Evolutionary origins of two-barrel RNA polymerases and site-specific transcription initiation

Thomas Fouqueau*, Fabian Blombach* & Finn Werner

Institute of Structural and Molecular Biology, Division of Biosciences, University College London, London WC1E 6BT, UK.
email: f.blombach@ucl.ac.uk; f.werner@ucl.ac.uk

* These authors contributed equally to this work

Abstract

Evolutionary related multi-subunit RNA polymerases (RNAPs) carry out RNA synthesis in all domains life. While their catalytic cores and fundamental mechanisms of transcription elongation are conserved, the initiation stage of the transcription cycle differs substantially between bacteria and archaea/eukaryotes in terms of the requirements for accessory factors and details of the molecular mechanisms. This review focuses on recent insights into the evolution of the transcription apparatus with regard to (i) the surprisingly pervasive double-Ψ β-barrel active site configuration among different nucleic acid polymerase families, (ii) the origin and phylogenetic distribution of TBP, TFB and TFE transcription factors, and (iii) the functional relation between transcription- and translation initiation mechanisms in terms of TSS selection and RNA structure.

Keywords

Multisubunit RNA polymerases, Evolution, LUCA, Translation initiation

Contents

Introduction

Nucleic acid polymerases carry out key functions in DNA replication, -repair and –recombination, as well as RNA transcription. The latter is the first step in gene expression, and provides both the templates for protein synthesis (mRNA) as well as the structural RNAs forming the essential components of the translation machinery (rRNA and tRNA). The two most important superfamilies of nucleic acids polymerases are the single-subunit 'right-handed' polymerases encompassing the thumb-, finger- and palm motifs, and the 'two-barrel'-type polymerases characterized by an active site formed at the interface between two double-Ψ β-barrels (DPBB) motifs. The single-subunit polymerase superfamily includes almost all replicative DNA polymerases, bacteriophage single-subunit RNAPs including the mitochondrial RNAP, and reverse transcriptases. As such they are considered to be the most versatile nucleic acid polymerase family since different members can utilize DNA or RNA templates to synthesise DNA or RNA, in any combination. The two-barrel nucleic acid polymerase superfamily comprises the multisubunit RNAPs (msRNAPs) that carry out transcription of the cellular genomes of bacteria, archaea and eukaryotes as well as the chloroplast genome. Recently some surprising additions have been made to this family that increase its functional breadth, not only in terms of template specificity but also with respect to the mechanisms of site-specific transcription initiation.

PolD and Qde1 contain double-Ψ β-barrels

The bulk of a cellular msRNAP is provided by the two large, catalytic subunits called β' and β, and Rpo1 and Rpo2 in the single bacterial- and archaeal RNAP, respectively, and RPB1 and RPB2 in eukaryotic RNAPII. The subunits show striking sequence- and structural similarities which are highest in the active site microenvironment (21; 32; 51; 86). These include the Trigger loop and Bridge helix elements, which are essential for nucleotide translocation cycle (24; 40; 44; 60; 65; 83; 99; 100). The catalytic centre is formed at the interface between two six-stranded double-Ψ β-barrel domains coined DPBB-A and -B (i.e. the two barrels), with each catalytic subunit contributing one DPBB domain (36). The DPBB-A of the largest β', Rpo1 and RPB1 RNAP subunit contributes three invariant aspartic acid residues in the highly conserved NADFDGD motif to the active centre that coordinates the catalytic Magnesium-A ion (Figure 1A) (81). The DPBB-B of the second largest subunit (β, Rpo2 and RPB2) provides two invariant lysine residues involved in substrate binding (22). The pervasiveness of the DPBBs architecture in msRNAPs is well known and was thought to be restricted to DNA-dependent transcription. Remarkably, the structures of two unorthodox nucleic acid polymerases revealed that this structural framework also can support RNA-dependent RNA synthesis and DNA-dependent DNA synthesis (73; 75). Both the eukaryotic RNA-dependent RNAP Qde-1 that facilitates RNA-silencing in the fungus *Neurospora crassa*, and the catalytic subunit DP2 of the replicative archaeal DNA polymerase D (PolD) are two-barrel polymerases (73; 75) (Figure 1B). The DPBB-A type barrels (named DPBB-2) with catalytic carboxylate residues are located in the

C-terminal regions of Qde-1 and PolD DP2. The DPBB-2 domain of Qde-1 contains the consensus DxDGD motif and binds the catalytic Mg^{2+} ion, whereas the DPBB-2 domain of PolD has only two of the canonical aspartic acid residues (NxDGD), except in species in the phylum of Thaumarchaeota, where the canonical three aspartic acid residues are conserved (i.e *Nitrosopumilus maritimus* in Figure1C). The X-ray structure of PolD lacks the catalytic Mg^{2+} ion but the two aspartic residues in DPPB-2 were shown to be essential for PolD activity consistent with a role in Mg^{2+} ion coordination (80). According to the paradigm of two-barrel polymerases the second, DPBB-B type barrel (named DPBB-1), harbours two canonical lysine residues in both Qde-1 and PolD DP2 (73; 75). Beyond the two DPBB domains no further structural similarity was detected between msRNAPs, Qde-1 and PolD.

Despite the fact that the catalytic subunits of msRNAPs are highly conserved in all three domains of life, the gene encoding the largest subunit is split into two ORFs encoded by adjacent genes in Archaea (Rpo1) and chloroplast plastids (β') (64; 94). A closer look on RNA polymerase gene organisation in archaea reveals even higher levels of complexity. In methanobacteria and halobacteria (both archaea despite the misleading names) the largest (Rpo1) and second largest RNAP subunit (Rpo2) are split, whereas Rpo1 is encoded by a single ORF in Thaumarchaeota and Korarchaeota (15). Those observations would suggest that multiple split and fusion events of catalytic core subunits occurred over the time. Interestingly, insertion of the corresponding split sites into the genes encoding the catalytic core subunits of *E. coli* RNAP does result in active enzymes (77). In contrast to msRNAPs, Qde-1 and

in PolD contain the two DPBB domains within a single polypeptide chain, which could be the result of a fusion of DPBB encoding genes. The genes encoding the largest- and second largest RNAP subunit are encoded in a polycistronic operon, and this organisation is conserved between bacteria and archaea, whether the genes for the two subunits are split, or not. In fact, ε–proteobacteria harbour a single fused catalytic subunit encompassing both β and β' (50), and the fusion of *rpoB* (encoding for β) and *rpoC* (encoding for β') genes in *E. coli* results in a functional RNAP in vitro and in vivo (76). The fused single catalytic core subunit encompasses DPBB-A in its C-terminal half and DPBB-B in its N-terminal half in the same order as found in its counterparts Qde-1 and PolD. In summary, the two DPBBs at the catalytic heart of msRNAP reveal an intriguing structural and functional conservation across a very broad range of 'two-barrel' polymerases, and the remarkable variation of the arrangement of the genes encoding the DPBB subunits bears witness to several split and fusion events during evolution.

Evolutionary insights from viral two-barrel RNAPs

Next to the two catalytic core subunits, all cellular msRNAPs include universally conserved subunits that play an important role for the efficient assembly of the two large catalytic subunits of msRNAPs (31; 55). These include the α_2 homodimer in bacteria that is homologous to Rpo3/11 and RPB3/11 in archaea and eukarya, respectively) and ω (Rpo/RBP6). The former constitute the RNAP assembly platform that in archaea and eukaryotes also includes the Rpo/RPB10 and Rpo/RPB12 subunits. Due to the universal nature of assembly platform subunits it was assumed

that they were essential for the correct and stable folding of the catalytic subunits of two-barrel msRNAPs. Recently, Minakhin and co-workers identified and biochemically characterised the first two-barrel msRNAP encoded by a giant bacteriophage. ϕKZ non-virion RNAP (nvRNAP) is evolutionarily related to the msRNAP of the host it infects, and is likely the result of a horizontal transfer of the genes encoding the two catalytic subunits (95). Interestingly, ϕKZ nvRNAP does not include any classical assembly platform subunits. However, ϕKZ nvRNAP harbours gp68, a subunit without any similarity to other proteins other than its homologues in related giant bacteriophages, which may play a role in nvRNAP assembly. Similarly, several msRNAPs from different eukaryotic virus families appear to lack assembly platform subunits. Insect baculoviruses encode nvRNAPs composed of only four subunits: the two largest subunits (LEF-8 and LEF-9) sharing sequence similarity with the DPBB-A and –B of msRNAPs, and two additional subunits (LEF-4 and p47) with no sequence homology with any known msRNAPs subunits (28; 71). Thus, these nvRNAPs lack distinguishable assembly subunits, which altogether demonstrates that these are not required for the efficient assembly of DPBB msRNAPs per se. Detailed structural and functional analysis of the nvRNAPs and its kindred is sure to reveal many surprises in the coming years.

