BEAN and HABAS: Polyphyletic insertions in RNAP that point to deep evolutionary divergence of bacteria

Claudia Alvarez-Carreño[1]*, Angela T. Huynh[2], Christine Orengo[1], Loren Dean Williams[2,3]*

[1]Department of Structural and Molecular Biology, University College London, London, United Kingdom

[2]School of Chemistry and Biochemistry, Georgia Institute of Technology, 901 Atlantic Dr, Atlanta, GA 30332, USA

[3]NASA Center for the Origin of Life, Georgia Institute of Technology, Atlanta, GA 30332-0400, USA

* Corresponding authors:

Claudia Alvarez-Carreño

Email: c.carreno@ucl.ac.uk

* Loren Dean Williams

Email: loren.williams@chemistry.gatech.edu

Abstract

The β and β' subunits of the RNA polymerase (RNAP) are large proteins with complex multi-domain architectures that include several insertional domains. Here, we analyze the multi-domain organizations of bacterial RNAP-β and RNAP-β' using sequence, experimentally determined structures and AlphaFold structure predictions. We observe that bacterial lineage-specific domains in RNAP-β belong to a group of domains that we call BEAN (Broadly Embedded ANnex) and that in RNAP-β', bacterial lineage-specific domains are HAmmerhead/BArrel-Sandwich Hybrid (HABAS) domains. The BEAN domain has a characteristic three-dimensional structure composed of two square bracket-like elements that are antiparallel relative to each other. The HABAS domain contains a four-stranded open β-sheet with a GD-box-like motif in one of the β-strands and the adjoining loop. The BEAN domain is identified not only in the bacterial RNAP-β', but also in the archaeal version of universal ribosomal protein L10. The HABAS domain is observed as an insertional domain in several metabolic proteins. The phylogenetic distributions of bacterial lineage-specific insertional domains of β and β' subunits of RNAP follow the Tree of Life. The presence of insertional domains can help establish a relative timeline of events in the evolution of a protein because insertion is inferred to post-date the base domain. We discuss mechanisms that might account for the discovery of homologous insertional domains in non-equivalent locations in bacteria and archaea.

Introduction

Transcription is the Central Dogma process in which the RNA polymerase (RNAP) transcribes DNA into RNA (Hurwitz, et al. 1961). mRNA is then translated into protein in the ribosome. RNAP contains five subunits called α1, α2, β, β', and ϖ. Each of these RNAP subunits has bacterial archaeal, and eukaryotic orthologs. The β and β' subunits of RNAP, the focus of this work, are large proteins with complex multi-domain architectures. Two double-Ψ-β-barrel (DΨBB) domains, one on the β subunit and the other on the β' subunit, compose the catalytic core of RNAP (Castillo, et al. 1999; Iyer, et al. 2003). Although the β and β' subunits of RNAP are universally distributed, their multi-domain architectures vary significantly between archaea and bacteria, and among bacteria. Here, we use sequences and structures to retrace the deep evolution of RNAP-β and RNAP-β' in bacteria.

RNAP-β and RNAP-β' contain universal sequence motifs and domains, which are conserved in all three branches of life (Archaea, Bacteria, and Eukarya) (Sweetser, et al. 1987; Jokerst, et al. 1989; Lane and Darst 2010a, b) and also contain bacteria- and archaea -specific domains (Lane and Darst 2010a, b). Archaea-specific domains are conserved in Eukaryotes (Supplementary Figures 2-3). Some bacteria-specific domains are observed only in certain bacterial lineages (Borukhov, et al. 1991; Severinov, et al. 1992; Iyer, et al. 2004; Lane and Darst 2010b, a; Huang, et al. 2015; Qayyum, et al. 2024). Domains of RNAP-β and RNAP-β' that occur in all archaea but not in bacteria are called here "a-specific". Domains that occur in all bacteria but not in archaea are called here "b-specific". Domains that occur in some bacterial lineages but not others are called here "b/lineage-specific".

Proteins most commonly acquire new domains by terminal addition (Weiner, et al. 2006; Marsh and Teichmann 2010), generating tandem multidomain architectures. Yet, both RNAP-β and RNAP-β' contain domains acquired by internal insertion within a preexisting base domain. Insertional domains are less frequent than terminally-added domains (Manriquez-Sandoval and Fried 2022), and can be particularly useful in establishing relative timelines of evolutionary events; an insertional domain is inferred to have been added more recently than the domain that hosts it.

Most of the a-, b- and b/lineage-specific domains of RNAP-β and RNAP-β' are insertional. In multiple sequence alignments (MSAs), a-, b- and b/lineage-specific domains are observed as "blocks" (Vishwanath, et al. 2004). In MSAs of RNAP-β and RNA -β', universal domains are broken by these blocks (Supplementary Figure 1).

The results here suggest that bacterial lineages independently acquired insertional domains at multiple distinct locations of RNAP-β and RNAP-β' genes. We observe a b/lineage-specific insertional domain with variable position in RNAP-β. We call this domain BEAN (Broadly Embedded ANnex). We observe a b/lineage-specific insertional domain with variable position in RNAP-β'. We call this domain HABAS (HAmmerhead/BArrel-Sandwich Hybrid). This acquisition occurred before or soon after the last universal common ancestor.

The diversity in position and type of insertional domain in two ancient components of the transcription apparatus (RNAP-β and -β'), suggests these proteins, during early evolution, had high propensity to acquire and retain insertions. The results allow us to establish a classification system for bacterial RNAP-β and RNAP-β' subunits, based on type, location and chronology of their insertional domains.

Results

Domain organizations of RNAP β and β'.

The MSAs of RNAP-β and RNAP-β' display block structures indicating universal as well as a-specific, b-specific and b/lineage-specific elements (Figure 1a). We analyzed the RNAP-β and RNAP-β' domain organizations using orthologous sequences from a subsample of a reference set of evenly sampled bacterial genomes (Zhu, et al. 2019). The subsample used here includes representatives from all known major bacterial species and has been adapted from (Moody, et al. 2022). The blocks within the MSA were annotated using CATH (Sillitoe, et al. 2021) (Tables 1 and 2). The location and type of insertion were verified using experimentally determined structures (Berman, et al. 2000) and AlphaFold structure predictions (Jumper, et al. 2021; Varadi, et al. 2022). We observe small insertions (< 50 residues) that lack sequence similarity to each other or to domain entries in three classification databases: CATH (Sillitoe, et al. 2021), ECOD (Schaeffer, et al. 2017) and SCOPe (Chandonia, et al. 2017). These insertions are omitted from the analysis here.

Our analysis is facilitated by our naming system for RNAP domains. In this scheme, the subunit is indicated by β or β', followed by a hyphen and the letter "u" to indicate universal conservation, or "a" to indicate a-specific, or "b" to indicate b-specific. The domains (D) are numbered in order of appearance in the sequence. For example, the N-terminal RNAP β-subunit domain, which is universal, is called β-uD1.

RNAP-β multi-domain architectures.

RNAP-β has six universal domains that are conserved in archaeal and bacterial orthologs (universal core, β-uD1, β-uD2,… β-uD6; Figure 1). Universal insertional domains were identified in three different locations of RNAP-β: β-uD2 is insertional in β-uD1; β-uD3 is insertional in β-uD1; and β-uD5 (HABAS domain) is insertional in β-uD4 (DΨBB domain).

