Genome-wide protein–DNA interaction site mapping using a double strand DNA-specific cytosine deaminase

Larry A. Gallagher[1], Elena Velazquez[1,2], S. Brook Peterson[1], James C. Charity[3], FoSheng Hsu[1], Matthew C. Radey[1], Michael J. Gebhardt[3], Marcos H. de Moraes[1], Kelsi M. Penewit[4], Jennifer Kim[1], Pia A. Andrade[1], Thomas LaFramboise[5], Stephen J. Salipante[4], Victor de Lorenzo[2], Paul A. Wiggins[1,6,7], Simon L. Dove[3,*] and Joseph D. Mougous[1,8,9,*]

[1]Department of Microbiology, University of Washington, Seattle, WA 98109, USA

[2]Systems Biology Department, National Center of Biotechnology CSIC, Madrid, 28049 Spain

[3]Division of Infectious Diseases, Boston Children's Hospital, Harvard Medical School, Boston, MA 02115, USA

[4]Department of Laboratory Medicine and Pathology, University of Washington, Seattle, WA 98195, USA

[5]Department of Genetics and Genome Sciences, Case Western Reserve University, Cleveland, OH 44106, USA

[6]Department of Physics, University of Washington, Seattle, WA 98195, USA

[7]Department of Bioengineering, University of Washington, Seattle, WA 98195, USA

[8]Department of Biochemistry, University of Washington School of Medicine, Seattle, WA 98195, USA

[9]Howard Hughes Medical Institute, University of Washington, Seattle, WA 98195, USA

* To whom correspondence should be addressed:

Email – simon.dove@childrens.harvard.edu, mougous@uw.edu

DNA–protein interactions (DPIs) are central to such fundamental cellular processes as transcription and chromosome maintenance and organization. The spatiotemporal dynamics of these interactions dictate their functional consequences; therefore, there is great interest in facile methods for defining the sites of DPI within cells. Here, we present a general method for mapping DPI sites *in vivo* using the double stranded DNA-specific cytosine deaminase toxin DddA. Our approach, which we term DddA-sequencing (3D-seq), entails generating a translational fusion of DddA to a DNA binding protein of interest, inactivating uracil DNA glycosylase, modulating DddA activity via its natural inhibitor protein, and DNA sequencing for genome-wide DPI detection. We successfully applied this method to three *Pseudomonas aeruginosa* transcription factors that represent divergent protein families and bind variable numbers of chromosomal locations. 3D-seq offers several advantages over existing technologies including ease of implementation and the possibility to measure DPIs at single-cell resolution.

Main

Advances in DNA sequencing have promoted rapid expansion in DPI mapping technologies and their applications. Chromatin immunoprecipation sequencing (ChIP-seq) became an early standard for studying both prokaryotic and eukaryotic systems[1]. In this approach, DPIs are identified through chemical crosslinking of DNA–protein complexes, DNA fragmentation, immunoprecipitation of a DNA binding protein (DBP) of interest, crosslink reversal, DNA purification, and DNA sequencing. Sample preparation is technically challenging and requires approximately one week to implement. More recently, Cut&Run and related technologies have gained popularity as alternatives to ChIP-seq[2,3]. These techniques offer several advantages relative to ChIP-seq including low starting material quantities that permit single cell measurements, the absence of crosslinking and its associated artifacts, and reduced sequencing with improved signal-to-noise[4-6].

Although powerful, ChIP-seq and Cut&Run-related approaches are fundamentally ex vivo technologies and cannot capture DPIs in living cells. A method that overcomes this limitation is DNA adenine methyltransferase identification (DamID), where the DBP of interest is fused to DAM and DPI site identification occurs through restriction enzyme or antibody mediated methylation site enrichment[7]. However, the utility of this technique is limited by low resolution (1 kb) owing to the frequency of DAM recognition sites (GATC) and by toxicity resulting from widespread adenine methylation. A second approach that facilitates the mapping of DPIs *in vivo* employs mapping the sites of insertion of so-called self-reporting transposons (SRTs). In this technique, a transposase is fused to the DBP of interest, and DPIs are identified by DNA or RNA sequencing to

38 determine sites of transposon insertion[8,9]. A major limitation to this approach is that
39 transposon insertions occur at low frequency within individual cells (15-100 events per
40 cell), and thus the technology it is not amenable to single cell studies[8]. Additionally, the
41 accumulation of transposon insertions within a population may cause phenotypic
42 consequences through gene disruption.

43 Nucleic acid-targeting deaminases are a diverse group of proteins that have found
44 a number of biotechnological applications due to their ability to introduce mutations in
45 DNA or RNA. Fusion of the single-stranded DNA (ssDNA) cytosine deaminase
46 APOBEC to catalytically inactive or nickase variants of Cas9 led to the development of
47 the first precision base editor capable of introducing single nucleotide substitutions (C•G-
48 to-T•A) *in vivo*[10]. This breakthrough technology inspired the repurposing of several other
49 ssDNA and RNA-targeting deaminases as base editing tools, including editors that
50 catalyze A•T-to-G•C substitutions in DNA, and RNA transcript editors that induce C to
51 U or A to I modifications[11]. RNA-targeting deaminases have additionally been employed
52 for the identification of RNA–protein complex sites[12]. As the only deaminase known to
53 act preferentially on dsDNA, the bacterial toxin-derived cytosine deaminase, DddA, is
54 unique. We previously capitalized on this feature to develop DddA-derived cytosine base
55 editors (DdCBEs), composed of DddA–TALE fusions that edit the human mitochondrial
56 genome in a programmable fashion[13]. In the current study, we harnessed the dsDNA-
57 targeting capability of DddA in the development of 3D-seq, a new technique for genome-
58 wide DPI mapping.

59 In DdCBEs, DddA activity is localized to particular sites on DNA by
60 reconstitution of the enzymatic domain of the toxin (amino acids 1264-1427) from split

forms fused to sequence-specific targeting proteins[13]. We envisioned an inverse approach whereby fusion of the intact deaminase domain of DddA, referred to herein as DddA, to DBPs with unknown binding sites could be used to define sites of interaction (Fig. 1a). To test the feasibility of this approach, we selected the candidate DNA binding protein GcsR of *P. aeruginosa*. GcsR is a sigma 54-dependent transcription activator of an operon encoding the glycine cleavage system (*gcvH2*, *gcvP2*, and *gcvT2*) and auxiliary glycine and serine metabolic genes (*glyA2* and *sdaA*)[14]. By analogy with closely related sigma 54-dependent regulators, also referred to as bacterial enhancer binding proteins (bEBPs), glycine binding to GcsR is thought to activate transcription of the operon by triggering conformational changes among subunits bound to three 18-bp tandem repeat binding sites in the *gcvH2* promoter region. RNA-seq analyses of *P. aeruginosa* Δ*gcsR* suggest that the *gcvH2* operon may encompass the only genes subject to direct regulation by GcsR[14].

To capture physiologically relevant DNA binding, we sought to generate a GcsR–DddA translational fusion encoded at the native *gcsR* locus. These efforts revealed that even in the context of fusion to transcription factors under native regulation, DddA exhibits sufficient toxicity to interfere with strain construction. To circumvent this, we inserted the gene encoding the DddA cognate immunity determinant, $dddA_I$, at the Tn7 attachment site under control of an arabinose inducible promoter (pAra). In this background, and with induction of immunity, we successfully replaced *gcsR* with an open reading frame encoding GcsR bearing an unstructured linker at its C-terminus fused to the deaminase domain of DddA (GcsR–DddA). Activation of the *gcvH2* operon by GcsR is required for *P. aeruginosa* growth using glycine as a sole carbon source[14]. We found

that unlike a strain lacking GcsR, strains expressing GcsR–DddA effectively utilize glycine as a growth substrate, suggesting the fusion retains functionality (Fig. 1b).

Our prior work demonstrated that uracil DNA glycosylase (Ung) effectively inhibits uracil accumulation in cells exposed to DddA[15]. Reasoning this DNA repair factor would limit our capacity to detect DddA activity, we deleted *ung* in the GcsR–DddA-expressing strain. Next, we passaged this strain in the presence and absence of arabinose and performed Illumina-based whole genome sequencing (WGS). Data from replicate experiments was minimally filtered to remove positions with low coverage or hypervariability (see methods) and the average frequency of C•G-to-T•A transition events within 5′-TC-3′ contexts were visualized across the *P. aeruginosa* genome (Fig. 1c-f). Other dinucleotide contexts were excluded based on the known strong preference of DddA for thymidine at the -1 position[13]. Remarkably, in samples propagated in the absence of arabinose, we observed a single apparent peak of DddA activity in this minimally filtered data, which was localized to the promoter region of *gcvH2* (Fig. 1f,g). This peak was not observed in samples containing arabinose, nor was it present in parallel studies using a strain containing Ung (Fig. 1d,e).

While a single peak of GcsR::DddA-dependent activity was readily apparent in our minimally processed data, we reasoned that additional filtering to remove background signal would improve the sensitivity and accuracy of our technique. The filters we employed are detailed in the methods and include *i*) accounting for sequencing errors by applying a minimum read count threshold for mutation events (~1%), *ii*) eliminating positions lacking a neighboring transition event within the approximate length window likely to be accessible to a bound DBP-DddA fusion protein (100 bp), and *iii*) removing

transitions representing SNPs present in the parent strain. Most significantly, given our prior observation that modifications catalyzed by free DddA are randomly distributed across genomes, we reasoned that substantial noise reduction could be achieved by removing transitions not reproduced in independent replicates. Visualization of four GcsR–DddA replicate datasets showed that transition events observed in at least three of the samples were highly enriched in the peak region associated with the *gcvH2* promoter (Fig. 2a,b), and therefore this criterion was added to our filtering workflow.

In parallel, we sought to develop a statistical analysis able to provide a quantitative means of distinguishing specific DPIs from background noise in 3D-seq data. Our approach employed a null hypothesis test and is described in detail in the methods. Briefly, a null hypothesis consisting of only background enzyme activity was compared to an alternative hypothesis in which a single putative peak was fit by maximum likelihood analysis. The null hypothesis was then either accepted or rejected at a confidence level of 95% using a Generalized Likelihood Ratio Test. If the null hypothesis was rejected, the model containing the peak replaced the null hypothesis and the test was repeated for another putative peak until no more peaks could be detected. P values for each detected peak are estimated and reported (Table S1). The application of these filtering criteria and statistical analyses to our GcsR 3D-seq data dramatically improved the apparent signal-to-noise and placed the major GcsR–DddA binding site centered within the 200 bp region containing the three known binding sites for GcsR[14] (Fig. 2c,d).

