Article:

Title

Full title: The archetypal gene transfer agent (RcGTA) is regulated via direct interaction with the enigmatic RNA polymerase omega subunit

Short Title: *Regulation of RcGTA production by GafA and Rpo-ω*

Authors

David Sherlock,[1] Paul C. M. Fogg[1,2]*

Affiliations

[1] Biology Department, University of York, York, United Kingdom. YO10 5DD

[2] York Biomedical Research Institute (YBRI), University of York, York, United Kingdom. YO10 5NG

* Corresponding author and lead contact: Dr Paul C.M. Fogg at paul.fogg@york.ac.uk

Summary

Gene transfer agents (GTAs) are small virus-like particles that indiscriminately package and transfer any DNA present in their host cell, with clear implications for bacterial evolution. The first transcriptional regulator that directly controls GTA expression, GafA, was recently discovered but its mechanism of action remained elusive. Here we demonstrate that GafA controls GTA gene expression by direct interaction with the RNA polymerase omega subunit (Rpo-ω) and also positively autoregulates its own expression by an Rpo-ω independent mechanism. We show that GafA is a modular protein with distinct DNA and protein binding domains. The functional domains we observe in Rhodobacter GafA also correspond to two-gene operons in Hyphomicrobiales pathogens. Together these data allow us to produce the most complete regulatory model for a GTA, and point towards an atypical mechanism for RNA polymerase recruitment and specific transcriptional activation in the alpha-proteobacteria.

INTRODUCTION

29 Horizontal gene transfer by viruses and other mobile genetic elements is the major driver
30 of rapid bacterial adaptation and spread of traits such as antibiotic resistance. Gene
31 Transfer Agents (GTAs) are virus-like genetic elements that are similar to viruses but
32 instead of prioritizing the spread of their own genes, they package and disseminate any
33 DNA within the host cell (Hynes *et al.*, 2016; Lang *et al.*, 2012; Shakya *et al.*, 2017;
34 Sherlock *et al.*, 2019; Tamarit *et al.*, 2018). Although GTAs usually package and transfer
35 'random' fragments of DNA from their host to compatible recipients in headful fragments
36 (Berglund *et al.*, 2009; Esterman *et al.*, 2021; Freese *et al.*, 2017; Hynes *et al.*, 2012;
37 Sherlock *et al.*, 2019), some species do exhibit bias towards certain regions of the
38 genome (Berglund *et al.*, 2009; Tomasch *et al.*, 2018). Significantly, GTAs have been
39 implicated in high frequency spread of genes between bacteria (McDaniel *et al.*, 2010)
40 and an extensive survey of the function of thousands of bacterial genes indicated that
41 GTA genes convey significant fitness benefits in multiple species under stress conditions
42 (Kogay *et al.*, 2019, 2020; Price *et al.*, 2018).

43 The true prevalence of GTAs is not currently known, however, a recent study identified
44 homologues of the model *Rhodobacter capsulatus* GTA (RcGTA) is present in at least
45 50% of sequenced alpha-proteobacteria genomes – many of which had been mis-
46 annotated as remnant prophages (Kogay *et al.*, 2019, 2020; Shakya *et al.*, 2017). The
47 GTA genes are often dispersed at multiple genomic locations (Hynes *et al.*, 2016; Motro
48 *et al.*, 2009), and co-ordinated expression initiates from a small subset of the bacterial
49 population (Fogg, 2019; Fogg *et al.*, 2012; Hynes *et al.*, 2012; Québatte and Dehio, 2019).
50 Timing and regulation of GTA production is tightly controlled by interlinked host regulatory
51 circuits including quorum sensing (Koppenhöfer *et al.*, 2019; Leung *et al.*, 2012), stringent
52 response (Québatte *et al.*, 2017; Westbye *et al.*, 2017), SOS response (Kuchinski *et al.*,
53 2016), c-di-GMP (Pallegar *et al.*, 2020b, 2020a) and the pleiotropic transcription factor
54 CtrA (Lang and Beatty, 2000; Westbye *et al.*, 2018). In *R. capsulatus*, these complex
55 pathways are integrated via a specific GTA transcriptional regulator, GafA (Fogg, 2019),
56 and an RTX-domain extracellular repressor, *rcc00280* (Ding *et al.*, 2019; Westbye *et al.*,
57 2018). However, the precise mechanism of action for these proteins is not fully known.

58 It has been suggested that *Bartonella* GTAs are produced by the fittest cells in a given
59 population in response to cytosolic ppGpp levels (Québatte *et al.*, 2017), and that RcGTA
60 production is also influenced by ppGpp via the RNA polymerase omega sub-unit (Rpo-ω)
61 (Westbye *et al.*, 2017). *R. capsulatus* Rpo-ω is not required for growth but is essential for
62 RcGTA production (Westbye *et al.*, 2017). In other species, Rpo-ω is thought to play
63 several roles including stabilization of the RNAP holoenzyme and modulation of
64 transcription profiles via recruitment of alternative sigma factors (Gunnelius *et al.*, 2014;
65 Paget, 2015; Ross *et al.*, 2013; Weiss *et al.*, 2017). One study showed that *E. coli* Rpo-
66 ω can facilitate transcriptional activation when covalently linked to DNA binding proteins

67 (Dove and Hochschild, 1998), but to our knowledge no native interaction between Rpo-ω
68 and a transcriptional regulator has ever been demonstrated. Here, we examine the
69 relationship between RNAP-ω and the RcGTA activator protein, GafA. We explore the
70 protein:protein and protein:DNA binding activities of GafA domains, identify putative *gafA*
71 genes in pathogenic Hyphomicrobiales species and speculate on the overall mechanism
72 of action for the GafA regulator.

RESULTS

Rpo-ω is required for activation of GTA production by GafA

GafA is the only known direct activator of Gene Transfer Agent expression in *R. capsulatus* (Fogg, 2019). The omega sub-unit of RNA polymerase (Rpo-ω), encoded by the *rpoZ* gene, is also required for RcGTA production (Westbye *et al.*, 2017) but the relationship between the two has not been established. Introduction of the plasmid pCMF180, containing *gafA* together with its native promoter, into wild-type *R. capsulatus* SB1003 leads to an increase in RcGTA production - presumably due to increased copy number (Fig. 1A & B); meanwhile deletion of *rpoZ* completely eliminates GTA production (Fig. 1A & B), both of which corroborate previous findings (Fogg, 2019; Westbye *et al.*, 2017). The pCMF180 plasmid was introduced into SB1003 Δ*rpoZ* to test whether moderate *gafA* overexpression can overcome the loss of RcGTA production phenotype but no GTA gene transfer was detected (Fig. 1B). Western blots of concentrated supernatants using an α-RcGTA capsid antibody also failed to detect any capsid protein accumulation in the supernatant of the SB1003 Δ*rpoZ* + *gafA* strains (Fig. 1A). As *gafA* was expressed from its own promoter, it is possible that Rpo-ω acts to regulate expression of *gafA* and consequently the GTA genes indirectly. To confirm that expression of *gafA* was not affected in any way by the loss of *rpoZ*, RNA was extracted for transcript quantification by qPCR. The transcription of *gafA* in SB1003 Δ*rpoZ* was equivalent to wild-type, and *gafA* transcript abundance was actually higher for SB1003 Δ*rpoZ* + pCMF180 compared to the *rpoZ* replete background (Fig. 1C). Finally, a construct was created containing *gafA* expressed from a non-native cumate inducible promoter – pCMF254. Overexpression of *gafA* in SB1003 led to an ~20-fold increase in GTA production, however, overexpression in SB1003 Δ*rpoZ* produced no detectable GTA production and was thus indistinguishable from the empty plasmid control (Fig. 1D). Taken together, these data indicate that Rpo-ω is not required for expression of *gafA* but instead regulates RcGTA production downstream.

Rpo-ω directly interacts with GafA

In multiple species Rpo-ω is thought to influence RNAP sigma factor preference and consequently global gene expression (Gunnelius *et al.*, 2014; Kurkela *et al.*, 2021; Mathew and Chatterji, 2006; Yamamoto *et al.*, 2018). We hypothesized that GafA acts by

104 binding to Rpo-ω to alter promoter preferences of the RNAP holoenzyme, and hence
105 deletion of *rpoZ* abolishes the influence of GafA. pUT18 bacterial-2-hybrid plasmids were
106 created with each of the *R. capsulatus* RNAP sub-units – α, β, β' and -ω, and tested for
107 binding to T25-GafA. In this assay, a successful interaction between two proteins brings
108 together the T18 and T25 domains of adenylate cyclase and ultimately leads to production
109 of β-galactosidase, which can be measured using colorimetric substrates such as X-gal
110 or O-nitrophenol (Karimova *et al.*, 1998). The α, β and β' sub-units all gave no detectable
111 signal for interaction with GafA, while Rpo-ω produced a strong positive signal in a β-
112 galactosidase assay (Fig. 2A). To confirm this result, MBP-GafA (Fogg, 2019) was bound
113 to amylose magnetic beads and used as bait for capture of purified H6-Rpo-ω. Mock bait
114 beads were simultaneously prepared by identical treatment but with GafA protein omitted.
115 Addition of the Rpo-ω protein to mock beads produced no detectable binding, whereas
116 Rpo-ω was detected in the eluate from the GafA pre-bound beads (Fig. 2B). These data
117 confirm that GafA directly interacts with Rpo-ω, which is then likely to lead to changes in
118 RNAP promoter selection and specific expression of RcGTA genes.

**119 GafA homologues are present throughout the Rhodobacterales and
120 Hyphomicrobiales**

121 The GafA protein shares little primary sequence similarity with any well characterized
122 proteins, but does possess localized similarity with DnaA and sigma factor DNA-binding
123 domains at the N and C-terminal regions of the protein (Fogg, 2019). We performed a
124 BLASTp sequence similarity search using the *R. capsulatus* GafA protein as a query,
125 which revealed hits to genes annotated as DUF6456 domain or helix-turn-helix domain-
126 containing proteins from widespread Rhodobacterales species (Table S1A). This agrees
127 with a previous finding by that GafA homologues were present in all twenty-one complete
128 Rhodobacterales genomes that were available at that time (Hynes *et al.*, 2016). A recent
129 study by Kogay *et al.* (2019) proposed that 60% of the 730 available Hyphomicrobiales
130 (formerly Rhizobiales) genome sequences contained putative RcGTA genes, however,
131 this study only focused on genes within the core structural gene cluster and so did not
132 include *gafA*. We performed additional PSI-BLAST and BLASTp sequence similarity
133 searches with an *R. capsulatus* GafA protein query, but limited the results to the
134 Hyphomicrobiales. Matches were produced to a wide variety species, though sequence
135 similarity was localized to the C-terminal portion of the protein (Fig. S1, Table S1B & C),
136 with the closest sequence similarity found in the final ~18 kDa. Notably, the majority of
137 Hyphomicrobiales homologues were ~22-32 kDa, compared to the 42 kDa *R. capsulatus*
138 GafA, but in most cases the "*gafA*" gene was preceded by a small gene predicted to
139 encode a DnaA-like DNA-binding protein (Fig. S2). Local synteny was also conserved in
140 the Hyphomicrobiales genomes with a downstream gene predicted to encode a cysteine
141 desulfuration enzyme (*sufE*) and an upstream transcriptional regulator annotated as
142 *mucR* or as an HTH-containing gene (Fig. S2). Occasional exceptions appear to be either

143 full length Rhodobacterales-type *gafA* genes with associated Rhodobacterales synteny
144 or Hyphomicrobiales synteny but without a DnaA-like gene (Fig. S2C, Table S1D). Further
145 BLASTp searches with taxonomic limits set for more distantly related Hyphomicrobiales
146 pathogen species (Agrobacteria and Brucellaceae) produced similar results in terms of
147 local synteny and sequence identity (Table S1E & F), suggesting that these genes and
148 gene organization are common throughout the Order.

The GafA central region is important for protein:protein interactions

150 Meanwhile, we predicted the GafA structure from the primary protein sequence using the
151 AlphaFold program (Jumper *et al.*, 2021). All five AlphaFold models placed the two
152 putative DNA-binding domains in equivalent positions and orientations, linked by a central
153 domain of unknown function (Fig. 3A&B). Informed by the structural model and the
154 alignments to Hyphomicrobiales genes (Fig. S1), three bacterial-2-hybrid constructs were
155 produced using truncated *gafA* gene fragments that encode residues 1-226 (N^{1-226}), 87-
156 382 (Cx^{87-382}) and 221-382 ($C^{221-382}$) (Fig. 3A). The three constructs were tested for an
157 interaction with Rpo-ω and both GafA-N^{1-226} and GafA-Cx^{87-382} produced a positive signal,
158 but GafA-$C^{221-382}$ did not (Fig. 3C). To confirm this result, purified MBP-GafA-Cx^{87-382}
159 protein was bound to amylose magnetic beads and used as bait for capture of H6-Rpo-ω
160 in solution. Binding of Rpo-ω to the immobilized MBP-GafA-Cx^{87-382} protein was
161 successfully detected (Fig. 3D). The GafA-N^{1-226} and GafA-Cx^{87-382} constructs overlap in
162 central region of the protein, which suggests that this is the location of GafA:Rpo-ω
163 binding. Additional bacterial-2-hybrid constructs were made to isolate the central region
164 of GafA i.e. amino acids 87-212 (Cen2) and 87-226 (CenN). Both were positive for binding
165 with Rpo-ω (Fig. 3C). These data indicate that GafA is comprised of two distal DNA
166 binding domains and a central protein binding domain. The AlphaFold model (Fig. 3B &
167 Fig. S3) predicted that the central region contains a beta-sheet motif (~amino acids 129-
168 181) that is presented in the opposite direction to the DNA binding motifs, and we
169 hypothesize that this is the interaction interface for Rpo-ω.

170 No experimental structures are available for Rpo-ω proteins from species that are closely
171 related to *R. capsulatus.* An HHPRED search, using *R. capsulatus* Rpo-ω as a query,
172 identified structural similarity matches across the first ~70 amino acids of the protein
173 (Table S2A). AlphaFold models of *R. capsulatus* Rpo-ω closely matched *E. coli* Rpo-ω,
174 but also lacked sufficient confidence at the C-terminal (Fig. S4A-C). The AlphaFold
175 models of Rpo-$ω^{1-71}$ and GafA-CenN were submitted to the LZerD Web Server for protein
176 docking prediction (Christoffer *et al.*, 2021). The results were not conclusive (highest rank
177 sum score = 57) but 6 of the top 10 models predicted that binding occurs with the beta-
178 sheet (Fig. S3D). Further experimental confirmation will be required to definitively pinpoint
179 the binding interface.

GafA NT is not required for autoregulation but essential for GTA activation

To test whether different domains of GafA play different regulatory roles, three regions identified in Fig. 3 (GafA-N$^{1-226}$, C$^{221-382}$ and Cx$^{87-382}$) were cloned into the cumate inducible expression vector pQF. The pQF vectors were introduced into SB1003 wild-type and SB1003 Δ*gafA* strains and tested for various RcGTA production phenotypes. In a RcGTA gene transfer bioassay, GafA-N$^{1-226}$ and GafA-C$^{221-382}$ were unable to induce any RcGTA production in either genetic background (Fig. 4A & S5A). Interestingly, overexpression of both GafA-Cx$^{87-382}$ and full length GafA in wild-type cells stimulated ~80-100-fold greater gene transfer frequencies than the vector only control (Fig. 4A), however, neither were able to complement the *gafA* knock-out (Fig. S5A). These data were corroborated by visualization of intracellular ~4kb RcGTA DNA accumulation by gel electrophoresis (Fig. S5B), detection of characteristic bacteriochlorophyll absorbance peaks in cell-free supernatant indicative of cell lysis (Fig. S5C & D), and western blots to assess accumulation of the RcGTA capsid in the supernatant (Fig. S5E). In all cases, full length GafA and GafA-Cx$^{87-382}$ induced RcGTA production and lysis in wild-type cells but no RcGTA production was detected for Δ*gafA* strains complemented with any *gafA* overexpression constructs. The GafA DnaA-like helix-turn-helix DNA binding motif is very close the N-terminus of the protein (~aa15-55, Fig. 3) and so it is possible that extra residues at this end of the protein interfere with DNA binding (Fogg, 2019). Previous work showed that the full length *gafA* ORF overexpressed from the *puf* photosynthesis promoter effectively complemented the Δ*gafA* mutant (Fogg, 2019), therefore, we produced comparable *puf*-GafA-Nc$^{1-86}$ and N$^{1-226}$ constructs and introduced them into SB1003 wild-type and SB1003 Δ*gafA* strains. Gene transfer bioassays showed that neither GafA-Nc$^{1-86}$ nor N$^{1-226}$ could complement the *gafA* knock-out and neither could induce RcGTA overexpression in the SB1003 wild-type (Fig. 4B & C). Meanwhile, *in trans* expression of full length GafA from the *puf* promoter complemented the SB1003 Δ*gafA* strain and increased SB1003 wild-type gene transfer frequencies by 43.5-fold, compared to the SB1003 plus empty vector control (Fig 4B & C).

