Fleming Prize Lecture 2016

The unexpected complexity of bacterial genomes

David C. Grainger

Institute of Microbiology and Infection, School of Biosciences, University of Birmingham, Edgbaston, Birmingham B15 2TT, UK

Correspondence
David C. Grainger
d.grainger@bham.ac.uk

Received 6 May 2016
Accepted 17 May 2016

Gene organization and control are described by models conceived in the 1960s. These models explain basic gene regulatory mechanisms and underpin current genome annotation. However, such models struggle to explain recent genome-scale observations. For example, accounts of RNA synthesis initiating within genes, widespread antisense transcription and non-canonical DNA binding by gene regulatory proteins are difficult to reconcile with traditional thinking. As a result, unexpected observations have often been dismissed and downstream consequences ignored. In this paper I will argue that, to fully understand the biology of bacterial chromosomes, we must embrace their hidden layers of complexity.

Introduction

Bacterial chromosomes primarily comprise genes encoding mRNA. These genes can be grouped into operons and transcribed as a single mRNA (Fig. 1). Synthesis of this mRNA initiates in intergenic DNA adjacent to the operon and is controlled by a regulatory protein. Hence, despite accounting for only a fraction of the genome, intergenic DNA has been studied intensely (Keseler *et al*., 2013). Consequently, over many decades, rules defining transcription initiation, and its control by regulatory proteins, have been defined (Browning & Busby, 2004). In contrast, regions encoding mRNA have been regarded as inert with respect to transcription initiation and its control.

Technical advances now permit unbiased study of transcription and its control on a genome-wide scale (Wade & Grainger, 2014). As expected, such work confirms that mRNA synthesis is indeed subject to regulation at intergenic regions. This is also true for genes encoding untranslated tRNA and rRNA species. However, hidden layers of complexity, superimposed upon expected transcriptional events, are also evident. Hence, many genes contain internal transcription start sites, antisense transcription is pervasive, and DNA binding by gene regulatory proteins is not restricted to intergenic regions (Wade & Grainger, 2014). In this paper, I will outline how the operon model came to dominate opinion and, in light of recent observations, argue that simplistic genome annotation conceals the true sophistication of bacterial DNA.

Abbreviations: CRP, cAMP receptor protein; sRNA, small RNA.

Understanding genes and their regulation: the operon model

The operon model rose to prominence in the early 1960s (Jacob *et al*., 1960; Jacob & Monod, 1961). The concept, which describes a group of genes, under the transcriptional control of a regulatory protein, transformed our understanding of gene expression (Fig. 1). Consequently, the terms promoter (a DNA sequence that stimulates transcription initiation), regulator (a protein that modulates promoter activity) and operator (a DNA binding target for a regulator) entered the scientific lexicon (Jacob & Monod, 1959; Jacob *et al*., 1964; Cohen & Jacob, 1959). Whilst the work of Jacob and Monod provided a conceptual framework to describe genes and their control, many decades elapsed before the underlying molecular events were understood (Browning & Busby, 2004). In this regard, the ability to define nucleic acid sequences was a major breakthrough (Wu, 1972; Gilbert & Maxam, 1973; Sanger *et al*., 1977). Hence, similarities in DNA sequence between regulatory regions were identified (Dickson *et al*., 1975; Maniatis *et al*.,

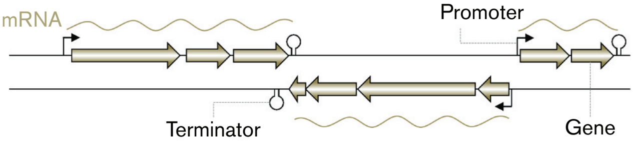

Fig. 1. The operon model. In bacteria, many genes encoding mRNA (block arrows) are organized into operons. These are transcription units demarked by a promoter (bent line arrow) and a terminator ('lollipop') of transcription. Thus, when transcribed, operons result in mRNA species (wavy line) of a precise length that map to the template strand of the DNA.

1975; Musso *et al.*, 1977; Smith & Schleif, 1978). In particular, it was noted that promoters share two regions of resemblance (Hawley & McClure, 1983). We now understand these as the −35 (5′-TTGACA-3′) and −10 (5′-TATAAT-3′) hexamers which interact with the housekeeping RNA polymerase (Zhang *et al.*, 2012; Zou & Steitz, 2015). In contrast, regulator binding sites have diverse sequences, but are often palindromic (Pabo & Sauer, 1984). This reflects the need to accommodate homodimeric regulatory proteins.

