

## RESEARCH ARTICLE

# RNA polymerases in strict endosymbiont bacteria with extreme genome reduction show distinct erosions that might result in limited and differential promoter recognition

Cynthia Paola Rangel-Chávez<sup>1</sup>, Edgardo Galán-Vásquez<sup>2</sup>, Azucena Pescador-Tapia<sup>1</sup>, Luis Delaye<sup>3</sup>, Agustino Martínez-Antonio<sup>1\*</sup>

**1** Biological Engineering Laboratory, Genetic Engineering Department, Center for Research and Advanced Studies of the National Polytechnic Institute, Irapuato Gto, México, **2** Departamento de Ingeniería de Sistemas Computacionales y Automatización, Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, Universidad Nacional Autónoma de México, Coyoacán, Ciudad de México, CDMX, México, **3** Evolutionary Genomics Laboratory, Genetic Engineering Department, Center for Research and Advanced Studies of the National Polytechnic Institute, Irapuato Gto, México

\* [agustino.martinez@cinvestav.mx](mailto:agustino.martinez@cinvestav.mx)



## OPEN ACCESS

**Citation:** Rangel-Chávez CP, Galán-Vásquez E, Pescador-Tapia A, Delaye L, Martínez-Antonio A (2021) RNA polymerases in strict endosymbiont bacteria with extreme genome reduction show distinct erosions that might result in limited and differential promoter recognition. PLoS ONE 16(7): e0239350. <https://doi.org/10.1371/journal.pone.0239350>

**Editor:** Maria Sola, Instituto de Biología Molecular de Barcelona, SPAIN

**Received:** September 2, 2020

**Accepted:** June 22, 2021

**Published:** July 29, 2021

**Copyright:** © 2021 Rangel-Chávez et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Data Availability Statement:** All relevant data are within the manuscript and its [Supporting Information](#) files.

**Funding:** The authors received no specific funding for this work.

**Competing interests:** The authors have declared that no competing interests exist.

## Abstract

Strict endosymbiont bacteria present high degree genome reduction, retain smaller proteins, and in some instances, lack complete functional domains compared to free-living counterparts. Until now, the mechanisms underlying these genetic reductions are not well understood. In this study, the conservation of RNA polymerases, the essential machinery for gene expression, is analyzed in endosymbiont bacteria with extreme genome reductions. We analyzed the RNA polymerase subunits to identify and define domains, subdomains, and specific amino acids involved in precise biological functions known in *Escherichia coli*. We also perform phylogenetic analysis and three-dimensional models over four lineages of endosymbiotic proteobacteria with the smallest genomes known to date: *Candidatus Hodgkinia cicadicola*, *Candidatus Tremblaya phenacola*, *Candidatus Tremblaya Princeps*, *Candidatus Nasuia deltocephalinicola*, and *Candidatus Carsonella ruddii*. We found that some *Hodgkinia* strains do not encode for the RNA polymerase  $\alpha$  subunit. The rest encode genes for  $\alpha$ ,  $\beta$ ,  $\beta'$ , and  $\sigma$  subunits to form the RNA polymerase. However, 16% shorter, on average, respect their orthologous in *E. coli*. In the  $\alpha$  subunit, the amino-terminal domain is the most conserved. Regarding the  $\beta$  and  $\beta'$  subunits, both the catalytic core and the assembly domains are the most conserved. However, they showed compensatory amino acid substitutions to adapt to changes in the  $\sigma$  subunit. Precisely, the most erosive diversity occurs within the  $\sigma$  subunit. We identified broad amino acid substitution even in those recognizing and binding to the -10-box promoter element. In an overall conceptual image, the RNA polymerase from *Candidatus Nasuia* conserved the highest similarity with *Escherichia coli* RNA polymerase and their  $\sigma^{70}$ . It might be recognizing the two main promoter elements (-10 and -35) and the two promoter accessory elements (-10 extended and UP-element). In *Candidatus Carsonella*, the RNA polymerase could recognize all the promoter elements except the -10-box extended. In *Candidatus Tremblaya* and *Hodgkinia*, due to the  $\alpha$  carboxyl-terminal

domain absence, they might not recognize the UP-promoter element. We also identified the lack of the  $\beta$  flap-tip helix domain in most *Hodgkinia*'s that suggests the inability to bind the -35-box promoter element.

## Introduction

Until 2006, scientists thought that the minimum quantity of genes necessary to support life would be around 500. However, this view changed soon after when genomes from obligate endosymbiotic bacteria began to be published. Several of these genomes contained less than 500 genes, with extreme cases with less than 200 genes [1]. Some clues on how these minimal genomes managed to provide all the necessary functions to sustain life begin to emerge. For instance, transcriptome analysis of *Buchnera aphidicola* offers evidence that its genome has a limited ability to respond to environmental fluctuations [2]. Thus, gene expression in obligate endosymbiont is somehow stable and just 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 reductions [3].

DNA transcription is an essential molecular process through which organisms decode genetic information into cellular functions [4]. RNA polymerase (RNAP) is the enzyme responsible for transcribing DNA to RNA. It consists of a multi-subunit protein complex present in all living organisms, from bacteria to eukaryotes [5]. In bacteria, the RNAP is responsible for synthesizing all RNAs, including messenger, ribosomal, transfer, and small RNAs. In free-living bacteria, the RNAP holoenzyme consists of six subunits ( $\alpha_2\beta\beta'\omega\sigma$ ), encoded by five different genes (this includes two copies of the  $\alpha$  subunit). The ordered assembly of these five proteins constitutes the holoenzyme with a molecular mass of around 400 kDa. Previous studies on *E. coli* found that the  $\omega$  subunit is not essential for RNAP activity [6]. Later studies indicate that the  $\omega$  subunit is absent in all endosymbionts analyzed herein [7]. The rest of the subunits are considered essential core components of RNAP ( $\alpha_2$ ,  $\beta$ , and  $\beta'$ ). They are well conserved in bacteria [5,8]. This RNAP core is catalytically active (transcribes DNA to RNA) but cannot initiate DNA transcription by itself. Instead, the RNAP core must bind to an additional subunit called the sigma factor ( $\sigma$ ) to initiate transcription. The  $\sigma$  is responsible for recognizing and binding to gene promoters [8]. Once transcription begins, and after synthesizing a short fragment of RNA, the  $\sigma$  release and the RNAP core continue transcribing until it reaches a transcription terminator.

Since  $\sigma$  is responsible for binding and discriminating among gene promoters, it is common to find different types of  $\sigma$ . Those fall into two evolutionary families. One of these is called the  $\sigma^{54}$ , which has a single member in free-living bacteria and is absent in endosymbionts. On the contrary, the  $\sigma^{70}$  family has several copies per genome (from 1 in endosymbionts to around 60 in free-living bacteria). One member of the  $\sigma^{70}$  family is also known as the "housekeeping  $\sigma$  factor," an essential gene present in all bacteria [9]. In *E. coli*, this housekeeping gene is precisely the  $\sigma^{70}$  or *rpoD* gene. One ortholog of this gene probably should encode the housekeeping  $\sigma$  in strict endosymbionts [7]. In *E. coli*, there are seven  $\sigma$  factors encoded in the *E. coli* genome, six of them correspond to the  $\sigma^{70}$  family ( $\sigma^{70}$ ,  $\sigma^{38}$ ,  $\sigma^{32}$ ,  $\sigma^{28}$ ,  $\sigma^{24}$  and  $\sigma^{19}$ ), and the remaining one corresponds to  $\sigma^{54}$ . The architecture of transcription units and the consensus promoter sequences for each  $\sigma$  in *E. coli* was proposed previously [10].

### A brief functional description of the RNAP essential subunits in *E. coli*

**The  $\alpha$  subunit.** In *E. coli*, *rpoA* encodes for the RNAP  $\alpha$  subunit. This protein has a molecular weight of ~37 kDa with two folded domains ( $\alpha$  carboxy and  $\alpha$  amino-terminal domains,

also known as  $\alpha$ -CTD and  $\alpha$ -NTD) connected by two flexible linkers [11]. This subunit performs three biological functions: i) it initiates the assembly of the RNAP complex through the interaction of its  $\alpha$ -NTD with the  $\beta$ - and  $\beta'$ -subunits; ii) it participates in promoter recognition through the interaction of its  $\alpha$ -CTD with the DNA UP promoter element, and iii) their  $\alpha$ -CTD is also a target for the binding of many transcriptional activators. These transcription factors are bound in the following architectures: 1) where the  $\alpha$ -CTD situates between a bound activator and the rest of the RNAP, 2) the  $\alpha$ -CTD located upstream of a bound activator; 3) the  $\alpha$ -CTD is flanking a bound activator, and 4) combinations of the above arrangements in the presence of multiple activators [12–18]. For instance, dimers of the cAMP receptor protein (dual transcriptional regulator CRP) interact with one of the two  $\alpha$ -CTD [14,15]. In addition, the integration host factor (IHF) interacts with  $\alpha$ -CTD to activate XylR [16]. Also, FIS activates transcriptionally through the  $\alpha$ -CTD that is  $\beta'$ -associated, linked to the DNA on the promoter distal side of FIS [17].

**The  $\beta$  and  $\beta'$  subunits.** *rpoB* and *rpoC* encode for the  $\beta$  and  $\beta'$  subunits of RNAP in *E. coli*. The  $\beta$  subunit forms a pincer, called "clamp." The  $\beta'$  subunit constitutes the other pincer. In between give place to a 27 Å-wide internal channel, where the catalytic site of the RNAP enzyme is located [18]. The  $\beta$  subunit binds to the  $\alpha$  and  $\beta'$  subunits through 24 and 73 amino acids distributed in the region between the amino acids 540 to 1340 [18]. The interaction of  $\sigma$  to the RNAP core is via the  $\beta$  subunit also, through their  $\beta$  flap-tip helix domain [19].

In the  $\beta'$  subunit, the amino acids present in the region from 917 to 1361 are essential for their interaction with the  $\beta$  subunit. And their  $\beta'$  coiled-coil domain for their interaction with the  $\sigma$  factor. Additionally, in the  $\beta'$  amino-terminal domain, several functional amino acids, typical of Zinc fingers, form the RNAP catalytic site and stabilizes the RNAP-DNA complex. The active site is composed of around 500 amino acids retaining two  $Mg^{2+}$  ions. However, only thirteen amino acid residues are the most conserved conforming to the catalytic site containing three aspartic amino acids. On the other hand, in the  $\beta'$  CTD domain, three polar residues and the G-loop domain conform to the cavity where the DNA fits and contact the RNAP [20].

Once assembled the RNAP core ( $\alpha_2\beta\beta'$ ),  $\sigma$  joins it by interacting with the  $\beta$  and  $\beta'$  subunits. First,  $\sigma$  binds to  $\beta'$  through the  $\beta'$  coiled-coil domain. Such binding places  $\sigma$  to contact the -10-box element to form the open promoter complex [21]. Likewise, the  $\beta$  subunit binds to the  $\sigma$  through the  $\beta$  flap-tip helix domain at the  $\sigma_4$  region (see below). This contact gives  $\sigma$  the capacity to adapt to variation in nucleotide distances between the -10 and the -35-box promoter elements [19].

**The  $\sigma$  subunit.** The  $\sigma$  factor is not a permanent component of the RNAP core but a transitory-associated subunit. The  $\sigma$  factor is necessary for promoter recognition and transcription initiation. In *E. coli*,  $\sigma^{70}$  consists of a protein with four helical domains ( $\sigma_1$ ,  $\sigma_2$ ,  $\sigma_3$ , and  $\sigma_4$ ). Each of them interacts with different promoter elements and domains of the different RNAP core subunits. The  $\sigma_1$  domain prevents  $\sigma^{70}$  from interacting with the DNA strand without a complete RNAP core [22,23]. The  $\sigma_2$  domain is the most conserved of the  $\sigma^{70}$  family and consist of four subdomains:  $\sigma_{2.1}$ ,  $\sigma_{2.2}$ ,  $\sigma_{2.3}$ , and  $\sigma_{2.4}$ . The  $\sigma_{2.1}$  and  $\sigma_{2.2}$  subdomains are involved in the binding to the RNAP core [24]. The  $\sigma_{2.2}$  subdomain contains sites for binding the  $\beta'$  coiled-coil domain. In contrast, the  $\sigma_{2.3}$  subdomain participates in DNA melting. It has seven conserved aromatic amino acids, whose replacement results in defects in DNA melting [25,26]. The specific recognition of the -10-box promoter element is a function of the  $\sigma_{2.4}$  subdomain. The amino acids involved in DNA melting and binding of the -10-box promoter element interact on the same DNA helices' faces. Deletion analyses have determined that the  $\sigma_2$  domain seems necessary for the correct functioning of the RNAP [26,27]. The  $\sigma_3$  region interacts with the -10-box extended promoter element. It stabilizes the short nascent DNA-RNA

hybrids during the early stages of gene transcription [28] because of their interaction with the 5'-triphosphate of the nascent RNA [29]. Finally, the  $\sigma_4$  domain is formed by two subdomains,  $\sigma_{4.1}$  and  $\sigma_{4.2}$ . The  $\sigma_{4.1}$  subdomain interacts with the  $\beta$  flap-tip helix of the  $\beta$  subunit to allow their correct binding to the -35-box promoter element. More specifically, the amino acids R518 and R516 in  $\sigma_{4.2}$  recognize the guanine and cytosine at the -34 and -32 nucleotide positions concerning the transcription start (+1) [27]. Additionally, the  $\sigma_{4.1}$  subdomain is a point of contact with transcriptional activators that bind upstream of the -35-box promoter element. One study reports that the binding of the  $\sigma_{4.1}$  subdomain of  $\sigma^{38}$  with the  $\beta$  flap-tip helix domain is more potent than with this subdomain in  $\sigma^{70}$  [30].

