

RNA polymerases in strict endosymbiont bacteria with extreme genome reduction show distinct erosions that could result in limited and differential promoters' recognition.

Cynthia Paola Rangel-Chávez¹, Edgardo Galán-Vásquez², Azucena Pescador-Tapia¹, Luis Delaye³ and Agustino Martínez-Antonio^{1*}

¹Biological Engineering Laboratory. Genetic Engineering Department. Center for Research and Advanced Studies of the National Polytechnic Institute. Km. 9.6 Libramiento Norte Carr. Irapuato-León 36824, Irapuato Gto, México.

²Departamento de Ingeniería de Sistemas Computacionales y Automatización, Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, Universidad Nacional Autónoma de México, Circuito Escolar 3000, C.U., Coyoacán, 04510 Ciudad de México, CDMX

¹Evolutionary Genomics Laboratory. Genetic Engineering Department. Center for Research and Advanced Studies of the National Polytechnic Institute. Km. 9.6 Libramiento Norte Carr. Irapuato-León 36824, Irapuato Gto, México.

*Corresponding author: agustino.martinez@cinvestav.mx

Keywords

RNA polymerase, Endosymbionts, Genome reduction, Transcription, Sigma factor

Short title:

RNA polymerases erosions in strict endosymbionts bacteria

Abstract

Strict endosymbiont bacteria with high degree of genome reduction retain smaller proteins and, in certain cases, lack complete functional domains compared to their free-living counterparts. Until now, the mechanisms underlying these genetic reductions are not well understood. However, it is thought that, in order to compensate for gene reduction, somehow hosts take over those vital functions that endosymbionts cannot perform. In the present study, the conservation of RNA polymerases, the essential machinery for gene expression, is analysed in bacteria with extreme genome reductions. For this purpose, comparative genomics, phylogenetic analysis and three-dimensional models of RNA polymerase subunits were done over four lineages of endosymbiotic proteobacteria with the smallest genomes known to date. Amino acids under positive selection in the α subunit and loss of motifs in other subunits of RNA polymerase were observed. According to three-dimensional models, sites under positive selection might compensate the loss of motifs in α subunit. In addition, variations in the σ subunit were identified, some of them already studied in *E. coli* as a result of random mutagenesis. Amino acid changes in RNA polymerase suggest a possible modification in the binding specificity of the canonical -10 box (TATAAT) in some of these organisms. Furthermore, the β -flap helix domain is absent in some *Hodgkinia* strains, as observed in RNA pol II of Archaea, thus lacking the capacity to bind to the -35

box. Here, we propose several RNA polymerases models for endosymbiont bacteria with extremely reduced genomes. Evidence suggests that RNA polymerases of each endosymbiont bacteria follow a unique evolutionary path, without necessarily following the same path as a lineage, this is probably influenced by the intimate interactions sustained with other endosymbionts and its hosts.

Introduction

Until 2006, there was a scientific consensus indicating that the minimum quantity of genes to support life would be around 500, even for tiny organisms like bacteria. This view changed soon after that date, when genomes of strict endosymbiotic bacteria began to be reported, and it was found that several genomes contained less than 500 genes, with some extreme cases reporting on genomes as small as 200 genes [1]. Transcriptome analysis of *Buchnera*, an obligate endosymbiont that has lost the ability to persist without its host, provided evidence indicating that, unlike *E. coli*, its genome has a limited ability to respond to environment fluctuations [2], which possibly indicates that gene expression in obligate endosymbiont bacteria is somehow stable and active at basal levels. According to this, the density of promoter-like signals, characteristic of free-living bacteria is not present in organisms exhibiting extreme genome-reduction [3].

DNA transcription is an essential molecular process through which organisms decode genetic information to cellular functions [4]. RNA polymerase (RNAP) is an enzyme responsible for the transcription of DNA to RNA, which consists of a multi-subunit protein complex present in all living organisms from bacteria to eukaryotes [5]. In bacteria, RNAP is responsible for the synthesis of all types of RNAs, including messenger RNAs, ribosomal RNAs, transference RNAs, and small RNAs. In free-living bacteria, RNAP consists of six subunits ($\alpha_2\beta\beta'\omega\sigma$), encoded by five different genes (the two α_2 subunits are encoded by the same gene), the specific assembly of these proteins constitutes a holoenzyme with a molecular mass around 400 kDa. Previous studies on *Escherichia coli* found that the ω subunit is

not essential for RNAP activity [6, 7], it was later found that ω subunit is absent in all the endosymbionts here analysed.

The rest of the subunits considered as essential core components of RNAP (α_2 , β and β' subunits) are well conserved in bacteria [5, 8]. This RNAP core is catalytically active (transcribing DNA to RNA) but unable to initiate DNA transcription by itself. For transcription to be initiated, the core enzyme must bind to an additional subunit called sigma factor (σ). The σ is responsible for recognizing and binding to specific sites of transcription initiation, formally known as promoters; binding of sigma factors enables RNAP to initiate transcription [8]. Once transcription is initiated, and after the synthesis of a short fragment of RNA, σ is released and the core protein complex continues transcribing until it reaches a transcription terminator. Since σ is responsible of binding and discriminating among gene's promoters, it is common to find different types (variants) of σ , in such a way that we can classify them in two evolutionary families: one of them called the σ^{54} family, that normally has one single gene per genome; and the σ^{70} family, that usually has several genes/members per genome (from 1 to 60 in the largest bacterial genomes). One member of the σ^{70} family is also known as the “housekeeping σ factor”, which is an essential gene present in all bacteria [9]. There are seven sigma factors encoded in the *E. coli* genome, six of them correspond to the σ^{70} family (σ^{70} , σ^{38} , σ^{32} , σ^{28} , σ^{24} and σ^{19}), and the remaining one corresponds to σ^{54} . In a previous work we analysed around 8500 promoters annotated in the *E. coli* genome and proposed the architecture of transcription units and the consensus sequences of the promoters for each sigma factor [10].

Endosymbionts live inside other organisms in a relationship estimated to be several million years old [11]. Strict endosymbiotic bacteria have lost most of their genes and large fragments of proteins due to a global process known as genome reduction [10, 6]. This dramatic process can be illustrated with the comparison of the *E. coli* genome, with approximately 4.5 thousand genes, and their relative endosymbiotic bacterium *Candidatus Carsonella ruddii*, which has only retained around 150 genes [12,

13]. Almost all genes found in *C. ruddii* are considerably shorter than their free-living orthologues. Furthermore, genes encoding proteins with multiple domains in strict endosymbionts commonly have lost some of these domains, which in some cases have been identified as essential for its activity in free-living bacteria [14, 15]. Endosymbiotic bacteria with extreme genome reduction a single copy of σ^{70} is maintained and none σ^{54} and in a previous work, we identified that obligate endosymbionts have lost all the transcription factors that interact with the promoters and RNAPs to activate or inhibit the transcription process [16]. Additionally, partial loss of the subunits α and σ have been reported previously in *Hodgkinia* sp. and *Carsonella ruddii*, respectively [15, 17], both subunits are essential for the proper operation of RNAP in free-living bacteria.

The phenomenon of gene erosion, including the transcriptional machinery in endosymbionts with extremely reduced genomes, raises questions concerning how gene transcription takes place in these bacteria. For instance: if the loss of RNAP subunits is happening in the same way in these bacteria, how is that some domains are dispensable in obligate endosymbionts while being essential in free-living bacteria? And, if we have elements to speculate, how the transcription process has been adapted in endosymbiont bacteria with extremely reduced. Here, we address some of these questions from the perspective of comparative genomics. We investigated the evolution of RNAP subunits in four bacterial groups of endosymbionts exhibiting extreme genome-reduction, all of them belonging to the proteobacteria phylum.

Materials and Methods

Dataset selection

The genetic information for 37 genomes of endosymbionts belonging to four lineages of proteobacteria with characteristics of extreme genome reduction were obtained from the NCBI database (<http://www.ncbi.nlm.nih.gov/>). These are *Candidatus Hodgkinia cicadicola*, that belongs to α -proteobacteria; *Candidatus Tremblaya phenacola*, *Candidatus Tremblaya princeps*, and *Candidatus*

Nasuia deltocephalinicola, which belong to β -proteobacteria; and *Candidatus Carsonella ruddii*, that belongs to Gamma-proteobacteria. The genomes of these bacteria have been sequenced completely. Further characteristics of these genomes are provided in the supplementary material S1 Table. From these, in order to study the most reduced RNAP proteins (see below), we selected 16 bacterial endosymbionts genomes displayed the largest modifications at the protein domain level (Table 1).

Table 1. Genomic characteristics of the analysed endosymbionts compared to *E. coli*.

Organism name	Lifestyle	Genome size*	Genes*	%GC	Reference
<i>Escherichia coli</i> k-12 MG1655	Free-living	4.63897	4405	50.8	[13]
α proteobacteria					
<i>Candidatus Hodgkinia cicadicola</i> TETUND1	Endosymbiont	0.133698	121	46.8	[18]
<i>Candidatus Hodgkinia cicadicola</i> TETUND2	Endosymbiont	0.14057	140	46.2	[18]
<i>Candidatus Hodgkinia cicadicola</i> Tetuln	Endosymbiont	0.14057	140	45.4	[18]
<i>Candidatus Hodgkinia cicadicola</i> Dsem	Endosymbiont	0.143795	169	58.4	[19]
β proteobacteria					

<i>Candidatus Tremblaya phenacola</i> PAVE	Endosymbiont	0.1715	183	42.1	[20]
<i>Candidatus Tremblaya princeps</i> PCIT	Endosymbiont	0.138927	172	58.8	[21]
<i>Candidatus Tremblaya princeps</i> PCVAL	Endosymbiont	0.138931	174	58.8	[22]
<i>Candidatus Nasuia deltocephalinicola</i> str. NAS-ALF	Endosymbiont	0.112091	137	17.1	[23]
<i>Candidatus Nasuia deltocephalinicola</i> str. PUNC	Endosymbiont	0.112031	137	16.6	[24]
Gamma proteobacteria					
<i>Candidatus Carsonella ruddii</i> HT isolate Thao2000	Endosymbiont	0.157543	166	14.6	[14]
<i>Candidatus Carsonella ruddii</i> PV	Endosymbiont	0.159662	175	16.6	[25]
<i>Candidatus Carsonella ruddii</i> PC isolate NHV	Endosymbiont	0.159923	174	15.6	[14]

<i>Candidatus Carsonella ruddii</i> CS isolate Thao2000	Endosymbiont	0.162504	181	14.0	[14]
<i>Candidatus Carsonella ruddii</i> CE isolate Thao2000	Endosymbiont	0.162589	182	14.0	[14]
<i>Candidatus Carsonella ruddii</i> HC isolate Thao2000	Endosymbiont	0.166163	182	14.2	[14]
<i>Candidatus Carsonella ruddii</i> DC	Endosymbiont	0.174014	196	17.6	[26]

*Genome size in Mega base pairs, **Open reading frame.

Comparative analysis

The amino acid sequences for each RNAP subunit were aligned against their corresponding *E. coli* ortholog. For simplicity, the amino acid positions will hereinafter refer to its location in the corresponding *E. coli* protein or subunit, accompanied by the abbreviation “ECO”. For each endosymbiont, we used the protein sequences annotated for each RNAP subunit and made a comparative analysis to define domains and amino acid sites. The following tools were used for this purpose: Pfam database [31], the proteins superfamily classification database [32], the NCBI's conserved domains search [33] and, additionally, we did a bibliographic research to gather relevant information regarding the sites and regions of each of the RNAP subunits and their functionality.

Structural Analysis

To analyse protein structures, we first recreated 3D structural models for each RNAP subunit using the I-TASSER server [34]. For a given sequence, TASSER generates 3D models by collecting high-score structural templates from PDB, with full-length atomic models constructed by iterative template-based fragment assembly simulations. The resulting structures were compared with the crystal structure of the holoenzyme RNAP of ECO (4YLP) [35]. Graphic representations for each structure were prepared with the PyMOL Molecular Graphic System software version 1.3 [36].

To investigate possible dimer association of α subunit monomers of *Hodgkinia* TETUND2 (this strain has lost a fragment of the α subunit amino terminal and it is involved in dimer formation), we used the ClusPro v.2.0 tool [37-40]. This is an automatic protein docking tool based on CAPRI (Critical Assessment of Predicted Interactions) [38, 41]. Three models of α dimer subunits with different cluster sizes were obtained through ClusPro and were used to map sites under putative positive selection and for evaluating free-energy changes in protein-protein binding. With this strategy, we recreated single mutations along the amino acid residues positively selected using the BindProfX server [42]. We exchanged the putative positive selected amino acid on the dimers predicted by the amino acid from different strains of *Hodgkinia* and determined the changes in protein binding affinity. The binding affinities between each pair of proteins were measured as Gibbs free-energy change $\Delta G = G(\text{complex}) - G(\text{monomers})$ when two monomers form a complex, the more negative a ΔG is, the more stable the complex is. The effect of mutations on binding affinity were measured by the differences on free-energy changes between the mutant and the wild type $\Delta\Delta G_{\text{wt} \rightarrow \text{mut}} = \Delta G_{\text{mut}} - \Delta G_{\text{wt}}$. A strongly favourable mutation was considered to have $\Delta\Delta G \leq -1 \text{ kcal/mol}$.

