Control of a gene transfer agent cluster in *Caulobacter crescentus* by transcriptional activation and anti-termination

*Ngat T. Tran & Tung B.K. Le**
Department of Molecular Microbiology
John Innes Centre, Norwich, NR4 7UH, United Kingdom

**Correspondence: tung.le@jic.ac.uk*

ABSTRACT
Gene Transfer Agents (GTAs) are domesticated prophages that cannot self-multiply and be infectious but might have been co-opted to perform biological functions for the host bacteria. *Caulobacter crescentus*, a bacterium best known as a model organism to study bacterial cell biology and cell cycle regulation, has recently been demonstrated to produce *bona fide* GTA particles (CcGTA). Two direct activators of the CcGTA biosynthetic gene cluster, GafY and GafZ, have been identified, however, it is unknown how GafYZ controls transcription mechanistically or how they coordinate gene expression of the CcGTA gene cluster with other accessory genes elsewhere on the genome for complete CcGTA production. Here, we show that the CcGTA gene cluster is transcriptionally co-activated by GafY, integration host factor (IHF), and by GafZ-mediated transcription anti-termination. We present evidence that GafZ is a novel transcription anti-terminator that likely forms an anti-termination complex with RNA polymerase, NusA, NusG, and NusE to bypass transcription terminators within the 14 kb CcGTA cluster. Overall, we reveal a two-tier regulation that coordinates the synthesis of GTA particles in *C. crescentus*.

INTRODUCTION

Viruses and mobile genetic elements are major drivers of evolution in bacteria[1,2]. In some cases, there is evidence that these elements might be co-opted to perform biological functions for the host, thus providing selective advantages for the producing organisms[3–7]. Virus-like Gene Transfer Agents (GTAs) are among these cases[3,4]. GTAs were first discovered and characterized in the non-sulfur photosynthetic bacterium *Rhodobacter capsulatus*[8], but have now been found in a wide range of bacterial species and archaea[3,4,9,10]. Homologs of GTA core genes are common in abundant environmental organisms, and it has been speculated that ~10^6 virus-like particles per milliliter in nature could be GTAs[9,11–14]. GTAs are thought to be domesticated prophages that can no longer self-multiply and be infectious, but can still transfer DNA via a DNA-filled capsid head[3,15–20]. In contrast to canonical phages, which preferably package and transfer their own DNA, genomic DNA from the host is packaged into GTA particles in a relatively non-specific manner, although some packaging bias exists in certain species[3,4,18,21,22]. Most notably, the length of DNA packaged into GTA particles is not sufficient to contain the entire gene cluster encoding GTA components[3,20]. Moreover, multiple accessory genes necessary for GTA synthesis are distributed across the host genome, sometimes megabases away from the main GTA gene cluster[3,19,21,22], thus GTAs cannot transfer themselves horizontally nor be self-replicative and infectious.

Since host bacteria ultimately die to release GTA particles, the production of GTA particles must be tightly regulated and integrated with the host physiology to prevent a collapse in cell population. In *R. capsulatus*, RcGTA gene expression and production are under the control of multiple host regulatory systems including the CckA-ChpT-CtrA phosphorelay, quorum sensing, and a sigma/anti-sigma-like mechanism[23–33]. Recently, a direct and dedicated activator of the RcGTA gene cluster, GafA, was identified[34]. Further characterization of this multi-domain activator by Sherlock and Fogg (2022) demonstrated that GafA controls RcGTA gene expression via a direct interaction with the RNA polymerase omega subunit (RpoZ-ω), potentially recruiting RNA polymerase to GTA-specific promoters for transcriptional activation[35].

More recently, we and colleagues demonstrated that the α-proteobacterium *Caulobacter crescentus*, best known as a model organism to study cell cycle regulation[36], produces *bona fide* GTA particles (CcGTA)[21]. In this bacterium, a 14-kb long gene cluster, spanning from CCNA02880 (here named *gtaT*) to CCNA02861 (here named *gtaB*), encodes the majority of genes required for the synthesis of GTA particles[21] (Fig. 1a). CcGTAs were shown to encapsulate host genomic DNA (~8.3 kb on average)[21], most likely through head-full packaging[15]. Unlike RcGTA, CcGTA synthesis is repressed under normal laboratory conditions by the transcriptional repressor RogA[21] (Fig. 1a), but CcGTA particles are produced in a subpopulation of cells once *rogA* is removed genetically[21]. RogA exerts its repression by binding directly to the core promoter region of the *gafYZ* operon (Fig. 1a), the two direct activators of the CcGTA cluster[21]. Notably, *C. crescentus* GafY and GafZ show sequence similarity to the N-terminal and C-terminal domains of *R. capsulatus* GafA, respectively[10,21]. *C. crescentus* GafY and GafZ were suggested to co-regulate the promoter upstream of *gtaT* (Fig. 1a), the first gene of the CcGTA gene cluster, to transcriptionally activate the cluster[21]. However, it is unknown how GafYZ controls transcription mechanistically, and how they coordinate expression of the main CcGTA gene cluster with the accessory genes elsewhere on the genome for CcGTA production.

To investigate, we combined a genetic screen, genome-wide deep sequencing, and biochemical approaches, and demonstrate that the *C. crescentus* GTA gene cluster is transcriptionally co-activated by integration host factor (IHF) and by GafYZ-mediated transcription anti-termination. We show that, in the absence of the repressor RogA, IHF directly binds to the promoter region of the *gafYZ* operon, and together with GafY, co-activates *gafYZ* transcription. Furthermore, IHF also binds upstream of the CcGTA cluster, and together with GafY and GafZ, mediates transcription activation and anti-termination to express the entire CcGTA cluster. We present evidence that GafZ is a novel transcription anti-

terminator that likely forms an anti-termination complex with RNA polymerase, NusA, NusG, and NusE to bypass transcription terminators identified within the CcGTA cluster. Overall, we reveal an exquisite two-tier regulation that coordinates the synthesis of GTA particles in *C. crescentus*.

RESULTS

Activation of the GTA gene cluster requires the integration host factor IHF

Previously, Gozzi *et al* (2022) showed that the production of *C. crescentus* GTA is tightly regulated by a repressor, RogA, which represses the transcription of the *gafYZ* operon, encoding the direct activators of GTA synthesis[21] (Fig. 1a). To further understand the control of the GTA cluster, we devised a screen based on a β-galactosidase reporter to identify additional factors that might act together with or independently of GafYZ. We engineered a transcriptional fusion of *lacA* to *gtaM* (encoding the GTA major capsid gene) in a Δ*lacA* background (Fig. 1a). The resulting strain produced white colonies on agar media containing X-gal (Fig. 1b), consistent with the previous finding that the GTA cluster is transcriptionally silent under normal lab conditions[21]. As expected, deleting *rogA* alleviated transcription repression of the GTA cluster[21], producing blue colonies (Fig. 1b). We transposon-mutagenized this Δ*rogA* i.e. GTA-on indicator strain and looked for rare white colonies. Confirming that the screen worked, we found transposon insertions within *gafY* and in the promoter of *gafYZ*, the operon encoding the two known GTA activators[21] (Fig. 1b). In addition, we found a single insertion in the small (291-bp) *ihfB* gene, which encodes the β subunit of integration host factor (IHF) (Fig. 1b). Complementation of this strain by expressing *ihfB* ectopically from the *xyl* locus restored β-galactosidase activity (Fig. 1b), suggesting that IHF is required to activate the GTA cluster. IHF is known to be a heterodimer, consisting of an α and a β subunit[37]. To investigate further, we deleted *ihfA* and *ihfB* individually from the GTA-on Δ*rogA* strain, then assayed for the presence of packaged GTA DNA and the head-tail connector protein (GtaL) as proxies for the synthesis of GTA particles (Fig. 1c-d). In the GTA-on Δ*rogA* strain, packaged GTA DNA can be seen as a distinct ~9-kb band in undigested total DNA samples[21] (Fig. 1c). Deletion of either *ihfA* or *ihfB* prevented detectable GTA DNA (Fig. 1c). Also, no detectable GtaL protein was produced in Δ*ihfA/B* Δ*rogA* strains, as assessed by an immunoblot using an α-GtaL antibody (Fig. 1d). GTA production was restored when *ihfA/B* was expressed ectopically from the *xyl* locus (Fig. 1c-d). Altogether, our data show that IHF is required for GTA synthesis in *C. crescentus*.

IHF binds the promoter regions of *gafYZ* and the GTA gene cluster to activate gene expression

IHF is a known DNA-binding and bending transcriptional factor[37–40], but its regulon has not been fully characterized in *C. crescentus*[41,42]. To further understand the roles IHF might play in GTA synthesis, we performed chromatin immunoprecipitation with deep sequencing (ChIP-seq) to map the genome-wide binding sites of FLAG-tagged IHFα and IHFβ in the GTA-off background (stationary phase WT *C. crescentus*) as well as the GTA-on background (stationary phase Δ*rogA C. crescentus*) (Supplementary fig. 1a). Non-tagged WT and Δ*rogA* strains were employed as negative controls to eliminate false signals that might arise from cross-reaction with the α-FLAG antibody (Supplementary fig. 1a). ChIP-seq revealed (i) there are numerous enriched IHF binding sites on the *C. crescentus* chromosome (430 peaks with $\log_2$(fold enrichment) > 2, -$\log_{10}$(q value) > 350), (ii) the targets of IHFα and IHFβ entirely overlap, consistent with IHF functioning as an αβ heterodimer, and (iii) the binding of IHF to DNA is independent of RogA (Supplementary fig. 1a). A closer inspection of α-FLAG-IHF ChIP-seq datasets revealed a single peak in the promoter region of the *gafYZ* operon (Fig. 2a), and five peaks in the GTA gene cluster in both Δ*rogA* and WT backgrounds (one of which, IBE1, is in the promoter region of the GTA cluster) (Fig. 2b). This suggested that IHF might directly regulate transcription of *gafYZ* and of the GTA gene cluster, a possibility that we confirmed by quantitative reverse transcriptase PCR (qRT-PCR) (Fig. 2c). Deletion of *rogA* caused an upregulation of *gafY* by ~30-fold, and of *gtaT* by ~600-fold, ultimately leading to the synthesis of GTA particles (Fig. 2c). When either *ihfA* or *ihfB* was deleted, the elevated gene expression of *gtaT* in the *ΔrogA* background was abolished, whereas that of *gafY* was reduced ~5-fold (Fig. 2c). The upregulation of *gtaT* and *gafY* in the *ΔrogA* background was restored when *ihfA* or *ihfB* was ectopically expressed from the *xyl* locus (Fig. 2c).

MEME analysis using the hundred most enriched sites in α-FLAG-IHF ChIP-seq datasets allowed us to propose a consensus DNA-binding motif for *C. crescentus* IHF (Supplementary fig. 1b), in turn allowing us to locate the IHF binding elements (IBEs) in the GTA promoter (IBE 1) and the *gafYZ* promoter (IBE 6) (Fig. 2a-b). We experimentally verified these IHF binding elements by replacing the conserved TT in the putative IHF binding site with GG, and then monitoring the binding of FLAG-tagged IHF by ChIP-seq. The TT→GG mutations selectively eliminated the enrichment of IHF at the DNA region under investigation (Fig. 2a-b), and nowhere else on the chromosome. When the IHF binding element in either the *gafYZ* promoter or the GTA promoter was mutated, GTA was no longer produced, as shown by the absence of both the packaged GTA DNA and GtaL protein (Fig. 2d-e).

We also observed four additional IHF binding elements (IBE 2-3-4-5) within the GTA cluster, in between the large terminase-encoding gene *gtaT* and the portal protein-encoding gene *gtaP* (Fig. 2b). To investigate the contribution of these sites to GTA synthesis, we eliminated them by making TT→GG/TG mutations while maintaining the coding sequence of the underlying *gtaQS* genes (Supplementary fig. 2a). ChIP-seq of FLAG-IHF confirmed that such mutations indeed eliminated IHF binding to the four elements internal to the GTA cluster (Supplementary fig. 2a), however, the resulting strains were still able to produce GTA particles, as suggested by the presence of packaged GTA DNA and GtaL protein (Supplementary fig. 2b-c). We concluded that these four internal IHF binding sites play little or no role in the transcription activation of the GTA cluster. Altogether, our results suggest that IHF activates transcription of the GTA phage cluster by binding directly to the promoter of the GTA cluster, and to the promoter of the *gafYZ* operon, which encodes the two direct activators of the GTA cluster.

IHF binds to GafY-regulated promoters

Our previous α-GafY ChIP-seq showed that GafY binds 18 sites on the chromosome, in addition to the promoter regions of *gafYZ* and the GTA cluster itself[21]. The ChIP-seq datasets showed that 18 out of 20 GafY-binding DNA regions are also co-occupied by IHF (Supplementary fig. 3). This suggested that IHF and GafY might work together to activate transcription, a possibility that we investigated using the promoter region of the GTA cluster.

Multiple binding elements at the GTA cluster promoter region are required for transcriptional activation

To investigate how IHF and GafYZ might work together to activate the transcription of the GTA cluster, we used RNA-seq to locate the possible transcription start sites (TSS) of the GTA cluster in the GTA-on Δ*rogA* strain (Fig. 3a). We discovered a single TSS, labeled A+1, in the upstream region of the GTA cluster (Fig. 3a). This identified TSS, however, lies 110 bp downstream of the computationally annotated start codon (TTG) of *gtaT*, suggesting mis-annotation (Fig. 3a). Indeed, mutating this TTG to CTG did not affect GTA synthesis, as judged by the presence of packaged GTA DNA in total DNA purification (Fig. 3b). Furthermore, deleting a DNA region ranging from +90 to +214, which would have rendered the remaining downstream region of *gtaT* out of frame (if TTG was a correct start codon), did not affect GTA production either (Fig. 3b). Downstream of *gtaT* TSS, we identified three possible in-frame ATG start codons (Fig. 3a). Individually mutating these candidates ATG codons to a stop codon (TGA) showed that only ATG number 3 (at position +251) was required for the production of packaged GTA DNA, suggesting it represents the true *gtaT* start codon (Fig. 3b).

The identification of the TSS also allowed us to identify the likely -10 (TATCTTT) and -35 (AATCAT) of the GTA promoter (Fig. 3a). Using a *lacZ* reporter construct containing DNA from -153 to +229 relative to the *gtaT* TSS, we introduced mutations into the putative -10 and -35 elements (Fig. 3c). The β-galactosidase activities were abolished or reduced in our mutational analysis (Fig. 3c), thus establishing the importance of the -10 and -35 elements for the promoter activity of *gtaT*.

Next, we sought to define a consensus GafY binding motif and the position of the GafY binding site in relation to the IHF binding element (IBE1) within the GTA promoter. Due to the small number of enriched regions in the α-GafY ChIP-seq dataset, it was difficult to predict a cognate GafY binding motif from MEME analysis. To overcome this, we sought to identify the minimal region of the GTA promoter that was responsive to GafY. To do so, we constructed a series of GTA promoter-*lacZ* fusions that contained progressively truncated DNA flanking the TSS (Fig. 4a). A DNA region ranging from -153 to +229 relative to the *gtaT* TSS, when fused to *lacZ*, produced a background level of β-galactosidase activity in the GTA-off WT background (Fig. 4a). In contrast, the same construct gave a high β-galactosidase activity in the GTA-on Δ*rogA* background (the activators GafYZ are produced in the absence of the RogA repressor, thus switching on GTA transcription) (Fig. 4a). This result suggested that the -153 to +229 region contains all the elements necessary for GafYZ-mediated transcriptional activation. In the Δ*rogA* background, the β-galactosidase activities did not change drastically when 10, 20, or 30 bp were truncated from the upstream end of the P_{gtaT} (-153 to +229)-*lacZ* fusion constructs (Fig. 4a). However, truncation of 40 bp reduced the β-galactosidase activity by ~20-fold (Fig. 4a), almost to the background level, suggesting a key regulatory element, possibly a GafY binding element, is in the DNA region from -123 to -113 relative to the *gtaT* TSS (Fig. 4a). Indeed, when mutations (P_{gtaT} YBE*: ACTATG→TGACCC) were introduced into this putative GafY binding element (YBE), GafY binding was eliminated as assessed by α-GafY ChIP-seq (Fig. 4b), providing further evidence that this represents the site of GafY binding.