The origins of the barrels in the RNA world?

While msRNAPs chiefly function as DNA-dependent RNAPs they can utilise RNA templates in some special cases in vitro and in vivo; e.g. human RNAPII facilitates the replication of the Hepatitis Virus D genome by RNA template-dependent RNA

synthesis (19; 67; 88). X-ray structures of yeast RNAPII with RNA scaffold templates show that it can accommodate an RNA duplex in similar manner to the RNA-DNA hybrid formed during DNA-dependent transcription. These results highlight the potentially ancient RNA-dependent activity of msRNAPs (53) and are in line with the idea of an 'RNA-protein world' preceding the modern era of cells employing DNA as genetic material. The common ancestor of extant msRNAPs likely evolved from a primordial RNA-dependent two-barrel RNAP, which consisted primarily of the DPBB motifs. In msRNAPs, the DPBB may later have contributed to its adaption to utilise double-stranded DNA as templates. It is thought that the primordial RNA-dependent RNAP appeared at the RNA world era and was a self-replicating RNA ribozyme (Figure 2, yellow panel). If that was the case, the processivity and fidelity of the primal ribozyme must have been sufficiently high to self-replicate, something that has not been achieved yet with synthetic ribozymes in vitro (38). Following the emergence of templated protein synthesis, binding of a RNA-binding proteinaceous cofactor containing a DPBB domain to the catalytic core of the RNAP ribozyme may have increased its stability, processivity and fidelity - all critical factors for efficient and faithful transcription. Given that the two DPBB domains of two-barrel polymerases are evolutionary related, the ancestor of those enzymes most likely functioned as a homodimer. Duplication followed by divergent evolution resulted in functional specialisation of the two DPBB domains: acquisition of metal chelating aspartates by DPBB-A/2 and acquisition of basic residues by DPBB-B/1 (Figure 2). Crucially, the template specificity changed from RNA to DNA, and in an unexplained fashion the catalysis was usurped by the DPBB proteins and the now obsolete

ancestral RNA was lost. The RNAP evolved by increasing its bulk by the acquisition of modules/domains into the DPBB-containing large subunits, and increasing its subunit repertoire via accretion of additional, reversible associated factors around the conserved core (37; 89). The ultra-minimal active site of two-barrel msRNAPs appears to be composed of the two DPBB domains, the switch 2 element that interacts with the template DNA strand, and the secondary channel that allows for entrance of the nucleotide substrates (71). The reduced subunit repertoire of the different viral msRNAPs described above supports the idea that primordial msRNAPs was mainly composed of the two catalytic core subunits.

The search for the evolutionary origins of the general transcription factors

In contrast to single-subunit RNAPs such as bacteriophage T7 RNAP that are able to initiate transcription without additional factors, all cellular msRNAPs strictly rely on basal transcription factors. Basal transcription factors facilitate promoter recognition, local melting of DNA and template strand loading into the RNAP active site to form the open complex (OC) ready for transcription initiation. Bacterial RNAP rely chiefly on a single σ^{70}–related transcription factor, whereas archaeal transcription initiation involves three basal transcription factors, TBP, TFB and TFE (29; 66; 92). TFE appears to have evolved originally as a heterodimeric factor with α- and β-subunits, but many archaea retained only the α-subunit (11). In eukaryotes TBP- and TFB-related factors are required for transcription initiation by the three canonical nuclear RNAPs. Similar to the archaea, the combination of TBP and TFIIB is necessary and sufficient for site specific transcription initiation of eukaryotic RNAPII in vitro on

strong promoters using a negatively supercoiled DNA template topology (29; 61; 66). Archaeal TFE has counterparts in the RNAPII and III transcription machineries (TFIIE and the RNAPIII sub-complex C82/34, respectively) that appear to carry out similar functional roles in OC formation next to additional functions in the recruitment of basal transcription factors or RNAP itself (11; 17; 33; 58; 63; 92). All other eukaryotic basal transcription initiation factors appear to be specific to this domain of life.

None of the three archaeo-eukaryotic basal transcription factors have clear homologous counterparts in bacteria but nevertheless some intriguing clues about their deep evolutionary origin are emerging. (i) TBP is a highly symmetrical saddle-shaped protein that consists of two β-sheet domains called 'TBP domains'. Interestingly, individual TBP domains are present in some bacterial nucleases (RNaseH III) and DNA glycosylases that are ubiquitous in all three domains of life, which demonstrates that the ancestry of the TBP domain predates the Last Universal Common Ancestor (LUCA) (14) (Figure 3A). (ii) The C-terminal core domain of TFB/TFIIB includes two multihelical bundle helix-turn-helix motifs (HTH) that are distantly related on structural level to tri-helical HTH motifs present in bacterial σ^{70} (1; 34) (Figure 3B). The internal symmetry of the TFB/TFIIB core domain dictates that the evolutionary ancestor must have contained a single multihelical-type HTH motif. (iii) The two subunits of TFE/TFIIE contain winged helix-turn-helix motifs (WH), a structural motif abundant in all three domains of life. Sequence analysis of the WH motifs from both TFE subunits suggest that they share a common ancestry with transcription regulators of the MarR family (1; 10; 12) (Figure 3C).

Are bacterial sigma and archaeo-eukaryotic TFB/TFIIB factors evolutionary related?

It has been proposed by Iyer and Burton that bacterial σ^{70} and archaeal and eukaryotic TFB/TFIIB share a common ancestry based on the fact that the C-terminal HTH motifs in both cases are involved in promoter recognition at similar position relative to the transcription start sites (TSS) (18; 34). Their binding mode follows the canonical way of DNA-binding HTH-motifs, i.e. intercalating α-helix 3 into the major groove of the DNA template. Recent high-resolution structures of bacterial and eukaryotic closed- and open complexes (5; 30; 63; 102) have provided us with a more detailed picture of transcription initiation of σ^{70}- and TBP/TFB-related transcription machineries. But even with this additional structural insight, the apparent homology between σ^{70} and TFB remains limited to the canonical HTH recognition of promoter DNA.

Analogous initiation mechanisms in the three domains of life

Despite the lack of robust homology between the basal transcription factors, there are several features shared between the bacterial and archaeal/eukaryotic transcription initiation machineries. Firstly, the extent of the DNA bubble formed during OC formation is near identical in the archaeal and the bacterial OC based on permanganate foot-printing assays (7; 11; 57; 74) and recent high-resolution structures of the bacterial OC (102). In bacteria, the border of the DNA bubble is marked by the -10 promoter element, an AT-rich promoter element, while archaeal

promoters generally show a preference for A/T around position -10 without constituting a sequence motif, such as found in proper promoter elements facilitating specific interaction with basal transcription factors (12). Secondly, transcription initiation universally prefers purine residues as initial nucleotides preceded by a pyridine in the corresponding -1 position on the non-template strand as RNA-seq TSS mapping data from bacteria and archaea revealed (4; 20; 42; 93). *E. coli* in vitro transcription experiments using a library of randomized sequences confirmed the YR nucleotide preference at position -1/+1 (87). The same preference appears to be echoed in the consensus of human and *Drosophila* initiator promoter elements (39). Thirdly, TF(II)B and σ^{70} both facilitate transcription initiation by stabilising the template strand through the TF(II)B B-reader element and σ^{70} region 3.2, respectively (3; 5; 48; 72; 102). And lastly, σ^{70} and TFE/TFIIE both prevent the universally conserved transcription elongation factor NusG/Spt5 and its bacterial paralogue RfaH from associating with the RNA polymerase during the initiation stage (9; 27; 54; 78). It is important to stress that a shared feature between the bacterial and archaeal transcription initiation machineries does not necessarily mean that the feature is evolutionarily conserved. The YR preference at position -1/+1 appears to rather reflect the ability for a template strand purine at position -1 to stabilise the incoming +1 nucleotide via base stacking interactions and this feature might be common to all types of RNAPs. This was identified not only in the initially transcribing complex of an msRNAP from *Thermus thermophilus* (6) but also in the evolutionarily unrelated single-subunit bacteriophage N4 RNAP (25).