The above data indicate that the presence of a short N-terminal Flag tag in the pQF vector impaired complementation, and that full-length GafA is required to induce RcGTA production. However, the fact that overexpression of a truncated *gafA* completely lacking the N-terminal DNA binding motif still induces high level RcGTA production in the presence of a full-length chromosomal copy of *gafA*, indicates that the GafA-Cx$^{87-382}$ region (Fig. 3) can perform at least some of the functions of the full-length protein. We hypothesized that the GafA-Cx$^{87-382}$ portion of the protein can activate the *gafA* promoter independent of the N-terminal DNA binding domain but the full protein is required for wider transcriptional activation of other RcGTA genes. To differentiate the effect of full length GafA and GafA-Cx$^{87-382}$ on RcGTA gene expression in SB1003 wild-type and SB1003 Δ*gafA* cells, transcript abundance of the RcGTA capsid, endolysin and *gafA* genes were

measured by qPCR (Fig. 4D). We used *gafA* primers that bind within the region encoding GafA-Cx and, therefore, the qPCR measured the total combined transcripts of chromosomal and plasmid-borne *gafA* or *gafA*-Cx genes. As expected, overexpression of *gafA* or *gafA*-Cx[87-382] in either the wild-type or knock-out strain produced similar levels of *gafA* transcripts and, consistent with the phenotypic data, this only led to increased RcGTA capsid and endolysin production in wild-type cells. To quantify the activity of the chromosomal *gafA* promoter, we used primers designed to amplify the 5'-UTR that is present only on the chromosome and is also retained in the Δ*gafA* mutant. Transcription from the native promoter was upregulated more than 10-fold when *gafA* or *gafA*-Cx[87-382] were overexpressed in either wild-type or Δ*gafA* cells (Fig. 4D).

Mutation of key residues near the GafA N-terminus impairs RcGTA activation

In agreement with previous work (Fogg, 2019), an HHPRED search for structural homologues of GafA identified tentative hits against numerous sigma factors for both the predicted N- and C-terminal GafA DNA-binding domains (Table S2B-D); C-terminal DBD: E-value>0.84, N-terminal DBD: E-value ≥ 0.0017). However, the N-terminal DBD also produced hits against three DnaA proteins from diverse species in the PDB database, two of which produced E-values ≥ 8.5E-07, as well the DnaA entry from the NCBI conserved domain database (Table S2C). Alignment of the *R. capsulatus* GafA N-terminal region with *E. coli*, *Mycobacterium tuberculosis* and *Aquifex aeolicus* DnaA proteins showed poor primary sequence conservation overall but patches of increased sequence similarity particularly around the residues predicted to bind in the major groove of DNA (Fig. 5A) (Blaesing *et al.*, 2000; Fujikawa *et al.*, 2003).

Ten amino acid locations in the GafA protein were chosen based on sequence conservation or predicted involvement in DNA-binding (Fig. 5A-C). Each position was changed to alanine in the *gafA* complementation plasmid, pCMF180, by site directed mutagenesis. The mutated plasmids were introduced into SB1003 Δ*gafA* to assess the relative ability of each to restore RcGTA production. All mutations had a strong impact on protein function with average gene transfer frequencies at approximately 20% or less compared to the unmutated version of *gafA* (Fig. 5D). The plasmids were also introduced into a *gafA*-null derivative of the RcGTA overproducer strain, *R. capsulatus* DE442. The *gafA* gene is known to be expressed at much higher levels in DE442 than the wild-type SB1003 strain (Fogg, 2019); we hypothesized that a higher dose of some GafA mutants in DE442 might overcome the impaired RcGTA phenotype and reveal which mutations have the greatest effect on function. Most DE442 *gafA* mutants also failed to complement RcGTA production in gene transfer bioassays, with the exception of L34A and L46A (Fig. 5E). In our alignment, L34 and L46 correspond to *E. coli* DnaA I425 and L438 - neither of which directly bind DNA. In the predicted protein structure L34 is also located on a separate helix to the major groove DNA-binding residues (Fig. 5B). *E. coli* DnaA T435

257   (equivalent to GafA S43) binds to specific DNA bases and R442 & K443 (GafA R50 &
258   R51) interact with the DNA backbone (Fig. 5A-C) (Fujikawa *et al.*, 2003). DnaA V437 &
259   Q446 (GafA I45 & Q54) are not predicted to bind DNA but they do sit on the same helix
260   as the residues described above and, therefore, mutations in this region may affect the
261   general conformation of the binding site. DnaA R399 & S400 (GafA E8 & S9) bind in the
262   minor groove of DNA (Fujikawa *et al.*, 2003). The AlphaFold model for GafA placed E8 &
263   S9 at a location unlikely to bind DNA (Fig. 5B), however, this could be due to poor multiple
264   sequence alignment coverage at the protein N-terminus (Fig. S3).

265   Taken together, these data indicate that the truncated GafA-Cx$^{87\text{-}382}$ protein can
266   effectively induce expression from the native *gafA* promoter, but full length GafA is
267   required to induce the various other RcGTA loci. It is likely to be the N-terminal DnaA-like
268   DNA binding domain that is essential for activation of the RcGTA promoters.

N- and C-terminal regions of GafA bind DNA

270   Previous work showed that GafA binds to the RcGTA promoter at a location 75-125 bases
271   upstream of the start codon of TerS (RcGTA *g1*/*rcc01682*) (Fogg, 2019; Sherlock *et al.*,
272   2019). The C-terminal 162 aa of GafA was expressed from a T7 expression vector with
273   an N-terminal MBP tag. The protein was purified to homogeneity and used for
274   electrophoretic mobility shift (EMSA) assays (Fig. 6). As predicted, MBP-GafA-C$^{221\text{-}382}$
275   bound the RcGTA promoter at the previously identified location (Fig. 6A & B), which
276   contains both the -10 and -35 promoter elements plus the transcription start site (TSS)
277   (Fig. 6A). It is also known that *gafA* binds to its own promoter in a 270 base region
278   upstream from the start codon (Fogg, 2019), however, the precise location was not
279   identified. To refine the binding site, we used three 50 bp Cy5-labelled dsDNA oligos
280   covering 150 bases upstream of the start codon for an EMSA binding assay. GafA-C$^{221\text{-}}$
281   ^{382} bound to *gafA* promoter fragment #3, which contains the predicted -35 element (Fig.
282   6D & E).

283   Similar protein expression constructs were also made for the GafA-Nc$^{1\text{-}86}$ and GafA-N$^{1\text{-}}$
284   ^{226} regions with N-terminal His and MBP tags but, as expected, no binding was detected
285   for any DNA targets tested; consistent with data shown in Fig. 4 that N-terminal
286   modifications impair activity of the protein probably by interfering with DNA binding. To
287   resolve this, the affinity purification tag was removed from the MBP-GafA-Nc$^{1\text{-}86}$ protein
288   by digestion with 3c protease, and EMSAs were performed with the tag-free protein.
289   GafA-Nc$^{1\text{-}86}$ produced DNA mobility shifts consistent with non-specific binding to most
290   templates (Fig. 6). Of the five RcGTA promoter fragments tested, four were bound by
291   GafA-Nc$^{1\text{-}86}$ with similar affinities (pGTA #1-4) but only two shifts remained in the
292   presence of an unlabelled non-specific dsDNA competitor (Fig. 6C). The two promoter
293   fragments that produced specific binding were located on either side of the GafA-C$^{221\text{-}382}$

294   binding site (Fig. 6A), which suggests that GafA could bind as a dimer. Analytic gel

295   filtration confirmed that GafA is dimeric in solution, and dimerization is retained for the

296   truncated GafA-Cx$^{87\text{-}382}$ and GafA-C$^{221\text{-}382}$ proteins (Fig. S6). Of the three *gafA* promoter

297   fragments tested, two were bound by GafA-Nc$^{1\text{-}86}$ (pGafA #1 & 2) but neither were specific

298   (Fig. 6F) – consistent with the observation that the GafA N-terminal 86 amino acids are

299   not required for stimulation of the *gafA* promoter.

DISCUSSION

301   Production of Gene Transfer Agents (GTAs) is indirectly controlled by various global

302   regulators in response to environmental stimuli and the disparate signals are integrated

303   via a single transcription factor, GafA. GafA shares little sequence or structural similarity

304   with proteins of known function, but short regions in the N- and C-termini of the protein

305   have tentative structural similarity to DnaA and sigma factor proteins, respectively (Tables

306   S8-10). These regions of similarity are tightly centred around predicted DNA binding

307   domains. The central portion of GafA, between these putative DNA binding domains, is

308   of unknown function. In this paper we sought to refine the mechanism of action for GafA

309   and to assign functions to the various domains. Interestingly, we have identified a direct

310   interaction between GafA and the RNA polymerase ω sub-unit.

The interaction between GafA and Rpo-ω

312   Bacterial DNA-dependent RNA polymerase (RNAP) is responsible for the production of

313   all RNA within a given cell. RNAP is a multi-protein holoenzyme comprised of two identical

314   α-subunits, catalytic β and β' sub-units, and an ω sub-unit encoded by the *rpoZ* gene.

315   The Rpo-ω subunit has been studied in a wide variety of species (Kurkela *et al.*, 2021),

316   where it is thought to stabilize the overall RNAP holoenzyme via direct interactions with

317   the β and β' sub-units (Glyde *et al.*, 2018; Lin *et al.*, 2019; Vassylyev *et al.*, 2002). *R.*

318   *capsulatus* Rpo-ω shares <50% sequence identity with its *E. coli* counterpart, but the

319   MAR ppGpp binding motif and all five conserved residues known to be important for

320   RNAP stabilization are present in both proteins (Kurkela *et al.*, 2021). With the exception

321   of *Mycobacterium tuberculosis* (Mao *et al.*, 2018), deletion of the *rpoZ* gene is not lethal

322   but instead results in various growth defects or alternative phenotypes (Kurkela *et al.*,

323   2021). Indeed, Westbye *et al.* (2017) showed that the growth rate of *R. capsulatus* Δ*rpoZ*

324   is slower than the wild-type and RcGTA production is abrogated (Westbye *et al.*, 2017),

325   the latter of which was confirmed here (Fig. 1).

326   Evidence from multiple species indicates that deletion of Rpo-ω decreases transcription

327   of some housekeeping genes and influences global transcription profiles by promoting

328   RNAP preference for alternative sigma factors (Paget, 2015; Shimada *et al.*, 2014; Weiss

329   *et al.*, 2017; Yamamoto *et al.*, 2018). The role of Rpo-ω in sigma factor selection has

330   largely been inferred from transcriptome data showing expression profiles characteristic

331 of certain sigma factors in wild-type versus Rpo-ω deletion strains, or by the efficiency of
332 sigma factor incorporation into RNAP *in vivo/in vitro* (Geertz *et al.*, 2011; Gunnelius *et al.*,
333 2014; Weiss *et al.*, 2017). Although sigma factors bind to promoter DNA at the -10 and -
334 35 sites, binding is not usually possible *in vitro* in the absence of the RNAP core (Feklístov
335 *et al.*, 2014). Data presented here and elsewhere shows that purified GafA does bind *in
336 vitro* to RcGTA promoters close to the -10/-35 regions (Fig. 6) and that it is the presence
337 of Rpo-ω rather than its absence that leads to expression of GafA regulated genes (Fig.
338 1) (Fogg, 2019; Westbye *et al.*, 2017). Structural data for RNAP complexes from various
339 species show that the Rpo-ω and sigma factor subunits both primarily interact with Rpo-
340 β' but are spatially separated in the stable holoenzyme (Geertz *et al.*, 2011; Glyde *et al.*,
341 2018; Mao *et al.*, 2018; Vassylyev *et al.*, 2002; Weiss *et al.*, 2017). Meanwhile
342 transcription factors usually bind upstream of the -35 element and interact with RNAP via
343 the α-subunit (Browning and Busby, 2004). No binding was detected between GafA and
344 the Rpo-α, Rpo-β or Rpo-β' subunits in a bacterial-2-hybrid assay (Fig. 2).

345 Possible scenarios are: a) GafA first binds to RcGTA promoters and recruits RNAP via
346 an interaction with Rpo-ω; b) GafA pre-recruits RNAP in solution and enhances its affinity
347 for RcGTA promoters, similar to the mechanism thought to be used by the MarA/SoxS
348 family (Griffith *et al.*, 2002; Martin *et al.*, 2002). Perhaps binding to Rpo-ω mimicks the
349 action of ppGpp (Westbye *et al.*, 2017); or c) GafA first binds to Rpo-ω, which then
350 mediates subsequent interactions between GafA and the Rpo-α, β or β' subunits that are
351 not apparent in one-on-one *in vitro* experiments.

Regulation of the RcGTA operons

353 We have sought to update our previous GafA-centric model for RcGTA (Fogg, 2019)
354 regulation with recent discoveries made here and elsewhere (Fig. 7). An important pre-
355 requisite for RcGTA production is high cell density and transition to stationary phase of
356 growth. The response to cell density is mediated by two contrasting influences - a
357 secreted RTX-domain protein represses expression by an unknown mechanism while a
358 quorum sensing signal molecule (homoserine lactone or HSL) promotes RcGTA gene
359 expression (Brimacombe *et al.*, 2013; Ding *et al.*, 2019; Leung *et al.*, 2012; Westbye *et
360 al.*, 2018). Quorum sensing is also important for regulation of the *Dinoroseobacter shibae*
361 GTA (DsGTA) where deactivation of one HSL synthase abolishes any DsGTA gene
362 expression while disruption of another leads to DsGTA overproduction (Koppenhöfer *et
363 al.*, 2019; Tomasch *et al.*, 2018; Wang *et al.*, 2014). Meanwhile, a RelA/SpoT homologue
364 responds to amino acid starvation by increasing intra-cellular concentrations of (p)ppGpp,
365 which is likely to interact directly with RNAP via Rpo-ω, or an alternative binding site, to
366 alter promoter preference (Westbye *et al.*, 2017). It is worth noting that *Bartonella* GTA
367 (BaGTA) production appears to occur in response to low ppGpp concentration, leading
368 to the hypothesis that BaGTA production actually occurs in the fittest cells in a population

369 rather than those under the most stress (Québatte and Dehio, 2019; Québatte *et al.*,
370 2017).

371 The pleiotropic regulator CtrA is absolutely essential for any detectable RcGTA
372 production, and its phosphorylation state controls the transition from RcGTA assembly
373 and DNA packaging to adornment and lysis (Farrera-Calderon *et al.*, 2021; Lang and
374 Beatty, 2000; Mercer *et al.*, 2010, 2012; Westbye *et al.*, 2013, 2018). Hence effective
375 production and release of mature RcGTA particles is dependent upon an intact
376 phosphorylation cascade from the response regulator CckA to CtrA via ChpT (Farrera-
377 Calderon *et al.*, 2021; Wang *et al.*, 2014; Westbye *et al.*, 2018). High levels of intracellular
378 c-di-GMP stimulate the phosphatase activity of CckA and help to maintain a higher
379 concentration of unphosphorylated CtrA(Farrera-Calderon *et al.*, 2021; Pallegar *et al.*,
380 2020b, 2020a). In its unphosphorylated state CtrA is required for transcription of *gafA*
381 (Fogg, 2019). Rpo-ω is not required to activate expression of *gafA* and only the C-terminal
382 region of GafA is required for autoregulation (Fig.1 & 4), which indicates that different
383 mechanisms regulate transcription of *gafA* and the core RcGTA structural locus. It is likely
384 that GafA works together with CtrA to recruit RNAP to the *gafA* promoter but works alone
385 at the core RcGTA promoter via interaction with Rpo-ω (Fig. 7) (Fogg, 2019).

386 Low c-di-GMP levels stimulate CckA kinase activity leading to CtrA
387 phosphorylation(Farrera-Calderon *et al.*, 2021; Pallegar *et al.*, 2020b, 2020a). CtrA-P also
388 increases expression of the PAS domain protein DivL, which further enhances CckA
389 kinase activity (Fogg, 2019; Westbye *et al.*, 2018). CtrA-P then acts in concert with GafA
390 to trigger the various late stage RcGTA genes – head spike (*rcc01079/80* aka *ghsA/B*),
391 tail fibres (*rcc00171*) and lysis genes (*rcc00555/6*) (Fogg, 2019). Putative CtrA half sites
392 were detected in the promoters of all three of these loci and putative GafA binding sites
393 in two out of three (Fig. S7). The housekeeping protease ClpXP degrades both forms of
394 CtrA and is important for maintenance of a proper equilibrium of phosphorylation states
395 (Westbye *et al.*, 2018). Interestingly, deletion of ClpX leads to tailless immature RcGTA
396 particles, reminiscent of DNA packaging mutants(Sherlock *et al.*, 2019).