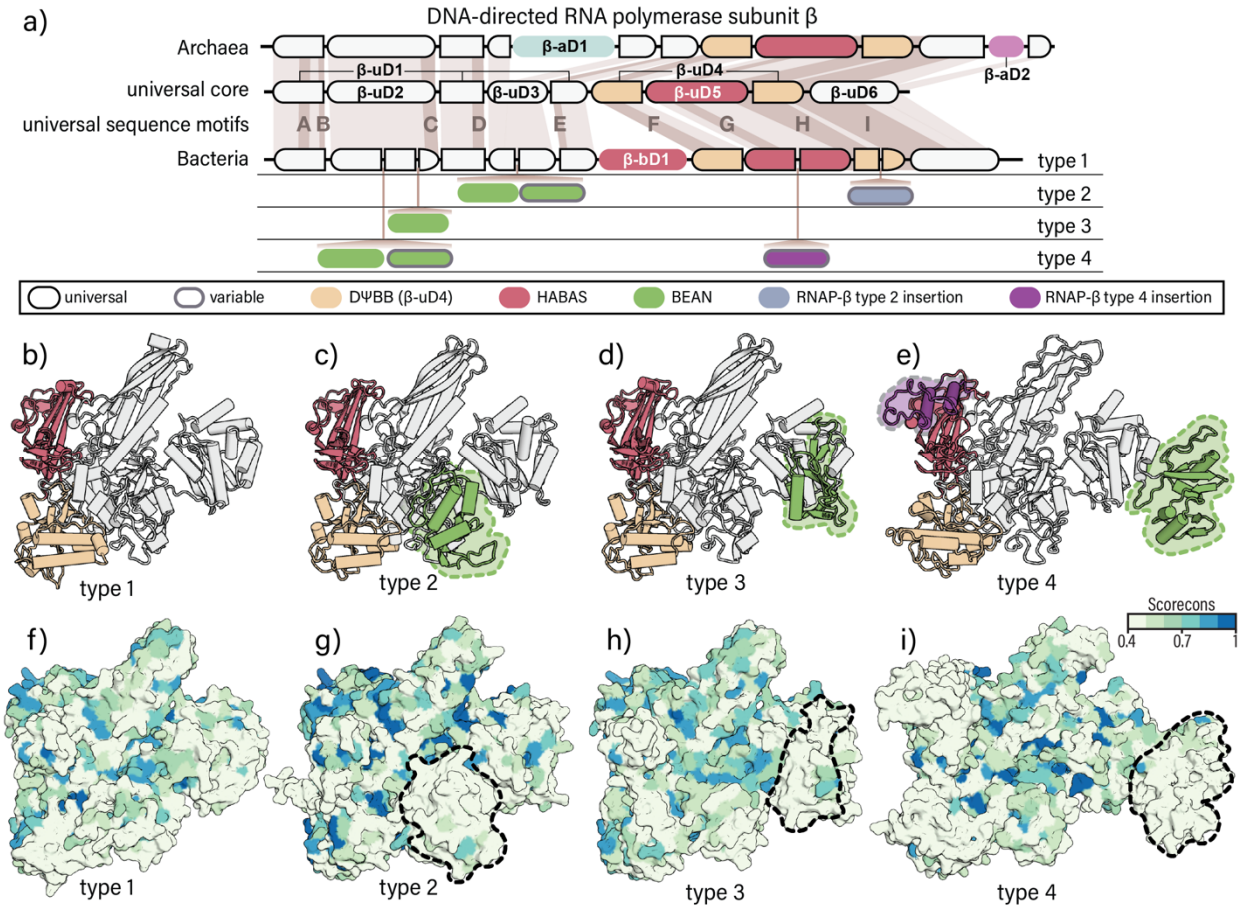

Figure 1. Multi-domain organizations of RNAP-β. (a) Domain organization of RNAP-β. First row: domains of archaeal RNAP-β. Second row: universally shared domains between archaeal and bacterial orthologs and universal sequence motifs described in (Sweetser, et al. 1987). Third row: domains of bacterial type 1 RNAP-β. Fourth row: location of bacterial type 2 insertions. Fifth row: location of bacterial type 3 insertions. Sixth row: location of bacterial type 4 insertions (b to e) Three-dimensional structure of the bacterial RNAP-β. Domains are coloured as in (a). (b) type 1 (AlphaFold DB: AF-A2BT61-F1), (c) type 2 (AlphaFold DB: AF-A9B6J3-F1), (d) type 3 (AlphaFold DB: AF-Q8ETY8-F1), and (e) type 4 (PDB: 4IGC, chain C). (f to i) Bacterial RNAP-β coloured by residue conservation. The conservation score ranges from 0 (not conserved) to 1 (highly conserved). Valdar01 score was calculated on the multiple sequence alignment of sequence representatives for each type using Scorecons. For clarity, N- and C-terminal residues that extend beyond the shared core of bacteria were masked.

Archaea and bacteria have additional domains at different locations in RNAP-β (Figure 1a). In archaea, β-aD1 is inserted within β-uD3 and β-aD2 is inserted within β-uD6. In bacteria, β-bD1 is inserted between β-uD2 and β-uD4.

RNAP-β contains b/lineage-specific domains at multiple positions (Figure 1a). These domains are inserted at five distinct sites of RNAP-β: in the i) central region of β-uD2; ii) C-terminal half of β-uD2; iii) N-terminal half of β-uD3; iv) C-terminal half of β-uD5; and v) C-terminal half of β-uD4. The b- and b/lineage-specific domains in RNAP-β are less conserved in sequence than universal domains (Figure 1f to 1i).

The b/lineage-specific insertions in β-uD2 and in β-uD3 share sequence and structure similarity (Figures 2c and 2d), suggesting common ancestry. We call these domains BEAN (Broadly Embedded ANnex).

Sites of BEAN insertion within RNAP-β define b/lineage-specific RNAP-β architecture (Figure 1a). Differential b/lineage insertion in RNAP-β allows us to distinguish four types of RNAP-β. Type 1 RNAP-β lacks b/lineage-specific insertions (Figure 1a and 1b). Type 2 RNAP-β has a BEAN domain inserted within β-uD3 (Figure 1a and 1c). Type 3 RNAP-β has a BEAN domain inserted within β-uD2 (Figure 1a and 1d). Type 4 RNAP-β has a BEAN domain inserted within β-uD2 (Figure 1a and 1e). Certain RNAP-β proteins display a type 4 multidomain architecture with additional insertions; these are referred to as type 4*.

The BEAN domain

The BEAN domain has a characteristic three-dimensional structure composed of two square bracket-like elements that are anti-parallel relative to each other (Figure 2d). Each bracket-like element is formed by an α-helix and two β-strands. The orientation between consecutive secondary elements is 90° within each bracket. The first bracket is formed by α1⊥β2⊥β3 and the second bracket by β3⊥β4⊥α5. The BEAN domain maps to CATH superfamily 3.90.105.10.

We have identified BEAN domains in bacterial and archaeal proteins other than RNAP-β (Figure 2a to 2d). Using sequence and structure similarities, we find the BEAN domain in bacterial RNAP-β'; the archaeal version, but not the bacterial version, of universal ribosomal protein L10 (uL10a); molybdenum cofactor biosynthesis protein MoeA; Orn/Lys/Arg decarboxylases; a putative ferredoxin; and in one uncharacterized protein.

A cluster analysis of sequences based on BLASTP P-values shows that b/lineage-specific BEAN domains in RNAP-β are more similar to each other than to BEAN domains in other

proteins (Figure 2a). Additionally, MSAs of b/lineage-specific BEAN domains in RNAP-β show a conserved TGD/E (threonine, glycine, aspartic/glutamic acid) sequence motif (Figure 2d). This motif is absent from other BEAN domains. Thus, BEAN domains in RNAP-β appear to share more recent ancestry with each other than with BEAN domains of other proteins.