GafY interacts directly with the sigma factor RpoD

In the promoter of the GTA cluster, the IHF binding element is positioned between the core -10 -35 promoter and the far upstream GafY binding element, consistent with the classical role of IHF in bending DNA to facilitate the recruitment of RNA polymerase to the core promoter by an upstream transcriptional activator. We hypothesized that GafY might interact directly with a component of the RNA polymerase (RNAP) holoenzyme to activate transcription[43,44]. To identify potential interacting partners of GafY, we conducted an *in silico* AlphaFold2-Multimer-based protein interaction screen[45,46] between GafY and 23 core components of RNAP and sigma factors in *C. crescentus* (Fig. 5a). Based on confidence metrics (ipTM) generated by AlphaFold2, sigma factor 70 (RpoD), specifically its domains 3-4, were predicted to interact with GafY (Fig. 5a). To investigate this potential GafY-RpoD interaction, we performed a co-immunoprecipitation (co-IP) experiment using FLAG-tagged RpoD as bait. We observed that GafY was indeed pulled down with FLAG-RpoD, while another DNA-binding protein, ParB[47], serving as a negative control, was not (Fig. 5b). Furthermore, to assess whether GafY and RpoD form a direct complex, we co-overexpressed non-tagged GafY and His6-tagged *C. crescentus* RpoD in *Escherichia coli* (*E. coli*) (Fig. 5c). GafY was found to co-purify with His-tagged RpoD on a nickel affinity column (Fig. 5c). Similarly, GafY co-purified with domains 3-4 of RpoD alone, showing that GafY and RpoD interact directly via domains 3-4 of the sigma factor (Fig. 5c). Consistent with these observations, the promoter of the GTA cluster, as well as in the majority of the other GafY target promoters, were enriched in α-FLAG-RpoD ChIP-seq experiments (Fig. 5d). Altogether, we suggest that GafY interacts with RpoD to recruit RNAP holoenzyme to the GTA promoter.

Evidence that GafY and GafZ form a transcription elongation complex with RNAP, NusA, NusG, and NusE

We noted that the majority of α-GafY ChIP-seq peaks are symmetrical and confined to intergenic regions (Supplementary fig. 2). By contrast, the α-GafY ChIP-seq peak at the GTA cluster locus is strikingly asymmetric, extending downstream of the promoter for more than 14 kb (Fig. 6a). Similarly, the α-GafZ ChIP-seq peak at the GTA promoter (its only target) showed the same pattern, extending downstream across the entire GTA cluster[21] (Fig. 6a). We hypothesized that GafY and GafZ associate with RNAP as it transcribes the GTA cluster. To further investigate this possibility, we performed α-FLAG ChIP-seq experiments in the GTA-on Δ*rogA* strain harboring a FLAG-tagged β' (RpoC) subunit of RNAP and compared the results to that from a negative control of a non-tagged Δ*rogA* strain (Fig. 6a). In parallel,

we also performed α-VSVG ChIP-seq experiments with Δ*rogA* cells expressing VSVG epitope-tagged NusA, NusG, or NusE, using the appropriate non-tagged negative controls (Fig. 6a). In all cases, the ChIP-seq profiles are highly similar and correlated, with signals spreading into and across the entire GTA gene cluster (Fig. 6a). These data suggest that GafYZ, together with RNAP and Nus proteins, might form an elongation complex that transcribes the main GTA gene cluster. To investigate this possibility further, we engineered Δ*rogA* FLAG-*gafZ* strains that harbor individually VSVG-tagged NusA, NusG, NusE, and performed co-IP using FLAG-tagged GafZ as bait (Fig. 6b). We observed an enrichment of Nus proteins in the IP fraction compared to the pre-IP input control (Fig. 6b). Another DNA-binding protein, ParB, was not pulled down in the IP fraction (Fig. 6b), suggesting that the enrichment of Nus proteins was specific.

The putative GafZ binding element (ZBE) is located in between the -10 and -35 promoter elements of the GTA gene cluster

Despite the previous observation that GafY and GafZ interact with each other[21], unlike GafY, GafZ has only one target - the promoter of the GTA cluster[21] (Fig. 6a). This suggested there must be a GafZ binding element that recruits GafZ to the promoter of the GTA cluster. Further, AlphaFold2 and FoldSeek[45,48] predict that GafZ contains a sigma-factor-like helix-turn-helix motif, consistent with direct DNA binding (Supplementary fig. 4). We inspected the sequence of the GTA promoter for inverted repeats that might represent a GafZ binding element (ZBE), and identified a near-perfect palindrome (GCGCCCG-CTGGCGC) in between the -10 and -35 elements (Fig. 7a). Next, we introduced mutations (P_{gtaT} ZBE*: CTGGCGC→TTTTTTC) to the right half of this palindrome and assayed for FLAG-GafZ binding by α-FLAG ChIP-seq (Fig. 7a). The enrichment of FLAG-GafZ was reduced to background level, while the enrichment of GafY in the upstream region of the promoter (as assessed by α-GafY ChIP-seq on the same strain) was unaffected (Fig. 7a). Furthermore, the P_{gtaT} ZBE* mutations also eliminated the production of GTA-packaged DNA and GtaL protein (Fig. 7b). Taken together, these findings demonstrated that the inverted repeat positioned between the -10 and -35 elements of the *gtaT* promoter is critical for GafZ binding and GTA synthesis.

GafYZ allows RNAP to bypass a transcription terminator located downstream of the first gene in the GTA gene cluster

We noted a sharp reduction in signal in ChIP-seq datasets of α-FLAG-GafZ, α-GafY, α-FLAG-β' RNAP, and α-VSVG-NusAEG immediately downstream of *gtaT*, the first gene of the GTA cluster (Fig. 6a and Fig. 8a). Focusing on this region, we discovered three 150-bp long GC-rich direct repeats, each of which contains imperfect inverted repeats. We hypothesized that these repeats might form a putative transcription terminator (ter_{GTA}) (Fig. 8a and Supplementary fig. 5a). Given the association of GafYZ with the RNAP elongation complex and the strict requirement of GafZ for transcription of the GTA cluster, we reasoned that GafYZ might enable RNAP to read through this putative ter_{GTA} terminator. To investigate these possibilities, we determined the effect of this putative ter_{GTA} on transcription activities of a strong GafYZ-independent promoter (P_{rsaA}, driving the expression of the most abundant S-layer protein-encoding gene, *rsaA*, in *C. crescentus*[49,50]) and of a GafYZ-dependent *gtaT* promoter, using promoter-*lacZ* fusion reporters (Fig. 8b). We fused a DNA region ranging from -153 to +229 relative to the *gtaT* TSS to *lacZ*. To assay for β-galactosidase activity, reporter plasmids were introduced into three genetic backgrounds, namely a WT GTA-off strain (GafY- GafZ-), a Δ*rogA* GTA-on strain (GafY+ GafZ+), and a Δ*rogA*Δ*gafZ* GTA-off strain (GafY+ GafZ-) (Fig. 8b).

A P_{rsaA}-*lacZ* fusion construct produced high levels of β-galactosidase activity in all three genetic backgrounds (Fig. 8b), consistent with P_{rsaA} being independent of GafYZ (no GafY or GafZ binding sites were found in the upstream region of *rsaA* by ChIP-seq). Insertion of the three 150-bp repeats in between P_{rsaA} and *lacZ* reduced β-galactosidase activity to background levels regardless of the presence of GafYZ (Fig. 8b), confirming that the tandem repeats constitute a strong transcription terminator.

A P_{gtaT} (-153 +229)-*lacZ* fusion was active in a GTA-on Δ*rogA* (GafY+ GafZ+) background but was reduced to the background level when both GafY and GafZ were absent, and ~4-fold reduced when only GafY was present i.e. in the Δ*rogA*Δ*gafZ* GTA-off (GafY+ GafZ-) background (Fig. 8b). An insertion of the putative ter_{GTA} in between P_{gtaT} (-153 +229) and *lacZ* in the GTA-on Δ*rogA* (GafY+ GafZ+) background reduced the β-galactosidase activity by ~8-fold but, crucially, did not eliminate all transcriptional activity. On the other hand, only background level β-galactosidase activity was detected for the same construct when GafZ was absent (Fig. 8b). We further observed a similar anti-termination response by GafZ when we replaced the putative ter_{GTA} by a well-characterized T1 transcription terminator from the *E. coli rrnB*[51] (Supplementary fig. 5b). Overall, our results demonstrated that the tandem repeats function as a transcription terminator, and that GafZ acts as a transcription anti-terminator, allows some (but not all) RNAP to read through and transcribe the entire GTA cluster.

DISCUSSION

In this study, we demonstrate that the *C. crescentus* GTA gene cluster is transcriptionally co-activated by IHF and GafY, and by a GafYZ-mediated transcription anti-termination. We show that, in the absence of the repressor RogA, IHF and GafY directly bind to the promoter region of the *gafYZ* operon, the promoter of the main CcGTA cluster, and ~18 promoters of accessory GTA genes, to activate transcription (Fig. 9a). Our findings support a model that GafY interacts directly with domain 3-4 of the housekeeping sigma factor 70 (RpoD) to recruit RNAP to the promoter, and that the recruitment of RNAP is facilitated by IHF binding (Fig. 9b). We also present evidence that GafZ is a novel transcription anti-terminator that likely forms an anti-termination complex with RNA polymerase, NusA, NusG, and NusE to bypass transcription terminators downstream of the first gene in the CcGTA cluster, resulting in transcription of the entire CcGTA cluster gene (Fig. 9b). The mapping of the GafZ binding element (ZBE) between the -10 and -35 core promoter elements of the GTA cluster suggests a model in which GafZ binding to the ZBE enables this anti-terminator to load onto RNAP, allowing RNAP to bypass the downstream transcriptional terminators (Fig. 9b).

GafY and GafZ show sequence homology to the N-terminal and C-terminal domain, respectively, of *R. capsulatus* GafA, the direct activator of RcGTA[34]. Sherlock and Fogg (2022) demonstrated that the central region in between the N-terminal and C-terminal domains of GafA interacts with the omega (RpoZ-ω) subunit of RNAP, thereby recruiting RNAP to activate the RcGTA cluster[35]. In *C. crescentus*, however, we have not observed the enrichment of VSVG-tagged RpoZ-ω in a co-IP using FLAG-tagged GafZ as bait (Supplementary fig. 6). A sequence alignment of GafA and GafY-Z shows that the central region of GafA has the least similarity to a fusion of GafY and GafZ (Supplementary fig. 7). Furthermore, several predicted loops in the central region of GafA are missing in the *C. crescentus* GafYZ fusion (Supplementary fig. 7). This likely explains why *C. crescentus* GafYZ do not appear to interact with RpoZ-ω. In *R. capsulatus*, (p)ppGpp, which is likely to interact directly with RNAP via RpoZ-ω, contributes to the synthesis of GTA particles[35,52]. For example, deletion of *rpoZ* or *relA/spoT* (responsible for the synthesis of (p)ppGpp) in *R. capsulatus* reduced GTA synthesis by five-fold[52], while deletion of the sole *relA/spoT* homolog in *C. crescentus*[53,54] only reduces CcGTA production two-fold (Supplementary fig. 6). Future work is necessary to better understand the possible role of ppGpp(p) and/or RpoZ-ω in CcGTA production in *C. crescentus*.

Although the exact binding elements for *R. capsulatus* GafA have not been mapped at nucleotide resolution, Sherlock and Fogg (2022) showed using EMSA assays that C-terminal domain of GafA (equivalent to *C. crescentus* GafZ) binds to a DNA fragment covering the -10, -35 and TSS of the RcGTA cluster[35]. Given that *R. capsulatus* GafA and *C. crescentus* GafZ are homologous, and the similar location of their DNA-binding elements, it is possible that *R. capsulatus* GafA might have previously unrecognized transcription anti-termination activity.

In this study, we have also identified a DNA element (ZBE) that is important for the regulation of the main CcGTA cluster by GafZ. This putative ZBE is located between the -10 and -35 elements of the CcGTA cluster promoter, similar to the location of the binding elements for well-characterized processive anti-terminators such as Q from bacteriophage λ and AlpA from *Pseudomonas. aeruginosa*[55,56]. The ChIP-seq profiles of *C. crescentus* GafZ are also reminiscent of the binding patterns of RNAP and transcription anti-terminators such as *P. aeruginosa* AlpA, *E. coli* RfaH, and λQ[55,57–59]. The λQ binding element (QBE) and the AlpA binding element (ABE) help the direct loading of Q and AlpA onto RNAP[55,56,60–63], and it seems likely that the putative GafZ binding element (ZBE) in the GTA promoter may help loading of GafZ onto RNAP in a similar way. The cryo-EM structures of λQ- and AlpA-bound RNAP have been solved, showing that these anti-terminators form a molecular nozzle near the RNA exit channel of RNAP to prevent the formation of terminator hairpin structures that would otherwise form in the nascent RNA and thereby impede or stop transcription elongation[56,60,64]. AlphaFold-predicted structures of GafZ show no sequence or structural similarity to these known anti-terminators, future work will be required to determine how GafZ mechanistically modifies RNAP to mediate transcription anti-termination.

Concluding remarks

The discovery that *C. crescentus* produces *bona fide* GTA particles offered a new and highly tractable organism to dissect the function, biosynthesis, and regulation of these enigmatic genetic elements[21]. Here, we revealed that GTA cluster gene expression is controlled by both transcriptional activation and by anti-termination. While most of the control, as revealed in this study, is at the cluster-specific level, it is interesting to note that a host factor, IHF, is co-opted in the activation of the CcGTA gene cluster. In future work, we hope to gain further insight into the extent to which GTA is domesticated and integrated with the host's physiology, potentially shedding light on the evolution of such domestication. Lastly, *C. crescentus* GTA is not produced under normal laboratory conditions[21], and so it was necessary for us to exploit a Δ*rogA* strain to induce CcGTA production in this study. Finding the environmental or physiological signals that naturally de-repress RogA or activate GafYZ expression, if they exist, will be illuminating in understanding the benefit of GTA to the host and how such exaptation can evolve.

MATERIALS AND METHODS

Strains, media, and growth conditions

E. coli and *C. crescentus* were grown in LB and PYE, respectively. When appropriate, media were supplemented with antibiotics at the following concentrations (liquid/solid media for *C. crescentus*; liquid/solid media for *E. coli* [μg/mL]): kanamycin (5/25; 30/50); spectinomycin (25/100; 50/50); oxytetracycline (1/2; 12/12). All strains, plasmids, and oligonucleotides used in this study are listed in Table S2. Details on constructions of plasmids and strains are in the Supplementary Materials and Methods.

Transposon (Tn5) mutagenesis

The Tn5 transposon delivery plasmid (pMCS1-Tn5-ME-R6Kγ-kanamycin^{R}-ME)[65] was conjugated from an *E. coli* S17-1 donor into *C. crescentus ΔrogA ΔlacA gtaM::gtaM-lacA* cells. Briefly, *E. coli* S17-1 was transformed with the transposon delivery plasmid and plated out on LB plates supplemented with kanamycin. On the next day, colonies forming on LB + kanamycin were scraped off the plates and resuspended in PYE to OD_{600} of 1.0. Cells were pelleted down and resuspended in fresh PYE twice to wash off residual antibiotics. 100 μl of cells were mixed with 1000 μl of exponentially growing *C.crescentus ΔrogA ΔlacA gtaM::gtaM-lacA* cells then the mixture was centrifuged at 13 000 rpm for 1 min. The cell pellet was subsequently resuspended in 50 μl of fresh PYE and spotted on a nitrocellulose membrane resting on a fresh PYE plates. PYE plates with nitrocellulose disks were incubated at 30°C

for 5 h before being resuspended by vortexing vigorously in fresh PYE liquid to release bacteria. Resuspended cells were plated out on Petri disks containing PYE agar supplemented with kanamycin and carbenicillin and 40 µg/ml X-gal, and incubated for 3 days at 30°C. After 3-day incubation, white colonies were picked and restruck on PYE + kanamycin + carbenicillin + X-gal to purify. To locate the Tn5 insertion point, genomic DNA from mutants of interest was extracted. 4 µg of extracted genomic DNA was partial digested with Sau3AI in 50 µL reaction and re-ligated with T4 DNA ligase. The ligation mixture were ethanol precipitated and introduced into *E.coli pir116* cells by electroporation. Colonies carrying the re-ligated Tn5 plasmid grew on kanamycin and the plasmid was subsequently extracted and sequenced with oligo KAN-2 FP-1: ACCTACAACAAAGCTCTCATCAACC and R6KAN-2 RP-1:CTACCCTGTGGAACACCTACATCT to map the position of Tn5 insertion on the *C. crescentus* chromosome.

Chromatin immunoprecipitation with deep sequencing (ChIP-seq)

C. crescentus cell cultures (50 mL) were grown in PYE to a stationary phase and fixed with formaldehyde to a final concentration of 1%. Fixed cells were incubated at room temperature for 30 min, then quenched with 0.125 M glycine for 15 min. Cells were washed three times with 1x PBS (pH 7.4) and resuspended in 1 mL of buffer 1 (20 mM K-HEPES pH 7.9, 50 mM KCl, 10% glycerol, and Roche EDTA-free protease inhibitors). Subsequently, the cell suspension was sonicated on ice using a Soniprep 150 probe-type sonicator (11 cycles, 15s ON, 15s OFF, at setting 8) to shear the chromatin to below 1 kb, and the cell debris was cleared by centrifugation (20 min at 13,000 rpm at 4ºC). The supernatant was then transferred to a new 2 mL tube and the buffer conditions were adjusted to 10 mM Tris-HCl pH 8, 150 mM NaCl and 0.1% NP-40. Fifty microliters of the supernatant were transferred to a separate tube for control (the input fraction) and stored at -20ºC. In the meantime, antibodies-coupled beads were washed off storage buffers before being added to the above supernatant. We employed α-VSV-G antibody coupled to sepharose beads (Merck) for ChIP-seq of NusG-VSVG, VSVG-NusA, VSVG-NusE, and α-FLAG antibody coupled to agarose beads (Merck) for ChIP-seq of RpoC-FLAG and FLAG-GafZ.

