Molecular BioSystems

REVIEW

Cite this: *Mol. BioSyst.*, 2017, **13**, 665

Consensus architecture of promoters and transcription units in *Escherichia coli*: design principles for synthetic biology†

Cynthia Rangel-Chavez, Edgardo Galan-Vasquez and Agustino Martinez-Antonio*

Received 22nd November 2016,
Accepted 17th February 2017

DOI: 10.1039/c6mb00789a

rsc.li/molecular-biosystems

Genetic information in genomes is ordered, arranged in such a way that it constitutes a code, the so-called *cis* regulatory code. The regulatory machinery of the cell, termed *trans*-factors, decodes and expresses this information. In this way, genomes maintain a potential repertoire of genetic programs, parts of which are executed depending on the presence of active regulators in each condition. These genetic programs, executed by the regulatory machinery, have functional units in the genome delimited by punctuation-like marks. In genetic terms, these informational phrases correspond to transcription units, which are the minimal genetic information expressed consistently from initiation to termination marks. Between the start and final punctuation marks, additional marks are present that are read by the transcriptional and translational machineries. In this work, we look at all the experimentally described and predicted genetic elements in the bacterium *Escherichia coli* K-12 MG1655 and define a comprehensive architectural organization of transcription units to reveal the natural genome-design and to guide the construction of synthetic genetic programs.

Introduction

With the sequences of complete genomes available, we can study the arrangement of genetic information. The genome of the bacterium *E. coli* strain K-12 MG1655 was one of the first to be sequenced and there has been continuous refinement of gene annotations since then.[1] At present, this bacterium is the model organism for genetic, molecular, and biochemical studies. Therefore, it is not surprising that the molecular functions of this bacterium are likely known in more detail than for any other organism. Indeed, one of the most comprehensive electronic encyclopedias for gene regulation is RegulonDB.[2] It contains documented information, experimentally supported, about genes and their regulation in *E. coli*, and particularly in the K-12 MG1655 strain. Likewise, scientific and technological progress has increased the usage of technological approaches such as synthetic biology, an interdisciplinary framework aimed at biological engineering acceleration.[3,4] At the core of synthetic biology are the functional genetic bioparts used as building blocks to construct biological devices, circuits, and systems for specific purposes (http://parts.igem.org).[5] To date, engineering approaches to construct genetic circuits have been subject to much trial and error, particularly regarding the composition of and distance between genetic elements. These elements include genes, promoters, enhancers, ribosome-binding sites (RBS), operator regions, and translation and transcription terminators.[6–8]

For this work, we performed a global survey on the genetic architecture of transcription units (TUs) in the *E. coli strain* K-12 MG1655 to propose a consensus architectural model for the construction of genetic circuits. We hope this study might contribute to the design of genetic circuits in synthetic biology and biological engineering.

Biological Engineering Laboratory, Genetic Engineering Department, Center for Research and Advanced Studies of the National Polytechnic Institute (Cinvestav), Campus Irapuato, Km. 9.6 Libramiento Norte Carr, Irapuato-León 36821, Irapuato Gto, Mexico. E-mail: agustino.martinez@cinvestav.mx

† Electronic supplementary information (ESI) available: Table S1. Genetic elements in the natural TUs of *Escherichia coli K-12* MG1655. Table S2. Genetic elements in strain K-12 W3110 and BL21 (DE3) of *E. coli*, *Salmonella typhimurium* SL1344, and *Pseudomonas aeruginosa* PA14 and PAO1 strains. Table S3. Gene Sequences of K-12 W3110 and BL21 (DE3) *E. coli*, *Salmonella typhimurium* SL1344, and *Pseudomonas aeruginosa* PA14 and PAO1 strains considered in this study. Fig. S1. Architecture of genetic elements in other bacteria considered in this study. See DOI: 10.1039/c6mb00789a

Experimental

Genetic data

Annotated operons and their genetic components were obtained from RegulonDB version 9.0.[2] From this database, we also obtained data on all the genetic elements that function as signaling marks for the transcriptional and translational machinery in *E. coli*: *e.g.* promoters, RBS, stop codons, *etc.* (Table 1). All the