Selective pressure analysis

To study the putative mechanisms of molecular evolution of the RNAP subunits, first we need to know the selective pressure acting on each of the subunits encoding for these genes. With this purpose, we performed D_N/D_S tests on gene/protein sequence, calculated as the ratio D_N/D_S of the number of non-

synonymous substitutions per non-synonymous sites (D_N) divided by the number of synonymous substitutions per synonymous sites (D_S). In such a way that D_N/D_S could take one of three values, i) if non-synonymous mutations are deleterious, *i.e.* purifying selection will reduce their fixation rate and D_N/D_S value will be less than 1; ii) if non-synonymous mutations are advantageous they will be fixed at a higher rate than synonymous mutations, and consequently the D_N/D_S value will be greater than 1; and iii) if the D_N/D_S ratio is equal to 1, it indicates that the evolution is neutral [43]. Selection analysis was performed for RNAP subunits in each group of bacteria.

Each set of genes was implemented in the codeml from PAML v.4.6 package [44], which uses an alignment by codons that was created using the Codon Alignment PAL2NAL v.2.1.0 program [45]. Phylogenetic trees were built using PhyML [46], the set of amino acids sequences of each subunit were first aligned with T-Coffee [47] and the informative blocks were recovered using Gblocks [48]. Two approaches, branch and branch-site were used to identify genes and amino acids under directional positive selection. For branches, three models were used; “M0” one-ratio model (D_N/D_{S0}), Free model (D_N/D_{S1}) and two-ratio model D_N/D_{S2} . The D_N/D_{S0} model assumes the same D_N/D_S ratio for all the branches, the D_N/D_{S1} assumes an independent D_N/D_S for each branch and the D_N/D_{S2} assumes that the interest branch (foreground branch) has a D_N/D_{S2} ratio different than the background ratio [49].

The level of significance for the Likelihood Ratio Test (LRT) was estimated using the χ^2 distribution with degrees of freedom (df), number equal to the difference of the number of parameters between the models and the calculated statistic was defined as twice the difference of log-likelihood between the models ($2\Delta\ln L = 2[\ln L_1 - \ln L_0]$; where L_1 and L_0 are the likelihoods for the alternative and null models, respectively) [50]. One ratio and free ratio models were compared in order to know whether D_N/D_S were different among the lineages, while one ratio and two-ratio models were compared in order to examine whether the lineage of interest has a different ratio compared to other lineages.

To detect positive or negative selection in specific lineages displaying D_N/D_S values greater or smaller than one, we approached through other models, in which model two and the D_N/D_S ratio were fixed to 1 ($D_N/D_{S=1}$), 0.2 ($D_N/D_{S=0.2}$) and 1.2 ($D_N/D_{S=1.2}$) for the foreground branch. First, we compared the models D_N/D_{S2} against the $D_N/D_{S=1}$, where the null hypothesis is that in which the models are not significantly different. If the null hypothesis is rejected ($p < 0.05$) and the two-ratio model is greater than 1, it indicates the possibility of positive selection in the foreground, otherwise, if the two-ratio model estimate is smaller than 1, it is indicative of negative selection. On the other hand, if the null hypothesis is accepted, it is evidence that the foreground branch is under neutral selection (without selection). Additionally, we compared the models D_N/D_{S2} with the $D_N/D_{S=0.2}$, where the null hypothesis is that in which the models are not significantly different. If the null hypothesis is rejected ($p < 0.05$) and the two-ratio model estimate is greater than $D_N/D_S = 0.2$, it indicates a weaker negative selection, while a value smaller than $D_N/D_S = 0.2$ indicates a stronger negative selection. Finally, when D_N/D_{S2} is greater than 1 and it is significantly different to $D_N/D_{S=1}$, it indicates positive selection. In order to get more evidence about possible positive selection we compared D_N/D_{S2} against $D_N/D_{S=1.2}$, the null hypothesis was defined as non-significant difference between D_N/D_{S2} and $D_N/D_{S=1.2}$, acceptance of the null hypothesis indicates that the foreground branch is possibly under positive selection, while a rejection indicates that the foreground branch might be subjected to a relaxed selection.

To identify individual codons under positive selection along specific branches, we performed a branch-site test for positive selection [50]. In these models, positive selection was allowed on a specific, “foreground” branch, and the LRTs ($df=1$) were performed against null models that assume no positive selection is happening. This test has four classes of sites: 0, 1, 2a and 2b; for the site classes 0 and 1, all codons are under purifying selection ($0 < D_N/D_{S0} < 1$) and neutral evolution ($D_N/D_{S1} = 1$) for all branches, respectively. For the sites in classes 2a and 2b, positive selection is allowed on the foreground branches ($D_N/D_{S2} > 1$), but the rest, the “background branches”, are under purifying selection ($0 < D_N/D_{S0} < 1$) and

neutral evolution ($D_N/D_{S1}=1$), respectively. For the null model, D_N/D_{S2} is fixed to 1. In our analysis, all the RNAP subunits in each endosymbiont group were tested that way. Phylogenies were reconstructed by testing each branch as foreground. We compared the two models using LRT and the significance between the models was assessed by calculating twice the log-likelihood difference following a χ^2 distribution, with a df number equal to the difference of the number of parameters between the models. Positive selected amino acid sites are identified based on Empirical Bayes and posterior probabilities were employed in codeml [51]. We did not test the *Nasuia* RNAP subunits, because there were only three sequenced strains but codeml requires at least 4 in order to get reliable results.

Results and Discussions

To evaluate the putative evolution of RNAP's in endosymbiotic bacteria exhibiting extreme genome reduction, we identified orthologous genes of the *E. coli* RNAP genes present in the 37 available genomes of endosymbiotic bacteria (S1 Table). These include 9 strains of *C. ruddii*, 17 strains of *H. cicadicola*, one strain of *T. phenacola*, 7 strains of *T. princeps*, and 3 strains of *N. deltocephalinicola*. Of those, 16 genomes displayed the largest modifications at the protein domain level. For this reason, we decided to describe the RNAP subunits of these 16 organisms (Table 1, Fig 1).

Fig 1. Conserved domains in RNAP subunits of endosymbionts compared to *E. coli*. Conservation of α subunit domains; black fragments represent N-terminal domains, light blue domains represent dimer interfaces, dark blue domains interact with the β subunit, turquoise domains interact with β' , and grey fragments represent the carboxy-terminal end. In the β subunit; green domains interact with β' , orange domains interact with α subunit 1, dark blue domains interact with α subunit 2. In the β' subunit; light blue domains interact with β subunit, catalytic sites are shown in gold, sites interacting with sigma subunits in beige, zinc fingers in orange, active sites in turquoise, black fragments represent carboxy-terminal domains, β prime interface interaction sites are shown in purple, DNA binding sites in yellow,

and G-loop domains in red. σ subunit domains; region 1 in black, region 1.2 in grey, region 2 in red, region 3 in mint, and region 4 in blue.

All bacteria conserve orthologous genes for each of the α , β , β' subunits, as well as a gene coding for a σ subunit. On the other hand, the ω subunit, with in *E. coli* assists during the final step of RNAP core assembly by helping to bind the β' subunit to the $\alpha 2\beta$ sub-complex, it is no longer conserved in any of the analysed endosymbiotic bacteria. Previous studies revealed that ω is not an essential subunit in free-living *E. coli* and therefore it is not surprising that this gene has been lost in all these endosymbionts [15].

Previous works report that proteins in endosymbionts with reduced genomes tend to be shorter compared to free-living bacteria, which leads to the loss of protein regions [10, 14]. We calculate that the DNA sequences encoding for each of the RNAP subunits exhibit a reduction of 16% on average in comparison to those in *E. coli*. In addition, we identified that these genes have lost DNA regions that code for functional protein domains in *E. coli*. That is the case of the α subunit, which lacks the α -carboxy terminal domain in both *Tremblaya* and *Hodgkinia* strains as well as a part of the domain for dimer formation domain in *Hodgkinia* TETUND2 (Fig 1a). The β and β' subunits conserve all the important functional domains (Fig 1b and Fig 1c) and sigma has lost regions 1 and 1.2 in all endosymbionts and region 3 in *Carsonella* strains (Fig 1d). In the next sections, we will describe in more detail the structure and variations found in each RNAP subunit of these endosymbionts.

The α subunit

In *E. coli*, the gene *rpoA* codes for the RNAP α subunit, this protein has a molecular weight of ~ 37 kDa with two independently folded domains connected by two flexible linkers (α carboxy terminal and α amino terminal domains, corresponding to the grey and black boxes in Fig 1a, respectively) [48, 49, 50]. This subunit has three described biological functions: i) it initiates the assembly of the RNAP complex

through the interaction of its amino-terminal domain (α NTD) with β and β' subunits (Fig 2, black upper bar), ii) it participates in promoter recognition through the interaction of its carboxy-terminal domain (α CTD) (Fig 2, grey upper bar) with the DNA UP promoter element, and iii) its carboxy-terminal domain (α CTD) is also a target for the binding of a large number of transcription factors acting as transcriptional activators [52, 53, 54].

Fig 2. Functional domains and amino acid sites of the α -subunit. Black and grey upper bars on the amino acid' alignment represent the domains corresponding to amino- and carboxy-terminal ends. Light blue represents the α subunits dimerization region, dark blue represents the region and the main site of interaction with β subunit, and turquoise represents the main sites of interaction with the β' subunit. Those functional residues that differ from the reference organism (*E. coli*) are shown in darker backgrounds.

The α NTD domain in turn consists of two subdomains: subdomain 1 is called the dimerization domain, and subdomain 2 contains the interface sequences for the β and β' subunits. These subdomains are well conserved among all the α subunits in prokaryotes, archaea and eukaryotes. The subdomain 1 consists of four anti-parallel β -sheets and two orthogonal α helices (H1 and H3). The subdomain 2 is conformed of one α helix and seven β strands in an antiparallel arrangement and are determinant for the interaction of α subunits with the β and β' subunits [55, 56, 57].

The first step towards RNAP core formation is the homodimerization between two monomers of α subunits. This dimer is formed by the interaction of H1 and H3 helices; such interaction takes place in subdomain1 of each monomer. There are nine residues conserved in the α NTD, five of them are in the H1 and H3 helices (ECO: 35F, 38T, 39L, 46I, 227Q). Other three residues (ECO: 32E, 8F, 50S) are in

the loop regions connecting the two α helices and the β sheets, and L31 is localized in the strand two of the C-terminal end (Fig 2, light blue sites).

Punctual mutations affecting residues 45R, 48L or 80E in *E. coli* have shown to prevent the binding of α_2 to the $\beta\beta'$ subunits (Fig 2, dark blue), but do not impede the homodimerization between α subunits. Furthermore, mutations affecting residues 86K, 173V, 180V or 200I (Fig 2, turquoise sites) in *E. coli* prevent the formation of $\alpha_2\beta\beta'$ complex, but don't do so for the $\alpha_2\beta$ complex; hence, these results suggest that these residues are important for the binding of α and β' subunit [57].

In the case of *Carsonella* and *Nasuia* strains, their α subunits conserved all the functional domains as described in *E. coli*, while *Hodgkinia* and *Tremblaya* strains just conserved the α NTD (dimerization of β and β' interaction domains) (Fig 2 light blue, dark blue and turquoise sites and S1 Fig). *In vitro* and *in vivo* experiments revealed that α NTD is essential for RNAP to achieve basal transcription [58, 59]. On the other hand, the α CTD is unnecessary for RNAP assembly and basal transcription but it is required for the interaction of DNA UP-promoter elements and transcription activators in *E. coli* [60, 61]. As mentioned, all these bacteria exhibiting extreme-genome reduction do not conserve any transcription factor in their genomes; and therefore, the absence of the α CTD domain would not impact the basal functionality of RNAP.

The loss of α CTD and other protein domains has also been reported in other studies conducted on bacteria, most of them belonging to the phylum *Parcubacteria*, which are ectosymbionts that live in mixed groups of bacterial communities, and therefore may be subject to a relaxed selection process [62]. The absence of α CTD is also common in some microalgae chloroplasts. In the case of genes encoded by chloroplasts, they are transcribed by two kinds of RNA polymerases, one of these is related to one-single protein phage RNA polymerase encoded in the nuclei genome of microalgae. The other RNAP is more related to the multi-subunit bacterial RNAP and is encoded by the in-chloroplast genome. The absence of α CTD in plant chloroplasts RNAP suggests that recent changes had happened in this clade [63].

A particular case is *H. cicadicola* TETUND 2, which partially conserves an α NTD, precisely located for its interaction with the β' subunit and preserving a few amino acid residues for homodimerization (Fig 2, light blue sites). With the 3D predicted model of this subunit, we observed that the α subunit does not conserve the H1 and H3 helices to form the dimers interface (Fig 5a, white structures). The absence of these two α helices suggests a loss of capacity to form that dimer, which is essential for the RNAP core formation. *H. cicadicola* TETUND 2 and *H. cicadicola* TETUND 1 share the same host, but they are losing different genes. The *rpoA* gene is not the unique case of genes losing a large fragment of their protein products in *H. cicadicola* TETUND 2; the *dnaQ* gene that encodes the ϵ subunit of DNA polymerase III has also lost large fragments, in such a way that it is now considered as a non-functional gene in Tetund 2. A recent analysis reported additional cases of other *Hodgkinia* strains undergoing triple and sixfold coexistence with other *Hodgkinia* strains, which suggests gene complementation [17]. On the other hand, the genome of *Hodgkinia* doesn't code for transporters or other proteins involved in cell envelope biosynthesis that could be mediating the exchange of proteins with other strains [18], hence, it is unclear how these bacteria maintain a functional metabolism.

