump in research in secondary structure evolution and this will lead to better predictive software for ncRNA searching (especially for non-human and non-Arabidopsis miRNAs) and new software for ncRNA phylogenetics.

When it comes to the universal “Tree of Life,” we need to be much more careful and be much more “Darwinian” than many modern commentators. The reason is that Darwin almost universally used the phrase “the theory of descent with modification,” not the “tree of life” (a concept that is biblical in origin and has strong mystical overtones). He did suggest at one stage that the “tree of life” was a useful simile, and that is certainly a constructive way of expressing the situation. Thus, although “tree of life” (written in lowercase) is a useful analogy, “descent with modification” is much more accurate and inclusive. We know that both mitochondria and chloroplasts arose by endosymbiosis, as did some other organelles, including the nitrogen-fixing organelle in the interesting diatom *Rhopalodia gibba* (Precht et al., 2004). With molecular data, hybridization is being found much more often in both plants and animals (Mallet, 2008). Bacteria regularly “beg, borrow, and steal” genes from relatives or more distant organisms (Dagan et al., 2008), and indeed even the concept of bacterial “species” needs updating away from the eukaryote expectation that every individual has more or less the full genome of the “species” (Lan and Reeves, 2000). With bacteria, that concept is not appropriate, and strains can be relatively different in their composition of genes (but still able to regain genes from related strains). This is only one way in which we have changed our thinking of genome evolution over the past 10 years. We still see this muddle of prokaryote evolution because bacteria and archaea exchange DNA faster than their researchers. However, most prokaryotic evolutionists can consider a “network” nature of bacterial genome evolution means that accurately working out the finer details of bacterial tree rooting is not going to be easy (Dagan et al., 2008).

## 1.4 FRED, THE LAST COMMON ANCESTOR OF MODERN EUKARYOTES

Of course, we are on one side of a eukaryotic origins’ argument (Kurland et al., 2006), but is diving directly into origins really the way to go? Consider that instead of going for the origin, we focus on the later step, the last eukaryotic common ancestor. The deep phylogeny of eukaryotes splits into five or six deep groups (e.g., Keeling et al., 2005) that are very much accepted, though certainly not the deeper rooting (Roger and Hug, 2006). Trees based on the



**Figure 1.2** Stem lineages and crown groups. The crown group consists of all descendants (living and extinct) of the last common ancestor of all (in this case) eukaryotes; we call this organism “Fred.” In contrast, the stem lineages are all the earlier organism, most of which will probably not be on the direct lineage to extant eukaryotes. In principle, some early members of the stem lineage might be difficult to recognize as being eukaryotic. We might, for example, choose the origin of the mitochondrion, by endosymbiosis, as the event that “defined” eukaryotes, and this is indicated as a large dot as “identifiable.” This event predates Fred, by an unknown amount. The choice of endosymbiosis to define eukaryotes is arbitrary, though not unreasonable, and other choices might be a nucleus, or the existence of a spliceosome, and so on. The “sister group” has its usual meaning of including the crown group of the next most closely related group, possibly archaea in the case of eukaryotes. However, in the case of eukaryotes, it is conceivable (possibly) that there was no “sister group” that the protoeukaryote lineage was formed by a fusion of an archaeal and a bacterial cell—each of those would be equally related. Even in this scenario, there would almost certainly be a stem lineage of “protoeukaryotes.”

fusion of a single gene do not take gene fission into account; trees based on SSU rRNA create problems for species with longer branch lengths. We are left with tongue-twisting groupings (e.g., Opisthokonts for fungamals (fungi plus animals)) and a variety of paths by which the “first” eukaryotes arose.

Given the uncertainty about deep phylogeny from aligned sequences, our current approach for eukaryote origins (Collins and Penny, 2005) is to temporarily put aside the ultimate origin of eukaryotes and to concentrate on the properties of the last common ancestor of modern eukaryotes—Fred.<sup>1</sup> Figure 1.2 illustrates the relative difference between these two key concepts. The more we learn about the biochemical and subcellular properties of Fred, the more we are basing our inferences on real data. Because we have huge amounts of genomic information from present-day eukaryotes, we are able to infer many aspects of the biology of past eukaryotes. Our strategy has been to search for features that occur in all deep lineages of eukaryotes. We call such features “general” (or “ancestral”) for eukaryotes, rather than universal; this allows some groups to have lost an ancestral feature. A general feature is expected to be present in the last common ancestor of the eukaryotic cell. Contrary to tradition, the younger of our authors refuses to call this hypothetical beastie LECA (last eukaryote common ancestor). The name could imply a close relationship to LUCA (the last universal common ancestor) and at this stage how close they are in time is hypothetical, so instead the name became “Fred.” With a neutral name, we can approach independently the characteristics of Fred and begin with an obvious question: what did Fred look like?