397 The overall model presented here appears to be mostly complete, with a few notable
398 exceptions. The SoS response regulator LexA is required for RcGTA production but its
399 precise mechanism is unknown, although it appears to act via CckA (Kuchinski *et al.*,
400 2016). There is an SoS box in the LexA promoter and deletion of *lexA* leads to increased
401 expression of *cckA*. This dysregulation presumably unbalances CtrA phosphorylation
402 and/or degradation. Notably, LexA, c-di-GMP, CckA and the phosphorylation of CtrA are
403 all associated with the regulation of DsGTA (Koppenhöfer *et al.*, 2019), but more work is
404 required to fully determine common mechanisms between the species. Another enigmatic
405 protagonist is *rcc01866*, which is located adjacent to the *gafA* gene and is expressed
406 divergently. The Δ*1866* mutant has a phenotype similar to Δ*cckA* i.e. RcGTA particles are

produced but are not fully mature and no detectable lysis occurs (Hynes *et al.*, 2016). We were unable to predict any putative function for the 1866 protein by primary sequence similarity or structural homology searches.

The *gafA* genes beyond the Rhodobacterales

Through bioinformatics analysis, we have identified *gafA* regions with conserved local synteny in the Hyphomicrobiales Order (Fig. S2 & Table S1B-F). The *gafA* homologues are found in a wide variety of species throughout the Order including several important pathogens such as those from the *Brucella* (Chain *et al.*, 2011) and *Agrobacterium* genera (Scholz *et al.*, 2008). The *Brucella gafA* genes have also previously been implicated as virulence/fitness factors of unknown function in high-throughput studies (He, 2012; Salmon-Divon *et al.*, 2019). Intriguingly, the Hyphomicrobiales *gafA* is split into two separate genes that roughly correspond to the GafA-Nc and GafA-Cx constructs used in this study, supporting the hypothesis that these domains have distinct biological roles.

Overall our data suggest that GafA either acts as an alternative sigma factor or as a transcription factor that is recruited by Rpo-ω via a direct protein:protein interaction (Lane and Darst, 2006; Lin *et al.*, 2019; Li *et al.*, 2019), and this interaction occurs via the central domain of the protein (Fig. 2 & 3). GafA has two mechanisms of action, one Rpo-ω dependent and one Rpo-ω independent, and the GafA N-terminal DNA binding domain is essential only for the former. Further research is ongoing to determine the precise mechanism of GafA and to establish how widespread this mechanism is in bacteria.

Limitations of the study

The DNA sequence bound by GafA was predicted from the short regions of the GafA and RcGTA promoters identified by EMSA analysis, however, a more extensive experimental approach will be required to confirm and refine these predictions e.g. systematic DNA mutagenesis. Although we demonstrated that GafA interacts with RNAP via the Omega sub-unit to co-ordinate RcGTA expression, we did not present a definitive mechanism for how RNAP promoter preference is altered. We envisage that this will be resolved via biochemical/structural approaches for the whole RNAP holoenzyme in complex with GafA and DNA.

Author Contributions

Conceptualization, P.C.M.F; Methodology, D.S. and P.C.M.F.; Investigation, D.S. and P.C.M.F.; Writing, P.C.M.F.; Visualization, P.C.M.F.; Funding Acquisition, P.C.M.F.; Resources, P.C.M.F.; Supervision, P.C.M.F.

Acknowledgements

450 **Declaration of Interests**

451 The authors declare no competing interests

Figure Titles and Legends

Figure 1. The *rpoZ* gene is essential for RcGTA production. For all panels the following strains were used – *R. capsulatus* SB1003 (WT) and a *rpoZ* KO derivative (ΔrpoZ*)*. Strains were complemented *in trans* with empty pCM66T vector (WT and ΔrpoZ), *rpoZ* expressed from its native promoter (+*rpoZ*), *gafA* expressed from its native promoter (+gafA), or *gafA* overexpressed from a cumate inducible promoter (+*gafA* OX). **A.** Representative western blot of *R. capsulatus* concentrated supernatants using an α-RcGTA capsid antibody. **B.** Bar chart showing the frequency of rifampicin gene transfer by the indicated strains, N=6. **C.** Quantitative PCR data showing *gafA* transcript abundance in the indicated strains relative to the *R. capsulatus* SB1003 (WT) control. Expression levels shown were calculated using the ΔΔCt method (*uvrD* reference gene) and a log2 transformation to give fold-differences. N=3. **D.** Bar chart of the frequency of rifampicin gene transfer by the annotated strains. N=4. Statistical significance is indicated on each graph as calculated by one-way ANOVA with the Holm-Sidak method for pairwise multiple comparison (***= $p < 0.001$, **= $p < 0.01$, *= $p < 0.05$, n.s.= $p > 0.05$). See also Data S1.

Figure 2. GafA directly interacts with Rpo-ω. A. Quantification of bacterial-2-hybrid interactions between T25-GafA vs T18-Rpo-α, T18-Rpo-β, T18-Rpo-β' and T18-Rpo-ω. Negative control is T25-gafA vs pUT18 empty vector (-ve). N=3. Statistical significance is indicated on the graph as calculated by one-way ANOVA with the Holm-Sidak method for pairwise multiple comparison (***= $p < 0.001$). **B.** Silver stained SDS PAGE gel of a pull-down assay using MBP-GafA as bait and H6-Rpo-ω as prey. Amylose magnetic beads that should only bind to MBP-tagged proteins were used to capture the proteins. Presence or absence of each protein in the assay is indicated by '-' or '+' symbols above the gel. Abcam broad range protein marker is included for reference. See also Data S1.

Figure 3. The domain structure of the *R. capsulatus* GafA. A. Amino acid sequence of GafA, colour coded to highlight the different regions used for subsequent characterization. Green = N-terminal concise (Nc, residues 1-86), Green & Blue = N-terminal (N, residues 1-226), Turquoise = C-terminal (C, residues 221-382), Blue-Turquoise = C-terminal extended (Cx, residues 87-382), Blue = Central region 2 (Cen2, residues 87-212), Blue & Purple = Central region N (CenN, residues 87-226). **B.** Alphafold structure prediction for GafA, regions used for subsequent characterization are colour coded as in part A and annotated above and below the image. The two predicted DNA binding domains (DBD) are annotated with arrows. **C.** Quantification of bacterial-2-hybrid interactions between T18-Rpo-ω and various T25-GafA constructs (defined above) by β-galactosidase assay, N=3. Statistical significance is indicated on the graph as calculated by one-way ANOVA with the Holm-Sidak method for pairwise multiple comparison (***= $p < 0.001$, n.s.= $p > 0.05$). **D.** InstantBlue stained SDS PAGE gel of a pull-down assay using MBP-GafA-Cx as bait and H6-Rpo-ω as prey. Amylose magnetic beads were used to capture the proteins. Presence or absence of each protein in the assay is indicated by '-' or '+' symbols above the gel. Abcam broad range protein marker and a lane showing the Rpo-ω protein input are included for reference. See also Figure S1-4, Table S1 & 2, Data S1.

14

Figure 4. Characterization of GafA domain function. A-C. Bar charts of the relative frequency of rifampicin gene transfer from **A.** *R. capsulatus* SB1003 wild-type donor strains complemented *in trans* with empty pQF vector (WT), full length *gafA* (Σ), or truncated regions of *gafA* described in Fig. 3 (N, C and Cx), N = 3. **B.** *R. capsulatus* SB1003 wild-type donor strains complemented *in trans* with empty pCM66T vector (WT, N=3) or with the *puf* promoter driving expression of full length *gafA* (Σ, N=4), or truncated regions of *gafA* (Nc and N, N=4) or **C.** *R. capsulatus* SB1003 Δ*gafA* donor strains complemented *in trans* with empty pCM66T vector (WT, N=8) or with the *puf* promoter driving expression of full length *gafA* (Σ, N=4), or truncated regions of *gafA* (Nc and N, N=7). **D.** Transcript abundance of RcGTA genes in *gafA* overexpression strains. The bar chart shows relative changes in transcript abundance measured using quantitative PCR and the ΔΔCt method. The *R. capsulatus* strains tested are annotated in the legend – SB1003 containing empty pQF (WT), SB1003 complemented with pCMF254 (WT + *gafA* OX), SB1003 complemented with pCMF264 (WT + *gafA* Cx OX) and SB1003 *gafA* knock-out complemented with pCMF264. Transcripts of *gafA*, RcGTA capsid (*rcc01687*), RcGTA endolysin (*rcc00555*) and the *gafA* 5' UTR (*pGafA*) were measured. Statistical significance for all panels is indicated above each graph, and was calculated by one-way ANOVA with the Holm-Sidak method for pairwise multiple comparison (***= p<0.001, n.s.= p>0.05). See also Figure S5.

Figure 5. Mutagenesis of the GafA N-terminal DNA binding domain. A. Alignment of *R. capsulatus* GafA (Rc_GafA) residues 1-59 with the DnaA DNA binding domains from *E. coli* (PDB: 1J1V), *Mycobacterium tuberculosis* (PDB: 3PVV), *Aquifex aeolicus* (PDB: 1L8Q) and *R. capsulatus* (Rc_DnaA). Conserved amino acids are coloured using the CLUSTLx scheme and mutated positions are indicated by black boxes. **B.** Alphafold structure prediction for the GafA N-terminal DNA binding domain. Side chains are shown for the amino acid positions mutated in this study, and each are coloured according to predicted interaction with DNA; Red – specific base interaction, Blue – nonspecific interaction with DNA backbone, Green – no direct interaction. **C.** DNA binding domain from *E. coli* DnaA (PDB: 1J1V). Equivalent amino acids to those mutated in GafA are coloured using the same scheme as in panel B. R399 and S400 are annotated as they sit in the minor groove of DNA, whereas their GafA counterparts (E8/S9) were predicted to have no proximity to the DNA – probably due to limitations of the model at the sequence extremity. **D & E.** Relative gene transfer frequencies for *gafA* gene knock-outs of **D** the wild-type strain *R. capsulatus* SB1003 (N=4, except E8 where N=3) and **E** the RcGTA overproducer strain DE442 (N=4, except S9 where N=3). Each strain was complemented *in trans* with either empty pCM66T vector (Δ) or the *gafA* gene with single point mutations as indicated on the X-axis. Frequencies shown are normalized to complementation of the respective strains (SB1003 or DE442) with unmodified *gafA*. Statistical significance tested by one-way ANOVA with the Holm-Sidak method for pairwise multiple comparison – all *gafA* point mutations were statistically different from wild-type *gafA* (p<0.001) except DE442 complemented with *gafA* L34A or L46A (p>0.05).

Figure 6. Binding of GafA domains to the RcGTA and *gafA* promoters. A. Schematic of the RcGTA promoter region with the start codon (ATG), transcription start site (TSS) and -10/-35 promoter elements annotated. The locations of DNA fragments used for EMSA band shifts are shown and labelled #1 to #4. **B.** Representative EMSA of GafA-C[221-382] binding to RcGTA promoter fragment #2. **C.** Representative EMSA of GafA- Nc[1-86] binding specifically to RcGTA

539   promoter fragments #1 & 3 and non-specifically to #2 & 4. Protein concentration is labelled above
540   the image. N = excess of non-specific competitor DNA added, S = excess of specific competitor
541   DNA added. **D.** Schematic of the *gafA* promoter region with the start codon (ATG), transcription
542   start site (TSS), CtrA binding site (CtrA) and -10/-35 promoter elements annotated. The locations
543   of DNA fragments used for EMSA band shifts are shown and labelled #1 to #3. **E.** Representative
544   EMSA of GafA-C$^{221\text{-}382}$ binding to *gafA* promoter fragment #3. **F.** Representative EMSA of GafA-
545   Nc$^{1\text{-}86}$ binding non-specifically to *gafA* promoter fragments #1 & 2. Protein concentration is
546   labelled above the image. N = excess of non-specific competitor DNA added, S = excess of
547   specific competitor DNA added. See also Figure S6&7.
548

549   **Figure 7. Model of RcGTA regulation.** Known contributors to RcGTA regulation are indicated
550   and broadly classified based on whether their major influence is on early production of structural
551   proteins (Stage 1) or late stage maturation and lysis (Stage 2). Arrows indicate positive regulation
552   and flat headed arrows indicate repression. Black arrows represent transcriptional regulation, blue
553   arrows represent post-translational or ligand:protein regulation, red arrows represent biosynthesis
554   or degradation, dashed arrows indicate uncertain mechanism. Arrows representing GafA
555   regulation that requires Rpo- ω are annotated with 'ω', and Rpo- ω independent regulation by the
556   GafA Cx domain is labelled 'Cx'.

STAR Methods

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Paul Fogg (paul.fogg@york.ac.uk).

Materials availability

All unique reagents or materials generated in this study will be made available on request by the lead contact, but we may require a completed materials transfer agreement if there is potential for commercial application.

Data and code availability

- All data reported in this paper will be shared by the lead contact upon request
- This paper does not report original code
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Bacterial Strains

Three *Rhodobacter capsulatus* strains were used in this study: rifampicin sensitive wild-type strain B10 (Wall *et al.*, 1975), a rifampicin resistant derivative SB1003 (ATCC BAA-309) and an RcGTA overproducer strain DE442 (Ding *et al.*, 2014; Fogg *et al.*, 2012). All *R. capsulatus* cultures were grown at 30°C either aerated in the dark or in anoxic sealed tubes under constant illumination. Two growth media were used – YPS complex broth (0.3 % w/v yeast extract, 0.3% w/v peptone, 2 mM $MgCl_2$, 2 mM $CaCl_2$) or RCV defined broth (10 mM potassium phosphate buffer, 0.4% w/v L-malic acid, 0.1% w/v $(NH_4)_2SO_4$, 0.020% w/v $MgSO_4 \cdot 7H_2O$, 0.0075% w/v $CaCl_2 \cdot 2H_2O$, 0.0012% w/v $FeSO_4 \cdot 7H_2O$, 0.0020% w/v Na_2EDTA, 0.0001% w/v thiamine hydrochloride. Plus 1 ml of trace element solution - 0.07% w/v H_3BO_3, 0.040% w/v $MnSO_4 \cdot H_2O$, 0.019% w/v $Na_2MoO_4 \cdot 2H_2O$, 0.006% w/v $ZnSO_4 \cdot 7H_2O$, 0.001% w/v $Cu(NO_3) \cdot 3H_2O$. The pH was adjusted to 6.8 with NaOH). For agar plates, 1.5% w/v agar was added to the above broth recipes. The *E. coli* S17-1 strain (DSM 9079), which contains chromosomally integrated *tra* genes, was used as a donor for all conjugations. NEB 10-beta Competent *E. coli* cells (New England Biolabs) were used for standard cloning and plasmid maintenance; T7 Express

17

Competent *E. coli* cells (New England Biolabs) were used for overexpression of proteins for purification.

METHOD DETAILS

Plasmid Construction

Cloning and Site Directed Mutagenesis

A full list of all plasmids and oligonucleotides used in the study can be found in Tables S3 and S4. All oligonucleotides were obtained from Integrated DNA Technologies (IDT) and designed with an optimal annealing temperature of 60°C when used with Q5 DNA Polymerase (New England Biolabs). Plasmid DNA was purified using the Monarch Plasmid Miniprep Kit (New England Biolabs). The destination plasmids pCM66T, pKT25, pUT18 and pUT18C were linearized by digestion with BamHI restriction enzyme (New England Biolabs), pETFPP_1 & 2 was linearized by PCR using inverse primers CleF and CleR. Inserts were amplified using primers with 15 bp 5' overhangs that have complementary sequence to the plasmid DNA with which it was to be recombined. All cloning reactions were carried out with the NEBuilder Cloning Kit (New England Biolabs). Site-directed mutagenesis was achieved by inverse PCR using Q5 DNA polymerase overlapping primers (offset by 8 to 10 bp) containing the desired mutation in the centre of the overlap region. Amplified DNA was cleaned using the Monarch PCR & DNA Cleanup Kit (New England Biolabs), then digested with DpnI restriction endonuclease (New England Biolabs) overnight at 37°C and introduced without further treatment into chemically competent *E. coli* by transformation.

Transformation and conjugation

Plasmids were introduced into *E. coli* by standard heat shock transformation (Maniatis *et al.*, 1982), and into *Rhodobacter* by conjugation. One millilitre aliquots of overnight cultures of the *E. coli* S17-1 donor and *Rhodobacter* recipient strains were centrifuged at 5,000 x g for 1 min, washed with 1 ml YPS broth, centrifuged again and resuspended in 100 µl YPS broth. 10 µl of concentrated donor and recipient cells were mixed and spotted onto YPS agar or spotted individually as negative controls. Plates were incubated o/n at 30°C. Spots were scrapped, suspended in 100 µl YPS broth and plated on YPS + 100 µg/ml rifampicin (counter-selection against *E. coli*) + 10 µg/ml kanamycin or 1 µg/ml tetracycline (plasmid selection). Plates were incubated o/n at 30°C then restreaked onto fresh selective agar to obtain pure single colonies.