β and β' subunits

The *rpoB* and *rpoC* genes encode the β and β' subunits of RNAP, respectively. The β' subunit forms a pincer, called “clamp”, and the β subunit constitutes the other pincer, giving place to a 27Å-wide internal channel, where the catalytic site of the RNAP enzyme is located. For the folding of RNAP, it is necessary for the β subunit to bind to the α and β' subunits. Crystallographic patterns obtained from RNAP of different organisms suggests the participation of 24 (ECO: 810-1224) and 73 (ECO: 546-1340) amino acid residues in the β subunit is necessary for its binding to α and β' subunits, respectively [64] (Fig 1b Green, orange and dark blue domains, respectively).

On the other hand, on β' subunit, the interaction sites with the β subunit (at positions ECO: 917-1361, Fig 1c, light blue domain) and ω subunit (ECO: 907-1316) are located on the amino-terminal domain. Additionally, in the amino-terminal domain of the β' subunit there are several functional residues that are typical of Zinc fingers (ECO: C70, C72, C85 and C88) (Fig 1c, orange region), involved in forming the catalytic site and giving stability to the RNAP during its interaction with DNA. The active site is formed by around 500 amino acids retaining two Mg^{2+} ions. However, only thirteen amino acid residues have been identified to be the most conserved ones (ECO: P121, R311, K334, R339, R346, R351, R425, A426, P427, T790, A791, G794 and Y795, Fig 1c, turquoise sites); they are adjacent to the catalytic site, which is composed by three aspartates (ECO: D460, D462 y D464, Fig 1c, golden sites) [65]. In the β' C-terminal domain three polar residues (ECO: R1148, K1311 and E1327, Fig 1c, yellow region) and the G-loop domain (ECO: 1238-1250, Fig 1c, red region) form the cavity where the DNA fits and contact the RNAP.

β and β' amino acid sequences are some of the most conserved among RNAP subunits in endosymbionts. Their critical role in RNAP complex formation may be the reason why these subunits are well conserved even in bacteria containing extremely reduced genomes (S2 Fig and S3 Fig).

Once the RNAP core is formed, σ and β' bind together by means of a protein region contained in the β' subunit, known as the β' coiled-coil domain (ECO: 262-309, Fig 1c, beige region), which interacts with the σ_2 domain of the σ factor (see below). Such binding puts the σ factor in contact with the -10 box, forming an open promoter complex. Different *in vitro* studies involving punctual mutations in the β' coiled-coiled domain residues (ECO: R275Q, E295K and A302D) (Fig 3b, beige sites), resulting in a deficient holoenzyme formation and a subsequent lack of promoter specificity [66]. In addition, the β' -zipper region also interacts with the -10 box element of DNA promoters; this phenomenon largely involves the amino acids residues ECO: Y34 and R35, which form polar interactions with the polyphosphate of the nucleotide in position -17 of the promoter [45, 46].

Likewise, the β subunit binds to the σ factor through the β -flap helix domain (ECO: 889-898) (S2 Fig, grey sites) and the σ_4 region of the σ factor (Fig 4c). The capacity of RNAP to adapt to variation in nucleotide distance between the promoter boxes (-10 and -35) depends on the aforementioned binding [67].

Except for *T. phenacola* PAVE and *C. ruddii* CE, which conserve these domains, the rest of endosymbionts display substitutions in at least one of the three important β' -coiled-coil amino acids (Fig 3b yellow sites and S3 Fig). The β -flap helix domain is conserved in all endosymbionts here studied, including *Hodgkinia* Dsem. However, other *Hodgkinia*'s strains lack this domain (Fig 3c grey sites and S2 Fig). Mutants of the *E. coli* β subunit involving β -flap helix removal result in an inability of the σ subunit to bind the -35 box, without affecting the ability of the RNAP core to bind DNA. Furthermore, these mutants adequately recognize the -10 box and -10 extend promoter elements. The absence of β -flap domain has not been observed in bacteria previously; however, this condition is common in archaeal RNAPs. It has been proposed that the β subunit of archaeal RNAPs recognizes both the TATAAT Box (TB) and an accessory DNA motif located upstream from it [68]. This could suggest that with to exception of *Hodgkinia* Dsem, the RNAP of *Hodgkinia*'s strains cannot bind the -35 box, and instead it would only recognize the -10 box and the -10 extended promoter region during transcription initiation.

Fig 3. Sigma regions and sites involved in promoter recognition and binding. a) Sigma factor regions: σ_2 (red), $\sigma_{2.1}$ (pink), $\sigma_{2.2}$ (magenta), $\sigma_{2.3}$ (violet), $\sigma_{2.4}$ (orange), σ_3 (green) and, σ_4 (dark blue), $\sigma_{4.1}$ (cyan) and $\sigma_{4.2}$ (light blue) b) Regions and important sites of the β' subunit forming the β' -coiled coil domain (yellow) by interacting with a sigma factor c) Region and important sites in the β subunit to form the β -flap tip helix domain (grey).

σ subunit

Sigma factors (σ) are not permanent components of RNAPs core and are the last subunit to be associated with the RNAP complex to initiate DNA transcription. As mentioned, σ is needed for promoter recognition in *E. coli*. This subunit dissociates from the core enzyme once RNAP begins with the RNA synthesis process. In the case of *E. coli*, it encodes a total of seven sigma factors; however, the most essential one is the σ^{70} factor, encoded by the *rpoD* gene. σ^{70} is formed by four helical domains (σ_1 , σ_2 , σ_3 , and σ_4 ; Fig 1c black, red, mint and blue boxes, respectively), each of them interacts with different promoter elements as well as with domains of the subunits forming the core enzyme. The $\sigma_{1.1}$ subdomain, once inserted into the apo-holoenzyme, is located into the downstream channel, in contact with both sides of the RNA cleft [69]. $\sigma_{1.1}$ prevents σ^{70} from interacting with the DNA strand in the absence of the RNAP core [70].

In the meanwhile, the σ_2 domain (Fig 3a red sites) is the most conserved of the σ^{70} family. This region is divided into four sub regions: $\sigma_{2.1}$ (Fig 3a, light pink bar), $\sigma_{2.2}$ (Fig 3a, magenta bar), $\sigma_{2.3}$ (Fig 3a, violet bar) and $\sigma_{2.4}$ (Fig 3a, orange bar). While the $\sigma_{2.2}$ contains the sites involved in the RNAP core binding (ECO: D403, Q406, E407 y N409) (Fig 3a, magenta sites), the $\sigma_{2.3}$ y $\sigma_{2.4}$ participate in DNA melting and the recognition of the -10 box, respectively (Fig 3c). Likewise, the $\sigma_{2.3}$ is characterized by the presence of seven conserved aromatic amino acids (ECO: F419, Y421, Y425, F427, Y430, W433 and W434) (Fig 3a, violet sites), and the replacement of ECO: Y425, Y430, W433 and W434 by alanine results in defects in DNA melting activity. In addition, previous studies showed that residues F419 and F427 play an important role for the precise folding of the α helices of the σ_2 region. Furthermore, the specific recognition of -10 boxes has been attributed to the $\sigma_{2.4}$ region (Fig 3a, orange bar). Multiple studies on σ^{70} have shown that the residues ECO: Q437 and T440 have implications in the recognition of the thymine in position -12 of the core promoter [71, 72]. On the other hand, the position -13 in the σ^{70} promoters is not conserved, and the substitution of residue ECO: 441R in σ has shown to be implicated

in the recognition of nucleotides at this position. Amino acids involved in the DNA melting and recognition of the -10 box are exposed on the same face of the DNA helix. Deletion analyses have determined that the residues ECO: 361-390, comprising the most of σ_2 region, seem to be necessary for RNAP correct functioning [72, 73].

The σ_3 region (Fig 3a, green bar) interacts with the -10 extended promoter element. Additionally, the σ_3 loop might have a role in stabilizing the short nascent DNA-RNA hybrids during the early stage of transcription initiation [74]. More specifically, this is the result of an interaction between the 5'-triphosphate of the nascent RNA with the $\sigma_{3,2}$ loop of the σ factor [75].

Finally, the σ_4 region (Fig 3a dark blue upper bar) is formed by two sub regions, $\sigma_{4,1}$ and $\sigma_{4,2}$ (Fig 3a, cyan and sky blue respectively), comprising four helices. The $\sigma_{4,1}$ region interacts with the β -flap domain (Fig 4c) and, therefore, the distance between the $\sigma_{2,4}$ and $\sigma_{4,2}$ increases and allows their binding to the -35 box of a promoter. Additionally, this region is also a point of contact with activators that bind upstream of the -35 box. The strength of the interaction between the β -flap domain and the $\sigma_{4,1}$ region has effect in promoter specificity. As an example, one study reports that the binding of σ^{38} with the β -flap domain is stronger than with σ^{70} , so the interaction of σ^{38} with the -35 box is stronger than that of σ^{70} , with independence of the promoter sequence. Finally, in the $\sigma_{4,2}$ region the third and fourth helices form a helix-turn-helix structure that interacts with the promoter at the -35 box, specifically, the amino acids R518 and R516 (Fig 3a, sky blue sites) recognize the guanine and cytosine in the -34 and -32 positions, respectively [73].

The σ subunit is the one that exhibits the most differentiated preservation pattern among endosymbionts with extreme genome reduction. In the case of *T. phenacola* PAVE, it conserves the σ_2 and σ_4 regions. On the other site, *Hodgkinia*, *N. deltocephalinicola* and *C. ruddii* only conserve the σ_4 region and to a lesser extent the σ_2 (Fig 3a). The main variations were observed in the σ_2 region, where the residues that

participate in the interaction with the -10 box and define the promoter specificity are located, particularly in the $\sigma_{2.2}$ and $\sigma_{2.4}$ sub regions.

For the $\sigma_{2.2}$ sub-region, *Hodgkinias* exhibit substitutions at all four residues involved in the interaction with the β' coiled-coil domain. The 17 strains have the same substitutions for the sites ECO E407A and 14 of them ECO N409R, while each one retains distinct amino acids at the positions ECO: 403 and ECO 406 (Fig 3a, magenta sites). These variations in the $\sigma_{2.2}$ region might prevent binding to the core enzyme since there would be no recognition of the β' coiled-coil domain. However, mutagenesis of these residues in σ has shown that there is just a weakening in their binding to the coiled-coil domain [66]. In addition, thermal denaturation experiments indicate that mutant subunits are as stably folded as the wild types, suggesting that the principal effect of these mutations is the allosteric regulation of the regions $\sigma_{2.3}$ and $\sigma_{2.4}$, involved in melting and recognition of the -10 box [66]. On the other hand, all the endosymbionts have distinct variations in the $\sigma_{2.3}$ region; however, all of them conserve the essential aromatic residues necessary for DNA melting and for the correct folding of the σ_2 region (Fig 3a, violet sites).

The $\sigma_{2.4}$ region presents exclusive substitutions of the residues involved in the recognition of the nucleotide in position -12 of the -10 box. Similar changes can be seen in all *Hodgkinia* strains, which conserve a histidine instead of glutamine in the position ECO: 437. Likewise, *C. Carsonella ruddii* PV, PC, HT, HC, DC, YCCR and BC replaced the ECO: T440 by a leucine or an isoleucine. Unlike the latter, *Tremblaya* strains and *T. phenacola* PAVE conserve these two sites (ECO Q437 and T440) (Fig 3a, orange sites). Previous studies using punctual mutagenesis in the region $\sigma_{2.4}$ of *E. coli* σ^{70} (70 KDa) and *Bacillus subtilis* SigA (42.9 KDa), the latter is homologous to σ^{70} but smaller, indicated that changes in amino acids at these positions change affect specificity of each σ for its respective promoters [50]. In *E. coli* σ^{70} the substitutions in two independent residues ECO: Q437H and T440I were performed, and the results showed that σ conserves the capacity to recognize the nucleotide at -12, however, the specificity for promoters with a cytosine at this position was significantly higher [69]. Additionally, in SigA (the *B.*

subtilis housekeeping σ equivalent to σ^{70} of *E. coli*) the replacement of ECO: Q196R (corresponding to ECO: Q437 in *E. coli*), had a similar result as in *E. coli* [77].

Substitutions observed in the $\sigma_{2,4}$ region might be related to variations in the specificity of σ for different promoters. According to previous studies in DNA-binding proteins, it has been observed that the substitution of amino acids such as lysine, asparagine, serine, methionine and aromatics like phenylalanine do not difficult their affinity to DNA [78]. However, these changes might exert mild effects on σ affecting its specificity for promoters. Based on structural predictions of σ factors for these endosymbionts, we observed structural similarities compared to *E. coli* σ^{70} (Fig 4). The changes observed in the $\sigma_{2,4}$ region of *C. Hodgkinia* and *C. Nasuia* correspond to mutations already experimentally made in the *E. coli* σ^{70} and *Bacillus* SigA. Nevertheless, it is not possible to ensure the same specificity and recognition of the promoter on σ of endosymbionts because these are the result of a set of complex interactions between distinct regions of σ with other subunits of RNAP and with DNA promoters. So far, the results indicate that RNAP of *Hodgkinia* to exception of the Dsem strain, recognizes only the -10 box and the -10 extended promoter elements, because they lack the fragment required to form the β -flap domain that recognizes the -35 box (Fig 4 d, e, f and g).