Briefly, 50 µL of α-VSVG beads or 100 µL α-FLAG beads was washed off storage buffer by repeated centrifugation and resuspension in IPP150 buffer (10 mM Tris-HCl pH 8, 150 mM NaCl and 0.1% NP-40). Beads were then introduced to the cleared supernatant and incubated with gentle shaking at 4ºC overnight. For α-GafY ChIP-seq experiments, protein A sepharose beads (Merck) were incubated with the cleared supernatant for an hour to remove non-specific binding. Afterward, the cleared supernatant was retrieved and incubated with 50 µL of α-GafY polyclonal antibody overnight. On the next day, protein A sepharose beads were added and incubated for 4 h to capture GafY-DNA complexes. Beads were then washed five times at 4°C for 2 min each with 1 mL of IPP150 buffer, then twice at 4ºC for 2 min each in 1x TE buffer (10 mM Tris-HCl pH 8 and 1 mM EDTA). Protein-DNA complexes were then eluted twice from the beads by incubating the beads first with 150 µL of the elution buffer (50 mM Tris-HCl pH 8, 10 mM EDTA, and 1% SDS) at 65ºC for 15 min, then with 100 µL of 1X TE buffer + 1% SDS for another 15 min at 65ºC. The supernatant (the ChIP fraction) was then separated from the beads and further incubated at 65ºC overnight to completely reverse crosslinks. The input fraction was also de-crosslinked by incubation with 200 µL of 1X TE buffer + 1% SDS at 65ºC overnight. DNA from the ChIP and input fraction were then purified using a PCR purification kit (Qiagen) according to the manufacturer's instruction, and then eluted out in 40 µL water. Purified DNA was then constructed into libraries suitable for Illumina sequencing using the NEXT Ultra II library preparation kit (NEB). ChIP libraries were sequenced on the Illumina Hiseq 2500 or Nextseq 550 at the Tufts University Genomics facility.

Processing ChIP-seq data

For analysis of ChIP-seq data, Hiseq 2500 or NextSeq 550 Illumina short reads (50 bp/75bp) were mapped back to the *C. crescentus* NA1000 reference genome (NCBI Reference Sequence: NC-011916.1) or appropriate reference genomes with mutations at the IBE/YBE/ZBE, using Bowtie 1[66] and

the following command: bowtie -m 1 -n 1 –best –strata -p 4 –chunkmbs 512 NA1000-bowtie –sam *.fastq > output.sam. Subsequently, the sequencing coverage at each nucleotide position was computed using BEDTools[67] using the following command: bedtools genomecov -d -ibam output.sorted.bam -g NA1000.fna > coverage_output.txt. When necessary, MACS2 were employed to call peaks[68]. Finally, ChIP-seq profiles were plotted with the x-axis representing genomic positions and the y-axis is the number of reads per base pair per million mapped reads (RPBPM) or number of reads per kb per million mapped reads (RPKPM) using custom R scripts. For the list of ChIP-seq datasets in this study, see Supplementary Table S3.

Co-immunoprecipitation (Co-IP)

C. crescentus cells (25 mL) were grown at 28°C to a stationary phase before cells were harvested by centrifugation. Cell pellets were washed with 1× PBS (pH 7.4), resuspended in 1 mL of lysis buffer (50 mM Tris-HCl pH8, 150 mM NaCl, 1% Triton X-100, EDTA-free protease inhibitors, 10 mg/mL lysozyme, and 1 µL of Benzonase), and incubated at 37°C for 20 min. Subsequently, the cell suspension was sonicated on ice using a probe-type sonicator (6 cycles of 15 s with 15 s resting on ice, amplitude setting 8). The lysate was cleared from the cell debris by centrifugation (13,000 rpm for 20 min at 4°C). 50 µL of this supernatant (the input fraction) was kept for downstream immunoblot analysis. The remaining supernatants were adjusted so that they had the same amount of total protein and mixed with 25 µL α-FLAG magnetic bead as instructed in the µMACS Epitope Tag Protein Isolation Kit (Miltenyi Biotec). From here, all the subsequent steps were performed according to the instructions from the µMACS kit. The immunoprecipitated proteins (the IP fraction) were eluted using 50 µL of elution buffer (50 mM Tris-HCl pH 6.8, 50 mM DTT, 1% SDS, and 1 mM EDTA).

For western blot analysis related to NusA, NusG and NusE, 10 µg of the input fraction, or 20 µL of the IP fraction for α-VSVG immunoblots, 10 µL of the IP fraction for α-ParB, 5 µL of the IP fraction for α-FLAG immunoblots were loaded on a 4-20% Novex WedgeWell SDS-PAGE gels (Thermo Fisher Scientific). For western blot analysis related to RpoD, 10 µg of the input fraction, 10 µL of the IP fraction for α-GafY immunoblots, 5 µL of the IP fraction for α-ParB, 5 µL of the IP fraction for α-FLAG immunoblots were loaded on a 4-20% Novex WedgeWell SDS-PAGE gels (Thermo Fisher Scientific). Resolved proteins were transferred to polyvinylidene fluoride (PVDF) membranes using the Trans-Blot Turbo Transfer System (BioRad), and the membrane was incubated with a 1:5000 dilution of an α-VSVG antibody, 1:2,500 dilution of an α-FLAG antibody (Merck), or a 1:5,000 dilution of an α-ParB polyclonal antibody (custom antibody, Cambridge Research Biochemicals, UK), or a 1:300 dilution of an α-GafY polyclonal antibody (custom antibody, Cambridge Research Biochemicals, UK). Subsequently, the membranes were washed twice in a 1× TBS + 0.005% Tween-20 buffer before being incubated in a 1:10,000 dilution of an HRP-conjugated secondary antibody. Blots were imaged using an Amersham Imager 600 (GE Healthcare).

Growth conditions for IHF-related experiments

C. crescentus cells (20 mL) were grown in PYE in the presence or absence of 0.3% xylose, as appropriate for each experiment, for 3 hrs. Cell pellets were collected from 5 mL cultures for RNA extraction and RT-qPCR analysis. Another 5 mL was collected for immunoblot analysis using α-CCNA03882 (GtaL, a GTA head-tail connector protein) polyclonal antibody. Another 5 mL was collected for total genomic DNA extraction.

Genomic DNA extraction

Cell pellets were resuspended in 300 µL cell lysis solution (Qiagen) and lysed by incubation at 50°C for 10 min. A total of 50 µg of RNaseA was added to the cell lysate, and the lysate was incubated at 37°C for an hour to remove cellular RNA. Proteins were precipitated by adding 100 µL of a protein precipitation solution (Qiagen). Samples were centrifuged for 5 min at 13000 rpm, the resulting supernatant was then

mixed with 600 µL of isopropanol, and the tubes were mixed by gentle inversions to precipitate genomic DNA. Genomic DNA was pelleted via centrifugation at 13,000 rpm for 5 min, washed once with 70% ethanol, and resuspended in 200 µL of water.

Immunoblot to detect GtaL

Cell pellets were resuspended in 300 µL of buffer 1 (20 mM K-HEPES pH 7.9, 50 mM KCl, 10% glycerol and Roche EDTA-free protease inhibitors). Samples were sonicated on ice using a probe-type sonicator (4 cycles of 15 s with 15 s resting on ice, amplitude setting 8). The lysate was cleared from the cell debris by centrifugation at 13,000 rpm for 20 min at 4°C. Bradford assay was used to determine the total protein concentration in each sample so that an equal amount of total proteins were loaded on each well of a 4-20% Novex WedgeWell SDS-PAGE gels (Thermo Fisher Scientific). Resolved proteins were transferred to PVDF membranes using the Trans-Blot Turbo Transfer System (BioRad), and the membrane was incubated with a 1:1000 dilution of an α-GtaL polyclonal antibodies (custom antibody, Cambridge Research Biochemicals, UK). Subsequently, the membranes were washed twice in a 1×TBS + 0.005% Tween-20 buffer before being incubated in a 1:10,000 dilution of an HRP-conjugated secondary antibody. Blots were imaged using an Amersham Imager 600 (GE Healthcare).

β-galactosidase assay

C. crescentus cultures (20 mL), inoculated from single colonies, were grown at 28°C to stationary phase and cooled on ice before β-galactosidase assay. Cultures were diluted fourfold before measuring OD_{600} and 200 µL of diluted cultures were used in the assay. B-galactosidase assays were carried out essentially as described previously[69]. Assays were performed at least twice in duplicates on two different days.

Total RNA extraction and quantitative reverse transcriptase PCR (qRT-PCR)

C. crescentus cells (20 mL) were grown at 28°C to stationary phase and cell pellets from 5 mL cultures were collected for total RNA extraction using a Direct-zol RNA miniprep kit (Zymo Research). 10 µg of isolated total RNA was subjected to DNaseI treatment with 20 units of Turbo DNaseI (Invitrogen) for an hour at 37ºC. DNaseI was subsequently removed from total RNA using an RNA clean and concentrator-25 (Zymo Research). Purified RNA isolated from wild-type and *ΔrogA C. crescentus* was sent to Azenta (UK) for RNA-seq. For qRT-PCR, 1 µg of DNaseI-treated total RNA was converted to cDNA using an Invitrogen SuperScript III First-Strand Synthesis SuperMix for qRT-PCR according to the manufacturer's instructions. PCR cycling was performed at 25°C for 10 min, 42°C for 120 min, 50°C for 30 min, 55°C for 30 min, then 5 min at 85°C. Following RNaseH treatment, samples were diluted 1:2 with water and 1 µL was used for qRT-PCR using a SYBR® Green JumpStart™ *Taq* ReadyMix™ in a BIORAD CFX96 instrument. Results were analyzed using BIORAD CFX96 software. Transcript quantities for *gafY and gtaT* were determined relative to the amount of *ruvA* transcript, which was selected for being constitutively expressed in the cell growth conditions. Relative expression values were calculated by using the comparative Ct method (ΔΔCt) and were the average of two biological replicates. Error bars represent the relative expression values calculated from plus or minus one standard deviation from the mean ΔΔCt values. qRT-PCR oligos for quantifying *gafY* transcription are 5'-GCAGCTCGCCATCTACC-3' and 5'- GCAGATCCTCGATCTTGCG-3', for that of *gtaT* (CCNA02880) are 5'-GGCCCTGTACGAGCAAG-3' and 5'-GGCTGTGTTCCAGATCTCC -3', for that of *ruvA* are 5'-ATGGGCGTCGGCTATCT-3' and 5'-CGAGTGAGGAAGCCGTAGA-3'.

Protein co-overexpression and purification of 6xHis-tagged RpoD, 6xHis-tagged RpoD (domain 3+4), and GafY

Plasmid pCOLA-Duet1::6xhis-*rpoD-gafY*, or pCOLA-Duet1::6xhis-*rpoD* (domain 3+4) and pET15:: *gafY* were introduced into *E. coli* Rosetta (BL21 DE3) competent cells (Merck) by heat-shock transformation or by electroporation. A 10 mL overnight culture was used to inoculate 1 L of LB medium supplemented with kanamycin and chloramphenicol. Cells were grown at 37°C with shaking at 210 rpm to an OD_{600} of

~0.4. The culture was then left to cool down to 28°C before isopropyl-β-D-thiogalactopyranoside (IPTG) was added to a final concentration of 1 mM. The culture was left shaking for an additional 3 hrs at 28°C before cells were harvested by centrifugation. Pelleted cells were resuspended in a buffer containing 100 mM Tris-HCl pH 8.0, 300 mM NaCl, 10 mM imidazole, 5% (v/v) glycerol, 1 µL of Benzonase nuclease (Merck), and an EDTA-free protease inhibitor tablet (Merck). The resuspended cells were then lysed by sonication (10 cycles of 15 s with 10 s resting on ice in between each cycle). The cell debris was pelleted by centrifugation at 28,000 g for 30 min and the supernatant was filtered through a 0.45 µm filter disk. The lysate was then incubated with 2 mL pre-washed HIS-Select Cobalt Affinity Gel (Merck, UK) with rotation for an hour. After an hour, the resin was washed three times with 25 mL buffer A (100 mM Tris-HCl pH 8.0, 300 mM NaCl, 10 mM imidazole, and 5% glycerol). Proteins were eluted from the gel using 2 mL of buffer B (100 mM Tris-HCl pH 8.0, 300 mM NaCl, 500 mM imidazole, and 5% glycerol).

An *in silico* screen for protein-protein interactions

A pairwise screen for possible interactions between GafY and *C. crescentus* sigma factors and components of RNA polymerase was conducted using AlphaFold2 Multimer[45] via ColabFold[46]. The confidence metrics (ipTM) for the top model from each pairwise interaction were tabulated, with ipTM > 0.7 indicating a possible protein-protein interaction[70].

ACKNOWLEDGEMENTS

We thank members of our laboratory, Mark Buttner, Dave Grainger, and Paul Fogg for helpful discussion and comments on this manuscript, and Tom McLean for assistance with an AlphaFold2-based screen. This work is supported by the Royal Society University Fellowship Renewal URF\R\201020, the Lister Institute fellowship, the Wellcome Trust Investigator grant 221776/Z/2/Z (to T.B.K.L), and the BBSRC-funded Institute Strategic Program Harnessing Biosynthesis for Sustainable Food and Health (HBio) (BB/X01097X/1).

CONFLICTS OF INTEREST

Non declared

DATA AVAILABILITY

All relevant data are within the manuscript and supplementary files. ChIP-seq data are available at GEO (accession code: GSE247216).

AUTHOR CONTRIBUTIONS

N.T.T and T.B.K.L conceived the project. N.T.T carried out all experiments and conducted data analysis. T.B.K.L contributed to ChIP-seq analysis. T.B.K.L procured funding and supervised the project. N.T.T and T.B.K.L wrote the manuscript.

REFERENCES

1. Soucy, S. M., Huang, J. & Gogarten, J. P. Horizontal gene transfer: building the web of life. *Nat Rev Genet* **16**, 472–482 (2015).
2. Daubin, V. & Szöllősi, G. J. Horizontal Gene Transfer and the History of Life. *Cold Spring Harb Perspect Biol* **8**, a018036 (2016).
3. Lang, A. S., Westbye, A. B. & Beatty, J. T. The Distribution, Evolution, and Roles of Gene Transfer Agents in Prokaryotic Genetic Exchange. *Annu Rev Virol* **4**, 87–104 (2017).
4. Lang, A. S., Zhaxybayeva, O. & Beatty, J. T. Gene transfer agents: phage-like elements of genetic exchange. *Nat Rev Microbiol* **10**, 472–482 (2012).
5. Nagakubo, T. Biological Functions and Applications of Virus-Related Bacterial Nanoparticles: A Review. *Int J Mol Sci* **23**, 2595 (2022).
6. Heiman, C. M., Vacheron, J. & Keel, C. Evolutionary and ecological role of extracellular contractile injection systems: from threat to weapon. *Front Microbiol* **14**, 1264877 (2023).
7. Hantak, M. P., Einstein, J., Kearns, R. B. & Shepherd, J. D. Intercellular Communication in the Nervous System Goes Viral. *Trends Neurosci* **44**, 248–259 (2021).
8. Solioz, M. & Marrs, B. The gene transfer agent of Rhodopseudomonas capsulata. Purification and characterization of its nucleic acid. *Arch Biochem Biophys* **181**, 300–307 (1977).
9. Lang, A. S. & Beatty, J. T. Importance of widespread gene transfer agent genes in alpha-proteobacteria. *Trends Microbiol* **15**, 54–62 (2007).
10. Kogay, R. *et al.* Machine-Learning Classification Suggests That Many Alphaproteobacterial Prophages May Instead Be Gene Transfer Agents. *Genome Biol Evol* **11**, 2941–2953 (2019).
11. Paul, J. H. Prophages in marine bacteria: dangerous molecular time bombs or the key to survival in the seas? *ISME J* **2**, 579–589 (2008).
12. Rohwer, F. & Thurber, R. V. Viruses manipulate the marine environment. *Nature* **459**, 207–212 (2009).
13. Kristensen, D. M., Mushegian, A. R., Dolja, V. V. & Koonin, E. V. New dimensions of the virus world discovered through metagenomics. *Trends Microbiol* **18**, 11–19 (2010).
14. Suttle, C. A. Marine viruses--major players in the global ecosystem. *Nat Rev Microbiol* **5**, 801–812 (2007).
15. Esterman, E. S., Wolf, Y. I., Kogay, R., Koonin, E. V. & Zhaxybayeva, O. Evolution of DNA packaging in gene transfer agents. *Virus Evol* **7**, veab015 (2021).
16. Hynes, A. P. *et al.* Functional and Evolutionary Characterization of a Gene Transfer Agent's Multilocus "Genome". *Molecular Biology and Evolution* **33**, 2530–2543 (2016).
17. Shakya, M., Soucy, S. M. & Zhaxybayeva, O. Insights into origin and evolution of α-proteobacterial gene transfer agents. *Virus Evolution* **3**, (2017).
18. Sherlock, D., Leong, J. X. & Fogg, P. C. M. Identification of the First Gene Transfer Agent (GTA) Small Terminase in Rhodobacter capsulatus and Its Role in GTA Production and Packaging of DNA. *J Virol* **93**, e01328-19 (2019).
19. Tamarit, D., Neuvonen, M.-M., Engel, P., Guy, L. & Andersson, S. G. E. Origin and Evolution of the Bartonella Gene Transfer Agent. *Mol Biol Evol* **35**, 451–464 (2018).
20. Bárdy, P. *et al.* Structure and mechanism of DNA delivery of a gene transfer agent. *Nat Commun* **11**, 3034 (2020).
21. Gozzi, K., Tran, N. T., Modell, J. W., Le, T. B. K. & Laub, M. T. Prophage-like gene transfer agents promote Caulobacter crescentus survival and DNA repair during stationary phase. *PLOS Biology* **20**, e3001790 (2022).
22. Québatte, M. *et al.* Gene Transfer Agent Promotes Evolvability within the Fittest Subpopulation of a Bacterial Pathogen. *Cell Syst* **4**, 611-621.e6 (2017).