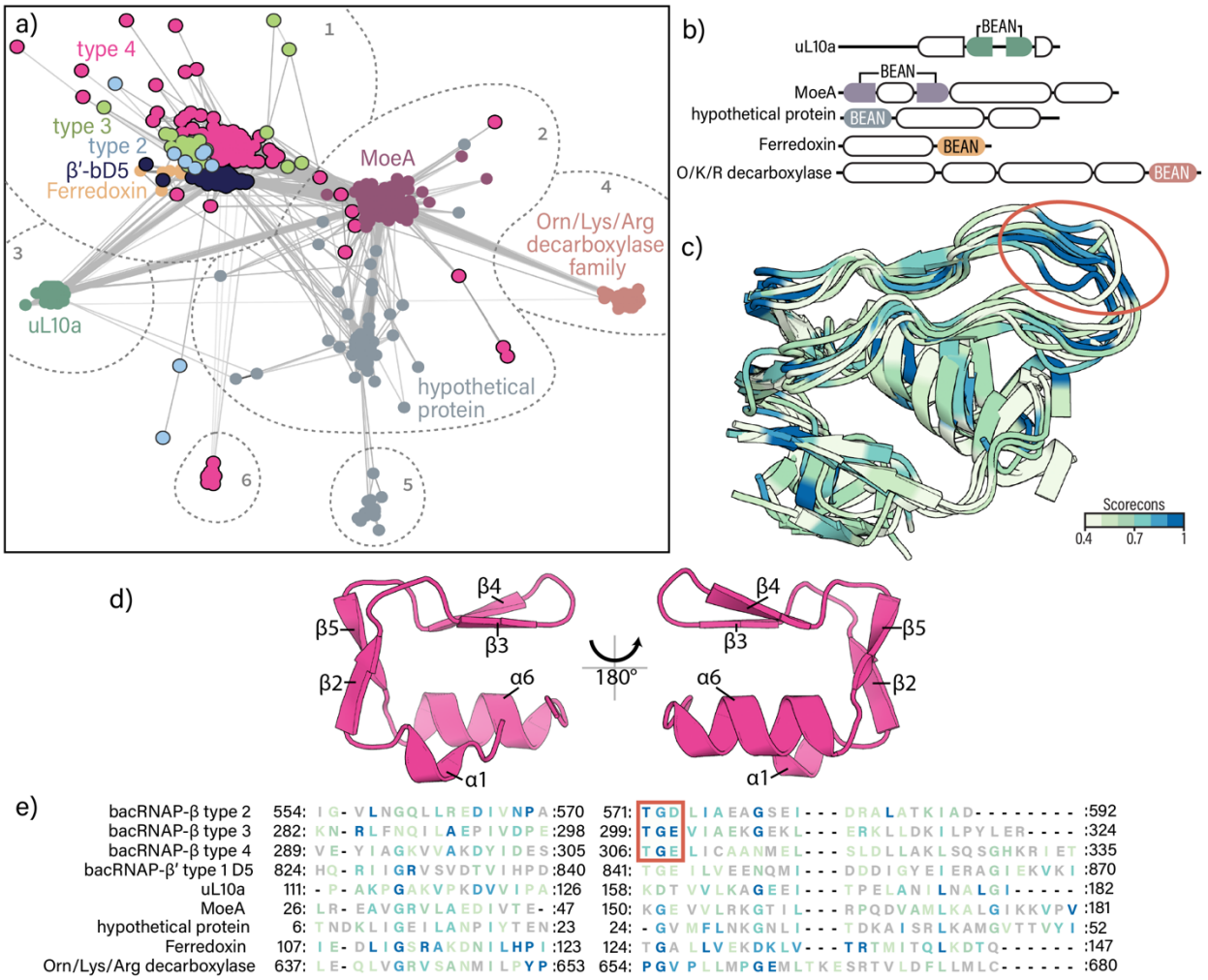

Figure 2. The BEAN domain. (a) BEAN domain sequences clustered by similarity at a P-value threshold of 1×10^{-11}. Dotted lines encircle sequence clusters that remain at a P-value of 1×10^{-15}. (b) Multi-domain organization of representative proteins containing BEAN domains. (c) Structure superimposition of BEAN domains from representative predicted structures. Structures are colored by conservation. (d) Structure-derived multiple sequence alignment. Above 0.5, residues are colored by conservation score. Below 0.4 residues are shown in gray.

RNAP-β' multi-domain architectures.

Bacterial RNAP-β' domains map to two distinct proteins in archaea (Rpo1N and Rpo1C). Rpo1N and Rpo1C assemble to form a complete RNAP-β' in archaea and eukaryotes, which is also called Rpo1 (Korkhin, et al. 2009). The naming system for domains in the archaeal RNAP-β' follows a continuous order from Rpo1N to Rpo1C. Domains β'-uD1 to β'-uD4 are common to bacterial RNAP-β' and archaeal Rpo1N; and domains β'-uD5 and β'-uD6 are common to bacterial RNAP-β' and archaeal Rpo1C (Figure 3a). The DΨBB domain of RNAP-β' is β'-uD2.

We identified small a-specific insertions in Rpo1N in β'-uD1 and β'-uD4, and a-specific insertional domains within β'-uD2 and β'-uD3. We also identified three a-specific domains in Rpo1C: β'-aD3, β'-aD4 and β'-aD5. β'-aD3 is an N-terminal addition to Rpo1C; β'-aD4 is inserted into β'-uD5, and β'-aD5 is inserted into β'-aD4.

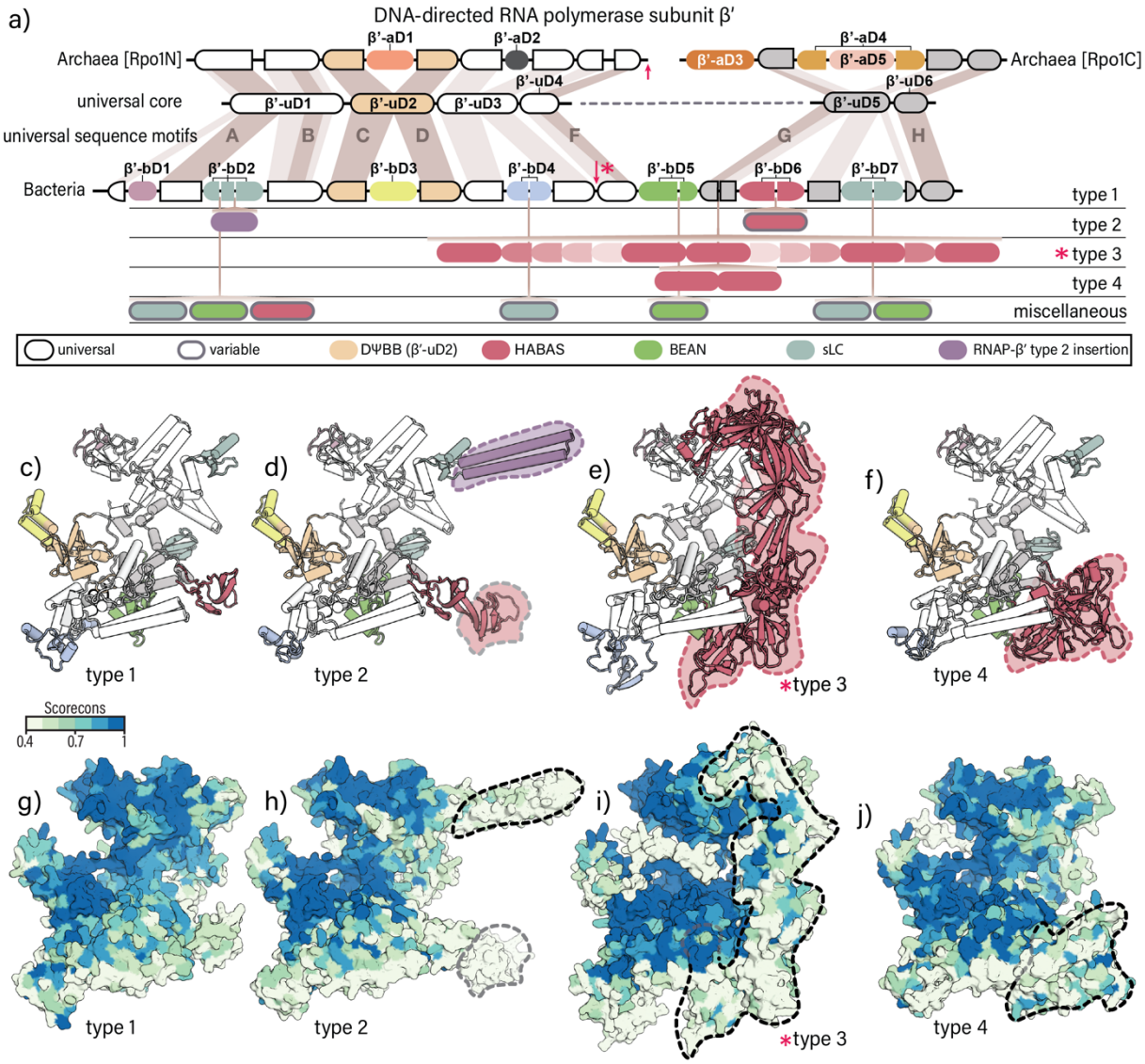

Figure 3. Multi-domain organization of RNAP-β'. (a) Multi-domain organization diagrams of RNAP-β' in archaea and bacteria. First row: domains of archaeal orthologs. Second row: universally shared domains between archaeal and bacterial orthologs and universal sequence motifs described in (Jokerst, et al. 1989). Third row: domain organization of the bacterial type 1 RNAP-β'. Fourth row: location of bacterial type 2 insertions. Fifth row: location of bacterial type 3 insertions. Sixth row: location of bacterial type 4 insertions. Red arrows: sites of split of RNAP-β' into sub-subunits. (b to e) Three-dimensional structures of bacterial RNAP-β' colored by domains as in (a). (b) type 1 (AlphaFold DB: AF-Q0AUH3-F1), (c) type 2 (AlphaFold DB: AF-Q3Z8V3-F1), (d) type 3, N-terminal fragment (AlphaFold DB: AF-A2BT59-F1) and C-terminal fragment (AlphaFold DB: AF-A2BT60-F1), and (e) type 4 (AlphaFold DB: AF-A7IKQ1-F1). *Site of truncation of bacterial type 3 RNAP-β. (f to i) Bacterial RNAP-β' coloured by residue conservation. The conservation score ranges from 0 (not conserved) to 1 (highly conserved). Valdar01 score was calculated on the multiple sequence alignment of sequence representatives for each type using Scorecons. For clarity, N- and C-terminal residues that extend beyond the shared core of bacteria were masked.