To benchmark the 3D-seq approach, we performed a comparative study using ChIP-seq – a current standard for assessing DPIs genome-wide in bacteria[16]. In place of the *dddA* translational fusion at the 3′ end of *gcsR*, we inserted a sequence encoding the

VSV-G epitope to facilitate the necessary immunoprecipitation step of ChIP-seq. Similar to 3D-seq, the most strongly supported candidate binding site for GcsR identified by ChIP-seq localized at the expected region upstream of *gcvH2* (Table S2). We noted in the course of this work that following strain construction, the 3D-seq workflow is considerably streamlined relative to that of ChIP-seq. The hands-on time to process a ChIP-seq sample to the point of sequencing library preparation is approximately one-week in our laboratory, whereas 3D-seq sample preparation constitutes only a genomic DNA preparation that occupies a portion of one day and requires little training.

Given that our initial experiment for detecting GcsR–DddA-catalyzed mutagenesis involved growth for multiple passages, we examined whether a peak of C•G-to-T•A transition frequency in the vicinity of the GcsR binding site could be detected after a shorter period of growth. In continuously growing cultures of *P. aeruginosa* Δ*ung* expressing GcsR–DddA in the absence of $DddA_I$ induction we observed a small peak at 9 hrs of propagation and robust DddA–GcsR activity was detected at 20 hrs of growth (Fig. S1). This latter incubation period was thus implemented for subsequent experiments.

We found that Ung inactivation is critical for the detection of GcsR-DNA interactions by 3D-seq (Fig. 1e,f). As an alternative to a *ung* knockout, we considered whether expression of the Ung inhibitor protein, UGI, could achieve sufficient Ung inactivation to reveal GcsR DPIs[17]. This approach is potentially advantageous for 3D-seq in organisms that are difficult to modify genetically. To determine whether expression of UGI could substitute for genetic inactivation of *ung*, we supplied *P. aeruginosa* expressing GcsR–DddA and $DddA_I$ with a plasmid possessing Ugi under control of the

153 *lac*UV5 promoter to allow orthogonal modulation of $DddA_I$ (arabinose) and Ugi (IPTG).
154 As when Ung was inactivated genetically, we found that inhibition of Ung by UGI
155 expression yielded a high significant peak of C•G-to-T•A transition events centered on
156 the known GcsR binding site upstream of *gcvH2* (Fig. S2, Table S1). This peak was not
157 observed in the empty vector control strain.

158 To begin to probe the versatility of 3D-seq, we next sought to determine whether
159 it could be successfully applied to the mapping of DPIs for a DBP that is structurally and
160 functionally divergent from GcsR. For this analysis, we selected the regulator GacA,
161 which belongs to a large group of transcription factors known as response regulators.
162 Canonically, phosphorylation of these proteins by cognate histidine kinases enhances
163 their interaction with promoter elements, leading to modulation of transcription [18]. In the
164 case of GacA, phosphorylation by the sensor kinase GacS promotes binding of GacA to
165 the promoter regions of two small RNA genes, *rsmY* and *rsmZ*[19]. GacS is itself regulated
166 by a second sensor kinase, RetS, which strongly inhibits GacS phosphotransfer to
167 GacA[20]. To further evaluate the capacity of 3D-seq to capture the effects of
168 posttranslational regulation of a transcription factor, we performed our studies in both
169 Δ*gacS* and Δ*retS* backgrounds of *P. aeruginosa*.

170 During preliminary testing of the 3D-seq protocol with GacA, we found that
171 repressing $DddA_I$ production by removing arabinose did not lead to detectable DddA
172 activity. We reasoned that leaky expression of $DddA_I$, which is well documented to occur
173 from pAra in *P. aeruginosa*, might be itself sufficient to effectively inhibit DddA in this
174 instance. After exploring alternative promoters without success, we tested a $DddA_I$
175 mutant in which the interaction with DddA is weakened by a C-terminal FLAG epitope

fusion ($DddA_I$–F, Fig. S3). At high arabinose levels, $DddA_I$–F provided sufficient protection against DddA to permit strain construction and under lower arabinose levels, DddA-dependent C•G-to-T•A transitions were observed.

Consistent with prior studies, 3D-seq revealed GacA binding sites upstream of *rsmY* and *rsmZ* in the Δ*retS* background of *P. aeruginosa* (Fig. 3a-c, Table S1). These peaks were the only significant GacA bindings sites detected and they were not found in the Δ*gacS* strain (Fig. 3d, Table S1). Huang et al. recently reported GacA binds 1125 sites across the *P. aeruginosa* genome, as measured by ChIP-seq[21]. Given the large discrepancy between this result and our findings by 3D-seq, we performed ChIP-seq analysis of GacA in-house. Rather than over-express GacA, which was the strategy adopted by Huang et al., we introduced an epitope-tagged allele of the regulator at its native locus in the Δ*retS* background of *P. aeruginosa*. Consistent with our 3D-seq results and an earlier ChIP-ChIP study[22], this approach identified regions upstream of *rsmY* and *rsmZ*, enriched 215- and 212-fold, respectively, as the two major bindings sites of GacA (Table S2). A third site located in the promoter region of PA4648 was the only additional site that surpassed our three-fold enrichment significance cut-off. These results added to our confidence in 3D-seq-based DPI site identification and they showed that the methodology can be applied to regulators of different binding modalities and with multiple interaction sites. Finally, they show that 3D-seq can potentially be used to assess PDI dynamics under different regulatory states.

Although they represent different transcription factor families, our findings show that GcsR and GacA both interact with a limited number of sites on the *P. aeruginosa* chromosome. To gauge the performance of 3D-seq when applied to a DBP with many

predicted sites of interaction, we selected the regulator FleQ. This protein is an unusual member of the bEBP family, as it can act as both an activator and repressor, it regulates transcription from both σ^{54} and σ^{70}-dependent promoters, and its regulatory functions appear to be modulated by interaction with an additional protein that does not bind DNA directly, FleN [23-26]. In its capacity as a σ^{54}-dependent transcription activator, studies have shown FleQ binds the promoters of several flagellar gene operons; as a σ^{70}-dependent regulator, it interacts with binding sites adjacent to or overlapping with transcription start sites for several genes involved in exopolysaccharide biosynthesis and can serve as both a repressor and activator depending on availability of the second messenger molecule cyclic-di-GMP[23,25,27]. To date, there are no published studies describing the full complement of genes directly regulated by FleQ in *P. aeruginosa.* FleQ was included in the study referenced above that utilized over-expressed transcription factors, but a list of FleQ sites was not provided, and our GacA ChIP-seq and 3D-seq results suggest the general workflow adopted by the authors is problematic[21].

3D-seq analysis employing FleQ–DddA expressed from its native promoter identified 14 peaks with a significantly elevated frequency of C•G-to-T•A transition events (Fig 3e-h, Table S1). Many of these peaks were localized to previously identified FleQ binding sites. Consistent with expectations for *P. aeruginosa* growing exponentially in liquid media, these included sites upstream of both exopolysaccaride biosynthesis and cell autoaggregation genes known to be repressed by FleQ (e.g. *pelA*, *pslA, siaA*) and flagellar motility genes known to be activated by the protein (e.g. *flhF, fliL, motD*) [23,25,27]. Interestingly, we also identified significant peaks upstream of several uncharacterized genes, including a homolog of the motility gene *fimV* (PA3340), a gene encoded

upstream of a c-di-GMP biosynthetic enzyme (PA2869), and a gene with no predicted links to other FleQ-regulated functions (PA3440) (Table S1). These results illustrate the capacity of 3D-seq to sensitively and specifically identify DPIs for proteins that bind at many sites across the genome.

To our knowledge, 3D-seq represents the first method for high-resolution genome-wide recording of DPIs in living cells. In addition to this unique capability of 3D-seq, we found the method offers several advantages over commonly employed technologies for DPI mapping. Key among these is its ease in implementation. Once the appropriate genetic elements are in place, which can in principle be reduced to transformation by a single plasmid, 3D-seq involves simply growing a small volume of the strain under examination followed by genomic DNA preparation and standard WGS. In contrast, ChIP-seq requires a number of specialized reagents, including highly purified antibodies targeting the DBP of interest or an associated epitope tag, and the subsequent technically demanding immunoprecipitation procedure requires several days to complete[1]. Another distinct advantage of 3D-seq is the minimal starting material required. Owing to handling challenges and sample loss occurring at each step of the ChIP-seq protocol, these experiments must generally be initiated with ~40-80 mL of bacterial culture [28,29] . The lower limit on material for a 3D-seq study is defined only by the terminal DNA sequencing technology being utilized. Indeed, we predict that in many circumstances, the genome of a single cell would be adequate for revealing DPIs by 3D-seq.

As performed in this study, 3D-seq exploits the small size of bacterial genomes to cost-effectively obtain high coverage (>100-fold) that can be translated into semi-

quantitative measures of DBP occupancy. In eukaryotic organisms with substantially larger genomes, an approach such as this is impractical and enrichment strategies are preferred. Nevertheless, we anticipate that 3D-seq will find application in organisms with large genomes. If experiments are conducted in a manner that permits mutations introduced by the DBP–DddA fusion of interest to approach 100% frequency in the population, far less sequencing depth is required. In another variation, candidate sites could be amplified by PCR and amplicon sequencing would be used to reveal lower frequency modifications.

Despite the strong performance of 3D-seq, there is ample opportunity for optimizing the technology. The straightforward genetic manipulation of *P. aeruginosa* allowed us to generate chromosomally-encoded DBP–DddA fusions and $DddA_I$ expression constructs. These sequences, along with that necessary for Ugi expression, could readily be incorporated into a single plasmid, thus eliminating the need for chromosomal manipulations. Future work will compare the performance of 3D-seq by the two approaches.

The resolution of 3D-seq is currently limited by the frequency of cytosines found in the sequence context preferred by DddA, 5′-TC-3′. In *P. aeruginosa*, this dinucleotide motif occurs on average every 12 bp, allowing sufficient resolution to accurately identify DPI sites. Although the average frequency of 5′-TC-3′ is expected to remain relatively consistent across organisms with varying GC content, within particular genomic regions, the frequency of 5′-TC-3′ could diminish substantially and limit resolution. DddA derivatives or novel dsDNA-targeting deaminases with alternative or relaxed sequence specificity hold great promise as a solution to this limitation of 3D-seq[15].