Application of the operon model on a genomic scale

Whilst the study of transcription initiation and regulation became focused on a few favoured intergenic regions, DNA sequencing approaches began to target genome-scale problems. A combination of cost reduction and automation underpinned publication of the *Haemophilus influenzae* genome in 1995 (Adams *et al.*, 1994; Fleischmann *et al.*, 1995). Further bacterial genome sequences followed in quick succession (Fraser *et al.*, 1995; Blattner *et al.*, 1997; Kunst *et al.*, 1997; Cole *et al.*, 1998). The availability of such sequences demanded new computational tools. In particular, it became necessary to annotate genome sequences to provide a standardized point of reference for future researchers. Annotation methods scan the DNA sequence for ORFs and cluster these into operons according to the principles of Jacob and Monod (Overbeek *et al.*, 2007). Similarly, attempts can be made to identify promoters and operators on the basis of the underlying DNA sequence (Gelfand *et al.*, 2000). Intriguingly, many researchers noted that regulatory sequences sometimes occurred inside genes (Robison *et al.*, 1998). However, such observations were routinely dismissed as unimportant (Blattner *et al.*, 1997; Li *et al.*, 2002; Madan Babu & Teichmann, 2003; Pavesi *et al.*, 2004; Wei & Yu, 2007).

Beyond the operon: pervasive transcription of bacterial chromosomes

Immobilized DNA oligonucleotides, arrayed on a solid surface, offered the first opportunity to study genes and their regulation on a genomic scale (Ramsay, 1998). Typically, such DNA microarrays were designed so that each oligonucleotide probe comprised a section of an annotated gene. Following RNA extraction and reverse transcription, cDNA hybridization to a cognate probe revealed transcript abundance. However, opportunities to detect anything other than mRNAs, rRNAs and tRNAs were limited; DNA microarrays were designed according to genome annotation. Eventually, unbiased transcript detection became possible as the resolution of DNA microarrays improved and, ultimately, the approach was superseded by massively parallel DNA sequencing (Selinger *et al.*, 2000; Grainger *et al.*, 2005; Reppas *et al.*, 2006; Sharma *et al.*, 2010; Nicolas *et al.*, 2012). Remarkably, unbiased analysis suggests that transcription of bacterial chromosomes is pervasive. This catch-all term describes RNA synthesis not constrained by the operon model (Fig. 2). Such transcripts include mRNAs with large non-coding appendages, stable non-coding RNAs (ncRNA) and unstable ncRNAs. The different types of transcript, and their modes of synthesis, are discussed below.

Transcription of stable ncRNAs

Small RNA (sRNA) species are a common class of non-coding transcript. They are stable and encoded by transcription units, typically <250 bp in length, with a defined promoter and terminator (Fig. 2). These transcription units can occur anywhere in the genome, but are found most frequently between mRNA encoding genes (Rivas *et al.*, 2001; Bak *et al.*, 2015; Rivers *et al.*, 2016). Hence, sRNA species contribute substantially to the complex patterns of transcription observed in bacteria. Prior to the widespread use of unbiased transcriptome analysis, only a handful of sRNAs had been identified (Storz *et al.*, 2011). However, application of genomic tools has demonstrated that hundreds of sRNAs may be encoded by any given bacterial genome (Gómez-Lozano *et al.*, 2015). If oppositely orientated to a protein encoding gene, an sRNA may be co-classified as an antisense RNA (asRNA) (Fig. 2, orange label).

Transcription of extended mRNAs

Recent transcriptome analyses have identified many mRNAs with large non-coding appendages (Sesto *et al.*, 2013; Conway *et al.*, 2014). In some cases, these RNAs act simply as templates for protein synthesis (Brown *et al.*, 2014). In other instances, the transcript may act as both an mRNA and a regulatory RNA (Sesto *et al.*, 2013). For simplicity, all such transcripts will be referred to as extended mRNAs. Often, the non-coding segment of the RNA is antisense with respect to adjacent genes. For example, 75 % of convergent operons in *Escherichia coli* produce transcripts with 3′ ends overlapping by a mean of 286 nt (Fig. 2, maroon label; Conway *et al.*, 2014). Similarly, 35 % of divergent operons generate transcripts with overlapping 5′ ends (Fig. 2, green label; Conway *et al.*, 2014). Non-coding

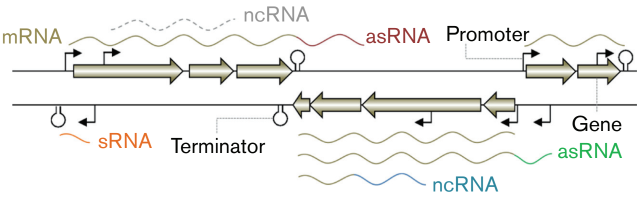

Fig. 2. Hidden layers of transcriptional complexity. Many RNA species are synthesized outside of the constraints imposed by the operon model (Wade & Grainger, 2014). Hence, small RNA (sRNA) may be derived from coding DNA sequences, and mRNAs with antisense (asRNA) or other non-coding (ncRNA) appendages are widely observed. If untranslated, the RNA is likely to be unstable (wavy grey dashed line).