Gene Knockouts

18

621 Knockouts were created by RcGTA transfer. pCM66T plasmid constructs were created
622 with a gentamicin resistance cassette flanked by 500-1000 bp of DNA from either side of
623 the target gene. Assembly was achieved by a one-step, four component NEBuilder (NEB)
624 reaction and transformation into NEB 10-beta cells. Deletion constructs were introduced
625 into the RcGTA hyperproducer strain and a standard GTA bio-assay (see below) was
626 carried out to replace the intact chromosomal gene with the deleted version.

627 *Rhodobacter* **Gene Transfer Assays**

628 *Rhodobacter* assays were carried out essentially as described in (Leung and Beatty,
629 2013). RcGTA donor cultures were grown photosynthetically (anoxic) with illumination in
630 YPS for ~72 h and recipient cultures were grown under chemotrophic conditions in RCV
631 for ~24 h. Cells were cleared from donor cultures by centrifugation and the supernatant
632 filtered through a 0.45 μm syringe filter. Recipient cells were concentrated 3-fold by
633 centrifugation at 5,000 x g and resuspension in 1/3 volume of G-Buffer (10 mM Tris-HCl
634 pH 7.8, 1 mM $MgCl_2$, 1 mM $CaCl_2$, 1 mM NaCl, 0.5 mg/ml BSA). Reactions were carried
635 out in polystyrene culture tubes (Starlab) containing 400 μl G-Buffer, 100 μl recipient cells
636 and 100 μl filter donor supernatant, then incubated at 30°C for 1 h. 900 μl YPS was added
637 to each tube and incubated for a further 3 h. Cells were harvested by centrifugation at
638 5,000 x g and plated on YPS + 100 μg/ml rifampicin (for standard GTA assays) or 3 μg/ml
639 gentamicin (for gene knock-outs).

640 **Nucleic Acid Purification**

641 One millilitre samples of relevant bacterial cultures were taken for each nucleic acid
642 purification replicate. Total DNA was purified according to the "Purification of Nucleic
643 Acids by Extraction with Phenol:Chloroform" protocol (Maniatis *et al.*, 1982). Cells were
644 resuspended in 567 μl TE then 30 μl 10% (w/v) SDS and 3 μl 20 μg/ml Proteinase K were
645 added. Cells were incubated at 37°C for 1 h. To each tube, 100 μl of 5 M NaCl was added
646 and thoroughly mixed by inversion. Eighty microlitres of 1% (w/v) CTAB was added,
647 mixed thoroughly by inversion and the cells were incubated at 65°C for 10 minutes. An
648 equal volume of Phenol:chloroform:isoamyl alcohol (25:24:1, pH 8) was added and mixed
649 vigorously. The tubes were centrifuged at 15,000 x g for 10 min. The upper aqueous layer
650 was removed to a fresh tube and the Phenol:chloroform:isoamyl alcohol treatment was
651 repeated at least two times or until the white interphase was no longer visible. An equal
652 volume of chloroform was added and mixed vigorously. The tubes were centrifuged at
653 15,000 x g for 2 min. The upper aqueous layer was transferred to a fresh tube and DNA
654 was precipitated by addition of 0.6 volumes of ice-cold isopropanol. Precipitation was
655 allowed to proceed at -20°C for 1 h. DNA was harvested by centrifugation at 15,000 x g
656 for 15 min, and the supernatant was discarded. The pellet was washed with 70% ethanol,
657 centrifuged at 15,000 x g for 15 min and the supernatant was discarded. The pellet was

658 allowed to air dry for ~15 min then resuspended in TE buffer. Total RNA was purified
659 using the NucleoSpin RNA Kit (Macherey-Nagel) and DNAseI treated on column
660 according to the recommended protocol. RNA was quantified using a Nanodrop
661 spectrophotometer. 1 µg of total RNA was converted to cDNA using the LunaScript RT
662 SuperMix Kit (NEB).

Quantitative Reverse Transcriptase PCR

664 One in fifty dilutions of the cDNA template were prepared and 1 µl used per reaction.
665 Reactions contained Fast Sybr Green Mastermix (Applied Biosystems), cDNA and
666 primers (500 nM). Standard conditions were used with an annealing temperature of 60°C.
667 All primer efficiencies were calculated as between 90 and 110%. Relative gene
668 expression was determined using the ΔΔCt method (Livak and Schmittgen, 2001). For
669 each sample, variance was calculated for three independent biological replicates, which
670 were each the mean of three technical replicates. QuantStudio 3 Real-Time PCR System
671 was used for all experiments (Applied Biosystems).

GafA Overexpression in *Rhodobacter*

673 Gene overexpression in *Rhodobacter* was achieved by a transcriptional fusion of the
674 genes of interest to a cumate inducible promoter in the plasmid pQF (Kaczmarczyk *et al.*,
675 2013) or to the *R. capsulatus puf* photosynthesis promoter in pCM66T (Fogg, 2019; Fogg
676 *et al.*, 2012). For overexpression experiments using the *puf* promoter, donor cultures were
677 first grown chemotrophically in the presence of oxygen to stationary phase then diluted
678 1:1 in fresh media and switched to anoxic photosynthetic growth for 6 h. pQF was a gift
679 from Julia Vorholt (Addgene plasmid #48095). Overexpression was induced by addition
680 of cumate to late log growth phase cultures at a final concentration of 50 µM.

Bacterial-two-hybrid (B2H) assays

682 The procedure and the resources were as described in (Karimova *et al.*, 1998). Plasmids
683 encoding T18 (pUT18C and derivatives) and the compatible plasmids encoding T25
684 (pKT25 and derivatives) were introduced pairwise into competent BTH101 by co-
685 transformation. Selection was using LB agar containing 50 µg/ml kanamycin, 100 µg/ml
686 ampicillin, 1 mM IPTG and 80 µg/ml X-Gal, and plates were incubated at 30°C for 48 h to
687 allow blue colour to develop. Colonies obtained from the B2H plate assay were used to
688 inoculate 5 ml of LB supplemented with 50 µg/ml kanamycin, 100 µg/ml ampicillin and 1
689 mM IPTG in a 96-well plate. Plates were incubated for 16 h at 30°C with agitation.
690 Absorbance (OD_{600}) readings of culture density were taken. Meanwhile, 80 µl aliquots of
691 permeabilization solution (100 mM Na_2HPO_4, 20 mM KCl, 2 mM $MgSO_4$, 0.5 mg/ml
692 lysozyme) were mixed with 20 µl of each bacterial culture, then incubated at room
693 temperature for 30 min. Six hundred microliters of substrate solution (60 mM Na_2HPO_4,

40 mM NaH$_2$PO$_4$, 1 mg/ml ONPG) was added and the mixture was incubated at room temperature. Once sufficient colour had developed, stop solution (1 M sodium carbonate) was added and the precise time noted. Cell debris was removed by centrifugation and absorbance (OD$_{420}$) readings were taken. Miller units were calculated according to the formula MU=1,000(Abs420/(Abs600*0.02 ml*time in min)).

Protein Purification

For His6-tagged proteins, 500 ml cultures of *E. coli* containing the relevant expression plasmid were induced at mid-exponential growth phase with 0.2 mM IPTG overnight at 20°C (Fogg and Wilkinson, 2008). Concentrated cells were lysed in 20 ml binding buffer (0.5 M NaCl, 75 mM Tris; pH 7.75) plus 0.2 mg ml^{-1} lysozyme and 500 U Basemuncher Endonuclease (Expedeon Ltd.) for 30 min on ice and then sonicated. Cleared supernatant was applied to a 5 ml HisTrap FF crude column (Cytiva) and the bound, his-tagged protein was eluted with 125 mM imidazole. Eluted protein was desalted on a HiPrep 26/10 desalting column (Cytiva) and then further separated by size exclusion chromatography on a HiLoad 16/60 Superdex 200 preparative grade gel filtration column. All chromatography steps were carried out on an AKTA Prime instrument (Cytiva). Purified proteins were concentrated in a Spin-X UF Centrifugal Concentrator (Corning) and quantified by the nanodrop extinction co-efficient method (Thermo Scientific). Samples were stored at -80 °C in binding buffer plus 50% glycerol. MBP-tagged proteins were purified as above except MBP binding buffer was used (200 mM NaCl, 20 mM Tris, 1 mM EDTA; pH 7.4), the lysate was applied to a 5 ml MBPTrap FF column (Cytiva) and purified protein was eluted with 10 mM maltose in binding buffer.

Analytical Gel Filtration

Protein multimeric states were estimated using a Superdex 200 increase 10/200 GL column (Cytiva). MBP binding buffer was used for all filtration runs. A protein molecular weight standard (1.3–670 kDa, Bio-Rad Laboratories) was run through the column at 0.75 ml/min and the peaks produced were used to construct a standard curve (R^2=1, predicted error for 17-670 kDa is <2%). Samples of each protein were sequentially run on the column and molecular weights were estimated from the elution volume and the equation derived from the standard curve.

Electrophoretic motility shift assays (EMSA)

For all 50 bp binding substrates, 50 base Cy5 5'-labelled oligos (IDT) were annealed to unlabelled complimentary oligos (IDT). Both oligos were mixed to a final concentration of 40 µM in annealing buffer (1 M Potassium Acetate, 300 mM HEPES; pH 7.5) and heated to 98°C for 5 min then allowed to cool to room temperature. Ten microliter EMSA mixtures contained 80 nM annealed Cy5-dsDNA, standard binding buffer (25 mM HEPES, 50 mM

K-glutamate, 50 mM MgSO$_4$, 1 mM dithiothreitol, 0.1 mM EDTA, 0.05% Triton X-100; pH 8.0) (Wiethaus *et al.*, 2008), 1 μg poly dI:dC, 4% glycerol and the specified concentrations of purified protein (Wiethaus *et al.*, 2006). 500-fold excess of competitor DNA was added to control mixtures – specific competitor was unlabelled but otherwise identical to the binding substrate and the non-specific competitor was an unlabelled 50 bp annealed oligo matching an arbitrary location elsewhere in the *R. capsulatus* genome. All assays were incubated for 30 min at room temperature then immediately loaded onto a 7 % Acrylamide gel (1 x TBE) without loading dye. Gels were run at 80 V for 90 min at room temperature in 1 x TBE. Fluorescence was imaged using a Typhoon Biomolecular Imager (Amersham) and analysed using ImageQuant (Amersham) and FIJI (Schindelin *et al.*, 2012) software.

Protein ligand pull down assays

One hundred microliters of 2 mg/ml MBP-tagged protein in binding buffer (200 mM NaCl, 20 mM Tris, 1 mM EDTA; pH 7.4) was incubated with 100 μl of amylose magnetic beads (New England Biolabs) at 4°C for 1 hour on a rolling platform. Mock beads were created by an identical method but using 100 ul of binding buffer without protein. Beads were washed 5 times with 500 μl of wash buffer (binding buffer + 0.05% Tween20) and resuspended in a final volume of 100 μl. For pull downs, 25 μl of prepared beads were harvested in a magnetic stand and the supernatant was replaced with either 100 μl of binding buffer alone or binding buffer containing 2 mg/ml H6-RpoZ. The beads were incubated for 2 h at 4°C on a rolling platform, then washed five times with wash buffer. To elute proteins, 50 μl of elution buffer was added (binding buffer + 10 mM maltose). LDS buffer (Abcam) was added to the eluate and heated to 90°C for 10 min. Twenty microliters of each sample were run on a 4-20% TruPAGE denaturing gradient gel (Merck Life Science Ltd) and the bands were visualized using Pierce silver stain for mass spectrometry (Thermo Scientific) or InstantBlue Coomassie stain (Abcam). Five microliters of extra broad molecular weight prestained protein ladder was used for size estimation (Abcam).

Western Blotting

Rhodobacter capsulatus supernatants were concentrated 10-fold using a SpeedVac (Thermo Scientific). Fifteen microliter samples were mixed with 5 μl LDS sample buffer (Abcam). heated to 90°C for 10 min and then run on 4-20% TruPAGE denaturing gradient gel (Merck Life Science Ltd). Proteins were transferred to a PVDF membrane using a Mini-PROTEAN Tetra Cell blotting module (Bio-Rad Laboratories) in 1X transfer buffer (25 mM tris base, 0.2 M glycine, 20% methanol; pH8.5), 100 V for 1 h. The membrane was blocked in 5% (w/v) skimmed milk powder in 1X TBS for 1 h at room temperature. The anti-RcGTA major capsid protein antibody (Agrisera Ltd) was used at 1:1000 dilution in blocking buffer overnight at 4°C, followed by four 10 min washes in TBST. The

767 secondary HRP-antibody conjugate was used at 1:2500 dilution in blocking buffer for 2 h
768 at room temperature, followed by four 10 min washes in TBST. SuperSignal west femto
769 maximum sensitivity substrate (Thermo Scientific) was used to develop the western and
770 the signal was detected using an iBRIGHT chemi-imager (Thermo Scientific).

Sequence similarity analysis

772 NCBI BLASTp and PSI-BLAST searches for GafA homologues were performed using the
773 default parameters - expect threshold=0.05, word size=6 or 3 (respectively), blosum62
774 similarity matrix, gap costs of existence=11 and extension=1. Queries were made against
775 the non-redundant protein sequences database (nr; posted:May 5th 2022). Where
776 indicated, taxonomic constraints were applied to limit results to the Hyphomicrobiales
777 (taxid:356), Brucellaceae (taxid:118882) and Agrobacterium (taxid:357). A tBLASTn
778 search was made using a GafA homologue from Roseibium sp. RKSG952 as a query
779 and using the default parameters - expect threshold=0.05, word size=6, blosum62
780 similarity matrix, gap costs of existence=11 and extension=1. The nucleotide collection
781 database was used (nr/nt; May 9th 2022 update). A summary of the full outputs can be
782 found in Table S1.

783 HHPRED analysis of GafA was carried out using the "pdb_mmcif70_14_Apr" and
784 "NCBI_Conserved_Domains(CD)_v3.18" databases accessed on the 8th May 2022
785 (Gabler *et al.*, 2020; Zimmermann *et al.*, 2018). Full length GafA protein sequence and
786 two shorter sequences, focused on the two predicted DNA binding domains, were used
787 as queries. The default parameters were used in each case i.e. HHBlits UniRef30 MSA
788 generation method, maximal generation steps = 3 and an E-value threshold of 1e-3.
789 Minimum coverage was 20%, minimum sequence identity was 0%. Secondary structure
790 scoring was done during alignment. A summary of the full outputs can be found in Table
791 S2.

Protein structure and function prediction

793 Three-dimensional structures for the *R. capsulatus* GafA and Rpo-ω proteins were
794 predicted using the Alphafold co-lab server using the msa_method:jackhammer and all
795 other parameters set to default (Jumper *et al.*, 2021). GafA predictions were made on 30th
796 Sept 2021 and RpoZ predictions were made on 3rd February 2022. Protein structures
797 were visualized using the UCSF ChimeraX version 1.1 (Goddard *et al.*, 2018).
798 Protein:protein interaction predictions were produced using the LZerD protein docking
799 algorithm on the LZerD web server using default parameters (Christoffer *et al.*, 2021).
800 Helix-turn-helix predictions were carried out using NPS@ (Combet *et al.*, 2000; Dodd and
801 Egan, 1990) and Gym2.0 (Narasimhan *et al.*, 2002) using the default settings. Promoter
802 -10/-35 elements were predicted with BPROM (Solovyev and Salamov, 2011). Clustal-ω
803 (Sievers *et al.*, 2011) was used for DNA/protein alignments and Jalview version: 2.11.2.2

804   (Waterhouse *et al.*, 2009) was used to visualize these alignments; relevant
805   similarity/identity colour schemes are indicated in the figure legends.

QUANTIFICATION AND STATISTICAL ANALYSIS

807   CorelDraw 2018 (Corel Corporation) was used for figure preparation Statistical analysis
808   was carried out using Sigmaplot software version 13 (Systat Software Inc.) and, for each
809   use, the test parameters are indicated in the figure legends and, where appropriate, in
810   the main text. All graphs present the means as a bar chart and the individual data points
811   are overlaid as discrete dots. All N values quoted refer to distinct biological replicates.