While we have demonstrated the utility of 3D-seq for the population-level mapping of DPIs involving bacterial transcription factors during in vitro growth, we envision its unique features will catalyze additional applications of the technology going forward. One such feature is the ability to modulate DddA activity through $DddA_I$ expression, which enables 3D-seq to capture a snapshot of the protein-DNA landscape during a fixed period of time. This can be particularly advantageous for identifying DPIs during growth under physiological conditions inaccessible to other mapping methods, such as during colonization of a host. The capacity to inducibly inhibit DddA also raises the intriguing possibility of employing 3D-seq to map DPIs within single cells. In this derivative of the technique, we envision a bacterial population would be grown under a condition of interest in the absence of $DddA_I$ expression, and subsequently individual clones would be isolated (e.g. as colonies) from media containing the inducer for $DddA_I$. Sequencing of these clones, which contain a mutational record of the activity and location of the DBP of interest, will provide heretofore unobtainable genome-wide insights into cell-cell heterogeneity in DPIs. In summary, we anticipate the simplicity of 3D-seq will greatly improve the accessibility of genome-wide DPI mapping studies and its unique attributes will help usher in a new era of DPI measurements in physiological contexts.

Methods

Bacterial strains, plasmids, and growth conditions

Detailed lists of all strains and plasmids used in this study can be found in Tables S3 and S4. All *P. aeruginosa* strains were derived from the sequenced strain PAO1[30] and were grown on Luria-Bertaini (LB) medium at 37°C supplemented as appropriate with 30 µg ml^{-1} gentamicin, 25 µg ml^{-1} irgasan, 5% (*w/v*) sucrose, 1.0 mM IPTG (isopropyl β-D-1-thiogalactopyranoside),and arabinose at varying concentrations. *Escherichia coli* was grown in LB medium supplemented as appropriate with 15 µg ml^{-1} gentamicin, 50 µg ml^{-1} trimethoprim, and 1% rhamnose. *E. coli* strains DH5α was used for plasmid maintenance and SM10 (Novagen, Hornsby Westfield, Australia) HB101 (pRK2103) and S17-1 were used for conjugative transfer.

Plasmid construction

Details of plasmid construction and primer sequences are provided in Tables S5 and S6. Plasmid pEXG2 was used to make the in-frame deletion constructs pEXG2-Δ*gcsR* as well as the VSV-G insertion constructs pEXG2-GcsR-V and pEXG2-GacA-V and the DddA fusion constructs pEXG2-*gcsR::dddA,* pEXG2-*gacA–dddA* and pEXG2-*fleQ–dddA*[31] . Plasmid pEXG2-Δ*gcsR*, was constructed by amplification of ~400 bp regions of genomic DNA flanking *gcsR*, with primers containing restriction sites, followed by digestion and ligation into pEXG2 that had been digested with the appropriate restriction enzymes. C-terminal VSV-G insertion constructs for GcsR–V and GacA–V were made by amplifying ~400 bp regions flanking each insertion site using primers that contained an in-frame sequence encoding the VSV-G epitope tag. Constructs for generating DddA

fusions encoded a protein in which DddA was fused to the C-terminus via a 32aa linker (SGGSSGGSSGSETPGTSESATPESSGGSSGGS). To generate these constructs, primers with 3′ overlapping regions were used to amplify both the linker and *dddA*, as well as 500 bp regions flanking the C-terminus of each gene. Gibson assembly[32] was then used for the generation of the pEXG2 plasmids containing each construct, and assembly mixes were transformed into *E. coli* DH5α expressing $DddA_I$ from pSCrhaB2-$DddA_I$ to avoid DddA-mediated toxicity. Construction of pEXG2-derived plasmids for deletion of *gacS, retS* and *ung* was previously described[15,33,34]. Site-specific chromosomal insertions of the immunity gene $dddA_I$ (with or without a FLAG tag at encoded at the C-terminus) were generated using pUC18T-miniTn7T-Gm-pBAD-*araE*. The genes encoding $DddA_I$ or $DddA_I$-FLAG were amplified and cloned into the KpnI/HindIII sites of this vector through Gibson assembly, to generate pUC18-miniTn7T-Gm-pBAD-araE-$dddA_I$ and pUC18T-miniTn7T-Gm-pBAD-araE-$dddA_I$-*FLAG*.

Strain construction

P. aeruginosa strains containing in-frame deletions of *gcsR, ung, retS or gacS* were constructed by allelic replacement using the appropriate pEXG2-derived deletion construct and were verified by PCR and site specific or genomic sequencing as described previously[31]. *P. aeruginosa* cells synthesizing GcsR with a C-terminal VSV-G epitope tag from the native chromosomal location were made by allelic replacement using vector pEXG2-GcsR-V. *P. aeruginosa* Δ*retS* mutant cells synthesizing GacA with a C-terminal VSV-G epitope tag from the native chromosomal location (*P. aeruginosa* Δ*retS* GacA-V) were made by allelic replacement using vector pEXG2-GacA-V. The *P. aeruginosa*

Δ*gcsR*, GcsR-V, and Δ*retS* GacA-V strains were verified by PCR and production of the GcsR-V and GacA-V fusion proteins was verified by Western blotting using an antibody against the VSV-G epitope tag. *P. aeruginosa* strains producing DddA fusion proteins were generated by first engineering the parent strain to express $DddA_I$ or $DddA_I$-FLAG from the chromosome under arabinose-inducible control by introduction of pUC18T-miniTn7T-Gm-pBAD-araE-*$dddA_I$* or pUC18T-miniTn7T-Gm-pBAD-araE-*$dddA_I$-FLAG* and helper plasmids pTNS3 and pRK2013 via tetraparental mating [35]. After chromosomal integration the GmR marker was removed from these cassettes by Flp/FRT recombination using plasmid pFLP2, which was then cured by sucrose counterselection[36]. *P. aeruginosa* strains synthesizing GcsR-DddA, GacA-DddA or FleQ-DddA from the native chromosomal loci of each regulator were then generated by two-step allelic exchange using the relevant pEXG2 construct. Rhamnose (0.1%, for *E. coli*) or arabinose (0.1%, for *P. aeruginosa*) were maintained during the DddA-fusion expressing strain construction process to minimize DddA toxicity and off-target activity. Fusion-expressing strains were verified by PCR and by assembly of complete genome sequences obtained during 3D-seq analyses.

Assessing the functionality of the GcsR-DddA fusion protein

To determine the functionality of the GcsR-DddA fusion protein cells were grown in biological triplicate in No Carbon E (NCE) minimal media[37] containing arabinose (1%) and glycine (20 mM), or arabinose (1%) and succinate (20 mM), at 37°C with aeration for 48 hours. Growth was determined by measuring the culture OD_{600}.

3D-seq sample preparation and sequencing

Culturing of DddA-fusion expressing strains

To generate genomic DNA for 3D-Seq analysis, strains carrying specific DddA fusion constructs and attTn7::araC-P_{BAD}-*dddI* (GcsR) or attTn7::araC-P_{BAD}-*dddI-FLAG* (GacA, FleQ) were grown for varying amounts of time and with variable levels of arabinose to induce $DddA_I$ or $DddA_I$-FLAG expression and/or IPTG to induce UGI production from pPSV39-UGI. In each case, the strains were initially streaked for single colonies on LB containing 0.1% or 1% arabinose, and single colonies were used to inoculate quadruplicate liquid cultures containing 0.1% or 1% arabinose. After ~16 hrs growth, these cultures were then washed with LB and used to inoculate fresh cultures. For GcsR-DddA in Δ*ung* and *ung*+ backgrounds and for the Δ*ung* strain without a dddA fusion contruct, washed cultures were inoculated into LB containing 0.1% (negative control) or no (experimental) arabinose at OD_{600} = 0.02, then grown for 8hrs before diluting back to OD_{600} = 0.02. After an additional ~16 hrs, cultures were again washed and diluted to OD_{600} = 0.02, then grown a final 8 hrs before samples were collected for genomic DNA preparation. For *gacA-dddA* (with Δ*retS* or Δ*gacS*) and *fleQ-dddA*, washed cultures were inoculated into LB containing 0.0005% arabinose at OD_{600} = 0.02, then grown for 6.5 hrs before samples were collected for genomic DNA preparation.

Genomic DNA preparation and sequencing

Genomic DNA was isolated from bacterial pellets using DNEasy Bood and Tissue Kit (Qiagen). Sequencing libraries for whole-genome sequencing were prepared from 200-300 ng of DNA using DNA Prep Kit (Illumina), with KAPA HiFi Uracil+ Kit (Roche)

used in place of Enhanced PCR Mix for the amplification step. Libraries were sequenced in multiplex by paired-end 150-bp reads on NextSeq 550 and iSeq instruments (Illumina).

ChIP-Seq sample preparation and library construction

200 mL cultures of the *P. aeruginosa* GcsR-V, wild-type, Δ*retS* and Δ*retS* GacA-V strains were grown in biological triplicate to an OD_{600} of 1.5 in LB at 37°C with aeration. 80 mL of culture was crosslinked with formaldehyde (1%) for 30 minutes at room temperature with gentle agitation. Crosslinking was quenched by the addition of glycine (250 mM) and cells were incubated at room temperature for 15 minutes with gentle agitation. Cells were pelleted by centrifugation, washed three times with phosphate buffered saline, and stored at -80°C prior to subsequent processing. Cell pellets were resuspended in 1 mL Buffer 1 (20 mM KHEPES, pH 7.9, 50 mM KCl, 0.5 mM dithiothreitol, 10% glycerol) plus protease inhibitor (complete-mini EDTA-free (Roche); 1 tablet per 10 mL), diluted to a total volume of 5.2 mL and divided equally among four 15 mL conical tubes (Corning). Cells were subsequently lysed and DNA sheared in a Bioruptor water bath sonicator (Diagenode) by exposure to two 8-minute cycles (30 seconds on, 30 seconds off) on the high setting. Cellular debris was removed by centrifugation at 4°C for 20 minutes at 20,000 xg. Cleared lysates were adjusted to match the composition of the immunoprecipitation (IP) buffer (10 mM Tris-HCl, pH 8.0, 150 mM NaCl, 0.1% NP-40 alternative (EMD-Millipore product 492018). The adjusted lysates were combined with anti-VSV-G agarose beads (Sigma) that had been washed once with IP buffer and reconstituted to a 50/50 bead/buffer slurry. For IP, 75 µL of the washed anti-VSV-G beads were added to each of the four aliquots for a given sample. IP

was performed overnight at 4°C with gentle agitation. Beads were then washed 5 times with 1 mL IP buffer and 2 times with 1X TE buffer (10 mM Tris-HCl, pH 7.4, 1 mM EDTA). Immune complexes were eluted from beads by adding 150 µL of TES buffer (50 mM Tris-HCl pH 8.0, 10 mM EDTA, 1% Sodium Dodecyl Sulfate (SDS)) and heating samples to 65°C for 15 minutes. Beads were pelleted by centrifugation (5 minutes at 16,000xg) at room temperature and a second elution was performed with 100 µL of 1X TE + 1% SDS. Supernatants from both elution steps were combined and incubated at 65°C overnight to allow cross-link reversal. DNA was then purified with a PCR purification kit (QIAGEN), eluted in 55 µL of 0.1X Elution Buffer and quantified on an Agilent Bioanalyzer. ChIP-Seq libraries were prepared from 1-40 ng of DNA using the NEBNext Ultra II DNA Library Prep Kit for Illumina (NEB). Adaptors were diluted 10-fold prior to ligation. AMPure XP beads (Beckman Coulter) were used to purify libraries, which were subjected to 7 rounds of amplification without size selection. Libraries were sequenced by the Biopolymers Facility (Harvard Medical School) on an Illumina HiSeq2500 producing 75-bp paired-end reads[38].