The apparent absence of universally conserved basal transcription factors facilitating transcription initiation is in contrast to the universal conservation of the transcription elongation factor NusG/Spt5 (90). This prompted us to speculate that (i) the regulation of elongation preceded the regulation of initiation in the primordial transcription system of LUCA – the 'elongation-first' hypothesis, and (ii) that initiation could have been relatively non start site-specific prior to the emergence of dedicated initiation factors (91). It remains impossible to infer whether the basal transcription machinery of LUCA contained TFB/TBP-like or σ^{70}-like factors or a combination of both or none (91). Nevertheless, it is worth considering the possible scenarios in the context of other basal transcription machineries in extant life forms and their viruses. The focus on TFB/TFIIB and σ^{70} blends out the real complexity of the different transcription initiation pathways that evolved in cellular life as well as in the virosphere. In fact, a third, phylogenetically unrelated basal transcription factor evolved in bacteria: σ^{54}. While σ^{54} and σ^{70} are composed of multiple domains with similar functions, these domains are not homologous (96). It is generally thought that an evolutionary advantage of σ^{54} may lay in tighter gene regulation as σ^{54}-mediated transcription initiation is fully dependent on the ATPase activity of bacterial enhancer binding proteins (bEBPs). σ^{54} and σ^{70} are able to regulate transcription of the same genes by using alternative promoters with different TSS (16). The patchy, but phylogenetically broad distribution of σ^{54} suggests that two different types of basal transcription factors that co-evolved with their own sets of transcriptional regulators have coexisted in bacteria since the early stages of bacterial evolution.

Within the boundaries of cellular life, the strict separation of TBP/TFB and σ^{70}-based transcription systems in archaea and bacteria, respectively, was recently challenged by the discovery of genes encoding σ^{70} homologs in several novel archaeal species by single-cell genomics (69). Phylogenetic analysis of these genes suggests that they are derived from horizontal gene transfer from bacteria. While the genome sequences of these archaeal species are still incomplete, it appears that at all these species also possess the canonical archaeal basal transcription factors (69). Whether the archaeal σ factors actually play a role in transcription in these species remains to be functionally verified.

Clues from unorthodox RNAPs from bacteriophages and eukaryotic viruses

The ability of msRNAPs to evolve an alternative basal 'support' machinery, unrelated to TFB and σ factors, was recently highlighted by the biochemical characterization of transcription initiation by ϕKZ nvRNAP (95). ϕKZ nvRNAP appears to be required for transcription from late promoters in the bacteriophage genome. While the full context of promoter elements directing transcription initiation is not yet fully understood, a TATG motif stretching from -3 to +1 relative to the TSS is essential. Transcription initiation of ϕKZ nvRNAP is not dependent on additional basal transcription factors, however, it is possible that the gp68 subunit plays a role in transcription initiation in vivo.

The discovery of giant viruses belonging to the proposed order Megavirales may bring yet more surprises about the evolution of msRNAPs in the virosphere and their mechanisms of transcription initiation. Members of Megaviridae and Poxviridae

families are double-stranded DNA viruses that encode msRNAPs that related to eukaryotic RNAPII (45; 56; 79; 98) as well as in some cases divergent homologues of basal transcription factors TBP and TFIIB in their genomes (35; 97). The African Swine Fever Virus (ASFV) is an extremely potent pathogen that causes haemorrhagic fever in domesticated pigs. ASFV genomes encode seven genes that are related to RNAPII subunits including the two large DPBB-containing catalytic subunits, and a fusion protein containing the two RPB3 and RPB11 assembly platform subunits. But maybe most surprisingly, while AFSV encodes a protein that is distantly related to TFIIB, no TBP homologues could be identified (70). Extracts prepared from ASF viroids are transcription competent (49), and since AFSV is propagated in two very different host environments (wild pigs such warthogs and bushpigs, and argasidae ticks), it is likely that the viral genome indeed encodes all components required for transcription without the need to coopt factors from the host cell. Only a few AFSV promoters have been partially characterised, none of which include classical RNAPII-like promoter elements such as BRE or TATA motifs (ie. binding sites for TFIIB and TBP) at a meaningful distance to the mapped transcription start sites (70).

Despite the increasing volume of information into the molecular mechanisms of transcription initiation in bacteria, archaea and eukaryotes, the lack of extended homology between TFB/TFIIB and σ^{70} makes it challenging if not impossible to draw persuasive conclusions about the nature of basal transcription factors in LUCA. Meanwhile, an increasing amount of genomic and biochemical data from microbial 'dark matter', bacteriophages and eukaryotic viruses draw a more complex picture with alternative modes of transcription initiation and inter-domain gene transfer of

both RNAP subunits and basal transcription factors. The example of σ^{54} and σ^{70} as two basal transcription factors co-existing in the same organism and co-regulating transcription might suggest that TFB/TBP- and σ^{70} have possibly co-evolved in LUCA from independent origins, rather than both factors evolving from the same proto-transcription factor present in LUCA before their structural and functional divergence in bacteria and archaea/eukaryotes (12).

The connection between transcription initiation and translation initiation

The functional and structural diversity of basal transcription initiation mechanisms in cellular life make it difficult to draw conclusions on the nature of the basal transcription machinery in LUCA. However, some of its functional properties can be deducted. To this end, it is worth to consider the products of transcription, coding and non-coding RNA in regard to their specific requirements of TSS selection. All three domains of life share two conserved translation initiation factors: IF1 and IF2 in bacteria (aeIF-1a and aeIF5B in archaea/eukaryotes, respectively). Their conserved role is thought to be guiding the aminoacylated initiator tRNA to the P-site (8). Additional non-homologous translation initiation factors are present in archaea and bacteria and the two primary domains especially diverged regarding selection of the aminoacylated initiator tRNA (8). Two conserved modes of translation initiation can be distinguished in bacteria and archaea: 70S ribosome initiation on leaderless mRNA and initiation starting with binding of the 30S ribosomal subunit to ribosomal binding site (RBS) present in mRNAs with 5'-UTR as well as in downstream cistrons of polycistronic mRNAs (8). In bacteria, translation initiation from leaderless mRNA

can occur in a factor-independent manner (85). The molecular basis of leaderless translation initiation in archaea is not yet understood. RBS-dependent translation initiation generally requires the aid of initiation factors. Based on their broad occurrence across the two prokaryotic domains of life, it is highly likely that both leaderless and RBS-dependent translation initiation mechanisms were operating in LUCA (59; 101). It has been argued, however, that leaderless translation initiation is evolutionary more ancient (8). This is based on the fact that leaderless mRNAs can be utilized in all three domains of life: more generally in archaea and bacteria (20; 93; 101), but also in the protozoan *Giardia lamblia* (23) as well as in a rabbit reticulocyte in vitro translation system (26). Since leaderless translation initiation requires that the TSS and the start codon overlap, the universal preference of msRNAPs to initiate transcription with guanine nucleotides and the choice of ATG/GTG as start codons could be functionally linked.

RBS-dependent and leaderless translation initiation have distinct advantages in terms of gene regulation. RBS-dependent translation is thought to aid the coordinated expression of genes organized in operons (Figure 4A) (101). This is of critical importance especially for larger heterooligomeric complexes such as ribosomes and msRNAPs themselves. Indeed, in organisms that preferably use leaderless mRNAs such as *Mycobacterium* and the archaeon *Sulfolobus* 5'-UTRs are still retained in the mRNAs of ribosomal protein encoding genes (20; 93). The operon encoding the two catalytic subunits of RNAP is a rare example of gene organisation being conserved between bacteria and archaea testifying the

importance of operons and RBS-dependent translation initiation in the coordinated expression of components of large heterooligomeric complexes. In both primary domains, transcription and translation are physically coupled and RBS-dependent translation initiation might facilitate the coordination between the two processes. Indeed, recent NET-seq data from *E. coli and Bacillus subtilis* RNAPs show that they tend to pause at translation start sites possibly ensuring maintenance of coupling (52). RBS-dependent translation initiation also allows for multiple promoters/TSS to be used to regulate transcription of a gene (Figure 4A). Lastly, RBS-dependent translation from 5'-UTR containing mRNAs can be regulated by small RNAs either by blocking access to the RBS or enabling access to it through changes in secondary structure (82). On the other hand, leaderless mRNAs are thought to allow for tighter regulation and preventing gene expression from spurious transcription or read-through from transcriptional units placed upstream in sense orientation (Figure 4B) (13; 101).