Over the past few years, we are getting some good information (e.g., Kurland et al., 2006). Fred almost certainly had a mitochondrial-type organelle and functional splicing (implying both introns and the complicated apparatus to remove them, Collins and Penny, 2005). Other facets of RNA processing especially within the transcription–translation

<sup>1</sup>Some people refer to Fred as a “fairly remote eukaryotic daddy.” We just use Fred.

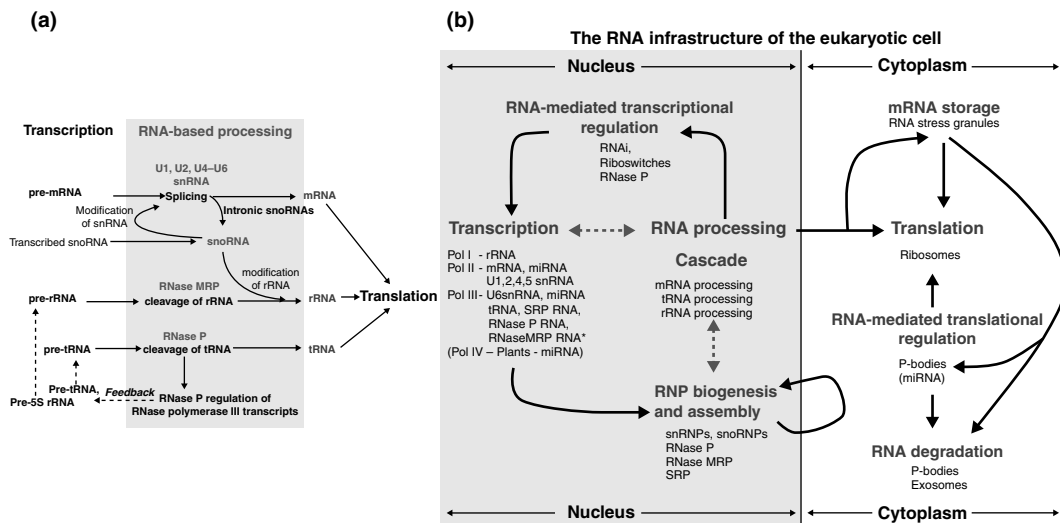
system are under investigation (Collins and Chen, 2009). There are immediate questions that should be solvable very soon (with perhaps a few more key eukaryote genomes). How much of RNA-based regulation (such as RNAi) can be traced back to Fred? Riboswitches are found in both eukaryotes and prokaryotes (Cheah et al., 2007), so we could make an assumption that they, as a general mechanism, extend even back to LUCA, but before making this great step, we have to check that the mechanisms are sufficiently similar to infer common ancestry. Recent work makes it likely that alternative splicing occurred relatively early within eukaryotes (Irimia et al., 2007) and therefore was available to be recruited into development of later multicellularity. But do the proteins of the nuclear pore complex (D'Angelo and Hetzer, 2008) occur through eukaryotes? It is interesting that the assumptions originally applied just to sequences (ancestral sequences or convergent evolution) and now being applied to entire mechanisms. To investigate the generality of mechanisms, we have to apply evolutionary models and concepts not only to sequences, but also to expression and pathway information, cell structure, and more complex metabolic information. Thus, we begin to step beyond evolutionary genomics and into the realm of evolutionary systems biology.

For example, there are many proteins that are only found in eukaryotes—eukaryote signature proteins (ESPs) (e.g., Hartman and Fedorov, 2002; Kurland et al., 2006). Clearly, an archaean plus bacterial fusion does not explain ESPs directly, though a very long period between fusion and Fred would help. This is a point where considering stem and crown groups does help (see Figure 1.2). Similarly, a fusion model does not either predict or explain the origin of the spliceosome/intron/exon structure of Fred (Collins and Penny, 2005). Certainly, some models attempt to explain the intron/exon structure as an invasion of type II introns at the time of mitochondrial acquisition by endosymbiosis. However, this certainly does not explain the origin of the spliceosome, a bigger structure even than the ribosome—this is one reason we emphasize the origin of the spliceosome as part of the exon/intron processing apparatus. Beware “the invasion of the introns” that some models propose, but those models (by themselves) appear to us unlikely. To us it would slow down RNA processing completely, and any members of the population that did not suffer intron invasion would strongly outcompete those poor individuals suffering from the invasion. Nevertheless, it serves as a point to introduce the widespread involvement of RNA in eukaryotes.