sense extensions at the 5′ end of an mRNA can result from promoters within genes (Fig. 2, blue label). For example, a promoter within the *E. coli rlmD* gene results in the production of an mRNA for the adjacent *relA* with a non-coding 667 nt extension at the 5′ end (Brown *et al*., 2014; Bonocora *et al*., 2015). Note that non-coding sense extensions at the 3′ end of an mRNA are unlikely. For example, if RNA polymerase were to progress past a termination element, and transcribe a downstream gene in the correct orientation, a polycistronic mRNA would be produced.

Transcription of unstable intragenic RNAs

In *E. coli*, hundreds of intragenic promoters also drive RNA production independently of mRNA synthesis (Dornenburg *et al*., 2010; Singh *et al*., 2014). These promoter sequences are indistinguishable from those found upstream of protein encoding genes, but are not associated with a canonical transcription unit (Fig. 2; grey label) (Hawley & McClure, 1983; Dornenburg *et al*., 2010). Hence, the transcripts that are produced are likely to be untranslated, unstable and rapidly terminated (Iost & Dreyfus, 1995; Wade & Grainger, 2014). Transcription of this type may be antisense with respect to overlapping genes, but sense transcription is more common (Singh *et al*., 2014).

Functions and consequences of pervasive transcription

Arguably, sRNA species are the best candidates for regulatory function; they tend to be stable and structured (Storz *et al*., 2011). Hence, an individual sRNA may interact with a target protein or base pair with numerous mRNAs (Storz *et al*., 2011). These interactions can be regulatory. For example, an sRNA may control stability, translation or termination of an mRNA (Storz *et al*., 2011). Regulation can also be a feature of antisense transcription. For example, expression of *rplJ* is downregulated by an overlapping antisense transcript in *E. coli* (Dornenburg *et al*., 2010). Similar effects can also be associated with those mRNAs that have a large antisense appendage (Sesto *et al*., 2013).

The production of unstable transcripts is poorly conserved and may represent transcriptional noise (Raghavan *et al*., 2012; Wade & Grainger, 2014). Interestingly, such transcripts align frequently with horizontally acquired sections of bacterial genomes where multiple mechanisms act to reduce their synthesis (Chintakayala *et al*., 2013; Singh & Grainger, 2013; Singh *et al*., 2014). For example, the histone-like nucleoid structuring (H-NS) protein hinders intragenic transcription initiation and elongation (Peters *et al*., 2012; Singh *et al*., 2014), the Rho factor stimulates transcription termination (Cardinale *et al*., 2008) and RNAses can degrade any transcripts that are synthesized (Durand *et al*., 2012). Horizontally acquired DNA suffers disproportionately from intragenic transcription initiation because of its high AT content (Singh *et al*., 2014). Bacterial promoters are also AT-rich DNA and occur frequently by chance within such genes (Hawley & McClure, 1983; Landick *et al*., 2015).

Complex patterns of transcription factor binding

The first gene regulatory proteins identified were shown to bind loci close to the 5′ end of known operons (Ptashne, 1967; Gilbert & Müller-Hill, 1967). Subsequent studies focused on such DNA targets and found further regulatory interactions (Schleif, 1969; Hua & Markovitz, 1975; Webster *et al*., 1987). These observations reinforced the original view that regulators principally target such regions. This circular reasoning led to a dogmatic application of early observations with many researchers excluding the possibility that gene regulatory proteins may bind elsewhere (i.e. within genes or close to the 3′ end of a gene) (Li *et al*., 2002; Madan Babu & Teichmann, 2003; Pavesi *et al*., 2004; Wei & Yu, 2007). Recent unbiased studies of regulator–DNA interactions show that whilst many regulators behave in accordance with dogma (Grainger *et al*., 2004; Yamamoto *et al*., 2011) others deviate substantially from expected behaviour (Wade *et al*., 2007). Thus, some regulators primarily bind targets within genes (Shimada *et al*., 2008), whilst others bind a combination of mRNA encoding and regulatory DNA (Grainger *et al*., 2006; Efromovich *et al*., 2008). Indeed, a recent study of 154 transcription factors in *Mycobacterium tuberculosis* showed a continuum of binding; the number of intergenic targets for a given regulator varied from 0 to 100 % of all targets (Minch *et al*., 2015). Since interactions with coding DNA have been overlooked we have little understanding of these targets. Recent examples of how transcription factor binding in unexpected locations can influence transcription are provided below.