SUPPLEMENTAL ITEM TITLES AND LEGENDS

Table S1. Full BLAST search results for **S1A** *R. capsulatus* SB1003 GafA full length protein query (BLASTp), **S1B** *R. capsulatus* SB1003 GafA full length protein query (BLASTp; hits limited to Hyphomicroberales), **S1C** *R. capsulatus* SB1003 GafA full length protein query (PSI-BLAST - 2 iterations; hits limited to Hyphomicroberales), **S1D** *Roseibium sp. RKSG952* GafA full length protein query (tBLASTn), **S1E** *R. capsulatus* SB1003 GafA full length protein query (BLASTp; hits limited to *Agrobacteria*), and **S1F** *R. capsulatus* SB1003 GafA full length protein query (BLASTp; hits limited to Brucellaceae). Related to Figure 3 and STAR methods.

Table S2. Full HHPRED search results for **S2A** full length *R. capsulatus* SB1003 RpoZ query, **S2B** full length *R. capsulatus* SB1003 GafA query, **S2C** *R. capsulatus* SB1003 GafA N-terminal DNA-binding domain query, and **S2D** *R. capsulatus* SB1003 GafA C-terminal DNA-binding domain query. Related to Figure 3 and STAR methods.

Table S3. A complete list of all plasmids used in this study. Related to STAR methods.

Table S4. A complete list of all oligonucleotides used in this study. Related to STAR methods.

Data S1. Uncropped western blots, SDS PAGE and agarose gel images. Related to Figures 1, 2, 3 and S5.

References

Berglund, E.C., Frank, A.C., Calteau, A., Vinnere Pettersson, O., Granberg, F., Eriksson, A.-S., Näslund, K., Holmberg, M., Lindroos, H., and Andersson, S.G.E. (2009). Run-off replication of host-adaptability genes is associated with gene transfer agents in the genome of mouse-infecting *Bartonella grahamii*. PLoS Genet. *5*, e1000546.

Blaesing, F., Weigel, C., Welzeck, M., and Messer, W. (2000). Analysis of the DNA-binding domain of *Escherichia coli* DnaA protein. Mol. Microbiol. *36*, 557–569.

Brimacombe, C.A., Stevens, A., Jun, D., Mercer, R., Lang, A.S., and Beatty, J.T. (2013). Quorum-sensing regulation of a capsular polysaccharide receptor for the *Rhodobacter capsulatus* gene transfer agent (RcGTA). Mol. Microbiol. *87*, 802–817.

Browning, D.F., and Busby, S.J. (2004). The regulation of bacterial transcription initiation. Nat. Rev. Microbiol. *2*, 57–65.

Chain, P.S.G., Lang, D.M., Comerci, D.J., Malfatti, S.A., Vergez, L.M., Shin, M., Ugalde, R.A., Garcia, E., and Tolmasky, M.E. (2011). Genome of *Ochrobactrum anthropi* ATCC 49188 T, a versatile opportunistic pathogen and symbiont of several eukaryotic hosts. J. Bacteriol. *193*, 4274–4275.

Christoffer, C., Bharadwaj, V., Luu, R., and Kihara, D. (2021). LZerD Protein-Protein docking webserver enhanced with *de novo* structure prediction. Front. Mol. Biosci. *8*, 724947.

Combet, C., Blanchet, C., Geourjon, C., and Deléage, G. (2000). NPS@: network protein sequence analysis. Trends Biochem. Sci. *25*, 147–150.

Ding, H., Moksa, M.M., Hirst, M., and Beatty, J.T. (2014). Draft genome sequences of six *Rhodobacter capsulatus* strains, YW1, YW2, B6, Y262, R121, and DE442. Genome Announc. *2*.

Ding, H., Grüll, M.P., Mulligan, M.E., Lang, A.S., and Beatty, J.T. (2019). Induction of *Rhodobacter capsulatus* gene transfer agent (RcGTA) gene expression is a bistable stochastic process repressed by an extracellular calcium-binding RTX protein homologue. J. Bacteriol. *201*, e00430-19.

Dodd, I.B., and Egan, J.B. (1990). Improved detection of helix-turn-helix DNA-binding motifs in protein sequences. Nucleic Acids Res. *18*, 5019–5026.

Dove, S.L., and Hochschild, A. (1998). Conversion of the omega subunit of *Escherichia coli* RNA polymerase into a transcriptional activator or an activation target. Genes Dev. *12*, 745–754.

865 Esterman, E.S., Wolf, Y.I., Kogay, R., Koonin, E.V., and Zhaxybayeva, O. (2021).
866 Evolution of DNA packaging in gene transfer agents. Virus Evol. *7*, veab015.

867 Farrera-Calderon, R.G., Pallegar, P., Westbye, A.B., Wiesmann, C., Lang, A.S., and
868 Beatty, J.T. (2021). The CckA-ChpT-CtrA phosphorelay controlling *Rhodobacter*
869 *capsulatus* Gene Transfer Agent production is bidirectional and regulated by Ccclic di-
870 GMP. J. Bacteriol. *203*.

871 Feklístov, A., Sharon, B.D., Darst, S.A., and Gross, C.A. (2014). Bacterial sigma factors:
872 a historical, structural, and genomic perspective. Annu. Rev. Microbiol. *68*, 357–376.

873 Fogg, P.C.M. (2019). Identification and characterization of a direct activator of a gene
874 transfer agent. Nat. Commun. *10*, 595.

875 Fogg, M.J., and Wilkinson, A.J. (2008). Higher-throughput approaches to crystallization
876 and crystal structure determination. Biochem. Soc. Trans. *36*, 771–775.

877 Fogg, P.C.M., Westbye, A.B., and Beatty, J.T. (2012). One for all or all for one:
878 heterogeneous expression and host cell lysis are key to gene transfer agent activity in
879 *Rhodobacter capsulatus*. PLoS ONE *7*, e43772.

880 Freese, H.M., Sikorski, J., Bunk, B., Scheuner, C., Meier-Kolthoff, J.P., Spröer, C.,
881 Gram, L., and Overmann, J. (2017). Trajectories and drivers of genome evolution in
882 surface-associated marine *Phaeobacter*. Genome Biol. Evol. *9*, 3297–3311.

883 Fujikawa, N., Kurumizaka, H., Nureki, O., Terada, T., Shirouzu, M., Katayama, T., and
884 Yokoyama, S. (2003). Structural basis of replication origin recognition by the DnaA
885 protein. Nucleic Acids Res. *31*, 2077–2086.

886 Gabler, F., Nam, S.-Z., Till, S., Mirdita, M., Steinegger, M., Söding, J., Lupas, A.N., and
887 Alva, V. (2020). Protein sequence analysis using the MPI bioinformatics toolkit. Curr.
888 Protoc. Bioinformatics *72*, e108.

889 Geertz, M., Travers, A., Mehandziska, S., Sobetzko, P., Chandra-Janga, S.,
890 Shimamoto, N., and Muskhelishvili, G. (2011). Structural coupling between RNA
891 polymerase composition and DNA supercoiling in coordinating transcription: a global
892 role for the omega subunit? MBio *2*.

893 Glyde, R., Ye, F., Jovanovic, M., Kotta-Loizou, I., Buck, M., and Zhang, X. (2018).
894 Structures of bacterial RNA polymerase complexes reveal the mechanism of DNA
895 loading and transcription initiation. Mol. Cell *70*, 1111-1120.e3.

896 Goddard, T.D., Huang, C.C., Meng, E.C., Pettersen, E.F., Couch, G.S., Morris, J.H.,
897 and Ferrin, T.E. (2018). UCSF ChimeraX: Meeting modern challenges in visualization
898 and analysis. Protein Sci. *27*, 14–25.

899 Griffith, K.L., Shah, I.M., Myers, T.E., O'Neill, M.C., and Wolf, R.E. (2002). Evidence for
900 "pre-recruitment" as a new mechanism of transcription activation in *Escherichia coli*: the
901 large excess of SoxS binding sites per cell relative to the number of SoxS molecules per
902 cell. Biochem. Biophys. Res. Commun. *291*, 979–986.

903 Gunnelius, L., Hakkila, K., Kurkela, J., Wada, H., Tyystjärvi, E., and Tyystjärvi, T.
904 (2014). The omega subunit of the RNA polymerase core directs transcription efficiency
905 in cyanobacteria. Nucleic Acids Res. *42*, 4606–4614.

906 He, Y. (2012). Analyses of *Brucella* pathogenesis, host immunity, and vaccine targets
907 using systems biology and bioinformatics. Front. Cell. Infect. Microbiol. *2*, 2.

908 Hynes, A.P., Mercer, R.G., Watton, D.E., Buckley, C.B., and Lang, A.S. (2012). DNA
909 packaging bias and differential expression of gene transfer agent genes within a
910 population during production and release of the *Rhodobacter capsulatus* gene transfer
911 agent, RcGTA. Mol. Microbiol. *85*, 314–325.

912 Hynes, A.P., Shakya, M., Mercer, R.G., Grüll, M.P., Bown, L., Davidson, F., Steffen, E.,
913 Matchem, H., Peach, M.E., Berger, T., *et al.* (2016). Functional and evolutionary
914 characterization of a gene transfer agent's multilocus "genome". Mol. Biol. Evol. *33*,
915 2530–2543.

916 Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O.,
917 Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., *et al.* (2021). Highly accurate
918 protein structure prediction with AlphaFold. Nature *596*, 583–589.

919 Kaczmarczyk, A., Vorholt, J.A., and Francez-Charlot, A. (2013). Cumate-inducible gene
920 expression system for sphingomonads and other Alphaproteobacteria. Appl. Environ.
921 Microbiol. *79*, 6795–6802.

922 Karimova, G., Pidoux, J., Ullmann, A., and Ladant, D. (1998). A bacterial two-hybrid
923 system based on a reconstituted signal transduction pathway. Proc Natl Acad Sci USA
924 *95*, 5752–5756.

925 Kogay, R., Neely, T.B., Birnbaum, D.P., Hankel, C.R., Shakya, M., and Zhaxybayeva,
926 O. (2019). Machine-learning classification suggests that many alphaproteobacterial
927 prophages may instead be Gene Transfer Agents. Genome Biol. Evol. *11*, 2941–2953.

928 Kogay, R., Wolf, Y.I., Koonin, E.V., and Zhaxybayeva, O. (2020). Selection for reducing
929 energy cost of protein production drives the GC content and amino acid composition
930 bias in gene transfer agents. MBio *11*, e01206-20.

931 Koppenhöfer, S., Wang, H., Scharfe, M., Kaever, V., Wagner-Döbler, I., and Tomasch,
932 J. (2019). Integrated transcriptional regulatory network of quorum sensing, replication
933 control, and SOS response in *Dinoroseobacter shibae*. Front. Microbiol. *10*, 803.

934 Kuchinski, K.S., Brimacombe, C.A., Westbye, A.B., Ding, H., and Beatty, J.T. (2016).
935 The SOS response master regulator LexA regulates the Gene Transfer Agent of
936 *Rhodobacter capsulatus* and represses transcription of the signal transduction protein
937 CckA. J. Bacteriol. *198*, 1137–1148.

938 Kurkela, J., Fredman, J., Salminen, T.A., and Tyystjärvi, T. (2021). Revealing secrets of
939 the enigmatic omega subunit of bacterial RNA polymerase. Mol. Microbiol. *115*, 1–11.

940 Lane, W.J., and Darst, S.A. (2006). The structural basis for promoter -35 element
941 recognition by the group IV sigma factors. PLoS Biol. *4*, e269.

942 Lang, A.S., and Beatty, J.T. (2000). Genetic analysis of a bacterial genetic exchange
943 element: the gene transfer agent of *Rhodobacter capsulatus*. Proc Natl Acad Sci USA
944 *97*, 859–864.

945 Lang, A.S., Zhaxybayeva, O., and Beatty, J.T. (2012). Gene transfer agents: phage-like
946 elements of genetic exchange. Nat. Rev. Microbiol. *10*, 472–482.

947 Leung, M., and Beatty, J. (2013). *Rhodobacter capsulatus* Gene Transfer Agent
948 transduction assay. Bio Protoc *3*.

949 Leung, M.M., Brimacombe, C.A., Spiegelman, G.B., and Beatty, J.T. (2012). The GtaR
950 protein negatively regulates transcription of the *gtaRI* operon and modulates gene
951 transfer agent (RcGTA) expression in *Rhodobacter capsulatus*. Mol. Microbiol. *83*, 759–
952 774.

953 Lin, W., Mandal, S., Degen, D., Cho, M.S., Feng, Y., Das, K., and Ebright, R.H. (2019).
954 Structural basis of ECF-σ-factor-dependent transcription initiation. Nat. Commun. *10*,
955 710.

956 Livak, K.J., and Schmittgen, T.D. (2001). Analysis of relative gene expression data
957 using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. Methods *25*,
958 402–408.

959 Li, L., Fang, C., Zhuang, N., Wang, T., and Zhang, Y. (2019). Structural basis for
960 transcription initiation by bacterial ECF σ factors. Nat. Commun. *10*, 1153.

961 Maniatis, T., Fritsch, E.F., and Sambrook, J. (1982). Molecular cloning: A laboratory
962 manual. Cold Spring Harbor laboratory press.

963 Mao, C., Zhu, Y., Lu, P., Feng, L., Chen, S., and Hu, Y. (2018). Association of ω with
964 the C-Terminal region of the β' Subunit Is essential for assembly of RNA polymerase in
965 *Mycobacterium tuberculosis*. J. Bacteriol. *200*.

966 Martin, R.G., Gillette, W.K., Martin, N.I., and Rosner, J.L. (2002). Complex formation
967 between activator and RNA polymerase as the basis for transcriptional activation by
968 MarA and SoxS in *Escherichia coli*. Mol. Microbiol. *43*, 355–370.

969 Mathew, R., and Chatterji, D. (2006). The evolving story of the omega subunit of
970 bacterial RNA polymerase. Trends Microbiol. *14*, 450–455.

971 McDaniel, L.D., Young, E., Delaney, J., Ruhnau, F., Ritchie, K.B., and Paul, J.H. (2010).
972 High frequency of horizontal gene transfer in the oceans. Science *330*, 50.

973 Mercer, R.G., Callister, S.J., Lipton, M.S., Pasa-Tolic, L., Strnad, H., Paces, V., Beatty,
974 J.T., and Lang, A.S. (2010). Loss of the response regulator CtrA causes pleiotropic
975 effects on gene expression but does not affect growth phase regulation in *Rhodobacter*
976 *capsulatus*. J. Bacteriol. *192*, 2701–2710.

977 Mercer, R.G., Quinlan, M., Rose, A.R., Noll, S., Beatty, J.T., and Lang, A.S. (2012).
978 Regulatory systems controlling motility and gene transfer agent production and release
979 in *Rhodobacter capsulatus*. FEMS Microbiol. Lett. *331*, 53–62.

980 Motro, Y., La, T., Bellgard, M.I., Dunn, D.S., Phillips, N.D., and Hampson, D.J. (2009).
981 Identification of genes associated with prophage-like gene transfer agents in the
982 pathogenic intestinal spirochaetes *Brachyspira hyodysenteriae*, *Brachyspira pilosicoli*
983 and *Brachyspira intermedia*. Vet. Microbiol. *134*, 340–345.

984 Narasimhan, G., Bu, C., Gao, Y., Wang, X., Xu, N., and Mathee, K. (2002). Mining
985 protein sequences for motifs. J. Comput. Biol. *9*, 707–720.

986 Paget, M.S. (2015). Bacterial sigma factors and anti-sigma factors: structure, function
987 and distribution. Biomolecules *5*, 1245–1265.

988 Pallegar, P., Peña-Castillo, L., Langille, E., Gomelsky, M., and Lang, A.S. (2020a).
989 Cyclic di-GMP-mediated regulation of gene transfer and motility in *Rhodobacter*
990 *capsulatus*. J. Bacteriol. *202*.

991 Pallegar, P., Canuti, M., Langille, E., Peña-Castillo, L., and Lang, A.S. (2020b). A two-
992 component system acquired by horizontal gene transfer modulates gene transfer and
993 motility via cyclic dimeric GMP. J. Mol. Biol. *432*, 4840–4855.

994 Price, M.N., Wetmore, K.M., Waters, R.J., Callaghan, M., Ray, J., Liu, H., Kuehl, J.V.,
995 Melnyk, R.A., Lamson, J.S., Suh, Y., *et al.* (2018). Mutant phenotypes for thousands of
996 bacterial genes of unknown function. Nature *557*, 503–509.

997 Québatte, M., and Dehio, C. (2019). *Bartonella* gene transfer agent: Evolution, function,
998 and proposed role in host adaptation. Cell. Microbiol. *21*, e13068.

999 Québatte, M., Christen, M., Harms, A., Körner, J., Christen, B., and Dehio, C. (2017).
1000 Gene Transfer Agent promotes evolvability within the fittest subpopulation of a bacterial
1001 pathogen. Cell Syst. *4*, 611-621.e6.

1002 Ross, W., Vrentas, C.E., Sanchez-Vazquez, P., Gaal, T., and Gourse, R.L. (2013). The
1003 magic spot: a ppGpp binding site on *E. coli* RNA polymerase responsible for regulation
1004 of transcription initiation. Mol. Cell *50*, 420–429.

