Mechanisms of Evolutionary Innovation Point to Genetic Control Logic as the Key Difference Between Prokaryotes and Eukaryotes

Citation: Bains, William, and Dirk Schulze-Makuch. "Mechanisms of Evolutionary Innovation Point to Genetic Control Logic as the Key Difference Between Prokaryotes and Eukaryotes." Journal of Molecular Evolution 81.1–2 (2015): 34–53.

As Published: http://dx.doi.org/10.1007/s00239-015-9688-6

Publisher: Springer US

Persistent URL: http://hdl.handle.net/1721.1/103317

Version: Author's final manuscript: final author's manuscript post peer review, without publisher's formatting or copy editing

Mechanisms of evolutionary innovation point to genetic control logic as the key difference between prokaryotes and eukaryotes

William Bains[1,4], Dirk Schulze-Makuch[2,3]

1) Department of Earth, Atmospheric and Planetary Science, MIT, 77 Mass. Ave., Cambridge, MA 02139, USA bains@mit.edu

2) School of the Environment, Washington State University, Pullman, WA 99164, USA

3) Center of Astronomy and Astrophysics, Technical University Berlin, 10623 Berlin, Germany

4) To whom correspondence should be addressed

Abstract

The evolution of life from the simplest, original form to complex, intelligent animal life occurred through a number of key innovations. Here we present a new tool to analyse these key innovations by proposing that the process of evolutionary innovation may follow one of three underlying processes, namely a Random Walk, a Critical Path, or a Many Paths process, and in some instances may also constitute a "Pull-up the Ladder" event. Our analysis is based on the occurrence of *function* in modern biology, rather than specific structure or mechanism. A function in modern biology may be classified in this way either on the basis of its evolution or the basis of its modern mechanism. Characterising key innovations in this way helps to identify the likelihood that an innovation could arise. In this paper, we describe the classification, and methods to classify functional features of modern organisms into these three classes based on the analysis of how a function is implemented in modern biology. We present the application of our categorization to the evolution of eukaryotic gene control. We use this approach to support the argument that there are few, and possibly no basic chemical differences between the *functional* constituents of the machinery of gene control between eukaryotes, bacteria and archaea. This suggests that the difference between eukaryotes and prokaryotes that allows the former to develop the complex genetic architecture seen in animals and plants is something other than their chemistry. We tentatively identify the difference as a difference in control logic, that prokaryotic genes are by default 'on' and eukaryotic genes are by default 'off". The Many Paths evolutionary process suggests that, from a 'default off' starting point, the evolution of the genetic complexity of higher eukaryotes is a high probability event.

Keywords: gene control, cellular evolution, innovation, transition, complexity, cell function

1 Introduction

Progression of life from the earliest forms to humans is often characterised as a series of major steps or key innovations, each providing a significant new capability to the newly evolved organisms that was lacking in more primitive forms. Debate on such major steps or innovations was re-ignited in recent times by Maynard-Smith (Smith and Szathmary 1995), who focussed on 'Major Transitions' as defined by changes in the nature of the individual that was the principle subject of selection, and the nature of information transfer within and between individuals. Maynard-Smith and Szathmary's work has been followed by many analyses of what the key steps or innovations are on the path from LCA (Last Common Ancestor) to man, the evolutionary mechanism that lead to them, and how likely they would happen again if the 'tape of life' were rewound (Gould 1989) or if life evolved on another world. But what is a key innovation, and why do they occur?

There is extensive discussion about why many transitions or key innovations in the evolution of life happened, but these discussions are usually framed in terms that are specific to those key innovations. In this paper we suggest a broad approach that classifies the explanation for the appearance of a key innovation into one of three hypotheses. The approach is applicable to any major evolutionary step. Our approach addresses not the specifics of an evolutionary advance in its geological and environmental context, but rather the potential paths to life's acquisition of a new capability. We apply this approach to reviewing one of the key innovations in the evolution of complex life, the evolution of the complex genetic controls of eukaryotes. Eukaryotes alone have developed complex, developmentally regulated multicellularity. This paper addresses the key genetic differences between eukaryotes and all other domains of life, which we will refer to here by the rather old-fashioned term "prokaryotes" for convenience unless referring specifically to the bacteria or archaea. We show that our approach can suggest which aspect(s) of eukaryotic gene control circuitry are the critical differences between prokaryotes and eukaryotes. We tentatively identify control logic rather than any feature of control chemistry as being the distinguishing difference between prokaryotes and eukaryotes. In a subsequent paper we will apply the approach more broadly to the appearance and evolution of life on Earth.

2 Model

A major step in evolution is by definition a rare event. Such an event could be the result of a single, highly unlikely step, or a series of steps. Examples of this second category are well-known as "Multiple Hit" processes from the causation of cancer (Hanahan and Weinberg 2011; Loeb et al. 2003) and other diseases (Bains 2000; Leblond et al. 2012; Polyzos et al. 2012; Swerdlow 2012), and are likely familiar to the reader. Less familiar is that Multiple Hit processes are not uniform. The path to an innovation might require a number of specific innovations or steps. This is functionally equivalent to an innovation that requires a single, highly improbable step to occur, and the probability of a multiple-serial-innovation event is the product of the probabilities of its component steps. By contrast, the path to an innovation may require multiple steps that are selected from a larger pool of possible steps, and not a specific combination of steps. Such a multiple hit process has different probabilities, and hence different kinetics, from a process that requires one, specific combination of steps. To avoid confusion with the general term "Multiple Hit", we term this second class of a Multiple Hit process "Many Paths" - several different paths can lead to the same outcome. The development of cancer is an example of such a process – several genes need to be mutated to cause a cell to become malignant, but many different combinations of mutations can cause malignancy, and can (within limits) be acquired in different orders.

The different types of explanation have significantly different implications for how evolutionary change occurs with time, and the nature of the innovation (and hence the chance that it would occur again if we 'rewound the tape of life' (Gould 1989)) can therefore be inferred from innovation timing, frequency and mechanism, as we will discuss below. Under our schema, an explanation may be:-

1. A Critical Path Hypothesis. The major event or innovation requires preconditions that take time to develop. However the time is (at least mostly) determined by the nature of the event and the geological and environmental conditions of the planet, and so once the necessary preconditions exist on the planet then the event will occur in a well-defined timescale.
2. A Random Walk Hypothesis. The major event or innovation is highly unlikely to occur in a specific time step, and the likelihood does not change (substantially) with time. This may be because the event requires a highly improbable event to occur, or a number of highly improbable steps that have to occur in sequence. Thus, substantial time has to elapse before chance events allow the innovation to be made. Once life exists on a planet, ultimately the innovation will occur, but when it occurs is up to chance, and whether it occurs before the planet's sun leaves the main sequence and renders the planet uninhabitable is not knowable.
3. A Many Paths Hypothesis. The major event or innovation requires many random events to create a complex new function, but many combinations of these can generate the same *functional* output, even though the genetic or anatomical details of the different outputs are **not** the same. So once life exists the chance that the innovation will occur in a given time period is high, but the exact time is not knowable.

Each of these may also fall into a fourth category, termed as "Pulling Up The Ladder". In this class of explanation, an innovation is likely (either because it is a Critical Path or a Many Paths process), but the results of the innovation destroy the preconditions for its own occurrence. The new organisms "pull up the ladder after themselves". The endosymbiotic origin of eukaryotic organelles could be a 'pulling up the ladder' process, because once the eukaryote ancestor had acquired a proto-mitochondrion, there was no opportunity for it to acquire another.

An example of the Critical Path hypothesis might be the argument that complex animal life depends on aerobic metabolism, and hence an oxygen atmosphere. Oxygenating the atmosphere and crust of a planet, so that an atmosphere with a high oxygen content can accumulate, takes a long time, perhaps a billion years on Earth (Schulze-Makuch and Irwin 2008). Thus a geological amount of time might have to elapse between the appearance of oxygenic photosynthesis and the rise of complex animal life (for example, see (Catling et al. 2005)). Once oxygenic photosynthesis has evolved, the evolution of large, complex animals is highly likely, but after a long delay.

An example of the Random Walk hypothesis might be the argument that the rise of the mammals required that the Therapsid precursors of mammals exist and that a diverse set of open ecological niches existed for them to radiate into. The former was true in the Triassic (Bi et al. 2014), but it took a random event (the Chicxulub impact combined with prior rapid climate change at the end of the Cretaceous) to make the latter happen (reviewed in (Archibald 2011)). That impact could have happened at the end of the Jurassic, or during the Eocene, or could not have happened yet.

An example of the Many Paths hypothesis is the evolution of imaging vision (Land and Nilsson 2012). Many genes are involved, and a small number (such as Pax6 (Komik 2005) and the opsins (Collin et al. 2003)) are common to many or even all imaging vision systems, speaking to a common, pre-existing light detection apparatus. However the parallel evolution of the insect, cephalopod and vertebrate eyes using generally different genetic programmes to produce very different anatomical structures shows that functionally equivalent structures for complex imaging can be generated from very different anatomy and genetics, and hence different evolutionary paths.

The reason for classifying innovations in this way is that the three classes of hypothesis have different implications for the likely timing of the events. Fig 1 illustrates the implications of these classes of hypothesis.

1. Critical Path Hypothesis. One set of preconditions is needed for that transition. Once those preconditions ("causes") are satisfied, the innovation will arise quickly, and will occur on all occasions that the preconditions are satisfied. The preconditions take only time to fulfil, there is no (major) random element in it (however the time may be very substantial). As a consequence, if an innovation occurs through a Critical Path process more than once, it is likely to follow a

similar evolutionary path in the different examples. Thus independent evolution of a common function in the descendants of a common ancestor are likely to use similar mechanisms.

2. Random Walk Hypothesis. There are no preconditions other that prior existence of life that can achieve the innovation (e.g. nervous systems cannot evolve without cells). The innovation will occur at random, but since it is highly improbable it will not likely occur twice even if the preconditions are satisfied many times.
3. Many Paths Hypothesis. There are no preconditions other that prior existence of life that can achieve the innovation. However once that precondition is met, the innovation will occur at a fairly reliable time (in generations) afterwards, and so will eventually occur on all occasions that the preconditions are satisfied. If an innovation occurs through a Many Paths process more than once, it is likely to use different mechanisms each time.

A Many Paths process is not functionally the same as a Random Walk process, and, as mentioned above, is only one example of a Multiple Hit process. It is well-known (but still surprising) that if many random events have to occur to cause an output, but many combinations of random events can cause the output, then the timing of the output is more predictable than the timing of any of its component, individual events (see (Bains 2000) and (Bains 2004) for examples of the biological implications of this effect). This is a stronger statement of de Duve's aphorism that "chance does not exclude inevitability" (de Duve 2005). Suggesting a Many Paths hypothesis also has the implication that, if an event can be caused by many combinations of random events, then it will inevitably happen, and it will have a high probability of happening in a defined time period. We can see a way of discriminating between the hypotheses from this formulation. If a major innovation has occurred only once, we might favour the Random Walk hypothesis. If it occurred many times spread through evolutionary time, we may prefer the Many Paths hypothesis. If it occurs many times with a very diverse set of mechanisms, we might also prefer a Many Paths hypothesis. If it occurred many times and in a closely defined time horizon, we may prefer the Critical Path hypothesis. If several independent evolutionary origins result is similar mechanisms (anatomical, molecule, genetic or other), then the evidence is even stronger for a Critical Path hypothesis. Thus we can decouple the overall likelihood of a transition in an evolutionary period from the specifics of how that transition actually occurred, providing we can either identify the time course over which multiple examples of the transition occurred (and match those to Fig 1), or identify multiple paths by which specific function has been acquired (and match these to one of the three models above).

We realize that many real world processes can share features of more than one of these. Thus acquisition of chloroplasts by the Archaeplastida clearly is, to an extent, a critical path process (life had to have evolved the necessary cell types for it to happen), a Random Walk process (the evolution of oxygenesis may be a unique, unreplicated event (Blankenship and Hartman 1998; Holland 2006), a Multiple Path process (any one of many combinations of endosymbionts could have evolved, and indeed many did),

and a Pulling Up the Ladder process. The key for discriminating which interpretation is relevant is, in our view, to focus on function, not structure. Throughout this paper we follow (Smith and Szathmary 1995) in being concerned with *function*, not *mechanism* or *structure*. We are seeking to distinguish analogous structures from homologous ones. For example, the mammalian placenta evolved only once (indeed the placental mammals are defined by their placentae). However placental vivapary has evolved many times (Pollux et al. 2009; Wourms and Lombardi 1992). Some placental reptiles show erosion and invasion of maternal tissue by fetal tissue, resulting in direct fetal contact with the maternal blood stream (Blackburn and Flemming 2012), a feature that used to be thought to be exclusively mammalian. If we define a placenta as the anatomical structure that occurs in mammals, then (by definition) it has only evolved once. But if (after (Mossman 1937)) we define a placenta as "an intimate apposition or fusion of maternal and fetal tissues for sustenance and physiological exchange", then it has evolved many times.

In this paper we illustrate this approach with an analysis of the appearance of the complex architecture of eukaryotic genetic control within the framework of the three classes of hypothesis. We show that the evolution of the major components of eukaryotic genetics are Many Paths processes, which may have relied on a single, specific functional difference between the ur-eukaryote and other life. We argue that the appearance of the specifics of chemistry of eukaryotic gene control, such as RNA modulation of chromatin architecture or multiple alternative splicing, is not the key to its functional capability. The key to the eukaryotic genome is its logic, not its chemistry. To simplify the difference greatly, eukaryotic genes are by default 'off', whereas prokaryotic genes are by default 'on'. Their default 'off' logic allows eukaryotic genomes to be expanded and complexified much more easily than prokaryotic ones, allowing eukaryotes to develop the staggeringly complex genome control architecture we see today.

In order to substantiate this hypothesis we need to show (1) that the difference in control logic is real, and (2) that *essentially all* aspects of the apparent differences in control chemistry between archaea, bacteria and eukaryotes are in fact different structural implementations of the same functionality. This second point requires us to review every one of the many modes of gene control in eukaryotes, which is (the authors admit) a dull and repetitive task. It is however essential to the logic of the argument. The reader who is willing to accept this point with less than an exhaustive demonstration is recommended to skip to section 4, which addresses the first part of our argument.

3 Genetic control and genome complexity

One of the surprises of the human genome project was that humans only have a few more protein coding genes than *Drosophila*, and only about 5 times as many as *E.coli*. This shock to our self-esteem

has been mitigated in the last decade by the realization that a lot of the DNA in complex organisms is related to gene control rather than to protein coding (Washietl et al. 2007), , and that much of the 'junk' DNA is actually functional (as suggested 30 years ago on basis even older data from mutation and radiation studies (Bains 1982); however, see also discussion by (Ponting and Hardison 2011)). All obligate multicellular organisms are eukaryotes, and so it might be postulated that eukaryotic gene control circuitry is uniquely exapted to the evolution of complex life, and as such its appearance represents a bottleneck (a Random Walk event) in the evolution of complex life.