## Endosymbiotic bacteria

Endosymbiotic bacteria live inside other organisms (usually eukaryotes). These show genomic features resulting from several million years of co-evolution [31]. Strict endosymbiotic bacteria cannot outside their host and have lost most of their genes and large fragments of amino acids in their remaining proteins [7,10]. The comparison of *E. coli* genome, with approximately 4.5 thousand genes to the endosymbiotic bacterium *Candidatus Carsonella ruddii*, which has retained just around 150 genes, exhibits this dramatic process [32,33]. In addition, almost all genes found in *Candidatus Carsonella ruddii* are considerably shorter than their free-living orthologues.

Furthermore, genes encoding proteins with multiple domains in obligate endosymbionts commonly have lost some regions or complete domains. These, in some cases, are essential for their activity in free-living bacteria [34,35]. For instance, in previous work, we determined that obligate endosymbionts had lost all the transcription factors that interact with the promoters and RNAP to activate or inhibit gene expression [36]. Additionally, partial loss of the  $\alpha$  and  $\sigma$  subunits is present in *Candidatus Hodgkinia* sp. and *Candidatus Carsonella ruddii* [35,37]. Both are considered essential components of the RNAP in free-living bacteria.

Gene erosion of the transcriptional machinery in endosymbionts with highly reduced genomes raises the critical question of how gene transcription could be happening in these bacteria. Here, we address this question from the perspective of comparative genomics. With this purpose, we investigated the RNAP subunits in four bacterial lineages of obligate endosymbionts: *Candidatus Hodgkinia cicadicola*, *Candidatus Tremblaya phenacola*, and *princeps*, *Candidatus Nasuia deltocephalinicola*, and *Candidatus Carsonella ruddii*, which exhibit extreme genome-reduction, within the proteobacteria phylum.

## Materials and methods

### Genome data of selected endosymbiont

To recover sequenced genomes, we use the NCBI database (<http://www.ncbi.nlm.nih.gov/>). In this database, we identified 37 endosymbionts with extreme genome reduction. All these pertain to the four lineages of proteobacteria. *Candidatus Hodgkinia cicadicola* belongs to  $\alpha$ -proteobacteria. *Candidatus Tremblaya phenacola*; *Candidatus Tremblaya princeps*; and *Candidatus Nasuia deltocephalinicola*, which belong to  $\beta$ -proteobacteria. Moreover, *Candidatus Carsonella ruddii*, which belongs to  $\gamma$ -proteobacteria. The genomes of these bacteria are completely sequenced. Further characteristics of these genomes are in the supplementary material [S1 Table](#).

### Amino acid alignment of the RNAP subunits in strict endosymbiotic bacteria and their comparison to *E. coli*

We compared the amino acid sequences for each RNAP subunit against their corresponding *E. coli* orthologous. We use the T-Coffee program in a multiple sequence alignment with

standard parameters [38]. For simplicity, the amino acid positions will from now on be referred to their location in the corresponding *E. coli* protein or subunit, accompanied by the abbreviation "ECO." With this alignment, we define domains, subdomains, and relevant amino acids. We also utilized the Pfam database [39], the proteins superfamily classification [40], and the NCBI's conserved domains [41]. In addition, we did bibliographic research to gather relevant information regarding the sites and amino acid regions for each of the RNAP [11–13,18–21,23–28,30,42–45].

### Structural analysis of the RNAP subunits and inference of their 3-D structural-functional models

We first recreated 3-D structural models for each RNAP subunit using the I-TASSER server with standard parameters [46]. I-TASSER generates 3-D models for a given sequence by collecting high-score structural templates from PDB (Protein Data Bank) with full-length atomic models constructed by iterative template-based fragment assembly simulations. The structural model of *Candidatus* Hodgkinia TETUND2 and Dsem strains was obtained from their homologous proteins in *E. coli* by using the multiple threading alignments.

Then we superpose the resulting 3-D models with the crystal structure of the RNAP ECO (4YLP) [47]. Graphic representations of each structure were prepared with the PyMOL Molecular Graphic System software version 1.3 [48]. For the interaction of RNAP with the promoter sequence, we used the predicted models of the RNAP subunits obtained by I-TASSER for *Candidatus* Hodgkinia Dsem and TETUND2. We did a structural alignment with the homologous structure of the holoenzyme RNAP of *E. coli* (ECO 4YLP) in PyMOL.

We find that *Candidatus* Hodgkinia TETUND2 has lost a fragment of the  $\alpha$ -NTD involved in dimer formation. To investigate the possible dimer association of  $\alpha$  subunit monomers in this bacterium, we used ClusPro v.2.0 [49–52]. This tool is an automatic protein docking tool based on CAPRI (Critical Assessment of Predicted Interactions) [50,53]. As a result, we get three models of  $\alpha$  dimer subunits with different cluster sizes. We map sites under putative positive selection with these models and evaluate free-energy changes in the protein-protein interactions. With this strategy, we recreated single mutations along with the amino acids positively selected using the BindProfX server [54]. Following this, we exchanged the putative positive selected amino acid on the dimers predicted in the different strains of Hodgkinia. Then we 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  $\Delta G$  is, the more stable the complex results. Finally, we calculate the effect of mutations on binding affinity by the differences in free-energy changes between the mutant and the wild type  $\Delta\Delta G_{\text{wt} \rightarrow \text{mut}} = \Delta G_{\text{mut}} - \Delta G_{\text{wt}}$ . The criteria to consider a strongly favorable mutation was to have  $\Delta\Delta G \leq -1 \text{ kcal/mol}$ .

### Natural selection analysis of the RNAP subunits from endosymbionts

To understand the putative mechanisms of RNAP subunits' molecular evolution, first, we need to know the selective pressure acting on each of the subunits of this protein. For this purpose, we estimated the  $D_N/D_S$  ratio ( $\omega$ ) on the protein-coding sequences studied here. The ratio considers non-synonymous substitutions per non-synonymous sites ( $D_N$ ), divided by the number of synonymous substitutions per synonymous sites ( $D_S$ ). The  $D_N/D_S$  ratio can result in three evolutionary processes: i) if  $D_N/D_S < 1$ , we infer purifying selection; ii) if  $D_N/D_S > 1$ , then we infer positive selection; and iii) if the  $D_N/D_S = 1$ , it indicates a neutral evolution [55]. We perform these selection analyses to RNAP subunits in each group of bacteria.

Natural selection was inferred with CodeML from PAML v.4.6 package [56]. This software requires codons alignment. For those, we used PAL2NAL v.2.1.0 program [57]. To graph the phylogenetic trees, we use PhyML [58]. We previously aligned the set of amino acid sequences of each subunit with T-Coffee [38]. We used Gblocks to recover the informative codons [59]. To identify specific genes and amino acids under positive selection, we use branch and branch-site models. In the case of a branch, we use three models: "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. The  $D_N/D_{S2}$  assumes that the branch of interest (foreground branch) has a  $D_N/D_{S2}$  ratio different than the background ratio [60].

The level of significance for the Likelihood Ratio Test (LRT) was estimated using the  $\chi^2$  distribution with degrees of freedom (*df*). These degrees of freedom are equal to the difference in the number of parameters between the models. The statistic considers 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 [61]. We compared one-ratio and free-ratio models to know whether  $D_N/D_S$  were different among the lineages. In contrast, we examine whether the lineage of interest has a different ratio than the other lineages with one-ratio and two-ratio models.

We approached a model where the  $D_N/D_S$  ratio was 1, 0.2, and 1.2 for the foreground branch to detect positive or negative selection in specific lineages. First, we compared  $D_N/D_{S2}$  against the  $D_N/D_S = 1$ , where the null hypothesis is that models are not significantly different. Suppose the null hypothesis is rejected ( $p < 0.05$ ) and the two-ratio model is greater than 1. In that case, 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 evolution. Additionally, we compared  $D_N/D_{S2}$  against the  $D_N/D_S = 0.2$ ; the null hypothesis is that models are not significantly different. For example, suppose the null hypothesis is ( $p < 0.05$ ), and the two-ratio model estimate is more significant 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 more substantial negative selection. Finally, when  $D_N/D_{S2}$  is greater than one and is significantly different from  $D_N/D_S = 1$ , it means positive selection. To get more evidence about the likely positive selection, we compared  $D_N/D_{S2}$  against  $D_N/D_S = 1.2$ ; the null hypothesis was a non-significant difference between  $D_N/D_{S2}$  and  $D_N/D_S = 1.2$ . Therefore, accepting the null hypothesis indicates that the foreground branch is possibly under positive selection. At the same time, a rejection means that the foreground branch might be subjected to a relaxed selection.

We performed a branch-site test for positive selection to identify individual codons under positive selection along specific branches [62]. In these models, positive selection was allowed on a particular "foreground" branch. We compared the LRTs ( $df = 1$ ) against null models that assume no positive selection is happening. This test results in 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. For sites in classes 2a and 2b, positive selection is allowed on the foreground branches ( $D_N/D_{S2} > 1$ ). For the rest, the "background branches" are under purifying selection ( $0 < D_N/D_{S0} < 1$ ) and neutral evolution ( $D_N/D_{S1} = 1$ ). For the null model,  $D_N/D_{S2}$  is 1. We test all the RNAP subunits in each endosymbiont in these ways. Each branch is considered as the foreground to reconstruct the phylogenies. We compared the two models using LRT. The calculus of significance between the models was twice the log-likelihood difference following an  $\chi^2$  distribution. With a *df* number equal to the difference of the number of parameters between the models. Positively selected amino acids were identified

based on Empirical Bayes, and posterior probabilities were employed in CodeML [63]. We did not test the *Nasuia* RNAP subunits because there were only three sequenced strains. However, CodeML requires at least 4 to get reliable results.

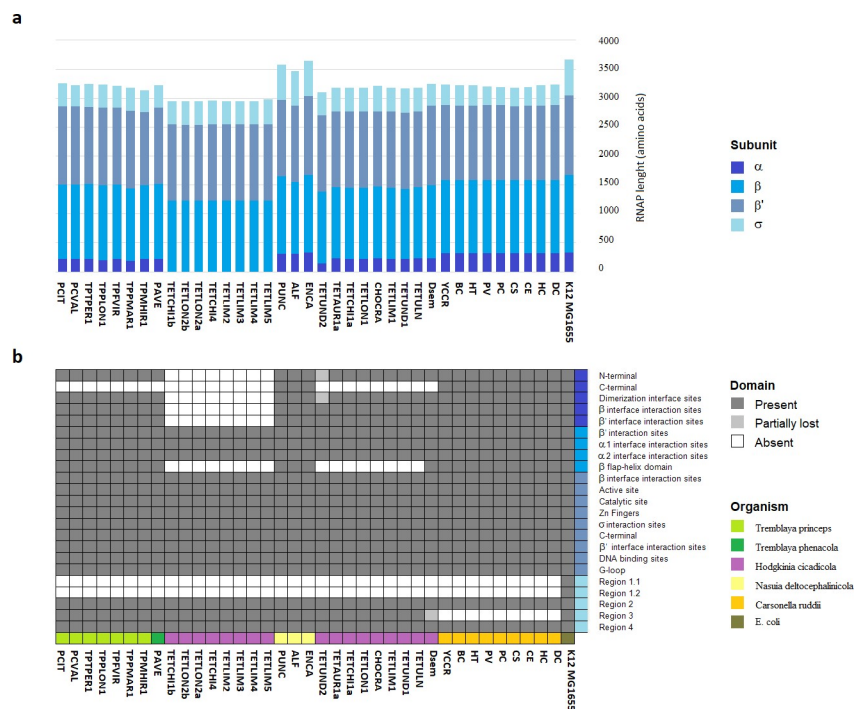
### Results

To study the evolution of RNAP subunits in genomes exhibiting extreme reduction, first, we identified orthologous to the *E. coli* RNAP in the 37 obligate endosymbiotic bacteria (Fig 1A and S1 Table). The initial genomes included nine strains of *Candidatus Carsonella ruddii*, seventeen of *Candidatus Hodgkinia cicadicola*, only one of *Candidatus Tremblaya phenacola*, seven of *Candidatus Tremblaya princeps*, and three of *Nasuia deltocephalinicola*.