Structural features of noncoding RNA genes may provide additional clues to these questions. There are four different types of universally conserved noncoding RNA genes/operons: transfer tRNA, ribosomal rRNA operons, 4.5S RNA (the RNA component of the signal recognition particle) and the RNA component of RNase P. The majority of these universally conserved ncRNA genes undergo 5' processing: The 16S rRNA gene is the first gene in the rRNA operon. The 5'-end of mature 16S rRNA is generated via the combined action of several RNases in bacteria (2). RNase P is required for 5' processing of tRNA as well as 4.5S RNA (41; 62; 84). The requirements for the maturation of the RNA components of RNase P itself are less

clear, but it has been reported that the productive transcription of M1 RNA, the RNA component of RNase P in *E. coli*, is driven from a proximal promoter that does not require 5′-end processing (46). Taken together, most of the universally conserved noncoding RNAs require 5′-end processing and this requirement could reflect the functional properties of the early transcription initiation machinery. However, it should be mentioned that 5′-end processing is also required for many domain-specific ncRNAs such as transfer-messenger RNA (tmRNA) and 6S RNA that evolved later in the bacterial domain (43; 47; 84). On the other hand it has been shown that the universal requirement for 5′-processing of tRNA by RNase P is not essential for life and has been overcome by transcription initiation at proper 5′-end in the archaeon *Nanoarchaeum equitans* (68). Independent of the 5′-end processing requirements for these universally conserved RNAs it can be inferred that the arguably evolutionary oldest genes probably have a relaxed requirement for transcription start site selection allowing for multiple promoters/TSSs to be utilised.

Conclusion

The discovery of viral msRNAPs with reduced subunit repertoire and basal transcription factor requirement and two-barrel DNA polymerases have advanced our understanding of the evolution of msRNAPs and the crucial role of the DPBB domains. Viral msRNAPs have evolved divergent catalytic subunit assembly pathways and mechanisms for site-specific transcription initiation that may provide clues to the evolution of transcription in cellular life. Transcription is the first step in gene expression towards protein synthesis and thereby the mechanisms of

transcription initiation and TSS selection directly affect the mechanism of translation initiation and vice versa. We argue that RBS-dependent translation initiation (and 5'-end processing of non-coding RNAs) might have contributed to the environment conducive for the evolution of alternative basal transcription factors such as σ^{70} and TBP/TFB in the same organism (Figure 4C). An alternative model, with precise selection of single TSS coupled to leaderless translation initiation would impose several restrictions on the organism in terms of the regulation of gene expression and the ability to evolve alternative basal transcription factors (Figure 4D). For these reasons, we consider the most likely scenario to be the early appearance of RBS-dependent translation initiation in evolution and parallel evolution of multiple basal transcription initiation factors.

Acknowledgments

This work was supported by Wellcome Trust Investigator Award WT096553MA to F.W.

Literature cited

1. Aravind L, Anantharaman V, Balaji S, Babu MM, Iyer LM. 2005. The many faces of the helix-turn-helix domain: transcription regulation and beyond. *FEMS microbiology reviews* 29:231-62
2. Arraiano CM, Andrade JM, Domingues S, Guinote IB, Malecki M, et al. 2010. The critical role of RNA processing and degradation in the control of gene expression. *FEMS microbiology reviews* 34:883-923
3. Artsimovitch I, Vassylyeva MN, Svetlov D, Svetlov V, Perederina A, et al. 2005. Allosteric modulation of the RNA polymerase catalytic reaction is an essential component of transcription control by rifamycins. *Cell* 122:351-63
4. Babski J, Haas KA, Nather-Schindler D, Pfeiffer F, Forstner KU, et al. 2016. Genome-wide identification of transcriptional start sites in the haloarchaeon Haloferax volcanii based on differential RNA-Seq (dRNA-Seq). *BMC Genomics* 17:629
5. Bae B, Feklistov A, Lass-Napiorkowska A, Landick R, Darst SA. 2015. Structure of a bacterial RNA polymerase holoenzyme open promoter complex. *Elife* 4
6. Basu RS, Warner BA, Molodtsov V, Pupov D, Esyunina D, et al. 2014. Structural basis of transcription initiation by bacterial RNA polymerase holoenzyme. *The Journal of biological chemistry* 289:24549-59
7. Bell SD, Jaxel C, Nadal M, Kosa PF, Jackson SP. 1998. Temperature, template topology, and factor requirements of archaeal transcription. *Proc Natl Acad Sci U S A* 95:15218-22
8. Benelli D, Londei P. 2009. Begin at the beginning: evolution of translational initiation. *Res Microbiol* 160:493-501
9. Blombach F, Daviter T, Fielden D, Grohmann D, Smollett K, Werner F. 2013. Archaeology of RNA polymerase: factor swapping during the transcription cycle. *Biochemical Society transactions* 41:362-7
10. Blombach F, Makarova KS, Marrero J, Siebers B, Koonin EV, van der Oost J. 2009. Identification of an ortholog of the eukaryotic RNA polymerase III subunit RPC34 in Crenarchaeota and Thaumarchaeota suggests specialization of RNA polymerases for coding and non-coding RNAs in Archaea. *Biol Direct* 4:39
11. Blombach F, Salvadori E, Fouqueau T, Yan J, Reimann J, et al. 2015. Archaeal TFEalpha/beta is a hybrid of TFIIE and the RNA polymerase III subcomplex hRPC62/39. *Elife* 4:e08378
12. Blombach F, Smollett KL, Grohmann D, Werner F. 2016. Molecular Mechanisms of Transcription Initiation-Structure, Function, and Evolution of TFE/TFIIE-Like Factors and Open Complex Formation. *J Mol Biol* 428:2592-606
13. Bolotin A, Wincker P, Mauger S, Jaillon O, Malarme K, et al. 2001. The complete genome sequence of the lactic acid bacterium Lactococcus lactis ssp. lactis IL1403. *Genome Res* 11:731-53
14. Brindefalk B, Dessailly BH, Yeats C, Orengo C, Werner F, Poole AM. 2013. Evolutionary history of the TBP-domain superfamily. *Nucleic Acids Research* 41:2832-45
15. Brochier-Armanet C, Gribaldo S, Forterre P. 2008. A DNA topoisomerase IB in Thaumarchaeota testifies for the presence of this enzyme in the last common ancestor of Archaea and Eucarya. *Biol Direct* 3:54

16. Brown DR, Barton G, Pan Z, Buck M, Wigneshweraraj S. 2014. Nitrogen stress response and stringent response are coupled in Escherichia coli. *Nature communications* 5:4115
17. Brun I, Sentenac A, Werner M. 1997. Dual role of the C34 subunit of RNA polymerase III in transcription initiation. *EMBO J* 16:5730-41
18. Burton SP, Burton ZF. 2014. The sigma enigma: bacterial sigma factors, archaeal TFB and eukaryotic TFIIB are homologs. *Transcription* 5:e967599
19. Cavanagh AT, Wassarman KM. 2014. 6S RNA, a global regulator of transcription in Escherichia coli, Bacillus subtilis, and beyond. *Annu Rev Microbiol* 68:45-60
20. Cortes T, Schubert OT, Rose G, Arnvig KB, Comas I, et al. 2013. Genome-wide Mapping of Transcriptional Start Sites Defines an Extensive Leaderless Transcriptome in Mycobacterium tuberculosis. *Cell Rep* 5:1121-31
21. Cramer P. 2002. Common structural features of nucleic acid polymerases. *Bioessays* 24:724-9
22. Cramer P, Bushnell DA, Kornberg RD. 2001. Structural basis of transcription: RNA polymerase II at 2.8 angstrom resolution. *Science* 292:1863-76
23. Forget D, Langelier MF, Therien C, Trinh V, Coulombe B. 2004. Photo-cross-linking of a purified preinitiation complex reveals central roles for the RNA polymerase II mobile clamp and TFIIE in initiation mechanisms. *Mol Cell Biol* 24:1122-31
24. Fouqueau T, Zeller ME, Cheung AC, Cramer P, Thomm M. 2013. The RNA polymerase trigger loop functions in all three phases of the transcription cycle. *Nucleic Acids Res* 41:7048-59
25. Gleghorn ML, Davydova EK, Basu R, Rothman-Denes LB, Murakami KS. 2011. X-ray crystal structures elucidate the nucleotidyl transfer reaction of transcript initiation using two nucleotides. *Proc Natl Acad Sci U S A* 108:3566-71
26. Grill S, Gualerzi CO, Londei P, Blasi U. 2000. Selective stimulation of translation of leaderless mRNA by initiation factor 2: evolutionary implications for translation. *EMBO J* 19:4101-10
27. Grohmann D, Nagy J, Chakraborty A, Klose D, Fielden D, et al. 2011. The initiation factor TFE and the elongation factor Spt4/5 compete for the RNAP clamp during transcription initiation and elongation. *Mol Cell* 43:263-74
28. Guarino LA, Xu B, Jin J, Dong W. 1998. A virus-encoded RNA polymerase purified from baculovirus-infected cells. *J Virol* 72:7985-91
29. Hausner W, Wettach J, Hethke C, Thomm M. 1996. Two transcription factors related with the eucaryal transcription factors TATA-binding protein and transcription factor IIB direct promoter recognition by an archaeal RNA polymerase. *The Journal of biological chemistry* 271:30144-8
30. He Y, Yan C, Fang J, Inouye C, Tjian R, et al. 2016. Near-atomic resolution visualization of human transcription promoter opening. *Nature* 533:359-65
31. Heyduk T, Heyduk E, Severinov K, Tang H, Ebright RH. 1996. Determinants of RNA polymerase alpha subunit for interaction with beta, beta', and sigma subunits: hydroxyl-radical protein footprinting. *Proc Natl Acad Sci U S A* 93:10162-6
32. Hirata A, Klein BJ, Murakami KS. 2008. The X-ray crystal structure of RNA polymerase from Archaea. *Nature* 451:851-4