One of the greatest surprises for many people is that eukaryotic molecular biology is so RNA based (Amaral et al., 2008; Costa, 2007). Eukaryotes abound in RNA-based processing of other RNAs (Figure 1.3): rRNA transcripts cleaved by MRP RNase and modified by guide RNAs such as snoRNAs (small nucleolar RNAs); mRNA cleaved by snRNAs (small nuclear RNAs) in the spliceosome, and tRNA transcripts cleaved by RNase P. There are others, but in addition there is also the very widespread regulation of RNA (RNAi, Munroe and Zhu, 2006), and even though riboswitches are much more frequent in bacteria, they still occur in eukaryotes (Montange and Batey, 2008).

Overall, there is a complex network of ncRNA-catalyzed and controlled processes, especially around transcription and translation (Figure 1.3). We find many systems such as transcription, splicing, and RNA export are so coordinated that they not only share protein components but also operate on the RNA at the same time. When we take into account the numerous biogenesis pathways to produce some of the components for these mechanisms (e.g., snRNAs, Matera et al., 2007), we find more RNA-based molecules linked to this pathway.

We quickly see what we call the RNA infrastructure (Collins et al., 2009), a network of ncRNA-based processes regulating RNA processes around the cell, both in time and in space. It is moving components in and out of the nucleus (Hopper, 2006), or in and out of



**Figure 1.3** The eukaryote RNA infrastructure. (a) A generalized RNA processing cascade showing ncRNA-based processing from transcription to translation, concentrating on the processing of mRNAs, rRNAs, and tRNAs. This is the central section of the overall RNA infrastructure network (b) where different RNA-based processes feed into and regulate others, including those in the central section. This is a generic model that can differ in detail within different lineages—for example, MRP RNA is transcribed by Pol III in humans but Pol II in *Saccharomyces cerevisiae*. Red dashed arrows indicate that processes within each group interact in either direction. Based on Woodhams et al. (2007). (See insert for color representation of this figure.)

RNA storage granules (Anderson and Kedersha, 2006). We could ask the question as to whether ncRNA is the biological “dark matter,” the previously unappreciated molecules that take a single stretch of DNA and produce a functional protein in the right place at the right time. Now that we are understanding more about such basic eukaryote features as RNA processing, we are now getting a good overview of the composition of Fred, and we need this before moving backward in time toward the origin of the eukaryote lineage. The widespread occurrence of RNA in eukaryotes will come up again in the next section.

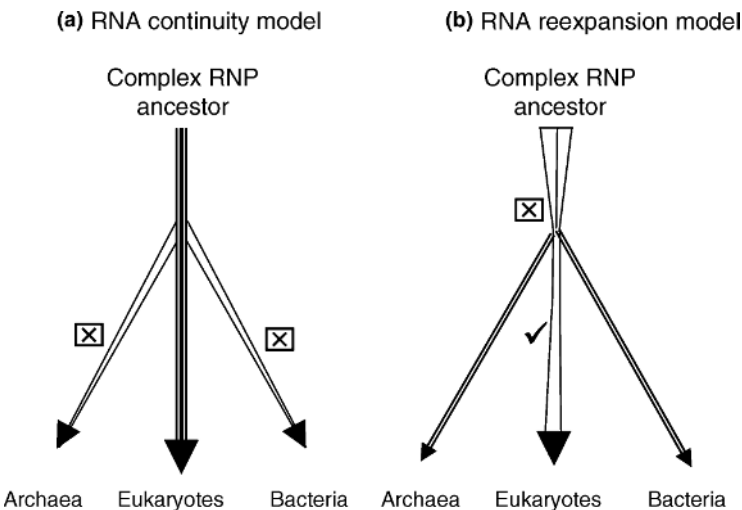
## 1.5 EUKARYOTE ORIGINS: CONTINUITY FROM THE RNA WORLD?