ChIP-Seq data analysis

ChIP-Seq data were analyzed as described previously[38]. Paired-end reads corresponding to fragments of 200 bp or less were mapped to the PAO1 genome (NCBI RefSeq NC_002516) using bowtie2 version 2.3.4.3 [39]. Only read 1 from each pair of reads was extracted and regions of enrichment were identified using QuEST version 2.4 [40]. Reads collected from the PAO1 replicates (i.e. IP from PAO1 cells that do not synthesize any VSV-G tagged protein) were merged and served as the mock control for the reads from

each of the PAO1 GcsR-V replicates. Merged reads from the PAO1 Δ*retS* replicates served as the mock control for the reads from the PAO1 Δ*retS* GacA-V replicates. The mock control data were used to determine the background for each corresponding ChIP biological replicate. The following criteria were used to identify regions of enrichment (peaks): (i) they must be 3.5-fold enriched in reads compared to the background, (ii) they are are not present in the mock control, (iii) they have a positive peak shift and strand correlation, and (iv) they have a q-value of less than 0.01. Peaks of enrichment for GcsR-V and GacA-V were defined as the maximal region identified in at least two biological replicates. Data were visualized using the Integrative Genomics Viewer (IGV) version 2.5.0[41]. Peak analyses used BEDtools version 2.27.1.

3D-seq data analysis

Fastq reads were first pre-processed using the HTStream pipeline v. 1.3.0 (https://s4hts.github.io/HTStream/), where the chain of programs is hts_SuperDeduper -> hts_SeqScreener -> hts_AdapterTrimmer -> hts_QWindowTrim -> hts_LengthFilter -> hts_Stats. In each case logging was enabled and default settings were used, with the following exceptions: 1) For hts_QWindowTrim a window size of 20bp was used with a minimum quality score of 10. 2) For hts_LengthFilter the minimum length was set to half the mean read length. Reads were subsequently aligned to the PAO1-UW reference sequence (https://www.ncbi.nlm.nih.gov/nuccore/NC_002516.2) using Minimap2 v. 2.17-r974-dirty (https://lh3.github.io/minimap2/) and the alignments were saved into sorted BAM files with SAMTools v. 1.10 (https://www.htslib.org/). Alignment position counts were then enumerated using PySAM v. 0.16.0.1

(https://pysam.readthedocs.io/en/latest/) using these settings: read_callback=′all′, quality_threshold=20. The reference genome was then surveyed using Biopython v. 1.78 (https://biopython.org/) to determine the proportion of high-quality read-pairs covering each 5′-TC-3′ site (the preferred DddA target sequence context; Mok et al.) on either strand that showed the alternative sequence 5′-TT-3′ (representing cytidine deamination), and corresponding base counts and allele frequencies were tabulated using Pandas v. 1.3.0. (https://pandas.pydata.org/).

To generate minimally filtered datasets, sites with sequence coverage of less than 15 read-pairs for that sample were ignored, as were a set of 52 sites within a phage region known to display hypervariability[42]. Average C•G-to-T•A transition frequency was then calculated using remaining positions for each set of quadruplicate samples per condition. To generate more stringently filtered data, sites with >95% C•G-to-T•A transition frequency in all four replicates of a given sample were considered parental SNPs and were ignored. The mean C•G-to-T•A transition frequency was then calculated for each position at which 3 of 4 replicate samples for a given condition exhibited at least 3 sequencing reads containing the mutation. Finally, positions were excluded for which the nearest neighboring position with an average C•G-to-T•A transition frequency >0 was within more than 100 bp. To generate the representations of the data shown in Fig. 1 and 2, this data was further processed by the calculation of a moving average employing a 75 bp window. For statistical analyses, we used data passing these criteria except we required a minimum of only 1 read to contain a given mutation. Additionally, we removed positions from any single sample with

Statistical analysis

See supplemental methods.

Data availability

Sequence data associated with this study is available from the Sequence Read Archive at BioProject PRJNA748760.

Code availability

Computer code generated for this study is available from GitHub at https://github.com/marade/3DSeqTools.

Acknowledgements

The authors wish to thank members of the Mougous laboratory for helpful suggestions. This work was supported by the NIH (AI080609 to JDM, AI143771 to S.L.D, GM128191 to P.A.W and DK089507 to SJS), and the Cystic Fibrosis Foundation (SINGH19R0 to SJS and a postdoctoral fellowship to M.J.G.), and the SynBio4Flav (H2020-NMBP-TR-IND/H2020-NMBP-BIO-2018-814650) and MIX-UP (MIX-UP H2020-BIO-CN-2019-870294) contracts of the European Union (to V.L.). E.V. was the recipient of a Fellowship from the Education Ministry, Madrid, Spanish Government (FPU15/04315). JDM is an HHMI Investigator.

Competing interests

The authors declare no competing interests.

Author contributions

L.A.G, E.V., S.B.P, M.H.D., V.L., S.L.D. and J.D.M designed the study, L.A.G, E.V., J.C.C, F.H., K.M.P., J.K., and P.A.A. performed experiments, L.A.G, E.V., S.B.P, M.C.R., M.J.G., T.L., S.J.S, P.A.W., S.L.D. and J.D.M analyzed data, and S.B.P and J.D.M wrote the manuscript with input from the other authors.

Author Information

Correspondence and requests for materials should be addressed to J.D.M. (mougous@u.washington.edu) or S.L.D. (simon.dove@childrens.harvard.edu).

References

1 Park, P. J. ChIP-seq: advantages and challenges of a maturing technology. *Nat Rev Genet* **10**, 669-680, doi:10.1038/nrg2641 (2009).

2 Henikoff, S., Henikoff, J. G., Kaya-Okur, H. S. & Ahmad, K. Efficient chromatin accessibility mapping in situ by nucleosome-tethered tagmentation. *Elife* **9**, doi:10.7554/eLife.63274 (2020).

3 Skene, P. J. & Henikoff, S. An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. *Elife* **6**, doi:10.7554/eLife.21856 (2017).

4 Kaya-Okur, H. S., Janssens, D. H., Henikoff, J. G., Ahmad, K. & Henikoff, S. Efficient low-cost chromatin profiling with CUT&Tag. *Nat. Protoc.* **15**, 3264-3283, doi:10.1038/s41596-020-0373-x (2020).

5 Kaya-Okur, H. S. *et al*. CUT&Tag for efficient epigenomic profiling of small samples and single cells. *Nat. Commun.* **10**, 1930, doi:10.1038/s41467-019-09982-5 (2019).

6 Meers, M. P., Bryson, T. D., Henikoff, J. G. & Henikoff, S. Improved CUT&RUN chromatin profiling tools. *Elife* **8**, doi:10.7554/eLife.46314 (2019).

7 van Steensel, B. & Henikoff, S. Identification of in vivo DNA targets of chromatin proteins using tethered dam methyltransferase. *Nat. Biotechnol.* **18**, 424-428, doi:10.1038/74487 (2000).

8 Moudgil, A. *et al*. Self-Reporting Transposons Enable Simultaneous Readout of Gene Expression and Transcription Factor Binding in Single Cells. *Cell* **182**, 992-1008 e1021, doi:10.1016/j.cell.2020.06.037 (2020).

9 Wang, H., Mayhew, D., Chen, X., Johnston, M. & Mitra, R. D. "Calling cards" for DNA-binding proteins in mammalian cells. *Genetics* **190**, 941-949, doi:10.1534/genetics.111.137315 (2012).

10 Komor, A. C., Kim, Y. B., Packer, M. S., Zuris, J. A. & Liu, D. R. Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. *Nature* **533**, 420-424, doi:10.1038/nature17946 (2016).

11 Porto, E. M., Komor, A. C., Slaymaker, I. M. & Yeo, G. W. Base editing: advances and therapeutic opportunities. *Nat Rev Drug Discov* **19**, 839-859, doi:10.1038/s41573-020-0084-6 (2020).

12 Brannan, K. W. *et al*. Robust single-cell discovery of RNA targets of RNA-binding proteins and ribosomes. *Nat Methods* **18**, 507-519, doi:10.1038/s41592-021-01128-0 (2021).

13 Mok, B. Y. *et al*. A bacterial cytidine deaminase toxin enables CRISPR-free mitochondrial base editing. *Nature* **583**, 631-637, doi:10.1038/s41586-020-2477-4 (2020).

14 Sarwar, Z. *et al*. GcsR, a TyrR-Like Enhancer-Binding Protein, Regulates Expression of the Glycine Cleavage System in Pseudomonas aeruginosa PAO1. *mSphere* **1**, doi:10.1128/mSphere.00020-16 (2016).

15 de Moraes, M. H. *et al*. An interbacterial DNA deaminase toxin directly mutagenizes surviving target populations. *Elife* **10**, doi:10.7554/eLife.62967 (2021).

16 Myers, K. S., Park, D. M., Beauchene, N. A. & Kiley, P. J. Defining bacterial regulons using ChIP-seq. *Methods* **86**, 80-88, doi:10.1016/j.ymeth.2015.05.022 (2015).

17 Mol, C. D. *et al*. Crystal structure of human uracil-DNA glycosylase in complex with a protein inhibitor: protein mimicry of DNA. *Cell* **82**, 701-708, doi:10.1016/0092-8674(95)90467-0 (1995).