Fig 4. Structural comparison between *E. coli* and *Hodgkinia* sigma factors. a) Crystallography structure of *E. coli* RNA polymerase bound to a promoter (4YLP). The figures b, d and f show the structure of regions involved in the recognition and binding of the -10 box of *E. coli* (4YLP), *C. Hodgkinia* Dsem and TETUND2 promoters respectively. Figures c, e and g show the structure of the regions involved in the recognition and binding of the -35 box in *E. coli*, *C. Hodgkinia* TETUND2 and Dsem. The colours in the structures correspond to the colours in the amino acid alignment of Fig 3.

Endosymbionts with reduced genomes also contain variable proportions of GC in their genomes. While *C. Carsonella* and *C. Nasuia* contain less than 18% GC, *C. Hodgkinia* and *C. Tremblaya* contain above 40%. In order to know if such variations of GC content in genomes correlate with changes in σ factors, we carried out a comparative analysis that included a phylogenetic tree. In this analysis we included (in addition to the 31 endosymbiotic bacteria) 13 homologues of σ^{70} coming from other endosymbionts and free-living bacteria with a lesser extent of genome reduction and containing similar, lesser or greater GC contents than the endosymbionts here studied (S2 Table).

In comparison to other endosymbionts with reduced genomes, we obtained that *T. phenacola* PAVE, princeps PCIT and PCVAL are the endosymbionts with the closest σ to *E. coli* σ^{38} , instead of the canonical σ^{70} (S4 Fig). σ^{38} pertains to the σ^{70} family but in *E. coli* it is dedicated to transcribe stationary phase genes. Between these two kinds of σ (70 vs 38), there are some differences in the region $\sigma_{4.1}$ that increase σ^{38} ability to bind to the β -flap domain region of the -35 box.

From this analysis, we were unable to determine any correlation between the GC content and changes in the σ factors. Regions of σ (such as σ^{70} from *E. coli*) are well preserved in all other bacteria, except for the case of *C. Zinderia insecticola* CARI, which is the organism with the lowest genomic GC content, but, unlike other endosymbionts with reduced genomes, it has a less preserved σ_4 region and contains less substitutions in the σ_2 region (Fig 3a).

Studies on selective pressure of RNA polymerase subunits

To analyse the effect of amino acid substitutions observed in genes that encode for RNAP subunits, we made two types of selection analyses. Considering that bacterial endosymbionts are subject to an accelerated rate of molecular evolution [79], a useful tool towards determining the strength and type of selection acting on the genes that encode proteins is the estimation of the ratio of non-synonymous to

synonymous substitutions (D_N/D_S) using phylogenetic codon-substitution models. Under neutral evolution, the rate of accumulation of D_N/D_S should be 1. While a $D_N/D_S > 1$ is indicative of positive selection and a value < 1 indicates negative selection [80]. We used this criterion to identify the kind of selection on genes that encode for the RNAP subunits from each group of endosymbiont bacteria (S3 Table).

Candidatus Hodgkinia cicadicola

The *Hodgkinia cicadicola* RNAP subunits reveal that *rpoB* and *rpoC* genes were subject to a purifying selection with lower D_N/D_S values (0.3) in almost all strains, this might be because most of the nucleotide substitutions in these genes were synonymous. On the other hand, *rpoA* genes had a high purifying selection process in most *Hodgkinia* strains, while in Tetund 2, TETLON and TETMLI1 strains show neutral selection. Finally, the *rpoD* gene show an increased D_N/D_S value, which means that the purifying selection is less rigorous in some *Hodgkinia* strains than in others, this lack of pressure results in greater diversity in protein regions, which possibly corresponds to the different coexistence conditions with other strains in each host (Fig 4).

Fig 5. Selective pressure by branch analysis on the RNAP subunits in obligate endosymbionts. The relation D_N/D_S for each of the subunits and for each endosymbiont was calculated. Here, we show the dispersion of the obtained values. Most genes are under negative selection, and in some cases, they display values greater than 1 (represented as 1.2), this does not mean that they are under positive selection (level of significance greater than 0.05), RNAP subunits shown a relaxed selection in all cases.

Candidatus Tremblaya

For the strain PAVE, the *rpoA*, *rpoB*, *rpoC* and *RpoD* genes had ω values less than 0.02. Conversely, the strain PCVAL shows a neutral selection for all those same subunits ($D_N/D_S > 0.8$). So, the relaxation of selective pressure is not the same for each strain of these bacteria (Fig 5).

Candidatus Carsonella ruddii

Same as for the other endosymbionts, the *rpoB* and *rpoC* genes of *Carsonella ruddii* strains are more conserved than other subunits of RNAP, they had D_N/D_S values less than 0.2 (except for DC and YCCR strains). Of these, the *rpoA* gene had a neutral selection in five strains (CE, CS, DC, HT and YCCR) and there is a strong purifying selection in the rest of *Carsonella* strains. On the contrary, the *rpoD* gene in all these strains had a neutral selection except for the BC strain, which might result in a σ factor containing variations in amino acid content (Fig 5).

Finally, we also performed a branch-site analysis for each subunit in each group of endosymbionts, where D_N/D_S is constant among branches but varies among sites, this approach is used to detect punctual positive selection. Here, our aim was to identify nucleotides in the genes under positive selection. Positively selected amino acids were predicted by a Bayes Empirical Bayes (BEB) analysis for all the branches to find putative significant positive selection. The relationships among these positively selected amino acid residues were evaluated by mapping them on the respective three-dimensional protein structures to determine if these have functional implication (Fig 6).

Fig 6. α -subunit structural comparison between *E. coli* and *C. Hodgkinia TETUND2*. 3D crystallographic structure of monomers and dimers of α -subunit in *E. coli* (4YPL) (a, b). 3D model of monomer α -subunit of *Hodgkinia TETUND2*, sites under positive selection are shown in red, H116 is shown in green (c). Predictions of α -subunit dimer formation in *C. Hodgkinia TETUND2*, sites under

positive selection are shown in red and yellow in each monomer, H116 is shown in green (d, e, f). White regions in a and b structures are not conserved in the α -subunit of *C. Hodgkinia* TETUND2 (g). Amino acids in blue dark and turquoise are important for RNAP core formation. In the b structure, the sites in dark green are the same sites in blue in the structure a.

Positively selected amino acids were detected in the α NTD of *C. Hodgkinia Dsem* and TETUND2 (S4 Table). We mapped the selected amino acids with high level of support (BEB $p > 0.95$) in the structural models of the α subunit of *Hodgkinia* TETUND2 and Dsem (Fig 6c, red amino acids and S5 Fig respectively). In the case of *C. Hodgkinia Dsem* the selected amino acids 55V, 95Q and 115H (S5 Fig) would be important for the correct folding of the α NTD. While, in *C. Hodgkinia* TETUND2 the amino acid residues 3A, 6K, 68E and 71T form a similar structure to that in the α NTD and, due to its localization, they could be involved in the interaction between the α subunit and both the β and β' subunits (Fig 6c).

As previously mentioned, the *Hodgkinia* TETUND2 α subunit has lost part of the α NTD involved in the dimer formation. So, we evaluated if *Hodgkinia* TETUND2 α subunit can form the dimer, which is essential for RNA core formation [30]. For this, we used the software ClustalPro to model the α dimer, we obtained tree models with different size clusters of amino acids involved in dimer formation (Fig 6 d, e, and f). In these models, we mapped the sites under positive selection (Fig 6 d, e, and f, amino acids in red and yellow). Then, we performed *in silico* amino acid substitutions of the sites under positive selection, changing them by other amino acids present in the same sites of other *Hodgkinia* strains and *E. coli*. Finally, we used the software Bindx to evaluate the free-energy changes in protein-protein binding for the three models. We observed that the substitution of H116 by a proline in the three models changed the dimer formation in an energetically unfavourable way (S5 Table). Although the site H116 is not statically significant to be considered under positive selection, it would have an important role for

stabilizing the dimer formation in this strain, however, experimental analysis is necessary to confirm this prediction.

Conclusions

Endosymbionts are highly specialized organisms, which seems to be the result of co-evolution with their hosts. This might lead endosymbionts to evolve its genomic content to produce proteins that are simplified versions of their free-living counterparts. In this work, we studied the changes of RNAP subunits, trying to get an idea on how these enzymes carry out DNA transcription in these organisms.

We found that the β and β' subunits tend to be conserved in all the studied endosymbionts, with just some differences in the regions involved in the interactions with the σ factor, possibly as result of major changes in σ . On the other side, the α subunit is more conserved in *C. ruddii* and *Nasuia* but than in other endosymbionts where the α subunit has lost its C-terminal domain. Furthermore, the total loss of the gene encoding this subunit has been reported in some strains of *Hodgkinia* [17], this may be due to complementation with genes from other bacteria, since these strains share host with other *Hodgkinia* strains. The strains Dsem, CHOCRA and TETULN are the unique strains known to not share their host with any other *Hodgkinia* strains and possibly that is why they preserve an α subunit, as we know for functional gene. On other hand, the rest of *Hodgkinia* strains share their host with one to five different strains, however, in these cases each *Hodgkinia* complex retains the α subunit genes [17]. We observed several amino acids under putative positive selection in the α subunit gene of *Hodgkinia* TETUND2, that would suggest a compensation in the loss of important regions for dimer formation in this subunit.

Some interesting events seem to be happening in all endosymbionts that do not occur in free-living bacteria. The σ factor encoding gene *rpoD* does not conserve the region 2, as does the conventional σ^{70} unit in *E. coli*; furthermore, we observed the loss of up to two domains (σ_1 and σ_3). Thus, drastic changes in endosymbiotic RNAPs coming from reduced genomes are related to sites and regions involved in promoter recognition, as we know in free-living organisms. These differences are not observed in any

other free-living bacteria, not even in endosymbionts with larger genomes (*e.g.*, with more than 170 Kpb genome size), which might suggest these changes as consequence of the drastic process of genome reduction that are experimenting these organisms.

In endosymbiont bacteria with reduced genomes it has not been possible to locate σ^{70} canonical promoters, not even for those the most conserved like those in ribosomal genes of bacteria [81], it might be, at least in part, due to the high A+T percentage in these genomes. However, *Hodgkinia* and *Tremblaya* have relatively high G+C% and lack intragenic regions, similarly to the rest of endosymbionts (more than 90% of their genome is composed of coding sequences) [6]. In addition, variations observed in σ factor regions are involved in the recognition of some promoter regions, hence, our results suggest that RNAP in endosymbionts conserve some capacity of sequence recognition, which is different to the σ^{70} consensus recognition sequence known for *E. coli*. However, experimental analysis must be done to corroborate this hypothesis.

Variations in recognition regions and sites of promoters might not be only ones in these bacteria. Previous reports indicate significant changes in the 16S ribosomal 3' tail and its binding sequence, as well as in the Shine-Dalgarno element localized upstream of the protein-encoding genes [82].

Finally, we proposed a model of RNAP for each endosymbiont according to our findings (Fig 7). We observed that the RNAP of *Nasuia* conserved greater similarity to *E. coli* σ^{70} in comparison to other endosymbionts. According to this, it is expected that *Nasuia* RNAP recognizes the two promoters boxes (-10 and -35) and the two promoter accessory zones (-10 extended and UP element), in contrast to *Carsonella* RNAP, which seems to be unable to recognize the -10 extended box.

Fig 7. Proposed structural RNAP models of bacterial endosymbionts with reduced genomes. The conserved domains in each group of bacteria are identified in each model: *E. coli* RNAP model (Eco),

Natsuia strains RNAP model, *Carsonella* strains RNAP model, *Hodgkinias* RNAP model (In orange, the TETULN, TETUND1 and TETUND1 strains exclusively using -10 extended region. In purple, Dsem strains exclusively using the -35 element), and *T. phenacola* PAVE, princeps PCIT and PCVAL RNAP model (Fig based on Busby and 2004 and Rangel-Chavez et al, 2019).

On the other hand, in the case of *Tremblaya* and *Hodgkinia* RNAP, due the absence of α C-terminal domain, they are unable of recognizing the UP element, and in the case of *C. Hodgkinia Dsem* neither the -10 extended promoter element. Contrary to *C. Hodgkinia Dsem*, other *Hodgkinia* strains are expected to recognize the -10 extended promoter element and, in this way, might compensate the loss of the β -flap-helix domain. The model of RNAP proposed for *Hodgkinia* suggests that it no longer conserves the mechanisms for transcription initiation common in free-living bacteria, as they lack the β flap-helix domain. In free-living bacteria, the β flap-helix domain is essential for the assembly of sigma to the RNAP core.

In a whole view, and regarding the study of RNAP subunits, we conclude that these four groups of endosymbionts with extreme-genome reduction do not follow the same evolutionary path, and that each strain has particular modifications, which may be a consequence of the conditions exerted by the host and/or the coexistence with other endosymbionts.

Conflict of interest.

Authors declare no conflict of interest.

Acknowledgments

CPR-C has a PhD fellowship (380338) from CONACYT. México. Authors thank Rafael Montiel for guiding on the selective pressure analyses and to Diego Andrés López Castro and Paola Angulo-Bejarano for reading the manuscript and to anonymous referees for useful comments to the document.