23. Farrera-Calderon, R. G. *et al.* The CckA-ChpT-CtrA Phosphorelay Controlling Rhodobacter capsulatus Gene Transfer Agent Production Is Bidirectional and Regulated by Cyclic di-GMP. *J Bacteriol* **203**, e00525-20 (2021).
24. Lang, A. S. & Beatty, J. T. Genetic analysis of a bacterial genetic exchange element: the gene transfer agent of Rhodobacter capsulatus. *Proc. Natl. Acad. Sci. U.S.A.* **97**, 859–864 (2000).
25. Mercer, R. G. *et al.* Loss of the response regulator CtrA causes pleiotropic effects on gene expression but does not affect growth phase regulation in Rhodobacter capsulatus. *J Bacteriol* **192**, 2701–2710 (2010).
26. Mercer, R. G. *et al.* Regulatory systems controlling motility and gene transfer agent production and release in Rhodobacter capsulatus. *FEMS Microbiol Lett* **331**, 53–62 (2012).
27. Westbye, A. B. *et al.* Phosphate concentration and the putative sensor kinase protein CckA modulate cell lysis and release of the Rhodobacter capsulatus gene transfer agent. *J Bacteriol* **195**, 5025–5040 (2013).
28. Westbye, A. B. *et al.* The Protease ClpXP and the PAS Domain Protein DivL Regulate CtrA and Gene Transfer Agent Production in Rhodobacter capsulatus. *Appl Environ Microbiol* **84**, e00275-18 (2018).
29. Ding, H., Grüll, M. P., Mulligan, M. E., Lang, A. S. & Beatty, J. T. Induction of Rhodobacter capsulatus Gene Transfer Agent Gene Expression Is a Bistable Stochastic Process Repressed by an Extracellular Calcium-Binding RTX Protein Homologue. *J Bacteriol* **201**, e00430-19 (2019).
30. Brimacombe, C. A., Ding, H. & Beatty, J. T. Rhodobacter capsulatus DprA is essential for RecA-mediated gene transfer agent (RcGTA) recipient capability regulated by quorum-sensing and the CtrA response regulator. *Mol. Microbiol.* **92**, 1260–1278 (2014).
31. Mercer, R. G. & Lang, A. S. Identification of a predicted partner-switching system that affects production of the gene transfer agent RcGTA and stationary phase viability in Rhodobacter capsulatus. *BMC Microbiol* **14**, 71 (2014).
32. Leung, M. M., Brimacombe, C. A., Spiegelman, G. B. & Beatty, J. T. The GtaR protein negatively regulates transcription of the gtaRI operon and modulates gene transfer agent (RcGTA) expression in Rhodobacter capsulatus. *Mol Microbiol* **83**, 759–774 (2012).
33. Koppenhöfer, S. *et al.* Integrated Transcriptional Regulatory Network of Quorum Sensing, Replication Control, and SOS Response in Dinoroseobacter shibae. *Front Microbiol* **10**, 803 (2019).
34. Fogg, P. C. M. Identification and characterization of a direct activator of a gene transfer agent. *Nature Communications* **10**, 595 (2019).
35. Sherlock, D. & Fogg, P. C. M. The archetypal gene transfer agent RcGTA is regulated via direct interaction with the enigmatic RNA polymerase omega subunit. *Cell Rep* **40**, 111183 (2022).
36. Barrows, J. M. & Goley, E. D. Synchronized Swarmers and Sticky Stalks: Caulobacter crescentus as a Model for Bacterial Cell Biology. *J Bacteriol* **205**, e0038422 (2023).
37. Swinger, K. K. & Rice, P. A. IHF and HU: flexible architects of bent DNA. *Curr Opin Struct Biol* **14**, 28–35 (2004).
38. Browning, D. F., Grainger, D. C. & Busby, S. J. Effects of nucleoid-associated proteins on bacterial chromosome structure and gene expression. *Curr Opin Microbiol* **13**, 773–780 (2010).
39. Freundlich, M., Ramani, N., Mathew, E., Sirko, A. & Tsui, P. The role of integration host factor in gene expression in Escherichia coli. *Mol Microbiol* **6**, 2557–2563 (1992).
40. Goosen, N. & van de Putte, P. The regulation of transcription initiation by integration host factor. *Mol Microbiol* **16**, 1–7 (1995).
41. Gober, J. W. & Shapiro, L. Integration host factor is required for the activation of developmentally regulated genes in Caulobacter. *Genes Dev* **4**, 1494–1504 (1990).
42. Muir, R. E. & Gober, J. W. Role of integration host factor in the transcriptional activation of flagellar gene expression in Caulobacter crescentus. *J Bacteriol* **187**, 949–960 (2005).
43. Browning, D. F. & Busby, S. J. W. Local and global regulation of transcription initiation in bacteria. *Nat Rev Microbiol* **14**, 638–650 (2016).

44. Chen, J., Boyaci, H. & Campbell, E. A. Diverse and unified mechanisms of transcription initiation in bacteria. *Nat Rev Microbiol* **19**, 95–109 (2021).
45. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
46. Mirdita, M. *et al.* ColabFold: making protein folding accessible to all. *Nat Methods* **19**, 679–682 (2022).
47. McLean, T. C. & Le, T. B. CTP switches in ParABS-mediated bacterial chromosome segregation and beyond. *Current Opinion in Microbiology* **73**, 102289 (2023).
48. van Kempen, M. *et al.* Fast and accurate protein structure search with Foldseek. *Nat Biotechnol* (2023) doi:10.1038/s41587-023-01773-0.
49. Malakooti, J., Wang, S. P. & Ely, B. A consensus promoter sequence for Caulobacter crescentus genes involved in biosynthetic and housekeeping functions. *J Bacteriol* **177**, 4372–4376 (1995).
50. Toporowski, M. C., Nomellini, J. F., Awram, P., Levi, A. & Smit, J. Transcriptional regulation of the S-layer protein type I secretion system in Caulobacter crescentus. *FEMS Microbiol Lett* **251**, 29–36 (2005).
51. Brosius, J., Dull, T. J., Sleeter, D. D. & Noller, H. F. Gene organization and primary structure of a ribosomal RNA operon from Escherichia coli. *J Mol Biol* **148**, 107–127 (1981).
52. Westbye, A. B., O'Neill, Z., Schellenberg-Beaver, T. & Beatty, J. T. The Rhodobacter capsulatus gene transfer agent is induced by nutrient depletion and the RNAP omega subunit. *Microbiology (Reading, Engl.)* **163**, 1355–1363 (2017).
53. Boutte, C. C. & Crosson, S. The complex logic of stringent response regulation in Caulobacter crescentus: starvation signalling in an oligotrophic environment. *Mol Microbiol* **80**, 695–714 (2011).
54. Lesley, J. A. & Shapiro, L. SpoT regulates DnaA stability and initiation of DNA replication in carbon-starved Caulobacter crescentus. *J Bacteriol* **190**, 6867–6880 (2008).
55. Peña, J. M. *et al.* Control of a programmed cell death pathway in Pseudomonas aeruginosa by an antiterminator. *Nat Commun* **12**, 1702 (2021).
56. Wen, A., Zhao, M., Jin, S., Lu, Y.-Q. & Feng, Y. Structural basis of AlpA-dependent transcription antitermination. *Nucleic Acids Res* **50**, 8321–8330 (2022).
57. Burmann, B. M. *et al.* An α helix to β barrel domain switch transforms the transcription factor RfaH into a translation factor. *Cell* **150**, 291–303 (2012).
58. Deighan, P. & Hochschild, A. The bacteriophage lambdaQ anti-terminator protein regulates late gene expression as a stable component of the transcription elongation complex. *Mol Microbiol* **63**, 911–920 (2007).
59. Roberts, J. W. *et al.* Antitermination by bacteriophage lambda Q protein. *Cold Spring Harb Symp Quant Biol* **63**, 319–325 (1998).
60. Yin, Z., Bird, J. G., Kaelber, J. T., Nickels, B. E. & Ebright, R. H. In transcription antitermination by Qλ, NusA induces refolding of Qλ to form a nozzle that extends the RNA polymerase RNA-exit channel. *Proc Natl Acad Sci U S A* **119**, e2205278119 (2022).
61. Krupp, F. *et al.* Structural Basis for the Action of an All-Purpose Transcription Anti-termination Factor. *Mol Cell* **74**, 143-157.e5 (2019).
62. Goodson, J. R. & Winkler, W. C. Processive Antitermination. *Microbiol Spectr* **6**, (2018).
63. Nickels, B. E., Roberts, C. W., Sun, H., Roberts, J. W. & Hochschild, A. The sigma(70) subunit of RNA polymerase is contacted by the (lambda)Q antiterminator during early elongation. *Mol Cell* **10**, 611–622 (2002).
64. Shi, J. *et al.* Structural basis of Q-dependent transcription antitermination. *Nat Commun* **10**, 2925 (2019).
65. Tran, N. T. *et al.* Permissive zones for the centromere-binding protein ParB on the Caulobacter crescentus chromosome. *Nucleic Acids Res* **46**, 1196–1209 (2018).
66. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**, R25 (2009).

67. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
68. Zhang, Y. *et al.* Model-based Analysis of ChIP-Seq (MACS). *Genome Biology* **9**, R137 (2008).
69. Dove, S. L. & Hochschild, A. A bacterial two-hybrid system based on transcription activation. *Methods Mol Biol* **261**, 231–246 (2004).
70. O'Reilly, F. J. *et al.* Protein complexes in cells by AI-assisted structural proteomics. *Mol Syst Biol* **19**, e11544 (2023).

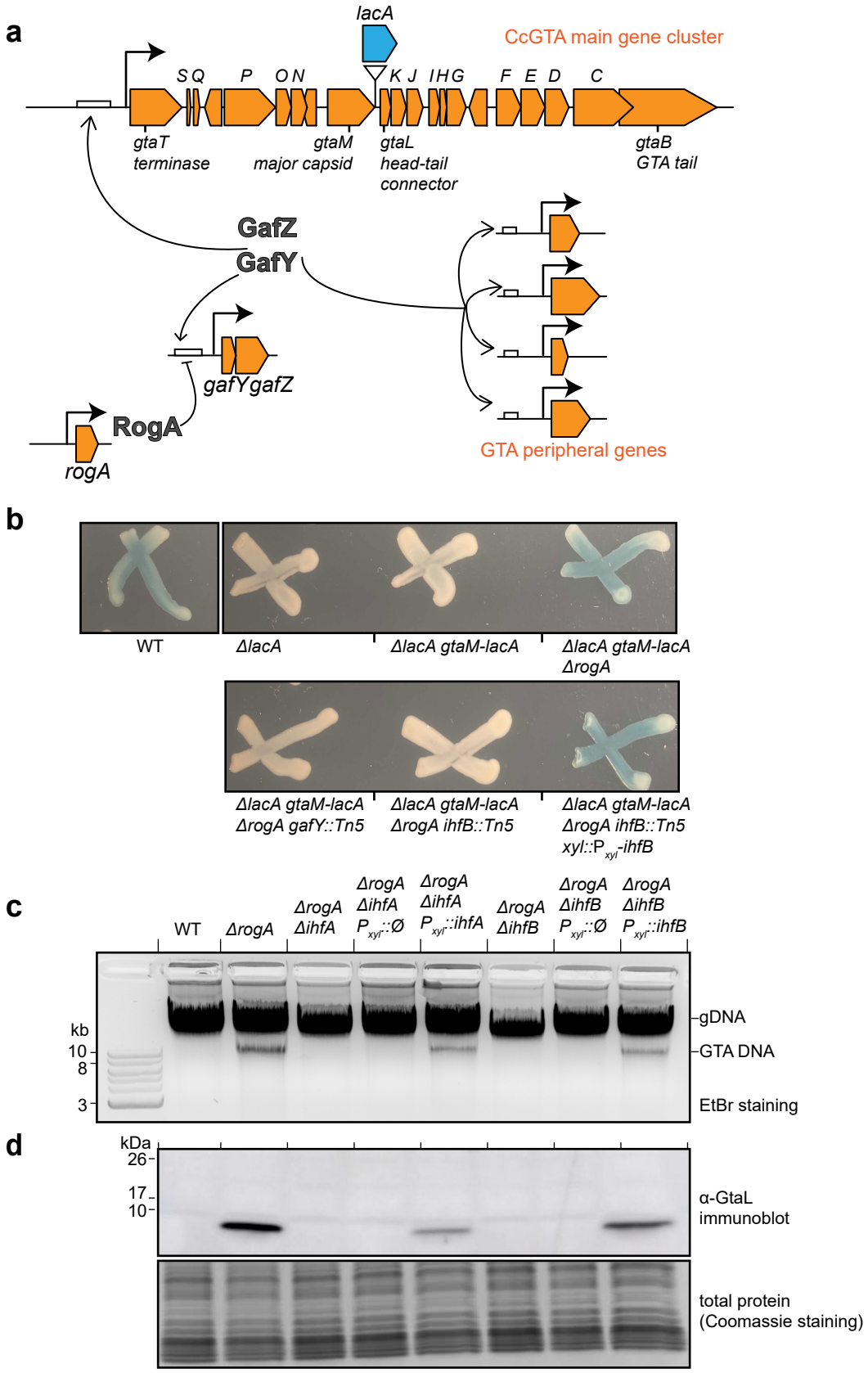


Fig. 1 The integration host factor (IHF) is required for GTA production. a Diagram of CcGTA regulation. RogA directly represses (flat-headed arrow) the expression of *gafYZ*. GafY and GafZ directly activate (arrows) the main GTA gene cluster (orange) and GTA peripheral genes (orange) on the *C. crescentus* chromosome. For the Tn5-based screen for additional GTA regulators, a dehydrogenase-encoding *lacA* gene (blue block arrow) was transcriptionally fused downstream of the GTA major capsid gene *gtaM*. For the full list of main CcGTA cluster genes and their annotated functions, see Table S1. **b** Phenotypes of indicated strains on agar media containing X-gal and 0.03% xylose. **c** Total DNA extraction from indicated strains grown up to a stationary phase. Total DNA was purified and separated by electrophoresis on a 1% agarose gel which was stained with ethidium bromide (EtBr) for DNA. **d** Immunoblot of total cell lysates of indicated strains using a polyclonal α-GtaL (GTA head-tail connector protein) antibody. A separate Coomassie-stained SDS-PAGE was loaded with the same volume of samples to serve as a loading control. Immunoblots were performed at least twice, and a representative image is shown.