Bacterial RNAP-β' is composed of six universal RNAP-β' domains and seven b-specific domains. Domains β'-bD1, β'-bD2, β'-bD3, β'-bD4, β'-bD6 and β'-bD7 are insertional. β'-bD5 is a BEAN domain and β'-bD6 is a HABAS domain. β'-bD3 is inserted into the DΨBB domain of the β' subunit and shows no sequence or structure similarity to the archaeal DΨBB domain

insertion (β'-aD1). β'-bD2 and β'-bD7 are homologous (CATH superfamily 3.30.60.280). Here, we call these domains sLC (small Left Claw).

The b/lineage-specific insertions in RNAP-β' were observed in seven distinct locations (Figure 2a): i) in the first half of β'-bD2; ii) near the middle of β'-bD2; iii) in the middle of β'-uD2; iv) in the middle of β'-uD3; v) in the second half of β'-bD5; vi) in the first half of β'-uD5; and vii) in the second half of β'-uD5. Based on the presence and location of b/lineage-specific insertions, we have defined four types of bacterial RNAP-β' subunits (Figure 2a). The simplest multi-domain organization of RNAP-β' in bacteria, called here type 1, has no b/lineage-specific insertions. Type 2, 3 and 4 bacterial RNAP-β' have b/lineage-specific insertional domains. Most of these b/lineage-specific insertions are HABAS domains, which share sequence and structural similarity (Figures 2d to 2f).

The location and number of b/lineage-specific insertions define the type of bacterial RNAP-β'. Bacterial type 2 RNAP-β' has an insertion in β'-bD2 and a variable HABAS domain insertion in RNAP-β'bD6. Bacterial type 3 and 4 RNAP-β' have HABAS domain insertions in RNAP-β'uD5; these insertions collocate in sequence but differ in number; bacterial type 4 RNAP-β' has one HABAS domain insertion whereas bacterial type 3 RNAP-β' has nine HABAS domain insertions. The HABAS domains in type 3 RNAP-β' are inserted on other HABAS domains (Chlenov, et al. 2005; Qayyum, et al. 2024) and exchange secondary structural elements (are domain swapped) forming a complex interdigitated structure (Figure 2e). Additional RNAP-β multi-domain organizations contain variable numbers of insertions in β'-bD2, β'-bD4, β'-bD5 and β'-bD7. We also identified a collection of unique RNAP-β' multidomain architectures containing variable numbers of HABAS, BEAN and sLC domain insertions.

Bacterial type 3 RNAP-β' is composed of two polypeptide chains. The N-terminal sub-subunit (RNAP-β'_{BacN}) ends with β'-uD3; and the C-terminal sub-subunit (RNAP-β'_{BacC}) starts at β'uD4 (Figure 2a). RNAP-β'_{BacN} and RNAP-β'_{BacC} assemble to form a complete RNAP-β (Qayyum, et al. 2024).

The HABAS domain

The b/lineage-specific insertions in RNAP subunit β' are HABAS domains (CATH superfamily 2.40.50.100). The HABAS domain is also observed in RNAP-β in archaea and bacteria as β-uD5, and as a b-specific domain (β-bD1). Sequence similarity reveals additional proteins containing HABAS domains, including proteins involved in metabolism (phosphatidylserine decarboxylase proenzyme, biotin carboxyl carrier protein, ion-translocating oxidoreductase complex subunit C, Cytochrome f), transport (major facilitator superfamily transporter protein, Resistance-Nodulation-Division family transporter protein) and genetic information processing (Transcription termination/antitermination protein NusG) (Figure 5).

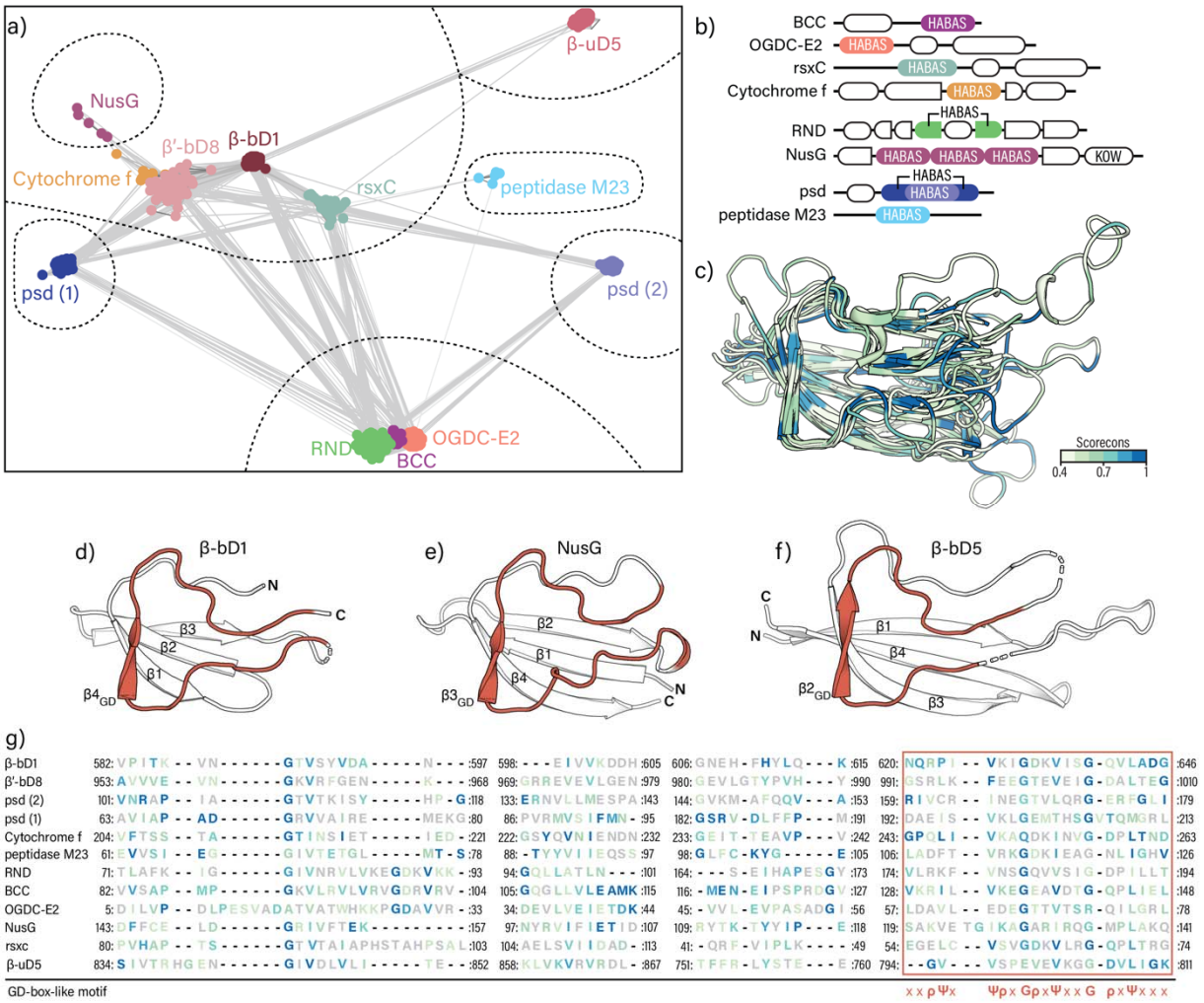

Figure 4. The HABAS domain. (a) HABAS domains clustered by sequences similarity at a P-value threshold of 5×10^{-10}. Dotted lines encircle sequence clusters that remain at a P-value of 1×10^{-12}. (b) Multi-domain organization of representative proteins containing HABAS domains. (c) Structure superimposition of HABAS domains from representative predicted structures. (d-e) Structure representation of HABAS domains. GD-box-like motif highlighted in dark orange. (d) HABAS domain $\beta4_{GD}$ topology

(AF-A2BPU4-F1). Residues 616-625 are masked. (e) HABAS domain $\beta3_{GD}$ topology (AF-P77611-F1). (f) HABAS domain $\beta2_{GD}$ topology (AF-C5CGE4-F1). Residues 711-725 and 747-775. (g) Structure-derived multiple sequence alignment of β-bD1 (A2BT61), β'-D8 (Q0AUH3), psd (A0L627), cytochrome f (A2BPU4), peptidase M23 (A0A0E3QUJ6), RND (A0A1T5M7J5), BCC (O59021), OGDC-E2 (P0AFG6), NusG (C5CGE4), rsxC (P77611) and β-uD5 (P11512).