Table 1 Genetic elements of transcription units

Genetic element	Definition	*E. coli* representative (strain K-12 MG1655)	Annotated in RegulonDB
Transcription unit	A sequence of nucleotides that encodes for a single RNA molecule. It includes informative signals for the cellular transcriptional apparatus such as the promoter, initiator, and terminator of transcription.[9]	The average size of a mRNA molecule is 2 kb ribonucleotides; mRNAs range from 48 (*greA*) to 15331 (*nuoABCEFGHIJKLMN*) ribonucleotides.	3549
Promoter	A DNA sequence to which RNA polymerase binds to initiate transcription. It dictates the direction of transcription and which of the two DNA strands should be read as the template.[10]	There are seven sigma factors from two families; the σ^{70} family (six members: σ^{19}, σ^{24}, σ^{28}, σ^{32}, σ^{38}, and σ^{70}) and σ^{54} (1 member). Only 36.38% of TUs have been assigned a sigma factor.	8597
Transcription start site (TSS)	A nucleotide that is the first transcribed nucleotide on RNA.[11]	Normally it is adenine (A, 43.88%) or guanine (G 24.31%).	8500
Terminator of transcription	The place where RNA polymerase dissociates from DNA.[12,13]	There are two classes of transcription termination signals: (i) intrinsic terminators, composed of a G + C-rich stem-loop (28 nt average) followed by a series of U residues, and (ii) Rho-dependent terminators, the activity of which relies on binding of Rho protein to a *rut* (Rho utilization) site on the nascent transcript, followed by interaction with RNA polymerase (RNAP).	279

annotated genetic elements were ordered in tables along with their positions in the genome and the information of the operons to which they pertain (Table S1, ESI†).

Data management

With the genetic position of elements in hand, the nucleotide distances between each pair of contiguous genetic elements on the linear DNA strand were calculated. The calculated distances correspond to the number of nucleotides between each pair of transcriptional or translational elements. We used the following rule: {distance = Elem2_start − (Elem1_finish + 1)}; where Elem1 and Elem2 represent the elements and the number identifies the order in which they appear in the genome. From these distances, a distribution was computed for each pair of annotated elements.[14]

The nucleotide consensuses were searched for each type of genetic element as well as for intragenic regions by using the RSA-tools consensus software.[15] Finally, we determined the efficiency of the *E. coli* terminator with the program TTEC (Transcriptional Terminator Efficiency Calculator) (http://www.terminatorefficiency.com/).

Results & discussion

Brief description of the organization of operons in *E. coli*

As described above, *E. coli* is the organism best studied as well as the main chassis for synthetic biology. There are currently genomes sequenced from 279 strains of *E. coli*, which vary in genome size from 3.9762 Mb to 5.8662 Mb (www.ncbi.nlm.nih.gov/genome/genomes). Despite the variability in nucleotide size among strains of *E. coli*, previous works have shown that the general structure between genetic elements in the genomes is conserved, for instance, by having a similar gene organization in operons.[16,17] There are many databases for *E. coli*, but the most complete database about transcription and related processes is RegulonDB; for this study we used Version 9.0 of this database, which describes the K-12 MG1655 strain with 4650 genes.[2] From these genes, 4526 are annotated in 2634 operons, and

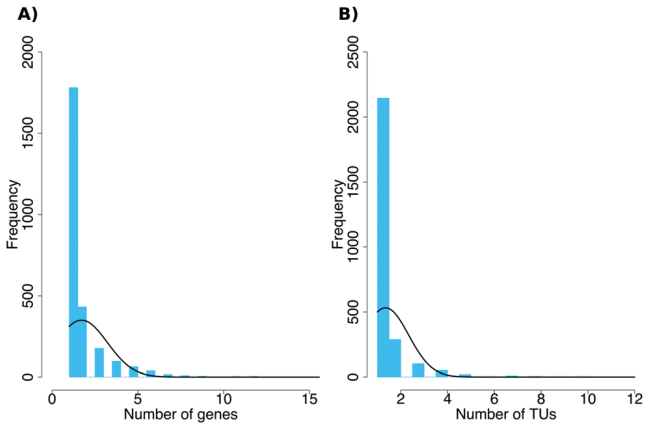

Fig. 1 Distribution of genes in the genome of *E. coli*. (A) Distribution of genes in operons; the greater proportion of operons contains only one gene (1782), the largest operon has 16 genes. (B) Distribution of TUs in operons; most of the operons contain only one TU, the largest operon has 12 TUs. A normal curve was drawn to fit each distribution.