Our objective is not to show that this complexity is not real. Rather, we argue that i) all the basic functionality in eukaryotic gene control has evolved several times with different chemistry in eukaryotes, bacteria and archaea, and ii) the difference between the ur-eukaryote and its prokaryotic contemporaries was genetic logic, not genetic chemistry. The ur-eukaryote probably had a genome not much larger than a modern-day prokaryote (Makarova et al. 2005). The establishment of the extremely complex control genetics of complex eukaryotes by expansion of this original control chemistry is therefore a Many Paths process.

It is now generally believed that the eukaryotic nuclear genome originated from an archaeal-like ancestor (see (Blackstone 2013; Brown and Doolittle 1997; Cavalier-Smith 2010; de Duve 2007; Spang et al. 2015; Weinzierl 2013; Williams et al. 2014) and references therein). Similarities in chemistry between archaea and eukaryotes are explained by common ancestry (i.e. are homologies). Similarly, archaea, prokarya and eukarya evolved from a presumed single ancestor, the Last Common Ancestor (LCA) (Doolittle et al. 1996; Glansdorff 2002), which must have had DNA, RNA and protein synthesis, the structure of a central metabolism, and mechanisms to control all of that machinery. While sequence homology may not have been preserved across 3 billion years, structural similarity (and hence presumed homology) may have been (See for example the actin, MreB and Ta0853 proteins of eukaryotes, prokaryotes and archaea respectively (Roeben et al. 2006)), so finding similar molecules performing similar functions in different branches of life may be evidence of common descent, not independent origin.

Here we do not attempt to add to the literature arguing homology from similarity between major domains of life. Rather, we emphasize similarity of *function* arising from *non-homologous* (and usually non-similar) chemistry in the domain of the control of genes. Thus finding a histone fold protein in archaea and eukaryotes is taken as evidence of their common ancestry. Finding proteins with no sequence or structural homology to histones forming the structural basis of kilobase nucleoprotein architecture in bacteria is very hard to explain by homology, and is more parsimoniously explained by independent evolution of that *function* from a different source chemistry. The presence of imaging lenses in the eyes in mammals and cephalopods does not prove that mammals are cephalopods, nor that their common ancestor had eyes. Rather, it suggests that the evolution of imaging vision is a process that can take

multiple paths, and hence is likely to evolve as a *function* even if the specifics of its implementation are unique to each lineage. We will argue similarly for the components and mechanisms of gene control.

Control of gene activity may be split, entirely for our convenience, into control of transcription, of protein synthesis, and of mRNA and protein turnover. Transcription can further be split into basic RNA polymerase activity, local modulation of transcription, and global control. We will follow this classification, but note that this is not a reflection of a biological hierarchy of control. There is no such hierarchy – all the different processes above are interlinked, such that (for example) ubiquitin-mediated protein turnover directly modulates RNA polymerase activity, the lincRNA-p21 lncRNA induces a wide range of genes, but also suppresses translation, and is degraded by a specific miRNA (Huarte et al. 2010; Yoon et al. 2012). lncRNAs that allosterically regulate cytoplasmic proteins and lncRNAs that are secreted as cell-cell signaling molecules (Geisler and Coller 2013) are known. We will return to the importance of this interlinking below.

Our argument rests on two pillars. We argue that the core functionality of the nucleoprotein structure of eukaryotic gene organization actually evolved independently at least three times, and so is a Many Paths process. We then argue that the types of function that control gene activity within the context of that nucleoprotein structure also have evolved multiple times, although the specific chemical structures that implement the control processes are completely different in each case. We argue therefore that this is also a Many Paths process. These two arguments illustrate the two types of evidence that may support the Many Paths hypothesis.

The next two sections address the first of these assertions, that the core nucleoprotein architecture of life has evolved similar solutions multiple times, and so its appearance is a Many Paths process.

3.1 Core chemical components

The basic chemical components of eukaryotic gene activity are basal to life, and we will only touch briefly on them here. DNA and RNA synthesis clearly are central to the existence of life, and their origin is part of the Origin Of Life problem; indeed, in the "RNA World" model (Gilbert 1986), RNA synthesis was the origin of life. In all domains of life, genes are controlled by proteins and RNA specifically binding to each other and to DNA, and by chemical modification of proteins, RNA and DNA. All of these features have evolved many times, and are applied in many combinations to control essentially the same input-output logic. Thus (for example) catabolic repression is a common 'logic circuit' in the genetic control of metabolism in all domains of life, but has evolved from different proteins and genetic elements in

bacteria, archaea and eukaryotes, and probably evolved several times in each lineage (Bini and Blum 2001).

The chemistry of the modification of the core components of gene control have evolved multiple times. DNA base modification is achieved in bacteria, archaea and eukaryotes by unrelated systems (Cao et al. 2003; Chan et al. 2004; Gaspin et al. 2000; Kumar et al. 1994). Modification of proteins to alter their interactions with DNA by protein methylation (Baumann et al. 1994; Eichler and Adams 2005; Martin and Zhang 2005; Reisenauer et al. 1999), acetylation (Eichler and Adams 2005; Yun et al. 2011), S-glutathionylation (Dalle-Donne et al. 2008), ADPribosylation (Fernando Bazan and Koch-Nolte 1997; Pallen et al. 2001) and poly-ADPribosylation (Haasa and Hottinger 2008) have appeared in diverse clades. RNA modification is similarly diverse and universal (addressed below). It is clear that all domains of life independently developed complex genetics from the same base chemical structures.

3.2 Nucleoprotein evolution: multiple independent origins

The structure of nucleoprotein performs two roles in eukaryotes. The first role is to compact a long genome into a small cell. The second is to control the transcription of that genome, through local and global structure. Below we show that both functions have evolved independently several times, and in some cases through co-opting similar chemistry.

3.2.1 DNA compactification

In all cells DNA condensation is essential to keep the large genome molecule(s) inside the relatively small cell (de Vries 2010; Luijsterburg et al. 2008; Zimmerman and Murphy 1996). The existence of at least three classes of DNA-compacting proteins in prokaryotes (Sandman and Reevem 2001) shows that solutions to this problem have evolved several times (Drlica and Rouviere-Yaniv 1987). Archaea and eukaryotes share DNA binding proteins with the 'histone fold' (Sandman and Reevem 2001), and all domains of life share Alba ("acetylation lowers binding affinity") proteins that bind and compact DNA (Sandman and Reeve 2005; White and Bell 2002). Some of the archaeal nucleic acid binding proteins are similar to proteins (presumed to be homologues) in eukaryotes, where they function as transcription factors (Mantovani 1999). Several unrelated classes of topiosomerases (Champoux 2001; Corbett and Berger 2004), and integrases and recombinases (Argos et al. 1986; Hallet and Sherratt 1997; Sauer 1994) have evolved to manage the resulting topological problems. In eukaryotes and a few bacteria the compaction solution has included an intracellular membrane-bounded compartment for the DNA (Fuerst

2005; Fuerst et al. 1998). While prokaryotic genomes are usually smaller than eukaryotic genomes, prokaryotic organisms can compact as large an amount of DNA into a prokaryotic cell as is compacted into eukaryotic cells, as shown by the existence of 12 magabase bacterial genomes (Chang et al. 2011) and by polyploid prokaryotic cells (Soppa 2014; Zerulla and Soppa 2014) containing hundreds (Griese et al. 2011) to thousands (Mendell et al. 2008) of copies of megabase genomes in expression-specific structures (Komaki and Ishikawa 2000).

Archaeal and bacterial small, basic, DNA-binding proteins can be deleted (Zhang et al. 1996) without inevitably killing the cells: this is not generally true of eukaryotes. However, Dinoflagellates do not have histones, although they have some histone-like proteins similar to those in bacteria (Wong et al. 2003). They seem to use DNA itself as a structural scaffold for very large chromosome-like structures (Bouligand and Norris 2001). Dinoflagellates are considered to be 'living histone knockouts' rather than relics of a primordial, pre-histone eukaryotic gene organization, as their stem groups all have conventional histone chromatin chemistry (Moreno Díaz de la Espina et al. 2005). Mammalian sperm also replace histone with protamines, although the resulting nuclei show minimal gene expression (Braun 2001; Ward and Coffey 1991). Both examples demonstrate that different routes to packaging a eukaryotic genome into a cell are possible.

We conclude that the DNA compaction solution found by eukaryotes is one of a number of equivalent solutions, and as such its evolution represents a likely Many Paths process. The extent to which the different solution are highly derived versions of an ancestral genome packaging chemistry present in LCA is unknown at the moment.

3.2.2 Control by nucleoprotein

Until recently, it was generally believed that eukaryotes control genes through modulation of chromatin structure, and prokaryotes control genes through binding of specific control factors in the classic Jacob and Monod model (Jacob and Monod 1961). This is now understood to be an over-simplification, and that all domains of life use nucleoprotein structure to control genes. This can be done through local or long-range interactions (Luijsterburg et al. 2008).

Eukaryotes have sophisticated mechanism for modulating chromatin structure, which we will discuss in more detail in section 3.3. In this section, we focus on the modification of nucleoprotein structure itself.

In eukaryotes, ATP-driven chromatin remodeling involves swapping a variety of chromatin components, including removing H2A/H2B histone dimers (Kireeva et al. 2002) by the complexes of the SNF2/SWI2 superfamily to open chromatin for transcription (Mizuguchi et al. 2004; Olave et al. 2002; Shen et al.

2000). Yeast have three such ATPases, mammals seven (Olave et al. 2002). Proteins with sequence similarity to yeast and human SWR1 complex proteins have been found in bacteria and archaea, many of which have been identified as helicases[1], i.e. similar, possibly homologous proteins have been coopted to different functions in different domains.

In eukaryotes, this machinery is targeted to a gene by local chromatin structure, especially methylation and acetylation of histones, primarily H3 and H4 (see reviews in (Bracken et al. 2006; Khalil et al. 2009; Martin and Zhang 2005; Mikkelsen et al. 2007)). This epigenetic code is then read by specific suites of recognition proteins, that direct other enzyme activities to the site (Geng et al. 2012).

This is analogous chemistry to DNA-binding-protein modification in prokaryotes, but its targeting is different. Archaeal histones lack the N-terminal tails that are methylated and acetylated in eukaryotes, but many other proteins that are not related to (and hence are presumably not homologues of) histones (Soppa 2010; Wardleworth et al. 2002), including Alba (Marsh et al. 2005), are acetylated *in vivo* in bacteria and archaea (White and Bell 2002).

While the complexity of eukaryotic lncRNA-mediated, long-range control is unique to eukaryotes, a much simpler version of the logic of coordination of structure, clustering nucleoprotein chemistry and gene control, implemented with a quite different chemistry, is found in bacteria. The *E. coli* genome (and a number of other bacteria genomes) is organized into loops of ~10kb by DNA binding proteins. One of the best characterized is the small, basic protein H-NS. H-NS organizes two groups of genes scattered around the *E.coli* chromosome into spatially close clusters of co-regulated genes. H-NS binds to DNA (Navarre et al. 2007; Navarre et al. 2006), oligomerizes to bring the distant genes together into one of two close physical clusters (Fang and Rimsky 2008; Wang et al. 2011), and co-ordinates their transcription by facilitating binding of RNA Polymerase (Pol) and accessory regulatory factors (Zhang et al. 1996). H-NS silencing is countered by a variety of mechanisms, including competition from similar proteins, but none involving modification of H-NS (Fang and Rimsky 2008). During DNA replication, the H-NS-coordinated loops are assembled fast, and in a specific order and position in the cell (Viollier et al. 2004). The similarity of all these features to the chromatin features of eukaryotes is obvious, albeit H-NS organizes far fewer genes over smaller distances and for simpler controls. Unrelated proteins appear to perform the same role as H-NS in *B. subtilis* (Smits and Grossman 2010).

In conclusion, nucleoprotein is found in bacteria, archaea and eukaryotes, and both its role as a DNA compactification system and its role in gene control is found in all kingdoms. The use of different proteins to achieve the same result shows that this was an independent origination of nucleoprotein function in

[1] WB personal observation from BLAST searches using NP_011365.1 (yeast INO80), EDN63720.1 (yeast SW1/SNF) BAG10015.1 (human INO80) and BAG10565.1 (human SW1/SNF) on NCBI protein database excluding Eukarya from the target database.

the different domains of life. We conclude that the appearance of nucleoprotein as a substrate for gene control is a Many Paths process.

In the next section, we address the more complex question of whether the mechanism of the control of eukaryotic nucleoprotein is a unique set of capabilities, i.e. probably the outcome of a Random Walk process, or whether it can be considered a Many Paths event as well.

3.3 Specific mechanisms of gene control in eukaryotes

Eukaryotic gene control is astonishingly complex. It is hard to imagine that its origination was not a uniquely unlikely event. In this section we argue that this complexity hides a wealth of duplication, independent origination, and shares a wide range of features with unrelated genetic control systems in prokaryotes. In short, the specifics of gene control in eukaryotes show all the features of the outcome of a Many Paths process.

In order to control a gene, chemical function must be targeted to that gene. For convenience, we will consider local, distant and global control systems in turn (i.e. systems that act at or within a few helical turns of the start of a gene, at hundreds or thousands of bases from a gene, or that affect every gene, respectively). We consider how gene control is achieved by considering what is doing the targeting. In every case, we will show that there are multiple, independently derived mechanisms that have evolved to achieve the same goal in different organisms, supporting a Many Paths process.

3.3.1 Local targeting by DNA

DNA is not used widely as a targeting moiety in 'normal' genetic function in any domain of life. DNA is usually considered as the target of genetic specificity, not the specifying agent (although this ultimately is a rather semantic distinction). Some integrating DNA viruses and transposons use DNA:DNA interactions as a way to target change to a specific region of the genome. Site-specific recombination in vertebrate immune systems uses short 'joining signals' that are necessary and sufficient to direct enzyme-catalyzed recombination of antigen-binding gene precursors (Lewis and Gellert 1989). Recombination systems, directed by DNA sequence, are the mechanism for chromosome terminus replication in some organisms (Levis et al. 1993; Vaillasante et al. 2008). Some bacterial Crispr/Cas systems use DNA targeting (Cao et al. 2003; Chan et al. 2004). So despite their rarity, DNA targeting has evolved independently several times.

3.3.2 Local targeting by protein.

Direct recognition of genetic elements by proteins has evolved many times in bacteria, archaea and eukaryotes. The different categories of RNA polymerase, the several classes of structurally distinct DNA-binding proteins, which are found in prokaryotes and eukaryotes (reviewed in (Landschulz et al. 1988; Schwabe et al. 1993)), DNA synthesis initiation factors from bacteria (Messer 2002), all attest to the multiple, parallel evolution of proteins with affinity for specific DNA sequences. We also note that DNA synthesis can be primed by specific proteins (i.e. DNA:protein targeting) in bacterial phages and eukaryotic viruses (Salas 1991), where it presumably evolved independently.

DNA can also be indirectly targeted through sequence-specific chemical modification and subsequent recognition of the modified DNA by proteins that have limited sequence specificity or are sequence agnostic. 5-methyl cytosine is the best known of these modifications, and is generated in all three domains of life by diverse, non-homologous enzymes (Kumar et al. 1994), but 5-hydroxymethylcytosine (Tahiliani et al. 2009), 5-hydroxythymidine (Cliffe et al. 2009), 6-methyl adenine (Wion and Casadesus 2006) are also common across the three domains, and are linked with a diverse range of species-specific genetic controls as well as general cell processes such as DNA replication (see eg (Wion and Casadesus 2006)).