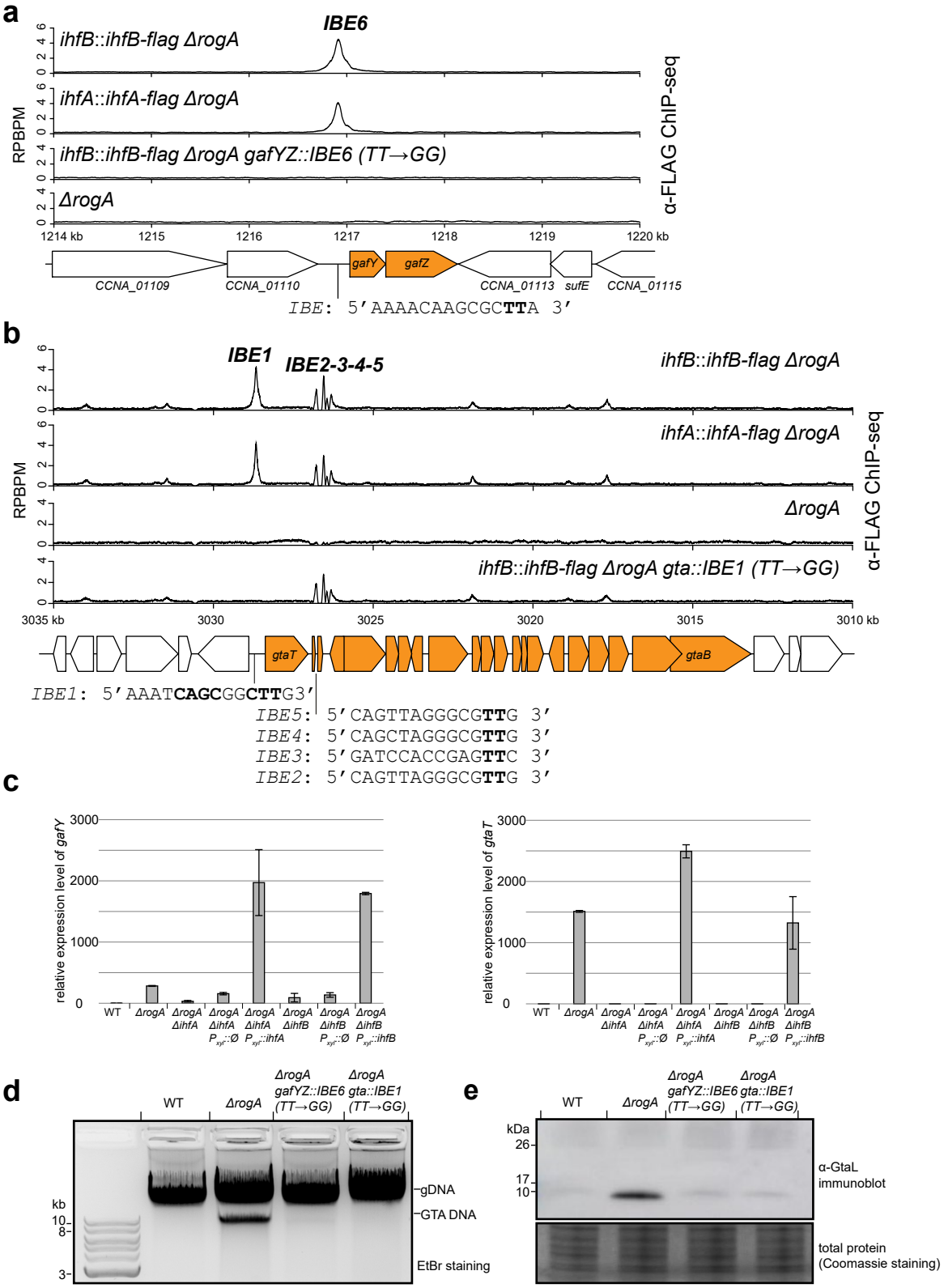


Fig. 2. IHF binds the promoter regions of *gafYZ* and the GTA gene cluster to activate gene expression. a-b α-FLAG ChIP-seq profiles show the enrichment of FLAG-tagged IHFα and β in different genetic backgrounds. Profiles were plotted with the x-axis representing genomic positions and the y-axis representing the number of reads per base pair per million mapped reads (RPBPM). The positions and sequences of identified IHF-binding elements (IBE) are shown beneath the schematic diagram of genes (GTA-related genes are depicted as orange block arrows). The conserved TT nucleotides in IBEs were highlighted in bold, and were mutated to GG to eliminate IHF binding. Note that the GTA main gene cluster is on the minus strand, but was inverted to run from left to right for a presentation purpose only. ChIP-seq experiments were performed twice using biological replicates, and a representative profile is shown. **c** Relative gene expression of *gafY* (left panel) and *gtaT* (right panel) from indicated strains as quantified by qRT-PCR (n=2, error bars indicate standard deviation). **d** Total DNA extraction from indicated strains grown up to a stationary phase. **e** Immunoblot of total cell lysates of indicated strains using a polyclonal α-GtaL (GTA head-tail connector protein) antibody. A separate Coomassie-stained SDS-PAGE was loaded with the same volume of samples to serve as a loading control. Immunoblots were performed at least twice, and a representative image is shown.

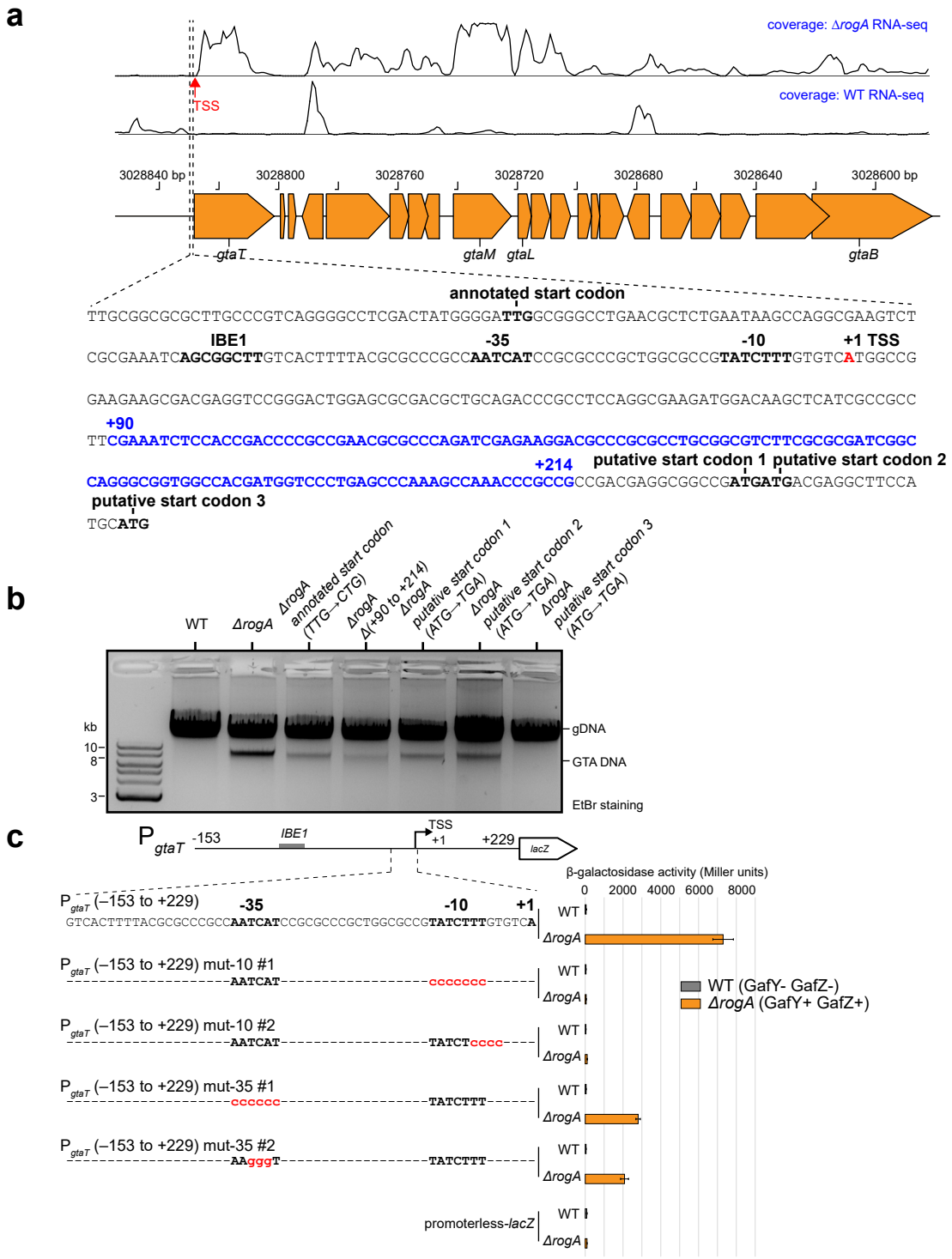


Fig. 3. Defining the transcription start site, core promoter elements, and the start codon for *gtaT*, the first gene in the GTA main cluster. a Plots of RNA-seq data for stationary-phase GTA-on Δ*rogA* cells vs. GTA-off WT cells. A single prominent TSS (red arrow) was found for the GTA main cluster in the GTA-on cells. A sequence covering the upstream region of *gtaT* is shown beneath the schematic diagram of genes, and features such as the annotated/putative start codons, IHF-binding element (IBE), core promoter elements (-10 -35), and the transcription start site (TSS, red) are indicated. The sequence from +90 to +214 (blue) was deleted for the experiment reported in panel b. **b** Total DNA extraction from indicated strains grown up to a stationary phase. **c** β-galactosidase activity (in Miller units) of the indicated promoter-*lacZ* reporter constructs in two *C. crescentus* genetic backgrounds (WT (grey) vs Δ*rogA* (orange)). The upstream region of *gtaT*, from -153 to +229 relative to the +1 TSS, was fused to *lacZ*. The -10 and -35 elements were mutated as indicated (red) on the schematic diagram. Cells containing an empty (promoterless) *lacZ* reporter plasmid served as a negative control. Values and error bars indicate mean ± standard deviations from four replicates.

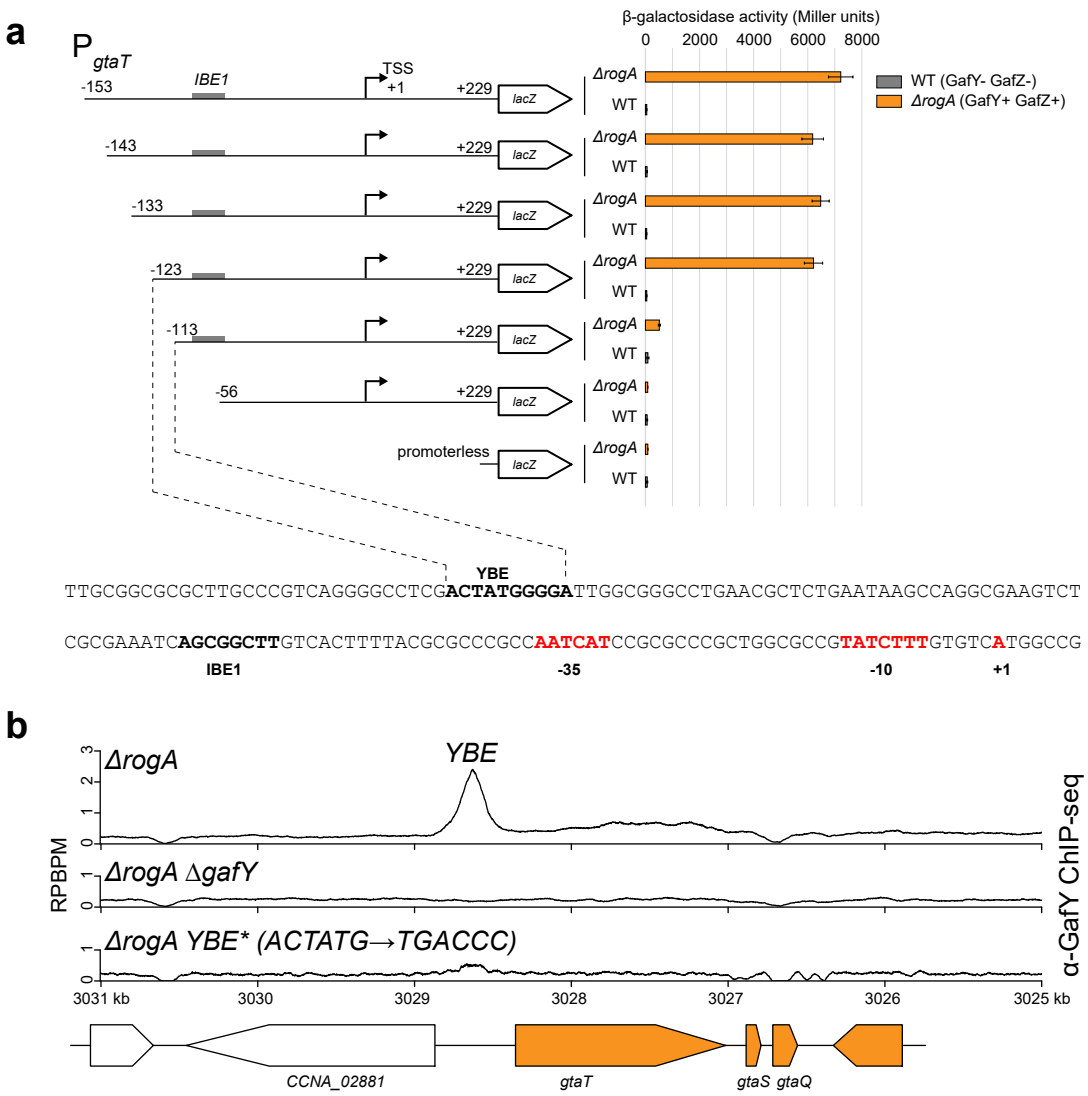


Fig. 4. A putative GafY-binding element (YBE) is located upstream of the IHF-binding element (IBE) on the promoter region of the main GTA cluster. a β-galactosidase activity (in Miller units) of progressively truncated promoter-*lacZ* reporter constructs in two *C. crescentus* genetic backgrounds (WT (grey) vs Δ*rogA* (orange)). Values and error bars indicate mean ± standard deviations from four replicates. The sequence of a putative GafY-binding element (YBE) is shown together with the IHF-binding element (IBE1), the core promoter elements (-10 -35) (red), and the TSS. **b** α-GafY ChIP-seq profiles show the enrichment of GafY at the upstream region of the GTA main cluster in different genetic backgrounds. Profiles were plotted with the x-axis representing genomic positions and the y-axis representing the number of reads per base pair per million mapped reads (RPBPM). ChIP-seq experiments were performed twice using biological replicates, and a representative profile is shown.