Put another way, by focusing on Fred, we can use highly detailed biological knowledge information from real (i.e., existing) organisms to infer ancestral properties, rather than just invoking magic to suggest something about the earlier origin of eukaryotes. We sometimes feel that “every man and her dog” has a theory about eukaryote origins (well summarized in Embley and Martin (2006)); it is just that we (all of us, ourselves included) don’t fully know all the features that need to be accounted for in the ancestral eukaryote. Thus, our preference at the moment is to define as fully as possible the properties of Fred, thereby helping understand what questions have to be answered in any theory about the origins of eukaryotes. We may delay facing it, but the question of the ultimate origin of eukaryotes will certainly not go away.

Although endosymbiosis is established for the origin of the mitochondrion (and its derivatives such as mitosomes and hydrogenosomes), this model has not by itself really helped understand the complexity of either the cell that at an early stage engulfed the

endosymbiont or the later stage of Fred (see Figure 1.2). It used to be fashionable to suggest that the eukaryote cell arose from a fusion of an archaeal and a bacterial cell; we call this the  $0 + 0 = \infty$  model. Because both archaea and bacteria lack important general features that are characteristic of eukaryotes (see Section 1.4), fusing an archeon and a bacterium certainly does *not* give a eukaryote! Fusion, by itself, does not explain the origin of eukaryotes. We need to be careful here—such an argument cannot establish that fusion did not occur, it is just that fusion by itself does not explain the origin of the many unique eukaryote features.

Now we get to a very fundamental and important, but difficult, question—is there continuity of the RNA processing of RNA from eukaryotes, past Fred, and all the way back to the predicted widespread processing of RNA by RNA in the earlier stages of life (Penny, 2005)? We expect there to have been an RNP world (an RNA plus protein world that must have preceded DNA) and an even earlier RNA world that would have preceded encoded proteins. For the ribosome, tRNAs, and mRNA, there is indeed little doubt that they are very ancient. What about the MRP RNase ribozyme that processes the rRNA transcript in eukaryotes? To what extent is RNA processing and regulation of RNA in modern eukaryotes continuous right back to an RNP world? Since 1998 (Jeffares et al., 1998), we have been exploring this rather unfashionable (some would perhaps say heretical) idea that eukaryotes retain some ancestral RNA processing features that have been lost in the highly streamlined (and efficient) “prokaryotes” (see Collins et al., 2009). Protein enzymes are far more effective than ribozymes (see comparative values of  $k_{\text{cat}}$  and  $k_{\text{cat}}/K_m$  in Table 1 of Jeffares et al. (1998)). Superficially, at least, it is unlikely that ribozymes will take over a catalytic or regulatory function that proteins are already doing. The simplest model is therefore an irreversible trend of ribozymes  $\rightarrow$  proteins for catalysis (see Figure 1.4). Thus, whatever the final decision, the idea of RNA continuity is well worth exploring.



**Figure 1.4** Comparison of two models for the origin of the high RNA complexity in eukaryotes. (a) Under the RNA continuity model, the complex system of RNA processing of RNA is *largely* continuous from an earlier ribonucleoprotein stage of the origin of life—an RNP world. The model involves two losses of complex RNA processing in the streamlined and efficient “prokaryote” groups. (b) Under the RNA reexpansion model, there was the same early complex system of RNA processing of RNA, but it was largely lost for the last universal common ancestor, and then subsequently reexpanded in eukaryotes. This model has one loss and one gain. (The order of branching of archaea, bacteria, and eukaryotes is not shown because it is not relevant to either model.)

In contrast to the RNA continuity concept, the dominant theory by far for the origin of eukaryotes is the pre-Darwinian theory of evolution by orthogenesis (for this concept, see Blomberg and Garland (2002)). This assumes some unknown (and possibly unknowable) “universal principle” of evolution going from the simple to the complex. In other words, it is “blindingly obvious” (except to a few heretics) that prokaryotes preceded eukaryotes! The idea that the smaller prokaryotic cells lead inexorably to the larger eukaryote cells is just one manifestation of this—“bigger is better” seems to be the motto.

An alternative way of thinking about prokaryotes and eukaryotes is to consider prokaryotes as elegant and efficient in both their genome structure and RNA functioning, and to consider eukaryotes as clumsy and inefficient in their genome organization and their RNA processing. Clumsy and inefficient maybe, but that very inefficiency and redundancy has allowed all sorts of complexities to develop, for which we multicellular eukaryotes are quite grateful. Indeed, we joke about the inordinate complexity of the RNA processing system in eukaryotes and say that “not even a University Committee could invent a system as clumsy and inefficient as the eukaryote genome.” In contrast, we might be quite proud to have been on a committee that designed intelligently a prokaryote genome! Again, it is a joke with a serious message. We cannot accept “*a priori*” that “because the RNA processing of RNA in eukaryotes is so complex, it *must* be advanced”! To us, eukaryote RNA processing is a just clumsy example of “unintelligent design.”