We consider it obvious that direct recognition of DNA and RNA sequences, and of proteins, by proteins has evolved many times, and that the evolution of any specific function achieved by DNA:protein or RNA:protein binding is a Many Paths process.

3.3.3 Local targeting by RNA in eukaryotes

Both prokaryotes and eukaryotes use RNA extensively in genetic control chemistry, with metazoa transcribing the majority of their genes into non-coding RNA that is believed to be associated with control function (Washietl et al. 2007). No class of small RNA has a unique function in any domain of life: all have been co-opted from their 'original' function to new ones. The development of RNAi in potential therapeutics (Vaishnaw et al. 2010) has accelerated the understanding of short regulatory RNAs, which therefore are classified into several functional classes (Joshua-Tor and Hannon 2010) (Cech and Steitz 2014). The longer transcripts are simple called Long Non-Coding RNAs (lncRNAs), a designation that everyone accepts is unsatisfactory, but is inevitable because the function of nearly all these transcripts is unknown. lncRNAs have diverse evolutionary origins (Ponting et al. 2009), some are strongly conserved

between species which suggests that they have an essential function (Guttman et al. 2010; Nagano and Fraser 2011; Ponting et al. 2009; Ulitsky et al. 2011) and are not 'junk DNA' (Pagel and Johnstone 1992) (although see (Rebollo et al. 2012)). The roles and mechanisms of a few lncRNAs have been identified (discussed below). It is generally believed that some, probably most of the lncRNAs are involved in genome control (see (Geisler and Coller 2013; Mattick and Gagen 2001; Meister and Tuschi 2004; Mello and Conte 2004; Rinn 2012; Wang and Chang 2011) for additional reviews of lncRNA biology). Regulatory RNAs are also being discovered in bacteria, both short transcripts (Waters and Storz 2009) as well as 'classical' longer antisense RNAs.

Because RNA-based gene controls are so much more extensive in complex eukaryotes than in other organisms, we will dwell more exhaustively on this category of control mechanism. RNA-based targeting systems that use small RNA molecules (<100 bases) are usually classified by their mechanism, related to the protein complexes associated with the RNAs. These broadly classify as follows.

piRNA (Piwi complex-associated). These primarily suppress transposon activity in the germ line of multicellular animals by directing DNA methylation (Malone and Hannon 2009), but are also used for sex determination in *Paramecium* (Singh et al. 2014) and silk moths (Kiucho et al. 2014), in replacement for the protein factors that determine sex in many other species. The Piwi proteins and associated target RNAs are widely expressed outside the germline in diverse organisms (Ross et al. 2014). Transposons are also silenced through the protein-mediated DNA modification system in eukaryotes (Tahiliani et al. 2009), which has also been repurposed as part of the antigenic switching machinery in some trypanosomes (Cliffe et al. 2009). piRNAs and associated proteins excise transposons from the ciliate macronucleus in a mechanism reminiscent of the bacteria CRISPR/Cas (Chalker and Yao 2011).

miRNA are short hairpin RNAs (Meister and Tuschi 2004) and are a major controller of metazoan mRNA stability via the RISC complex (Meister and Tuschi 2004; Nykänen et al. 2001). The Argonaut protein of the siRNA-processing DICER complex is closely similar to the Ago protein of archaea (Song et al. 2004): however, whether they are homologues is still contentious. The function of prokaryotic Ago is not known, but genomic context suggests it is part of a viral defense system parallel to CRISPR/Cas which is similar in role to siRNA in eukaryotes (Makarova et al. 2009). lncRNAs also input into this system (Ruthenburg et al. 2007).

The miRNA repertoire on plants and animals appear to have evolved independently from a common basic mechanism: subsequent bewildering complexity in both lineages has evolved by duplication and divergence of the various components (Shabalina and Koonin 2008). Ctenophores have no miRNA system (Moroz et al. 2014).

siRNA system degrades unwanted transcripts, primarily (in metazoa) viral sequences (Malone and Hannon 2009), but also some endogenous sequences in yeast called Cryptic Unstable Sequences (CUTS)

(Houseley 2012; Shabalina and Koonin 2008). The RNAse III component of the siRNA system has analogous proteins in proteobacteria. In *Cryptococcus* siRNA targets transposon transcripts, using a complex (SCANR) that has proteins similar to a spliceosome protein in eukaryotes (Dumesic et al. 2013). Bacteria have no RNAi system, but some small bacterial RNAs modulate mRNA stability through completely different mechanisms (Görke and Vogel 2008).

For controls through short RNA sequences in eukaryotes, we therefore see

- Multiple, independent evolution of similar systems (eg animal and plant miRNA)
- Chemically different systems achieving the same functional goal (eg sex determination on silk moths vs vertebrates)

Both are hallmarks of a Many Paths process.

3.3.4 Targeting by RNAs in prokaryotes

Prokaryotes are now understood to transcribe dozens or hundreds of non-coding RNAs (Livny et al. 2006; Rivas et al. 2001; Vockenhuber et al. 2011), most of which modulate translation. Most require protein factors such as Hfq to assist imperfect base pair recognition of target RNAs (Papenfort and Vogel 2010; Waters and Storz 2009). Some longer bacterial antisense RNAs span several genes or operons with related function, and provide operon-scale translational control from a single molecule (Sesto et al. 2013). A number of longer bacterial RNAs are now known to have multiple regulatory functions (Papenfort and Vogel 2010), analogous to some locally acting lncRNAs in eukaryotes.

The CRISPR/Cas bacterial system also uses small RNAs as guides for nucleic acid destruction (Jore et al. 2012), but use a different enzyme machinery than RNAi (Hale et al. 2009). Some target incoming ssRNA for destruction, exactly analogously to the RNAi system (Horvath and Barrangou 2010). Others CRISPR/Cas chemistries target dsDNA of invaders, in a mechanism reminiscent of (but evolutionarily unrelated to) small RNA-directed *de novo* DNA methylation in eukaryotes (Cao et al. 2003; Chan et al. 2004).

RNA also guides base modification enzymes in prokaryotes and eukaryotes. In general, pseudouridine is inserted into rRNA with site-specific enzymes in bacteria, but with broad-specificity enzymes guided by snoRNAs into rRNA (Lafontaine and Tollervey 1998) and mRNA (Carlile et al. 2014) in eukaryotes. Archaea use an intermediate system that comprises guide RNAs and specific proteins, including sequence relatives of eukaryotic snoRNA (Aittaleb et al. 2003), for rRNA O-methylation (Bachellerie et al. 2002; Gaspin et al. 2000). Some promoter-specific DNA modification is guided by siRNA (Cao et al. 2003; Mello and Conte 2004), although much is guided by chromatin structure (discussed below). Again, bacteria use

sequence-specific proteins for *de novo* methylation of DNA, which have regulatory roles as well as roles in the phage defense/restriction systems

Local RNA control is therefore not specific to eukaryotes but has evolved independently in prokaryotes. The elaborate local control systems found in eukaroytes carry out overlapping and mutually replaceable functions, which in at least some cases have evolved independently, and other chemical mechanisms to achieve the same role have evolved independently in different lineages.

3.3.5 Controls via long RNAs

Higher eukaryotes are unique for their extensive use of long RNAs (arbitrarily, >100 bases) that control gene expression by control of the short- and long-scale chromatin structures. lncRNA can coordinate gene activity locally, or over megabase distances through chromatin folding in eukaryotes (de Santa et al. 2009; Lettice et al. 2003; Nagano and Fraser 2011; Ørom et al. 2010; Smemo et al. 2014; Yao et al. 2010), and the strength of enhancement is not related to the length of the loop (Sanyal et al. 2012; Wang et al. 2013), so this is not an extension of local gene control to longer distances. lncRNA scaffolding targets enzymatic activities to different regions of the genome (reviewed in (Mercer and Mattick 2013; Nagano and Fraser 2011; Rinn 2012; Wilusz et al. 2009)). Typically lncRNAs will interact with many loci across the genome (Nagano and Fraser 2011), with gene activity requiring a combination of loop topology, protein binding and appropriate chromatin tags in a broad sequence context (Domené et al. 2013; Jin et al. 2013; Taher et al. 2011; Taher et al. 2012; Wang et al. 2013). This level of control interacts with more local levels of control, for example by directing chromatin modifying enzymes to specific regions of the genome (Mercer and Mattick 2013).

lncRNAs target chromatin modification enzymes to add 'tags' to chromatin: the chromatin tags in turn are targeted by protein, small and large RNAs. The Polycomb system proteins (PCGPs), that methylate histones over short (Müller and Kassis 2006) or long (Lee et al. 2006; Schwartz et al. 2006; Wang and Chang 2011) distances, are targeted by direct binding to promoters or repressors or recruited to chromatin by lncRNAs which bind both PGCPs and either other proteins or other RNAs (Khalil et al. 2009; Nagano and Fraser 2011; Wilusz et al. 2009). Although Polycomb is usually associated with gene repression, in some cases it has been recruited to gene activation pathways (Gao et al. 2014). Many of these systems are multi-component complexes or multifunctional molecules that recognize a combination of chromatin features (Ruthenburg et al. 2007). lncRNAs can also recruit histone modifying enzymes independently of PolyComb (Camblong et al. 2007; Houseley et al. 2008). After the transcription bubble has passed, Pol-II recruits proteins to epigenetically tag transcribed sequences, so as to repress promoter sequences occurring within the gene (Whitehouse et al. 2007; Yadon et al. 2010). As a side-effect of this,

transcription of one RNA blocks transcription of a downstream overlapping transcript, yet another role of an RNA controlled process (Thebault et al. 2011).

All of these interact and compete with each other for binding to target proteins and microRNAs (Tay et al. 2014). However this bewildering catalogue of complexity does not imply anything unique in eukaryotes, only the extraordinary expansion of capabilities seen to evolve in other systems (siRNA, simple repressor-operator systems) or other domains of life. All the processes which we summarize very briefly above have precedence in preceding paragraphs in terms of targeting DNA and protein modification, DNA transcription, and the associated enzymes through interactions of DNA, RNA and protein with each other, sometimes in complexes that link distant genes. What is different in metazoan genomes is the amount of this activity, not its nature.

3.3.6 Splicing and other RNA roles

RNA splicing is found in all domains of life and it is likely that splicing was a mechanism that LCA had already evolved. Self-splicing Group II introns are not present in eukaryotic nuclear genome (Edgell et al. 2011), splicesomal introns only in eukaryotes (see (Dumesic et al. 2013; Martin and Koonin 2006; Pyle 2012; Roy and Gilbert 2006; William and Gilbert 2006) for reviews on the origin of splicesomal introns). While RNA-catalysed self-splicing is the distinctive hallmark of these introns, their splicing *in vivo* require protein maturases that accelerate splicing chemistry and act as RNA chaperone proteins (reviewed in (Lambowitz and Zimmerly 2004; Meng et al. 2005)). Fourteen nuclear encoded proteins are required for splicing *Chlamydemonas* chloroplast Type II 'self-splicing' introns, most of which share no similarity with the components of the nuclear spliceosome (Rivier et al. 2001). Combined protein and RNA machinery to rearrange RNA appears to have evolved multiple times, with varying degrees of dependence on the protein component (Meng et al. 2005).

The unrelated mechanism of translational skipping has a similar net effect to RNA splicing; the generation of a protein from non-adjacent regions of a transcript. Short ribosomal frameshifting is found ubiquitously, with different chemistry in different domains speaking to different origins (Belew et al. 2014; Brierley et al. 1989; Chandler and Fayet 1993; Cobucci-Ponzano et al. 2005; Dinman 2012; Lang et al. 2014)

RNA is also central to priming DNA synthesis at specific sites in prokaryotes and eukaryotes. However a large number of phage and eukaryotic viral examples show that proteins can also prime DNA synthesis (Salas 1991). RNA plays structural roles in ribosomes, telomerase and other structures (reviewed in (Cech and Steitz 2014)), but as these are common to all the forms of life that use those structures, we assume these are homologies, not analogies.

RNAs can compete with DNA binding proteins, including RNA polymerase in *E.coli* (Wassarman 2007) but also factors such as steroid receptor proteins in mammals, thus titrating their activity (Martianov et al. 2007; Poliseno et al. 2010; Salmena et al. 2011). tRNAs can also play this role (Kino et al. 2010). The RNAs concerned share no sequence similarity.

Lastly, the interaction of RNA with small molecules to modulate RNA function ('riboswitches') has evolved independently many times in all three domains of life (Breaker 2012; Coppins et al. 2007). Riboswitches are built from a large number of distinct motifs with limited or no sequence similarity, some broadly distributed across bacteria and archaea, some quite specific to smaller groups of organisms (Weinberg et al. 2010).

We conclude two things from this very short survey of RNA-based gene control:

i) That RNAs that can bind to DNA, to other RNAs or to proteins to control transcriptional activity have evolved many times

ii) That the functions carried out in metazoa by specific classes of RNA can be carried out by many classes of RNA in different organisms, and in many cases their functions can be performed by proteins in bacteria.

From this we infer that the origin of *function* of the RNA-based chemistry of gene circuitry was a Many Path process, with many potential outcomes that would permit the subsequent Critical Path complexification of the full metazoan gene control circuitry.

3.4 Other expression controls

All branches of life have a wealth of (unrelated) transcription factors to control the process of initiation of RNA synthesis (Baliga et al. 2000; Bell and Jackson 2001). The transcription initiation complex in all domains shows DNA:protein as well as protein:protein interactions, with the overall architecture more similar between archaea and eukaryotes than between bacteria (with respect to the HTH Sigma family of proteins (Helmann and Chamberlin 1988)) and archaea (Soppa 2001; Weinzierl 2013), and consequent similarities in promoter sequences (Miller and Hahn 2006; Rhee and Pugh 2012). The dynamic initiation complex can be as large as a ribosome (Liu et al. 2013). There is some sequence similarity between the polymerases and some of the accessory factors between all domains (Bartlett et al. 2000; Bell and Jackson 2001), but others have evolved independently.

The transition for initiation to elongation can also be a point of control (Nechaev and Adelman 2011) as can termination. RNA polymerase can bind to a promoter and then 'stall' (Core et al. 2008; Muse et al. 2007; Nechaev et al. 2010) through several, different mechanisms (FitzGerald et al. 2006; Hendrix et al.

2008; Li and Gilmour 2013). RNA can interact with the translation process in a variety of ways to modulate translation – similar effects using completely different mechanisms have evolved for RNA modulation of translation in bacteria and eukaryotes (Grigg and Ke 2013; Valencia-Sanchez et al. 2006).

Elongation and termination require specific protein complexes, both have substantial differences in the three domains of life (see (Braglia et al. 2005; Jeong et al. 1995; Mischo and Proudfoot 2013) and refs therein): one eukaryotic termination system has similarities to PolyComb (Camblong et al. 2009), and formation of R-loops over G-rich terminators induce antisense transcription of the recently-transcribed gene, and hence recruitment of histone methylation and siRNA mechanisms (Skourti-Stathaki et al. 2014). RNA degradation is also controlled, often through polyadenylation. Although the core polynucleotide phosphorylase activity has similarity (and hence presumed homology) between archaea, bacteria and eukaryotes, the polyadenylation complexes differ, and yeast at least has two different polyadenylation-based RNA degradation systems for 'correct' and aberrantly folded RNA (reviewed in (Houseley and Tollervey 2008)).