Regulation of a promoter for an overlapping gene

The cAMP receptor protein (CRP) is a global regulator of transcription found in many bacteria (Green *et al*., 2014). Mapping of CRP binding in *E. coli* has identified numerous binding targets within genes or between convergent genes (Grainger *et al*., 2005; Haycocks *et al*., 2015). One such example, a target between the convergent *aatC* and *tnpA* genes in enterotoxigenic *E. coli*, has recently been examined (Haycocks & Grainger, 2016). The analysis revealed that CRP bound at this site activates the transcription of a small unannotated gene completely embedded within, and in the opposite orientation to, *aatC* (Haycocks & Grainger, 2016). Thus, whilst the position of CRP binding appears unusual, an unannotated gene is correctly positioned for regulation (Fig. 3a). Presumably, other intragenic regulator binding sites will have a similar function.

Regulation of a distal promoter

The pyrimidine utilization repressor (RutR) binds at least 20 different DNA targets across the *E. coli* genome (Shimada *et al*. 2008). The consensus RutR operator is a perfect palindrome and purified RutR binds tightly to its DNA targets *in vitro*. Of the 20 RutR operators, 16 are located within genes. Initial inspection detected RutR-mediated repression at only one such target; expression of *ves* was undetectable in the presence of RutR (Shimada *et al.* 2008). However, subsequent work has shown that RutR activity is controlled by deacetylation and autoproteolysis (Tu *et al.*, 2015). Thus, under appropriate conditions, repression of further RutR target genes is apparent. Therefore, repression of a distal upstream promoter can explain binding of regulators in some instances (Fig. 3b).

Regulation of a promoter within a defined operon

The *M. tuberculosis* genes Rv0250c and Rv0249c can be co-transcribed as part of an operon (Knapp *et al.*, 2015). Global analysis of CRP binding across the *M. tuberculosis* chromosome identified CRP binding at the 3′ end of Rv0250c. It was subsequently shown that CRP bound at this locus activates a promoter located between Rv0250c and Rv0249c (Fig. 3c). Hence, in some instances, promoters within operons require the binding of transcription factors to coding DNA. This phenomenon is likely to be common; Conway and co-workers recently noted that 36 % of *E. coli* operons contain internal promoters or terminators (Conway *et al.*, 2014).

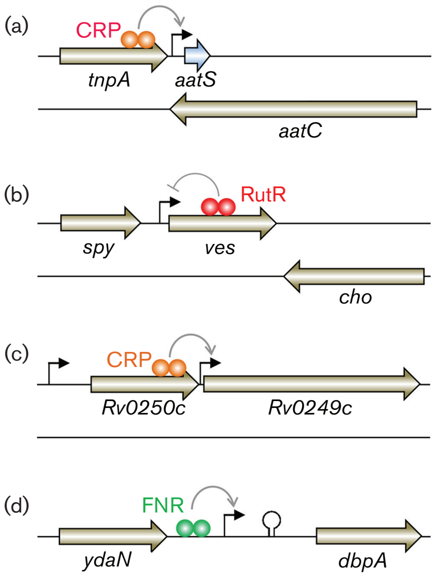

Fig. 3. Binding of transcription factors within genes and operons. Complex transcriptional events are associated with the binding of regulatory factors (coloured spheres) in unexpected locations. (a) In enterotoxigenic *E. coli*, the transcription factor CRP activates a promoter (bent arrow) between convergent genes (beige block arrows) to control expression of an unannotated gene (blue block arrow) (Haycocks & Grainger, 2016). (b) In *E. coli* K-12, the transcriptional repressor RutR binds within genes and prevents their expression (Shimada *et al.*, 2008; Tu *et al.*, 2015). (c) An *M. tuberculosis* operon contains an internal promoter that is activated by CRP (Knapp *et al.*, 2015). (d) The sRNA FnrS is encoded in the *dbpA* regulatory region. Binding of FNR to this regulatory region controls expression of *fnrS* but not *dbpA* expression (Grainger *et al.*, 2007; Boysen *et al.*, 2010; Durand & Storz, 2010).

Regulation of sRNA expression

Understanding sRNA regulation can also reveal hidden functions for transcription factor binding. The global regulator FNR binds upstream of the mRNA encoding gene *dbpA* in *E. coli*, but does not control its expression (Grainger *et al.*, 2007). Instead, FNR activates the expression of FnrS, an sRNA encoded within the *dbpA* regulatory region (Boysen *et al.*, 2010; Durand & Storz, 2010). Hence, apparently cryptic regulator binding can be associated with the control of sRNA expression (Fig. 3d). In some cases, an sRNA may overlap the 3′ untranslated region of an mRNA (Chao *et al.*, 2012). Hence, at such loci, control of sRNA transcription likely involves intragenic regulator binding.