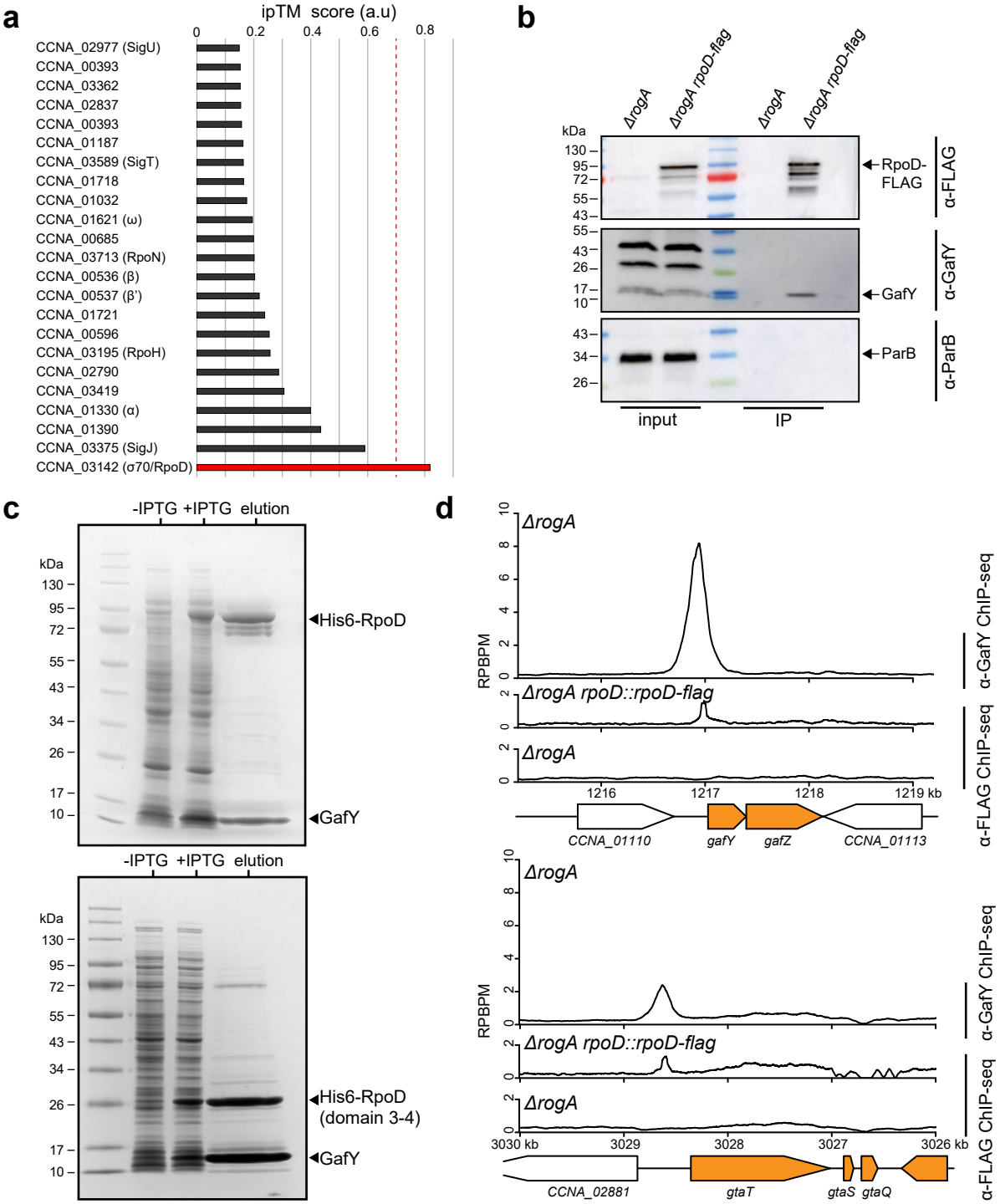


Fig. 5. GafY interacts directly with the sigma factor 70 RpoD. a An AlphaFold2-Multimer-based protein interaction screen between GafY and 23 core components of *C. crescentus* RNAP and sigma factors suggests a GafY-RpoD interaction. The likelihood of protein-protein interaction was ranked based on the confidence metrics (ipTM), with ipTM greater than 0.7 indicating likely protein-protein interaction. **b** Immunoblots analysis of co-immunoprecipitation of FLAG-tagged RpoD. ParB served as a non-associated protein control. The positions of bands corresponding to FLAG-tagged RpoD, GafY, and ParB are indicated with arrows. Co-IP experiments were performed at least twice, and a representative image is shown. **c** Co-expression and purification of a soluble RpoD-GafY complex (top panel) or RpoD (domain3-4 only)-GafY (bottom panel). RpoD (full-length or truncated) and GafY were co-overexpressed in *E. coli*, with only RpoD or RpoD (domain 3-4) carrying an N-terminal His tag. Following co-overexpression, the soluble extract was passed over a nickel column, and, after washing, bound proteins were eluted and analyzed on SDS-PAGE. **d** α-GafY and α-FLAG ChIP-seq profiles show the enrichment of GafY and FLAG-tagged RpoC (RNAP β' subunit), respectively, at the upstream region of *gafYZ* (top panel) and the GTA main cluster (bottom panel) in different genetic backgrounds. Profiles were plotted with the x-axis representing genomic positions and the y-axis representing the number of reads per base pair per million mapped reads (RPBPM). ChIP-seq experiments were performed twice using biological replicates, and a representative profile is shown.

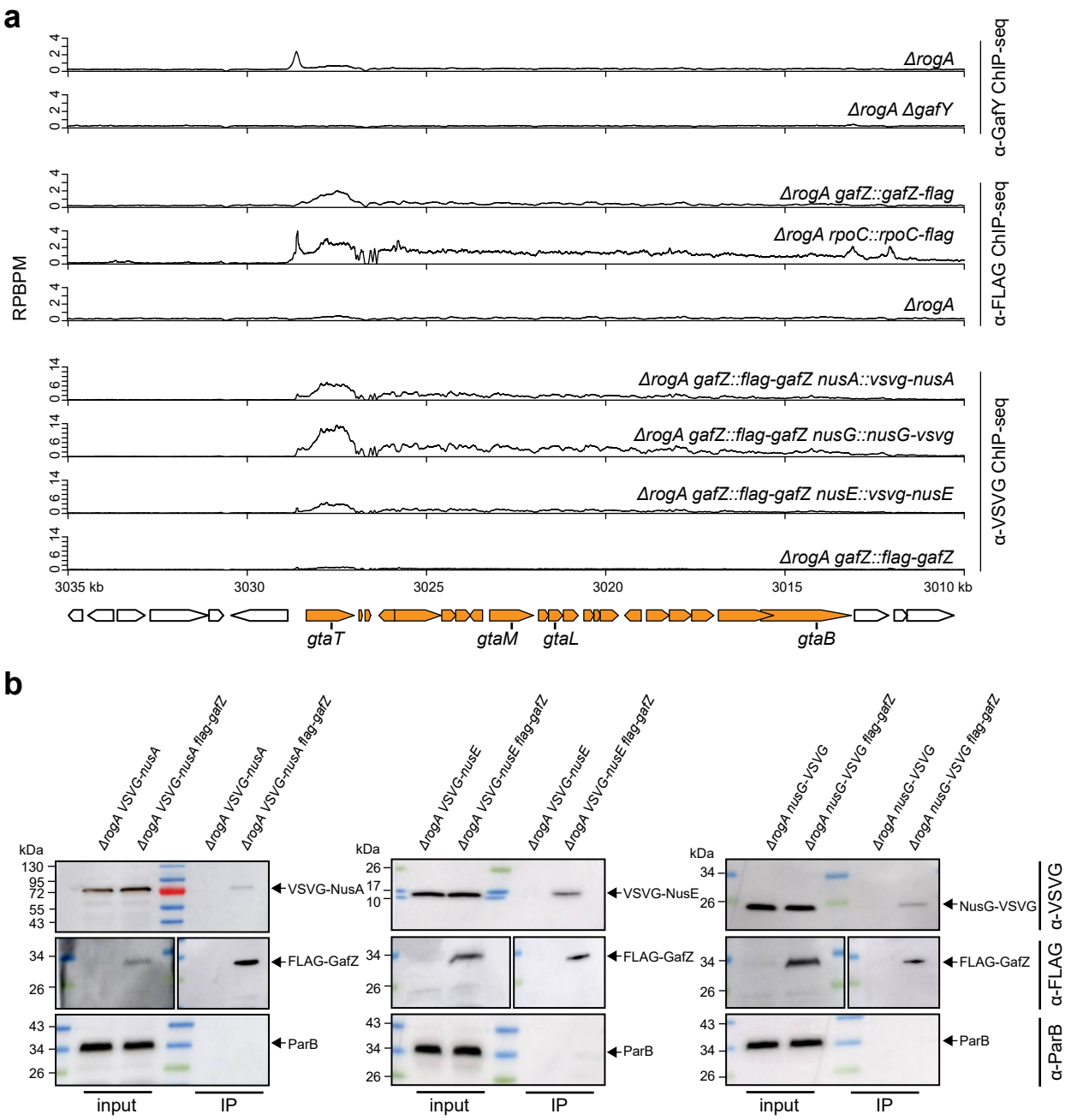


Fig. 6. GafY and GafZ might form a transcription elongation complex with RNAP, NusA, NusG, and NusE. a α-GafY, α-FLAG, and α-VSVG ChIP-seq profiles show the enrichment of GafY, FLAG-tagged GafZ, and VSVG-tagged NusAGE, respectively, at the GTA main cluster in different genetic backgrounds. Profiles were plotted with the x-axis representing genomic positions and the y-axis representing the number of reads per base pair per million mapped reads (RPBPM). ChIP-seq experiments were performed twice using biological replicates, and a representative profile is shown. **b** Immunoblots analysis of co-immunoprecipitation of FLAG-tagged GafZ. ParB served as a non-associated protein control. The positions of bands corresponding to VSVG-tagged NusA/E/G, FLAG-tagged GafZ, and ParB are indicated with arrows. The immunoblot for α-FLAG of the input fraction was exposed for a longer time than that from the IP fraction as FLAG-tagged GafZ was highly concentrated in the IP fraction compared to the input fraction. Co-IP experiments were performed at least twice, and a representative image is shown.

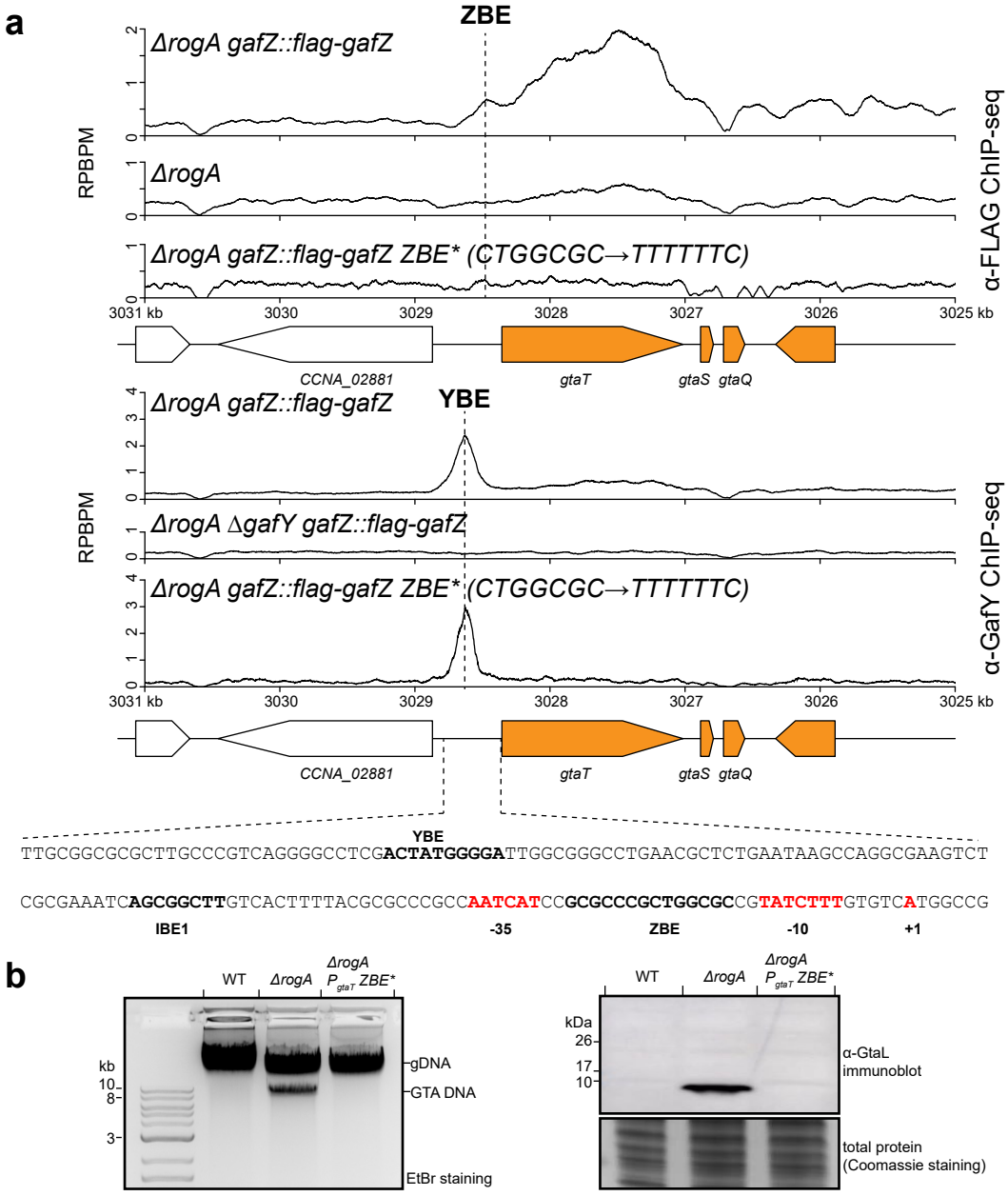


Fig.7. The putative GafZ binding element (ZBE) is located in between the -10 and -35 promoter elements of the GTA gene cluster. a α-FLAG and α-GafY ChIP-seq profiles show the enrichment of FLAG-tagged GafZ and GafY, respectively, at the main GTA cluster in different genetic backgrounds. Profiles were plotted with the x-axis representing genomic positions and the y-axis representing the number of reads per base pair per million mapped reads (RPBPM). The sequence of a putative GafZ-binding element (ZBE) is shown together with the IHF-binding element (IBE1), the GafY-binding element (YBE), the core promoter elements (-10 -35), and the TSS. ChIP-seq experiments were performed twice using biological replicates, and a representative profile is shown. **b** (left panel) Total DNA extraction from indicated strains grown up to a stationary phase. Total DNA was purified and separated by electrophoresis on a 1% agarose gel which was stained with ethidium bromide (EtBr) for DNA. (right panel) Immunoblot of total cell lysates of indicated strains using a polyclonal α-GtaL (GTA head-tail connector protein) antibody. A separate Coomassie-stained SDS-PAGE was loaded with the same volume of samples to serve as a loading control. Immunoblots were performed at least twice, and a representative image is shown.

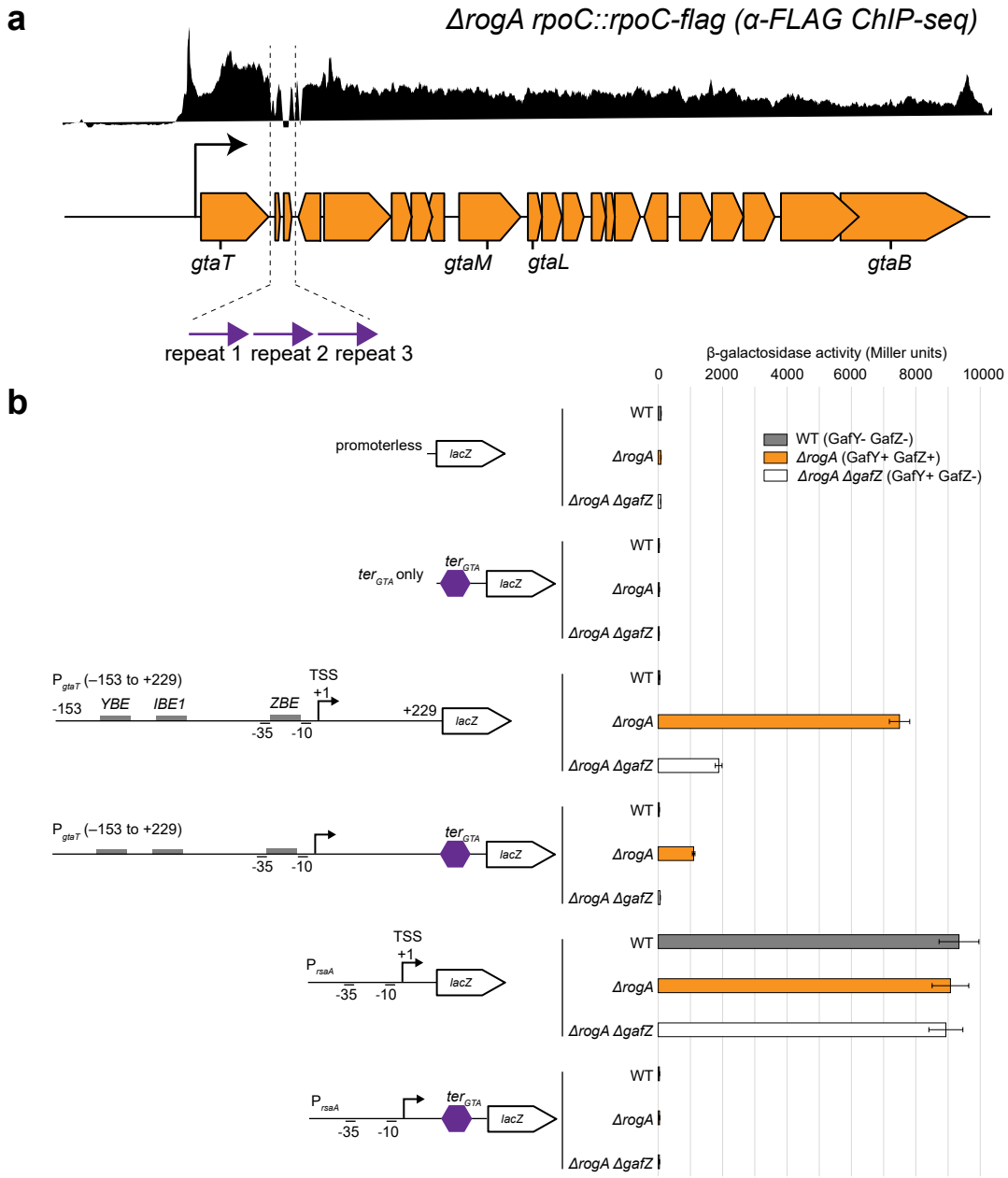


Fig. 8. GafYZ allows RNAP to bypass a transcription terminator located downstream of the first gene in the GTA gene cluster. a Identification of three 150-bp direct repeats (purple arrows) downstream of the *gtaT* gene in the GTA main cluster (see Supplementary fig. 5a for the sequences and alignment of the three repeats). **b** β-galactosidase activity (in Miller units) of indicated promoter-*lacZ* reporter constructs in three *C. crescentus* genetic backgrounds (WT (grey) vs Δ*rogA* (orange) vs. Δ*rogA*Δ*gafZ* (white)). The upstream region of *gtaT*, from -153 to +229 relative to the +1 TSS, was fused to *lacZ*, with or without the three direct repeats (ter_{GTA}, purple hexagon). Cells containing an empty (promoterless) *lacZ* reporter plasmid or plasmid with only ter_{GTA} fused to *lacZ* served as negative controls. Values and error bars indicate mean ± standard deviations from four replicates.