Protein turnover is modulated by a range of systems in prokaryotes (Battesti and Gottesman 2013) and eukaryotes (Geng et al. 2012; Pickart 2001). Protein abundance in eukaryotes (in mammals anyway) is controlled mostly at the level of translation, not protein breakdown (Schwanhausser et al. 2011). The role of the ubiquitin-proteasome system (now known to involve a range of protein tags) is mostly to clear degraded or misfolded proteins (Geiss-Friedlander and Melchior 2007; Schwartz and Hochstrasser 2003).

Thus the same general points apply – there are many systems, overlapping controls, and independent origins for many of them in different lineages.

4 The evolutionary step to eukaryotic gene control

In the previous section we have emphasised that

i) There are multiple types of control of gene activity in eukaryotes that overlap with each other
ii) That different control functions evolved many times, even if their specific chemistry is unique to each example, and that the same general type of genetic function is often carried out by different chemistries in different organisms
iii) That many types of control chemistry in eukaryotes have precedent in bacteria or archaea.

We argue that this shows that the evolution of the complex genome of (say) the metazoa is a Many Paths process, one that takes time but is highly likely to happen. If this is so, why are eukaryotes so obviously more genetically more complex than prokaryotes? Why are there not prokaryotes with as complex genomes as, for example, *C. elegans*?

We do not have a robust answer to this, but our analytical approach suggests a direction for hypothesising. One core feature of eukaryotic gene control apparently appeared once early in eukaryotic evolution, and has not appeared in other lineages. Archaeal chromatin and bacterial DNA compaction proteins do not (in general) block transcription (Weinzierl 2013; Xie and Reeve 2004), unlike eukaryotic nucleosomes. (Though some archaeal histone-like proteins inhibit transcription in *in vitro,* these systems are not exact models for the *in vivo* case (Chang and Luse 1997; Soares et al. 1998)). Transcription of prokaryotic genes is under the control of sequences that recruit RNA polymerase to a gene, or recruit polymerase-recruiting or blocking proteins in bacteria and in archaea, despite the latter's having RNA polymerase complexes similar to those in eukaryotes (Geiduschek and Ouhammouch 2005; Reeve 2003). By contrast, in eukaryotic organisms there is a global repression system for all genetic activity, and transcription of eukaryotic DNA requires relieving this global repression by energy-consuming modification of chromatin (Kireeva et al. 2002; Mizuguchi et al. 2004; Olave et al. 2002; Shen et al. 2000) as well as sequence-specific recruitment of specific transcription factors (Kireeva et al. 2005; Li et al. 2007). In short, the logic of eukaryotic chromatin is at a default 'off' state, whereas the nucleoprotein in other domains of life is at default 'on'.

This is reflected in several functional observations. Expression vectors are engineered with specific genetic elements to ensure high levels of gene transcription. Those that function in bacteria and archaea can rely on chromosomal promoters alone, even if they integrate into the genome (although the T7 phage promoter is also popular in *E.coli*), and other genetic elements are included only to block transcription through repressor/operator control circuits. By contrast, eukaryotic vectors almost always have to contain viral promoters that have evolved to abrogate chromosomal gene control, and also additional viral enhancer sequences or complex chromatin control elements to enhance transcription (McCarty et al. 2004; Miller 1992) even if they replicate as episomes (Mumberg et al. 1995) (summarised in Fig. 2). In bacteria, simply having promoter sequences that recruit transcription enzymes is sufficient to ensure transcription; in eukaryotes additional sequences to flag a sequence as transcribable are required.

Gene duplication is common in all domains of life, but in eukaryotes duplicate genes that have mutated to become pseudogenes are often retained in the genome (Mighell et al. 2000; Vanin 1985), whereas in bacteria and archaea they rarely are (Liu et al. 2004). In eukaryotes a high pseudogene load is the mark of a large genome, in prokaryotes it is the mark of the highly degraded genomes of evolving parasites, such as *Mycobacterium leprae* and *Yersinia pestis* (reviewed in (Bentley and Parkhill 2004)). Even r-strategist eukaryotes like yeast have ~5% of their genome as pseudogenes (Harrison et al. 2002). We see this as supporting evidence for basic differences in control logic. In eukaryotes a gene is by default 'off' unless specifically activated, so pseudogenes are almost all transcriptionally silent (Zheng and Gerstein 2007) and hence of little phenotypic relevance. In prokaryotes a gene with a promoter attached

is by default 'on' unless repressed, and so a mutated gene will have a significant chance of producing an aberrantly functional protein, and pseudogenes are observed to be efficiently selected against (Kuo and Ochman 2010).

Weaker but still intriguing support for the idea that mammalian genes are by default 'off' comes from somatic cell fusion experiments. Two decades of this now neglected area of research showed that if two cell lines that show different, differentiation-specific gene expression patterns are fused, the differentiation-specific genes that are expressed in only one originating cell line are usually not expressed in the hybrid (reviewed in (Gourdeau and Fournier 1990; Weiss 1982)). This phenomenon (termed extinction) suggests that in competition between the expression status of a particular gene between the two genome states ('on' in one cell, 'off' in the other), the 'off' state is usually dominant. Derivative cell lines that have lost chromosomes often show re-expression of the differentiated phenotype, showing that the epigenetic imprinting of the differentiation-specific genes is not over-written, it is just suppressed by more powerful 'off' signals. Immortalised cell lines and cell fusion are not physiologically normal states, but the observation supports our general thesis.

Why is this relevant, if the chemistry of gene control can evolve multiple times? In complex organisms, all of the control systems described above interact with each other to define cell- and tissue-specific gene expression patterns (and hence phenotypes). In examples ranging from the mammalian development of white and brown fat (Peirce et al. 2014), neurogenesis (Jobe et al. 2012; Schouten et al. 2012) to yeast mating type loci control (Buhler and Moazed 2007; Grewal and Rice 2004) we see all of miRNA, piRNA, protein transcription factors, specific DNA sequence elements, histone methylation and acetylation used in a spaghetti code of interactions to define the biological endpoint. Even apparently highly specific enzymes such as telomerase are found, on closer examination, to have multiple roles in gene control (Li and Tergaonkar 2014).

The complexity of genetic circuits is therefore not just a function of the number of coding and regulatory elements, but of the number of ways they can interact, so that the number of distinct genetic programs is a polynomial function of genome size. It was a well-known observation from the dawn of molecular genetics that most of the genome is not transcribed in most cells of a multicellular body, nor in single celled organisms most of the time (see for example (Chu et al. 1998; Ghaemmaghami et al. 2003; Menssen et al. 2011; Narlikar et al. 2010; Rabbani et al. 2003; Yamashita et al. 2000)). To add a new set of genes to a genome, not only must a unique control network for that gene set be created, but a way of *not* activating all the other genes in the genome must be implemented as well. If the default status of genes is 'off' then this second task is already achieved. If the default state of the genes is 'on', then the first task is easier, but the second requires modulation of the control system for every other gene in the

organism [2]. We note that in eukaryotes (animals, anyway) general release of the chromatin-mediated repression of genes is profoundly toxic ((Frost et al. 2014), and references therein).

Thus we postulate that the evolution of a genome in which the default expression status was 'off' was the key, and a unique, innovation that allowed eukaryotes to evolve the complex control systems that they show today, not the evolution of any of those control systems *per se*. Whether the evolution of a 'default off' logic was a uniquely unlikely, Random Walk event or a probable, Many Paths event is the subject of future work.

5 Summary and conclusions

In this paper we present a simple classification of evolutionary innovations based on what sort of process leads to the appearance of the *function* that those innovations provide. We suggest that the process of innovation may be classified into

- Random walk (improbable, unlikely to be duplicated)
- Many Paths (probable, likely to be duplicated through different mechanisms)
- Critical Path (probable, likely to occur multiple times in the same form)

We have sketched the vast field of gene control chemistry to show that all the key functions of gene control in eukaryotes are carried out by multiple classes of molecules, that similar molecules have adopted different functions in eukaryotes, prokaryotes and archaea, and that there is good evidence for the independent evolution of many control chemistries and processes in different lineages and domains. All these observations support the idea that the development of eukaryotic gene control circuitry was a Many Paths process. Many Paths processes are highly likely to occur within a defined time 'window' given suitable environmental conditions; the timescale depends on the pace of the underlying individual component innovations, and the width of the 'window' depends on the number of possible innovations and the number of actual innovations needed to achieve the overall function (discussed in more detail in (Bains 2000; Bains 2004)). As the timing of the appearance of both eukaryotes and of complex, multicellular genomes is controversial, it will be hard to constrain either timescale or window. However

[2] We realise that this is a rather simplified view of the constraints on the evolution of gene control mechanisms. One could, for example, imagine the evolution of a mechanism in a prokaryote from an RNA that interacted with no genes to one that interacted with only one, thence with two and so on. However a myriad of *in vitro* protein and RNA evolution experiments show us that it is easier to find a macromolecule that interacts weakly with many things, and then refine its specificity by selection, than it is to find a macromolecule that interacts specifically (and hence tightly) with just one molecule in one step.

the analysis does not depend on doing so, and does suggest that evolution of a complex genome comparable to a modern plant or animal was not an unlikely outcome given the origin of life.

We suggest that a key difference between prokaryotes and eukaryotes is that the nucleoprotein of prokaryotes is by default open to transcription ('on'), while that in eukaryotes is by default transcriptionally inactive ('off'). From the appearance of this 'default off' state, the evolution of complex genomes was a likely, Many Paths process.

We wish to emphasise that our analysis does not remove chance from large-scale evolution. The Chicxulub impact did have a profound impact on macro-faunal evolution (Archibald 2011). However we should not over-glamorise these unique events, nor postulate that other, unseen unique events are key to evolutionary innovation. The specifics of chemistry and topology of individual eukaryotic genomes is undoubtedly both unique and extremely unlikely to evolve twice. The evolution of complex genetic controls in eukaryotes was not deterministic. But the evolution of complex genomes was highly likely.

6 Acknowledgements

We are grateful to Janusz Petkowski (ETH Zurich, Switzerland) for helpful comments on an earlier formulation of this paper, David Simpson (Glythera Ltd, UK) for help with expression systems, to several anonymous readers and reviewers for their comments, and Mike Danson (University of Bath, UK) for pointers to the literature on archaeal expression systems.

The authors received no funding for this work, and have no conflict of interest to declare.

Fig 1: Timing implications of the three models

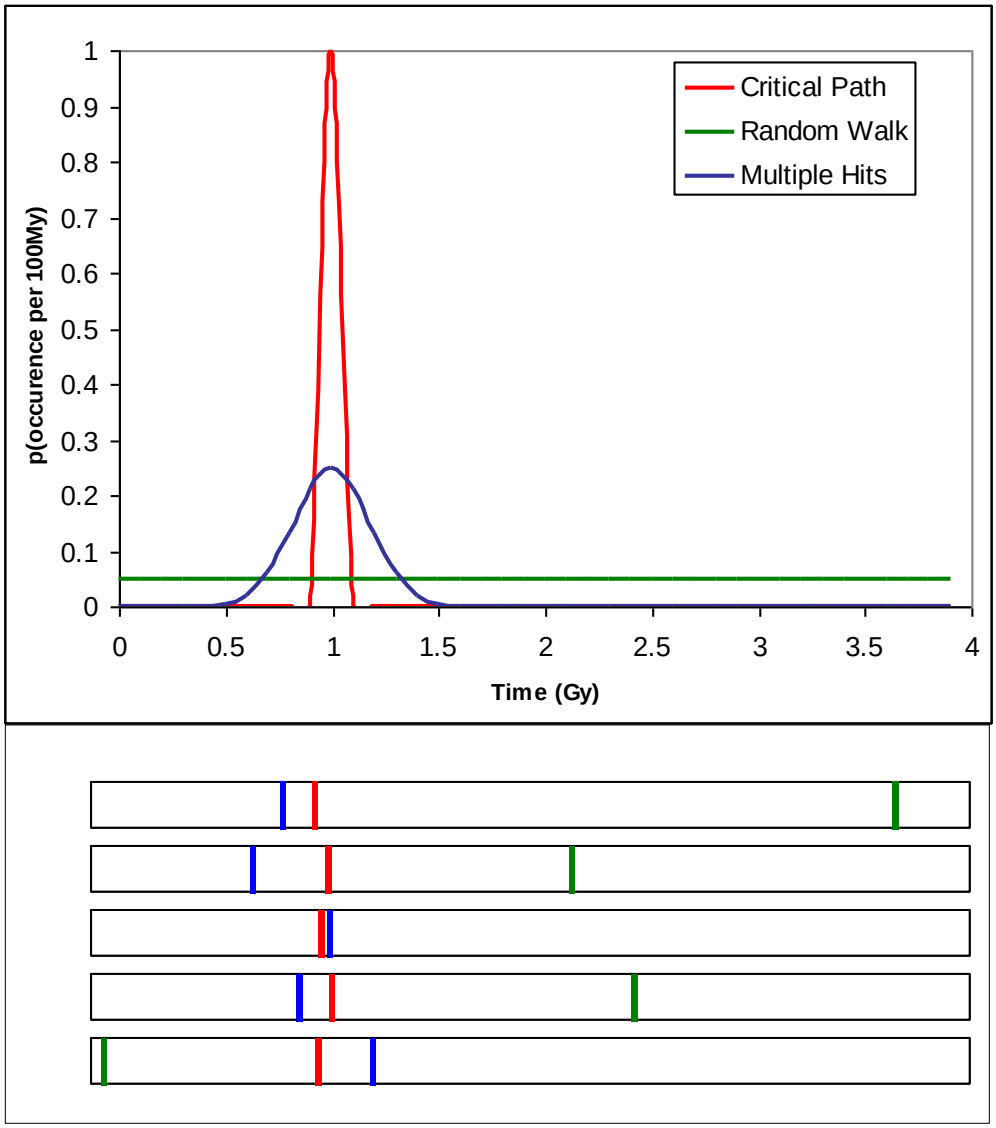
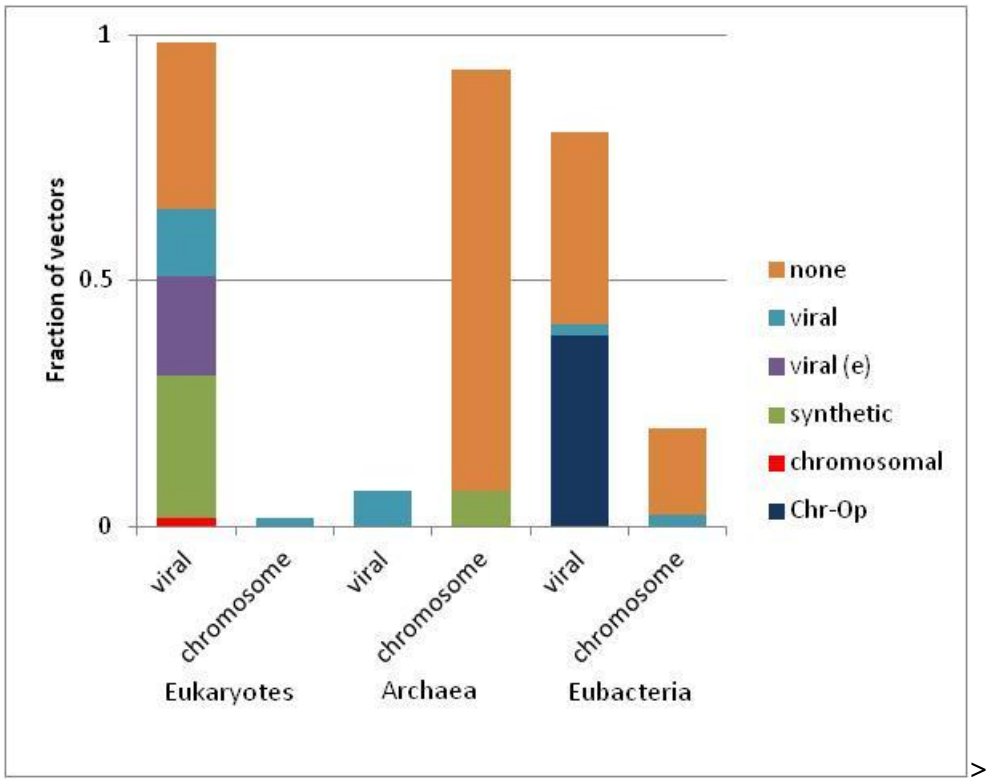


Illustration of our test for the hypotheses explaining the major innovations. The figure illustrates *one* specific transition, and its probability under *three models.*

Top panel, the probability that a transition will occur in any specific 100 My period (Y axis) after a given number of Gigayears (X axis). Critical Path (red) – will inevitably occur once a specific environmental threshold has been passed at 1Gy. Random walk (green) – equal probability of happening at any time. Many Paths (blue) - will probably happen around 1 Gya.