Nine of the seventeen *Hodgkinia* strains (52%) lack  $\alpha$  subunits but were considered in the remainder analyses for the rest of the RNAP subunits. The rest (28 genomes) conserve orthologous genes for each of the  $\alpha$ ,  $\beta$ ,  $\beta'$  subunits, as well as a single gene coding for the  $\sigma$  factor. The total amino acid sequences encoding for each of the RNAP subunits exhibit a reduction of 16% on average, compared to those in *E. coli*. Also, we observed that these genes had lost DNA regions encoding important functional protein domains in *E. coli*. In some cases, with the loss of total domains (see below) (Fig 1B). In the following sections, we describe the structure and amino acid diversity found in each of the RNAP subunits of these endosymbiotic bacteria.

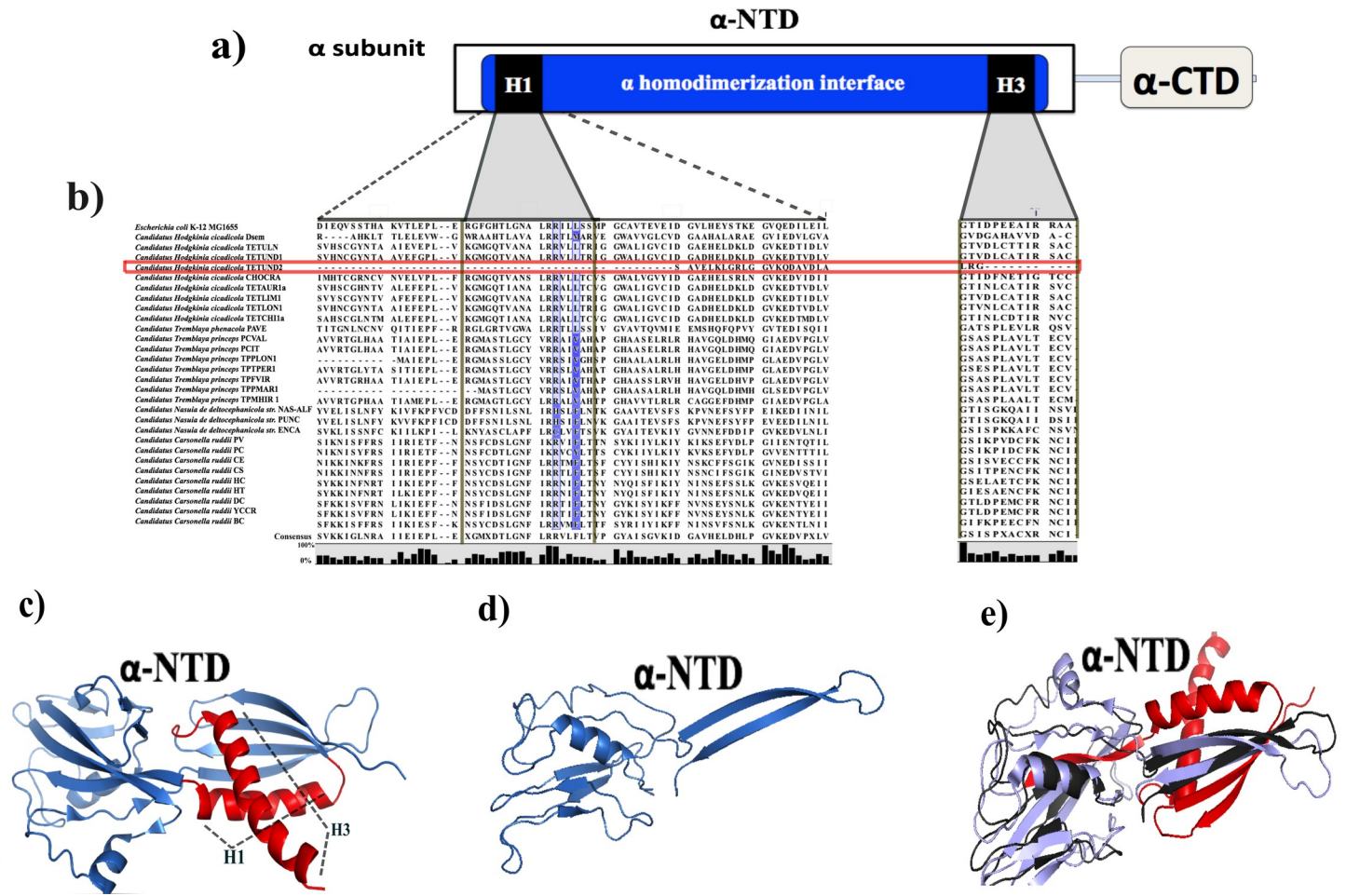
### The $\alpha$ -NTD is more conserved than $\alpha$ -CTD in the $\alpha$ subunit

In *Carsonella* and *Nasuia*, their  $\alpha$  subunits conserved all the functional domains known in *E. coli*. At the same time, *Hodgkinia* and *Tremblaya* mostly retain the  $\alpha$ -NTD (for self-homodimerization



**Fig 1. Conservation of RNAP subunits in endosymbiotic bacteria showing extremely reduced genomes.** a) The bars represent the size (in amino acids) of RNAP subunits found in each endosymbiont; the different blue colors represent the relative size contribution of each RNAP subunits. b) Colors squares represents the degree conservation of functional domains in the RNAP subunits of *Candidatus Tremblaya princeps* (lime green), *Candidatus Tremblaya phenacola* (dark green), *Candidatus Hodgkinia cicadicola* (magenta), *Candidatus Nasuia deltocephalinicola* (yellow), *Candidatus Carsonella ruddii* (orange), and *E. coli* (olive green).

<https://doi.org/10.1371/journal.pone.0239350.g001>



**Fig 2. Domains and amino acid conservation in the  $\alpha$ -subunit.** a) The white and grey rectangle represents the  $\alpha$ -NTD and  $\alpha$ -CTD. Dark blue represents the  $\alpha$  subunit dimerization region. The regions to form the H1 and H3 helices are shown in black rectangles. b) The amino acid alignment of the  $\alpha$  subunits shows the conserved, functional amino acids involved in the dimer formation and the H1 and H3  $\alpha$ -helices. Therefore, darker backgrounds are showing those amino acids that differ from (*E. coli*). c) 3-D *E. coli*  $\alpha$ -NTD (4YLP) crystal structure shows red regions absent in *Candidatus* Hodgkinia TETUND2. d) The predicted 3-D model obtained for the  $\alpha$  subunit of Hodgkinia TETUND2. e) Structural comparison between the *E. coli*  $\alpha$ -NTD (4YLP) crystal structure (violet) and the predicted 3-D model obtained for the  $\alpha$  subunit of Hodgkinia TETUND2 (black). In red, it shows the  $\alpha$ -NTD regions absent in the  $\alpha$  subunit of Hodgkinia TETUND2.

<https://doi.org/10.1371/journal.pone.0239350.g002>

and the interaction with  $\beta$  and  $\beta'$  subunits) (Fig 2A and S1 Fig). *In vitro* and *in vivo* experiments revealed that the  $\alpha$ -NTD is essential for RNAP to get basal transcription [64,65]. In addition, studies have shown that the  $\alpha$ -CTD is not necessary for RNAP assembly and basal transcription. However, the  $\alpha$  subunit requires  $\alpha$ -CTD to interact with the UP-promoter elements and transcriptional activators in *E. coli* [66,67]. The loss of the  $\alpha$ -CTD is also present in *Parcubacteria*. These are ecto-symbiont bacteria that live in mixed groups [68]. Lack of  $\alpha$ -CTD is also present in microalgae chloroplasts [69].

Two *Tremblaya princeps* strains, TPPLON1 and TPPMAR1, showed incomplete  $\alpha$ -NTD. However, they preserved the essential regions for homodimerization and those for interaction with the  $\beta$  and  $\beta'$  subunits. A particular case is *Candidatus* Hodgkinia cicadicola TETUND 2. This strain conserves just the region of  $\alpha$ -NTD for interaction with the  $\beta'$  subunit and only some amino acids for homodimerization (Fig 2B, red line). It is necessary to mention that the  $\alpha$ -NTD consists of two well-conserved subdomains. Subdomain 1 contains two orthogonal  $\alpha$ -helices (H1 and H3) called the homodimerization region (Fig 2C). Subdomain 2 includes the



interfaces for interactions with the  $\beta$  and  $\beta'$  subunits [41–43]. The first step towards RNAP core formation is the homodimerization between two monomers of  $\alpha$  subunits. This dimer proceeds by the interaction of H1 and H3 helices in the subdomains 1 of each monomer. The 3-D predicted model of this subunit of *Candidatus* Hodgkinia TETUND2 shows that it does not conserve the H1 and just some amino acids of H3 helices and other motifs necessary to form the dimers interface (Fig 2D and 2E). Based on *E. coli*, we cannot infer if homodimerization is happening in the  $\alpha$  subunits of *Candidatus* Hodgkinia TETUND2. The *rpoA* is not the unique case of genes losing a significant fragment in *Candidatus* Hodgkinia *cicadicola* TETUND 2. The DNA gene that encodes the  $\epsilon$  subunit of DNA polymerase III has also lost large fragments. In such a way that it is no more considered a functional protein [33]. Finally, rest the nine Hodgkinias strains that lack a complete  $\alpha$  subunit. The authors who reported these genomes say these bacteria were the most prevailing among several coexisting strains, all with fragmented genomes [66]. These nine bacteria conserve the other RNAP subunits, but it is difficult to infer if their RNAP remains functional.

### Strict endosymbionts conserve the $\beta$ and $\beta'$ subunits except for the $\beta$ flap-tip helix domain in Hodgkinias

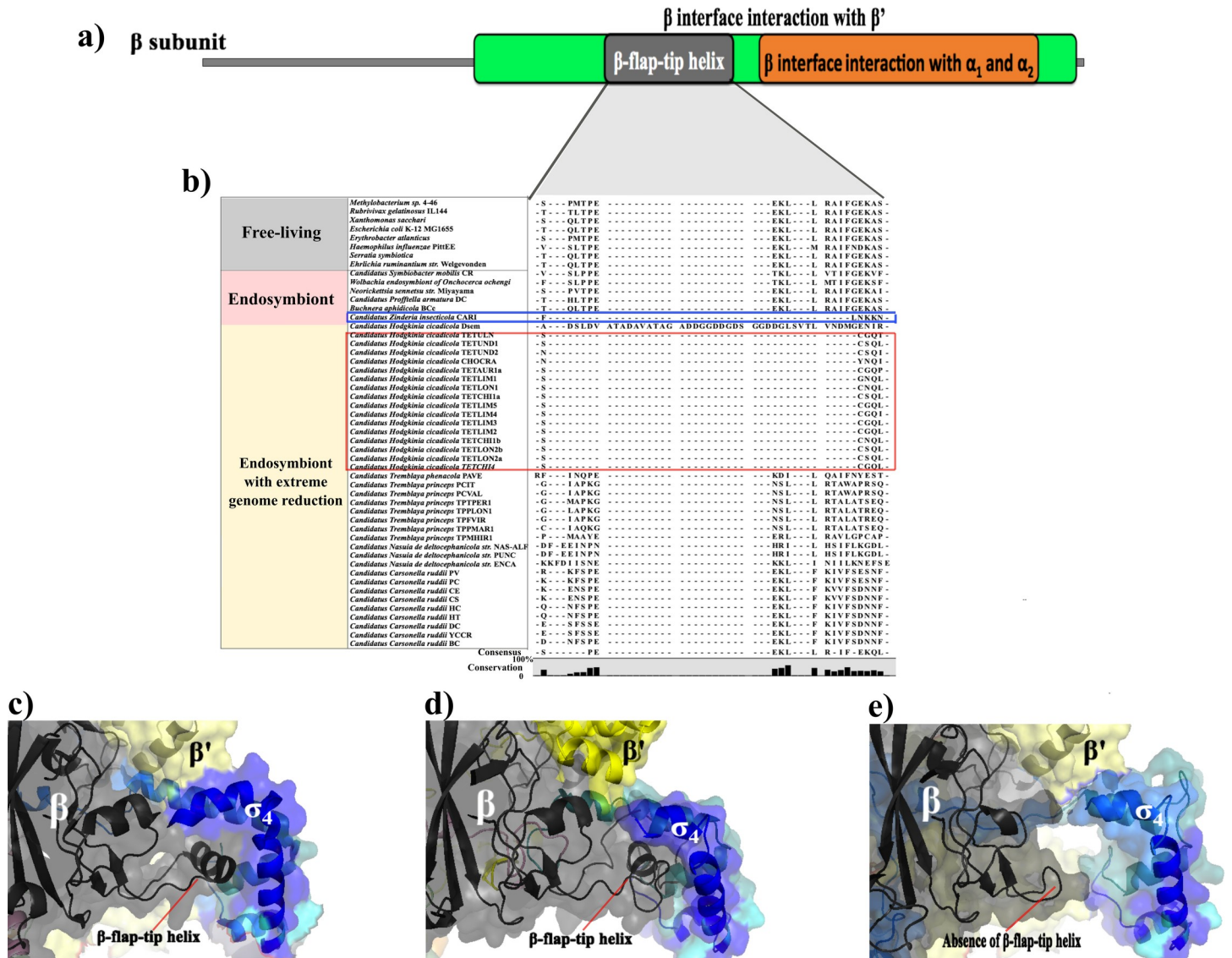
We identified that the  $\beta$ - and  $\beta'$ -subunits are the most conserved among all the RNAP subunits in these endosymbionts (Figs 3A and 4A). The reason could be their critical role in the RNAP complex formation and activity. Furthermore, all the endosymbionts preserve the catalytic core and its assembly domains within the  $\beta$ - and  $\beta'$ -subunits. Nevertheless, these present some changes in the domains involved in the binding with the  $\sigma$  and with other RNAP core subunits (S2 and S3 Figs).