67.65% (1782) of these operons contain a unique gene. In other words, around 40% of *E. coli* genes are encoded individually in operons (Fig. 1A). Operons are distributed almost uniformly between the DNA strands, with 1312 on the forward strand and 1322 on the reverse.

An operon can contain more than one TU, if we consider that different promoters can start and stop transcription at the same or at different termini. In *E. coli*, most operons have one TU (2146 cases, 81%), whereas the rest have more than one TU. The extreme case is one single operon (see below), which contains up to 12 TUs (Fig. 1B). If we dissect the anatomy of TUs encoded into operons, we can distinguish cases in which multiple starts of transcription finish at the same terminus of transcription (283 TUs), and conversely, there could exist a unique start of transcription that finishes in more than one terminus of transcription (115 TUs). Delimitations of TUs are defined experimentally by sequencing transcripts individually, although it is clearly not an exhaustive search as multiple

growth conditions could not be examined at the same time and therefore these numbers can increase even though proportions are maintained. TUs are more homogenous at their ends than at their beginnings, considering the diversity of genetic elements, possibly because gene regulation commonly happens at the promoter zones. The number of promoters in operons seems to be correlated with the number of encoded genes, as the average of promoters in TUs with single genes is 1 whereas in polycistronic operons it is 2. On the other hand, considering the signals for gene expression (*i.e.* operators and promoter regions recognized by transcription or sigma factors), there is a bias toward operons containing few regulatory elements since 966 (70.67%) of the operons (from 1367 with annotated promoters) have a single promoter whereas, at the other extreme, one operon contains up to 12 promoters. In the next sections, we will describe and report the *cis*-elements encoded in a TU (Table 1), and the most important *trans*-elements that interact with the TU are described in Table 2.

Genetic elements and punctuation marks for gene expression

Since our main purpose in this study is to analyze the natural architecture of functional genetic elements, we start by defining these genetic elements and the signaling marks most commonly present in TUs (Table 1). The informational marks on each TU direct the transcriptional and translational cellular activity.[27] A TU has three distinctive genetic elements: (i) a promoter region, which contains DNA-binding sites that are recognized for the RNA polymerase and other TFs; (ii) the translated message, which comprises the RNA that is translated from DNA by the RNA polymerase and includes ribosome binding sites (RBS) at the beginning of each polypeptide; (iii) the terminus of transcription, which is the sequence that marks the end of transcription; it can be of two types: Rho dependent or Rho independent (see below).[9] For information in mRNAs to be translated into polypeptides, the translational signals in mRNA should be decoded by the translational apparatus. This process starts with the binding of a ribosome at the Ribosomal Binding Site (RBS) on the mRNA. The ribosome then displaces over the first translatable codons, continues with peptide elongation, and finishes when the ribosome finds a stop codon (see below). This process is repeated for each protein encoded on a polycistronic mRNA. Furthermore, the mRNA molecule could be subject to post-transcriptional regulation such as attenuation, anti-termination, riboswitches, and by sRNAs, among others (Table 3).[34] Although all these information signals operatively could be divided into transcriptional and translational processes, all these hallmarks can be tracked at the level of DNA sequence. Therefore, we looked for all the annotated pairs of neighboring genetic elements at the level of DNA sequence in *E. coli*: *i.e.*, promoters, transcription initiations, RBS, translation initiations, *etc.*, and analyzed their distance distribution and consensus composition.

Description of the functional genetic elements or bioparts in the *E. coli* genome

Anatomy of promoters. Transcription is the process through which organisms transcribe genetic information to RNA molecules. In bacteria, a multi-subunit enzyme called RNA polymerase, which has five subunits, performs this process. Four of the subunits (α2, β, β′, and ω) constitute the core of the enzyme.[24,25,35] This core is catalytically active but it is incapable of initiating transcription by itself efficiently and specifically. For this to happen, the core must bind an additional subunit protein named sigma factor (σ), to form the RNA polymerase holoenzyme.