Lower panel – five independent lineages (clades within one planet or life on different planets) and the time of occurrence of that *one* transition under the Critical Path hypothesis (red), Random walk hypothesis (green) and Many Paths hypothesis (blue). Note that in lineage 3 the transition has not occurred at all in the time available under the Random Walk hypothesis.

Fig. 2: Overview of expression vector components



Summary of features included in 159 expression vectors that drive expression of inserted genes. Vectors were classified as to whether the primary promoter sequence was from viral or chromosomal. Vectors are also classified as to whether other, additional elements controlling expression levels were present – viral promoters, viral enhancers ('viral-e'), synthetic enhancer elements including complex chromatin modulating synthetic segments (Williams et al. 2005) ('synth-e'), chromosomal elements and chromosomal operator elements of the Lac-operon, negative regulatory type (Chr-Op), or no additional elements over the base promoter ('none'). Vectors are shown by the Domain in which they are designed to express protein. Data of mammalian and bacterial vectors from (EMBL 2015a; EMBL 2015b; Merck Millipore Inc 2015; Promega Corp. 2015), Archaeal vectors from (Albers et al. 2006; Allers 2010; Allers et al. 2010; Aravalli and Garrett 1997; Contursi et al. 2003; Lucas et al. 2002; Peng et al. 2012; Santangelo et al. 2008; Schreier et al. 1999; Stedman et al. 1999; Zheng et al. 2012). Multiple families of vectors with essentially identical control systems and differing only in gene insertion sites or selectable genes are counted as one entry.

7 References

Aittaleb M, Rashid R, Chen Q, palmer JR, Daniels CJ, Li H (2003) Structure and function of archaeal box C/D sRNP core proteins Nature Structural Biology 10:256 - 263

Albers S-V et al. (2006) Production of Recombinant and Tagged Proteins in the Hyperthermophilic Archaeon Sulfolobus solfataricus Applied and Environmental Microbiology 72:102-111 doi:10.1128/aem.72.1.102-111.2006

Allers T (2010) Overexpression and purification of halophilic proteins in Haloferax volcanii Bioengineered Bugs 1:288 - 290

Allers T, Barak S, Liddlell S, Wardell K, Mevarech M (2010) Improved Strains and Plasmid Vectors for Conditional Overexpression of His-Tagged Proteins in *Haloferax volcanii* Applied and Environmental Microbiology 76:1759 - 1769

Aravalli RN, Garrett RA (1997) Development of a Simvastatin Selection Marker for a Hyperthermophilic Acidophile, Sulfolobus islandicus Extremophiles 1:183 - 191

Archibald JD (2011) Extinction and radiation: how the fall of the dinosaurs led to the rise of the mammals. The Johns Hopkins University Press, Baltimore

Argos P et al. (1986) The integrase family of site-specific recombinases: regional similarities and global diversity. EMBO Journal 5:433 - 440

Bachellerie J-P, Cavaillé J, Hüttenhofer A (2002) The expanding snoRNA world Biochimie 84:775-790 doi:http://dx.doi.org/10.1016/S0300-9084(02)01402-5

Bains W (1982) The structure of cloned histone genes from *Xenopus borealis* (pp 223 - 238) available online at http://wrap.warwick.ac.uk/67748. University of Warwick

Bains W (2000) Statistical mechanic prediction of non-Gompertzian ageing in extremely aged populations Mechanisms of Aging and development 112:89 - 97

Bains W (2004) Paradoxes of Non-Trivial Gene Networks: How Cancer-Causing Mutations Can Appear to Be Cancer-Protective Rejuvenation Research 7:199 - 210

Baliga NS, Goo YA, Ng WV, Hood L, Daniels CJ, DasSarma S (2000) Is gene expression in Halobacterium NRC-1 regulated by multiple TBP and TFB transcription factors? Molecular Microbiology 36:1184-1185 doi:10.1046/j.1365-2958.2000.01916.x

Bartlett MS, Thomm M, Geiduschek EP (2000) The orientation of DNA in an archaeal transcription initiation complex Nat Struct Mol Biol 7:782-785

Battesti A, Gottesman S (2013) Roles of adaptor proteins in regulation of bacterial proteolysis Current Opinion in Microbiology 16:140-147 doi:http://dx.doi.org/10.1016/j.mib.2013.01.002

Baumann H, Knapp S, Lundback T, Landstein R, Hard T (1994) Solution structure and DNA-binding properties of a thermostable protein from the archaeon Sulfolobus solfataricus Nature Structural Biology 1:808 - 819

Belew AT et al. (2014) Ribosomal frameshifting in the CCR5 mRNA is regulated by miRNAs and the NMD pathway Nature 512:265-269 doi:10.1038/nature13429

http://www.nature.com/nature/journal/v512/n7514/abs/nature13429.html#supplementary-information

Bell SD, Jackson SP (2001) Mechanism and regulation of transcription in archaea Current Opinion in Microbiology 4:208-213 doi:http://dx.doi.org/10.1016/S1369-5274(00)00190-9

Bentley SD, Parkhill J (2004) COMPARATIVE GENOMIC STRUCTURE OF PROKARYOTES Annual Review of Genetics 38:771-791 doi:doi:10.1146/annurev.genet.38.072902.094318

Bi S, Wang Y, Guan J, Sheng X, Meng J (2014) Three new Jurassic euharamiyidan species reinforce early divergence of mammals Nature advance online publication doi:10.1038/nature13718

http://www.nature.com/nature/journal/vaop/ncurrent/abs/nature13718.html#supplementary-information

Bini E, Blum P (2001) Archael catabolite repression: a gene regulatory paradigm. In: Laskin AI, Bennett JW, Gadd GM (eds) Advances in Applied Microbiology, vol 50. Academic Press, San Diego, pp 339 - 362

Blackburn DG, Flemming AF (2012) Invasive implantation and intimate placental associations in a placentotrophic african lizard, Trachylepis ivensi (scincidae) Journal of Morphology 273:137-159 doi:10.1002/jmor.11011
Blackstone NW (2013) Why did eukaryotes evolve only once? Genetic and energetic aspects of conflict and conflict mediation Phil Trans Roy Soc B 368:http://dx.doi.org/10.1098/rstb.2012.0266
Blankenship RE, Hartman H (1998) The origin and evolution of oxygenic photosynthesis Trends in Biochemical Sciences 23:94 - 97
Bouligand Y, Norris V (2001) Chromosome separation and segregation in dinoflagellates andbacteria may depend on liquid crystalline states Biochimie 83:187-192 doi:http://dx.doi.org/10.1016/S0300-9084(00)01211-6
Bracken AP, Dietrich N, Pasini D, Hansen KH, Helin K (2006) Genome-wide mapping of Polycomb target genes unravels their roles in cell fate transitions Genes & Development 20:1123-1136 doi:10.1101/gad.381706
Braglia P, Percudani R, Dieci G (2005) Sequence Context Effects on Oligo(dT) Termination Signal Recognition by Saccharomyces cerevisiae RNA Polymerase III Journal of Biological Chemistry 280:19551-19562 doi:10.1074/jbc.M412238200
Braun RE (2001) Packaging paternal chromosomes with protamine Nature Genetics 28:10 - 12
Breaker RR (2012) Riboswitches and the RNA World Cold Spring Harb Perspect Biol 4:a003566
Brierley I, Digard P, Inglis SC (1989) Characterization of an efficient coronavirus ribosomal frameshifting signal: Requirement for an RNA pseudoknot Cell 57:537-547 doi:http://dx.doi.org/10.1016/0092-8674(89)90124-4
Brown JR, Doolittle WF (1997) Archaea and the prokaryote-to-eukaryote transition Microbiology and Molecular Biology Reviews 61:456-502
Buhler M, Moazed D (2007) Transcription and RNAi in heterochromatic gene silencing Nat Struct Mol Biol 14:1041-1048
Camblong J, Beyrouthy N, Guffanti E, Schlaepfer G, Steinmetz LM, Stutz F (2009) Trans-acting antisense RNAs mediate transcriptional gene cosuppression in S. cerevisiae Genes & Development 23:1534-1545 doi:10.1101/gad.522509
Camblong J, Iglesias N, Fickentscher C, Dieppois G, Stutz F (2007) Antisense RNA Stabilization Induces Transcriptional Gene Silencing via Histone Deacetylation in S. cerevisiae Cell 131:706-717 doi:http://dx.doi.org/10.1016/j.cell.2007.09.014
Cao X, Aufsatz W, Zilberman D, Mette MF, Huang MS, Matzke M, Jacobsen SE (2003) Role of the DRM and CMT3 Methyltransferases in RNA-Directed DNA Methylation Current Biology 13:2212-2217 doi:http://dx.doi.org/10.1016/j.cub.2003.11.052
Carlile TM, Rojas-Duran MF, Zinshteyn B, Shin H, Bartoli KM, Gilbert WV (2014) Pseudouridine profiling reveals regulated mRNA pseudouridylation in yeast and human cells Nature advance online publication doi:10.1038/nature13802

http://www.nature.com/nature/journal/vnfv/ncurrent/abs/nature13802.html#supplementary-information
Catling DC, Glein CR, Zahnle KJ, McCay CP (2005) Why O2 Is Required by Complex Life on Habitable Planets and the Concept of Planetary "Oxygenation Time" Astrobiology 5:415 - 438
Cavalier-Smith T (2010) Deep phylogeny, ancestral groups and the four ages of life Philosophical Transactions of the Royal Society B: Biological Sciences 365:111-132 doi:10.1098/rstb.2009.0161
Cech Thomas R, Steitz Joan A (2014) The Noncoding RNA Revolution—Trashing Old Rules to Forge New Ones Cell 157:77-94 doi:http://dx.doi.org/10.1016/j.cell.2014.03.008
Chalker DL, Yao M-C (2011) DNA Elimination in Ciliates: Transposon Domestication and Genome Surveillance Annual Review of Genetics 45:227-246 doi:10.1146/annurev-genet-110410-132432
Champoux JJ (2001) DNA TOPOISOMERASES: Structure, Function, and Mechanism Annual Review of Biochemistry 70:369-413 doi:10.1146/annurev.biochem.70.1.369
Chan SW-L, Ziberman D, Xie Z, Johansen LK, Carrington JC, Jacobsen SE (2004) RNA Silencing Genes Control de Novo DNA Methylation Science 303:1336
Chandler M, Fayet O (1993) Translational frameshifting in the control of transposition in bacteria Molecular Microbiology 7:497-503 doi:10.1111/j.1365-2958.1993.tb01140.x

Chang C-H, Luse DS (1997) The H3/H4 Tetramer Blocks Transcript Elongation by RNA Polymerase II in Vitro Journal of Biological Chemistry 272:23427-23434 doi:10.1074/jbc.272.37.23427
Chang Y-j et al. (2011) Non-contiguous finished genome sequence and contextual data of the filamentous soil bacterium Ktedonobacter racemifer type strain (SOSP1-21(T)) Standards in Genomic Sciences 5:97-111 doi:10.4056/sigs.2114901
Chu S, DeRisi J, Eisen M, Mulholland J, Botstein D, Brown PO, Herskowitz I (1998) The Transcriptional Program of Sporulation in Budding Yeast Science 282:699-705 doi:10.1126/science.282.5389.699
Cliffe LJ, Kieft R, Southern T, Birkeland SR, Marshall M, Sweeney K, Sabatini R (2009) JBP1 and JBP2 are two distinct thymidine hydroxylases involved in J biosynthesis in genomic DNA of African trypanosomes Nucleic Acids Research 37:1452-1462 doi:10.1093/nar/gkn1067
Cobucci-Ponzano B, Rossi M, Moracci M (2005) Recoding in Archaea Molecular Microbiology 55:339-348 doi:10.1111/j.1365-2958.2004.04400.x
Collin SP, Knight MA, Davies WL, Potter IC, Hunt DM, Trezise AEO (2003) Ancient colour vision: multiple opsin genes in the ancestral vertebrates Current Biology 13:R864-R865 doi:http://dx.doi.org/10.1016/j.cub.2003.10.044
Contursi P, Cannio R, Prato S, Fiorentino G, Rossi M, Bartolucci S (2003) Development of a genetic system for hyperthermophilic Archaea: expression of a moderate thermophilic bacterial alcohol dehydrogenase gene in *Sulfolobus solfataricus* FEMS Microbiology Letters 218 115 - 120
Coppins RL, Hall KB, Groisman EA (2007) The intricate world of riboswitches Current Opinion in Microbiology 10:176-181 doi:http://dx.doi.org/10.1016/j.mib.2007.03.006
Corbett KD, Berger JM (2004) STRUCTURE, MOLECULAR MECHANISMS, AND EVOLUTIONARY RELATIONSHIPS IN DNA TOPOISOMERASES Annual Review of Biophysics and Biomolecular Structure 33:95-118 doi:10.1146/annurev.biophys.33.110502.140357
Core LJ, Waterfall JJ, Lis JT (2008) Nascent RNA Sequencing Reveals Widespread Pausing and Divergent Initiation at Human Promoters Science 322:1845-1848 doi:10.1126/science.1162228
Dalle-Donne I, Rossi R, Colombo G, Giustarini D, Milzani A (2008) Protein S-glutathionylation: a regulatory device from bacteria to humans Trends in Biochemical Sciences 34:85 - 96
de Duve C (2005) Singularities. Landmarks on the pathways of life. Cambridge University Press, Cambridge
de Duve C (2007) The origin of eukaryotes: a reappraisal Nature Reviews Genetics 8:395 - 403
de Santa F et al. (2009) A Large Fraction of Extragenic RNA Pol II Transcription Sites Overlap Enhancers PLoS Biology 8: e1000384. doi:1000310.1001371/journal.pbio.1000384
de Vries R (2010) DNA condensation in bacteria: Interplay between macromolecular crowding and nucleoid proteins Biochimie 92:1715-1721 doi:http://dx.doi.org/10.1016/j.biochi.2010.06.024
Dinman JD (2012) Mechanisms and implications of programmed translational frameshifting Wiley Interdisciplinary Reviews: RNA 3:661-673 doi:10.1002/wrna.1126
Domené S, Bumaschny VF, de Souza FSJ, Franchini LF, Nasif S, Low MJ, Rubinstein M (2013) Enhancer turnover and conserved regulatory function in vertebrate evolution Philosophical Transactions of the Royal Society B: Biological Sciences 368 doi:10.1098/rstb.2013.0027
Doolittle WF, Feng D-F, Tsang S, Cho G, Little E (1996) Determining divergence times of the major kingdoms of living organisms with a protein clock Science 271:470 - 477
Drlica K, Rouviere-Yaniv J (1987) Histonelike proteins of bacteria. Microbiological Reviews 51:301 - 319
Dumesic Phillip A et al. (2013) Stalled Spliceosomes Are a Signal for RNAi-Mediated Genome Defense Cell 152:957-968 doi:http://dx.doi.org/10.1016/j.cell.2013.01.046
Edgell D, Chalamcharla V, Belfort M (2011) Learning to live together: mutualism between self-splicing introns and their hosts BMC Biology 9:22
Eichler J, Adams MWW (2005) Posttranslational Protein Modification in Archaea Microbiology and Molecular Biology Reviews 69:393-425 doi:10.1128/mmbr.69.3.393-425.2005
EMBL (2015a) Bacterial expression vectors. http://www.embl.de/pepcore/pepcore_services/strains_vectors/vectors/bacterial_expression_vectors/.