18 Gao, R., Bouillet, S. & Stock, A. M. Structural Basis of Response Regulator Function. *Annu. Rev. Microbiol.* **73**, 175-197, doi:10.1146/annurev-micro-020518-115931 (2019).

19 Lapouge, K., Schubert, M., Allain, F. H. & Haas, D. Gac/Rsm signal transduction pathway of gamma-proteobacteria: from RNA recognition to regulation of social behaviour. *Mol. Microbiol.* **67**, 241–253 (2008).

20 Goodman, A. L. *et al*. Direct interaction between sensor kinase proteins mediates acute and chronic disease phenotypes in a bacterial pathogen. *Genes Dev.* **23**, 249–259, doi:10.1101/gad.1739009 (2009).

21 Huang, H. *et al*. An integrated genomic regulatory network of virulence-related transcriptional factors in Pseudomonas aeruginosa. *Nat. Commun.* **10**, 2931, doi:10.1038/s41467-019-10778-w (2019).

22 Brencic, A. *et al*. The GacS/GacA signal transduction system of Pseudomonas aeruginosa acts exclusively through its control over the transcription of the RsmY and RsmZ regulatory small RNAs. *Mol. Microbiol.* **73**, 434–445 (2009).

23 Baraquet, C., Murakami, K., Parsek, M. R. & Harwood, C. S. The FleQ protein from Pseudomonas aeruginosa functions as both a repressor and an activator to control gene expression from the pel operon promoter in response to c-di-GMP. *Nucleic Acids Res.* **40**, 7207-7218, doi:10.1093/nar/gks384 (2012).

24 Dasgupta, N. *et al*. A four-tiered transcriptional regulatory circuit controls flagellar biogenesis in Pseudomonas aeruginosa. *Mol. Microbiol.* **50**, 809-824, doi:10.1046/j.1365-2958.2003.03740.x (2003).

25 Jyot, J., Dasgupta, N. & Ramphal, R. FleQ, the major flagellar gene regulator in Pseudomonas aeruginosa, binds to enhancer sites located either upstream or atypically downstream of the RpoN binding site. *J. Bacteriol.* **184**, 5251-5260, doi:10.1128/JB.184.19.5251-5260.2002 (2002).

26 Hickman, J. W. & Harwood, C. S. Identification of FleQ from Pseudomonas aeruginosa as a c-di-GMP-responsive transcription factor. *Mol. Microbiol.* **69**, 376–389, doi:MMI6281 [pii] 10.1111/j.1365-2958.2008.06281.x (2008).

27 Baraquet, C. & Harwood, C. S. FleQ DNA Binding Consensus Sequence Revealed by Studies of FleQ-Dependent Regulation of Biofilm Gene Expression in Pseudomonas aeruginosa. *J. Bacteriol.* **198**, 178-186, doi:10.1128/JB.00539-15 (2016).

28 Ramsey, K. M. *et al*. Ubiquitous promoter-localization of essential virulence regulators in Francisella tularensis. *PLoS Pathog*. **11**, e1004793, doi:10.1371/journal.ppat.1004793 (2015).

29 Singh, S. S. *et al*. Widespread suppression of intragenic transcription initiation by H-NS. *Genes Dev*. **28**, 214-219, doi:10.1101/gad.234336.113 (2014).

30 Stover, C. K. *et al*. Complete genome sequence of Pseudomonas aeruginosa PA01, an opportunistic pathogen. *Nature* **406**, 959–964 (2000).

31 Rietsch, A., Vallet-Gely, I., Dove, S. L. & Mekalanos, J. J. ExsE, a secreted regulator of type III secretion genes in Pseudomonas aeruginosa. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 8006–8011 (2005).

32 Gibson, D. G. *et al*. Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat Methods* **6**, 343–345, doi:10.1038/nmeth.1318 (2009).

33 LeRoux, M. *et al*. Kin cell lysis is a danger signal that activates antibacterial pathways of Pseudomonas aeruginosa. *Elife* **4**, doi:10.7554/eLife.05701 (2015).

34 Mougous, J. D. *et al*. A virulence locus of Pseudomonas aeruginosa encodes a protein secretion apparatus. *Science* **312**, 1526–1530 (2006).

35 Kulasekara, B. R. *et al*. c-di-GMP heterogeneity is generated by the chemotaxis machinery to regulate flagellar motility. *Elife* **2**, e01402, doi:10.7554/eLife.01402 (2013).

36 Hoang, T. T., Kutchma, A. J., Becher, A. & Schweizer, H. P. Integration-proficient plasmids for Pseudomonas aeruginosa: site-specific integration and use for engineering of reporter and expression strains. *Plasmid* **43**, 59–72 (2000).

37 Davis, R. W., Botstein, D. & Roth, J. R. *Advanced Bacterial Genetics: A Manual for Genetic Engineering*. (Cold Spring Harbor Laboratory, 1980).

613 38 Gebhardt, M. J., Kambara, T. K., Ramsey, K. M. & Dove, S. L. Widespread
614 targeting of nascent transcripts by RsmA in Pseudomonas aeruginosa. *Proc. Natl.*
615 *Acad. Sci. U. S. A.* **117**, 10520-10529, doi:10.1073/pnas.1917587117 (2020).

616 39 Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat*
617 *Methods* **9**, 357–359, doi:10.1038/nmeth.1923 (2012).

618 40 Valouev, A. *et al*. Genome-wide analysis of transcription factor binding sites
619 based on ChIP-Seq data. *Nat Methods* **5**, 829-834, doi:10.1038/nmeth.1246
620 (2008).

621 41 Thorvaldsdottir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics
622 Viewer (IGV): high-performance genomics data visualization and exploration.
623 *Brief Bioinform* **14**, 178-192, doi:10.1093/bib/bbs017 (2013).

624 42 Klockgether, J., Cramer, N., Wiehlmann, L., Davenport, C. F. & Tummler, B.
625 Pseudomonas aeruginosa Genomic Structure and Diversity. *Front. Microbiol.* **2**,
626 150, doi:10.3389/fmicb.2011.00150 (2011).

627 43 Boyer, H. W. & Roulland-Dussoix, D. A complementation analysis of the
628 restriction and modification of DNA in Escherichia coli. *J. Mol. Biol.* **41**, 459-
629 472, doi:10.1016/0022-2836(69)90288-5 (1969).

630 44 Simon, R., Priefer, U. & Puher, A. A Broad Host Range Mobilization System for
631 In Vivo Genetic Engineering: Transposon Mutagenesis in Gram Negative
632 Bacteria. *Nat. Biotechnol.* **1**, 784–791 (1983).

633 45 Cardona, S. T. & Valvano, M. A. An expression vector containing a rhamnose-
634 inducible promoter provides tightly regulated gene expression in Burkholderia
635 cenocepacia. *Plasmid* **54**, 219–228 (2005).

636 46 Silverman, J. M. *et al*. Haemolysin Coregulated Protein Is an Exported Receptor
637 and Chaperone of Type VI Secretion Substrates. *Mol. Cell* **51**, 584–593,
638 doi:10.1016/j.molcel.2013.07.025 (2013).

639 47 Figurski, D. H. & Helinski, D. R. Replication of an origin-containing derivative of
640 plasmid RK2 dependent on a plasmid function provided in trans. *Proc. Natl.*
641 *Acad. Sci. U. S. A.* **76**, 1648-1652, doi:10.1073/pnas.76.4.1648 (1979).

642 48 Choi, K. H. *et al*. Genetic tools for select-agent-compliant manipulation of
643 Burkholderia pseudomallei. *Appl. Environ. Microbiol.* **74**, 1064–1075 (2008).

644 49 Hoang, T. T., Karkhoff-Schweizer, R. R., Kutchma, A. J. & Schweizer, H. P. A
645 broad-host-range Flp-FRT recombination system for site-specific excision of
646 chromosomally-located DNA sequences: application for isolation of unmarked
647 Pseudomonas aeruginosa mutants. *Gene* **212**, 77-86, doi:10.1016/s0378-
648 1119(98)00130-9 (1998).

649 50 Schellenberger, V. *et al*. A recombinant polypeptide extends the in vivo half-life
650 of peptides and proteins in a tunable manner. *Nat*. *Biotechnol*. **27**, 1186-1190,
651 doi:10.1038/nbt.1588 (2009).
652