The usual hypothesis to explain genome organization under prokaryotes-first model implies that the extensive RNA processing and regulation that we expect in an early RNA and RNP world would be largely lost in prokaryotes, and then reappear by magic in eukaryotes (Figure 1.4b). This is not impossible (maybe just unlikely from our point of view, given the relatively poor catalytic power of ribozymes, Jeffares et al., 1998).

New discoveries are reinforcing the fundamental (and possibly ancient) role of RNA in the basic functioning of the cell. The most recent is the report that tRNA itself is involved in the catalysis of an amino acid (threonine) onto the tRNA (Minajigia and Francklyn, 2008)—it is not only protein involved in the catalysis. More research is expected here; if this involvement of tRNAs in amino acid charging is widespread in other tRNAs and in other organisms, then it strengthens even further the concept of an RNA world being “alive and well” in modern organisms. So our plea is really quite simple—keep an open mind about the relationships between eukarya, archaea, and bacteria. We need more evidence, and we need to consider earlier stages of evolution (especially before DNA).

## 1.6 MINIMAL GENOMES AND REDUCTIVE EVOLUTION

Certainly, since the work of Forterre (1995), it is an option that has to be taken seriously that prokaryotes have streamlined their genomes and RNA processing from a more complex earlier state—whether the selective forces were thermoreduction (Forterre, 1995), *k*-selection (Jeffares et al., 1998), or something else is a separate issue. An immediate test is whether reduction in genome size is an ongoing strategy in some organisms today. It is basic to a “Darwinian” approach to evolution that the same selective forces are there, even though they might operate at very different rates in different lineages. It is also fundamental to Darwinian evolution that evolution is never directed to long-term goals, anything that gives an immediate advantage will be selected—even if possibly deleterious in the longer term. Evolution is certainly not always “forward” toward increased complexity.

Thus, it is conceptually important that genome reduction is observed in a variety of cases, as discussed by Andersson (2006) and Ochman and Davalos (2006). For example,

many pathogens (both bacterial and eukaryote) have reduced genomes and rely on the host for many nutrients. *Buchnera* is a bacterium that lives internally in insects such as aphids and makes essential amino acids for the aphid, but has lost its genes for making the other amino acids (Moran and Baumann, 2000). The critical point here is that there are many examples among existing organisms where there are reduced genomes—there is certainly no *a priori* argument against genome reduction in archaea and bacteria (see Figure 1.4).

On a different subject, in some eukaryotes there has been selection against high intron numbers. For example, a plot of the average numbers of introns per gene versus life cycle times shows a strong negative correlation (Jeffares et al., 2006). Eukaryotes (such as yeast) with short life cycles have few introns per gene. Conversely, eukaryotes (such as humans) with a long life cycle have many introns per gene, around eight per gene in our case. Similarly, it appears that genes that are turned on and off quickly (“nimble genes”) have fewer introns (Jeffares et al., 2008)—even though it is not clear yet which is cause and which effect. The evolution of intron numbers has been well studied (Roy and Irimia, 2008), and it seems clear that early eukaryotes did have larger number of introns per gene (Roy, 2006). Certainly, population size factors must be important (Lynch, 2002), and selection strengths on the gain or loss of an individual intron (for example) will be small (Wagner 2005), but the results outlined above indicate that selective factors do appear to be important. The way we phrase it is, the evolution is “sideways, backward, and occasionally, forward.” There is certainly no universal tendency to becoming larger and more complex, though there is certainly a niche there for some organisms that do manage it. The widespread occurrence of reductive evolution throughout nature means that the RNA continuity model has to be considered seriously as an option for the origin of eukaryote RNA processing.

## 1.7 EVOLUTIONARY GENOMICS FOR THE FUTURE

Of course, the younger author’s undergraduate years were not quite so long ago, but long enough to have witnessed the rise of the bioinformatic and genomic era. Starting out in molecular biology at a time when it was cool to “ftp” rDNA sequences onto our new computer, the rise in computing power saw the introduction of breakthrough technologies such as PCR, automated sequencing, and microarrays. Thus, we began to investigate evolutionary principles seen in many genes, rather than the accidental examples explored previously. When will it stop?