The HABAS domain contains a four-stranded open β-sheet with a conserved sequence motif in one of the β-strands and the adjoining loop. The motif contains glycine (G), aliphatic (Ψ), and polar (ρ) amino acids as follows: ΨxΨρxGρxΨxxGρxΨxx. We call this motif the GD-box-like motif because it is similar but not identical to the GD-box sequence motif ΨxΨxxGρxΨxΨ (Alva, et al. 2009). We found three distinct topologies of secondary structural elements in HABAS domains (Figure 5). These topologies are related by circular permutation. We distinguish and name these topological variants by the locations of their GD-box-like motifs. The most frequently observed topology has a GD-box-like motif in strand β4 ($\beta4_{GD}$). β-bD1 has a $\beta4_{GD}$ topology (Figure 4d). β-uD5 has a $\beta2_{GD}$ topology (Figure 4e), and a HABAS domain of NusG has a $\beta3_{GD}$ topology (Figure 4f).

Phylogenetic distribution of RNAP-β and RNAP-β' types

The phylogenetic distribution of RNAP-β and RNAP-β' types as we define them here follows the deep divergence between Gracilicutes and Terrabacteria (Coleman, et al. 2021; Witwinowski, et al. 2022). We identified type 4 RNAP-β and type 4 RNAP-β' in most Gracilicutes; by contrast, we identified all types of RNAP-β and RNAP-β' in Terrabacteria.

A given bacterial lineage tends to have a single type of RNAP-β. Bacteria in the CMS group (Cyanobacteria and related bacteria) contain type 1 RNAP-β; Armatimonadetes contain type 4* RNAP-β; Actinobacteria contain type 1 RNAP-β; and Chloroflexi contain type 2 RNAP-β.

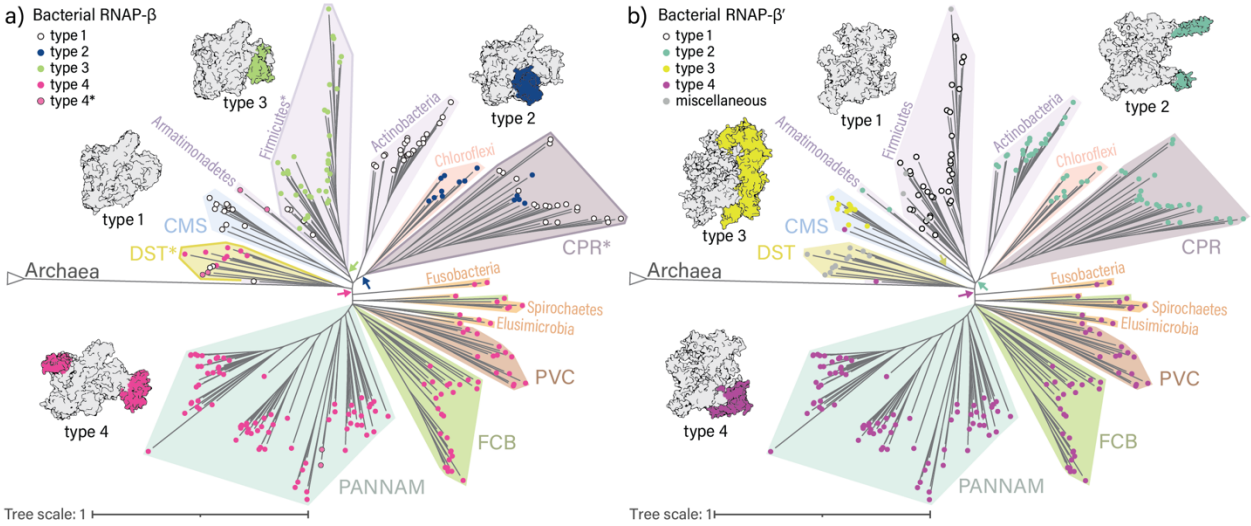

Figure 5. **Phylogenetic distribution of the different multi-domain organizations of RNAP-β and RNAP-β' in bacteria**. Domain organization types are shown in different colors. The tree of Bacteria was adapted from (Moody, et al., 2022). (a) Distribution of RNAP-β types in bacteria. Phylogenetic groups with a scattered distribution of type 1 RNAP-β are indicated by a darker outline. (b) Distribution of RNAP-β types in bacteria. CMS: Cyanobacteria, Margulisbacteria, Melainabacteria; CPR: Candidatus Phyla Radiation; DST: Deinococcus-Thermus, Synergistes, Thermotogae, Bipolaricaulota, Caldiserica, Coprothermobacterota; FCB: Fibrobacteres, Chlorobi, Bacteroides, Gemmatimonadetes, Candidatus Cloacimonetes, division KSB1, Eisenbacteria, Candidatus Fermentibacteria, Firestonebacteria, Candidatus Glassbacteria, Ignavibacteria, Kryptonia, Marinimicrobia, Raymondbacteria, Stahlbacteria, Zixibacteria; PANNAM: Bdellovibrio, Dependentia, Proteobacteria, Aquificae, Myxococcota, Nitrospinae, Nitrospirae, Acidobacteria, Chrysiogenetes, Deferribacteres, Schekmanbacteria and Thermodesulfobacteria; PVC: Planctomycetes, Verrucomicrobia, Chlamydiae, Kiritimatiellaeota, Lentisphaerae, Candidatus Desantisbacteria, Candidatus Omnitrophica.

However, some bacterial lineages contain a scattered distribution of type 1 RNAP-β. Bacteria in the DST group (Deinococcus-Thermus, Synergistetes, Thermotogae and related bacteria) generally contain type 4, but some DST bacteria contain type 1 RNAP-β. Firmicutes generally contain type 3, but some Firmicutes contain type 1 RNAP-β. Bacteria in the CPR group (Candidate Phyla Radiation) type 2 or type 1 RNAP-β.

We calculated the gene tree of RNAP-β using sites that are conserved in all bacteria to test whether the scattered distribution of type 1 RNAP can be attributed to HGT. We compared the gene tree of RNAP-β to a consensus tree of bacteria calculated previously using 27 vertically inherited genes (Moody, et al. 2022). Our phylogenetic analysis shows that, regardless of the type of insertion, the tree of RNAP-β follows the topology of the tree of bacteria (Figure 6 and Supplementary Figure 4). This correspondence suggests that type 1 RNAP-β, which has no b/lineage-specific BEAN insertions, evolved independently in different bacterial lineages. The ancestors of distinct bacterial groups may have convergently lost BEAN insertions of the protein coding gene for RNAP-β.