EMBL (2015b) Insect cell expression vectors. http://www.embl.de/pepcore/pepcore_services/strains_vectors/vectors/insectcell_expression_vectors/index.html.
Fang FC, Rimsky S (2008) New insights into transcriptional regulation by H-NS Current Opinion in Microbiology 11:113-120 doi:http://dx.doi.org/10.1016/j.mib.2008.02.011
Fernando Bazan J, Koch-Nolte F (1997) Sequence and Structural Links between Distant ADP-Ribosyltransferase Families. In: Haag F, Koch-Nolte F (eds) ADP-Ribosylation in Animal Tissues, vol 419. Advances in Experimental Medicine and Biology. Springer US, pp 99-107. doi:10.1007/978-1-4419-8632-0_12
FitzGerald PC, Sturgill D, Shyakhtenko A, Oliver B, Vinson. C (2006) Comparative genomics of Drosophila and human core promoters Genome Biology 7:R53
Frost B, Hemberg M, Lewis J, Feany MB (2014) Tau promotes neurodegeneration through global chromatin relaxation Nat Neurosci 17:357-366 doi:10.1038/nn.3639

http://www.nature.com/neuro/journal/v17/n3/abs/nn.3639.html#supplementary-information
Fuerst JA (2005) Intracellular compartmentalization inPlantomycetes Ann Rev Microbiology 59:299 - 328
Fuerst JA, Webb RI, Garson MJ, Hardy L, Reiswig HM (1998) Membrane-bounded nucleoids in microbial symbionts of marine sponges FEMS Microbiology Letters 166:29-34 doi:10.1111/j.1574-6968.1998.tb13179.x
Gao Z, Lee P, Stafford JM, von Schimmelmann M, Schaefer A, Reinberg D (2014) An AUTS2-Polycomb complex activates gene expression in the CNS Nature 516:349-354 doi:10.1038/nature13921

http://www.nature.com/nature/journal/v516/n7531/abs/nature13921.html#supplementary-information
Gaspin C, Cavaillé J, Erauso G, Bachellerie J-P (2000) Archaeal homologs of eukaryotic methylation guide small nucleolar RNAs: lessons from the Pyrococcus genomes Journal of Molecular Biology 297:895-906 doi:http://dx.doi.org/10.1006/jmbi.2000.3593
Geiduschek EP, Ouhammouch M (2005) Archaeal transcription and its regulators Molecular Microbiology 56:1397-1407 doi:10.1111/j.1365-2958.2005.04627.x
Geisler S, Coller J (2013) RNA in unexpected places: long non-coding RNA functions in diverse cellular contexts Nat Rev Mol Cell Biol 14:699-712 doi:10.1038/nrm3679
Geiss-Friedlander R, Melchior F (2007) Concepts in sumoylation: a decade on Nat Rev Mol Cell Biol 8:947-956 doi:http://www.nature.com/nrm/journal/v8/n12/suppinfo/nrm2293_S1.html
Geng F, Wenzel S, Tansey WP, Tansey WP (2012) Ubiquitin and Proteasomes in Transcription Ann Rev Biochem 81:177 - 201
Ghaemmaghami S et al. (2003) Global analysis of protein expression in yeast Nature 425:737-741 doi:http://www.nature.com/nature/journal/v425/n6959/suppinfo/nature02046_S1.html
Gilbert W (1986) The RNA World Nature 319:618
Glansdorff N (2002) About the last common ancestor, the universal life-tree and lateral gene transfer: a reappraisal. Molecular Microbiology 38:177 - 185
Görke B, Vogel J (2008) Noncoding RNA control of the making and breaking of sugars Genes & Development 22:2914-2925 doi:10.1101/gad.1717808
Gould SJ (1989) Wonderful Life. The Burgess Shales and the nature of history. W. W. Norton and Co,
Gourdeau H, Fournier REK (1990) Genetic Analysis of Mammalian Cell Differentiation Annual Review of Cell Biology 6:69-94 doi:doi:10.1146/annurev.cb.06.110190.000441
Grewal SIS, Rice JC (2004) Regulation of heterochromatin by histone methylation and small RNAs Current Opinion in Cell Biology 16:230-238 doi:http://dx.doi.org/10.1016/j.ceb.2004.04.002
Griese M, Lange C, Soppa J (2011) Ploidy in cyanobacteria FEMS Microbiology Letters 323:124-131
Grigg JC, Ke A (2013) Structural determinants for geometry and information decoding of tRNA by T box leader RNA. Structure 21:2025 - 2032
Guttman M et al. (2010) Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs Nat Biotech 28:503-510 doi:http://www.nature.com/nbt/journal/v28/n5/abs/nbt.1633.html#supplementary-information
Haasa PO, Hottinger MO (2008) The diverse biological roles of mammalian PARPs, a small but powerful family of poly-ADP-ribose polymerases Frontiers in Bioscience 13:3036 - 3082

Hale CR et al. (2009) RNA-Guided RNA Cleavage by a CRISPR RNA-Cas Protein Complex Cell 139:945-956 doi:http://dx.doi.org/10.1016/j.cell.2009.07.040
Hallet B, Sherratt DJ (1997) Transposition and site-specific recombination: adapting DNA cut-and-paste mechanisms to a variety of genetic rearrangements FEMS Microbiology Reviews 21:157-178 doi:10.1111/j.1574-6976.1997.tb00349.x
Hanahan D, Weinberg Robert A (2011) Hallmarks of Cancer: The Next Generation Cell 144:646-674 doi:http://dx.doi.org/10.1016/j.cell.2011.02.013
Harrison P, Kumar A, Lan N, Echols N, Snyder M, Gerstein M (2002) A small reservoir of disabled ORFs in the yeast genome and its implications for the dynamics of proteome evolution Journal of Molecular Biology 316:409-419 doi:http://dx.doi.org/10.1006/jmbi.2001.5343
Helmann JD, Chamberlin MJ (1988) Structure and Function of Bacterial Sigma Factors Annual Review of Biochemistry 57:839-872 doi:10.1146/annurev.bi.57.070188.004203
Hendrix DA, Hong J-W, Zeitlinger J, Rokhsar DS, Levine MS (2008) Promoter elements associated with RNA Pol II stalling in the Drosophila embryo Proceedings of the National Academy of Sciences 105:7762-7767 doi:10.1073/pnas.0802406105
Holland HD (2006) The oxygenation of the atmosphere and oceans Philosophical Transactions of the Royal Society B: Biological Sciences 361:903-915 doi:10.1098/rstb.2006.1838
Horvath P, Barrangou R (2010) CRISPR/Cas, the Immune System of Bacteria and Archaea Science 327:167-170 doi:10.1126/science.1179555
Houseley J (2012) Form and function of eukaryotic unstable non-coding RNAs Bioch Soc Trans 40:836 - 840
Houseley J, Rubbi L, Grunstein M, Tollervey D, Vogelauer M (2008) A ncRNA Modulates Histone Modification and mRNA Induction in the Yeast GAL Gene Cluster Molecular Cell 32:685-695 doi:http://dx.doi.org/10.1016/j.molcel.2008.09.027
Houseley J, Tollervey D (2008) The nuclear RNA surveillance machinery: The link between ncRNAs and genome structure in budding yeast? Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms 1779:239-246 doi:http://dx.doi.org/10.1016/j.bbagrm.2007.12.008
Huarte M et al. (2010) A Large Intergenic Noncoding RNA Induced by p53 Mediates Global Gene Repression in the p53 Response Cell 142:409-419 doi:http://dx.doi.org/10.1016/j.cell.2010.06.040
Jacob F, Monod J (1961) On the Regulation of Gene Activity Cold Spring Harbor Symposia on Quantitative Biology 26:193-211 doi:10.1101/sqb.1961.026.01.024
Jeong SW, Lang WH, Reeder RH (1995) The release element of the yeast polymerase I transcription terminator can function independently of Reb1p Molecular and Cellular Biology 15:5929-5936
Jin F et al. (2013) A high-resolution map of the three-dimensional chromatin interactome in human cells Nature 503:290-294 doi:10.1038/nature12644

http://www.nature.com/nature/journal/v503/n7475/abs/nature12644.html#supplementary-information
Jobe EM, McQuate AL, Zhao X (2012) Crosstalk among epigenetic pathways regulates neurogenesis Frontiers in Neuroscience 6 doi:10.3389/fnins.2012.00059
Jore MM, Brouns SJJ, van der Oost J (2012) RNA in Defense: CRISPRs Protect Prokaryotes against Mobile Genetic Elements Cold Spring Harb Perspect Biol 4:a003657
Joshua-Tor L, Hannon GJ (2010) Ancestral Roles of Small RNAs: An Ago-Centric Perspective Cold Spring Harb Perspect Biol 2011:doi: 10.1101/cshperspect.a003772
Khalil AM et al. (2009) Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression Proceedings of the National Academy of Sciences 106:11667-11672 doi:10.1073/pnas.0904715106
Kino T, Hurt DE, Ichijo T, Nader N, Chrousos GP (2010) Noncoding RNA Gas5 Is a Growth Arrest- and Starvation-Associated Repressor of the Glucocorticoid Receptor Sci Signal 3:ra8- doi:10.1126/scisignal.2000568
Kireeva ML, Hancock B, Cremona GH, Walter W, Studitsky VM, Kashlev M (2005) Nature of the Nucleosomal Barrier to RNA Polymerase II Molecular Cell 18:97-108 doi:http://dx.doi.org/10.1016/j.molcel.2005.02.027

Kireeva ML, Walter W, Tchernajenko V, Bondarenko V, Kashlev M, Studitsky VM (2002) Nucleosome Remodeling Induced by RNA Polymerase II: Loss of the H2A/H2B Dimer during Transcription Molecular Cell 9:541-552 doi:http://dx.doi.org/10.1016/S1097-2765(02)00472-0
Kiucho T et al. (2014) A single female-specific piRNa is the primary determiner of sex in the silkworm Nature 509:633 - 636
Komaki K, Ishikawa H (2000) Genomic copy number of intracellular bacterial symbionts of aphids varies in response to developmental stage and morph of their host Insect Biochemistry and Molecular Biology 30:253-258 doi:http://dx.doi.org/10.1016/S0965-1748(99)00125-3
Komik Z (2005) Pax genes in eye development and evolution. Current Opinion in Genetics and Development 15:430 - 438
Kumar S, Cheng X, Klimasauskas S, Mi S, Posfai J, Roberts RJ, Wilson GG (1994) The DNA (cytosine-5) methyltransferases. Nucleic Acids Research 22:1 - 10
Kuo C-H, Ochman H (2010) The Extinction Dynamics of Bacterial Pseudogenes PLoS Genetics 6:e1001050
Lafontaine DLJ, Tollervey D (1998) Birth of the snoRNPs: the evolution of the modification-guide snoRNAs Trends in Biochemical Sciences 23:383-388 doi:http://dx.doi.org/10.1016/S0968-0004(98)01260-2
Lambowitz AM, Zimmerly S (2004) Mobile Group II introns Ann Rev Genetics 38:1 - 35
Land MF, Nilsson D-E (2012) Animal Eyes, 2nd edition. Oxford University Press, Oxford, UK
Landschulz W, Johnson P, McKnight S (1988) The leucine zipper: a hypothetical structure common to a new class of DNA binding proteins Science 240:1759-1764 doi:10.1126/science.3289117
Lang BF et al. (2014) Massive programmed translational jumping in mitochondria Proceedings of the National Academy of Sciences 111:5926-5931 doi:10.1073/pnas.1322190111
Leblond CS et al. (2012) Genetic and Functional Analyses of SHANK2 Mutations Suggest a Multiple Hit Model of Autism Spectrum Disorders PLoS Genetics 8:e1002521
Lee TI et al. (2006) Control of Developmental Regulators by Polycomb in Human Embryonic Stem Cells Cell 125:301-313 doi:http://dx.doi.org/10.1016/j.cell.2006.02.043
Lettice LA et al. (2003) A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly Human Molecular Genetics 12:1725-1735 doi:10.1093/hmg/ddg180
Levis RW, Ganesan R, Houtchens K, Tolar LA, Sheen F-m (1993) Transposons in place of telomeric repeats at a Drosophila telomere Cell 75:1083-1093 doi:http://dx.doi.org/10.1016/0092-8674(93)90318-K
Lewis S, Gellert M (1989) The mechanism of antigen receptor gene assembly Cell 59:585-588 doi:http://dx.doi.org/10.1016/0092-8674(89)90002-0
Li B, Carey M, Workman JL (2007) The Role of Chromatin during Transcription Cell 128:707-719 doi:http://dx.doi.org/10.1016/j.cell.2007.01.015
Li J, Gilmour DS (2013) Distinct mechanisms of transcriptional pausing orchestrated by GAGA factor and M1BP, a novel transcription factor EMBO Journal 32:1829 - 1841
Li Y, Tergaonkar V (2014) Noncanonical Functions of Telomerase: Implications in Telomerase-Targeted Cancer Therapies Cancer Research 74:1639-1644 doi:10.1158/0008-5472.can-13-3568
Liu X, Bushnell DA, Kornberg RD (2013) RNA polymerase II transcription: Structure and mechanism Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms 1829:2-8 doi:http://dx.doi.org/10.1016/j.bbagrm.2012.09.003
Liu Y, Harrison PM, Kunin V, Gerstein M (2004) Comprehensive analysis of pseudogenes in prokaryotes: widespread gene decay and failure of putative horizontally transferred genes Genome Biology 5:R64: doi:10.1186/gb-2004-1185-1189-r1164
Livny J, Brencic A, Lory S, Waldor MK (2006) Identification of 17 Pseudomonas aeruginosa sRNAs and prediction of sRNA-encoding genes in 10 diverse pathogens using the bioinformatic tool sRNAPredict2 Nucleic Acids Research 34:3484-3493 doi:10.1093/nar/gkl453
Loeb LA, Loeb KR, Anderson JP (2003) Multiple mutations and cancer Proceedings of the National Academy of Sciences 100:776-781 doi:10.1073/pnas.0334858100