Author Contributions

Conceptualization: AM-A and CPR-C

Formal analysis: Cynthia Paola Rangel-Chávez, Edgardo Galán-Vásquez, Azucena Pescador-Tapia.

Funding acquisition: Agustino Martínez-Antonio.

Supervision: Agustino Martínez-Antonio.

Writing – original draft: Cynthia Paola Rangel-Chávez, Edgardo Galán-Vásquez, Agustino Martínez-Antonio.

Writing – review & editing: Agustino Martínez-Antonio, Luis Delaye Arredondo.

Supplementary material

S1 Table. Complete list of studied endosymbionts

S2 Table. Bacteria with similar %GC to endosymbiotic bacteria with reduced genomes

S3 Table. Results obtained of selective pressure analysis by branch model

S4 Table. Results of selective pressure obtained by branch-site model

S5 Table. Changes of free energy in selected sites of α subunit homodimer

S1 Fig. Interaction domains in α subunit alignment

S2 Fig. Interaction domains in the β subunit alignment

S3 Fig. Functional domains and sites in β' subunit alignment

S4 Fig. Phylogenetic Tree of sigma 70 between endosymbionts with extreme genome reduction and other bacteria

S5 Fig. Selected sites mapped in *Hodgkinia* Dsem α subunit

References

1. McCutcheon, J. P., & Moran, N. A. Extreme genome reduction in symbiotic bacteria. *Nature Reviews Microbiology*. 2012; 10(1), 13-26.
2. Moran, N. A., Dunbar, H. E., & Wilcox, J. L. Regulation of transcription in a reduced bacterial genome: nutrient-provisioning genes of the obligate symbiont *Buchnera aphidicola*. *Journal of Bacteriology*. 2005; 187(12), 4229-4237.
3. Huerta, A. M., Francino, M. P., Morett, E., & Collado-Vides, J. Selection for unequal densities of [[sigma]. sup. 70] promoter-like signals in different regions of large bacterial genomes. *PLoS Genetics*. 2006; 2(11), 1740-1751.
4. Ptashne M. Regulation of transcription: from lambda to eukaryotes. *Trends Biochem Sci*. 2005; 30 (6): 275-279.
5. Ebright RH. RNA polymerase: Structural similarities between bacterial RNA polymerase and eukaryotic RNA polymerase II. *J Mol Biol*. 2000; 304: 687-698.
6. Moran NA, Bennett GM. The tiniest tiny genomes. *Ann Rev Microbiol*. 2014; 68: 195-215.
7. Gentry D, Burgess RR. rpoZ, encoding the omega subunit of *Escherichia coli* RNA polymerase, is in the same operon as spot. *J Bacteriol*. 1989; 171(3): 1271-1277.
8. Sweetser D, Nonet M, Young RA. Prokaryotic and eukaryotic RNA polymerases have homologous core subunits. 1987; *Proc Natl Acad Sci USA*. 84: 1192.-1196.

9. Ishihama A. Functional modulation of Escherichia coli RNA polymerase. *Annu Rev Microbiol.* 2000; 54: 499-518.
10. Rangel-Chavez, C., Galan-Vasquez, E., & Martinez-Antonio, A. Consensus architecture of promoters and transcription units in Escherichia coli: design principles for synthetic biology. *Molecular bioSystems.* 2017; 13(4), 665–676.
11. Eleftherianos, I., Atri, J., Accetta, J., & Castillo, J. C. Endosymbiotic bacteria in insects: guardians of the immune system? *Frontiers in physiology.* 2013; 4, 46.
12. McCutcheon JP, Moran NA. Extreme genome reduction in symbiotic bacteria. *Nat Rev Microbiol.* 2012; 10 (1): 13-26.
13. Blattner, F. R., Plunkett, G., Bloch, C. A., Perna, N. T., Burland, V., Riley, M., ... & Gregor, J. The complete genome sequence of Escherichia coli K-12. *Science.* 1997; 277(5331), 1453-1462.
14. Sloan, D. B., & Moran, N. A. Genome reduction and co-evolution between the primary and secondary bacterial symbionts of psyllids. *Molecular biology and evolution.* 2012; 29(12), 3781-3792.
15. Tamames J, Gil R, Latorre A, Peretó J, Silva FJ, Moya A. The frontier between cell and organelle: genome analysis of *Candidatus Carsonella ruddii*. *BMC Evol Biol.* 2007; 7: 181.
16. Galán-Vásquez E, Sánchez-Osorio I, Martínez-Antonio A. Transcription Factors Exhibit Differential Conservation in Bacteria with Reduced Genomes. *PLOS ONE.* 2016; 11(1): e0146901.
17. Lukasik, P., Nazario, K., Van Leuven, J. T., Campbell, M. A., Meyer, M., Michalik, A., ... & McCutcheon, J. P. Multiple origins of interdependent endosymbiotic complexes in a genus of cicadas. *Proceedings of the National Academy of Sciences.* 2018. 115(2): E226-E235.

18. Van Leuven, J. T., Meister, R. C., Simon, C., & McCutcheon, J. P. Sympatric speciation in a bacterial endosymbiont results in two genomes with the functionality of one. *Cell*. 2014; 158(6): 1270–1280.
19. McCutcheon, J. P., McDonald, B. R., & Moran, N. A. Convergent evolution of metabolic roles in bacterial co-symbionts of insects. *Proceedings of the National Academy of Sciences*. 2009; 106(36): 15394-15399.
20. Husnik, F., Nikoh, N., Koga, R., Ross, L., Duncan, R. P., Fujie, M., ... & Von Dohlen, C. D. Horizontal gene transfer from diverse bacteria to an insect genome enables a tripartite nested mealybug symbiosis. *Cell*. 2013; 153(7): 1567-1578.
21. McCutcheon, J. P., & Von Dohlen, C. D. An interdependent metabolic patchwork in the nested symbiosis of mealybugs. *Current biology*. 2011; 21(16): 1366-1372.
22. López-Madrigal, S., Latorre, A., Porcar, M., Moya, A., & Gil, R. Complete genome sequence of “*Candidatus Tremblaya princeps*” strain PCVAL, an intriguing translational machine below the living-cell status. *Journal of Bacteriology*. 2011; 193 (19): 5587-5588.
23. Husnik, F., & McCutcheon, J. P. Repeated replacement of an intrabacterial symbiont in the tripartite nested mealybug symbiosis. *Proceedings of the National Academy of Sciences*. 2016; 113(37): E5416-E5424.
24. Gordon M. Bennett, Nancy A. Moran, Small, Smaller, Smallest: The Origins and Evolution of Ancient Dual Symbioses in a Phloem-Feeding Insect, *Genome Biology and Evolution*. 2013; 5(9): 1675–1688
25. Bennett, G. M., Abbà, S., Kube, M., & Marzachi, C. Complete Genome Sequences of the Obligate Symbionts "*Candidatus Sulcia muelleri*" and "*Ca. Nasuia deltocephalinicola*" from the Pestiferous Leafhopper *Macrostelus quadripunctulatus* (Hemiptera: Cicadellidae). *Genome announcements*. 2014; 4(1): e01604-15.

26. Mao, M., Yang, X., Poff, K., & Bennett, G. Comparative genomics of the dual-obligate symbionts from the treehopper, *Entylia carinata* (Hemiptera: Membracidae), provide insight into the origins and evolution of an ancient symbiosis. *Genome Biology and Evolution*. 2017; 9(6): 1803-1815.
27. Nakabachi, A., Yamashita, A., Toh, H., Ishikawa, H., Dunbar, H. E., Moran, N. A., & Hattori, M. The 160-kilobase genome of the bacterial endosymbiont *Carsonella*. *Science*. 2006; 314(5797): 267.
28. Nakabachi, A., Ueoka, R., Oshima, K., Teta, R., Mangoni, A., Gurgui, M., Oldham, N. J., van Echten-Deckert, G., Okamura, K., Yamamoto, K., Inoue, H., Ohkuma, M., Hongoh, Y., Miyagishima, S. Y., Hattori, M., Piel, J., & Fukatsu, T. Defensive bacteriome symbiont with a drastically reduced genome. *Current biology*. 2013; 23(15): 1478–1484.
29. Wu, F., Zheng, Z., Deng, X., Cen, Y., Liang, G., & Chen, J. Draft Genome Sequence of "Candidatus *Liberibacter asiaticus*" from *Diaphorina citri* in Guangdong, China. *Genome announcements*. 2015; 3(6): e01316-15.
30. Riley, A. B., Kim, D., & Hansen, A. K. Genome Sequence of "Candidatus *Carsonella ruddii*" Strain BC, a Nutritional Endosymbiont of *Bactericera cockerelli*. *Genome announcements*. 2017; 5(17): e00236-17.
31. Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, Potter SC, Punta M, Qureshi M, Sangrador-Vegas A, Salazar GA, Tate J, Bateman A. The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res*. 2016; 44(D1): D279-85.
32. Gough J, Karplus K, Hughey R, Chothia C. Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *J Mol Biol*. 2001; 13(4): 903-19.

33. Marchler-Bauer, A., Bo, Y., Han, L., He, J., Lanczycki, C. J., Lu, S., ... & Gwadz, M. CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. *Nucleic acids research*. 2016; 45(D1): D200-D203.
34. Zhang, Y. I-TASSER server for protein 3D structure prediction. *BMC bioinformatics*. 2008; 9(1): 40.
35. Zuo, Y., & Steitz, T. A. Crystal structures of the E. coli transcription initiation complexes with a complete bubble. *Molecular cell*. 2015; 58(3): 534-540.
36. The PyMOL Molecular Graphics System, Version 1.2r3pre, Schrödinger, LLC.
37. Kozakov, D., Hall, D. R., Xia, B., Porter, K. A., Padhorny, D., Yueh, C., ... & Vajda, S. The ClusPro web server for protein–protein docking. *Nature protocols*. 2017; 12(2): 255.
38. Vajda S, Yueh C, Beglov D, Bohnuud T, Mottarella SE, Xia B, Hall DR, Kozakov D. New additions to the ClusPro server motivated by CAPRI. *Proteins: Structure, Function, and Bioinformatics*. 2017; 85(3): 435-444.
39. Kozakov D, Beglov D, Bohnuud T, Mottarella S, Xia B, Hall DR, Vajda, S. How good is automated protein docking. *Proteins: Structure, Function, and Bioinformatics*. 2013; 81(12): 2159-66
40. Yueh C, Hall DR, Xia B, Padhorny D, Kozakov D, Vajda S. ClusPro-DC: Dimer Classification by the Cluspro Server for Protein–Protein Docking. *Journal of Molecular Biology*. 2017; 429(3): 372-381
41. Lensink, M. F., Méndez, R., & Wodak, S. J. Docking and scoring protein complexes: CAPRI 3rd Edition. *Proteins: Structure, Function, and Bioinformatics*. 2007; 69(4): 704-718.

42. Xiong, P., Zhang, C., Zheng, W., & Zhang, Y. BindProfX: assessing mutation-induced binding affinity change by protein interface profiles with pseudo-counts. *Journal of molecular biology*. 2017; 429(3): 426-434.
43. Ziheng Yang, Rasmus Nielsen, Estimating Synonymous and Nonsynonymous Substitution Rates Under Realistic Evolutionary Models, *Molecular Biology and Evolution*. 2000; 17(1): 32–43.
44. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Molecular biology and evolution*. 2007; 24(8): 1586-1591.
45. Suyama, M., Torrents, D., & Bork, P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic acids research*. 2006; 34(Web Server issue): W609–W612.
46. Guindon, S., Dufayard, J. F., Lefort, V., Anisimova, M., Hordijk, W., & Gascuel, O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic biology*. 2010; 59(3): 307-321.
47. Notredame, C., Higgins, D.G., Heringa, J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol*. 2000; 302(1): 205-217.
48. Castresana, J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular Biology and Evolution*. 2000; 17: 540-552.
49. Yang, Z., & Nielsen, R. Synonymous and nonsynonymous rate variation in nuclear genes of mammals. *Journal of molecular evolution*. 1998; 46(4), 409-418.
50. Zhang, J., Nielsen, R., & Yang, Z. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Molecular biology and evolution*. 2005; 22(12): 2472-2479.

51. Yang, Z., Wong, W. S., & Nielsen, R. Bayes empirical Bayes inference of amino acid sites under positive selection. *Molecular biology and evolution*. 2005; 22(4): 1107-1118.
52. Blatter EE, Ross W, Tang H, Gourse RL, Ebright RH. Domain organization of RNA polymerase α subunit: C-terminal 85 amino acids constitute a domain capable of dimerization and DNA binding. *Cell*. 1994; 78(5): 889-896.
53. Ross W, Ernst A, Gourse RL. Fine structure of E. coli RNA polymerase-promoter interactions: α subunit binding to the UP element minor groove. *Genes Dev*. 2001; 15(5): 491-506.
54. Murakami, K. S., Masuda, S., Campbell, E. A., Muzzin, O., & Darst, S. A. Structural basis of transcription initiation: an RNA polymerase holoenzyme-DNA complex. *Science*. 2002; 296(5571): 1285-1290.
55. Busby S, Ebright RH. Promoter structure, promoter recognition and transcription activation in prokaryotes. *Cell*. 1994; 79(5): 743-746.
56. Ebright RH, Busby S. The Escherichia coli RNA polymerase subunit: structure and function. *Curr Opin Genet Dev*. 1995; 5(2): 197-203.
57. Zhang G, Darst SA. Structure of the Escherichia coli RNA polymerase subunit Amino-terminal domain. *Science*. 1998; 281(5374): 262-266.
58. Kimura, M., Fujita, N., & Ishihama, A. Functional map of the α subunit of Escherichia coli RNA polymerase: deletion analysis of the amino-terminal assembly domain. *Journal of molecular biology*. 1994; 242(2), 107-115.
59. Igarashi, K., & Ishihama, A. Bipartite functional map of the E. coli RNA polymerase α subunit: involvement of the C-terminal region in transcription activation by cAMP-CRP. *Cell*. 1991; 65(6): 1015-1022.