The  $\beta$  flap-tip helix domain is incomplete in most *Hodgkinia* strains (Fig 3B, red box). This loss is evident in the comparisons of *Candidatus* Hodgkinia Dsem and TETUND2 with the 3-D crystal structure of the  $\beta$ - $\sigma$  subunits complex in *E. coli* (Fig 3C–3E). Unlike *Candidatus* Hodgkinia Dsem, we can observe that TETUND2 does not present a complete  $\beta$  flap-tip helix domain (Fig 3D and 3E). These changes in the  $\beta$  flap-tip helix suggest that the RNAP core in these *Hodgkinia* cannot bind to the  $\sigma_4$  domain. Consequently, the  $\sigma$  factor should not bind to the -35-box promoter element properly. Mutants in *E. coli* lacking the  $\beta$  flap-tip helix result in an inability of the  $\sigma$  subunit to attach to the -35-box promoter without affecting the RNAP core to bind to DNA. Furthermore, these mutants adequately recognize the -10-box and the -10-extend promoter elements [19]. The absence of the  $\beta$  flap-tip helix domain, although not observed in bacteria, is common in archaea [70].

*Candidatus* Tremblaya phenacola PAVE and Carsonella ruddii CE conserve the  $\beta'$  coiled-coil domain in the  $\beta'$  subunit. In contrast, the rest of the endosymbionts display substitutions in at least one of the three necessary amino acids in the  $\beta'$  coiled-coil (Fig 4B in brown, and S3 Fig). *In vitro* studies involving single amino acid substitutions in the  $\beta'$  coiled-coil domain in the three residues ECO: R275Q, E295K, and A302D result in a deficient holoenzyme formation and a subsequent lack of promoter specificity [26]. Unlike Hodgkinia Dsem, the rest of Hodgkinias strains show the same substitutions in the two positions ECO: 275 and 302, and a deletion in the residue ECO 295. Furthermore, we observed that these substitutions could not affect the  $\beta'$  coiled-coil domain formation in the re-created 3-D structures. However, the exposed residues and the orientation for the interaction with the  $\sigma_2$  domain are different from those in the *E. coli*  $\beta'$  coiled-coil domain (Fig 4C–4E).

### The $\sigma$ subunit shows the most erosive evolution in these endosymbionts

$\sigma$  is the subunit that exhibits the most differentiated conservation among endosymbionts with extreme genome reduction (Fig 5A and 5B). In the case of *Candidatus* Tremblaya phenacola

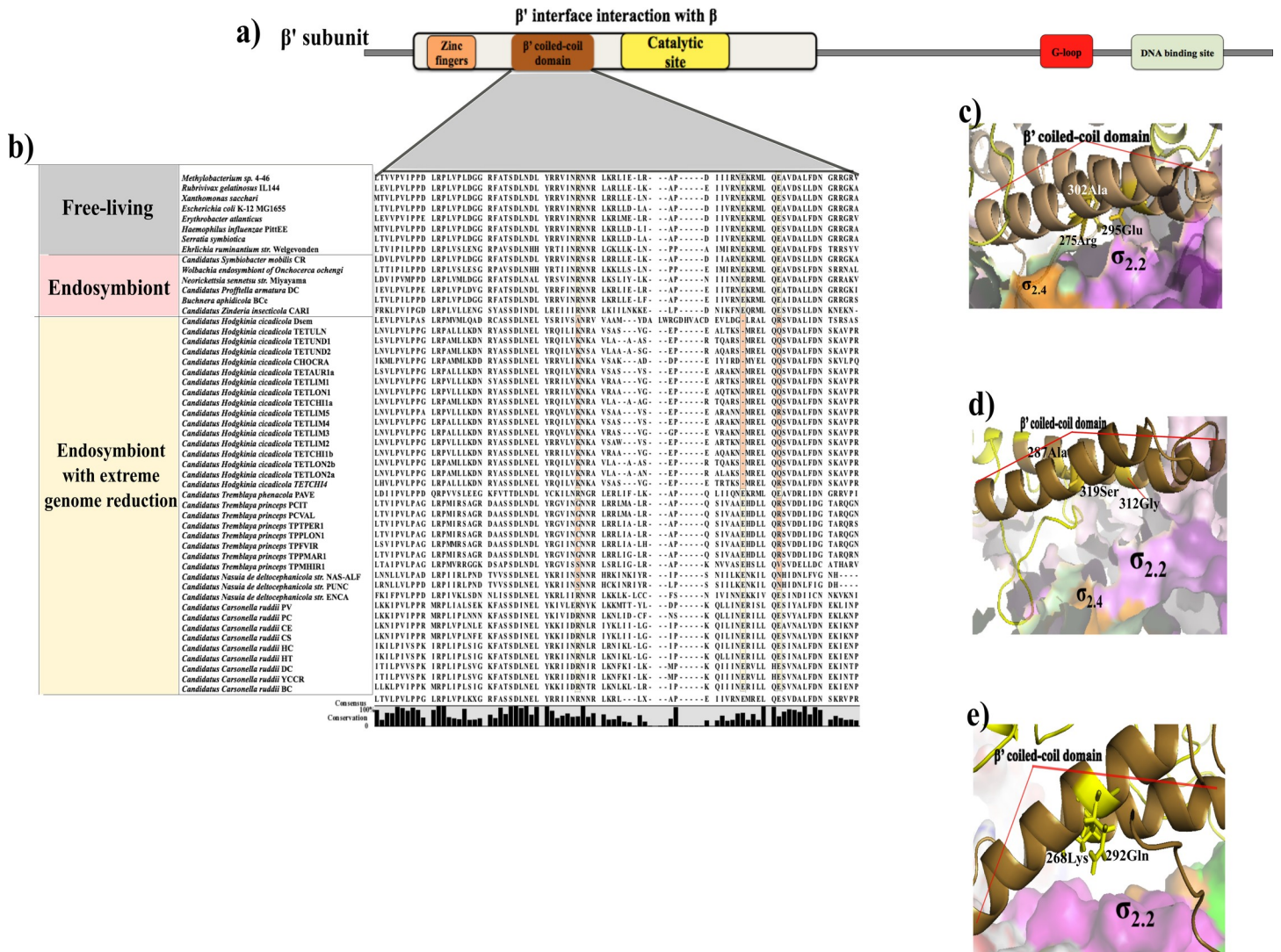


**Fig 3. Amino acids sequence conservation in the  $\beta$  subunit.** a) The upper figure represents the structural domains of the  $\beta$  subunit in *E. coli*. Besides the  $\beta$  flap-tip helix domain (grey region), the interaction regions with the two  $\alpha$  and the  $\beta'$  subunits (orange and pea-green). b) Amino acid alignment shows that most Hodgkinias lost the  $\beta$  flap-tip helix domain (red box). Also observed in *Candidatus Zinderia cicadicola* (blue box). c) Crystallographic structure of the *E. coli*  $\beta$  flap-tip helix and their interaction with  $\sigma_4$  (4YLP grey and blue). d) Predicted 3-D model for the  $\beta$  and  $\sigma$  subunits of Hodgkinia TETUND2 shows the interaction between the  $\beta$  flap-tip helix (grey) and the  $\sigma_4$  subdomain (blue). e) The predicted model for the  $\beta$  and  $\sigma$  subunits of Hodgkinia TETUND2 shows that, like the rest of Hodgkinias, the  $\beta$  flap-tip helix (grey) is not present. As a result, the interaction between the  $\beta$  subunit and the  $\sigma_4$  subdomain (blue) might be deficient.

<https://doi.org/10.1371/journal.pone.0239350.g003>

PAVE, it conserves whole the  $\sigma_2$  and  $\sigma_4$  domains. On the other side, *Candidatus Hodgkinia*, *Nasuia deltocephalinicola*, and *Carsonella ruddii* conserve the  $\sigma_4$  and, to a lesser extent, the  $\sigma_2$  domain (Fig 5B). The main variations happen inside the  $\sigma_2$  domain, whose amino acids interact with the -10-box promoter element and define the promoter specificity (Fig 5B, magenta, and orange amino acids).

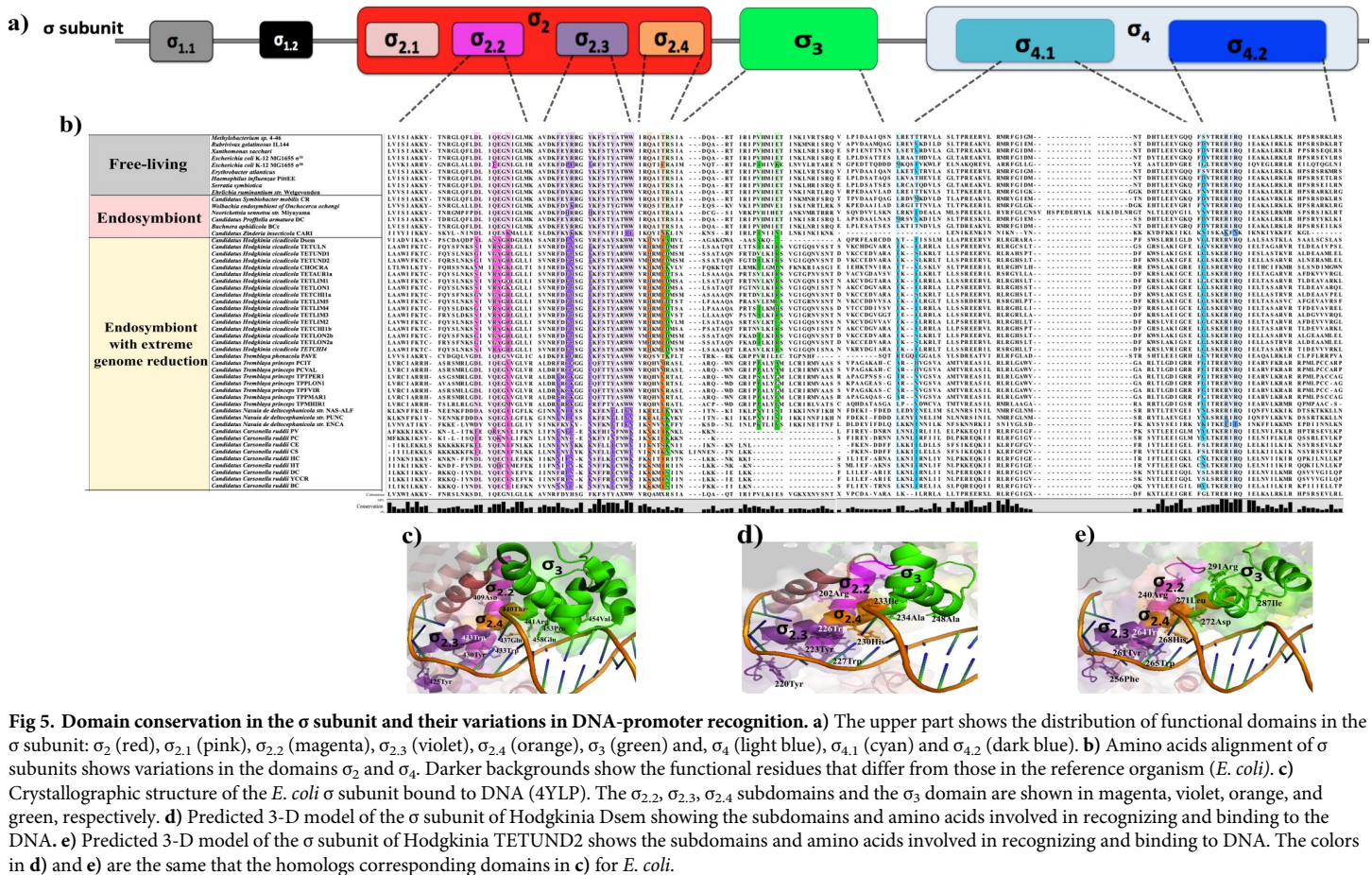
For the  $\sigma_{2,2}$  subdomain, the Hodgkinia strains exhibit substitutions in all the four amino acids involved in the interaction with the  $\beta'$  coiled-coil domain. The seventeen strains have the same substitutions for ECO E407A and 14 of them in ECO N409R. At the same time, each presents different amino acids at the positions ECO: 403 and ECO 406 (Fig 5B, magenta).