33. Holstege FC, Tantin D, Carey M, van der Vliet PC, Timmers HT. 1995. The requirement for the basal transcription factor IIE is determined by the helical stability of promoter DNA. *EMBO journal* 14:810-9
34. Iyer LM, Aravind L. 2012. Insights from the architecture of the bacterial transcription apparatus. *J Struct Biol* 179:299-319
35. Iyer LM, Balaji S, Koonin EV, Aravind L. 2006. Evolutionary genomics of nucleo-cytoplasmic large DNA viruses. *Virus Res* 117:156-84
36. Iyer LM, Koonin EV, Aravind L. 2003. Evolutionary connection between the catalytic subunits of DNA-dependent RNA polymerases and eukaryotic RNA-dependent RNA polymerases and the origin of RNA polymerases. *BMC Struct Biol* 3:1
37. Iyer LM, Koonin EV, Aravind L. 2004. Evolution of bacterial RNA polymerase: implications for large-scale bacterial phylogeny, domain accretion, and horizontal gene transfer. *Gene* 335:73-88
38. Johnston WK, Unrau PJ, Lawrence MS, Glasner ME, Bartel DP. 2001. RNA-catalyzed RNA polymerization: accurate and general RNA-templated primer extension. *Science* 292:1319-25
39. Kadonaga JT. 2012. Perspectives on the RNA polymerase II core promoter. *Wiley Interdiscip Rev Dev Biol* 1:40-51
40. Kaplan CD, Larsson KM, Kornberg RD. 2008. The RNA polymerase II trigger loop functions in substrate selection and is directly targeted by alpha-amanitin. *Mol Cell* 30:547-56
41. Kazantsev AV, Pace NR. 2006. Bacterial RNase P: a new view of an ancient enzyme. *Nature reviews. Microbiology* 4:729-40
42. Kim D, Hong JSJ, Qiu Y, Nagarajan H, Seo JH, et al. 2012. Comparative Analysis of Regulatory Elements between Escherichia coli and Klebsiella pneumoniae by Genome-Wide Transcription Start Site Profiling. *Plos Genet* 8
43. Kim KS, Lee Y. 2004. Regulation of 6S RNA biogenesis by switching utilization of both sigma factors and endoribonucleases. *Nucleic Acids Res* 32:6057-68
44. Kireeva ML, Nedialkov YA, Cremona GH, Purtov YA, Lubkowska L, et al. 2008. Transient reversal of RNA polymerase II active site closing controls fidelity of transcription elongation. *Mol Cell* 30:557-66
45. Knutson BA, Broyles SS. 2008. Expansion of poxvirus RNA polymerase subunits sharing homology with corresponding subunits of RNA polymerase II. *Virus Genes* 36:307-11
46. Ko JH, Han K, Kim Y, Sim S, Kim KS, et al. 2008. Dual function of RNase E for control of M1 RNA biosynthesis in Escherichia coli. *Biochemistry* 47:762-70
47. Komine Y, Kitabatake M, Yokogawa T, Nishikawa K, Inokuchi H. 1994. A tRNA-like structure is present in 10Sa RNA, a small stable RNA from Escherichia coli. *Proc Natl Acad Sci U S A* 91:9223-7
48. Kostrewa D, Zeller ME, Armache KJ, Seizl M, Leike K, et al. 2009. RNA polymerase II-TFIIB structure and mechanism of transcription initiation. *Nature* 462:323-30
49. Kuznar J, Salas ML, Vinuela E. 1980. DNA-dependent RNA polymerase in African swine fever virus. *Virology* 101:169-75
50. Lane WJ, Darst SA. 2010. Molecular evolution of multisubunit RNA polymerases: sequence analysis. *J Mol Biol* 395:671-85

51. Lane WJ, Darst SA. 2010. Molecular evolution of multisubunit RNA polymerases: structural analysis. *J Mol Biol* 395:686-704
52. Larson MH, Mooney RA, Peters JM, Windgassen T, Nayak D, et al. 2014. A pause sequence enriched at translation start sites drives transcription dynamics in vivo. *Science* 344:1042-7
53. Lehmann E, Brueckner F, Cramer P. 2007. Molecular basis of RNA-dependent RNA polymerase II activity. *Nature* 450:445-9
54. Liu B, Steitz TA. 2016. Structural insights into NusG regulating transcription elongation. *Nucleic Acids Res*
55. Minakhin L, Bhagat S, Brunning A, Campbell EA, Darst SA, et al. 2001. Bacterial RNA polymerase subunit omega and eukaryotic RNA polymerase subunit RPB6 are sequence, structural, and functional homologs and promote RNA polymerase assembly. *Proc Natl Acad Sci U S A* 98:892-7
56. Moss B, Ahn BY, Amegadzie B, Gershon PD, Keck JG. 1991. Cytoplasmic transcription system encoded by vaccinia virus. *The Journal of biological chemistry* 266:1355-8
57. Nagy J, Grohmann D, Cheung AC, Schulz S, Smollett K, et al. 2015. Complete architecture of the archaeal RNA polymerase open complex from single-molecule FRET and NPS. *Nature communications* 6:6161
58. Naji S, Grünberg S, Thomm M. 2007. The RPB7 orthologue E' is required for transcriptional activity of a reconstituted archaeal core enzyme at low temperatures and stimulates open complex formation. *The Journal of biological chemistry* 282:11047-57
59. Nakagawa S, Niimura Y, Miura K, Gojobori T. 2010. Dynamic evolution of translation initiation mechanisms in prokaryotes. *Proc Natl Acad Sci U S A* 107:6382-7
60. Nedialkov YA, Opron K, Assaf F, Artsimovitch I, Kireeva ML, et al. 2013. The RNA polymerase bridge helix YFI motif in catalysis, fidelity and translocation. *Biochim Biophys Acta* 1829:187-98
61. Parvin JD, Sharp PA. 1993. DNA topology and a minimal set of basal factors for transcription by RNA polymerase II. *Cell* 73:533-40
62. Peck-Miller KA, Altman S. 1991. Kinetics of the processing of the precursor to 4.5 S RNA, a naturally occurring substrate for RNase P from Escherichia coli. *J Mol Biol* 221:1-5
63. Plaschka C, Hantsche M, Dienemann C, Burzinski C, Plitzko J, Cramer P. 2016. Transcription initiation complex structures elucidate DNA opening. *Nature* 533:353-8
64. Puhler G, Leffers H, Gropp F, Palm P, Klenk HP, et al. 1989. Archaebacterial DNA-dependent RNA polymerases testify to the evolution of the eukaryotic nuclear genome. *Proc Natl Acad Sci U S A* 86:4569-73
65. Qiu C, Erinne OC, Dave JM, Cui P, Jin H, et al. 2016. High-Resolution Phenotypic Landscape of the RNA Polymerase II Trigger Loop. *Plos Genet* 12:e1006321
66. Qureshi SA, Bell SD, Jackson SP. 1997. Factor requirements for transcription in the Archaeon Sulfolobus shibatae. *EMBO J* 16:2927-36
67. Rackwitz HR, Rohde W, Sanger HL. 1981. DNA-dependent RNA polymerase II of plant origin transcribes viroid RNA into full-length copies. *Nature* 291:297-301
68. Randau L, Schroder I, Soll D. 2008. Life without RNase P. *Nature* 453:120-+

69. Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, et al. 2013. Insights into the phylogeny and coding potential of microbial dark matter. *Nature* 499:431-7
70. Rodriguez JM, Salas ML. 2013. African swine fever virus transcription. *Virus Res* 173:15-28
71. Ruprich-Robert G, Thuriaux P. 2010. Non-canonical DNA transcription enzymes and the conservation of two-barrel RNA polymerases. *Nucleic Acids Res* 38:4559-69
72. Sainsbury S, Niesser J, Cramer P. 2013. Structure and function of the initially transcribing RNA polymerase II-TFIIB complex. *Nature* 493:437-40
73. Salgado PS, Koivunen MR, Makeyev EV, Bamford DH, Stuart DI, Grimes JM. 2006. The structure of an RNAi polymerase links RNA silencing and transcription. *PLoS Biol* 4:e434
74. Sasse-Dwight S, Gralla JD. 1989. KMnO4 as a probe for lac promoter DNA melting and mechanism in vivo. *The Journal of biological chemistry* 264:8074-81
75. Sauguet L, Raia P, Henneke G, Delarue M. 2016. Shared active site architecture between archaeal PolD and multi-subunit RNA polymerases revealed by X-ray crystallography. *Nature communications* 7:12227
76. Severinov K, Mooney R, Darst SA, Landick R. 1997. Tethering of the large subunits of Escherichia coli RNA polymerase. *Journal of Biological Chemistry* 272:24137-40
77. Severinov K, Mustaev A, Kukarin A, Muzzin O, Bass I, et al. 1996. Structural modules of the large subunits of RNA polymerase. Introducing archaebacterial and chloroplast split sites in the beta and beta' subunits of Escherichia coli RNA polymerase. *The Journal of biological chemistry* 271:27969-74
78. Sevostyanova A, Svetlov V, Vassylyev DG, Artsimovitch I. 2008. The elongation factor RfaH and the initiation factor sigma bind to the same site on the transcription elongation complex. *Proc Natl Acad Sci U S A* 105:865-70
79. Sharma V, Colson P, Giorgi R, Pontarotti P, Raoult D. 2014. DNA-dependent RNA polymerase detects hidden giant viruses in published databanks. *Genome Biol Evol* 6:1603-10
80. Shen Y, Musti K, Hiramoto M, Kikuchi H, Kawarabayashi Y, Matsui I. 2001. Invariant Asp-1122 and Asp-1124 are essential residues for polymerization catalysis of family D DNA polymerase from Pyrococcus horikoshii. *The Journal of biological chemistry* 276:27376-83
81. Sosunov V, Zorov S, Sosunova E, Nikolaev A, Zakeyeva I, et al. 2005. The involvement of the aspartate triad of the active center in all catalytic activities of multisubunit RNA polymerase. *Nucleic Acids Res* 33:4202-11
82. Storz G, Vogel J, Wassarman KM. 2011. Regulation by small RNAs in bacteria: expanding frontiers. *Mol Cell* 43:880-91
83. Tan L, Wiesler S, Trzaska D, Carney HC, Weinzierl RO. 2008. Bridge helix and trigger loop perturbations generate superactive RNA polymerases. *J Biol* 7:40
84. Tous C, Vega-Palas MA, Vioque A. 2001. Conditional expression of RNase P in the cyanobacterium Synechocystis sp PCC6803 allows detection of precursor RNAs - Insight in the in vivo maturation pathway of transfer and other stable RNAs. *Journal of Biological Chemistry* 276:29059-66

85. Udagawa T, Shimizu Y, Ueda T. 2004. Evidence for the translation initiation of leaderless mRNAs by the intact 70 S ribosome without its dissociation into subunits in eubacteria. *The Journal of biological chemistry* 279:8539-46
86. Vassylyev DG, Sekine S, Laptenko O, Lee J, Vassylyeva MN, et al. 2002. Crystal structure of a bacterial RNA polymerase holoenzyme at 2.6 A resolution. *Nature* 417:712-9
87. Vvedenskaya IO, Zhang Y, Goldman SR, Valenti A, Visone V, et al. 2015. Massively Systematic Transcript End Readout, "MASTER": Transcription Start Site Selection, Transcriptional Slippage, and Transcript Yields. *Mol Cell* 60:953-65
88. Wagner SD, Yakovchuk P, Gilman B, Ponicsan SL, Drullinger LF, et al. 2013. RNA polymerase II acts as an RNA-dependent RNA polymerase to extend and destabilize a non-coding RNA. *EMBO J* 32:781-90
89. Werner F. 2007. Structure and function of archaeal RNA polymerases. *Mol Microbiol* 65:1395-404
90. Werner F. 2012. A nexus for gene expression-molecular mechanisms of Spt5 and NusG in the three domains of life. *J Mol Biol* 417:13-27
91. Werner F, Grohmann D. 2011. Evolution of multisubunit RNA polymerases in the three domains of life. *Nature reviews. Microbiology* 9:85-98
92. Werner F, Weinzierl RO. 2005. Direct modulation of RNA polymerase core functions by basal transcription factors. *Mol Cell Biol* 25:8344-55
93. Wurtzel O, Sapra R, Chen F, Zhu Y, Simmons BA, Sorek R. 2010. A single-base resolution map of an archaeal transcriptome. *Genome Res* 20:133-41
94. Xie WQ, Jager K, Potts M. 1989. Cyanobacterial RNA polymerase genes rpoC1 and rpoC2 correspond to rpoC of Escherichia coli. *J Bacteriol* 171:1967-73
95. Yakunina M, Artamonova T, Borukhov S, Makarova KS, Severinov K, Minakhin L. 2015. A non-canonical multisubunit RNA polymerase encoded by a giant bacteriophage. *Nucleic Acids Res* 43:10411-20
96. Yang Y, Darbari VC, Zhang N, Lu D, Glyde R, et al. 2015. TRANSCRIPTION. Structures of the RNA polymerase-sigma54 reveal new and conserved regulatory strategies. *Science* 349:882-5
97. Yutin N, Colson P, Raoult D, Koonin EV. 2013. Mimiviridae: clusters of orthologous genes, reconstruction of gene repertoire evolution and proposed expansion of the giant virus family. *Virol J* 10:106
98. Yutin N, Wolf YI, Koonin EV. 2014. Origin of giant viruses from smaller DNA viruses not from a fourth domain of cellular life. *Virology* 466-467:38-52
99. Yuzenkova Y, Bochkareva A, Tadigotla VR, Roghanian M, Zorov S, et al. 2010. Stepwise mechanism for transcription fidelity. *BMC Biol* 8:54
100. Zhang J, Palangat M, Landick R. 2010. Role of the RNA polymerase trigger loop in catalysis and pausing. *Nat Struct Mol Biol* 17:99-104
101. Zheng X, Hu GQ, She ZS, Zhu H. 2011. Leaderless genes in bacteria: clue to the evolution of translation initiation mechanisms in prokaryotes. *BMC Genomics* 12:361
102. Zuo Y, Steitz TA. 2015. Crystal structures of the E. coli transcription initiation complexes with a complete bubble. *Mol Cell* 58:534-40