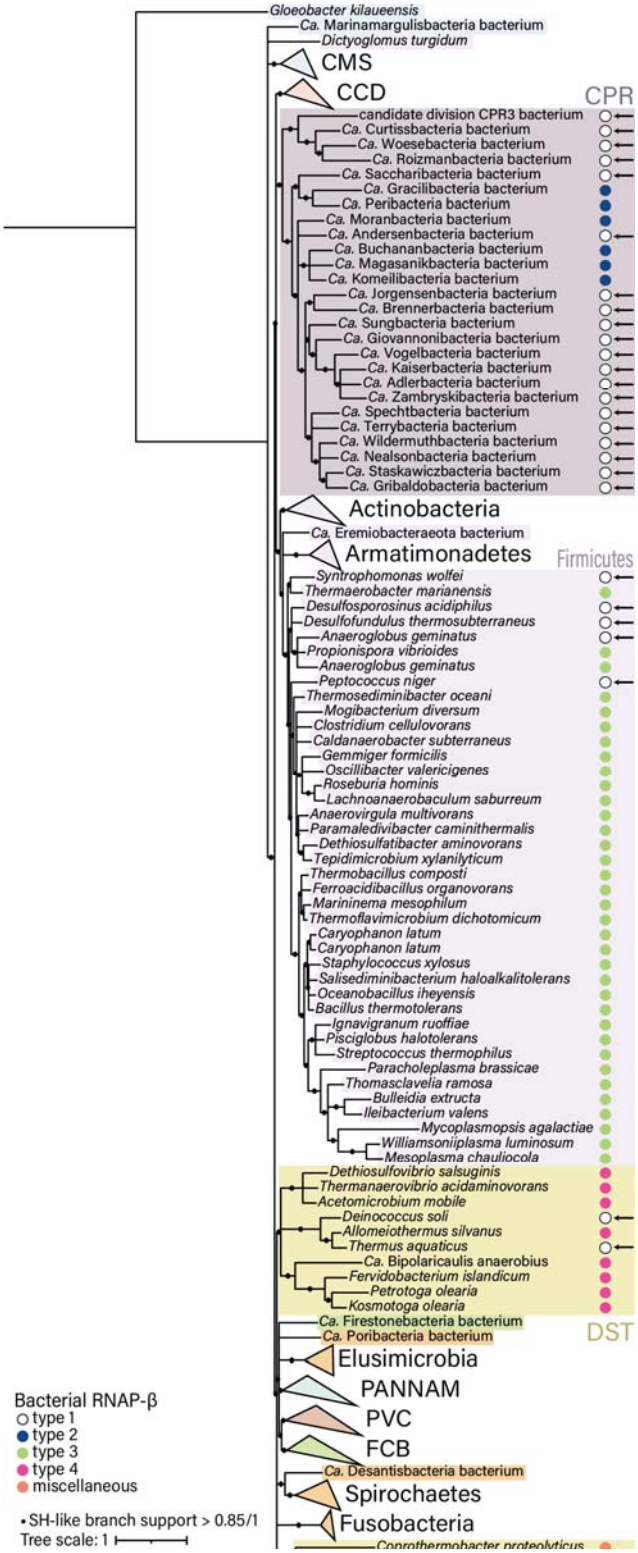

Figure 6. Maximum likelihood tree of RNAP-β. The ML tree (inferred with model Q.yeast +G+I) of RNAP-β was calculated using positions conserved in all bacteria. Arrows: type 1 RNAP-β.

Discussion

RNAP is a multi-subunit protein complex that contains RNAP-β and RNAP-β' subunits. RNAP-β and RNAP-β' are found in RNAPs in archaea, bacteria, eukarya, and nucleocytoplasmic large DNA viruses (Iyer, et al. 2001). Here we report that RNAP-β and RNAP-β' each contain distinctive homologous insertional domains with variable positions. These insertional domains are informative about evolutionary history but also pose difficult questions about evolutionary mechanisms.

We call these domains BEAN and HABAS. The BEAN domain has a characteristic three-dimensional structure composed of two square bracket-like elements that are antiparallel relative to each other. Each bracket-like element is formed by an α-helix and two β-strands. The orientation between consecutive secondary elements is 90° within each bracket. The HABAS domain contains a four-stranded open β-sheet with a GD-box-like motif in one of the β -strands and the adjoining loop.

We identify a b-specific BEAN insertion in RNAP-β' and b/lineage specific BEAN insertions in bacterial RNAP-β. In addition, we observe BEAN insertions in other apparently unrelated proteins. These proteins are the archaeal version of universal ribosomal protein L10 (uL10a); molybdenum cofactor biosynthesis protein MoeA; Orn/Lys/Arg decarboxylases; a putative ferredoxin; and one uncharacterized protein. We identify a universal HABAS and a b-specific HABAS in RNAP-β and b/lineage specific HABAS insertions in RNAP-β'. In addition, we observe HABAS insertions in proteins involved in metabolism (phosphatidylserine decarboxylase proenzyme, biotin carboxyl carrier protein, ion-translocating oxidoreductase complex subunit C, Cytochrome f), transport (major facilitator superfamily transporter protein, Resistance-Nodulation-Division family transporter protein) and genetic information processing (Transcription termination/antitermination protein NusG).

The genes for RNAP-β and -β' are adjacent to each other in the genomes of virtually all bacteria and most are in the neighbourhood of the genes that encode for NusG, universal ribosomal proteins uL1 and uL11 (Supplementary Tables 1-2). Similarly, in most archaea, uL10a is in the neighbourhood of the genes that encode for NusG (Spt4/5), uL1 and uL11 (Supplementary Table 3). The local gene neighbourhood may have increased the chances of

acquisition of BEAN and HABAS domains by insertion by a genetic mechanism that remains unclear.

Insertional domains in the evolution of translation and transcription

Our sequence similarity searches indicate that HABAS and BEAN are insertional in several proteins. BEAN insertional domains appear at different locations in archaea and in bacteria in RNAP subunits β and β' subunits and in archaea in ribosomal protein uL10. Transcription and translation are two of the central biological processes responsible for the encoding and synthesis of proteins. The patterns of insertion of HABAS and BEAN domains in universal and ancient proteins pose provocative questions regarding the timing and order of events during the early evolution of life. We do not understand mechanisms by which BEAN was universally inserted in bacterial RNAP but not archaeal RNAP and was universally inserted in archaeal uL10 but not bacterial uL10. Apparently, BEAN was mobile at or before LUCA and was acquired by each of these proteins during that realm. The combined data suggest that acquisition of BEAN domains in RNAP-β and archaeal uL10 and HABAS domains in RNAP-β' occurred in ancestral lineages and that the descendants retained these insertions. Our observations support deep roots, predating LUCA, of the bacterial and archaeal split and even of various bacterial lineages.

b/lineage-specific HABAS and BEAN domain insertions are polyphyletic

We grouped the β and β' subunits of RNAP within bacteria based on the number and type of insertional domains. Insertions occur in distinct locations in different types of RNAP-β and RNAP-β'. Insertions interrupt universal domains, and thus, post-date the establishment of the basic multi-domain architectures of RNAP-β and RNAP-β'. BEAN insertions specify the type of RNAP-β. HABAS domains specify the type of RNAP-β'.

Acquisition of BEAN domains may have been independent of acquisition of HABAS domains. We observe a mismatch in the distributions of RNAP-β and RNAP-β' (Figure 5). For instance, bacteria from the DST group with type 4 RNAP-β have RNAP-β' multi-domain architectures that fall outside of our classification, but Armatimonadetes with type 4* RNAP-β have type 2 RNAP-β', and Gracilicutes with type 4 RNAP-β have type 4 RNAP-β'. Below we discuss the phylogenetic distributions of RNAP-β and RNAP-β' separately.

The distribution of RNAP-β types suggests that BEAN domain insertions impacted the trajectories of at least three populations of ancestral bacterial lineages: i) the ancestor of Firmicutes, which acquired type 2 insertions; ii) the ancestor of Chloroflexi and CPR bacteria, which acquired type 3 insertions; and iii) the ancestor of Gracilicutes, which acquired type 4 insertions.

The presence of type 4 RNAP-β insertions in two Terrabacteria groups, DST (Deinococcus-Thermus, Synergistes, Thermotogae) bacteria and Armatimonadetes, is nevertheless, puzzling. Three scenarios could explain this distribution: i) vertical inheritance; ii) HGT from Gracilicutes; and iii) convergence in the insertion site. In the case of Armatimonadetes, a more recent common ancestor with Gracilicutes is not supported; bacterial phylogenies calculated from different gene markers consistently put Armatimonadetes within Terrabacteria (Coleman, et al. 2021; Megrian, et al. 2022; Witwinowski, et al. 2022). Neither is HGT from Gracilicutes is supported by our gene tree of RNAP-β (Figure 6 and Supplementary Figure 4), which places type 4 RNAP-β from Armatimonadates species as a sister group of type 3 RNAP-β from Firmicutes. In the case of RNAP-β type 4 in the DST group, our phylogenetic tree does not rule out a HGT from Gracilicutes.

HABAS insertions in RNAP-β' appear to have occurred in three ancestral populations: i) the ancestor of Armatimonadetes, Actinobacteria, Chloroflexi and CPR bacteria; ii) an ancestor within the CMS group; and iii) the ancestor of Gracilicutes. Independent insertions of HABAS, BEAN and sLC domains appear to have affected DTS bacteria and Firmicutes.