**Fig 4. Amino acids sequence conservation in the  $\beta'$  subunit.** a) The figure in the upper part represents the position of functional domains in the  $\beta'$  subunit, such as the  $\beta'$  coiled-coil (brown), Zinc fingers (orange), the catalytic site (yellow), the G-loop (red), and the DNA-binding site (olive). The figure also shows the  $\beta'$  interaction interface with the  $\beta$  subunit (beige) b) The alignment of  $\beta'$  subunits shows substitution in the essential amino acids ECO: R275, E295, and A302 in the  $\beta'$  coiled-coil domain. Darker backgrounds show those functional residues that differ from the reference *E. coli*. c) The crystallographic structure of the  $\beta'$  coiled-coil domain of *E. coli* (brown). It shows the amino acid residues (in yellow) involved in the interaction with the  $\sigma_{2.2}$  domain (magenta). d) Predicted 3-D model for the  $\beta'$  subunit in *Candidatus Hodgkinia Dsem* shows that the  $\beta'$  coiled-coil domain could occur (brown). The amino acids A287, G312, and S319 (yellow) are involved in the interaction with the  $\sigma_{2.2}$  domain (magenta). e) Predicted 3D model for the  $\beta'$  subunit in *Hodgkinia TETUND2* shows that the  $\beta'$  coiled-coil domain formation could also occur (brown region). The amino acids K268 and Q292 (yellow) could interact with the  $\sigma_{2.2}$  domain (magenta).

<https://doi.org/10.1371/journal.pone.0239350.g004>

Mutagenesis on these three amino acids in  $\sigma_{2.2}$  has shown that they cause just a weakening in their binding to the  $\beta'$  coiled-coil domain [21]. Besides, thermal denaturation experiments suggest that these mutants folded differently concerning the *E. coli* wild type. Suggesting that these mutations' principal effect is the allosteric regulation of the subdomains  $\sigma_{2.3}$  and  $\sigma_{2.4}$  who participate in DNA-melting and recognition of the -10-box promoter element [21]. On the other hand, all the endosymbionts have distinct variations in the  $\sigma_{2.3}$  subdomain. However, all of them conserve the essential aromatic residues necessary for DNA melting and the correct folding of the  $\sigma_2$  domain (Fig 5B, violet).



**Fig 5. Domain conservation in the  $\sigma$  subunit and their variations in DNA-promoter recognition.** **a)** The upper part shows the distribution of functional domains in the  $\sigma$  subunit:  $\sigma_2$  (red),  $\sigma_{2.1}$  (pink),  $\sigma_{2.2}$  (magenta),  $\sigma_{2.3}$  (violet),  $\sigma_{2.4}$  (orange),  $\sigma_3$  (green) and,  $\sigma_4$  (light blue),  $\sigma_{4.1}$  (cyan) and  $\sigma_{4.2}$  (dark blue). **b)** Amino acids alignment of  $\sigma$  subunits shows variations in the domains  $\sigma_2$  and  $\sigma_4$ . Darker backgrounds show the functional residues that differ from those in the reference organism (*E. coli*). **c)** Crystallographic structure of the *E. coli*  $\sigma$  subunit bound to DNA (4YLP). The  $\sigma_{2.2}$ ,  $\sigma_{2.3}$ ,  $\sigma_{2.4}$  subdomains and the  $\sigma_3$  domain are shown in magenta, violet, orange, and green, respectively. **d)** Predicted 3-D model of the  $\sigma$  subunit of Hodgkinia Dsem showing the subdomains and amino acids involved in recognizing and binding to the DNA. **e)** Predicted 3-D model of the  $\sigma$  subunit of Hodgkinia TETUND2 shows the subdomains and amino acids involved in recognizing and binding to DNA. The colors in **d)** and **e)** are the same that the homologs corresponding domains in **c)** for *E. coli*.

<https://doi.org/10.1371/journal.pone.0239350.g005>

The  $\sigma_{2.4}$  subdomain presents substitutions on the amino acids involved in recognizing nucleotides at position -12 of the -10-box promoter element. Hodgkinia strains show a different amino acid at position ECO: 437. It consists of histidine instead of glutamine. Likewise, *Candidatus Carsonella ruddii* PV, PC, HT, HC, DC, YCCR, and BC have replaced the ECO: T440 with leucine or isoleucine. *Tremblaya* strains and *Candidatus Tremblaya phenacola* PAVE, conserve these two amino acids (ECO Q437 and T440) (Fig 5B, orange). Previous works studied punctual mutations in these regions of  $\sigma$ . More precisely, in the subdomains  $\sigma_{2.4}$  of *E. coli*  $\sigma^{70}$  and SigA from *Bacillus subtilis* (homologous to  $\sigma^{70}$ ). Changes in amino acids at these positions affect the specificity for their respective promoters [25]. In *E. coli*, substitutions in ECO: Q437H and T440I of  $\sigma^{70}$  result in conserving the capacity to recognize the nucleotide at the -12 position. However, promoters with cytosine in this position were significantly better (in specificity) than with another nucleotide [21]. Compared with *E. coli*, the substitutions observed in the subdomains of the  $\sigma_2$  domain seem not to affect the conformation of this domain. Nevertheless, the amino acids responsible for contacting the promoter can differ from those in *E. coli* (Fig 5C–5E).

The  $\sigma_3$  domain is well preserved in *Tremblaya phenacola* PAVE. However, it presents amino acid substitutions in other endosymbionts and is absent in *Carsonella ruddii*. In Hodgkinia, this domain is partially present in the Dsem strain, although more conserved in the rest (Fig 5D and 5E). The partial or total loss of the  $\sigma_3$  domain might suggest a deficient or null binding of  $\sigma$  at the -10-box extended promoter element.

## Most of the endosymbionts show strong purifying selection on the RNAP subunits core

We made two selection pressure analyses to investigate the effects of amino acid substitutions observed in genes encoding for the RNAP subunits. First, considering that bacterial endosymbionts are subject to an accelerated rate of molecular evolution [71]. We estimated the ratio of non-synonymous to synonymous substitutions ( $D_N/D_S$ ) using phylogenetic codon-substitution models (S2 Table).

***Candidatus Hodgkinia cicadicola*.** The *Candidatus Hodgkinia cicadicola* shows that their *rpoB* ( $\beta$  subunit) and *rpoC* ( $\beta'$ ) were subject to purifying selection. With lower  $D_N/D_S$  values (0.3), most of the nucleotide substitutions in these genes were synonymous. On the other hand, the *rpoA* ( $\alpha$  subunit) genes had a higher purifying selection in most *Hodgkinia* strains ( $D_N/D_S < 0.2$ ). In contrast, Tetund 2, TETLON, and TETMLI1 strains present neutral selection ( $D_N/D_S = 1$ ). Finally, the *rpoD* gene ( $\sigma$  factor) shows an increased  $D_N/D_S$  value ( $D_N/D_S < 0.5$ ). It means that the purifying selection is less rigorous in some *Hodgkinia* strains. This not uniform selection pressure could result in greater diversity in this subunit (Fig 6).

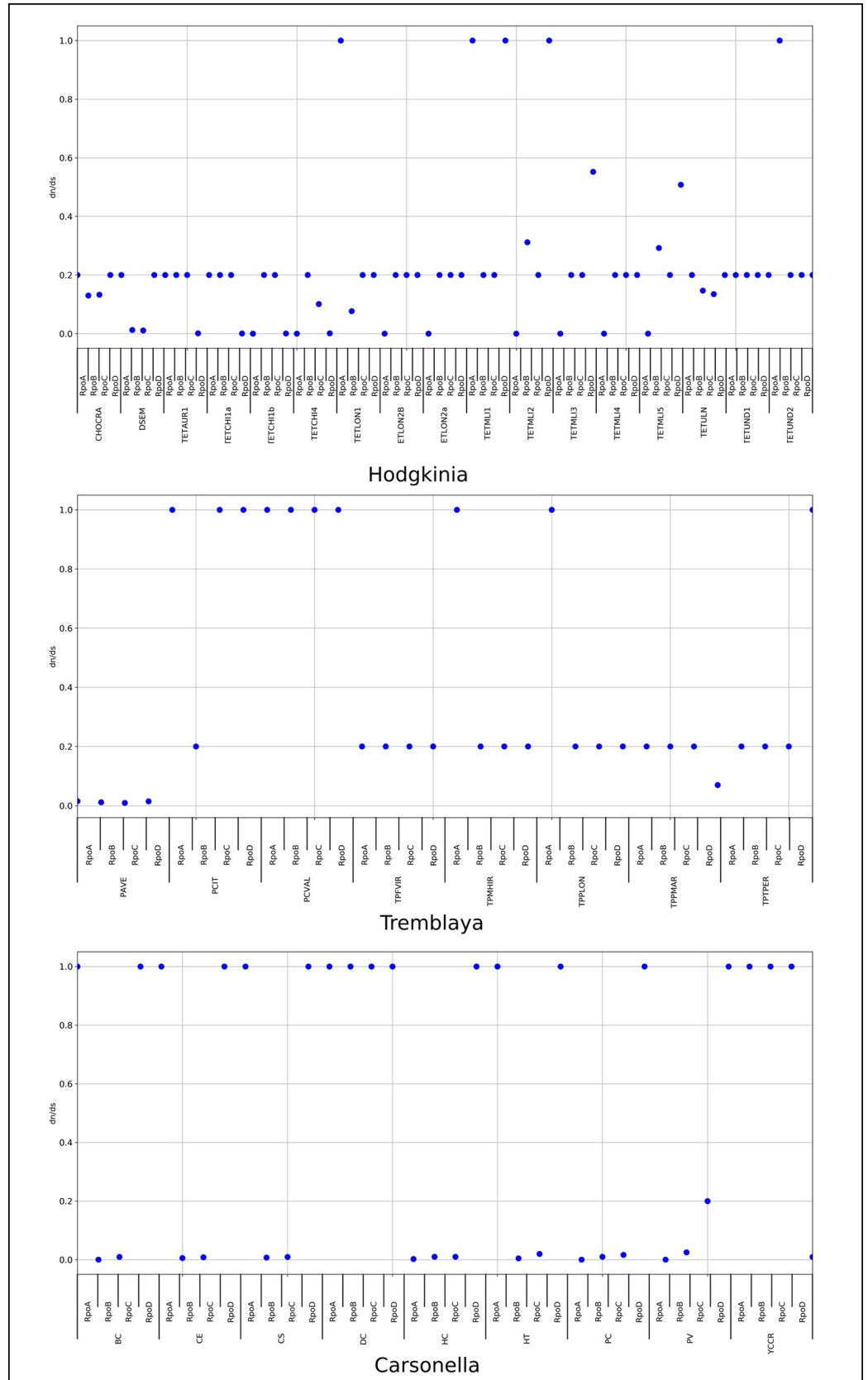
***Candidatus Tremblaya*.** The strain PAVE, *rpoA*, *rpoB*, *rpoC*, and *rpoD* genes had  $\omega$  values less than 0.02. Conversely, PCVAL shows a neutral selection for all the subunits ( $D_N/D_S = 1$ ). So, a generalized relaxation of selective pressure, to a different extent, is present in these strains (Fig 6).

***Candidatus Carsonella ruddii*.** The *rpoB* and *rpoC* genes in *Candidatus Carsonella ruddii* strains are more conserved than the other RNAP subunits. They had  $D_N/D_S$  values less than 0.2 (except for DC and YCCR strains). The *rpoA* gene had a neutral selection in five strains (CE, CS, DC, HT, and YCCR) and a strong purifying selection in the rest of the *Candidatus Carsonella* strains ( $D_N/D_S < 0.2$ ). On the contrary, the *rpoD* gene had a neutral selection except for the BC strain ( $D_N/D_S < 0.01$ ). This neutral selection might explain why the  $\sigma$  factor contains more amino acid variations concerning the other RNAP subunits (Fig 6).