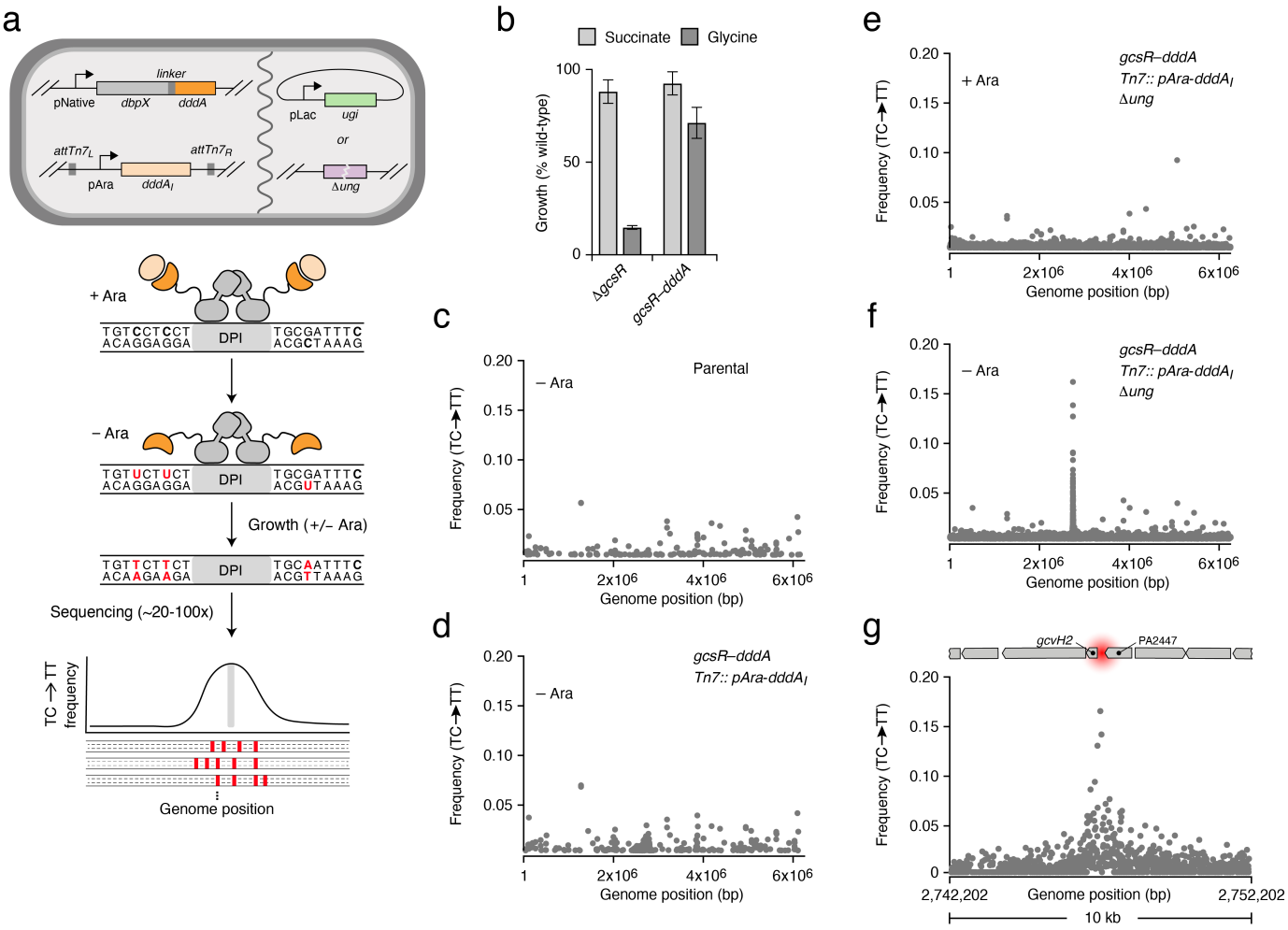

Fig. 1: 3D-seq is a method for *in vivo* DPI mapping and can be applied to *P. aeruginosa* GcsR. a, Diagram providing an overview of the 3D-seq method. Top, cell schematic containing the genetic elements required for 3D-seq. Elements may be integrated into the chromosome or supplied on plasmids. Middle, model depicting localized activity of DddA (dark orange) when fused to a DBP of interest (grey) and after growth in the absence of arabinose to limit production of $DddA_I$ (light orange). Bottom, schematized 3D-seq output indicating enrichment of C•G-to-T•A transitions (red) in the vicinity of a DPI site (grey). **b,** Growth yield (normalized to wild-type) of the indicated strains on minimal medium containing glycine or succinate as the sole carbon source. **c-f,** Average (n=4) C•G-to-T•A transition frequency by genome position after passaging cultures of *P. aeruginosa* bearing the indicated genotypes, in the presence or absence of arabinose (Ara) to induce $DddA_I$ expression. Data were filtered to remove a prophage hypervariable region and positions with low sequence coverage (<15-fold read depth), and positions with an average transition frequency <0.004 were removed ease of visualization. **g**, Zoomed view of a subset of the data depicted in (**f**). Location of the previously characterized GcsR binding site (red) and adjacent genetic elements shown to scale above.

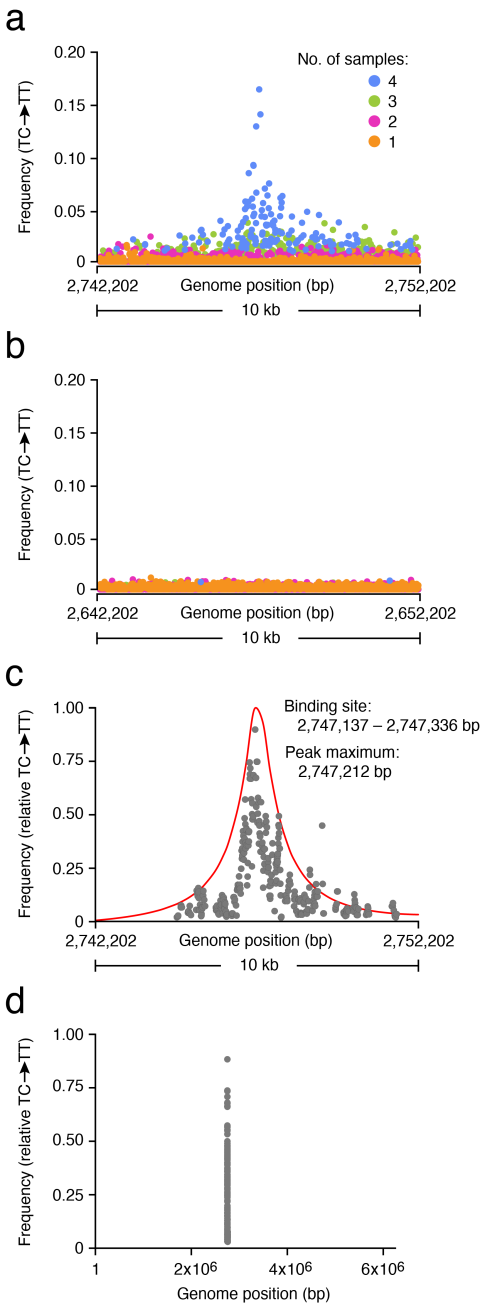

Fig. 2: Statistical analyses and data filtering enhance signal-to-noise and allow 3D-seq to precisely map DPIs. a,b Average (n=4) C•G-to-T•A transition frequency within the (**a**) primary GcsR 3D-seq peak region or (**b**) a control region located 100,000 bp upstream, with positions colored by the number of replicates in which a transition at that position was observed. **c**, Moving average (75 bp window) of C•G-to-T•A transition frequencies and the curve deriving from our statistical model (red line) calculated from filtered 3D-seq data for the GcsR peak region (see methods). Y-coordinates for the model curve are scaled arbitrarily. **d,** Genome-wide moving average (75 bp window) of C•G-to-T•A transition frequencies calculated for GcsR 3D-seq data after filtering as in (**c**).

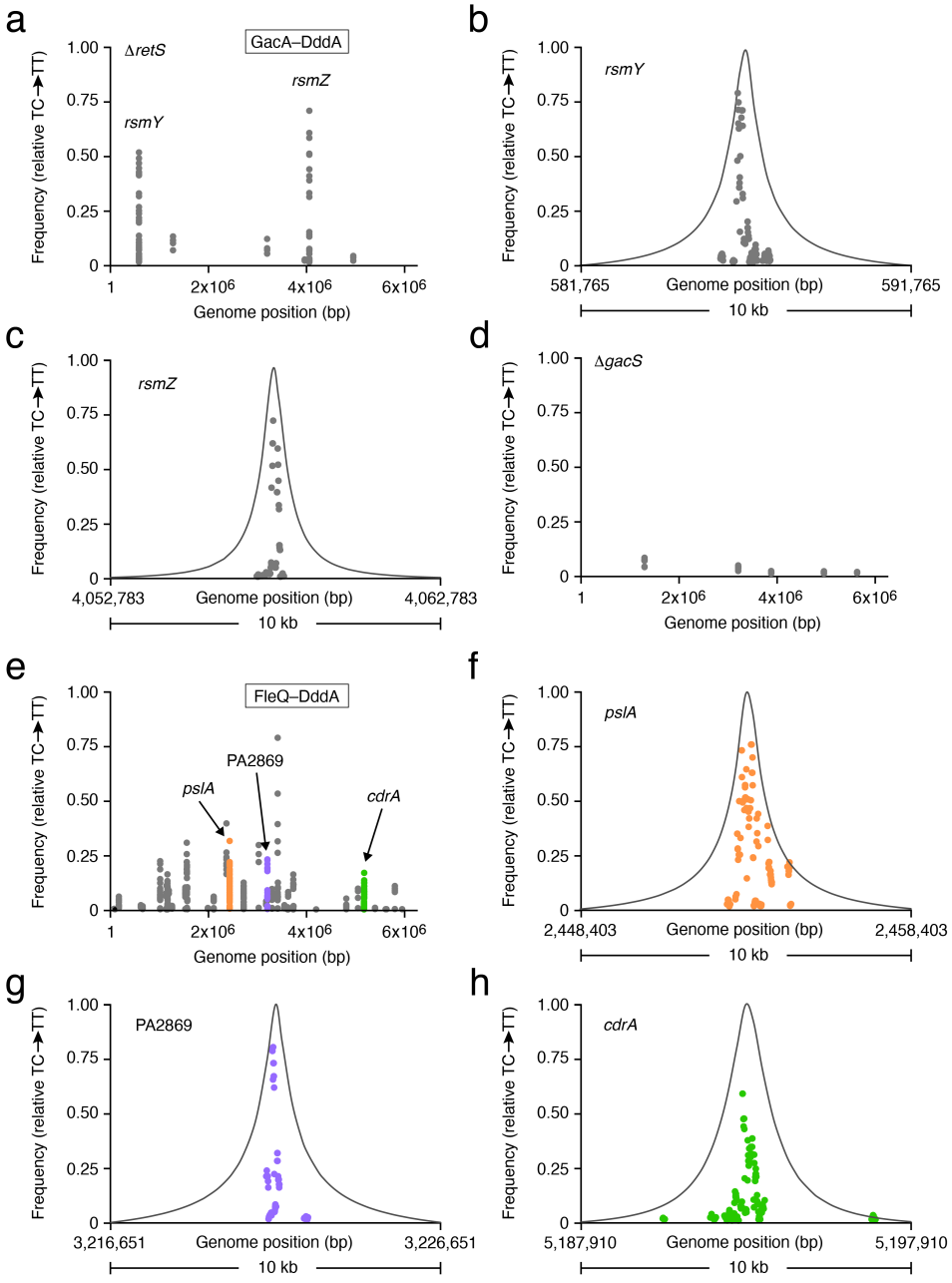

Fig. 3: 3D-seq maps DPIs for *P. aeruginosa* transcription factors belonging to different families and with varying numbers of binding sites. a-h, Moving average (n=4, 75 bp window) of C•G-to-T•A transition frequencies calculated from filtered 3D-seq data deriving from the indicated *P. aeruginosa* strains expressing GacA–DddA (**a-d**) or FleQ–DddA (**e-h**) grown with 0.0005% w/v arabinose for induction of $DddA_I$–F. Genome-wide (**a,d,e**) and zoomed (**b,c, f-h**) regions of the data shown in (**a**) or (**e**) are provided. Curves deriving from our statistical model (grey line) calculated from filtered 3D-seq data are shown in the zoomed regions. Y-coordinates for the model curves are scaled arbitrarily. Points in (**f-h**) are colored as in (**e**).