Conclusions

There is an overwhelming body of evidence, generated using multiple independent experimental approaches, that bacterial genomes are more complex than implied by either the operon model or genome annotation. On consideration, it was perhaps naive to expect that bacteria would conform entirely to our expectations. Natural selection, operating over an incomprehensible time scale, has eked out adaptations that enhance the fitness of bacterial cells. We should not be surprised if such adaptations allow bacteria to obtain ‘added value’ from their small genomes.

Acknowledgements

Work in my laboratory is currently funded by the BBSRC (grants BB/N014200/1, BB/N005961/1 and BB/J006076/1), the Leverhulme Trust (RPG-2013-147) and the Human Frontiers Science Program (RGP0014/2014). I should like to thank Joseph Wade for helpful comments on this manuscript and all colleagues, past and present, for their support.

References

Adams, M. D., Fields, C. & Venter, J. C. (1994). *Automated DNA Sequencing and Analysis*. San Diego, CA: Academic Press.

Bak, G., Lee, J., Suk, S., Kim, D., Young Lee, J., Kim, K. S., Choi, B. S. & Lee, Y. (2015). Identification of novel sRNAs involved in biofilm formation, motility, and fimbriae formation in *Escherichia coli*. *Sci Rep* **5**, 15287.

Blattner, F. R., Plunkett, G., Bloch, C. A., Perna, N. T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J. D., Rode, C. K. & other authors (1997). The complete genome sequence of *Escherichia coli* K-12. *Science* **277**, 1453–1462.

Bonocora, R. P., Smith, C., Lapierre, P. & Wade, J. T. (2015). Genome-scale mapping of *Escherichia coli* σ54 reveals widespread, conserved intragenic binding. *PLoS Genet* **11**, e1005552.

Boysen, A., Møller-Jensen, J., Kallipolitis, B., Valentin-Hansen, P. & Overgaard, M. (2010). Translational regulation of gene expression by an anaerobically induced small non-coding RNA in *Escherichia coli*. *J Biol Chem* **285**, 10690–10702.

Brown, D. R., Barton, G., Pan, Z., Buck, M. & Wigneshweraraj, S. (2014). Nitrogen stress response and stringent response are coupled in *Escherichia coli*. *Nat Commun* **5**, 4115.

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

Cardinale, C. J., Washburn, R. S., Tadigotla, V. R., Brown, L. M., Gottesman, M. E. & Nudler, E. (2008). Termination factor Rho and its cofactors NusA and NusG silence foreign DNA in *E. coli*. *Science* **320**, 935–938.

Chao, Y., Papenfort, K., Reinhardt, R., Sharma, C. M. & Vogel, J. (2012). An atlas of Hfq-bound transcripts reveals 3′ UTRs as a genomic reservoir of regulatory small RNAs. *EMBO J* **31**, 4005–4019.

Chintakayala, K., Singh, S. S., Rossiter, A. E., Shahapure, R., Dame, R. T. & Grainger, D. C. (2013). *E. coli* Fis protein insulates the *cbpA* gene from uncontrolled transcription. *PLoS Genet* **9**, e1003152.

Cohen, G. & Jacob, F. (1959). Sur la repression de la synthese des enzymes intervenant dans la formation du tryptophane chez *Escherichia coli*. *C R Acad Sci Paris* **248**, 3490–3492.

Cole, S. T., Brosch, R., Parkhill, J., Garnier, T., Churcher, C., Harris, D., Gordon, S. V., Eiglmeier, K., Gas, S. & other authors (1998). Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* **393**, 537–544.

Conway, T., Creecy, J. P., Maddox, S. M., Grissom, J. E., Conkle, T. L., Shadid, T. M., Teramoto, J., San Miguel, P., Shimada, T. & other authors (2014). Unprecedented high-resolution view of bacterial operon architecture revealed by RNA sequencing. *MBio* **5**, e01442-14.

Dickson, R. C., Abelson, J., Barnes, W. M. & Reznikoff, W. S. (1975). Genetic regulation: the Lac control region. *Science* **187**, 27–35.

Dornenburg, J. E., Devita, A. M., Palumbo, M. J. & Wade, J. T. (2010). Widespread antisense transcription in *Escherichia coli*. *MBio* **1**, e00024-10.