60. Hayward, R. S., Igarashi, K., & Ishihama, A. Functional specialization within the α -subunit of *Escherichia coli* RNA polymerase. *Journal of molecular biology*. 1991; 221(1): 23-29.
61. Ross W, Gosink KK, Salomon J, Igarashi K, Zou C, Ishihama A, Severinov K, Gourse RL. A third recognition element in bacterial promoters: DNA binding by the α subunit of RNA polymerase. *Science*. 1993; 262(5138): 1407-1413.
62. Nelson, W. C., & Stegen, J. C. The reduced genomes of Parcubacteria (OD1) contain signatures of a symbiotic lifestyle. *Frontiers in microbiology*. 2015; 6, 713.
63. Sheveleva, E. V., Giordani, N. V., & Hallick, R. B. Identification and comparative analysis of the chloroplast α -subunit gene of DNA-dependent RNA polymerase from seven *Euglena* species. *Nucleic acids research*. 2002; 30(5): 1247-1254.
64. Cramer, P., Bushnell, D. A., & Kornberg, R. D. Structural basis of transcription: RNA polymerase II at 2.8 Å resolution. *Science*. 2001; 292(5523): 1863-1876.
65. Vassylyev, D. G., Sekine, S. I., Laptenko, O., Lee, J., Vassylyeva, M. N., Borukhov, S., & Yokoyama, S. Crystal structure of a bacterial RNA polymerase holoenzyme at 2.6 Å resolution. *Nature*. 2002; 417(6890): 712-719.
66. Young, B. A., Anthony, L. C., Gruber, T. M., Arthur, T. M., Heyduk, E., Lu, C. Z., ... & Gross, C. A. A coiled-coil from the RNA polymerase β' subunit allosterically induces selective nontemplate strand binding by $\sigma 70$. *Cell*. 2001; 105(7): 935-944.
67. Geszvain, K., Gruber, T. M., Mooney, R. A., Gross, C. A., & Landick, R. A hydrophobic patch on the flap-tip helix of *E. coli* RNA polymerase mediates $\sigma 70$ region 4 function. *Journal of molecular biology*. 2004; 343(3): 569-587.
68. Hirata, A., & Murakami, K. S. Archaeal RNA polymerase. *Current opinion in structural biology*. 2009; 19(6): 724-731.

69. Bae, B., Davis, E., Brown, D., Campbell, E. A., Wigneshweraraj, S., & Darst, S. A. Phage T7 Gp2 inhibition of Escherichia coli RNA polymerase involves misappropriation of $\sigma 70$ domain 1.1. Proceedings of the National Academy of Sciences. 2013; 110(49): 19772-19777.
70. Camarero, J. A., Shekhtman, A., Campbell, E. A., Chlenov, M., Gruber, T. M., Bryant, D. A., ... & Muir, T. W. Autoregulation of a bacterial σ factor explored by using segmental isotopic labeling and NMR. Proceedings of the National Academy of Sciences. 2002; 99(13): 8536-8541.
71. Waldburger, C., Gardella, T., Wong, R., & Susskind, M. M. Changes in conserved region 2 of Escherichia coli $\sigma 70$ affecting promoter recognition. Journal of molecular biology. 1990; 215(2): 267-276.
72. Siegele, D. A., Hu, J. C., Walter, W. A., & Gross, C. A. Altered promoter recognition by mutant forms of the $\sigma 70$ subunit of Escherichia coli RNA polymerase. Journal of molecular biology. 1989; 206(4): 591-603.
73. Lesley, S. A., & Burgess, R. R. Characterization of the Escherichia coli transcription factor. sigma. 70: localization of a region involved in the interaction with core RNA polymerase. Biochemistry. 1989; 28(19): 7728-7734.
74. Malhotra, A., Severinova, E., & Darst, S. A. Crystal structure of a $\sigma 70$ subunit fragment from E. coli RNA polymerase. Cell. 1996; 87(1): 127-136.
75. Zuo, Y., & Steitz, T. A. Crystal structures of the E. coli transcription initiation complexes with a complete bubble. Molecular cell. 2015; 58(3), 534-540.
76. Kuznedelov, K., Minakhin, L., Niedziela-Majka, A., Dove, S. L., Rogulja, D., Nickels, B. E., ... & Severinov, K. A role for interaction of the RNA polymerase flap domain with the σ subunit in promoter recognition. Science. 2002; 295(5556): 855-857.

77. Kenney, T. J., York, K., Youngman, P., & Moran, C. P. Genetic evidence that RNA polymerase associated with sigma A factor uses a sporulation-specific promoter in *Bacillus subtilis*. *Proceedings of the National Academy of Sciences*. 1989; 86(23): 9109-9113.
78. Luscombe, N. M., Laskowski, R. A., & Thornton, J. M. Amino acid–base interactions: a three-dimensional analysis of protein–DNA interactions at an atomic level. *Nucleic acids research*. 2001; 29(13): 2860-2874.
79. Dufresne, A., Garczarek, L. & Partensky, F. Accelerated evolution associated with genome reduction in a free-living prokaryote. *Genome Biol* . 2005; 6(2): R14.
80. Jeffares, D. C., Tomiczek, B., Sojo, V., & dos Reis, M. A beginner’s guide to estimating the non-synonymous to synonymous rate ratio of all protein-coding genes in a genome. *Methods Mol Biol*. 2015; 1201: 65-90.
81. Clark, M. A., Baumann, L., Thao, M. L., Moran, N. A., & Baumann, P. Degenerative minimalism in the genome of a psyllid endosymbiont. *Journal of Bacteriology*. 2001; 183(6): 1853-1861.
82. Lim, K., Furuta, Y., & Kobayashi, I. Large variations in bacterial ribosomal RNA genes. *Molecular biology and evolution*. 2012; 29(10), 2937-2948.