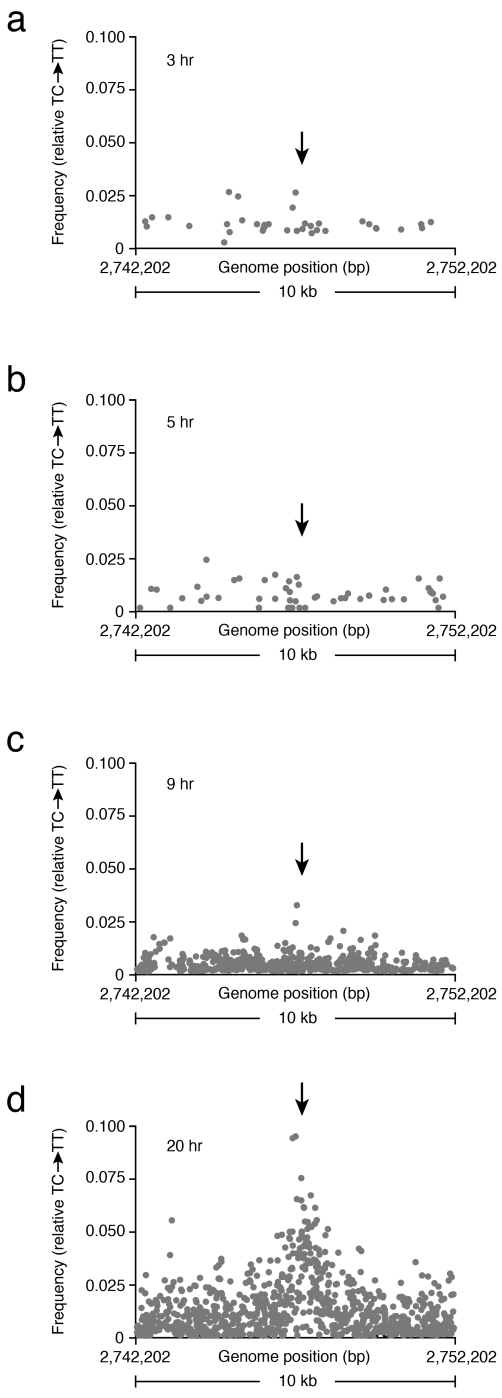

Figure S1. Transition mutations associated with GcsR:DddA activity accumulate over time. a-d, Average (n=4) C•G-to-T•A transition frequency within the primary GcsR 3D-seq peak region after the indicated growth period and in the absence of arabinose. Data were filtered as in Fig. 1. The arrow indicates the approximate position of the known GcsR binding site.

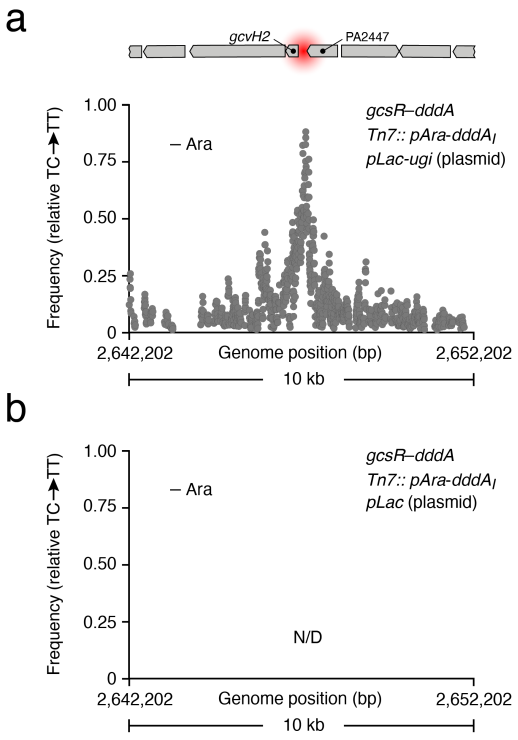

Figure S2. Ugi expression can substitute for genetic inactivation of *ung* in 3D-seq. a,b, Moving average (n=4, 75 bp window) of C•G-to-T•A transition frequencies calculated from filtered 3D-seq data deriving from the indicated *P. aeruginosa* strains grown in the absence of arabinose for 20 hrs. IPTG was included to induce the expression of Ugi throughout the growth period. The location of the previously characterized GcsR binding site (red) and adjacent genetic elements are shown to scale above.

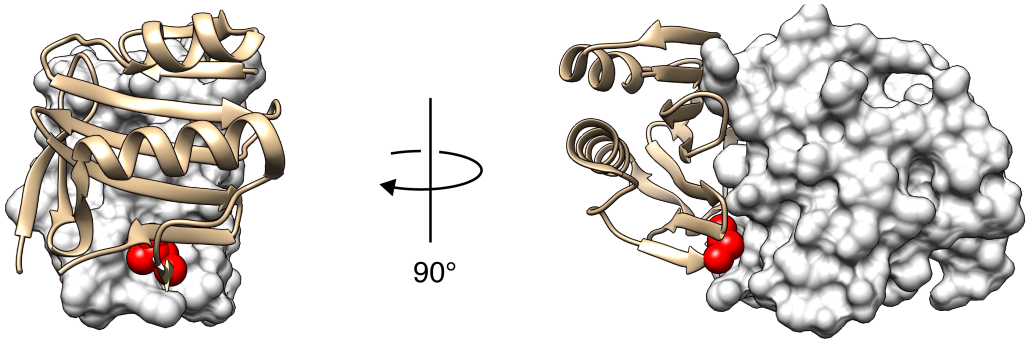

Figure S3. The C-terminus of $DddA_I$ abuts DddA. X-ray crystal structure of the $DddA_I$–DddA complex in ribbon and surface representation, respectively. The C-terminal amino acid of $DddA_I$ (Leu123) is colored red and shown in space filling representation to highlight its position against the surface of DddA.

Table S1. Significant peaks detected in this study by 3D-seq.

Peak number[1]	p-value[2]	Peak maximum position	Peak height[3]	Width parameter[4]	Closest annotated gene
GcsR (Δ*ung* Tn7::pAra-*dddA$_I$ gcsR–dddA*)					
1	<1.9E-999	2747212	0.0308	650	*gcvH2*
2	4.4E-6	2754622	0.0032	507	*gcvH2*
GacA (Δ*ung* Δ*retS* Tn7::pAra-*dddA$_I$ gacA–dddA*)					
1	2.0E-273	586705	0.0096	552	*rsmY*
2	4.5E-167	4057755	0.0090	449	*rsmZ*
FleQ (Δ*ung* Tn7::pAra-*dddA$_I$ fleQ–dddA*)					
1	1.9E-167	5192977	0.0182	828	*cdrA*
2	1.6E-157	2453477	0.0196	624	*pslA*
3	4.0E-25	3221641	0.0092	479	PA2869
4	5.9E-25	1549819	0.0079	402	*bdlA*
5	1.6E-21	3434343	0.0067	704	*pelA*
6	2.2E-16	1592133	0.0066	563	*motD*
7	2.2E-16	1187368	0.0042	480	*fleQ*
8	5.7E-14	2129043	0.0054	529	PA1945
9	2.4E-11	196854	0.0070	439	*siaA*
10	9.9E-11	1582818	0.0036	429	*flhF*
11	2.2E-09	2737599	0.0075	686	PA2440
12	1.4E-04	1572170	0.0030	394	*fliL*
13	1.6E-02	3751994	0.0027	316	PA3340
14	1.8E-02	1215200	0.0040	508	*yfiR*
GcsR (Tn7::pAra-*dddA$_I$ gcsR–dddA* pPSV39::pLac-*ugi*)					
1	<1.9E-999	2747582	0.0830	511	*gcvH2*
2	9.6E-113	6199779	0.0391	610	PA5508
3	2.1E-83	2744696	0.0344	204	*gcvH2*
4	1.5E-26	2141807	0.0153	1000	PA1957
5	5.0E-22	6151145	0.0104	934	PA5459
6	1.0E-21	4230516	0.0117	756	PA3372
7	1.8E-21	2754908	0.0127	283	PA2454
8	2.8E-17	1121658	0.0117	612	PA1032
9	4.0E-15	2749090	0.0463	26	PA2448
10	1.5E-12	6195909	0.0110	364	PA5503
11	9.0E-12	2751477	0.0155	213	PA2449
12	1.5E-09	2742380	0.0107	166	PA2443
13	8.9E-08	2750504	0.0168	109	*gcsR*

[1]Peaks are listed in order of increasing p-value.

[2]P-values calculated as described in the supplemental methods.

[3,4] The peak profile function represents cell-mean allele frequency as a function of genomic position. The amplitude parameter I represents the height of the peak of the profile function. The width parameter L controls its width. See supplemental methods.

Table S2. Significant peaks detected in this study by ChIP-seq.

Peak number[1]	Fold enrichment	Peak maximum	Peak start	Peak end	Closest annotated gene
GcsR (*gcsR*–VSV-G)					
1	92	2747441	2746047	2748292	*gcvH2*
2	13	5402788	5402462	5403174	*lipC*
3	9.1	945626	945487	945780	PA0864
4	5.6	2516215	2516073	2516373	PA2288
5	3.4	4373881	4373741	4374013	PA3904
6	3.4	3070313	3070176	3070417	PA2715
7	3.2	1205058	1204920	1205189	PA1112a
GacA (Δ*retS gacA*–VSV-G)					
1	22E01	586983	586145	587484	*rsmY*
2	21E01	4057808	4057399	4058161	*rsmZ*
3	3.6	5214833	5214674	5215156	PA4648

[1]Peaks are listed in order of decreasing fold enrichment.