A downside of breakthrough technologies is that the analysis procedures and software may be years behind. We saw this with microarrays where the initial analysis software left a lot to be desired. But it did spur the bioinformatics industry. Suddenly bioinformatics was not the odd biologist who knew how to program but instead were programmers working on biological issues, and it was these programmers who developed and published the solutions. But what of the odd biologists pushed into the “bioinformatics” niche. Like any clever organism they had to evolve; some moved into heavier programming to join the bioinformaticians, and others kept solidly in the biological realm and created the field of genomics.

Technology has now moved again in the form of “next-generation sequencing (NGS).” As a former bench molecular biologist, we remember in nostalgic times when it was great to get our “gene” sequenced; now the problem is quickly analyzing data from entire genomes. The data is produced in a week; but the analysis can take forever! On the positive side, gone are the days when in evolutionary research our organism of choice was too distant from an already sequenced genome. Now we just sequence it. Unfortunately, next-generation sequencing still comes at a cost, but even that is predicted to be reduced within a few

years. Instead of a complete genome, we will be sequencing a population of genomes. This rapid progression of high-throughput technologies pumping out genome-scale data set after data set is also enabling genomics to grow at a rate much faster than a molecular clock, and entire centers now analyze and integrate data from a wide range of species (Schuster, 2008). But smaller centers are not being left out as we piggyback on the protocols set by these large centers, shifting the focus from model species to every species.

The problem of course is that with all these data, we still have to analyze it . . . “ah but there’s the rub ” (to coin a Shakespearean phrase). Should we be dismayed and allow ourselves to be overwhelmed by the sea of genomes in which we see ourselves swimming. Actually, no. Those of us with NGS data are rather excited. We turn again to bioinformatics and although it has been a slow start, the “sea of data” problem has been recognized (Valdivia-Granda, 2008), and this is the critical first step. Solutions to sifting and filtering these data are being approached as we type, but now as evolutionary biologists we have to face our own problems. How do we compare the genomes not of one or two representatives of a species, but of individuals from an entire population? Knowing as we do now the differences in evolutionary rates within a gene, and between genes, how do we apply our tools and models to correct for rate variation within entire genomes and can we infer ancestral genomes (Muffato and Crollius, 2008) (can we have a bigger computer to do it on please)?

The last comment is in fact a harsh reality of the genomics and systems biology world. We require not only “hunky” computers and flash programming in order to correlate a genome’s worth of evolution but also those with biological and evolutionary knowledge to work out what questions to ask of the bioinformaticians and the computers. Working one without the other is pointless. Evolutionary genomics was born beyond the chalk on blackboard age, right onto the silicon chip. As the technology advances, we pull in those advances, often before the classical molecular biologists have realized that evolutionary genomicists have poached upon their territory.

We look to the future and see a time where along with coffee, a laptop, a server, and a fast wireless connection, we can sit in the sun to look at how the world and eukaryotes have evolved. We are confident that the availability of more genomes from deeply diverging eukaryotes will answer many of the questions about the nature of the ancestral eukaryotes and its origins. More important, we should be able to answer the fundamental evolutionary question—extending the analysis from the first section. Namely, is there anything in genome evolution, for any species, that is *not* the result of normal microevolutionary processes? Yes, the human genome appears 100% the product of natural processes—how long before we can claim that for all genomes? Better make that lots of coffee!

## REFERENCES

- AMARAL, P.P., DINGER, M.E., MERCER, T.R., and MATTICK, J.S., 2008. The eukaryotic genome as an RNA machine. *Science* **319**: 1787–1789.
- ANDERSON, P. and KEDERSHA, N., 2006. RNA granules. *J. Cell Biol.* **172**: 803–808.
- ANDERSSON, S.G.E., 2006. The bacterial world gets smaller. *Science* **314**: 259–260.
- BLOMBERG, S.P. and GARLAND, T., 2002. Tempo and mode in evolution: phylogenetic inertia, adaptation and comparative methods. *J. Evol. Biol.* **15**: 899–910.
- BOORE, J.L., 2006. The use of genome-level characters for phylogenetic reconstruction. *Trends Ecol. Evol.* **21**: 439–446.
- CAETANO-ANOLLÉS, G., KIM, H.S., and MITTENTHAL, J.E., 2007. The origin of modern metabolic networks inferred from phylogenomic analysis of protein architecture. *Proc. Natl. Acad. Sci. USA* **104**: 9358–9363.
- CHEAH, M.T., WACHTER, A., SUDARSAN, N., and BREAKER, R.R., 2007. Control of alternative RNA splicing and gene expression by eukaryotic riboswitches. *Nature* **447**: 497–507.