Durand, S. & Storz, G. (2010). Reprogramming of anaerobic metabolism by the FnrS small RNA. *Mol Microbiol* **75**, 1215–1231.

Durand, S., Gilet, L. & Condon, C. (2012). The essential function of *B. subtilis* RNase III is to silence foreign toxin genes. *PLoS Genet* **8**, e1003181.

Efromovich, S., Grainger, D., Bodenmiller, D. & Spiro, S. (2008). Genome-wide identification of binding sites for the nitric oxide-sensitive transcriptional regulator NsrR. *Methods Enzymol* **437**, 211–233.

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J. F., Dougherty, B. A. & other authors (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269**, 496–512.

Fraser, C. M., Gocayne, J. D., White, O., Adams, M. D., Clayton, R. A., Fleischmann, R. D., Bult, C. J., Kerlavage, A. R., Sutton, G. & other authors (1995). The minimal gene complement of *Mycoplasma genitalium*. *Science* **270**, 397–403.

Gelfand, M. S., Novichkov, P. S., Novichkova, E. S. & Mironov, A. A. (2000). Comparative analysis of regulatory patterns in bacterial genomes. *Brief Bioinform* **1**, 357–371.

Gilbert, W. & Maxam, A. (1973). The nucleotide sequence of the *lac* operator. *Proc Natl Acad Sci U S A* **70**, 3581–3584.

Gilbert, W. & Müller-Hill, B. (1967). The *lac* operator is DNA. *Proc Natl Acad Sci U S A* **58**, 2415–2421.

Grainger, D. C., Hurd, D., Harrison, M., Holdstock, J. & Busby, S. J. (2005). Studies of the distribution of *Escherichia coli* cAMP-receptor protein and RNA polymerase along the *E. coli* chromosome. *Proc Natl Acad Sci U S A* **102**, 17693–17698.

Grainger, D. C., Hurd, D., Goldberg, M. D. & Busby, S. J. (2006). Association of nucleoid proteins with coding and non-coding segments of the Escherichia coli genome. *Nucleic Acids Res* **34**, 4642–4652.

Grainger, D. C., Aiba, H., Hurd, D., Browning, D. F. & Busby, S. J. (2007). Transcription factor distribution in *Escherichia coli*: studies with FNR protein. *Nucleic Acids Res* **35**, 269–278.

Grainger, D. C., Overton, T. W., Reppas, N., Wade, J. T., Tamai, E., Hobman, J. L., Constantinidou, C., Struhl, K., Church, G. & other authors (2004). Genomic studies with *Escherichia coli* MelR protein: applications of chromatin immunoprecipitation and microarrays. *J Bacteriol* **186**, 6938–6943.

Green, J., Stapleton, M. R., Smith, L. J., Artymiuk, P. J., Kahramanoglou, C., Hunt, D. M. & Buxton, R. S. (2014). Cyclic-AMP and bacterial cyclic-AMP receptor proteins revisited: adaptation for different ecological niches. *Curr Opin Microbiol* **18**, 1–7.

Gómez-Lozano, M., Marvig, R. L., Molina-Santiago, C., Tribelli, P. M., Ramos, J. L. & Molin, S. (2015). Diversity of small RNAs expressed in *Pseudomonas* species. *Environ Microbiol Rep* **7**, 227–236.

Hawley, D. K. & McClure, W. R. (1983). Compilation and analysis of *Escherichia coli* promoter DNA sequences. *Nucleic Acids Res* **11**, 2237–2255.

Haycocks, J. R. J. & Grainger, D. C. (2016). Unusually situated binding sites for bacterial transcription factors can have hidden functionality. *PLoS One* **11**, e0157016.

Haycocks, J. R., Sharma, P., Stringer, A. M., Wade, J. T. & Grainger, D. C. (2015). The molecular basis for control of ETEC enterotoxin expression in response to environment and host. *PLoS Pathog* **11**, e1004605.

Hua, S. S. & Markovitz, A. (1975). Regulation of galactose operon at the gal operator-promoter region in *Escherichia coli* K-12. *J Bacteriol* **122**, 510–517.

Iost, I. & Dreyfus, M. (1995). The stability of *Escherichia coli lacZ* mRNA depends upon the simultaneity of its synthesis and translation. *EMBO J* **14**, 3252–3261.

Jacob, F. & Monod, J. (1959). Gènes de structure et gènes de regulation dans la biosynthèse des proteins. *C R Acad Sci Paris* **249**, 1282–1284.

Jacob, F. & Monod, J. (1961). Genetic regulatory mechanisms in the synthesis of proteins. *J Mol Biol* **3**, 318–356.