## Positively selected amino acids are present in the $\alpha$ -NTD of the $\alpha$ subunit in the *Hodgkinia* strains

Positively selected amino acids are present in the  $\alpha$ -NTD of *Candidatus Hodgkinia* Dsem and TETUND2 (S3 Table). We mapped the selected amino acids with a high level of support (BEB  $p > 0.95$ ) in the structural models of the  $\alpha$  subunit of *Candidatus Hodgkinia* TETUND2 and Dsem (Fig 7C, red amino acids, and S4 Fig, respectively). In *Candidatus Hodgkinia* Dsem, the selected amino acids V55, Q95, and H115 (S4 Fig) would be necessary for the correct folding of the  $\alpha$ -NTD. In *Candidatus Hodgkinia* TETUND2, the amino acid residues V68, S69, and E70 allow adopting a similar structure to maintain the  $\alpha$  subunit interactions with the  $\beta$  and  $\beta'$  subunits in *E. coli* (Fig 7A and 7C).

As previously mentioned, the *Candidatus Hodgkinia* TETUND2  $\alpha$  subunit has lost part of the  $\alpha$ -NTD involved in the homodimer formation. Therefore, we evaluated *in silico* if the *Candidatus Hodgkinia* TETUND2  $\alpha$  subunit can still form the homodimer, essential for RNAP core formation. We obtained three models with different clustered amino acids involved in homodimer formation (Fig 7D–7F). First, we mapped the sites under positive selection (Fig 7D–7F, amino acids red and yellow). Then, we performed *in silico* amino acid substitutions of the sites under positive selection. It was changing them by amino acids present in the same locations of different *Candidatus Hodgkinia* strains and *E. coli* (Fig 7B). We observed a substitution of histidine 116 by proline in the three models. This substitution changed the homodimer formation to be energetically unfavorable (S4 Table). Although H116 seems not under positive selection, it would have an essential role in stabilizing this strain's homodimer formation.

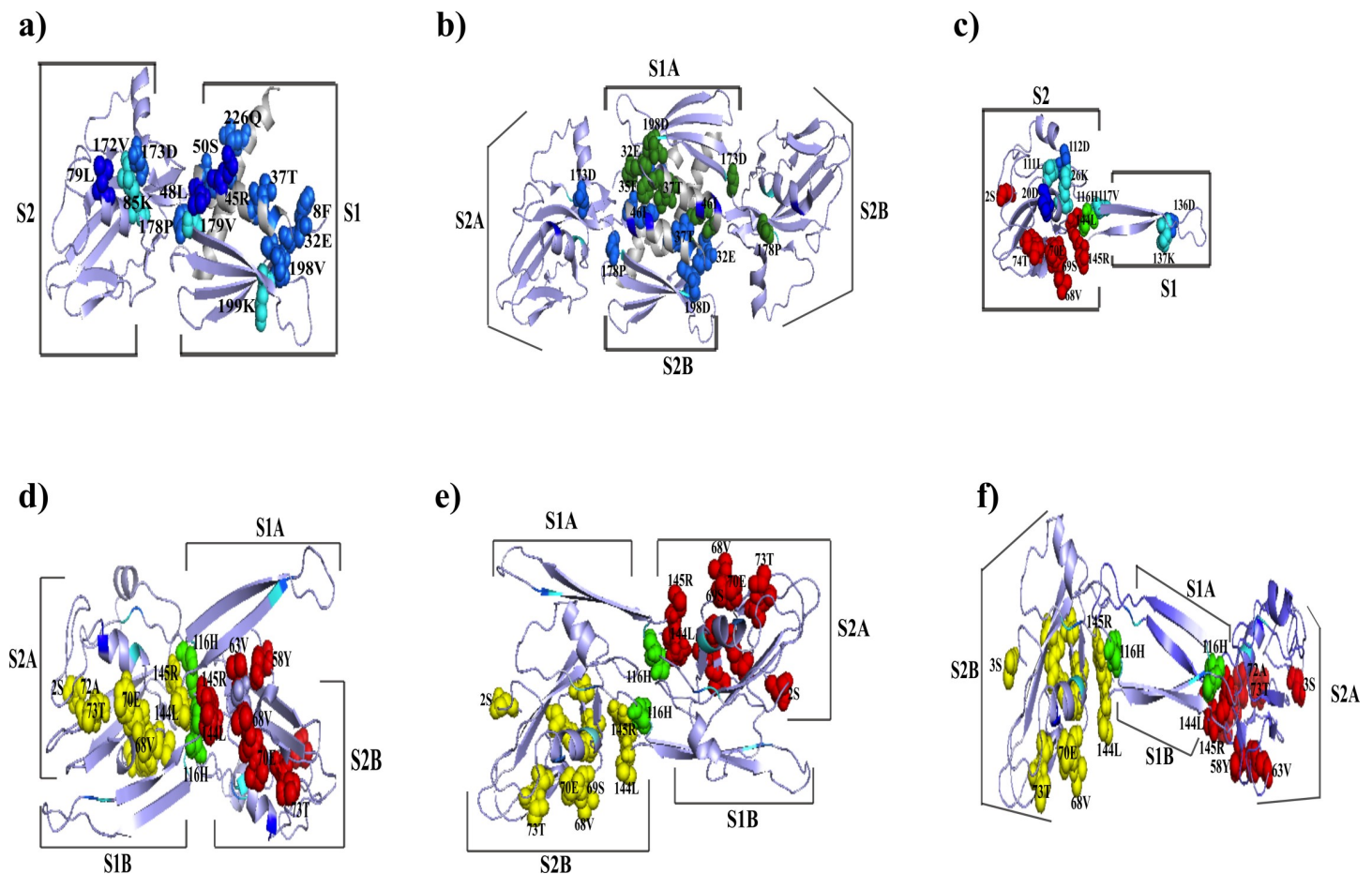


**Fig 6. Selective pressure on the RNAP subunits by branch analysis.** The figure shows the relationship  $D_N/D_S$  for each subunit. Most genes are under negative selection, and only in few cases, they display values greater than 1 (represented as 1.2). However, this does not mean that they are under positive selection (level of significance greater than 0.05); in fact, the RNAP subunits show a relaxed selection in all the cases.

<https://doi.org/10.1371/journal.pone.0239350.g006>

### Variations on the conservation of the subunits involved in promoter recognition are independent of the CG content

Endosymbionts with reduced genomes carry out variable proportions of GC in their genomes. For example, while *Candidatus Carsonella* and *Candidatus Nasuia* contain less than 18% GC, *Candidatus Hodgkinia* and *Candidatus Tremblaya* contain above 40%. We want to know if such variations of GC content in genomes relate to changes in  $\sigma$  factors. Then, we carried out a comparative analysis that involved a phylogenetic tree (S5 Fig). We include the 37 strict endosymbiont bacteria of this study. However, we also include 13 homologs of  $\sigma^{70}$  coming from six endosymbionts and seven free-living bacteria. These other bacteria have a lesser extent of genome reduction. Still, similar, less, or more extensive GC contents than the endosymbionts studied (S5 Table).



**Fig 7. Structural comparison between the *E. coli* and the *Candidatus Hodgkinia TETUND2*  $\alpha$ -subunits.** a) 3-D structure of monomers and b) for homodimers of the  $\alpha$ -NTD subunit in *E. coli* (4YLP). c) 3-D model of  $\alpha$ -subunit monomer in *Candidatus Hodgkinia TETUND2* with the amino acids under positive selection in red and the H116 in green. d), e), and f) show predictions of the  $\alpha$ -subunit homodimer formation in *Candidatus Hodgkinia TETUND2*. The amino acids under positive selection are in red and yellow in each monomer, H116 in green. White regions in a) and b) structures are not conserved in the  $\alpha$ -subunit of *Candidatus Hodgkinia TETUND2*. The amino acids in blue, dark, and turquoise are necessary for the RNAP core formation. In the b) structure, the dark green amino acids are the same as the blue in a). S1 and S2 indicate subdomains 1 and 2 in the  $\alpha$ -NTD. Moreover, the letter A and B correspond to each monomer that forms the homodimer.

<https://doi.org/10.1371/journal.pone.0239350.g007>

Except for *C. Zinderia cicadicola*, the rest of the bacteria preserve the  $\beta$  flap-tip helix and the  $\beta'$  coiled-coil domains. These also conserve the  $\sigma_{2.2}$  and  $\sigma_{2.2}$  subdomains involved in the promoter recognition in *E. coli* (Figs 3A, 4A and 5A). Thus, this analysis may suggest no relationship between the GC content and changes in the  $\sigma$  factors.

## Discussions

Here we approach the study of the evolution of RNAP subunits in vastly reduced bacteria genomes. We found that the  $\beta$  and  $\beta'$  subunits are the most conserved in all the studied endosymbionts. These have just some differences in the regions involved in the interactions with the  $\sigma$  factor, possibly because of significant changes in  $\sigma$ . On the other side, the  $\alpha$  subunit is more conserved in *Candidatus Carsonella ruddii* and *Nasuia*. In contrast, in the other endosymbionts, the  $\alpha$  subunit has lost its  $\alpha$ -CTD.

Furthermore, studies report the absence of a recognized gene encoding for the  $\alpha$  subunit in *Hodgkinia* strains [37]. It is unknown how to perform an RNAP without the  $\alpha$  subunit if the transcription is present in these *Hodgkinias*. It might mean that  $\alpha$  subunits can follow  $\omega$  as dispensable RNAP subunits. *Hodgkinia* strains inhabit their host as consortia with other bacteria. Then, it is attractive that the community consortia contribute to cellular activities [72]. Still, it isn't easy to know if these complementary activities include gene transcription.

Furthermore, *Candidatus Hodgkinia Dsem*, CHOCRA, and TETULN strains preserve an  $\alpha$  subunit and are not known to share their host with other *Hodgkinia* strains [37]. We also observed several amino acids under putative positive selection in the  $\alpha$  subunit of *Candidatus Hodgkinia TETUND2*. Thus, they suggest compensation for the loss of critical amino acid regions for homodimer formation in this subunit.

Given the importance of  $\sigma$  in promoter recognition for transcription initiation, the substitutions observed in the  $\sigma_{2.4}$  subdomain might correspond to variations in the specificity of  $\sigma$  for promoters. Previous studies indicate that the substitution of some amino acids does not compromise their affinity to DNA. These include lysine, asparagine, serine, methionine, and phenylalanine [73]. However, these changes might exert mild effects on  $\sigma$  affecting its specificity for promoters. The differences observed in the  $\sigma_{2.4}$  subdomain of *Candidatus Hodgkinia* and *Candidatus Nasuia* correspond to mutations already experimentally found in the *E. coli*  $\sigma^{70}$  and *Bacillus SigA* [25,26,52].

So far, we can suggest the functionality of the RNAP of the seventeen *Hodgkinia* strains that conserve an  $\alpha$  subunit, to the exception of the *Dsem* strain. These can recognize only the -10-box and the -10-box extended promoter elements. They lack the fragment required to form the  $\beta$  flap-tip helix domain that recognizes the -35-box promoter element and neither recognize the UP element (Figs 3D and 4D). *Candidatus Tremblaya phenacola PAVE*, *Principes PCIT*, and *PCVAL* are the endosymbionts with a  $\sigma$  nearest to *E. coli*  $\sigma^{38}$  instead of the housekeeping  $\sigma^{70}$  (S4 Fig).  $\sigma^{38}$  pertains to the  $\sigma^{70}$  family, but in *E. coli*, it transcribes stationary phase genes [9]. The  $\sigma_{4.1}$  subdomain present in  $\sigma^{38}$  has some amino acid changes concerning those  $\sigma^{70}$ . These changes make the  $\beta$  flap-tip helix domain of  $\sigma^{38}$  with more affinity and increased performance for the binding to -35-box promoter elements. This higher affinity to -35-box promoter elements and the stationary phase transcription factors could displace the main transcription activity from  $\sigma^{70}$  to  $\sigma^{38}$  in the stationary phase in *E. coli*.