Block structures of MSAs are not exclusive to RNAP-β and -β' and have been described for universal components of the translation system [ribosomal proteins (Vishwanath, et al. 2004), and aminoacyl tRNA synthetases (Alvarez-Carreño, et al. 2023)]. But block differences are generally observed between archaea and bacteria. Our combined results suggests that BEAN and HABAS domain insertions capture deep evolutionary events that affected the genes for RNAP-β and -β' in ancestral bacterial groups. Here, we observe a process that reshaped the multidomain architecture of bacterial and archaeal orthologs and then continued during early bacteria evolution.

Methods

Identification of RNAP-β and β' subunits in bacteria and archaea

The sequences of RNAP subunits β and β' from *Sulfolobus acidocaldarius* (UniProt IDs: P11513 and P11512) and *Bacillus subtitlis* (UniProt IDs: P37870 and P37871) were searched in a set of archaeal and bacterial proteomes derived from (Moody, et al. 2022) using phmmer from the HMMER3 suite (Eddy 2011). Sequences above threshold (E-value < 1×10^{-10}) were retrieved and aligned. Multiple sequence alignments (MSAs) were generated with the einsi option from MAFFT v7 (Katoh and Standley 2013).

Domain annotation of RNAP-β and β' and classification

The MSAs of RNAP subunits β and β' were converted each into a sequence profile and compared to CATH_S40, ECOD_F70 and SCOPe95 with HH-search (Steinegger, Meier, et al. 2019) on the MPI Bioinformatics Toolkit (Zimmermann, et al. 2018). CATH_S40 contains CATH domains clustered at 40% sequence identity; ECOD_F70 contains ECOD domains clustered at 70% sequence identity; and SCOPe95 contains domain sequences clustered at 95% sequence identity. The block patterns on the MSAs were used as reference to classify the multi-domain organization types of bacterial RNAP-β and RNAP-β' proteins (Figure S1). For each RNAP-β and RNAP-β' type, a representative was selected for structure analysis. All representatives have a experimentally determined structure in the PDB (Berman, et al. 2000) or a predicted structure in AlphaFold DB (Varadi, et al. 2022). Per-residue confidence score (pLDDT) and predicted aligned error (PAE) of the structure predictions are shown in Supplementary Figures 6-16. Sequences with unique insertion patterns were annotated individually.

Identification of HABAS and BEAN domains homologs

A-, b- and B/lineage-specific insertions were trimmed according to the blocks in the MSAs. Profiles were calculated for each trimmed MSA with hhmake from the hh-suite (Steinegger, Meier, et al. 2019), considering columns with fewer than 50% gaps match states.

The MSAs were converted to profile Hidden Markov Models using the HH-suite version 3.3.0 (Steinegger, Meier, et al. 2019). The profiles were searched against the BFD database (Steinegger, Mirdita, et al. 2019) of clustered genome and metagenome sequences using HHblits (3 iterations, Probability > 60). Significant matches (minimum probability: 60, minimum

coverage with master sequence 80%) were retrieved and clustered with CLANS (Gabler, et al. 2020) by all-against-all BLASTP sequence similarity (P-value 1×10-20). Groups with at least 30 homologs in the cluster map were extracted; realigned with MAFFT einsi (Katoh and Standley 2013); and converted to HMM profiles with HMMER version 3.3.2 (Eddy 2011). The HMM profiles were searched with phmmer in the same set of archaeal and bacterial proteomes (Moody, et al. 2022) that was used RNAP-β and RNAP-β' identification.

Structure based MSAs of HABAS and BEAN domains were calculated with MATRAS (Kawabata 2003).

Maximum likelihood tree of RNAP-β.

The MSA of bacteria RNAP-β was trimmed with trimAl v1.4 to remove positions with more than 10% gaps. The ML tree was calculated with PhyML (Guindon, et al. 2010) on the Montpellier Bioinformatics Platform. Model selection was determined with SMS (Lefort, et al. 2017).

Data availability

Sequence alignments and structure predictions associated with this manuscript have been deposited in the FigShare repository.

Acknowledgements

The authors thank Anton S. Petrov and Finn Werner for insightful discussions. This work was funded by the National Aeronautics and Space Administration grant 80NSSC18K1139. Claudia Alvarez-Carreño's research was supported by the Royal Society Newton International Fellowship.

Tables

Table 1. Multi-domain architecture of RNAP β subunit in representatives from Archaea and Bacteria

Short name	CATH ID	Archaea	Bacterial type 1	Bacterial type 2	Bacterial type 3	Bacterial type 4
		P11513 (AAY80073.1)	A2BT61 (ABM70972.1)	A9B6J3 (ABX02896.1)	Q8ETY8 (BAC12068.1)	P0A8V4 (AAC76961.1)
		Sulfolobus acidocaldarius	*Prochlorococcus marinus*	*Herpetosiphon aurantiacus*	*Oceanobacillus iheyensis*	*Escherichia coli*
β-uD1	3.90.1100.10	1-166; 338-476	1-132; 315-442; 524-581	1-159; 364-491; 662-719	1-142; 406-535; 618-675	1-152; 450-577; 675-714
β-uD2	3.90.1110.10	167-337	133-314	160-363	143-279; 371-407	153-226;343-449
Type 3 insertion	3.90.105 .10	-	-	-	286-370	-
Type 4 insertion (1)	3.90.105.10	-	-	-	-	240-338
β-uD3	2.30.150.10	471-476; 569-711	443-523	492-510; 602-661	536-617	578-674
β-aD1	3.90.1070.20	477-568	-	-	-	-
Type 2 insertion	3.90.105.10	-	-	511-601	-	-
β-bD1	2.40.50.100	-	582-650	720-788	676-752	715-790
β-uD4	2.40.270.10	712-747; 869-994	651-687; 819-886,887-913,916-923,925-958	789-825; 957-1081	753-789; 921-1045	791-827; 1059-1237
β-uD5	2.40.50.150	748-868	688-818	826-956	790-920	828-939; 1038-1058
Type 4 insertion (2)	6.10.140.1670	-	-	-	-	940-1037
β-uD6	3.90.1800.10	995-1130	959-1077	1082-1220	1046-1180	1238-1400

Table 2. Multi-domain architecture of RNAP β' subunit in representatives from Archaea and Bacteria

Short name	CATH	Archaea	Bacterial type 1	Bacterial type 2	Bacterial type 3 (N-ter)	Bacterial type 3 (C-ter)	Bacterial type 4
		P11512 (AAY80072.1)	Q0AUH3 (ABI69631.1)	Q3Z8V3 (AAW40096.1)	A2BT59 (ABM70970.1)	A2BT60 (ABM70971.1)	A7IKQ1 (ABS68594.1)
		Sulfolobus acidocaldarius	*Syntrophomonas wolfei subsp. wolfei*	*Dehalococcoides mccartyi*	*Prochlorococcus marinus*	*Prochlorococcus marinus*	*Xanthobacter autotrophicus*
β'-uD1	4.10.960.120	1-148; 176-317	10-64; 93-138; 178-343	1-56; 84-130; 237-402	1-70; 98-144; 192-357	-	1-67; 95-141; 182-348
β'-bD1	3.90.820.30	-	65-91	57-83	71-97	-	68-94
a-insertion (1)	-	149-175	-	-	-	-	-
β'-bD2	3.30.60.280	-	139-177	131-140; 211-236	145-191	-	142-181
Type 2 insertion	-	-	-	141-210	-	-	-
β'-uD2	2.40.40.20	318-346; 414-494	344-371; 414-495	403-430; 475-554	358-385; 428-509	-	349-376; 419-500
β'-bD3	1.10.40.90	-	372-413	431-474	386-427	-	477-418
β'-aD1	3.30.1490.180	347-413	-	-	-	-	-
β'-uD3	1.10.274.100	495-560; 599-644	496-512; 579-640	555-571; 618-679	510-526; 618-635	-	501-517; 572-637
β'-bD4	-	-	513-578	572-617	527-617	1-84	518-571
β'-aD2	2.60.40.2940	561-598	-	-	-	-	-
β'-uD4	1.10.132.30	645-832	641-798	680-837	-	85-244	638-806
		1054-1093					
a-insertion (2)	-	756-779	-	-	-	-	-
β'-bD5	3.90.105.10	-	799-879	838-917	-	245-318	807-888
β'-uD5	1.10.1790.20	-	880-949; 1013-1059; 1104-1113	918-987; 1118-1164; 1209-1282	-	319-369; 1009-1027; 1092-1154;	889-939; 1124-1143; 1208-

						1199-1208	1254; 1299-1308
β′-bD6	2.40.50.100	-	950-1012	988-1117	-	1028-1091	1144-1207
Type 3 insertion	2.40.50.100	-	-	-	-	370-1008*	-
Type 4 insertion	1.10.132.30	-	-	-	-	-	940-1123*
β′-bD7	3.30.60.280	-	1060-1103	1165-1208	-	1155-1198	1255-1298
β′-uD6	1.10.150.390	-	1114-1181	1219-1283		1209-1252	1309-1375

References

Alva V, Dunin-Horkawicz S, Habeck M, Coles M, Lupas AN. 2009. The GD box: a widespread noncontiguous supersecondary structural element. Protein Science 18:1961-1966.