Lucas S et al. (2002) Construction of a Shuttle Vector for, and Spheroplast Transformation of, the Hyperthermophilic Archaeon Pyrococcus abyssi Applied and Environmental Microbiology 68:5528 - 5535
Luijsterburg MS, White MF, Van Driel R, Dame RT (2008) The Major Architects of Chromatin: Architectural Proteins in Bacteria, Archaea and Eukaryotes Critical Reviews in Biochemistry and Molecular Biology 43:1 - 26
Makarova KS, Wolf YI, Mekhedov SL, Mirkin BG, Koonin EV (2005) Ancestral paralogs and pseudoparalogs and their role in the emergence of the eukaryotic cell Nucleic Acids Research 33:4626-4638 doi:10.1093/nar/gki775
Makarova KS, Wolf YI, van der Oost J, Koonin EV (2009) Prokaryotic homologs of Argonaute proteins are predicted to function as key components of a novel system of defense against mobile genetic elements Biol Direct 4:29 doi: 10.1186/1745-6150-1184-1129
Malone CD, Hannon GJ (2009) Small RNAs as Guardians of the Genome Cell 136:656 - 668
Mantovani R (1999) The molecular biology of the CCAAT-binding factor NF-Y Gene 239:15-27 doi:http://dx.doi.org/10.1016/S0378-1119(99)00368-6
Marsh VL, Peak-Chew SY, Bell SD (2005) Sir2 and the Acetyltransferase, Pat, Regulate the Archaeal Chromatin Protein, Alba Journal of Biological Chemistry 280:21122 - 21128
Martianov I, Ramadass A, Serra Barros A, Chow N, Akoulitchev A (2007) Repression of the human dihydrofolate reductase gene by a non-coding interfering transcript Nature 445:666-670 doi:http://www.nature.com/nature/journal/v445/n7128/suppinfo/nature05519_S1.html
Martin C, Zhang Y (2005) The diverse functions of histone lysine methylation Nat Rev Mol Cell Biol 6:838-849
Martin W, Koonin EV (2006) Introns and the origin of nucleus–cytosol compartmentalization Nature 440:41 - 45
Mattick JS, Gagen MJ (2001) The Evolution of Controlled Multitasked Gene Networks: The Role of Introns and Other Noncoding RNAs in the development of Complex Organisms Mol Biol Evol 18:1611 - 1630
McCarty DM, Young SM, Samulski RJ (2004) INTEGRATION OF ADENO-ASSOCIATED VIRUS (AAV) AND RECOMBINANT AAV VECTORS Annual Review of Genetics 38:819-845 doi:doi:10.1146/annurev.genet.37.110801.143717
Meister G, Tuschi T (2004) Mechanisms of gene silencing by double-stranded RNA Nature 431:343 - 349
Mello CC, Conte DJ (2004) Revealing the world of RNA interference Nature 431:338 - 341
Mendell JE, Clements KD, Choat JH, Angert ER (2008) Extreme polyploidy in a large bacterium Proceedings of the National Academy of Sciences 105:6730-6734 doi:10.1073/pnas.0707522105
Meng Q, Wang Y, Liu X-Q (2005) An Intron-encoded Protein Assists RNA Splicing of Multiple Similar Introns of Different Bacterial Genes Journal of Biological Chemistry 280:35085-35088 doi:10.1074/jbc.C500328200
Menssen A, Haupl T, Sittinger M, Delorme B, Charbord P, Ringe J (2011) Differential gene expression profiling of human bone marrow-derived mesenchymal stem cells during adipogenic development BMC Genomics 12:461
Mercer TR, Mattick JS (2013) Structure and function of long noncoding RNAs in epigenetic regulation Nat Struct Mol Biol 20:300-307
Merck Millipore Inc (2015) Expression vectors. http://www.emdmillipore.com/.
Messer W (2002) The bacterial replication initiator DnaA. DnaA and oriC, the bacterial mode to initiate DNA replication FEMS Microbiology Reviews 26:355-374 doi:10.1111/j.1574-6976.2002.tb00620.x
Mighell AJ, Smith NR, Robinson PA, Markham AF (2000) Vertebrate pseudogenes FEBS letters 468:109-114 doi:http://dx.doi.org/10.1016/S0014-5793(00)01199-6
Mikkelsen TS et al. (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells Nature 448:553-560 doi:http://www.nature.com/nature/journal/v448/n7153/suppinfo/nature06008_S1.html
Miller AD (1992) Retroviral Vectors. In: Muzyczka N (ed) Viral Expression Vectors, vol 158. Current Topics in Microbiology and Immunology. Springer Berlin Heidelberg, pp 1-24. doi:10.1007/978-3-642-75608-5_1

Miller G, Hahn S (2006) A DNA-tethered cleavage probe reveals the path for promoter DNA in the yeast preinitiation complex Nat Struct Mol Biol 13:603-610 doi:http://www.nature.com/nsmb/journal/v13/n7/suppinfo/nsmb1117_S1.html
Mischo HE, Proudfoot NJ (2013) Disengaging polymerase: Terminating RNA polymerase II transcription in budding yeast Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms 1829:174-185 doi:http://dx.doi.org/10.1016/j.bbagrm.2012.10.003
Mizuguchi G, Shen X, Landry J, Wu W-H, Sen S, Wu C (2004) ATP-Driven Exchange of Histone H2AZ Variant Catalyzed by SWR1 Chromatin Remodeling Complex Science 303:343-348 doi:10.1126/science.1090701
Moreno Díaz de la Espina S, Alverca E, Cuadrado A, Franca S (2005) Organization of the genome and gene expression in a nuclear environment lacking histones and nucleosomes: the amazing dinoflagellates European Journal of Cell Biology 84:137-149 doi:http://dx.doi.org/10.1016/j.ejcb.2005.01.002
Moroz LL et al. (2014) The ctenophore genome and the evolutionary origins of neural systems Nature advance online publication doi:10.1038/nature13400

http://www.nature.com/nature/journal/vaop/ncurrent/abs/nature13400.html#supplementary-information
Mossman HW (1937) Comparative morphogenesis of the fetal membranes and accessory uterine structures Carnegie Inst Contr Embryol 26:129 - 246
Müller J, Kassis JA (2006) Polycomb response elements and targeting of Polycomb group proteins in Drosophila Current Opinion in Genetics & Development 16:476-484 doi:http://dx.doi.org/10.1016/j.gde.2006.08.005
Mumberg D, Müller R, Funk M (1995) Yeast vectors for the controlled expression of heterologous proteins in different genetic backgrounds Gene 156:119-122 doi:http://dx.doi.org/10.1016/0378-1119(95)00037-7
Muse GW et al. (2007) RNA polymerase is poised for activation across the genome Nat Genet 39:1507-1511 doi:http://www.nature.com/ng/journal/v39/n12/suppinfo/ng.2007.21_S1.html
Nagano T, Fraser P (2011) No-Nonsense Functions for Long Noncoding RNAs Cell 145:178 - 181
Narlikar L, Sakabe NJ, Blanski AA, Arimura FE, Westlund JM, Nobrega MA, Ovcharenko I (2010) Genome-wide discovery of human heart enhancers Genome Research 20:381 - 392
Navarre WW, McClelland M, Libby SJ, Fang FC (2007) Silencing of xenogeneic DNA by H-NS—facilitation of lateral gene transfer in bacteria by a defense system that recognizes foreign DNA Genes & Development 21:1456-1471 doi:10.1101/gad.1543107
Navarre WW, Porwollik S, Wang Y, McClelland M, Rosen H, Libby SJ, Fang FC (2006) Selective Silencing of Foreign DNA with Low GC Content by the H-NS Protein in Salmonella Science 313:236-238 doi:10.1126/science.1128794
Nechaev S, Adelman K (2011) Pol II waiting in the starting gates: Regulating the transition from transcription initiation into productive elongation Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms 1809:34-45 doi:http://dx.doi.org/10.1016/j.bbagrm.2010.11.001
Nechaev S, Fargo DC, dos Santos G, Liu L, Gao Y, Adelman K (2010) Global Analysis of Short RNAs Reveals Widespread Promoter-Proximal Stalling and Arrest of Pol II in Drosophila Science 327:335-338 doi:10.1126/science.1181421
Nykänen A, Haley B, Zamore PD (2001) ATP Requirements and Small Interfering RNA Structure in the RNA Interference Pathway Cell 107:309-321 doi:http://dx.doi.org/10.1016/S0092-8674(01)00547-5
Olave IA, Peck-Peterson SI, Crabtree GR (2002) Nuclear actin and actin-related proteins in chromatin remodelling Annual Review of Biochemistry 71:755 - 781
Ørom UA et al. (2010) Long Noncoding RNAs with Enhancer-like Function in Human Cells Cell 143:46-58 doi:http://dx.doi.org/10.1016/j.cell.2010.09.001
Pagel M, Johnstone RA (1992) Variation across Species in the Size of the Nuclear Genome Supports the Junk-DNA Explanation for the C-Value Paradox Proceedings of the Royal Society of London Series B: Biological Sciences 249:119-124 doi:10.1098/rspb.1992.0093

Pallen MJ, Lam AC, Loman NJ, McBride A (2001) An abundance of bacterial ADP-ribosyltransferases — implications for the origin of exotoxins and their human homologues Trends in Microbiology 9:302-307 doi:http://dx.doi.org/10.1016/S0966-842X(01)02074-1
Papenfort K, Vogel J (2010) Regulatory RNA in Bacterial Pathogens Cell Host and Microbe 8:116 - 127
Peirce V, Carobbio S, Vidal-Puig A (2014) The different shades of fat Nature 510:76 - 83
Peng N et al. (2012) A Synthetic Arabinose-Inducible Promoter Confers High Levels of Recombinant Protein Expression in hyperthermophilic Archaeon *Sulfolobus islandicus* Applied and Environmental Microbiology 79:5630 - 5637
Pickart CM (2001) Mechanisms underlying ubiquitination Ann Rev Biochem 70:503 - 533
Poliseno L, Salmena L, Zhang J, Carver B, Haveman WJ, Pandolfi PP (2010) A coding-independent function of gene and pseudogene mRNAs regulates tumour biology Nature 465:1033-1038 doi:http://www.nature.com/nature/journal/v465/n7301/suppinfo/nature09144_S1.html
Pollux BJA, Pires MN, Banet AI, Reznick DN (2009) Evolution of Placentas in the Fish Family Poeciliidae: An Empirical Study of Macroevolution Annual Review of Ecology, Evolution, and Systematics 40:271-289 doi:doi:10.1146/annurev.ecolsys.110308.120209
Polyzos SA, Kountouras J, Zavos C, Deretzi G (2012) Nonalcoholic Fatty Liver Disease: Multimodal Treatment Options for a Pathogenetically Multiple-hit Disease Journal of Clinical Gastroenterology 46:272-284 210.1097/MCG.1090b1013e31824587e31824580
Ponting CP, Hardison RC (2011) What fraction of the human genome is functional? Genome Res 21:1769 - 1776
Ponting CP, Oliver PL, Reik W (2009) Evolution and Functions of Long Noncoding RNAs Cell 136:629-641 doi:http://dx.doi.org/10.1016/j.cell.2009.02.006
Promega Corp. (2015) Vectors. http://www.promega.co.uk/products/vectors/.
Pyle AM (2012) Group II intron architecture and its implications for the development of eukaryotic splicing systems FASEB Journal 26:217.213
Rabbani MA et al. (2003) Monitoring Expression Profiles of Rice Genes under Cold, Drought, and High-Salinity Stresses and Abscisic Acid Application Using cDNA Microarray and RNA Gel-Blot Analyses Plant Physiology 133:1755-1767 doi:10.1104/pp.103.025742
Rebollo R, Romanish MT, Mager DL (2012) Transposable Elements: An Abundant and Natural Source of Regulatory Sequences for Host Genes Annual Review of Genetics 46:21-42 doi:doi:10.1146/annurev-genet-110711-155621
Reeve JN (2003) Archaeal chromatin and transcription Molecular Microbiology 48:587-598 doi:10.1046/j.1365-2958.2003.03439.x
Reisenauer A, Kahng LS, McCollum S, Shapiro L (1999) Bacterial DNA Methylation: a Cell Cycle Regulator? Journal of Bacteriology 181:5135-5139
Rhee HS, Pugh F (2012) Genome-wide structure and organization of eukaryotic pre-initiation complexes Nature 483:295 - 301
Rinn JLC, Howard Y. (2012) Genome regulation by long noncoding RNAs Ann Rev Biochem 81:145 - 166
Rivas E, Klein RJ, Jones TA, Eddy SR (2001) Computational identification of noncoding RNAs in E. coli by comparative genomics Current Biology 11:1369-1373 doi:http://dx.doi.org/10.1016/S0960-9822(01)00401-8
Rivier C, Goldschmidt-Clermont M, Rochaix J-D (2001) Identification of an RNA–protein complex involved in chloroplast group II intron trans-splicing in Chlamydomonas reinhardtii The EMBO Journal 20:1765-1773 doi:10.1093/emboj/20.7.1765
Roeben A, Kofler C, Nagy I, Nickell S, Ulrich Hartl F, Bracher A (2006) Crystal Structure of an Archaeal Actin Homolog Journal of Molecular Biology 358:145-156 doi:http://dx.doi.org/10.1016/j.jmb.2006.01.096
Ross RJ, Weiner MM, Lin H (2014) PIWI proteins and PIWI-interacting RNAs in the soma Nature 505:353-359 doi:10.1038/nature12987
Roy SW, Gilbert W (2006) The evolution of spliceosomal introns: patterns, puzzles and progress Nature Reviews Genetics 7:211 - 221
Ruthenburg A, KLi H, Patel DJ, Allis CD (2007) Multivalent engagement of chromatin modifications by linked binding modules Nature Reviews: Molecular and Cell Biology 8:983 - 994