1005 Salmon-Divon, M., Zahavi, T., and Kornspan, D. (2019). Transcriptomic analysis of the
1006 *Brucella melitensis* Rev.1 vaccine strain in an acidic environment: Insights Into virulence
1007 attenuation. Front. Microbiol. *10*, 250.

1008 Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T.,
1009 Preibisch, S., Rueden, C., Saalfeld, S., Schmid, B., *et al.* (2012). Fiji: an open-source
1010 platform for biological-image analysis. Nat. Methods *9*, 676–682.

1011 Scholz, H.C., Pfeffer, M., Witte, A., Neubauer, H., Al Dahouk, S., Wernery, U., and
1012 Tomaso, H. (2008). Specific detection and differentiation of *Ochrobactrum anthropi*,
1013 *Ochrobactrum intermedium* and *Brucella* spp. by a multi-primer PCR that targets the
1014 recA gene. J. Med. Microbiol. *57*, 64–71.

1015 Shakya, M., Soucy, S.M., and Zhaxybayeva, O. (2017). Insights into origin and
1016 evolution of α-proteobacterial gene transfer agents. Virus Evol. *3*, vex036.

1017 Sherlock, D., Leong, J.X., and Fogg, P.C.M. (2019). Identification of the first gene
1018 transfer agent (GTA) small terminase in *Rhodobacter capsulatus*, its role in GTA
1019 production and packaging of DNA. J. Virol. *93*, e01328-19.

1020 Shimada, T., Yamazaki, Y., Tanaka, K., and Ishihama, A. (2014). The whole set of
1021 constitutive promoters recognized by RNA polymerase RpoD holoenzyme of
1022 *Escherichia coli*. PLoS ONE *9*, e90447.

1023 Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R.,
1024 McWilliam, H., Remmert, M., Söding, J., *et al.* (2011). Fast, scalable generation of high-
1025 quality protein multiple sequence alignments using Clustal Omega. Mol. Syst. Biol. *7*,
1026 539.

1027 Solovyev, V., A Salamov (2011) Automatic annotation of microbial genomes and
1028 metagenomic equences. In Metagenomics and its Applications in Agriculture,
1029 Biomedicine and Environmental Studies (Ed. R.W. Li), Nova Science Publishers, p.61-
1030 78.

1031 Tamarit, D., Neuvonen, M.-M., Engel, P., Guy, L., and Andersson, S.G.E. (2018). Origin
1032 and evolution of the *Bartonella* gene transfer agent. Mol. Biol. Evol. *35*, 451–464.

1033 Tomasch, J., Wang, H., Hall, A.T.K., Patzelt, D., Preusse, M., Petersen, J., Brinkmann,
1034 H., Bunk, B., Bhuju, S., Jarek, M., *et al.* (2018). Packaging of *Dinoroseobacter shibae*
1035 DNA into Gene Transfer Agent particles is not random. Genome Biol. Evol. *10*, 359–
1036 369.

1037 Vassylyev, D.G., Sekine, S., Laptenko, O., Lee, J., Vassylyeva, M.N., Borukhov, S., and
1038 Yokoyama, S. (2002). Crystal structure of a bacterial RNA polymerase holoenzyme at
1039 2.6 A resolution. Nature *417*, 712–719.

1040 Wall, J.D., Weaver, P.F., and Gest, H. (1975). Gene transfer agents, bacteriophages,
1041 and bacteriocins of *Rhodopseudomonas capsulata*. Arch. Microbiol. *105*, 217–224.

1042 Wang, H., Ziesche, L., Frank, O., Michael, V., Martin, M., Petersen, J., Schulz, S.,
1043 Wagner-Döbler, I., and Tomasch, J. (2014). The CtrA phosphorelay integrates
1044 differentiation and communication in the marine alphaproteobacterium *Dinoroseobacter*
1045 *shibae*. BMC Genomics *15*, 130.

1046 Waterhouse, A.M., Procter, J.B., Martin, D.M.A., Clamp, M., and Barton, G.J. (2009).
1047 Jalview Version 2--a multiple sequence alignment editor and analysis workbench.
1048 Bioinformatics *25*, 1189–1191.

1049 Weiss, A., Moore, B.D., Tremblay, M.H.J., Chaput, D., Kremer, A., and Shaw, L.N.
1050 (2017). The ω subunit governs RNA polymerase stability and transcriptional specificity
1051 in *Staphylococcus aureus*. J. Bacteriol. *199*.

1052 Westbye, A.B., Leung, M.M., Florizone, S.M., Taylor, T.A., Johnson, J.A., Fogg, P.C.,
1053 and Beatty, J.T. (2013). Phosphate concentration and the putative sensor kinase protein
1054 CckA modulate cell lysis and release of the *Rhodobacter capsulatus* gene transfer
1055 agent. J. Bacteriol. *195*, 5025–5040.

1056 Westbye, A.B., O'Neill, Z., Schellenberg-Beaver, T., and Beatty, J.T. (2017). The
1057 *Rhodobacter capsulatus* gene transfer agent is induced by nutrient depletion and the
1058 RNAP omega subunit. Microbiology *163*, 1355–1363.

1059 Westbye, A.B., Kater, L., Wiesmann, C., Ding, H., Yip, C.K., and Beatty, J.T. (2018).
1060 The protease ClpXP and the PAS domain protein DivL regulate CtrA and Gene Transfer
1061 Agent production in *Rhodobacter capsulatus*. Appl. Environ. Microbiol. *84*.

1062 Wiethaus, J., Wirsing, A., Narberhaus, F., and Masepohl, B. (2006). Overlapping and
1063 specialized functions of the molybdenum-dependent regulators MopA and MopB in
1064 *Rhodobacter capsulatus*. J. Bacteriol. *188*, 8441–8451.

1065 Wiethaus, J., Schubert, B., Pfänder, Y., Narberhaus, F., and Masepohl, B. (2008). The
1066 GntR-like regulator TauR activates expression of taurine utilization genes in
1067 *Rhodobacter capsulatus*. J. Bacteriol. *190*, 487–493.

1068 Yamamoto, K., Yamanaka, Y., Shimada, T., Sarkar, P., Yoshida, M., Bhardwaj, N.,
1069 Watanabe, H., Taira, Y., Chatterji, D., and Ishihama, A. (2018). Altered distribution of
1070 RNA polymerase lacking the Omega subunit within the prophages along the
1071 *Escherichia coli* K-12 genome. MSystems *3*.

1072 Zimmermann, L., Stephens, A., Nam, S.-Z., Rau, D., Kübler, J., Lozajic, M., Gabler, F.,
1073 Söding, J., Lupas, A.N., and Alva, V. (2018). A completely reimplemented MPI
1074 bioinformatics toolkit with a new HHpred server at its core. J. Mol. Biol. *430*, 2237–2243.

Figure 1

Figure 2

A

β-galactosidase activity (Miller Units)

T18-RNAP subunits

B

Elution

GafA
Rpo-ω

← GafA

← Rpo-ω

A Figure 3

B

C

D

A
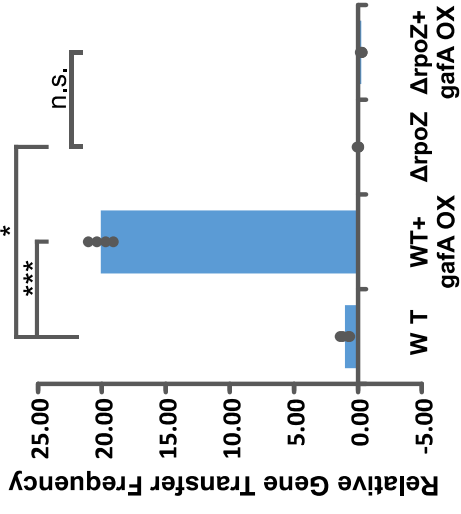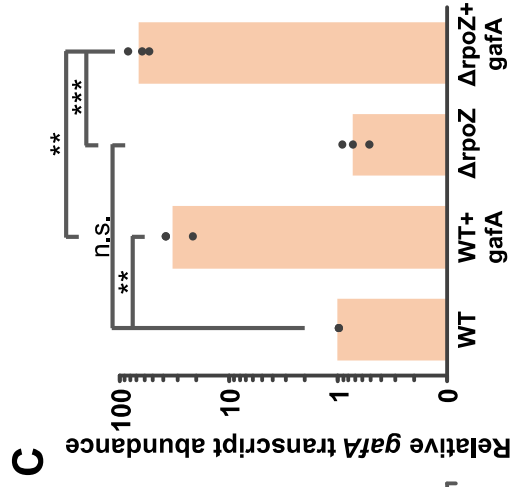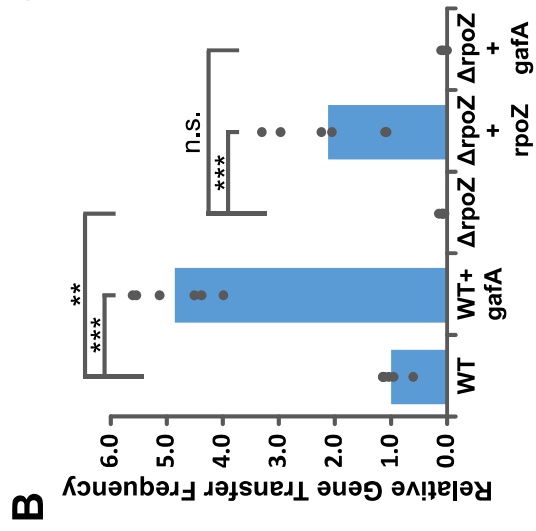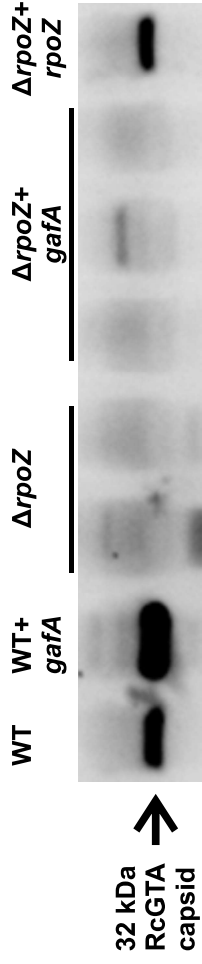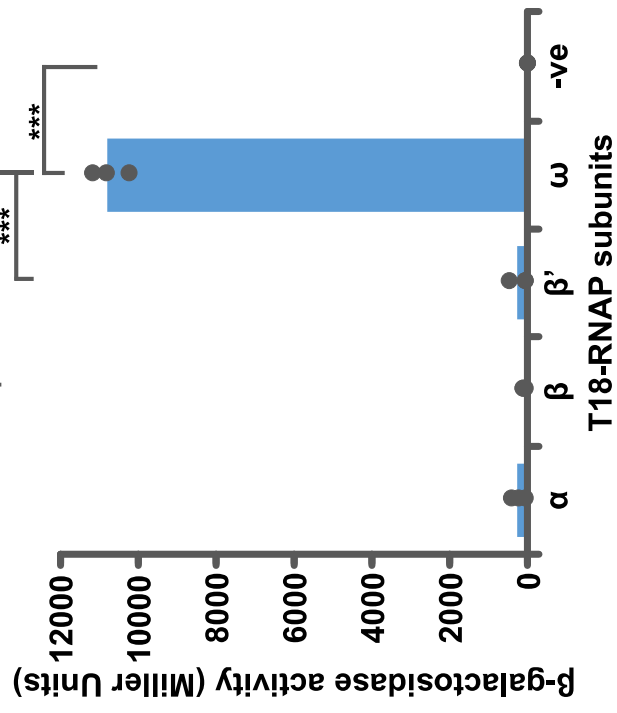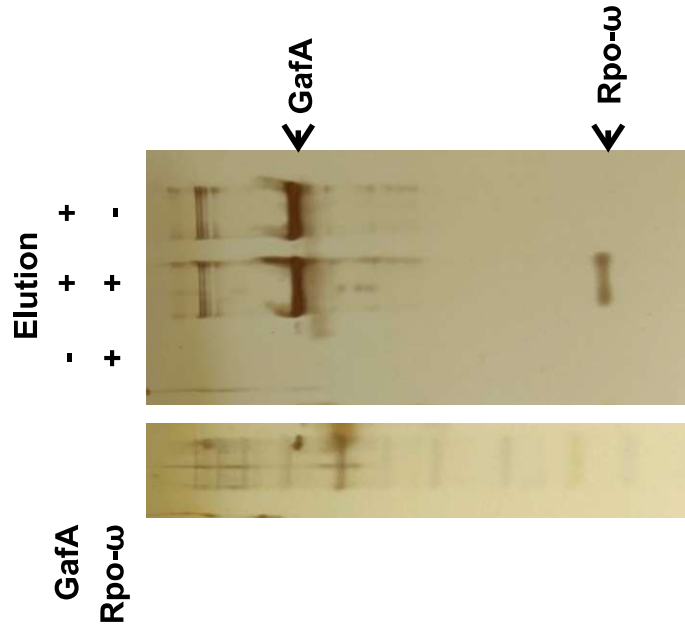
>GafA_1-382
MKTMQDRESLPDWLPDHARLYLRHVEEGVPIRQLARAEG
CHASTILRRVRRIEQRRDDPLVDEALTRLGRFAAAASAA
PPREDDPAMTAPIRPTAPQACPEAAEDPDSPDIATLSRE
GRRVLRRLAEPGALLIIAPDMEKAVVLRGTVRTAVVARE
VAQGFALNGWILVQHSGRVTSYELSATGRAALKRLLEAE
ALTAGRDPATAADNPHADRHRDWGERTVNEGQGRVTRMR
MNLAESPLGVLARRRDSDGRPFLSPDLVAAGERLREDFE
LAQMGPRVAQNWERFMTGGARGQYRPELGHGGPGGSDRA
RERVAAALCDLGPGLGDMVLRCCCFLEGLETAEKRMGWS
ARSGKIVLRIALMRLKRHYDETYGGAAPLIG
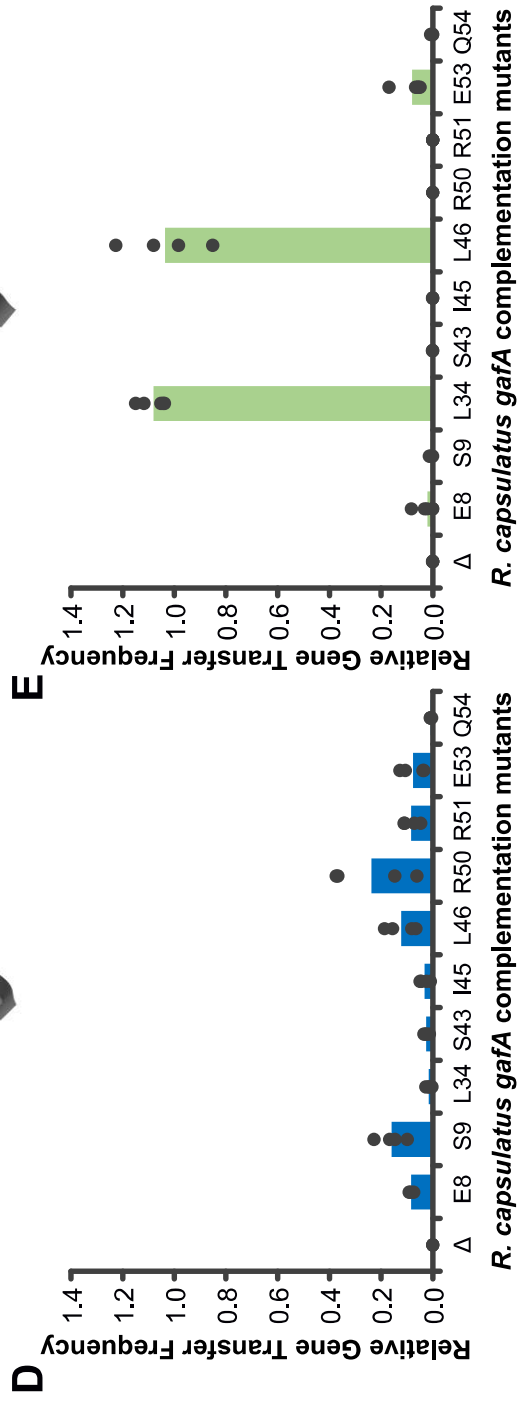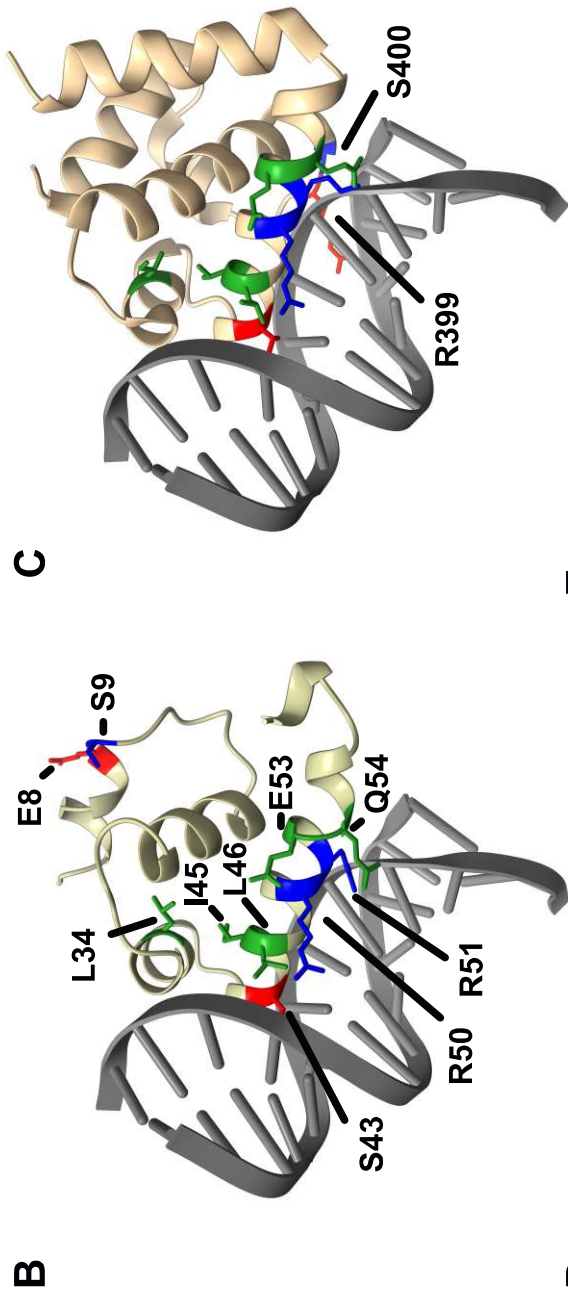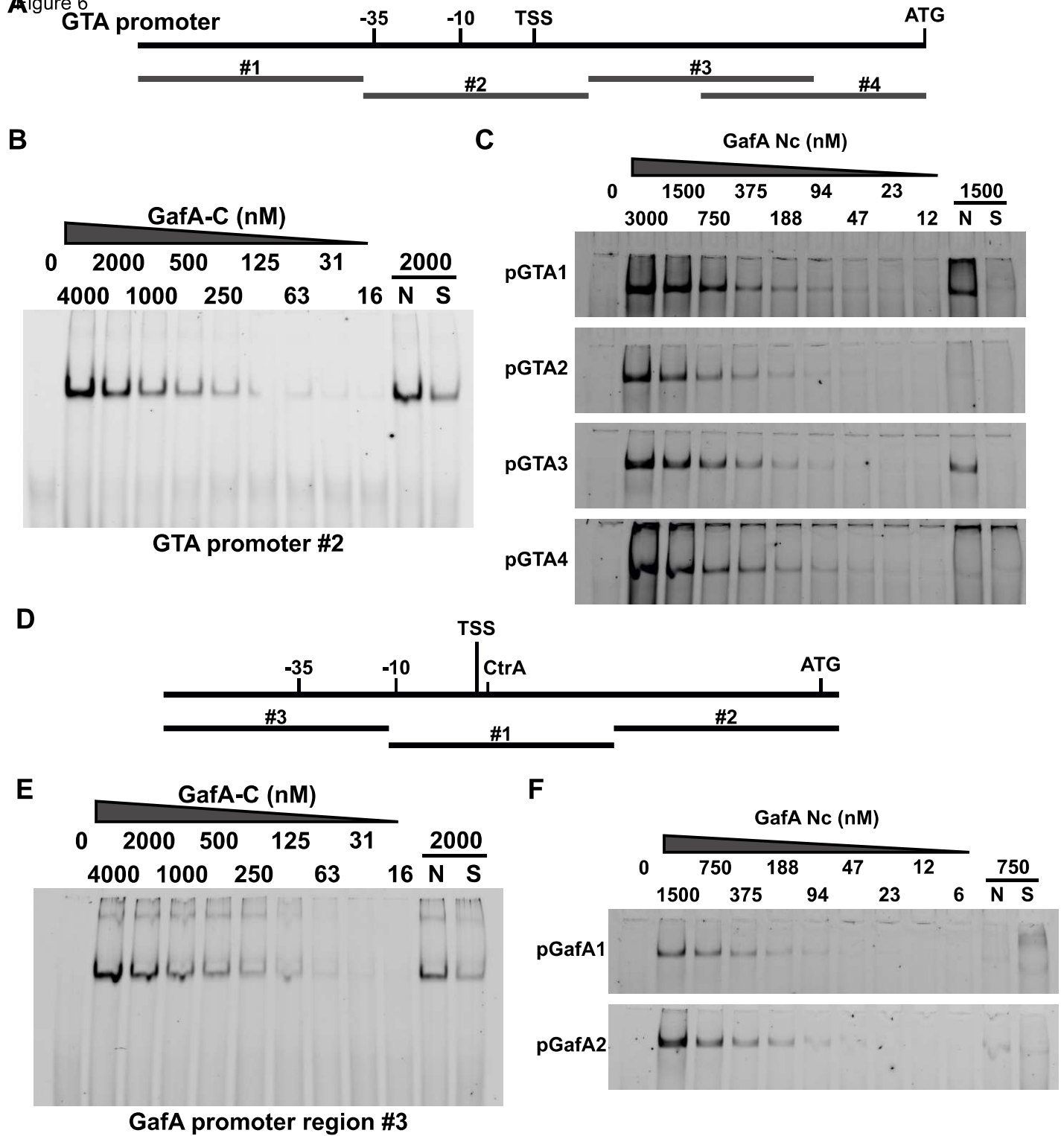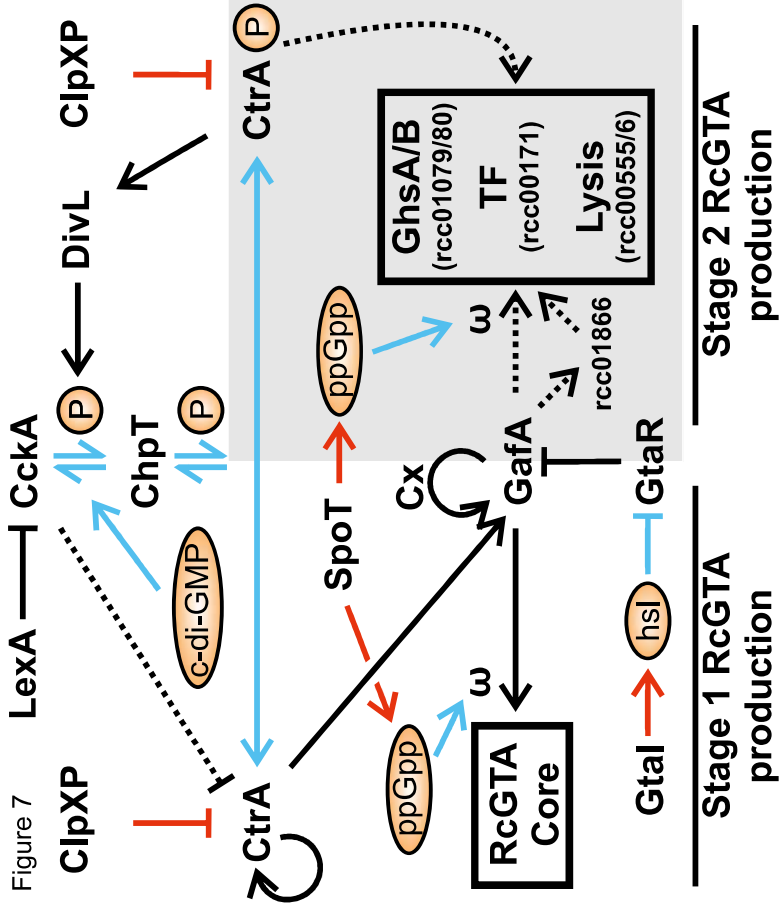
B