Figures

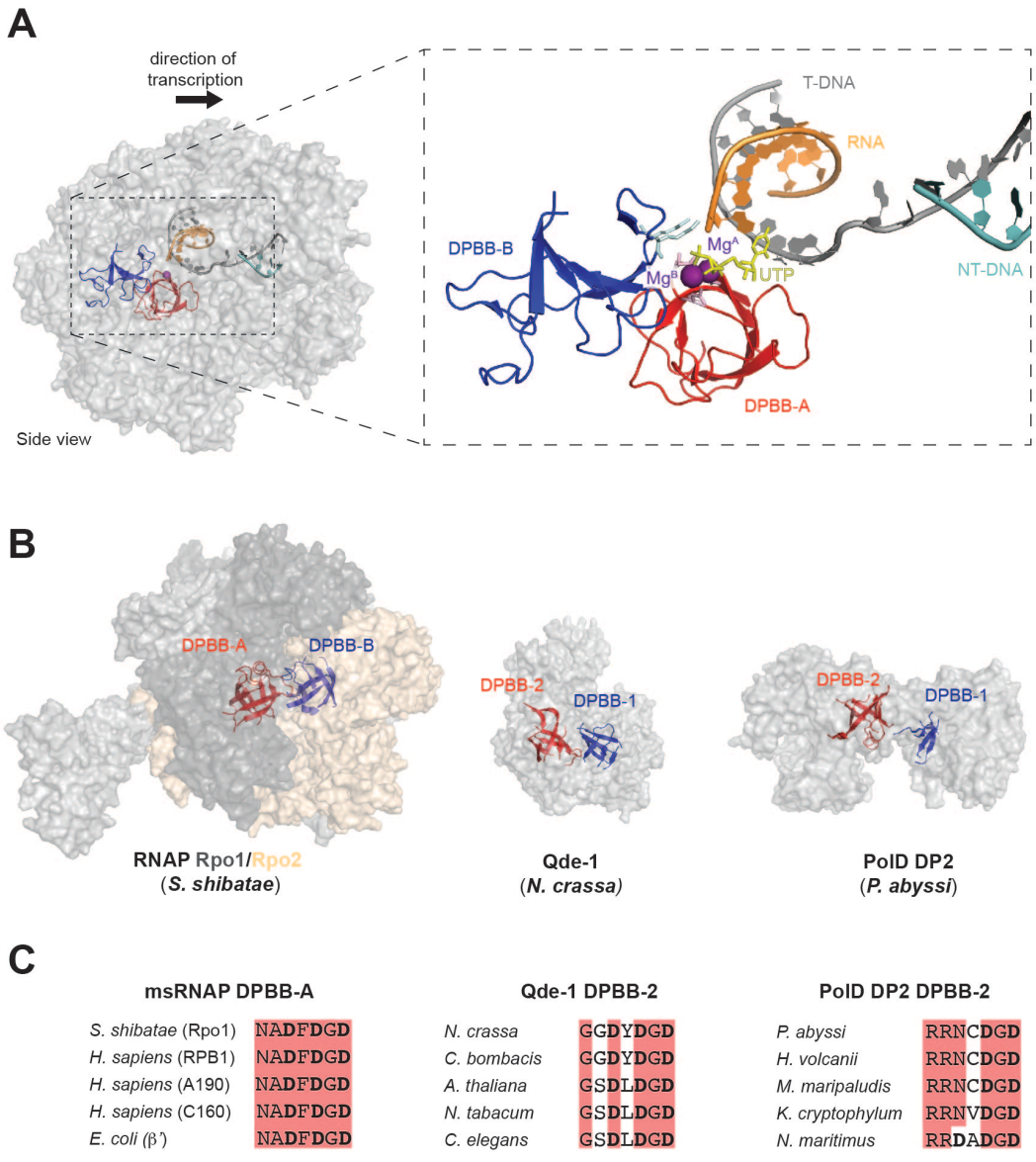

Figure 1: Evolution of the catalytic core of two-barrel polymerases.

(A) Structure of the conserved catalytic core of two-barrel msRNAPs. Three conserved aspartic acid residues of DPBB-A (stick representation in light pink) are coordinating the catalytic magnesium ion (Mg^A). The two conserved lysine residues of DPBB-B are shown as stick representation in light blue. The catalytic centre is occupied by UTP in complex with a second Magnesium ion (Mg^B). The schematic is based on the structure of *S. saccharomyces* RNAPII (PDB id: 2NVZ). (B) Structural

overview of the conserved catalytic core of two-barrel nucleic acid polymerases: msRNAP, RNA-dependent RNAP Qde-1 and DNA polymerase PolD. **(C)** Multiple sequence alignment of conserved catalytic motifs of (i) DNA-dependent RNAPs from *Sulfolobus shibatae* Rpo1 (ACL36488.1), *Homo sapiens* RPB1 (RNAP II: CAA45125.1); *Homo sapiens* A190 (RNAPII: AA126304.1), *Homo sapiens* C160 (RNAPIII: AAH41089.1) and *Escherichia coli* β' (AIX65985.1), (ii) RNA-dependent RNAPs from *Neurospora crassa* (EAA29811.1), *Ceraceosorus bombacis* (CEH11733.1), *Arabidopsis thaliana* (AEE29226.1), *Nicotiana tabacum* (CAA09697.1), and *Caenorhabditis elegans* (CAA88315.2), (iii) DNA polymerase PolD subunit DP2 from *Pyrococcus abyssi* (CAB49044.1), *Haloferax volcanii* (CAG38138.1), *Methanococcus maripaludis* (CAF29582.1), *Korearchaeaum cryptophylum* (ACB08273.1) and *Nitrosopumilus maritimus* (ABX 13690.1). Conserved motifs are highlighted in red. Catalytic aspartic acid residues are highlighted in bold.

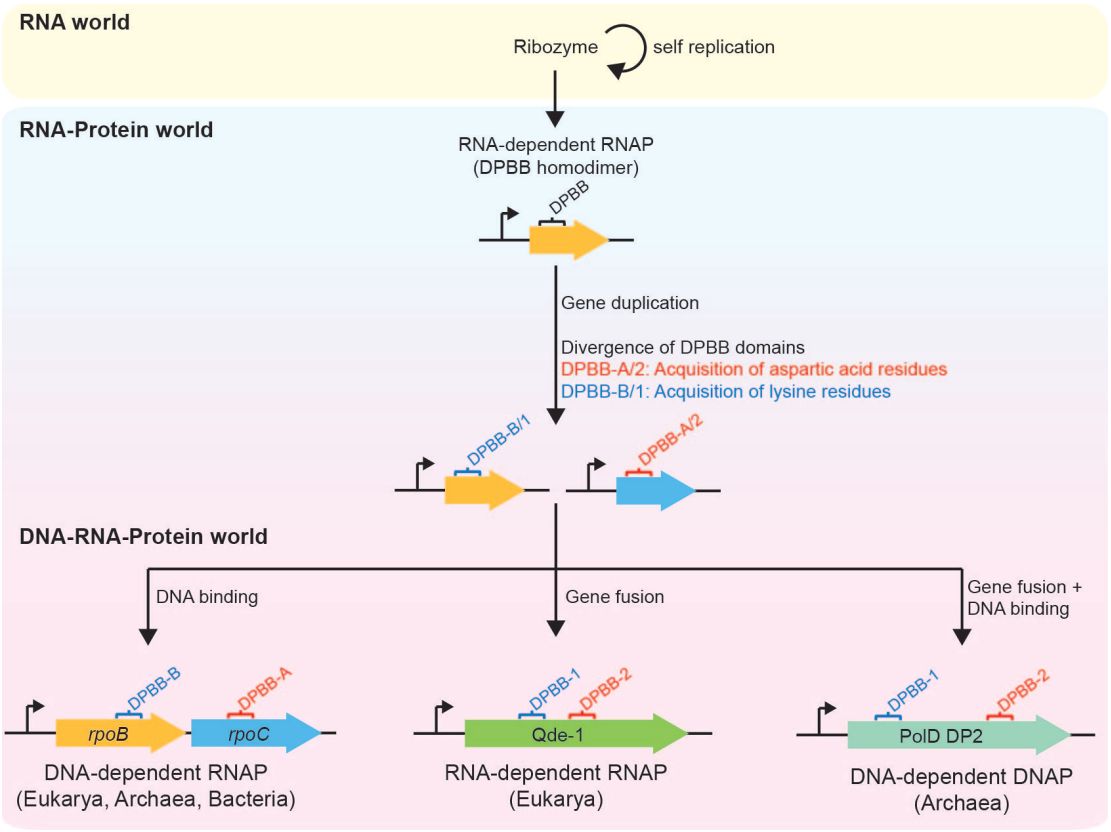

Figure 2: Hypothetical schematic for the evolution of the "two-barrel" nucleic acid polymerases.

The primordial RNAP was a self-replicating, RNA-dependent ribozyme which emerged in the RNA world era (in yellow). In the RNA-protein era, the ribozyme was invaded by cofactor containing a DPBB domain, forming ribonucleoprotein complex. The ultimate ancestor of two-barrel polymerases functioned as a homodimer enzyme. Duplication followed by divergent evolution resulted in acquisition of aspartic acid residues by DPBB-A/2, and acquisition of lysine residues by DPBB-B/1. In the modern DNA-RNA-protein era, the template specificity changed from RNA to DNA for msRNAPs and PolD. In Qde-1 and PolD, the two DPBB domains are in a single polypeptide (DPBB-1 at the N-terminus and DPBB-2 at the C-terminus), suggesting a fusion of DPBB encoding genes. msRNAP genes

containing the DPBB domains are encoded in a polycistronic operon, and the organisation is conserved between bacteria and archaea.