Jacob, F., Perrin, D., Sánchez, C. & Monod, J. (1960). L'opéron: groupe de gènes à expression coordonnée par un opérateur. *C R Acad Sci Paris* **250**, 1727–1729.

Jacob, F., Ullman, A., Sánchez, C. & Monod, J. (1964). Le promoteur, élément génétique necessaire à l' expression d' un opéron. *C R Acad Sci Paris* **258**, 3125–3128.

Keseler, I. M., Mackie, A., Peralta-Gil, M., Santos-Zavaleta, A., Gama-Castro, S., Bonavides-Martínez, C., Fulcher, C., Huerta, A. M., Kothari, A. & other authors (2013). EcoCyc: fusing model organism databases with systems biology. *Nucleic Acids Res* **41**, D605–612.

Knapp, G. S., Lyubetskaya, A., Peterson, M. W., Gomes, A. L., Ma, Z., Galagan, J. E. & McDonough, K. A. (2015). Role of intragenic binding of cAMP responsive protein (CRP) in regulation of the succinate dehydrogenase genes Rv0249c-Rv0247c in TB complex mycobacteria. *Nucleic Acids Res* **43**, 5377–5393.

Kunst, F., Ogasawara, N., Moszer, I., Albertini, A. M., Alloni, G., Azevedo, V., Bertero, M. G., Bessières, P., Bolotin, A. & other authors (1997). The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*. *Nature* **390**, 249–256.

Landick, R., Wade, J. T. & Grainger, D. C. (2015). H-NS and RNA polymerase: a love-hate relationship? *Curr Opin Microbiol* **24**, 53–59.

Li, H., Rhodius, V., Gross, C. & Siggia, E. D. (2002). Identification of the binding sites of regulatory proteins in bacterial genomes. *Proc Natl Acad Sci U S A* **99**, 11772–11777.

Madan Babu, M. & Teichmann, S. A. (2003). Functional determinants of transcription factors in *Escherichia coli*: protein families and binding sites. *Trends Genet* **19**, 75–79.

Maniatis, T., Ptashne, M., Backman, K., Kield, D., Flashman, S., Jeffrey, A. & Maurer, R. (1975). Recognition sequences of repressor and polymerase in the operators of bacteriophage lambda. *Cell* **5**, 109–113.

Minch, K. J., Rustad, T. R., Peterson, E. J., Winkler, J., Reiss, D. J., Ma, S., Hickey, M., Brabant, W., Morrison, B. & other authors (2015). The DNA-binding network of *Mycobacterium tuberculosis*. *Nat Commun* **6**, 5829.

Musso, R., Di Lauro, R., Rosenberg, M. & de Crombrugghe, B. (1977). Nucleotide sequence of the operator-promoter region of the galactose operon of *Escherichia coli*. *Proc Natl Acad Sci U S A* **74**, 106–110.

Nicolas, P., Mäder, U., Dervyn, E., Rochat, T., Leduc, A., Pigeonneau, N., Bidnenko, E., Marchadier, E., Hoebeke, M. & other authors (2012). Condition-dependent transcriptome reveals high-level regulatory architecture in *Bacillus subtilis*. *Science* **335**, 1103–1106.

Overbeek, R., Bartels, D., Vonstein, V. & Meyer, F. (2007). Annotation of bacterial and archaeal genomes: improving accuracy and consistency. *Chem Rev* **107**, 3431–3447.

Pabo, C. O. & Sauer, R. T. (1984). Protein-DNA recognition. *Annu Rev Biochem* **53**, 293–321.

Pavesi, G., Mauri, G. & Pesole, G. (2004). *In silico* representation and discovery of transcription factor binding sites. *Brief Bioinform* **5**, 217–236.

Peters, J. M., Mooney, R. A., Grass, J. A., Jessen, E. D., Tran, F. & Landick, R. (2012). Rho and NusG suppress pervasive antisense transcription in *Escherichia coli*. *Genes Dev* **26**, 2621–2633.

Ptashne, M. (1967). Specific binding of the lambda phage repressor to lambda DNA. *Nature* **214**, 232–234.

Raghavan, R., Sloan, D. B. & Ochman, H. (2012). Antisense transcription is pervasive but rarely conserved in enteric bacteria. *MBio* **3**, e00156-12.

Ramsay, G. (1998). DNA chips: state-of-the art. *Nat Biotechnol* **16**, 40–44.

Reppas, N. B., Wade, J. T., Church, G. M. & Struhl, K. (2006). The transition between transcriptional initiation and elongation in *E. coli* is highly variable and often rate limiting. *Mol Cell* **24**, 747–757.

Rivas, E., Klein, R. J., Jones, T. A. & Eddy, S. R. (2001). Computational identification of noncoding RNAs in *E. coli* by comparative genomics. *Curr Biol* **11**, 1369–1373.