Cx

CenN/Cen2

DBD

N   C

Nc

DBD

C

Relative β-galactosidase Activity

3.0 2.5 2.0 1.5 1.0 0.5 0.0

Full Con N Cen2 CenN C Cx

T25-GafA Fragments

n.s.

***

D

Input   Elution

GafA-Cx   -   +   -   +   +
Rpo-ω   +   -   -   +   +

Figure 4

Figure 5

Figure 6

A GTA promoter

B GafA-C (nM)
GTA promoter #2

C GafA Nc (nM)
pGTA1
pGTA2
pGTA3
pGTA4

D

E GafA-C (nM)
GafA promoter region #3

F GafA Nc (nM)
pGafA1
pGafA2

Figure 7

Figure S1. Alignment of the *R. capsulatus* GafA C-terminal extended domain with Hyphomicrobiales counterparts. Related to Figure 3. The top four hits against fully assembled Hyphomicrobiales genomes were chosen from separate BLASTp and PSI-BLAST sequence similarity searches with an *R. capsulatus* GafA query. Conservation is indicated with the Jalview percentage identity colour scheme. The predicted C-terminal DNA binding domain is boxed and annotated to highlight increased sequence conservation. The open box indicated the beginning of the C-terminal concise constructs.
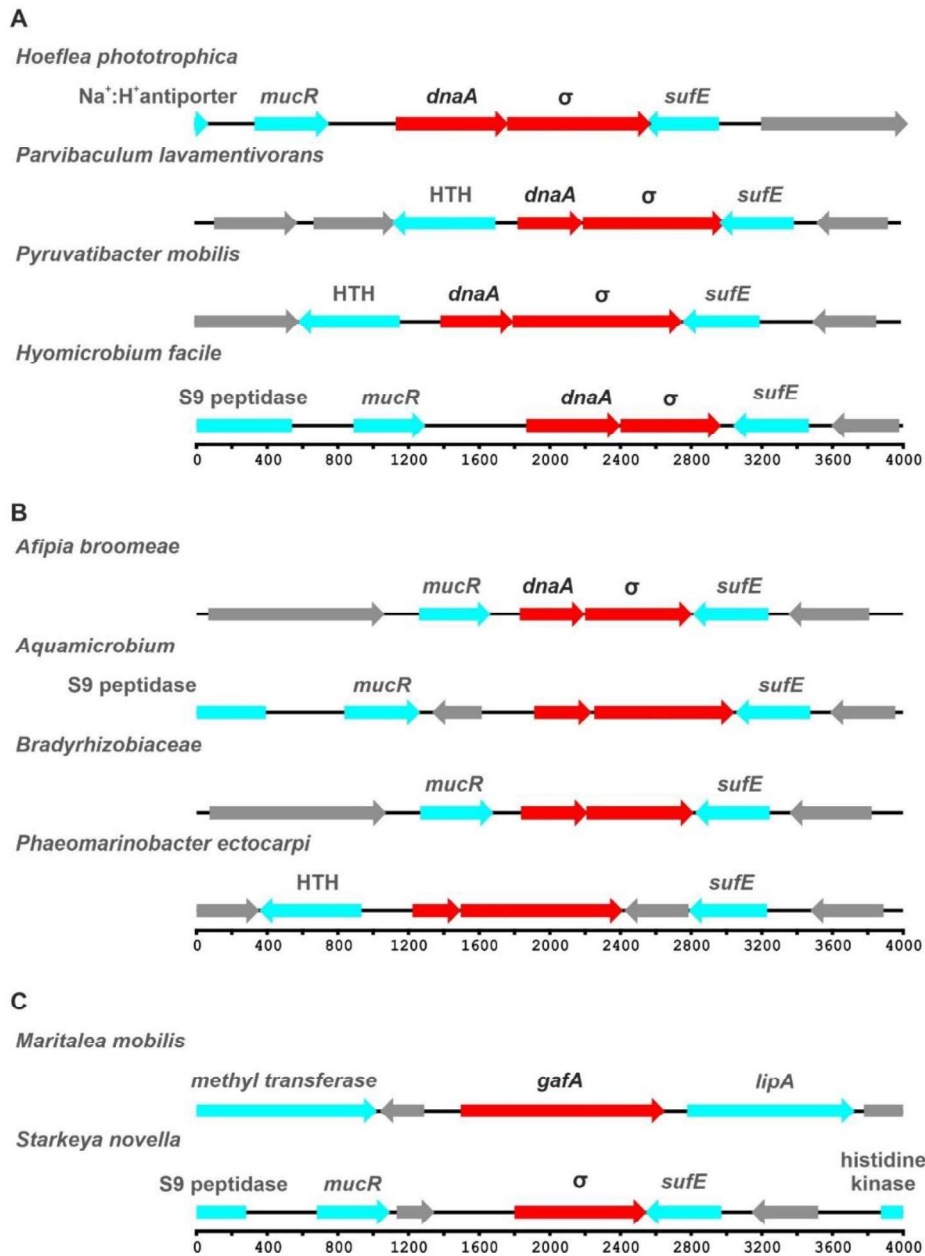
Figure S2. Synteny plots for Hyphomicrobiales GafA homologues. Related to Figure 3. The top four **A.** BLASTp and **B.** PSI-BLAST hits for *R. capsulatus* GafA against fully assembled Hyphomicrobiales genomes. Sequence matches mainly occurred for the GafA C-terminal region only with genes annotated as DUF6456 domain-containing proteins. The matched Hyphomicrobiales genes are annotated here using the HHPRED prediction of a Sigma factor-like domain (σ) and the ORF is coloured red. The upstream dnaA-like ORF is also coloured red. Flanking genes with predicted function are cyan, hypothetical proteins of unknown function are grey. **C.** Two exceptions are shown where either a full-length match was obtained but with Rhodoabcterales-like synteny (*Maritalea*) or the *dnaA* gene was absent with otherwise Hyphomicrobiale-like synteny (*Starkeya*). Scale bars are provided below each panel in bases.
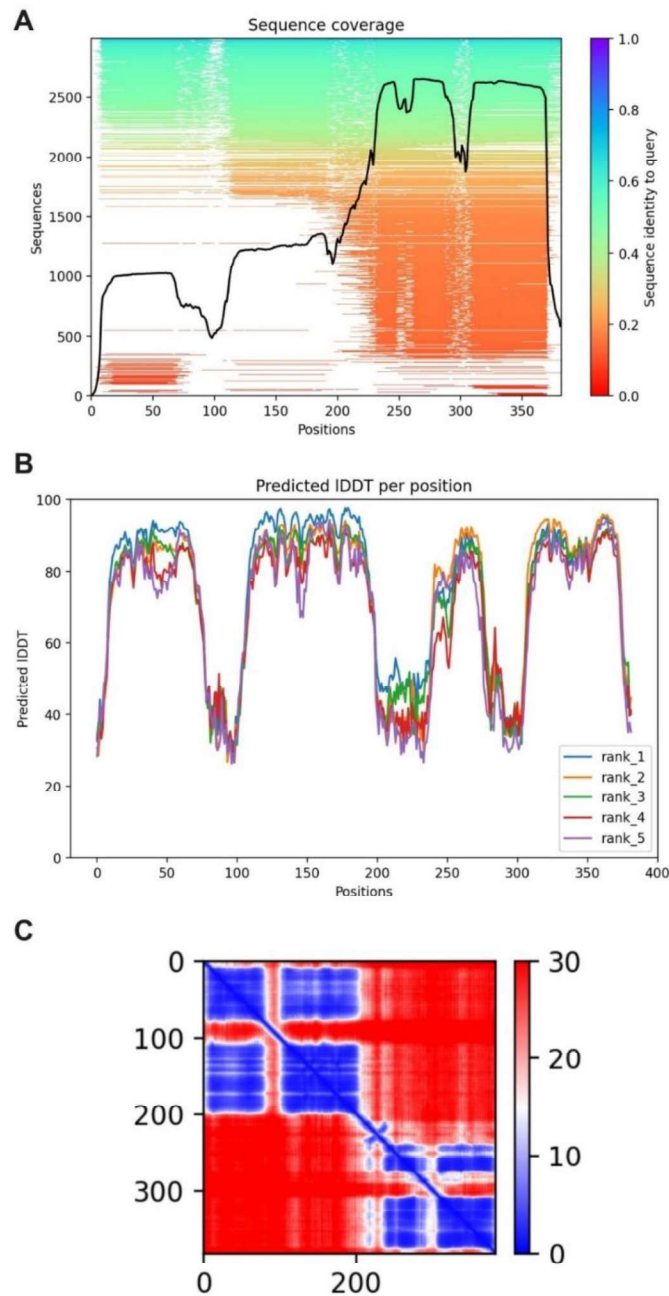
**Figure S3. Confidence outputs for *R. capsulatus* GafA structure prediction. Related to Figure 3. A.** The jackhmmer method was used on the Alphafold server to align GafA to related proteins and the multiple sequence alignment coverage plot is shown. Aligned sequence coverage is depicted as a line chart and sequence identity is colour coded as shown in the legend. **B.** AlphaFold output plot showing the predicted local Distance Difference Test score (pIDDT) confidence metric. Amino acid positions are shown on the X-axis. **C.** Predicted Aligned Error for each amino acid position labelled on the X and Y-axes. Error is shown on a scale of 0-30, and colour coded as shown in the legend. Clear drop-offs in model confidence can be seen between predicted domains, but each domain is has strong scores typically >80.
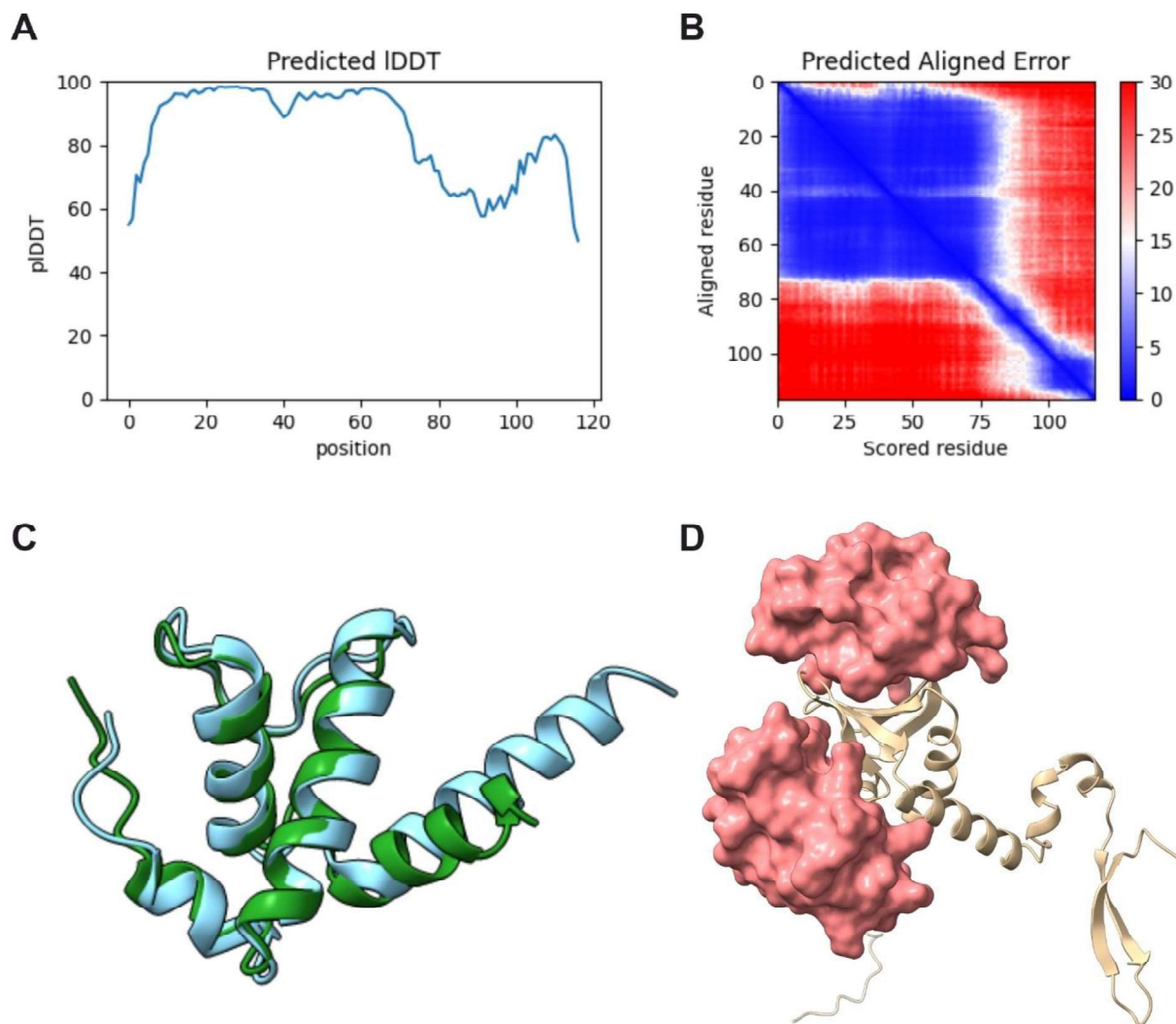
Figure S4. Predicted structure of *R. capsulatus* Rpo-ω protein and its interaction with GafA. Related to Figure 3. A & B. AlphaFold output plots showing the predicted local Distance Difference Test score (pIDDT) confidence metric and Predicted Aligned Error for each amino acid position. A clear drop-off in model confidence, domain packing and broader topology is observed from approximately residue 70 onwards. **C.** AlphaFold predicted *R. capsulatus* Rpo-ω structure trimmed to residues 1-71 (green) and overlaid with *E. coli* Rpo-ω, PDB: 6ALF (pale blue). **D.** LZerD protein docking predictions for GafA-CenN and Rpo-ω$^{1-71}$. The two Rpo-ω surface structures shown are representatives of the two centroid clusters that comprise the top ten interaction models. The upper location in contact with the β-sheet was favoured by 6 out of 10 models including the top ranked (rank sum = 47).