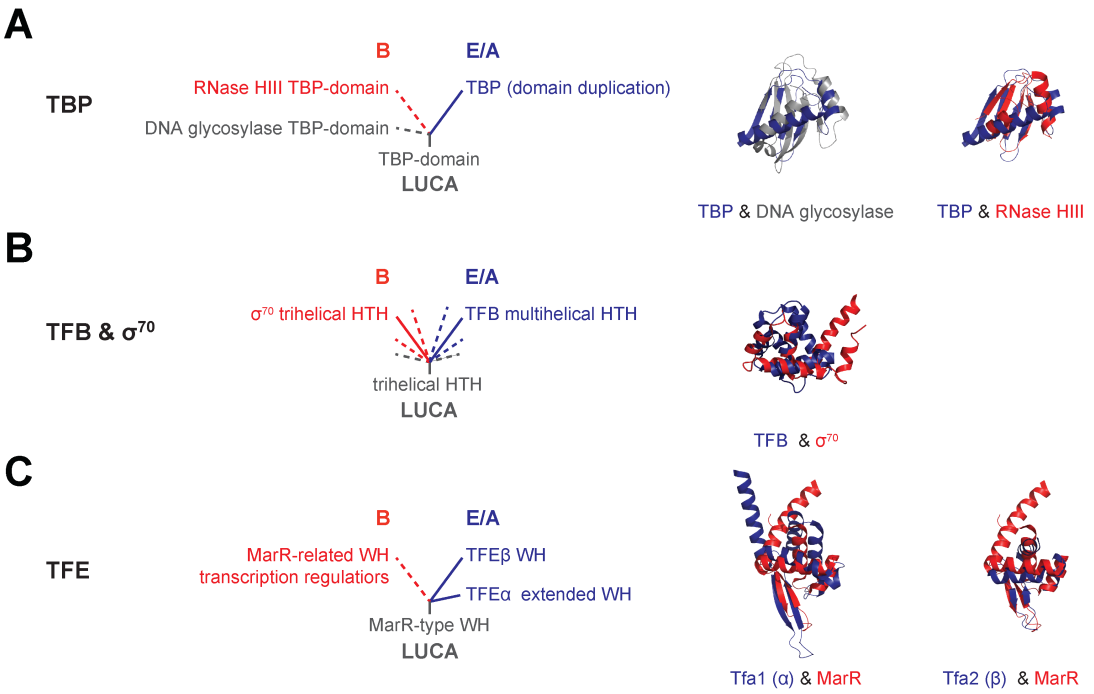

Figure 3: Evolutionary origins of basal transcription factors.

Schematic phylogenetic relationships of structural motifs present in bacterial (B, red branches) and eukaryotic/archaeal (E/A, blue branches) basal transcription factors based on (1; 10; 14). Structurally related proteins with different functions are indicated by dotted lines. Grey branches indicate presence of the respective structural motif in universally conserved proteins. **(A)** Proteins containing TBP-domains were already present in LUCA, but the domain duplication observed in TBP is unique to this protein. **(B)** In TFB/TFIIB two multihelical HTH motifs are found that are probably derived from trihelical HTH motifs (1). Trihelical HTH motifs are found in bacteria-specific and archaea-eukaryotic lineages as well as in several universally conserved proteins (1). **(C)** MarR-related WH motifs are abundant in bacterial transcription regulators and distantly related to the WH motifs present in the two subunits of archaeal TFE (TFEα and TFEβ) and eukaryotic TFIIE (termed Tfa1 and

Tfa2 in yeast). The overlay of the structural motifs is based on the following PDB entries: *Sulfolobus acidocaldarius* TBP (PDB id: 1MP9), *E. coli* DNA glycosylase II (1MPG), *Geobacillus stearothermophilus* RNase HIII (2D0A), *Saccharomyces cerevisiae* Tfa1 and Tfa2 (5FYW), *E. coli* MarR (1JGS), *Pyrococcus woesei* TFB (1D3U), *E. coli* σ^{70} (4YG2).

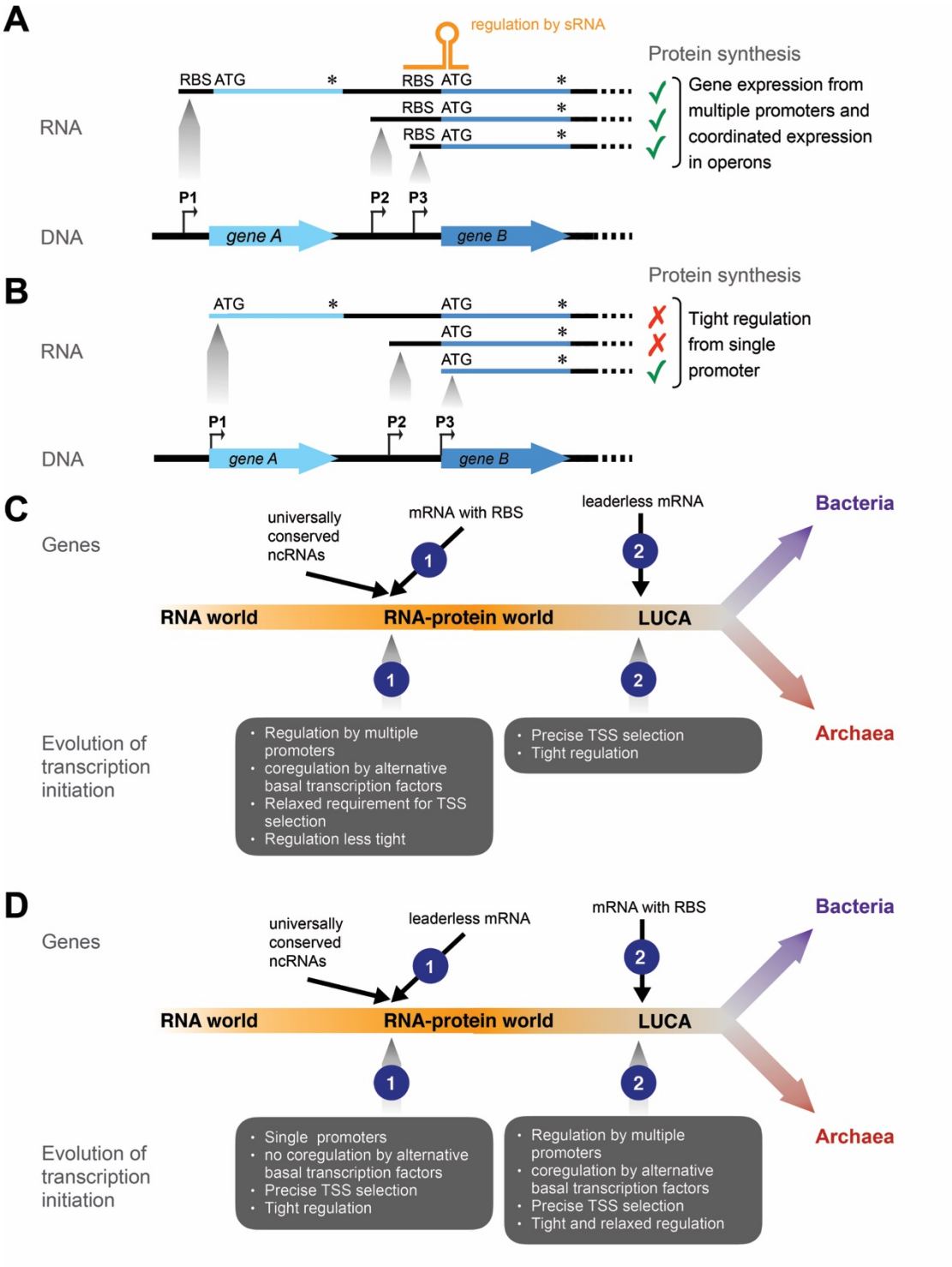

Figure 4: Interdependency of transcription and translation initiation.

A) and B) Consequences of 5'-UTR containing (A) and leaderless mRNAs (B) on the regulation and coordination of gene expression. While RBS-dependent translation initiation allows for protein expression from mRNAs synthesised from different TSSs

(A), leaderless translation initiation requires transcription to initiate from a single TSS (B). **C) and D)** The order of evolution of translation initiation on leaderless mRNA and RBS-dependent translation initiation has major implications on the evolution of transcription initiation concerning TSS selection and regulatory mechanisms. (C) An early evolution of RBS-dependent translation initiation (1) would allow for gene expression to be directed from multiple promoters with a relaxed requirement for precise TSS selection. It would facilitate also the coordinated expression of genes such as those coding for the components of heteroligomeric complexes. The evolution of translation initiation from leaderless RNAs (2) would add a mechanism that enables tight gene regulation and minimisation of gene expression arising from spurious transcription read-through. (D) A late appearance of RBS-dependent translation initiation (2) with translation initiation exclusively from leaderless mRNAs at early stages of evolution of cellular life (1) would require the early evolution of basal transcription factors enabling precise TSS selection and the use of single promoters.