Salas M (1991) Protein-Priming of DNA Replication Ann Rev Biochem 60:39 - 71
Salmena L, Poliseno L, Tay Y, Kats L, Pandolfi Pier P (2011) A ceRNA Hypothesis: The Rosetta Stone of a Hidden RNA Language? Cell 146:353-358 doi:http://dx.doi.org/10.1016/j.cell.2011.07.014
Sandman K, Reeve JN (2005) Archaeal chromatin proteins: different structures but common function? Current Opinion in Microbiology 8:656-661 doi:http://dx.doi.org/10.1016/j.mib.2005.10.007
Sandman K, Reevem JH (2001) Chromosome packaging by archael histones. In: Laskin aI, Bennett JW, Gadd gM (eds) Advances in Applied Microbiology, vol 50. Academic Press, San Diego, pp 73 - 100
Santangelo TJ, Čuboňová Lu, Reeve JN (2008) Shuttle Vector Expression in Thermococcus kodakaraensis: Contributions of cis Elements to Protein Synthesis in a Hyperthermophilic Archaeon Applied and Environmental Microbiology 74:3099-3104 doi:10.1128/aem.00305-08
Sanyal A, Lajoie BR, Jain G, Dekker J (2012) The long-range interaction landscape of gene promoters Nature 489:109-113 doi:http://www.nature.com/nature/journal/v489/n7414/abs/nature11279.html#supplementary-information
Sauer B (1994) Site-specific recombination: developments and applications Current Opinion in Biotechnology 5:521-527 doi:http://dx.doi.org/10.1016/0958-1669(94)90068-X
Schouten M, Buijink MR, Lucassen PJ, Fitzsimons CP (2012) New Neurons in Aging Brains: Molecular Control by Small Non-Coding RNAs Frontiers in Neuroscience 6:doi: 10.3389/fnins.2012.00025
Schreier H, Robinson-Bidle KA, Romashko AM, Patel G (1999) Heterologous expression in the Archaea: transcription from *Pyrococcus furiousus gdh* and *mlrA* promoters in *Haloferax volcanii* Extremophiles 3:11 - 19
Schulze-Makuch D, Irwin LN (2008) Life in the Universe: Expectations and Constraints. 2nd edition. Springer, Berlin, germany
Schwabe JWR, Chapman L, Finch JT, Rhodes D (1993) The crystal structure of the estrogen receptor DNA-binding domain bound to DNA: How receptors discriminate between their response elements Cell 75:567-578 doi:http://dx.doi.org/10.1016/0092-8674(93)90390-C
Schwanhausser B et al. (2011) Global quantification of mammalian gene expression control Nature 473:337-342 doi:http://www.nature.com/nature/journal/v473/n7347/abs/10.1038-nature10098-unlocked.html#supplementary-information
Schwartz DC, Hochstrasser M (2003) A superfamily of protein tags: ubiquitin, SUMO and related modifiers Trends in Biochemical Sciences 28:321-328 doi:http://dx.doi.org/10.1016/S0968-0004(03)00113-0
Schwartz YB, Kahn TG, Nix DA, Li X-Y, Bourgon R, Biggin M, Pirrotta V (2006) Genome-wide analysis of Polycomb targets in Drosophila melanogaster Nat Genet 38:700-705 doi:http://www.nature.com/ng/journal/v38/n6/suppinfo/ng1817_S1.html
Sesto N, Wurtzel O, Archambaud C, Sorek R, Cossart P (2013) The excludon: a new concept in bacterial antisense RNA-mediated gene regulation Nat Rev Micro 11:75-82
Shabalina SA, Koonin EV (2008) Origins and evolution of eukaryotic RNA interference Trends in Ecology & Evolution 23:578-587 doi:http://dx.doi.org/10.1016/j.tree.2008.06.005
Shen X, Mizuguchi G, Hamiche A, Wu C (2000) A chromatin remodelling complex involved in transcription and DNA processing Nature 406:541-544 doi:http://www.nature.com/nature/journal/v406/n6795/suppinfo/406541A0_S1.html
Singh DP et al. (2014) Genome-defence small RNAs exapted for epigenetic mating-type inheritance Nature 509:447-452 doi:10.1038/nature13318

http://www.nature.com/nature/journal/v509/n7501/abs/nature13318.html#supplementary-information
Skourti-Stathaki K, Kamieniarz-Gdula K, Proudfoot NJ (2014) R-loops induce repressive chromatin marks over mammalian gene terminators Nature 516:436-439 doi:10.1038/nature13787

http://www.nature.com/nature/journal/v516/n7531/abs/nature13787.html#supplementary-information
Smemo S et al. (2014) Obesity-associated varients within FTO form long-range functional connections with IRX3 Nature 507:371 - 375
Smith JM, Szathmary E (1995) The major transitions in evolution. W H Freeman, Oxford

Smits WK, Grossman AD (2010) The Transcriptional Regulator Rok Binds A+T-Rich DNA and Is Involved in Repression of a Mobile Genetic Element in Bacillus subtilis PLoS Genetics 6:e1001207
Soares D, Dahlke I, Li W-T, Sandman K, Hethke C, Thomm M, Reeve JN (1998) Archaeal histone stability, DNA binding, and transcription inhibition above 90°C Extremophiles 2:75-81 doi:10.1007/s007920050045
Song J-J, Smith SK, Hannon GJ, Joshua-Tor L (2004) Crystal Structure of Argonaute and Its Implications for RISC Slicer Activity Science 305:1434-1437 doi:10.1126/science.1102514
Soppa J (2001) Basal and regulated transcription in archaea. In: Laskin AI, Bennett JW, Gadd GM (eds) Advances in Applied Microbiology, vol 50. Academic Press, San Diego, pp 171 - 217
Soppa J (2010) Protein Acetylation in Archaea, Bacteria, and Eukaryotes Archaea 2010:doi:10.1155/2010/820681
Soppa J (2014) Polyploidy in Archaea and Bacteria: About Desiccation Resistance, Giant Cell Size, Long-Term Survival, Enforcement by a Eukaryotic Host and Additional Aspects Journal of Molecular Microbiology and Biotechnology 24:409-419
Spang A et al. (2015) Complex archaea that bridge the gap between prokaryotes and eukaryotes Nature 521:173-179 doi:10.1038/nature14447

http://www.nature.com/nature/journal/v521/n7551/abs/nature14447.html#supplementary-information
Stedman KM, Schleper C, Rumpf E, Zilig W (1999) Genetic Requirements for the Function of the Archaeal Virus SSV1 in Sulfolobus solfataricus: Construction and Testing of Viral Shuttle Vectors Genetics 152:1397 - 1405
Swerdlow R (2012) Alzheimer's Disease Pathologic Cascades: Who Comes First, What Drives What Neurotox Res 22:182-194 doi:10.1007/s12640-011-9272-9
Taher L et al. (2011) Genome-wide identification of conserved regulatory function in diverged sequences Genome Research 21:1139 - 1149
Taher L, Narlikar L, Ovcharenko I (2012) CLARE: Cracking the LAnguage of Regulatory Elements Bioinformatics 28:581-583 doi:10.1093/bioinformatics/btr704
Tahiliani M et al. (2009) Conversion of 5-Methylcytosine to 5-Hydroxymethylcytosine in Mammalian DNA by MLL Partner TET1 Science 324:930-935 doi:10.1126/science.1170116
Tay Y, Rinn J, Pandolfi PP (2014) The multilayered complexity of ceRNA crosstalk and competition Nature 505:344-352 doi:10.1038/nature12986
Thebault P, Boutin G, Bhat W, Rufiange A, Martens J, Nourani A (2011) Transcription Regulation by the Noncoding RNA SRG1 Requires Spt2-Dependent Chromatin Deposition in the Wake of RNA Polymerase II Molecular and Cellular Biology 31:1288-1300 doi:10.1128/mcb.01083-10
Ulitsky I, Shkumatava A, Jan Calvin H, Sive H, Bartel David P (2011) Conserved Function of lincRNAs in Vertebrate Embryonic Development despite Rapid Sequence Evolution Cell 147:1537-1550 doi:http://dx.doi.org/10.1016/j.cell.2011.11.055
Vaillasante A, de Pablos B, Mendez-Lago M, Abad JP (2008) Telomere maintenance in Drosophila Cell Cycle 7:2134 - 2138
Vaishnaw AK et al. (2010) Review A status report on RNAi therapeutics Silence 1:doi: 10.1186/1758-1907X-1181-1114.
Valencia-Sanchez MA, Liu J, Hannon GJ, Parker R (2006) Control of translation and mRNA degradation by miRNAs and siRNAs Genes & Development 20:515-524 doi:10.1101/gad.1399806
Vanin EF (1985) Processed pseudogenes: characteristics and evolution Ann Rev Genetics 19:253 - 272
Viollier PH, Thanbichler M, McGrath PT, West L, Meewan M, McAdams HH, Shapiro L (2004) Rapid and sequential movement of individual chromosomal loci to specific subcellular locations during bacterial DNA replication Proceedings of the National Academy of Sciences 101
Vockenhuber M-P et al. (2011) Deep sequencing-based identification of small non-coding RNAs in *Streptomyces coelicolor* RNA Biology 8:468-477
Wang D, Rendon A, Wernisch L (2013) Transcription factor and chromatin features predict genes associated with eQTLs Nucleic Acids Research 41:1450-1463 doi:10.1093/nar/gks1339
Wang KC, Chang HY (2011) Molecular Mechanisms of Long Noncoding RNAs Cell 43:904 - 914
Wang W, Li G-W, Chen C, Xie XS, Zhuang X (2011) Nucleoid-Associated Protein in Live Bacteria Science 333

Ward WS, Coffey DS (1991) DNA packaging and organization in mammalian spermatozoa: comparison with somatic cells Biology of Reproduction 44:569-574 doi:10.1095/biolreprod44.4.569
Wardleworth BN, Russell RJM, Bell SD, Taylor GL, White MF (2002) Structure of Alba: an archaeal chromatin protein modulated by acetylation The EMBO Journal 21:4654-4662 doi:10.1093/emboj/cdf465
Washietl S et al. (2007) Structured RNAs in the ENCODE selected regions of the human genome Genome Research 17:852-864 doi:10.1101/gr.5650707
Wassarman KM (2007) 6S RNA: a small RNA regulator of transcription Current Opinion in Microbiology 10:164-168 doi:http://dx.doi.org/10.1016/j.mib.2007.03.008
Waters LS, Storz G (2009) Regulatory RNAs in Bacteria Cell 136:615-628 doi:http://dx.doi.org/10.1016/j.cell.2009.01.043
Weinberg Z, Wang JX, Bogue J, Yang J, Corbino K, Moy RH, Breaker RR (2010) Comparative genomics reveals 104 candidate structured RNAs from bacteria, archaea, and their metagenomes Genome Biology 11:R31
Weinzierl ROJ (2013) The RNA Polymerase Factory and Archaeal Transcription Chemical Reviews 113:8350 - 8376
Weiss M (1982) Cell Hybridization: A Tool for the Study of Cell Differentiation. In: Caskey CT, Robbins DC (eds) Somatic Cell Genetics, vol 50. NATO Advanced Study Institutes Series. Springer US, pp 169-182. doi:10.1007/978-1-4684-4256-4_10
White MF, Bell SD (2002) Holding it together: chromatin in the Archaea Trends in Genetics 18:621-626 doi:http://dx.doi.org/10.1016/S0168-9525(02)02808-1
Whitehouse I, Rando OJ, Delrow J, Tsukiyama T (2007) Chromatin remodelling at promoters suppresses antisense transcription Nature 450:1031-1035 doi:http://www.nature.com/nature/journal/v450/n7172/suppinfo/nature06391_S1.html
William R, Scott, Gilbert W (2006) The evolution of spliceosomal introns: patterns, puzzles and progress Nat Rev Genet 7:211-221
Williams S et al. (2005) CpG-island fragments from the HNRPA2B1/CBX3 genomic locus reduce silencing and enhance transgene expression from the hCMV promoter/enhancer in mammalian cells BMC Biotechnology 5:doi:10.1186/1472-6750-1185-1117
Williams TA, Foster PG, Cox CJ, Embley TM (2014) An archeal origin of eukaryotes supports only two primary domains of life Nature 504:231 - 236
Wilusz JE, Sunwoo H, Spector DL (2009) Long noncoding RNAs: functional surprises from the RNA world Genes & Development 23:1494 - 1504
Wion D, Casadesus J (2006) N6-methyl-adenine: an epigenetic signal for DNA-protein interactions Nat Rev Micro 4:183-192
Wong JTY, New DC, Wong JCW, Hung VKL (2003) Histone-Like Proteins of the Dinoflagellate Crypthecodinium cohnii Have Homologies to Bacterial DNA-Binding Proteins Eukaryotic Cell 2:646-650 doi:10.1128/ec.2.3.646-650.2003
Wourms JP, Lombardi J (1992) Reflections on the Evolution of Piscine Viviparity American Zoologist 32:276-293 doi:10.1093/icb/32.2.276
Xie Y, Reeve JN (2004) Transcription by an Archaeal RNA Polymerase Is Slowed but Not Blocked by an Archaeal Nucleosome Journal of Bacteriology 186:3492-3498 doi:10.1128/jb.186.11.3492-3498.2004
Yadon AN, Van de Mark D, Basom R, Delrow J, Whitehouse I, Tsukiyama T (2010) Chromatin Remodeling around Nucleosome-Free Regions Leads to Repression of Noncoding RNA Transcription Molecular and Cellular Biology 30:5110-5122 doi:10.1128/mcb.00602-10
Yamashita T et al. (2000) Comprehensive Gene Expression Profile of a Normal Human Liver Biochemical and Biophysical Research Communications 269:110-116 doi:http://dx.doi.org/10.1006/bbrc.2000.2272
Yao H, Brick K, Evrard Y, Xiao T, Camerini-Otero RD, Felsenfeld G (2010) Mediation of CTCF transcriptional insulation by DEAD-box RNA-binding protein p68 and steroid receptor RNA activator SRA Genes & Development 24:2543-2555 doi:10.1101/gad.1967810

Yoon J-H et al. (2012) LincRNA-p21 Suppresses Target mRNA Translation Molecular Cell 47:648-655 doi:http://dx.doi.org/10.1016/j.molcel.2012.06.027
Yun M, Wu J, Workman JL, Li B (2011) Readers of histone modifications Cell Res 21:564-578
Zerulla K, Soppa J (2014) Polyploidy in haloarchaea: advantages for growth and survival Frontiers in Microbiology 5:274 doi:10.3389/fmicb.2014.00274
Zhang A, Rimsky S, Reaban ME, Buc H, Belfort M (1996) Escherichia coli protein analogs StpA and H-NS: regulatory loops, similar and disparate effects on nucleic acid dynamics. EMBO Journal 15:1340 - 1349
Zheng D, Gerstein MB (2007) The ambiguous boundary between genes and pseudogenes: the dead rise up, or do they? Trends in Genetics 23:219-224 doi:http://dx.doi.org/10.1016/j.tig.2007.03.003
Zheng T, Hunag Q, Zhang C, Ni J, She Q, Shen Y (2012) Development of a Simvastatin Selection Marker for a Hyperthermophilic Acidophile, Sulfolobus islandicus Applied and Environmental Microbiology 78:568 - 574
Zimmerman SB, Murphy LD (1996) Macromolecular crowding and the mandatory condensation of DNA in bacteria FEBS letters 390:245 - 248