- CLARK, A.G., EISEN, M.B., SMITH, D.R., et al., 2007. Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature* **450**: 203–218.
- COLLINS, L.J. and PENNY, D., 2005. Complex spliceosomal organization ancestral to extant eukaryotes. *Mol. Biol. Evol.* **22**: 1053–1066.
- COLLINS, L.J., and CHEN, X.S. 2009. Ancestral RNA: the RNA biology of the eukaryote ancestor. *RNA Biol.* **6**: 1–8.
- COLLINS, L.J., KURLAND, C.G., BIGGS, P., and PENNY, D., 2009. The modern RNP world of eukaryotes. *J. Hered.*, **100**: 597–604.
- COSTA, F.F., 2007. Non-coding RNAs: lost in translation? *Gene* **386**: 1–10.
- DAGAN, T., ARTZY-RANDRUP, Y., and MARTIN, W., 2008. Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution. *Proc. Natl. Acad. Sci. USA* **105**: 10039–10044.
- D'ANGELO, M.A. and HETZER M.W., 2008. Structure, dynamics and function of the nuclear pore complexes. *Trends Cell Biol.* **18**: 456–466.
- EMBLEY, T.M. and MARTIN, W., 2006. Eukaryote evolution, changes and challenges. *Nature* **440**: 623–630.
- FORTERRE, P., 1995. Thermoreduction, a hypothesis for the origin of prokaryotes. *CR Acad. Sci. Paris Life Sci.* **318**: 415–422.
- FREYHULT, E.K., BOLLBACK, J.P., and GARDNER, P.P., 2007. Exploring genomic dark matter: a critical assessment of the performance of homology search methods on noncoding RNA. *Genome Res.* **17**: 117–125.
- GOODMAN, M. et al., 1962. Immunochemistry of primates and primate evolution. *Ann. NY Acad. Sci.* **102**: 219–234.
- HARTMAN, H. and FEDOROV, A., 2002. The origin of the eukaryotic cell: a genomic investigation. *Proc. Natl. Acad. Sci. USA* **99**: 1420–1425.
- HENZ, S.R., HUSON, D.H., AUCH, A.F., NIESELT-STRUWE, K., and SCHUSTER, S.C., 2005. Whole-genome prokaryote phylogeny. *Bioinformatics* **21**: 2329–2335.
- HOPPER, A.K., 2006. Cellular dynamics of small RNAs. *Crit. Rev. Biochem. Mol. Biol.* **41**: 3–19.
- IRIMIA, M., RUKOV, J.L., PENNY, D., and ROY, S.W., 2007. Functional and evolutionary analysis of alternatively spliced genes is consistent with an early eukaryotic origin of alternative splicing. *BMC Evol. Biol.* **7**: 188.
- JEFFARES, D.C., POOLE, A.M., and PENNY, D., 1998. Relics from the RNA world. *J. Mol. Evol.* **46**: 18–36.
- JEFFARES, D.C., MOURIER, T., and PENNY, D., 2006. The biology of intron gain and loss. *Trends Genet.* **22**: 16–22.
- JEFFARES, D.C., PENKETT, C.J., and BÄHLER, J., 2008. Selection against introns in rapidly regulated genes. *Trends Genet.* **24**: 375–378.
- JEFFROY, O., BRINKMANN, H., DELSUC, F., and PHILIPPE, H., 2006. Phylogenomics: the beginning of incongruence? *Trends Genet.* **22**: 225–231.
- KEELING, P.J. et al., 2005. The tree of eukaryotes. *Trends Ecol. Evol.* **20**: 670–676.
- KURLAND, C.G., COLLINS, L.J., and PENNY, D., 2006. Genomics and the irreducible nature of eukaryote cells. *Science* **312**: 1011–1014.
- LAN, R. and REEVES, P.R., 2000. Intraspecies variation in bacterial genomes: the need for a species genome concept. *Trends Microbiol.* **8**: 396–401.
- LEVY, S. and STRAUSBERG, R.L., 2008. Individual genomes diversity. *Nature* **456**: 49–50.
- LI, W.H. and SAUNDERS, M.A., 2005. The chimpanzee and us. *Nature* **437**: 50–51.
- LOCKHART, P.J., LARKUM, A.W.D., STEEL, M.A., WADDELL, P.J., and PENNY, D., 1996. Evolution of chlorophyll and bacteriochlorophyll: the problem of invariant sites in sequence analysis. *Proc. Natl. Acad. Sci. USA* **93**: 1930–1934.
- LYNCH, M., 2002. Intron evolution as a population-genetic process. *Proc. Natl. Acad. Sci. USA* **99**: 6118–6123.
- MALLET, J., 2008. Hybridization, ecological races and the nature of species: empirical evidence for the ease of speciation. *Philos. Trans. R. Soc. B* **363**: 2971–2986.
- MATERA, A.G., TERNS, R.M., and TERNS, M.P., 2007. Non-coding RNAs: lessons from the small nuclear and small nucleolar RNAs. *Nat. Rev. Mol. Cell Biol.* **8**: 209–220.
- MINAJIGIA, A. and FRANCKLYN, C.S., 2008. RNA-assisted catalysis in a protein enzyme: the 2'-hydroxyl of tRNA<sup>Thr</sup> A76 promotes aminoacylation by threonyl-tRNA synthetase. *Proc. Natl. Acad. Sci. USA* **105**: 17748–17753.
- MONTANGE, R.K. and BATEY, R.T., 2008. Riboswitches: emerging themes in RNA structure and function. *Annu. Rev. Biophys.* **37**: 117–133.
- MORAN, N.A. and BAUMANN, P., 2000. Bacterial endosymbionts in animals. *Curr. Opin. Microbiol.* **3**: 270–275.
- MOSSEL, E. and STEEL, M., 2004. A phase transition for a random cluster model on phylogenetic trees. *Math. Biosci.* **187**: 189–203.
- MUFFATO, M. and CROLLIUS, H.R., 2008. Paleogenomics in vertebrates, or the recovery of lost genomes from the mist of time. *Bioessays* **30**: 122–134.
- MUNROE, S.H. and ZHU, J., 2006. Overlapping transcripts, double-stranded RNA and antisense regulation: a genomic perspective. *Cell. Mol. Life Sci.* **63**: 2102–2118.
- OCHMAN, H. and DAVALOS, L.M., 2006. The nature and dynamics of bacterial genomes. *Science* **311**: 1730–1733.
- PENNY, D., 2005. An interpretive review of the origin of life research. *Biol. Philos.* **20**: 633–671.
- PENNY, D. and PHILLIPS, M.J., 2004. The rise of birds and mammals: are microevolutionary processes sufficient for macroevolution. *Trends Ecol. Evol.* **19**: 516–522.
- PHILLIPS, M.J., DELSUC, F., and PENNY, D., 2004. Genome-scale phylogeny: sampling and systematic errors are both important. *Mol. Biol. Evol.* **21**: 1455–1458.
- PRECHTL, J., KNEIP, C., LOCKHART, P., WENDEROTH, K., and MAIER, U.G., 2004. Intracellular spheroid bodies of *Rhopalodia gibba* have nitrogen-fixing apparatus of cyanobacterial origin. *Mol. Biol. Evol.* **21**: 1477–1481.