Table S3: Bacterial strains used

Strain	Source
***P. aeruginosa* strains**	
PAO1	[30]
Δ*ung*	[15]
attTn7::*araC*-P_{BAD}-*dddA*$_I$	This study
Δ*ung* attTn7::*araC*-P_{BAD}-*dddA*$_I$	This study
attTn7::*araC*-P_{BAD}-*dddA*$_I$ *gcsR-dddA*	This study
Δ*ung* attTn7::*araC*-P_{BAD}-*dddA*$_I$ *gcsR-dddA*	This study
attTn7::*araC*-P_{BAD}-*dddA*$_I$ *gcsR-dddA* pPSV39(empty)	This study
attTn7::*araC*-P_{BAD}-*dddA*$_I$ *gcsR-dddA* pPSV39-*ugi*	This study
attTn7::*araC*-P_{BAD}-*dddA*$_I$-FLAG	This study
Δ*ung* attTn7::*araC*-P_{BAD}-*dddA*$_I$-FLAG	This study
Δ*ung* attTn7::*araC*-P_{BAD}-*dddA*$_I$-FLAG *gacA-dddA*	This study
Δ*ung* attTn7::*araC*-P_{BAD}-*dddA*$_I$-FLAG *gacA-dddA* Δ*retS*	This study
Δ*ung* attTn7::*araC*-P_{BAD}-*dddA*$_I$-FLAG *gacA-dddA* Δ*gacS*	This study
Δ*ung* attTn7::*araC*-P_{BAD}-*dddA*$_I$-FLAG *fleQ-dddA*	This study
Δ*gcsR*	This study
GcsR-V	This study
Δ*retS*	This study
Δ*retS* GacA-V	This study
***E. coli* strains**	
DH5α	
SM10	Novagen
HB101	[43]
S17-1	[44]

Table S4: Plasmids used

Plasmid	Reference
pEXG2	31
pSCrhaB2	45
pPSV39	46
pUC18T-miniTn7T-Gm-pBAD-araE	35
pRK2103	47
pTNS3	48
pFLP2	49
pEXG2-Δ*ung*	15
pEXG2-Δ*retS* (pEXG2_ΔPA4856)	34
pEXG2-Δ*gacS* (pEXG2_ΔPA0928)	33
pSCrhaB2-*dddA*$_I$	This study
pEXG2-*gcsR*-*dddA*	This study
pEXG2-*gacA*-*dddA*	This study
pEXG2-*fleQ*-*dddA*	This study
pEXG2-Δ*gcsR*	This study
pEXG2-GscR-V	This study
pEXG2-GacA-V	This study
pPSV39-*ugi*	This study
pUC18T-miniTn7T-Gm-pBAD-araE-*dddA*$_I$	This study
pUC18T-miniTn7T-Gm-pBAD-araE-*dddA*$_I$-FLAG	This study

Table S5: Plasmid construction.
The following plasmids were generated by combining the individual fragments listed using either Gibson cloning or overlap extension PCR. Primer sequences are listed in Table S6.

pSCrhaB2-*dddA$_I$*

Fragment	Template	F primer	R primer
pSCrhaB2/NdeI/XbaI	NA		
dddA$_I$	*dddA$_I$* [13]	pSCRhaB::*dddA*$_I$-F	pSCRhaB::*dddA$_I$*-R

pEXG2-*gcsR-dddA*

Fragment	Template	F primer	R primer
pEXG2/HindIII/XhoI	NA		
5' end of *gcsR*	PAO1 gDNA	*gcsR-dddA* 1	*gcsR-dddA* 2
linker+*dddA*	XTEN linker [50]	*gcsR-dddA* 3	*gcsR-dddA* 4
gcsR 3' flank	PAO1 gDNA	*gcsR-dddA* 5	*gcsR-dddA* 6

pEXG2-*gacA-dddA*

Fragment	Template	F primer	R primer
pEXG2/HindIII/XhoI	NA		
5' end of *gacA*	PAO1 gDNA	*gacA-dddA* 1	*gacA-dddA* 2
linker+*dddA*	XTEN linker [50]	*gacA-dddA* 3	*gacA-dddA* 4
gacA 3' flank	PAO1 gDNA	*gacA-dddA* 5	*gacA-dddA* 6

pEXG2-*fleQ-dddA*

Fragment	Template	F primer	R primer
pEXG2/HindIII/XhoI	NA		
5' end of *fleQ*	PAO1 gDNA	*fleQ-dddA* 1	*fleQ-dddA* 2
linker+*dddA*	XTEN linker [50]	*fleQ-dddA* 3	*fleQ-dddA* 4
fleQ 3' flank	PAO1 gDNA	*fleQ-dddA* 5	*fleQ-dddA* 6

pUC18T-miniTn7T-Gm-p$_{BAD}$-araE-*dddAI*

Fragment	Template	F primer	R primer
pUC18T-miniTn7T-Gm-pBAD-*araE*/KpnI/HindIII	NA		
dddA$_I$	*dddA$_I$* [13]	pUC18-*dddA$_I$*-F	pUC18-*dddA$_I$*-R

pUC18T-miniTn7T-Gm-p$_{BAD}$-araE-*dddAI*-FLAG

Fragment	Template	F primer	R primer
pUC18T-miniTn7T-Gm-pBAD-araE/KpnI/HindIII	NA		

dddA$_I$-FLAG	*dddA$_I$* [13]	pUC18-*dddA$_I$*-F	pUC18-*dddA$_I$*-FLAG-R

pPSV39-*ugi*

Fragment	Template	F primer	R primer
pPSV39/SacI/XbaI	NA		
ugi	ugi gBlock codon optimized	pPSV39-UGI-F	pPSV39-UGI-R

pEXG2-Δ*gcsR*

Fragment	Template	F primer	R primer
pEXG2/HindIII /BamHI	NA		
gcsR 5' flank	PAO1 gDNA	pEX.del.gcsR 1	pEX.del.gcsR 2
gcsR 3' flank	PAO1 gDNA	pEX.del.gcsR 3	pEX.del.gcsR 4

pEXG2-GcsR-V

Fragment	Template	F primer	R primer
pEXG2/HindIII /BamHI	NA		
gacA 5' flank	PAO1 gDNA	pEX-GcsR-V 1	pEX-GcsR-V 2
gacA 3' flank	PAO1 gDNA	pEX-GcsR-V 3	pEX-GcsR-V 4

pEXG2-GacA-V

Fragment	Template	F primer	R primer
pEXG2/HindIII /BamHI	NA		
fleQ 5' flank	PAO1 gDNA	pEX-GacA-V 1	pEX-GacA-V 2
fleQ 3' flank	PAO1 gDNA	pEX-GacA-V 3	pEX-GacA-V 4

Table S6: Primer sequences

Primer	Sequence
pSCRhaB::dddAI-F	TGAAATTCAGCAGGATCACATATGTACGCAGACGATTTCGAC
pSCRhaB::dddAI-R	TCATTTCAATATCTGTATATCTAGATTACAACTCGCTCCATGTC
gcsR-dddA 1	GGAAGCATAAATGTAAAGCAAGCTTGCAACCTGGAGAAGATG GTCGCCG
gcsR-dddA 2	TACCTCCAGAGGCGCGCGGACCGATGCC
gcsR-dddA 3	TCCGCGCGCCTCTGGAGGTAGCTCCGGC
gcsR-dddA 4	GCCCGCTTCAACAACCTCCTTTCGTGGG
gcsR-dddA 5	AGGAGGTTGTTGAAGCGGGCTCAGCCCT
gcsR-dddA 6	TTAAGGTACCGAATTCGAGCTCGAGCAATCCCAAGGAGTTCG AGCG
gacA-dddA 1	GGAAGCATAAATGTAAAGCAAGCTTCGGATGTCGTCCTGATG GAC
gacA-dddA 2	TACCTCCAGAGCTGGCGGCATCGACCAT
gacA-dddA 3	TGCCGCCAGCTCTGGAGGTAGCTCCGGC
gacA-dddA 4	CGCTCATCTAACAACCTCCTTTCGTGGG
gacA-dddA 5	AGGAGGTTGTTAGATGAGCGCCGTTTTCGACGC
gacA-dddA 6	TTAAGGTACCGAATTCGAGCTCGAGGGCCGCGTACGGTTGCG G
fleQ-dddA 1	GGAAGCATAAATGTAAAGCAAGCTTTCGCCCTGCTGCTCAACG
fleQ-dddA 2	TACCTCCAGAATCATCCGACAGGTCGTCG
fleQ-dddA 3	GTCGGATGATTCTGGAGGTAGCTCCGGC
fleQ-dddA 4	CGACCTGTCAACAACCTCCTTTCGTGGG
fleQ-dddA 5	AGGAGGTTGTTGACAGGTCGTTTCGCAACGCTTTG
fleQ-dddA 6	TTAAGGTACCGAATTCGAGCTCGAGCGCGCGGAGCGAAGCAG C
pPSV39-UGI-F	GATAACAATTTCAGAATTCGAGCTCACGGGAGGAAAGATGAC GAATCTCAGCGACAT
pPSV39-UGI-R	TCATTTCAATATCTGTATATCTAGATTAGAGCATCTTGATTTTG TTCTCGC
pUC18-dddAI-F	GGGCTAGCGAATTCGAGCTCGGTACCACGGGAGGAAAGATGT AC
pUC18-dddAI-R	CTCATCCGCCAAAACAGCCAAGCTTTCACAACTCGCTCCATGT C
pUC18-dddAI-FLAG-R	CTTCTCTCATCCGCCAAAACAGCCAAGCTTTCATTTGTCGTCGT CGTCTTTGTAGTCCAACT CGCTCCATGTCAG
pEX.del.gcsR 1	CATAAATGTAAAGCAAGCTTGGTACCGAGGCGGACT
pEX.del.gcsR 2	AGCCCGCTTCAGGCGCGCGGGATGCGCATGCGGGA
pEX.del.gcsR 3	CAGGTTCCCGCATGCGCATCCCGCGCGCCTGAAGC
pEX.del.gcsR 4	TCGAGCTCGAGCCCGGGGATCCTTCGATTACCCACCTGC
pEX-GcsR-V 1	CATAAATGTAAAGCAAGCTTACCTGTTCTACCGCCTCA
pEX-GcsR-V 2	CTTGCCGAGGCGGTTCATTTCGATGTCGGTGTAAGCGGCCGCG GCGCGCGGACCGATGC

pEX-GcsR-V 3	GCGGCCGCTTACACCGACATCGAAATGAACCGCCTCGGCAAG TGAAGCGGGCTCAGCCC
pEX-GcsR-V 4	TCGAGCTCGAGCCCGGGGATCCGAGTTCGAGCGCTTCAG
pEX-GacA-V 1	CATAAATGTAAAGCAAGCTTGAACTGAAGCCGGATGTC
pEX-GacA-V 2	CTTGCCGAGGCGGTTCATTTCGATGTCGGTGTAAGCGGCCGCG CTGGCGGCATCGACCA
pEX-GacA-V 3	GCGGCCGCTTACACCGACATCGAAATGAACCGCCTCGGCAAG TAGATGAGCGCCGTTTTC
pEX-GacA-V 4	TCGAGCTCGAGCCCGGGGATCCGCGCTCGGATAGGGACC