Rivers, A. R., Burns, A. S., Chan, L. K. & Moran, M. A. (2016). Experimental identification of small non-coding RNAs in the model marine bacterium *Ruegeria pomeroyi* DSS-3. *Front Microbiol* **7**, 380.

Robison, K., McGuire, A. M. & Church, G. M. (1998). A comprehensive library of DNA-binding site matrices for 55 proteins applied to the complete *Escherichia coli* K-12 genome. *J Mol Biol* **284**, 241–254.

Sanger, F., Nicklen, S. & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* **74**, 5463–5467.

Schleif, R. (1969). Induction of the L-arabinose operon. *J Mol Biol* **46**, 197–199.

Selinger, D. W., Cheung, K. J., Mei, R., Johansson, E. M., Richmond, C. S., Blattner, F. R., Lockhart, D. J. & Church, G. M. (2000). RNA expression analysis using a 30 base pair resolution *Escherichia coli* genome array. *Nat Biotechnol* **18**, 1262–1268.

Sesto, N., Wurtzel, O., Archambaud, C., Sorek, R. & Cossart, P. (2013). The excludon: a new concept in bacterial antisense RNA-mediated gene regulation. *Nat Rev Microbiol* **11**, 75–82.

Sharma, C. M., Hoffmann, S., Darfeuille, F., Reignier, J., Findeiss, S., Sittka, A., Chabas, S., Reiche, K., Hackermüller, J. & other authors (2010). The primary transcriptome of the major human pathogen *Helicobacter pylori*. *Nature* **464**, 250–255.

Shimada, T., Ishihama, A., Busby, S. J. & Grainger, D. C. (2008). The *Escherichia coli* RutR transcription factor binds at targets within genes as well as intergenic regions. *Nucleic Acids Res* **36**, 3950–3955.

Singh, S. S. & Grainger, D. C. (2013). H-NS can facilitate specific DNA-binding by RNA polymerase in AT-rich gene regulatory regions. *PLoS Genet* **9**, e1003589.

Singh, S. S., Singh, N., Bonocora, R. P., Fitzgerald, D. M., Wade, J. T. & Grainger, D. C. (2014). Widespread suppression of intragenic transcription initiation by H-NS. *Genes Dev* **28**, 214–219.

Smith, B. R. & Schleif, R. (1978). Nucleotide sequence of the L-arabinose regulatory region of *Escherichia coli* K12. *J Biol Chem* **253**, 6931–6933.

Storz, G., Vogel, J. & Wassarman, K. M. (2011). Regulation by small RNAs in bacteria: expanding frontiers. *Mol Cell* **43**, 880–891.

Tu, S., Guo, S. J., Chen, C. S., Liu, C. X., Jiang, H. W., Ge, F., Deng, J. Y., Zhou, Y. M., Czajkowsky, D. M. & other authors (2015). YcgC represents a new protein deacetylase family in prokaryotes. *Elife* **4**, e05322.

Wade, J. T. & Grainger, D. C. (2014). Pervasive transcription: illuminating the dark matter of bacterial transcriptomes. *Nat Rev Microbiol* **12**, 647–653.

Wade, J. T., Struhl, K., Busby, S. J. & Grainger, D. C. (2007). Genomic analysis of protein-–DNA interactions in bacteria: insights into transcription and chromosome organization. *Mol Microbiol* **65**, 21–26.

Webster, C., Kempsell, K., Booth, I. & Busby, S. (1987). Organisation of the regulatory region of the *Escherichia coli* melibiose operon. *Gene* **59**, 253–263.

Wei, W. & Yu, X. D. (2007). Comparative analysis of regulatory motif discovery tools for transcription factor binding sites. *Genomics Proteomics Bioinformatics* **5**, 131–142.

Wu, R. A. Y (1972). Nucleotide sequence analysis of DNA. *Nature* **236**, 198–200.

Yamamoto, K., Ishihama, A., Busby, S. J. & Grainger, D. C. (2011). The *Escherichia coli* K-12 MntR miniregulon includes *dps*, which encodes the major stationary-phase DNA-binding protein. *J Bacteriol* **193**, 1477–1480.

Zhang, Y., Feng, Y., Chatterjee, S., Tuske, S., Ho, M. X., Arnold, E. & Ebright, R. H. (2012). Structural basis of transcription initiation. *Science* **338**, 1076–1080.

Zuo, Y. & Steitz, T. A. (2015). Crystal structures of the *E. coli* transcription initiation complexes with a complete bubble. *Mol Cell* **58**, 534–540.

Edited by: T. Msadek