In endosymbiotic bacteria with exceedingly reduced genomes, it has not been possible to locate  $\sigma^{70}$  canonical promoters. Not even for the most conserved, like the ribosomal genes [74]. This inability to find promoters might be partly due to the high A+T percentage and the lack of intergenic regions in these genomes. However, this may not be the whole explanation since *Hodgkinia* and *Tremblaya* have relatively high G+C %. More



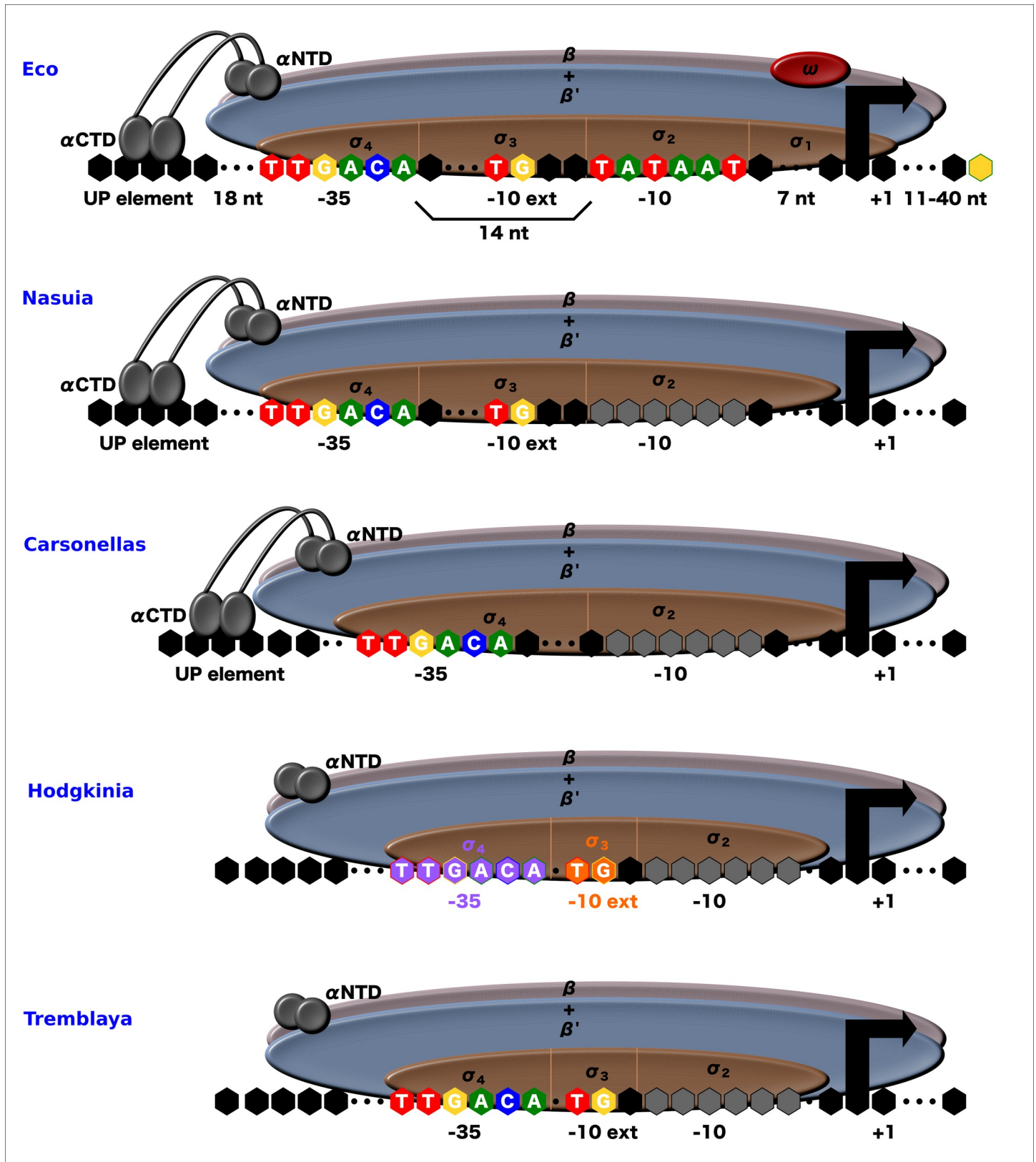


Fig 8. Proposed functional RNAP models of bacterial endosymbionts with reduced genomes. The conserved domains in each group of bacteria are present in each illustration. Figures correspond to *E. coli* RNAP (Eco), *Nasuia* strains RNAP, and *Carsonella* strains RNAP. In *Hodgkinia* strains in orange, the TETULN, TETUND1, and

TETUND1 strains use the -10 extended region. In purple, Dsem strains recognize the -35 element. *Candidatus* Tremblaya phenacola PAVE, Princeps PCIT, and PCVAL RNAP model (Based on [10, 76]).

<https://doi.org/10.1371/journal.pone.0239350.g008>

than 90% of their genome comprises coding sequences, and neither presents a recognized promoter [7]. Hence, these results suggest that RNAP in endosymbionts can conserve some sequence recognition capacity, but this should differ from the  $\sigma^{70}$  consensus promoters in *E. coli*. Thus, it seems that some promoter elements are unnecessary in endosymbiont. With this, shorter promoter sequences might be sufficient for gene transcription. This fact can explain the difficulty of recovering consensus promoter sequences as we know in free-living bacteria. Besides, transcription factors that assist in gene regulation are also absent in these bacteria with highly reduced genomes, being the last to be lost the nucleoid-associated proteins [36]. Therefore, it makes sense that regions for gene activation, such as the UP-promoter element and the contact region for these activators in the  $\alpha$  subunit, are absent. Variations in recognition regions of promoters might not be the only ones in these bacteria. For example, previous reports indicate significant changes in the 16S ribosomal 3' tail and its binding sequence with the corresponding changes in the Shine-Dalgarno element localized upstream of the protein-encoding genes [75]. Then, the observations of this study can be a more generalized phenomenon in these bacteria.

## Conclusions

DNA sequences encoding for each of the RNAP subunits exhibit a reduction of 16% on average compared to those in *E. coli*. The gene reductions present in RNAP subunits are independent of the CG content (18–40% GC in these genomes). Most endosymbionts experiment strong purifying selection on the RNAP subunit genes, particularly on the  $\beta$  and  $\beta'$  subunits. In the case of  $\sigma$ , the type of selection determined was less uniform among the endosymbionts.

A closer inspection in the  $\alpha$  subunit reveals that the  $\alpha$ -NTD is more conserved than  $\alpha$ -CTD. Additionally, some amino acid changes in homodimer assembly are under positive selection in the *Candidatus* Hodgkinia TETUND2 and Dsem strains. On the other hand, the  $\beta$  and  $\beta'$  subunits are more conserved in strict endosymbionts except for the  $\beta$  flap-tip helix domain in Hodgkinia strains. Furthermore, the  $\sigma$  subunit presents the more varied erosion in these endosymbionts. These unequal losses result in promoter elements' differential recognition.

To better illustrate our inferences, we present a functional conclusion based on the conservation of RNAP subunits. We offer drawing models with the inferred regions of promoters where RNAP for each endosymbiont should be recognizing (Fig 8). We can deduce that the RNAP of *Nasuia* conserved the more significant similarity to the *E. coli*  $\sigma^{70}$ . According to this, *Nasuia* RNAP should recognize the two main promoter elements (-10-box and -35-box) and the two promoter accessory elements (-10-box extended and even the UP element). Another way, *Carsonella* RNAP seems to maintain recognition of the promoter elements except for the -10-box extended element. This limited recognition can result in promoters with shorter regions between the -10-box and the -35-box promoter elements. In another case, the  $\sigma$  of Tremblaya resembles more to  $\sigma^{38}$  instead of the canonical  $\sigma^{70}$ . In Tremblaya and Hodgkinia, due to the absence of the  $\alpha$ -CTD, they might not recognize the UP element. And in the case of the strain *Candidatus* Hodgkinia Dsem neither the -10 extended promoter element. Additional studies, ideally experimental ones, should generate new knowledge about what is happening with the functioning of shorter proteins in this fascinating field of highly reduced genomes.

## Supporting information

**S1 Fig. Interaction domains in  $\alpha$  subunit alignment.**

(TIF)

**S2 Fig. Interaction domains in the  $\beta$  subunit alignment.**

(TIF)

**S3 Fig. Functional domains and amino acids in  $\beta'$  subunit alignment.**

(TIF)

**S4 Fig. Selected sites mapped in *Hodgkinia Dsem*  $\alpha$  subunit.**

(TIF)

**S5 Fig. Phylogenetic tree of sigma 70 proteins in endosymbionts with extreme genome reduction and other symbionts and free-living bacteria.**

(TIF)

**S1 Table. Complete list of studied endosymbionts.**

(PDF)

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

(PDF)

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

(PDF)

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

(PDF)

**S5 Table. Changes of free energy by in silico mutations in selected sites of  $\alpha$  subunit homo-dimer predicted for *Hodgkinia* TETUND2.**

(PDF)

## Acknowledgments

CPR-C has a Ph.D. fellowship (380338) from CONACYT. México. Thank Rafael Montiel for guiding the selective pressure analyses and Diego Andrés López Castro and Paola Isabel Angulo-Bejarano for reading the manuscript.

## Author Contributions

**Conceptualization:** Cynthia Paola Rangel-Chávez, Agustino Martínez-Antonio.

**Data curation:** Cynthia Paola Rangel-Chávez, Azucena Pescador-Tapia.

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

**Funding acquisition:** Agustino Martínez-Antonio.

**Investigation:** Cynthia Paola Rangel-Chávez.

**Methodology:** Cynthia Paola Rangel-Chávez, Edgardo Galán-Vásquez.

**Project administration:** Agustino Martínez-Antonio.

**Supervision:** Edgardo Galán-Vásquez, Agustino Martínez-Antonio.

**Validation:** Luis Delaye.

**Writing – original draft:** Cynthia Paola Rangel-Chávez.

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

## 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. <https://doi.org/10.1128/JB.187.12.4229-4237.2005> PMID: 15937185
3. Huerta A. M., Francino M. P., Morett E., & Collado-Vides J. Selection for unequal densities of  $\sigma$  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. <https://doi.org/10.1016/j.tibs.2005.04.003> PMID: 15950866
5. Ebright RH. RNA polymerase: Structural similarities between bacterial RNA polymerase and eukaryotic RNA polymerase II. *J Mol Biol*. 2000; 304: 687–698. <https://doi.org/10.1006/jmbi.2000.4309> PMID: 11124018
6. 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. <https://doi.org/10.1128/jb.171.3.1271-1277.1989> PMID: 2646273
7. Moran NA, Bennett GM. The tiniest tiny genomes. *Ann Rev Microbiol*. 2014; 68: 195–215. <https://doi.org/10.1146/annurev-micro-091213-112901> PMID: 24995872
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. <https://doi.org/10.1073/pnas.84.5.1192> PMID: 3547406
9. Ishihama A. Functional modulation of *Escherichia coli* RNA polymerase. *Annu Rev Microbiol*. 2000; 54: 499–518. <https://doi.org/10.1146/annurev.micro.54.1.499> PMID: 11018136
10. Rangel-Chavez C., Galan-Vasquez E., & Martínez-Antonio A. Consensus architecture of promoters and transcription units in *Escherichia coli*: design principles for synthetic biology. *Molecular bioSystems*. 2017; 13(4), 665–676. <https://doi.org/10.1039/c6mb00789a> PMID: 28256660
11. Blatter EE, Ross W, Tang H, Gourse RL, Ebright RH. Domain organization of RNA polymerase  $\alpha$  subunit: C-terminal 85 amino acids constitute a domain capable of dimerization and DNA binding. *Cell*. 1994; 78(5): 889–896. [https://doi.org/10.1016/s0092-8674\(94\)90682-3](https://doi.org/10.1016/s0092-8674(94)90682-3) PMID: 8087855
12. Ross W, Ernst A, Gourse RL. Fine structure of *E. coli* RNA polymerase-promoter interactions:  $\alpha$  subunit binding to the UP element minor groove. *Genes Dev*. 2001; 15(5): 491–506. <https://doi.org/10.1101/gad.870001> PMID: 11238372
13. Busby S, Ebright RH. Promoter structure, promoter recognition and transcription activation in prokaryotes. *Cell*. 1994; 79(5): 743–74. [https://doi.org/10.1016/0092-8674\(94\)90063-9](https://doi.org/10.1016/0092-8674(94)90063-9) PMID: 8001112
14. Murakami K., Owens J. T., Belyaeva T. A., Meares C. F., Busby S. J., & Ishihama A. Positioning of two alpha subunit carboxy-terminal domains of RNA polymerase at promoters by two transcription factors. *Proceedings of the National Academy of Sciences*. 1997; 94(21), 11274–11278. <https://doi.org/10.1073/pnas.94.21.11274> PMID: 9326599
15. Belyaeva T. A., Rhodius V. A., Webster C. L., & Busby S. J. Transcription activation at promoters carrying tandem DNA sites for the *Escherichia coli* cyclic AMP receptor protein: organisation of the RNA polymerase  $\alpha$  subunits. *Journal of molecular biology*. 1998; 277(4), 789–804. <https://doi.org/10.1006/jmbi.1998.1666> PMID: 9545373
16. Bertoni G., Fujita N., Ishihama A., & de Lorenzo V. Active recruitment of  $\sigma$ <sup>54</sup>-RNA polymerase to the Pu promoter of *Pseudomonas putida*: role of IHF and  $\alpha$ CTD. *The EMBO journal*. 1998; 17(17), 5120–5128. <https://doi.org/10.1093/emboj/17.17.5120> PMID: 9724648
17. McLeod S. M., Aiyar S. E., Gourse R. L., & Johnson R. C. The C-terminal domains of the RNA polymerase  $\alpha$  subunits: contact site with Fis and localization during co-activation with CRP at the *Escherichia coli* proP P2 promoter. *Journal of molecular biology*. 2002; 316(3), 517–529. <https://doi.org/10.1006/jmbi.2001.5391> PMID: 11866515
18. Cramer P., Bushnell D. A., & Kornberg R. D. Structural basis of transcription: RNA polymerase II at 2.8 Ångstrom resolution. *Science*. 2001; 292(5523): 1863–1876. <https://doi.org/10.1126/science.1059493> PMID: 11313498