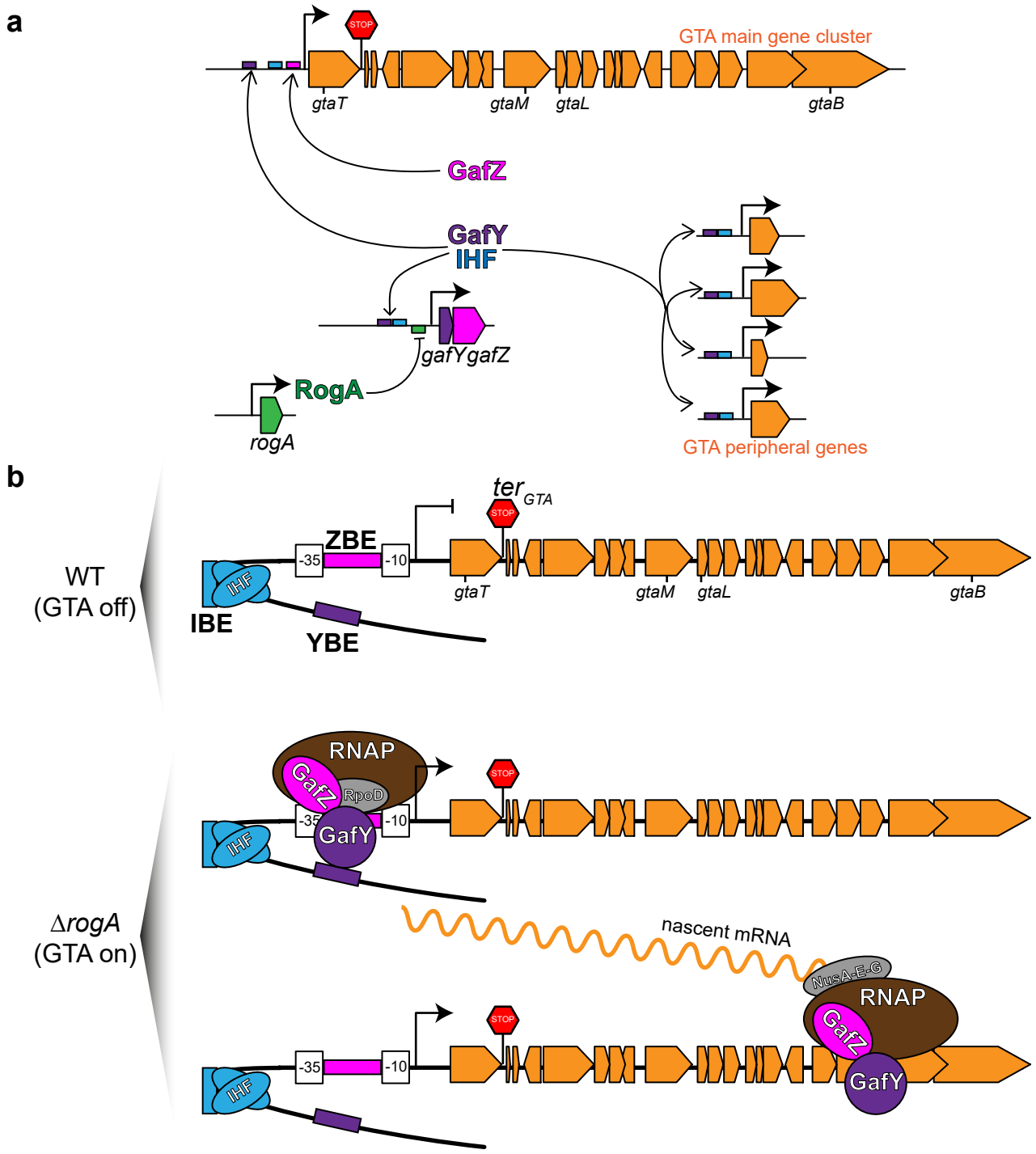
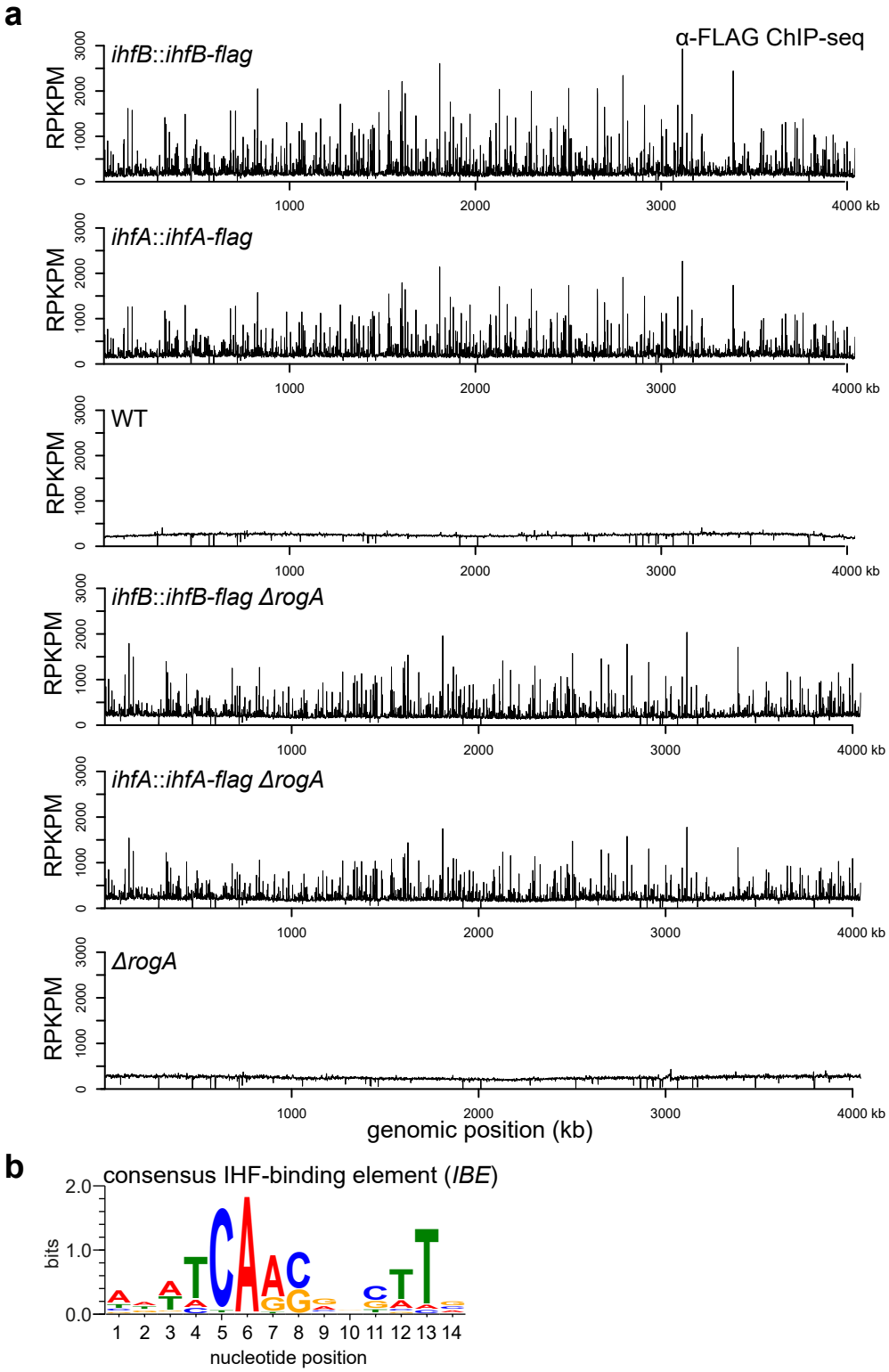
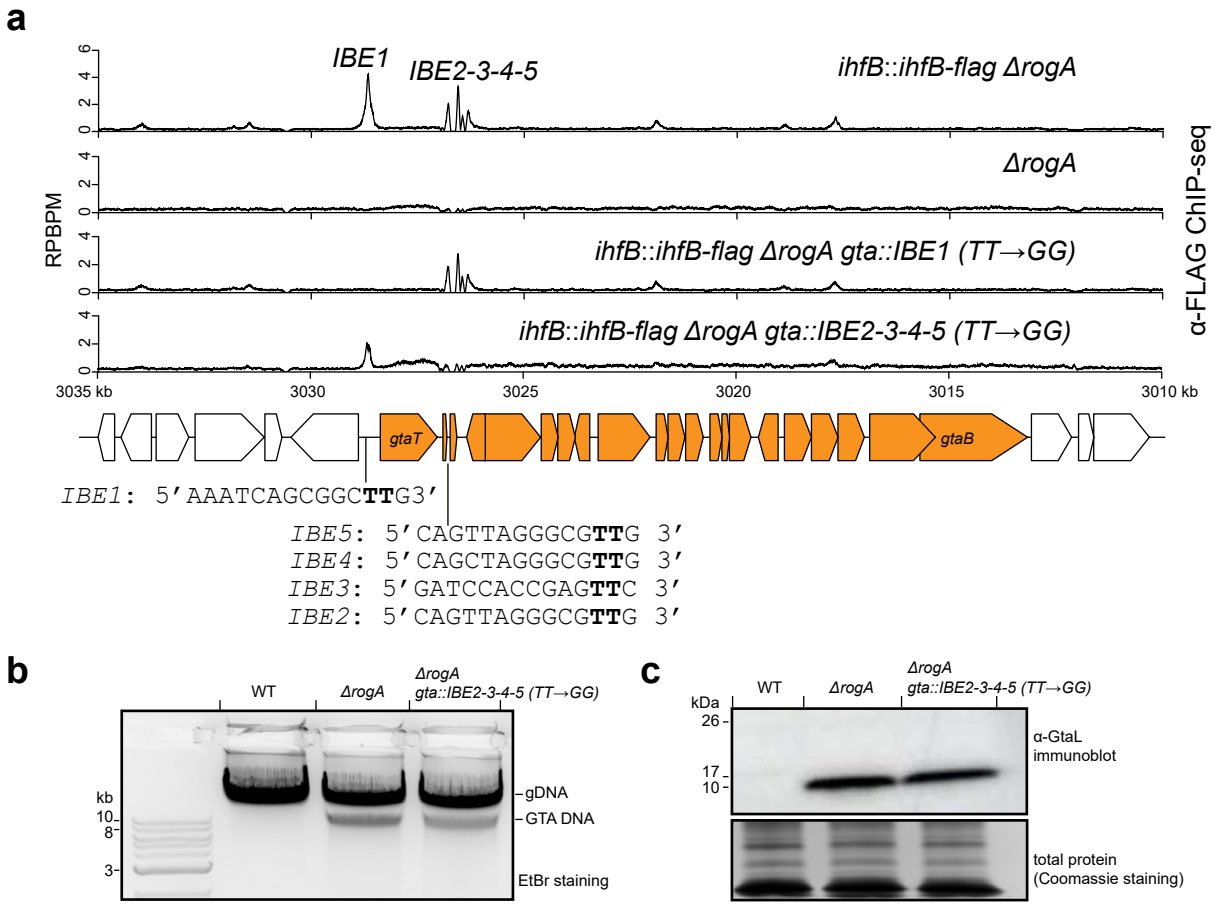


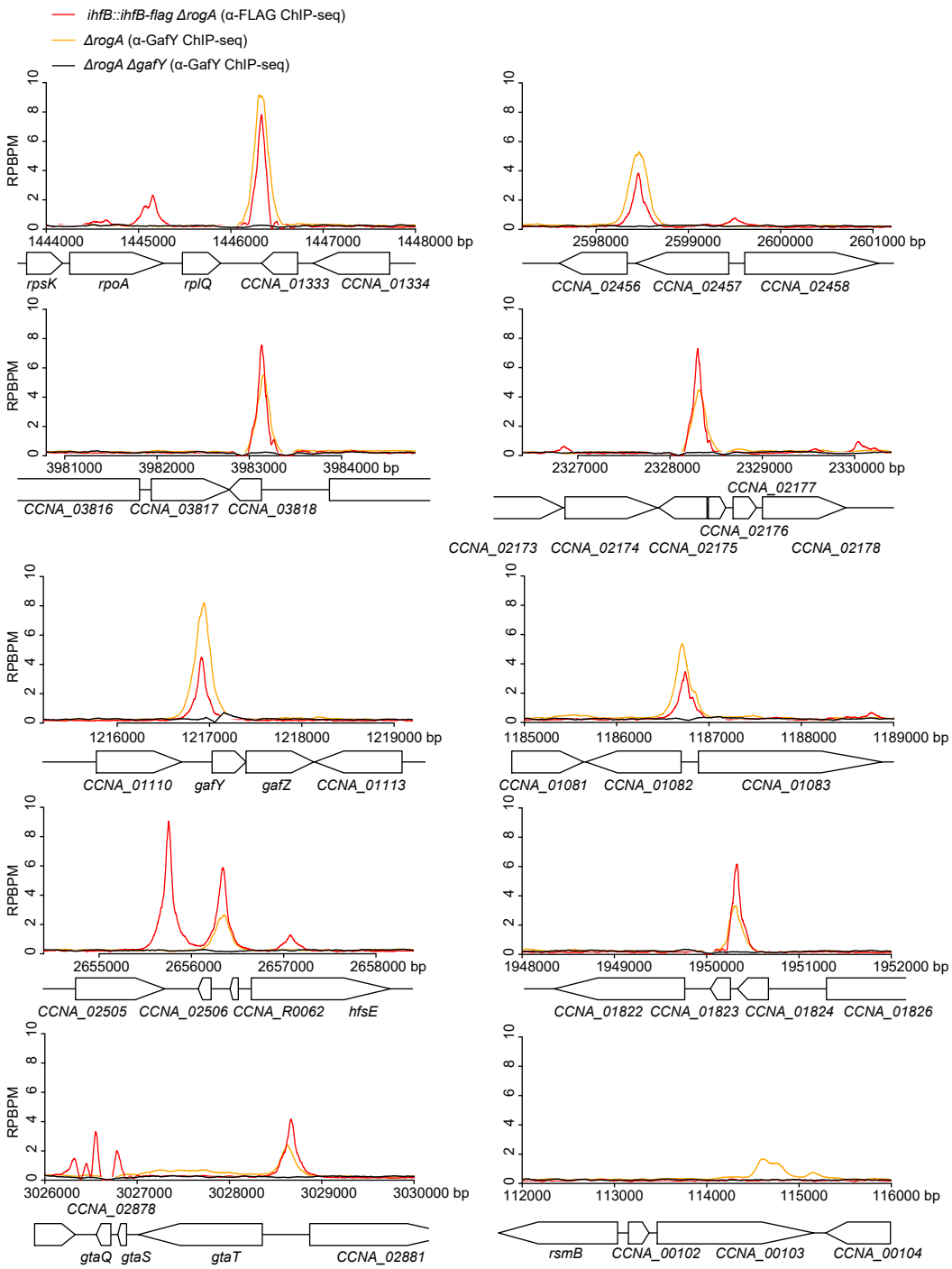
Fig. 9. Proposed model of how GafYZ and IHF function to activate the GTA gene cluster. a Diagram of CcGTA regulation. RogA (dark green) directly represses (flat-headed arrow) the expression of *gafYZ*. GafY (purple) and IHF (blue) co-activate (arrows) the *gafYZ* operon, the GTA main gene cluster (orange) as well as GTA peripheral genes (orange) on the *C. crescentus* chromosome. GafZ (magenta) together with GafY and IHF are required to activate the transcription of the GTA main cluster. **b** In GTA-off WT cells, RogA represses the expression of *gafYZ*. No GafY and GafZ were produced to activate the transcription of the GTA main gene cluster, even though IHF (blue) is prebound at the upstream region of the GTA main cluster at its IHF-binding element (IBE). In GTA-on Δ*rogA* cells, the promoter of *gafYZ* is derepressed, leading to the production of GafY (purple) and GafZ (magenta). GafY binding at the putative GafY-binding element (YBE), likely being assisted by the DNA-bending action of IHF (blue), binds RpoD (grey) to recruit RNAP (brown) to the core -10 -35 promoter elements. GafZ (magenta) binds the promoter at the putative GafZ-binding element (ZBE) positioned between the -10 and -35 elements of the promoter, then binds RNAP to allow a holo-complex of RNAP-GafZ-NusAEG to bypass the ter_{GTA} (red stop sign) positioned downstream of *gtaT* to synthesize nascent mRNA (wavy orange line). ChIP-seq profiles of GafY and biochemical data that show GafY-GafZ co-interaction suggest that GafY might travel together with an elongating RNAP-GafZ-NusAEG holo-enzyme complex.