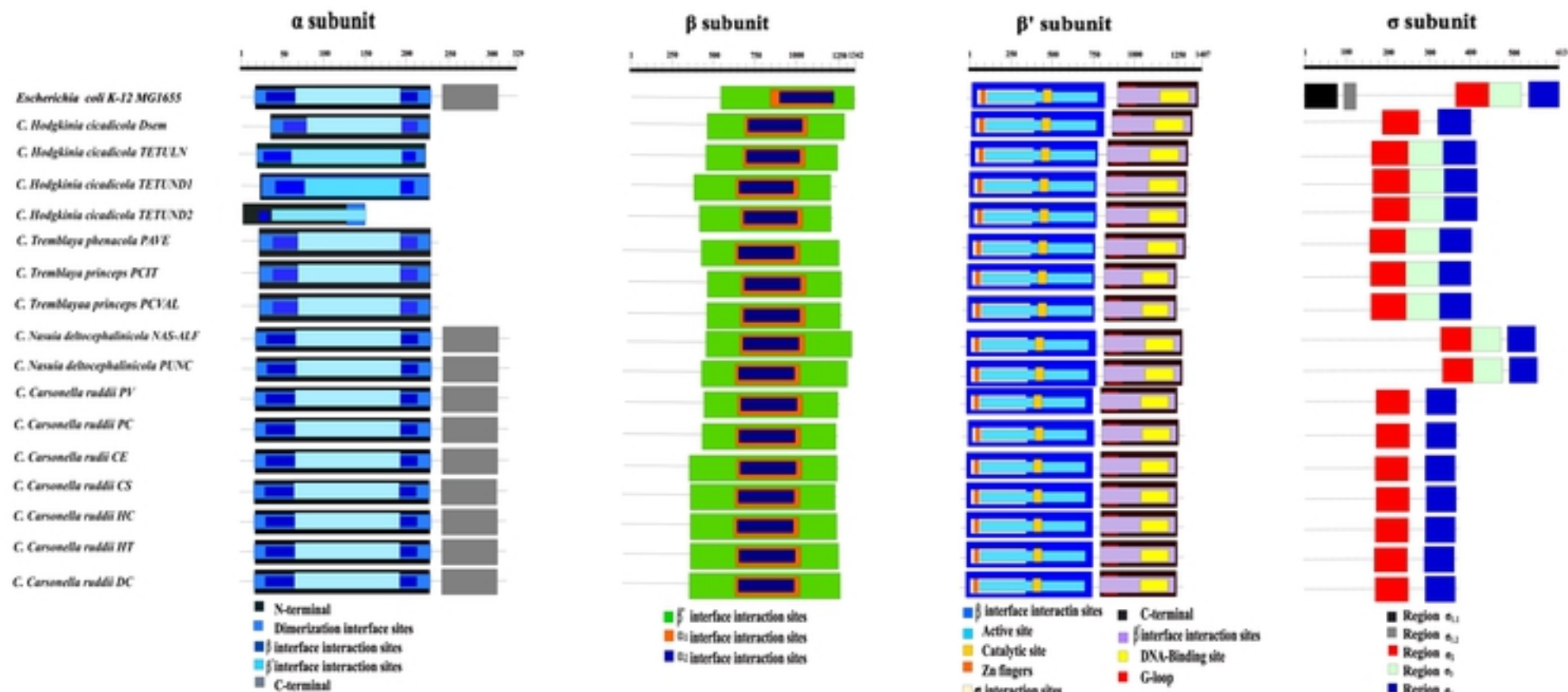


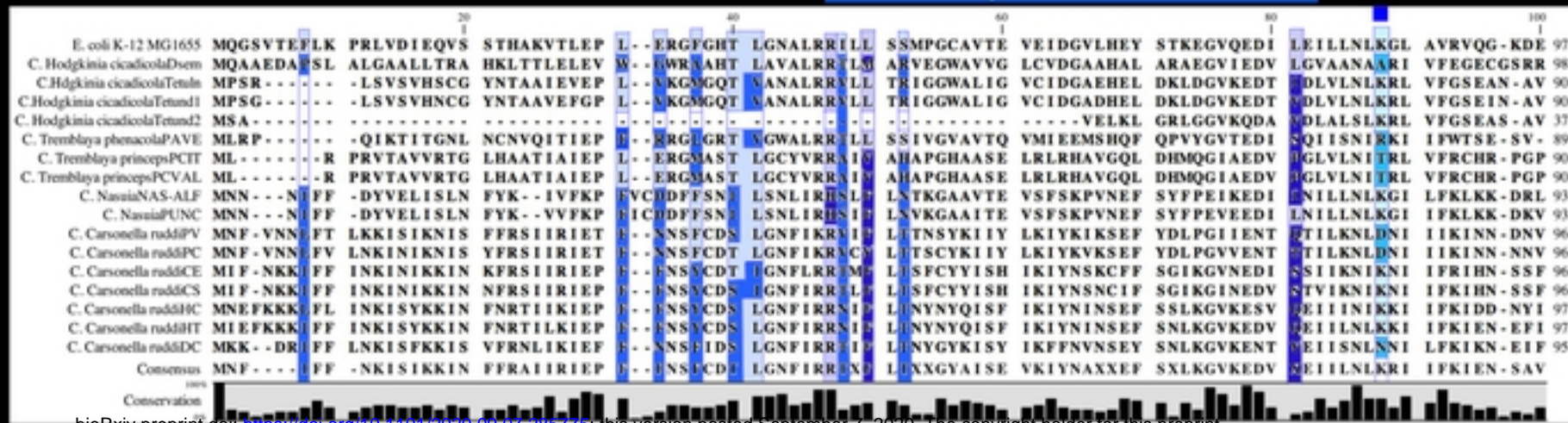
Figure 1

α -NTD

Dimer interface

β interaction site

β interaction site



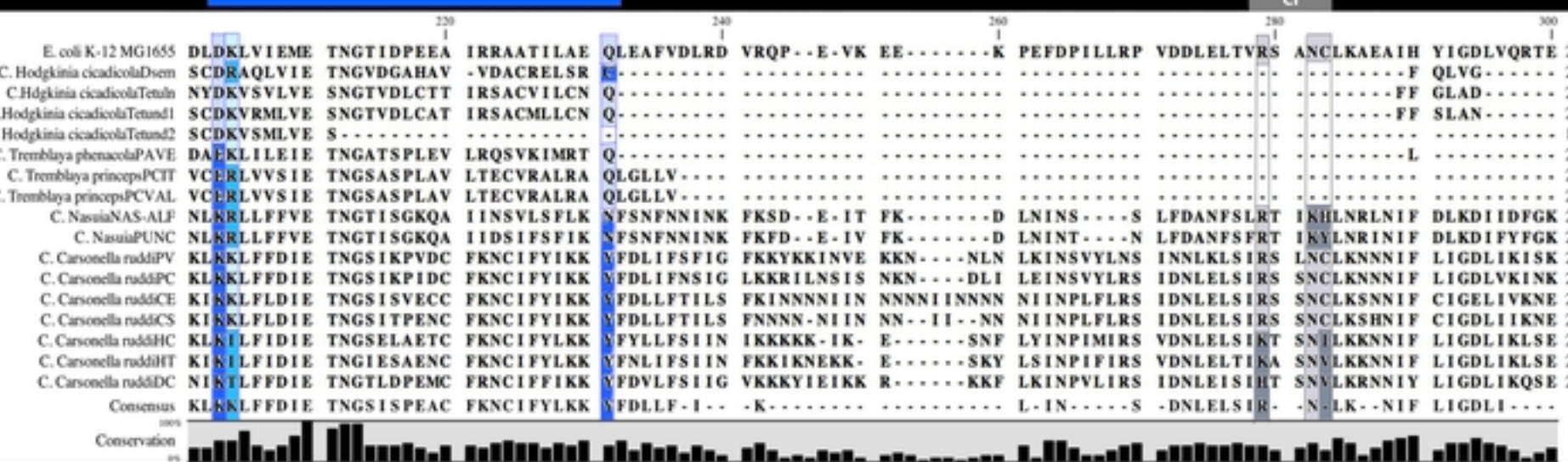
α -NTD



α -NTD

α -CTD

C1



α -CTD

CII

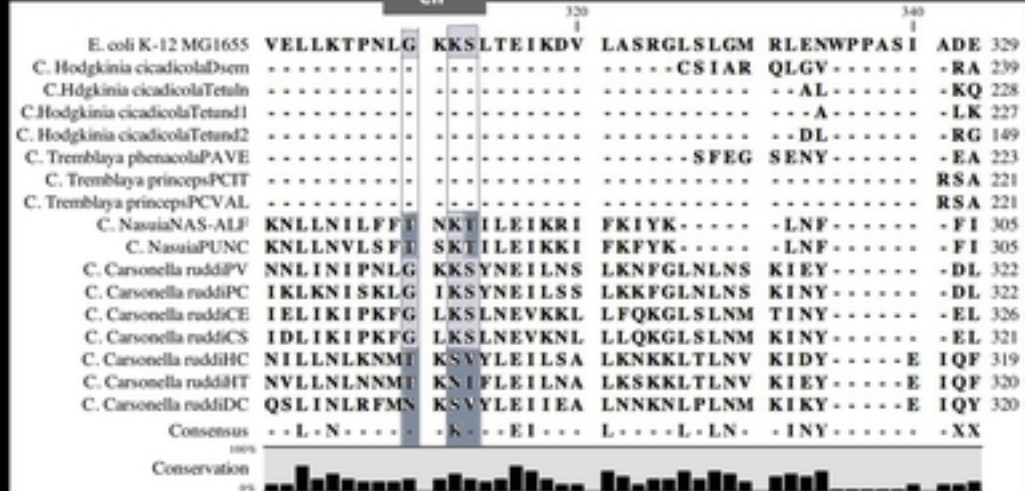
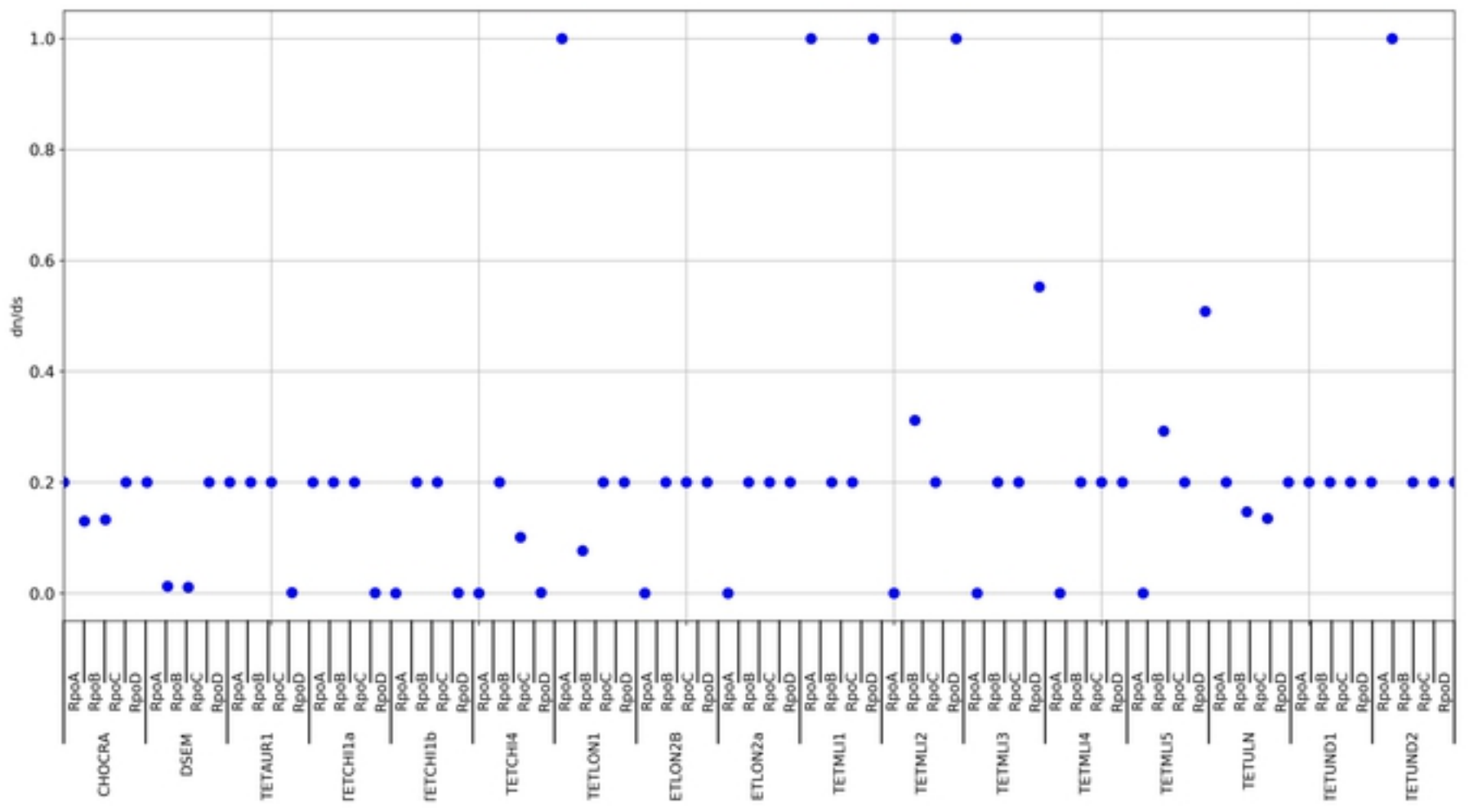
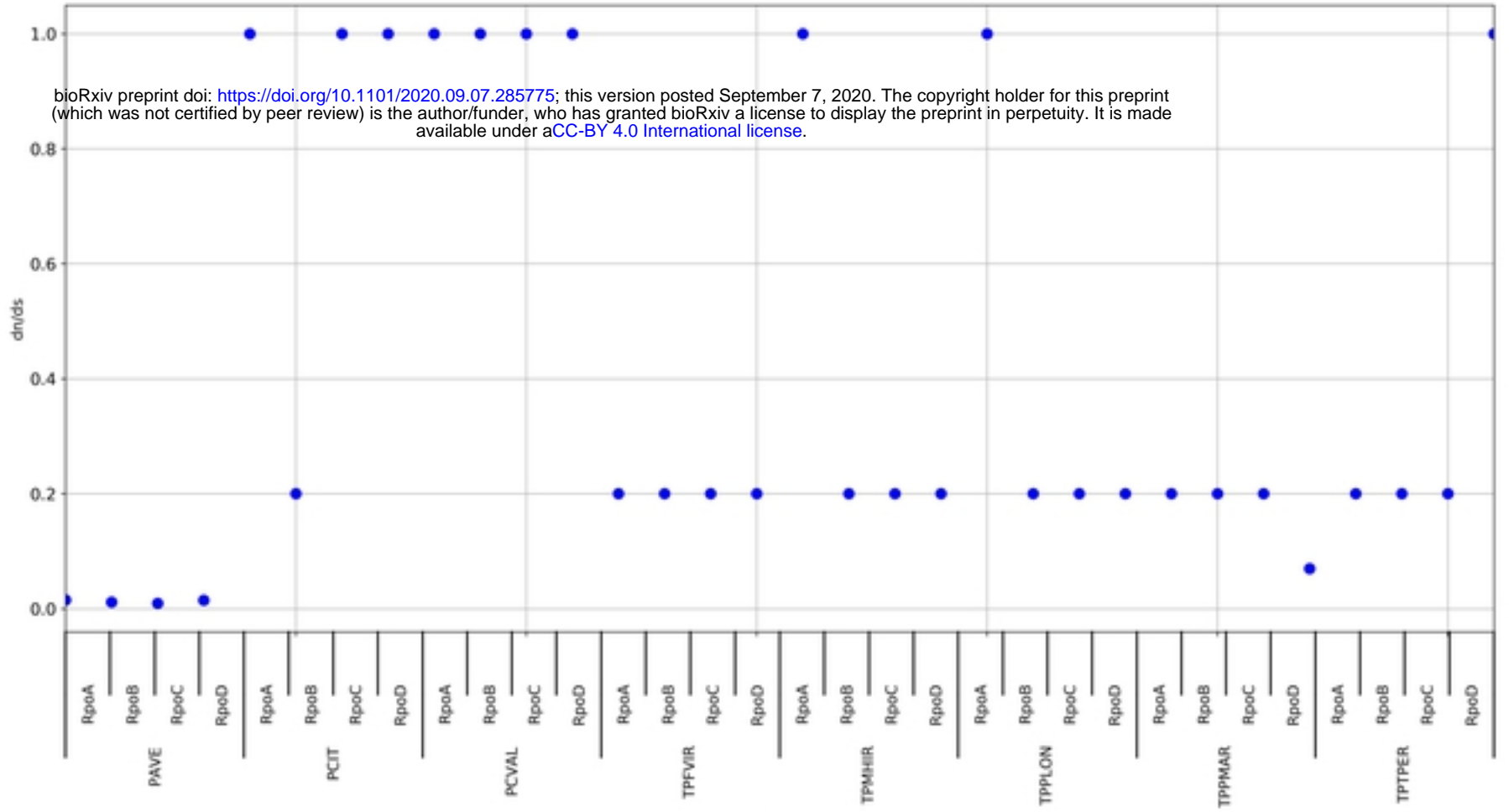


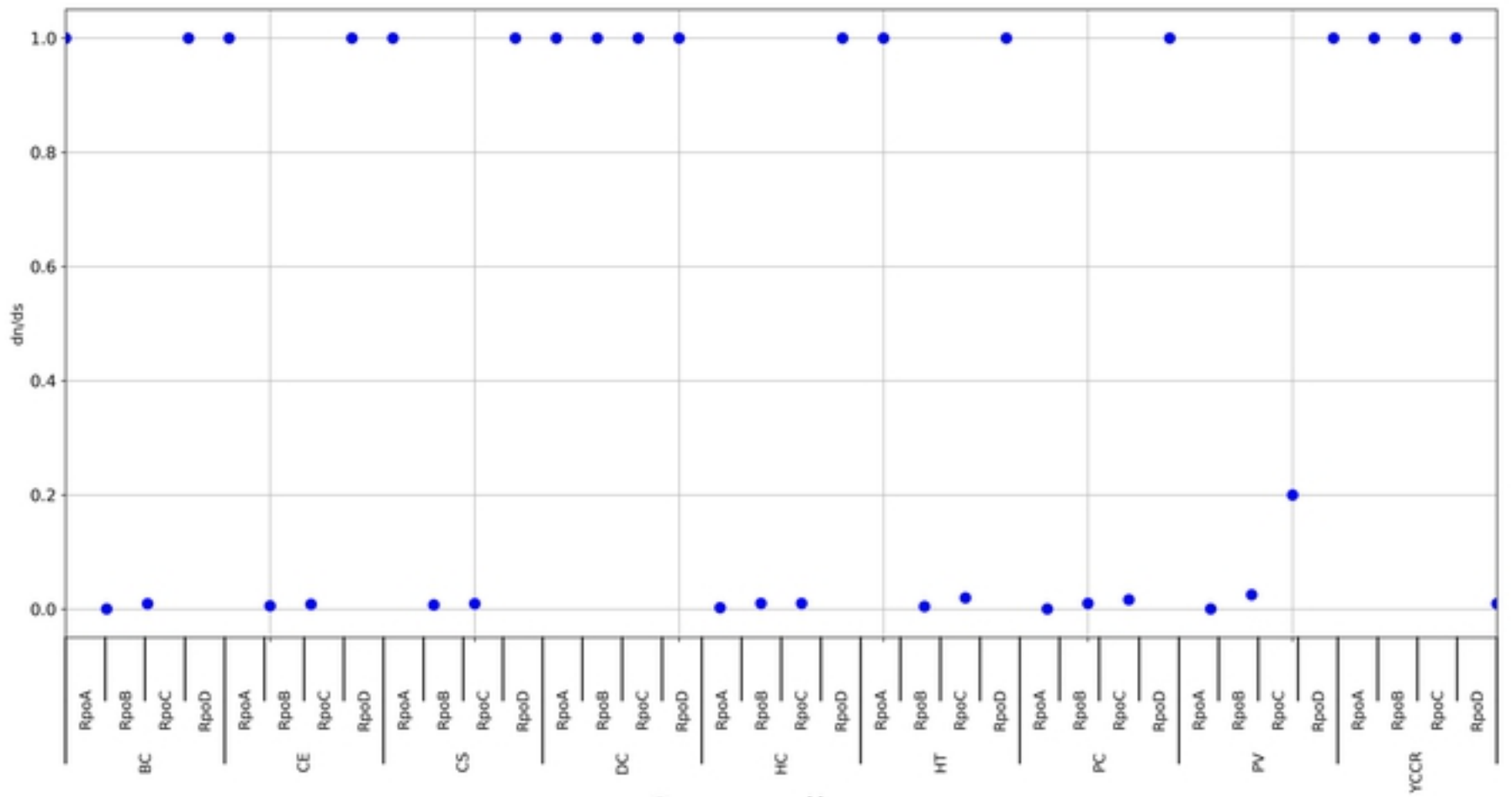
Figure 2



Hodgkinia



Tremblaya



Carsonella

Figure 5

bioRxiv preprint doi: <https://doi.org/10.1101/2020.09.07.285775>; this version posted September 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