4
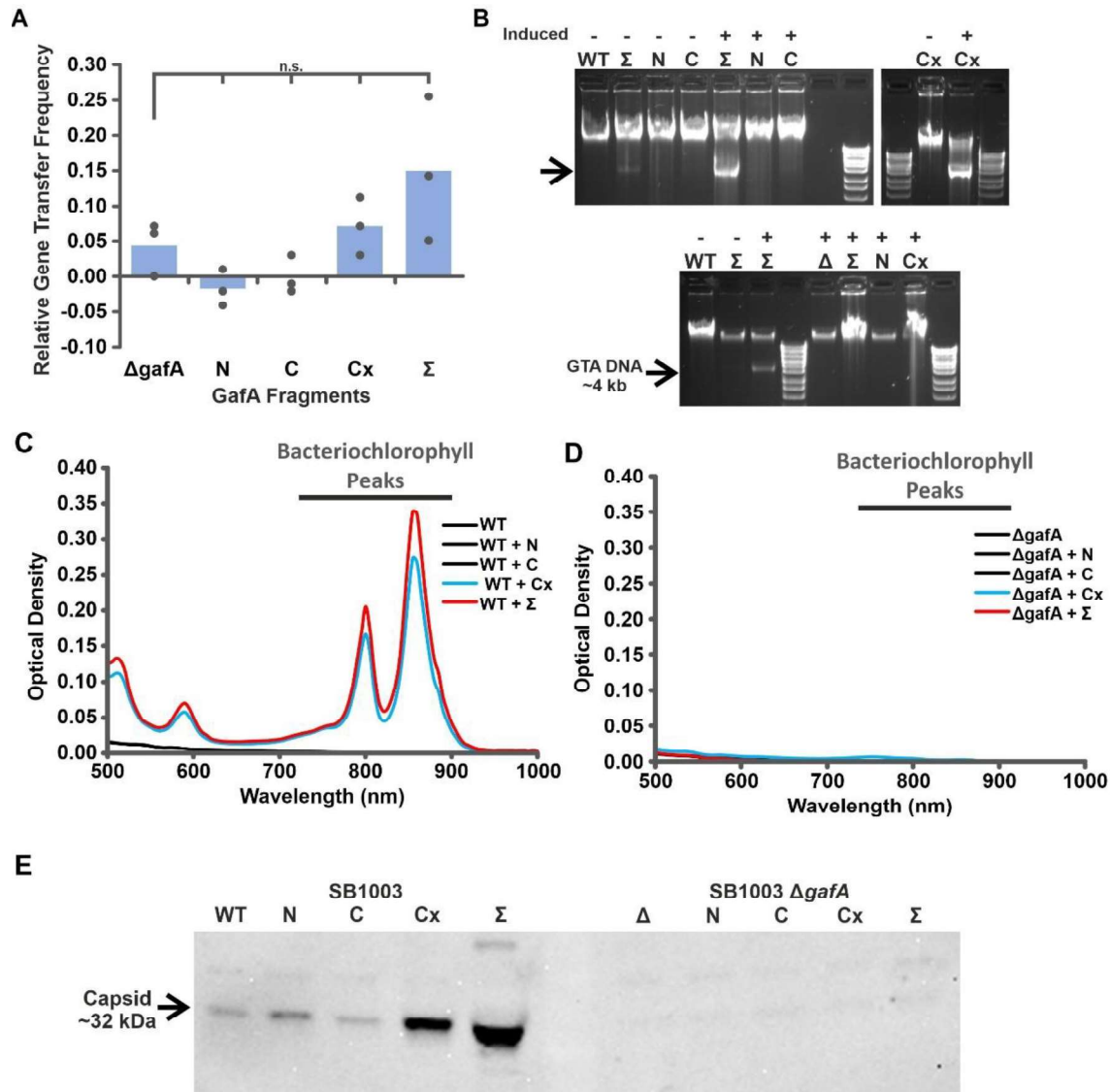
Figure S5. RcGTA production phenotypes after *in trans* expression of GafA full length and truncated proteins. Related to Figure 4. In all panels, SB1003 wild-type and a Δ*gafA* derivative were complemented with empty pQF vector (WT or Δ) or pQF containing truncated *gafA* genes as indicated (*gafA*-N, *gafA*-C, *gafA*-Cx, *gafA*-Σ). **A.** Chart of the frequency of rifampicin gene transfer from *R. capsulatus* SB1003 Δ*gafA* donor strains complemented *in trans* with the indicated pQF vectors, N = 3. **B.** Total intracellular DNA content showing the presence or absence of characteristic 4 kb RcGTA DNA. **C.** Mean absorbance trace of *R. capsulatus* SB1003 supernatant or **D.** SB1003 Δ*gafA* supernatants in the 500-1000 nm wavelength range. Complementation *in trans* the pQF plasmid containing full-length *gafA* is represented by a red line, with *gafA-Cx* is represented by a cyan line and all other constructs (pQF-empty, *gafA*-N and *gafA*-C) are shown in black. N=6 except Δ*gafA* + Cx N=4. Distinctive bacteriochlorophyll peaks indicating cells lysis are annotated. **E.** Representative western blot of concentrated supernatant from the indicated *R. capsulatus* strains using an α-RcGTA capsid antibody. See also Data S1.

A
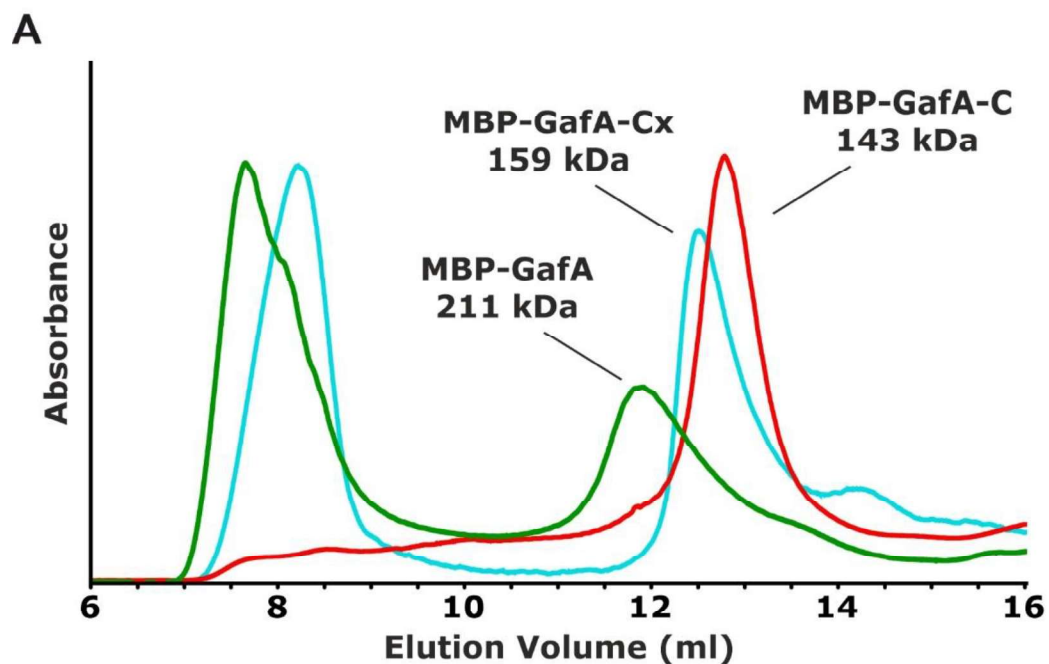
B

Protein	Elution Peak (ml)	Estimated MW (kDa)	Monomer Size (kDa)	Ratio
MBP-GafA-Cx	12.54	158,779	75,806	2.1
MBP-GafA-C	12.78	142,537	61,119	2.3
MBP-GafA	11.91	210,774	85,159	2.5

Figure S6. Analytical gel filtration of GafA proteins. Related to Figure 6. A. Representative traces showing absorbance of GafA (green), GafA-Cx (cyan) and GafA-C (red) at 280 nm versus elution time from the column. Absorbance values are omitted on the Y-axis because the traces are scaled differently to improve comparability. **B.** Summary table of values plotted in part A, the estimated MW of the protein peaks, the calculated MW of each monomer and the ratio of observed MW to that of the monomer.
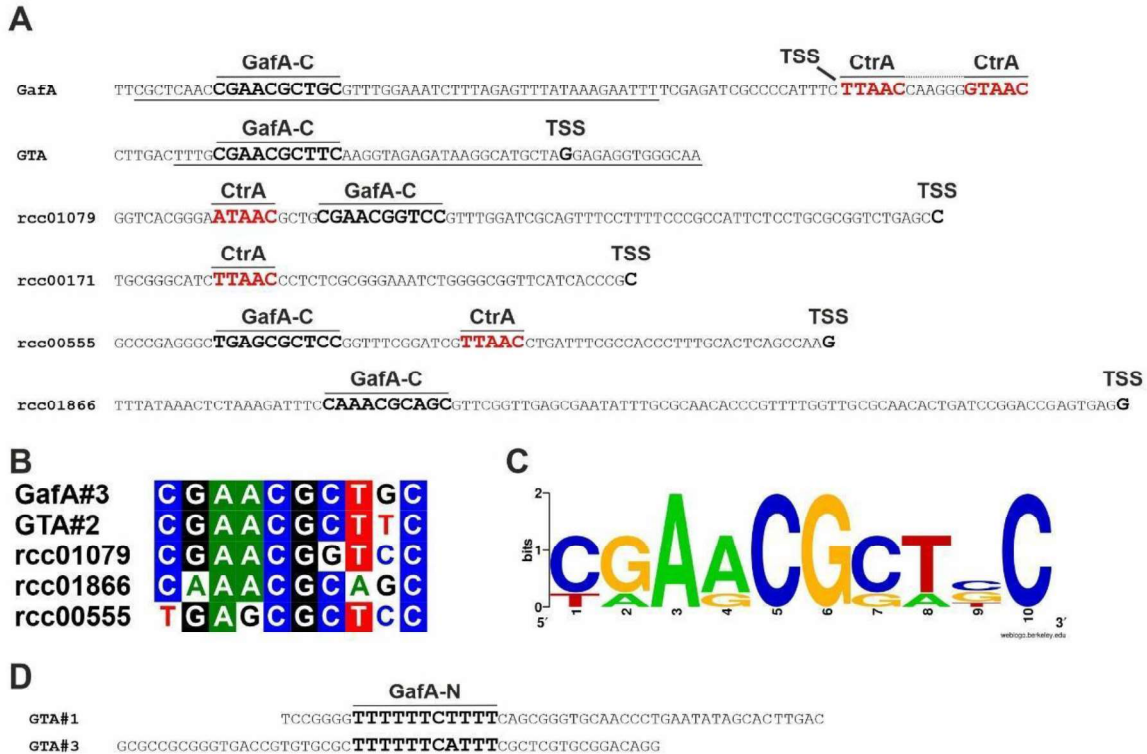
Figure S7. Predicted binding sites for GafA N/C-terminal DNA binding domains. Related to Figure 6.
A. Schematic of RcGTA related promoters. Transcription start sites (TSS) were estimated based on published RNAseq data. Predicted CtrA binding sites/half-sites are highlighted in bold red and annotated, predicted GafA C-terminal (GafA-C) DNA binding sites are highlighted in bold black and annotated. Underlined sequence indicates the region used for EMSA band shift assays. The five predicted GafA-C binding sites are depicted in **B.** an alignment and **C.** a Logo plot. **D.** The two oligo sequences that were specifically bound by the GafA N-terminal DNA binding domain (GafA-N) are shown with the putative binding site aligned, emboldened and annotated.