19. 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. <https://doi.org/10.1016/j.jmb.2004.08.063> PMID: 15465046
20. Vassylyev D. G., Sekine S. I., Laptenko O., Lee J., Vassylyeva M. N., Borukhov S., et al. Crystal structure of a bacterial RNA polymerase holoenzyme at 2.6 Å resolution. *Nature*. 2002; 417(6890): 712–719. <https://doi.org/10.1038/nature752> PMID: 12000971
21. Young B. A., Anthony L. C., Gruber T. M., Arthur T. M., Heyduk E., Lu C. Z., et al. A coiled-coil from the RNA polymerase  $\beta'$  subunit allosterically induces selective nontemplate strand binding by  $\sigma$ 70. *Cell*. 2001; 105(7): 935–944. [https://doi.org/10.1016/s0092-8674\(01\)00398-1](https://doi.org/10.1016/s0092-8674(01)00398-1) PMID: 11439189
22. 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. <https://doi.org/10.1073/pnas.1314576110> PMID: 24218560
23. Camarero J. A., Shekhtman A., Campbell E. A., Chlenov M., Gruber T. M., Bryant D. A., et al. Autoregulation of a bacterial  $\sigma$  factor explored by using segmental isotopic labeling and NMR. *Proceedings of the National Academy of Sciences*. 2002; 99(13): 8536–8541. <https://doi.org/10.1073/pnas.132033899> PMID: 12084914
24. Cathleen L Chan, Michael A Lonetto, Carol A Gross. Sigma domain structure: one down, one to go. *Structure*. 1996; 4(11):235–1238.
25. 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. [https://doi.org/10.1016/s0022-2836\(05\)80345-6](https://doi.org/10.1016/s0022-2836(05)80345-6) PMID: 2213883
26. Siegel 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. [https://doi.org/10.1016/0022-2836\(89\)90568-8](https://doi.org/10.1016/0022-2836(89)90568-8) PMID: 2661828
27. 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. <https://doi.org/10.1021/bi00445a031> PMID: 2692703
28. 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. [https://doi.org/10.1016/s0092-8674\(00\)81329-x](https://doi.org/10.1016/s0092-8674(00)81329-x) PMID: 8858155
29. 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. <https://doi.org/10.1016/j.molcel.2015.03.010> PMID: 25866247
30. Kuznedelov K., Minakhin L., Niedziela-Majka A., Dove S. L., Rogulja D., Nickels B. E., et al. A role for interaction of the RNA polymerase flap domain with the  $\sigma$  subunit in promoter recognition. *Science*. 2002; 295(5556): 855–857. <https://doi.org/10.1126/science.1066303> PMID: 11823642
31. Eleftherios I., Atri J., Accetta J., & Castillo J. C. Endosymbiotic bacteria in insects: guardians of the immune system? *Frontiers in physiology*. 2013; 4, 46. <https://doi.org/10.3389/fphys.2013.00046> PMID: 23508299
32. McCutcheon JP, Moran NA. Extreme genome reduction in symbiotic bacteria. *Nat Rev Microbiol*. 2012; 10(1): 13–26.
33. Blattner F. R., Plunkett G., Bloch C. A., Perna N. T., Burland V., Riley M., et al. The complete genome sequence of *Escherichia coli* K-12. *Science*. 1997; 277(5331), 1453–1462. <https://doi.org/10.1126/science.277.5331.1453> PMID: 9278503
34. 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. <https://doi.org/10.1093/molbev/mss180> PMID: 22821013
35. 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. <https://doi.org/10.1186/1471-2148-7-181> PMID: 17908294
36. Galán-Vázquez E, Sánchez-Osorio I, Martínez-Antonio A. Transcription Factors Exhibit Differential Conservation in Bacteria with Reduced Genomes. *PLOS ONE*. 2016; 11(1): e0146901. <https://doi.org/10.1371/journal.pone.0146901> PMID: 26766575
37. Lukasik P., Nazario K., Van Leuven J. T., Campbell M. A., Meyer M., Michalik A., et al. Multiple origins of interdependent endosymbiotic complexes in a genus of cicadas. *Proceedings of the National Academy of Sciences*. 2018. 115(2): E226–E235.

38. 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. <https://doi.org/10.1006/jmbi.2000.4042> PMID: 10964570
39. Finn RD, Coghill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, et al. The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.* 2016; 44(D1): D279–85. <https://doi.org/10.1093/nar/gkv1344> PMID: 26673716
40. 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. <https://doi.org/10.1006/jmbi.2001.5080> PMID: 11697912
41. Marchler-Bauer A., Bao Y., Han L., He J., Lanczycki C. J., Lu S., et al. CDD/SPARKLE: functional classification of proteins via subfamily domain architectures. *Nucleic acids research.* 2016; 45(D1): D200–D203. <https://doi.org/10.1093/nar/gkw1129> PMID: 27899674
42. 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. <https://doi.org/10.1126/science.1069595> PMID: 12016307
43. Ebricht RH, Busby S. The Escherichia coli RNA polymerase subunit: structure and function. *Curr Opin Genet Dev.* 1995; 5(2): 197–203. [https://doi.org/10.1016/0959-437x\(95\)80008-5](https://doi.org/10.1016/0959-437x(95)80008-5) PMID: 7613089
44. Zhang G, Darst SA. Structure of the Escherichia coli RNA polymerase subunit Amino-terminal domain. *Science.* 1998; 281(5374): 262–266. <https://doi.org/10.1126/science.281.5374.262> PMID: 9657722
45. 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. <https://doi.org/10.1073/pnas.86.23.9109> PMID: 2512576
46. Zhang Y. I-TASSER server for protein 3D structure prediction. *BMC bioinformatics.* 2008; 9(1): 40. <https://doi.org/10.1186/1471-2105-9-40> PMID: 18215316
47. 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. <https://doi.org/10.1016/j.molcel.2015.03.010> PMID: 25866247
48. The PyMOL Molecular Graphics System, Version 1.2r3pre, Schrödinger, LLC.
49. Kozakov D., Hall D. R., Xia B., Porter K. A., Pothorny D., Yueh C., et al. The ClusPro web server for protein-protein docking. *Nature protocols.* 2017; 12(2): 255. <https://doi.org/10.1038/nprot.2016.169> PMID: 28079879
50. Vajda S, Yueh C, Beglov D, Bohnuud T, Mottarella SE, Xia B, et al. New additions to the ClusPro server motivated by CAPRI. *Proteins: Structure, Function, and Bioinformatics.* 2017; 85(3): 435–444.
51. Kozakov D, Beglov D, Bohnuud T, Mottarella S, Xia B, Hall DR, et al. How good is automated protein docking. *Proteins: Structure, Function, and Bioinformatics.* 2013; 81(12): 2159–66.
52. Yueh C, Hall DR, Xia B, Pothorny 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. <https://doi.org/10.1016/j.jmb.2016.10.019> PMID: 27771482
53. 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.
54. 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. <https://doi.org/10.1016/j.jmb.2016.11.022> PMID: 27899282
55. Yang Ziheng, Nielsen Rasmus, Synonymous Estimating and Non-synonymous Substitution Rates Under Realistic Evolutionary Models, *Molecular Biology and Evolution.* 2000; 17(1): 32–43. <https://doi.org/10.1093/oxfordjournals.molbev.a026236> PMID: 10666704
56. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. *Molecular biology and evolution.* 2007; 24(8): 1586–1591. <https://doi.org/10.1093/molbev/msm088> PMID: 17483113
57. 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. <https://doi.org/10.1093/nar/gkl315> PMID: 16845082
58. 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. <https://doi.org/10.1093/sysbio/syq010> PMID: 20525638
59. 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. <https://doi.org/10.1006/jmbi.2000.4042> PMID: 10964570

60. Castresana J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular Biology and Evolution*. 2000; 17: 540–552. <https://doi.org/10.1093/oxfordjournals.molbev.a026334> PMID: 10742046
61. Yang Z., & Nielsen R. Synonymous and non-synonymous rate variation in nuclear genes of mammals. *Journal of molecular evolution*. 1998; 46(4), 409–418. <https://doi.org/10.1007/pl00006320> PMID: 9541535
62. 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. <https://doi.org/10.1093/molbev/msi237> PMID: 16107592
63. 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. <https://doi.org/10.1093/molbev/msi097> PMID: 15689528
64. Kimura M., Fujita N., & Ishihama A. Functional map of the  $\alpha$  subunit of *Escherichia coli* RNA polymerase: deletion analysis of the amino-terminal assembly domain. *Journal of molecular biology*. 1994; 242(2), 107–115. <https://doi.org/10.1006/jmbi.1994.1562> PMID: 8089834
65. Igarashi K., & Ishihama A. Bipartite functional map of the *E. coli* RNA polymerase  $\alpha$  subunit: involvement of the C-terminal region in transcription activation by cAMP-CRP. *Cell*. 1991; 65(6): 1015–1022. [https://doi.org/10.1016/0092-8674\(91\)90553-b](https://doi.org/10.1016/0092-8674(91)90553-b) PMID: 1646077
66. Hayward R. S., Igarashi K., & Ishihama A. Functional specialization within the  $\alpha$ -subunit of *Escherichia coli* RNA polymerase. *Journal of molecular biology*. 1991; 221(1): 23–29. [https://doi.org/10.1016/0022-2836\(91\)80197-3](https://doi.org/10.1016/0022-2836(91)80197-3) PMID: 1920407
67. Ross W, Gosink KK, Salomon J, Igarashi K, Zou C, Ishihama A, et al. A third recognition element in bacterial promoters: DNA binding by the  $\alpha$  subunit of RNA polymerase. *Science*. 1993; 262(5138): 1407–1413. <https://doi.org/10.1126/science.8248780> PMID: 8248780
68. Nelson W. C., & Stegen J. C. The reduced genomes of *Parcubacteria* (OD1) contain signatures of a symbiotic lifestyle. *Frontiers in microbiology*. 2015; 6, 713. <https://doi.org/10.3389/fmicb.2015.00713> PMID: 26257709
69. Sheveleva E. V., Giordani N. V., & Hallick R. B. Identification and comparative analysis of the chloroplast  $\alpha$ -subunit gene of DNA-dependent RNA polymerase from seven *Euglena* species. *Nucleic acids research*. 2002; 30(5): 1247–1254. <https://doi.org/10.1093/nar/30.5.1247> PMID: 11861918
70. Hirata A., & Murakami K. S. Archaeal RNA polymerase. *Current opinion in structural biology*. 2009; 19(6): 724–731. <https://doi.org/10.1016/j.sbi.2009.10.006> PMID: 19880312
71. Dufresne A., Garczarek L. & Partensky F. Accelerated evolution associated with genome reduction in a free-living prokaryote. *Genome Biol*. 2005; 6(2): R14. <https://doi.org/10.1186/gb-2005-6-2-r14> PMID: 15693943
72. Husnik F., Nikoh N., Koga R., Ross L., Duncan R. P., Fujie M., et al. Horizontal gene transfer from diverse bacteria to an insect genome enables a tripartite nested mealybug symbiosis. *Cell*. 2013; 153(7): 1567–1578. <https://doi.org/10.1016/j.cell.2013.05.040> PMID: 23791183
73. 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. <https://doi.org/10.1093/nar/29.13.2860> PMID: 11433033
74. 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. <https://doi.org/10.1128/JB.183.6.1853-1861.2001> PMID: 11222582
75. Lim K., Furuta Y., & Kobayashi I. Large variations in bacterial ribosomal RNA genes. *Molecular biology and evolution*. 2012; 29(10), 2937–2948. <https://doi.org/10.1093/molbev/mss101> PMID: 22446745
76. Browning DF, Busby SJ. The regulation of bacterial transcription initiation. *Nat Rev Microbiol*. 2004; 2(1):57–65. <https://doi.org/10.1038/nrmicro787> PMID: 15035009