- REDON, R., et al., 2006. Global variation in copy number in the human genome. *Nature* **444**: 444–454.
- ROGER, A.J. and HUG, L.A., 2006. The origin and diversification of eukaryotes: problems with molecular phylogenetics and molecular clock estimation. *Philos. Trans. R Soc. Lond. B Biol. Sci.* **361**: 1039–1054.
- ROY, S.W., 2006. Intron-rich ancestors. *Trends Genet.* **22**: 468–471.
- ROY, S.W. and IRIMIA, M., 2008. Spliceosomal introns as tools for genomic and evolutionary analysis. *Nucleic Acids Res.* **36**: 1703–1712.
- SCHUSTER, S.C., 2008. Next-generation sequencing transforms today's biology. *Nat. Methods* **5**: 16–8.
- VALDIVIA-GRANDA, W., 2008. The next meta-challenge for bioinformatics. *Bioinformatics* **2**: 358–62.
- WAGNER, A., 2005. Energy constraints on the evolution of gene expression. *Mol. Biol. Evol.* **22**: 1365–1374.
- WANG, J., et al., 2008. The diploid genome sequence of an Asian individual. *Nature* **456**: 60–65.
- WOODHAMS, M.D., STADLER, P.F., PENNY, D., and COLLINS, L.J., 2007. RNase MRP and the RNA processing cascade in the eukaryotic ancestor. *BMC Evol. Biol.* **7**: S13.