Alvarez-Carreño C, Arciniega M, Ribas de Pouplana L, Petrov AS, Hernández-González A, Dimas-Torres JU, Valencia-Sánchez MI, Williams LD, Torres-Larios A. 2023. Common evolutionary origins of the bacterial glycyl tRNA synthetase and alanyl tRNA synthetase. Protein Science 33:e4844.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. 2000. The Protein Data Bank. Nucleic Acids Research 28:235-242.

Borukhov S, Severinov K, Kashlev M, Lebedev A, Bass I, Rowland GC, Lim PP, Glass RE, Nikiforov V, Goldfarb A. 1991. Mapping of trypsin cleavage and antibody-binding sites and delineation of a dispensable domain in the beta subunit of Escherichia coli RNA polymerase. J Biol Chem 266:23921-23926.

Castillo RM, Mizuguchi K, Dhanaraj V, Albert A, Blundell TL, Murzin AG. 1999. A six-stranded double-psi beta barrel is shared by several protein superfamilies. Structure 7:227-236.

Chandonia JM, Fox NK, Brenner SE. 2017. SCOPe: Manual Curation and Artifact Removal in the Structural Classification of Proteins - extended Database. J Mol Biol 429:348-355.

Chlenov M, Masuda S, Murakami KS, Nikiforov V, Darst SA, Mustaev A. 2005. Structure and function of lineage-specific sequence insertions in the bacterial RNA polymerase beta' subunit. J Mol Biol 353:138-154.

Coleman GA, Davín AA, Mahendrarajah TA, Szánthó LL, Spang A, Hugenholtz P, Szöllősi GJ, Williams TA. 2021. A rooted phylogeny resolves early bacterial evolution. Science 372.

Eddy SR. 2011. Accelerated Profile HMM Searches. PLoS Comput Biol 7:e1002195.

Gabler F, Nam SZ, Till S, Mirdita M, Steinegger M, Söding J, Lupas AN, Alva V. 2020. Protein Sequence Analysis Using the MPI Bioinformatics Toolkit. Curr Protoc Bioinformatics 72:e108.

Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Systematic Biology 59:307-321.

Huang Y, Kendall T, Forsythe ES, Dorantes-Acosta A, Li S, Caballero-Pérez J, Chen X, Arteaga-Vázquez M, Beilstein MA, Mosher RA. 2015. Ancient Origin and Recent Innovations of RNA Polymerase IV and V. Mol Biol Evol 32:1788-1799.

Hurwitz J, Furth JJ, Anders M, Ortiz PJ, August JT. 1961. The enzymatic incorporation of ribonucleotides into RNA and the role of DNA. Cold Spring Harb Symp Quant Biol 26:91-100.

Iyer LM, Aravind L, Koonin EV. 2001. Common origin of four diverse families of large eukaryotic DNA viruses. Journal of Virology 75:11720-11734.

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.

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 Structural Biology 3:1.

Jokerst RS, Weeks JR, Zehring WA, Greenleaf AL. 1989. Analysis of the gene encoding the largest subunit of RNA polymerase II in Drosophila. Molecular and General Genetics 215:266-275.

Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, et al. 2021. Highly accurate protein structure prediction with AlphaFold. Nature.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30:772-780.

Kawabata T. 2003. MATRAS: A program for protein 3D structure comparison. Nucleic Acids Research 31:3367-3369.

Korkhin Y, Unligil UM, Littlefield O, Nelson PJ, Stuart DI, Sigler PB, Bell SD, Abrescia NG. 2009. Evolution of complex RNA polymerases: the complete archaeal RNA polymerase structure. PLoS Biology 7:e1000102.

Lane WJ, Darst SA. 2010a. Molecular evolution of multisubunit RNA polymerases: sequence analysis. J Mol Biol 395:671-685.

Lane WJ, Darst SA. 2010b. Molecular evolution of multisubunit RNA polymerases: structural analysis. J Mol Biol 395:686-704.

Lefort V, Longueville JE, Gascuel O. 2017. SMS: Smart Model Selection in PhyML. Mol Biol Evol 34:2422-2424.

Manriquez-Sandoval E, Fried SD. 2022. DomainMapper: Accurate domain structure annotation including those with non-contiguous topologies. Protein Science 31:e4465.

Marsh JA, Teichmann SA. 2010. How do proteins gain new domains? Genome Biology 11:126.

Megrian D, Taib N, Jaffe AL, Banfield JF, Gribaldo S. 2022. Ancient origin and constrained evolution of the division and cell wall gene cluster in Bacteria. Nat Microbiol 7:2114-2127.

Moody ERR, Mahendrarajah TA, Dombrowski N, Clark JW, Petitjean C, Offre P, Szöllősi GJ, Spang A, Williams TA. 2022. An estimate of the deepest branches of the tree of life from ancient vertically evolving genes. Elife 11.

Qayyum MZ, Imashimizu M, Leanca M, Vishwakarma RK, Riaz-Bradley A, Yuzenkova Y, Murakami KS. 2024. Structure and function of the Si3 insertion integrated into the trigger loop/helix of cyanobacterial RNA polymerase. Proc Natl Acad Sci U S A 121:e2311480121.

Schaeffer RD, Liao Y, Cheng H, Grishin NV. 2017. ECOD: new developments in the evolutionary classification of domains. Nucleic Acids Research 45:D296-d302.

Severinov K, Mustaev A, Kashlev M, Borukhov S, Nikiforov V, Goldfarb A. 1992. Dissection of the beta subunit in the Escherichia coli RNA polymerase into domains by proteolytic cleavage. J Biol Chem 267:12813-12819.

Sillitoe I, Bordin N, Dawson N, Waman VP, Ashford P, Scholes HM, Pang CSM, Woodridge L, Rauer C, Sen N, et al. 2021. CATH: increased structural coverage of functional space. Nucleic Acids Research 49:D266-d273.

Steinegger M, Meier M, Mirdita M, Vöhringer H, Haunsberger SJ, Söding J. 2019. HH-suite3 for fast remote homology detection and deep protein annotation. BMC Bioinformatics 20:473.

Steinegger M, Mirdita M, Söding J. 2019. Protein-level assembly increases protein sequence recovery from metagenomic samples manyfold. Nature Methods 16:603-606.

Sweetser D, Nonet M, Young RA. 1987. Prokaryotic and eukaryotic RNA polymerases have homologous core subunits. Proc Natl Acad Sci U S A 84:1192-1196.

Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G, Yuan D, Stroe O, Wood G, Laydon A, et al. 2022. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Research 50:D439-d444.

Vishwanath P, Favaretto P, Hartman H, Mohr SC, Smith TF. 2004. Ribosomal protein-sequence block structure suggests complex prokaryotic evolution with implications for the origin of eukaryotes. Mol Phylogenet Evol 33:615-625.

Weiner J, 3rd, Beaussart F, Bornberg-Bauer E. 2006. Domain deletions and substitutions in the modular protein evolution. FEBS Journal 273:2037-2047.

Witwinowski J, Sartori-Rupp A, Taib N, Pende N, Tham TN, Poppleton D, Ghigo JM, Beloin C, Gribaldo S. 2022. An ancient divide in outer membrane tethering systems in bacteria suggests a mechanism for the diderm-to-monoderm transition. Nat Microbiol 7:411-422.

Zhu Q, Mai U, Pfeiffer W, Janssen S, Asnicar F, Sanders JG, Belda-Ferre P, Al-Ghalith GA, Kopylova E, McDonald D, et al. 2019. Phylogenomics of 10,575 genomes reveals evolutionary proximity between domains Bacteria and Archaea. Nat Commun 10:5477.

Zimmermann L, Stephens A, Nam SZ, Rau D, Kübler J, Lozajic M, Gabler F, Söding J, Lupas AN, Alva V. 2018. A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core. J Mol Biol 430:2237-2243.