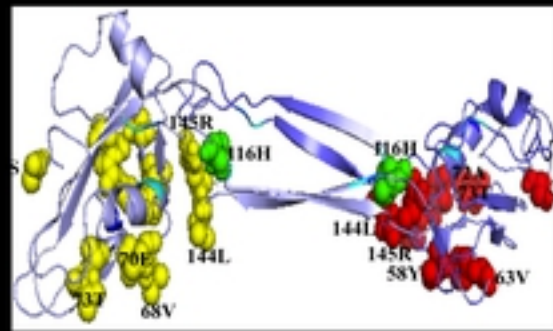
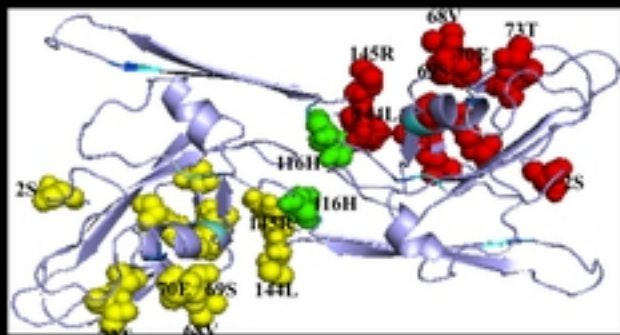
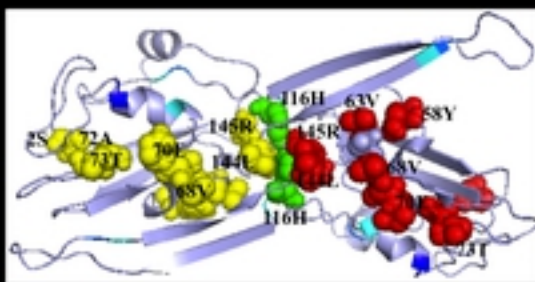
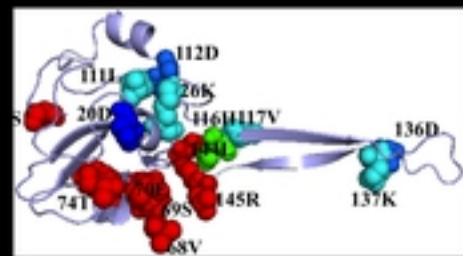
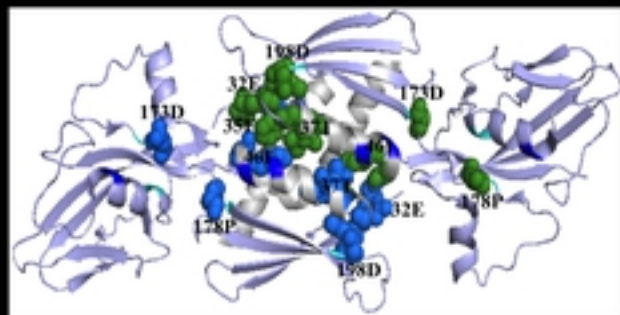
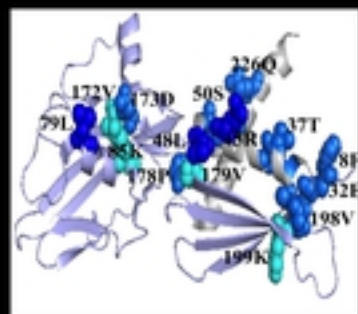
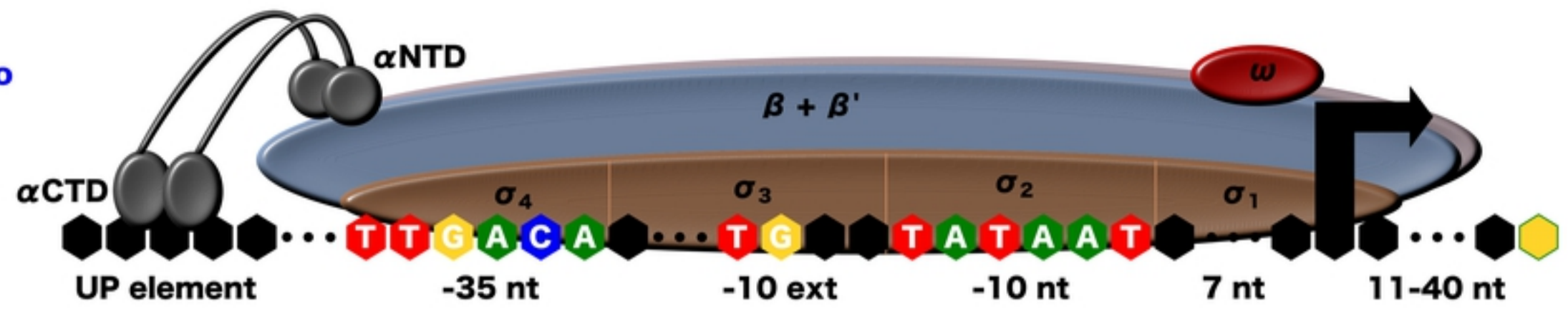


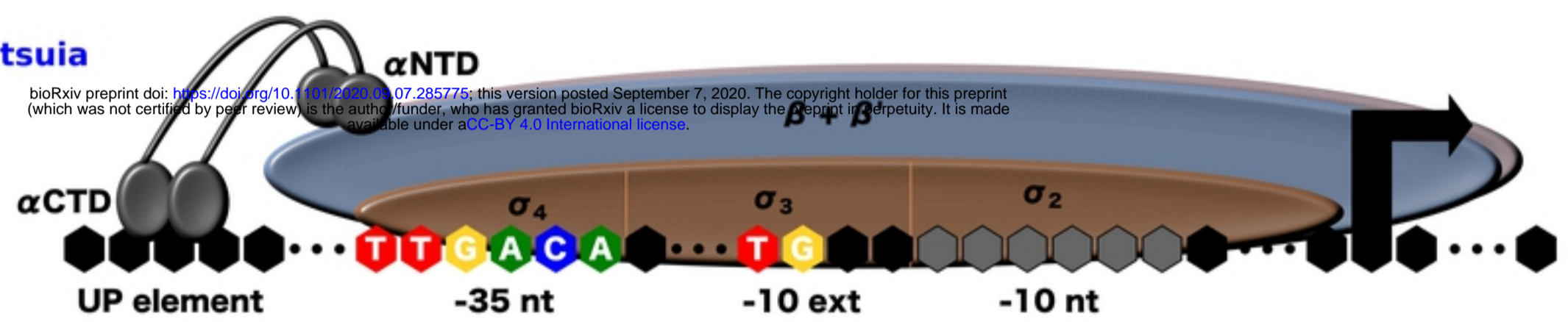
Figure 6

Eco

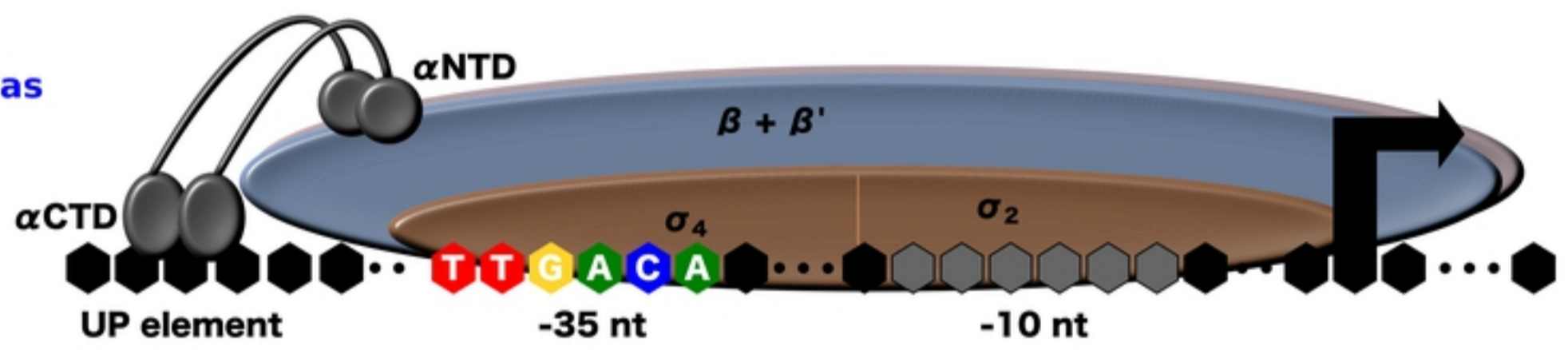


Natsuia

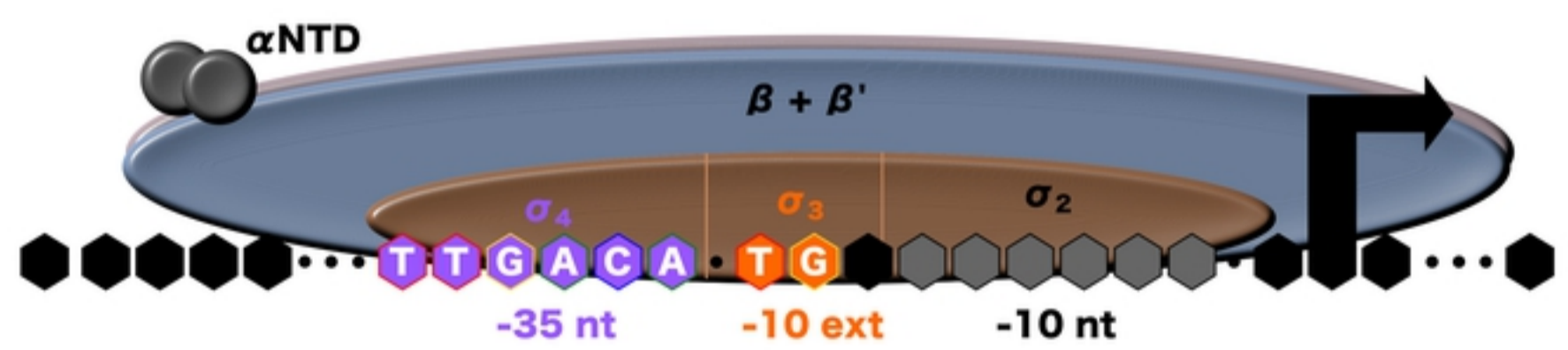
bioRxiv preprint doi: <https://doi.org/10.1101/2020.09.07.285775>; this version posted September 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



Carsonellas



Hodgkinia



Tremblaya

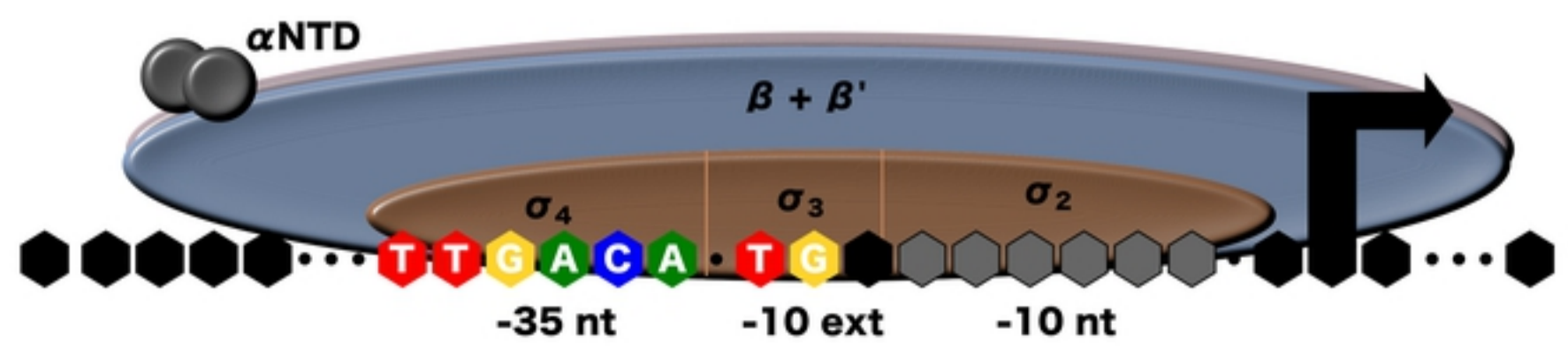


Figure 7