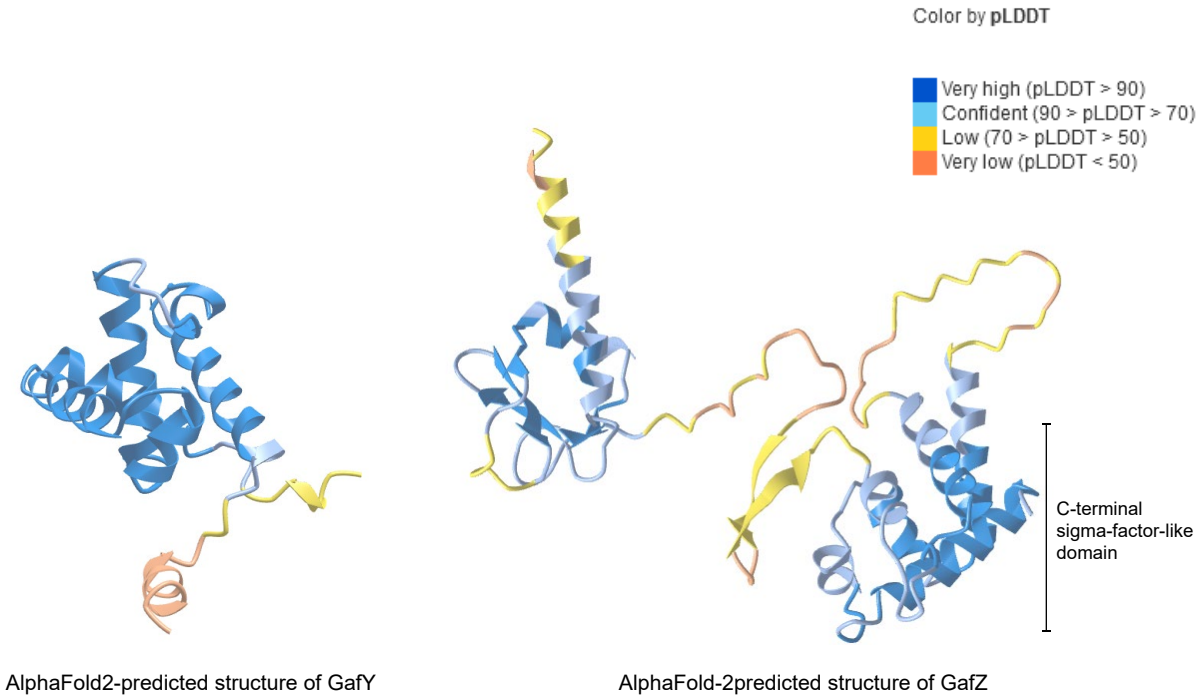
Supplementary fig. 1. IHF binding on *C. crescentus* chromosome is independent of RogA. a α-FLAG ChIP-seq profiles show the enrichment of FLAG-tagged IHFα and FLAG-tagged IHFβ on the chromosomes of either GTA-off WT or GTA-on Δ*rogA C. crescentus* cells in stationary phase. Profiles were plotted with the x-axis representing genomic positions and the y-axis representing the number of reads per kilobase pair per million mapped reads (RPKPM). ChIP-seq experiments were performed twice using biological replicates, and a representative profile is shown. **b** The consensus IHF-binding element (IHF), constructed from the hundred most enriched sites in α-FLAG-IHF ChIP-seq datasets.



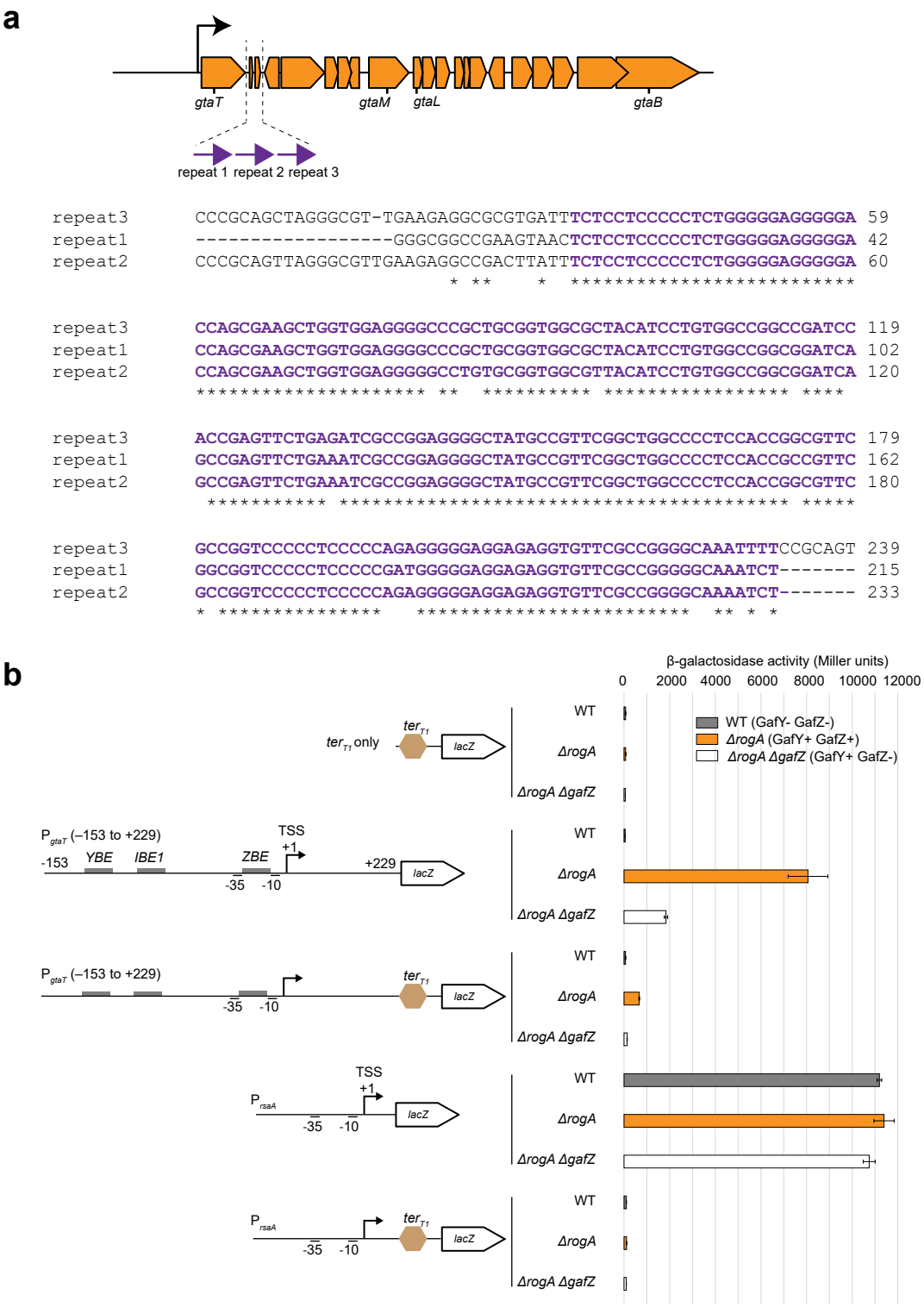
Supplementary fig. 2. Four IHF binding elements (IBE 2-3-4-5) play no or little role in the transcriptional activation of the GTA main gene cluster. a α-FLAG ChIP-seq profiles show the enrichment of FLAG-tagged IHFα and β in different genetic backgrounds. Profiles were plotted with the x-axis representing genomic positions and the y-axis representing the number of reads per base pair per million mapped reads (RPBPM). The positions and sequences of identified IHF-binding elements (IBE) are shown beneath the schematic diagram of the GTA gene cluster. Note that the main GTA cluster is on the minus strand, but was inverted to run from left to right for a presentation purpose only. ChIP-seq experiments were performed twice using biological replicates, and a representative profile is shown. **b** Total DNA extraction from indicated strains grown up to a stationary phase. **c** Immunoblot of total cell lysates of indicated strains using a polyclonal α-GtaL (GTA head-tail connector protein) antibody. A separate Coomassie-stained SDS-PAGE was loaded with the same volume of samples to serve as a loading control. Immunoblots were performed at least twice, a representative image is shown.



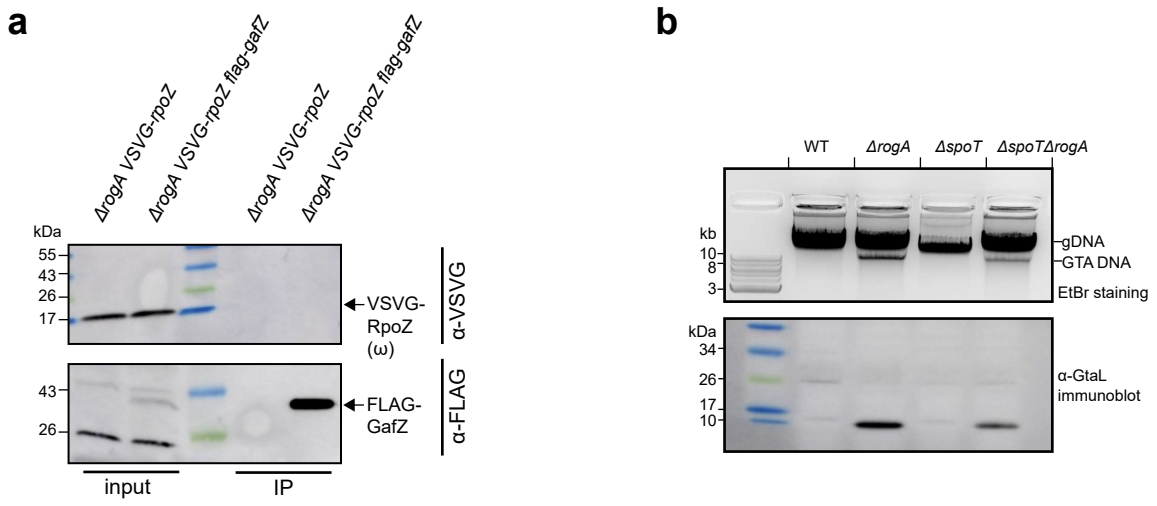
Supplementary fig. 3. IHF binds to GafY-regulated promoters. α-FLAG and α-GafY ChIP-seq profiles show the enrichment of FLAG-tagged IHF (red), GafY (orange), and a negative control (black), respectively, at the ten most GafY-enriched regions. Profiles were plotted with the x-axis representing genomic positions and the y-axis representing the number of reads per base pair per million mapped reads (RPBPM). ChIP-seq experiments were performed twice using biological replicates, and a representative profile is shown.



Supplementary fig. 4. AlphaFold-2 predicted structures of GafY (left panel) and GafZ (right panel). Predicted structures were colored according to their predicted local distance difference test scores (pLDDT).



Supplementary fig. 5. GafYZ allows RNAP to bypass a putative transcription terminator (*ter*$_{GTA}$) or an *E. coli rrnB* T1 transcription terminator (*ter*$_{T1}$) located downstream of *gtaT*. a Sequence alignment of three direct repeats (purple arrows) that constitute the putative terminator of the GTA cluster (*ter*$_{GTA}$). **b** β-galactosidase activity (in Miller units) of indicated promoter-*lacZ* reporter constructs in three *C. crescentus* genetic backgrounds (WT (grey) vs Δ*rogA* (orange) vs. Δ*rogA*Δ*gafZ* (white)). The upstream region of *gtaT*, from -153 to +229 relative to the +1 TSS, was fused to *lacZ*, with or without the *E. coli rrnB* T1 transcription terminator (*ter*$_{T1}$, light brown hexagon). Cells containing an empty (promoterless) *lacZ* reporter plasmid or plasmid with only *ter*$_{T1}$ fused to *lacZ* served as negative controls. Values and error bars indicate mean ± standard deviations from four replicates.



Supplementary fig. 6. *C. crescentus* GafZ did not co-immunoprecipitate with *C. crescentus* RpoZ (ω subunit of RNA polymerase). a Immunoblots analysis of co-immunoprecipitation of FLAG-tagged GafZ. The positions of bands corresponding to VSVG-tagged RpoZ and FLAG-tagged GafZ are indicated with arrows. **b** Deletion of *spoT* did not eliminate GTA synthesis in *C. crescentus*. (top panel) Total DNA extraction from indicated strains grown up to a stationary phase. Total DNA was purified and separated by electrophoresis on a 1% agarose gel which was stained with ethidium bromide (EtBr) for DNA. (bottom panel) Immunoblot of total cell lysates of indicated strains using a polyclonal α-GtaL (GTA head-tail connector protein) antibody. Immunoblots were performed at least twice, and a representative image is shown.

```
GafA                --MKT-----------------------------------MQDRESLPDWLPDHARLY  21
GafY                MLVQALEDIWAITPDPRRDRLTVGFVTHLVSLATGVPAVEIATPKRVSHTAVRARQLAIY  60
                      :::                                       .* *       :  :*
GafA                LRHVEEGVPIRQLARAEGCHASTILRRVRRIEQRRDDPLVDEALTRLGRFAAAASAAPPR  81
GafY                LTHITLHWPLARVAFAFGRDRTTCGHACRKIEDLREDEAFDRRLCELEAC---LRQAPSH  117
                    * *:    *: ::* * * . :*  :  *:**: *:*  .*. * .*         ** :
GafA                EDDPAMTAPIRPTAPQACPEAAEDPDSPDIATLSREGRRVLRRLAEPGALLIIAPDMEKA  141
GafY                AQ------------------------------------------------------------  119
                     :
GafA                VVLRGTVRTAVVAREVAQGFALNGWILVQHSGRVTSYELSATGRAALKRLLEAEALTAGR  201
GafY                ------------------------------------------------------------  119

GafA                DPATAADNPHADRHRDWGERTVNEGQGRVTRMRMNLAESPLGVLARRRDSDGRPFLSPDL  261
GafY                ------DLP-----------------------------------------------------  122
                          * *
GafA                VAAGERLREDFELAQMGPRVAQNWERFMTGGARGQYRPELGHGGPGGSDRARERVAAALC  321
GafY                ------------------------------------------------------------  122

GafA                DLGPGLGDMVLRCCCFLEGLETAEKRMGWSARSGKIVLRIALMRLKRHYDETYGGAAPLI  381
GafY                ------------------------------------------------------------  122

GafA                G     382
GafY                -     122
```

```
GafA                MKTMQDRESLPDWLPDHARLYLRHVEEGVPIRQLARAEGCHASTILRRVRRIEQRRDDPL  60
GafZ                ------------------------------------------------------------  0

GafA                VDEALTRLGRFAAAASAAPPREDDPAMTAPIRPTAPQACPEAAEDPDSPDIATLSREGRR  120
GafZ                --------------------------MTGGLQI-----HFEA--------ALDAERRAER  21
                                              **. ::        **            .*...*
GafA                VLRRLAEPGALLIIAPDMEKAVVLRGTVR------------TAVVAREV-AQGFALNGWI  167
GafZ                AMRLLARPGAVIEPHG--GGYGVRLGASRRRCVMLTLDEATFAVLAREATLKPRREGGWT  79
                    .:* **.***::          *  *: *             **:***.  :    .**
GafA                LVQHSGRVTSYELSATGRAALKRLLEAEALTAGRDPATAADNPHADRHRDWGERTVNEGQ  227
GafZ                MVARPE------------------------------VAAVPPP-GR-PGVIEASVETA-  105
                    :* :                                .**  *  .*  .  * :*: .
GafA                GRVTRMRMNLAESPLGVLARRRDSDGRPFLSPDLVAAGERLREDFELAQMGPRVAQNWER  287
GafZ                --DGVVRRNLGESPIAWLARRRDSHGRPWLAPAEIAAAERLREEFESLGTLGRMTMRWDA  163
                         :* **.***:. *******.***:*:*  :**.*****:**      *:: .*:
GafA                FMTGGA-RGQYRPELGHGGPGGSDRARERVAAALCDLGPGLGDMVLRCCCFLEGLETAEK  346
GafZ                TPRVDGGQTVLA------PAERDHATRQRIARALTAVGPGLRDILERVCLMGSALEAAED  217
                        .. :              .. :*:*:* **  :**** *:: * * : ..**:**.
GafA                RMGWSARSGKIVLRIALMRLKRHYDETYGGAAPLIG       382
GafZ                SLKLPRRAGKTVLKLALQALARHYRMA---------       244
                     :    *:** **::**  * ***  :
```

Supplementary fig. 7. Sequence alignment of *R. capsulatus* GafA to *C. crescentus* GafY and GafZ. The N-terminal domain, the central domain, and the C-terminal domain of GafA are colored in dark green, blue, and cyan, respectively. The sequence alignment shows poor amino acid similarity for the central region of *R. capsulatus* GafA and the N-terminal side of *C. crescentus* GafZ (shaded box).