The σ protein was discovered in 1969 in *E. coli* as a multi-domain protein, with four domains, where domain 1, 2, 3 and 4 were described to be involved in recognition of different regions of promoters.[20,36–38] The σ has three main functions: (1) to ensure the recognition of a specific promoter sequence, (2) to position the RNA polymerase holoenzyme at a target promoter, and (3) to facilitate the unwinding of the DNA duplex near the transcription start site. A promoter is a short DNA sequence (~40 bp) that recruits the RNA polymerase to transcribe a downstream DNA region. These sequences are recognized by subunits of the RNA polymerase and other DNA-binding proteins.[20,39–41]

In *E. coli*, there are two evolutionary families of sigma, the σ^{70} and σ^{54} families (the numbers 70 and 54 correspond

Table 2 *trans*-Factors that interact with the genetic elements of transcription units

Factor	Definition	*E. coli* representative	Annotated in RegulonDB
RNA polymerase holoenzyme	Responsible for mRNA synthesis from a DNA template.[18,19]	The RNA polymerase core enzyme subunits are encoded in *E. coli* as *rpoA* (α subunit, 329 aa residues), *rpoB* (β subunit, 1342 residues), *rpoC* (β′ subunit, 1407 residues), and *rpoZ* (ω subunit, 92 residues). There are seven σ factors.	7
Transcription factors	Proteins that bind to the promoter sequence and modulate the transcription start.[20]	In *E. coli* there are ~300 predicted proteins with DNA binding site.	175
Rho factor	A protein that recognizes and binds to C-rich sequences in mRNA and promotes transcription termination.[21,22] Rho is a homohexameric protein and has RNA-dependent ATPase and helicase activities.[23,24]	Encoded as the *rho* gene (419 residues).	1
Translation complex	The machinery necessary to correctly incorporate amino acids to form a polypeptide. The bacterial translation elements comprise 2 ribosomal subunits (50S and 30S), the aminoacyl-tRNA, GTP, and initiation factors.[25,26]	There are 7 operons to encode the genes of rRNA, 53 genes encoded for ribosomal proteins and 86 for tRNA.	53

Table 3 Genetic signals for the operation of the translation apparatus

Signal	Definition	*E. coli* representative	Annotated in RegulonDB[a]
Ribosome-binding site (RBS or Shine–Dalgarno)	A specific sequence on mRNA where the ribosome binds to start the translation of codons to a polypeptide.[28–30]	It is a sequence of four ribonucleotides with the consensus AGGG.	179
Encoded gene	A locatable region of genome corresponding to a unit of inheritance, which is associated with regulatory regions, transcribed regions, and/or other functional sequence regions.[31]	The typical size is 1000 nucleotides, ranging from 30 nt (*sokA*, a pseudogene) to 7104 (*yeeJ*).	4650
Translation start	Any of the codons AUG, GUG, or UUG – where the amino acid chain starts to polymerize.[32]	Typically, it is AUG that is read as formyl-methionine.	4237[a]
Translation stop	Any of three stop codons UAG, UGA, and UAA. Their presence on mRNA causes the nascent polypeptide to be released from the ribosome since there is no tRNA for these codons.[33]	The most used is UAA (63.9%) followed by UGA (28.69%) and UAG (7.31%).	4237[a]

[a] These include two incompletely annotated proteins, which have not been localized in the genome (AlaB and Dgd).

to their molecular weights in kDa). σ^{70} constitutes six members (σ^{70}, σ^{38}, σ^{32}, σ^{28}, σ^{24} and σ^{19}) and the family takes the name of the housekeeping σ^{70} in *E. coli.* All that is known about σ^{70} is inferred to happen with the rest of the members in this family.

The most common promoters employed in engineered genetic circuits are those for σ^{70}.[10] σ^{70} promoters in *E. coli* have four identifiable elements which are named: −10, −10 extended, −35, and UP elements.[24,42,43] The −10 (TATAA) box is recognized by domain 2 of the RNA polymerase, the −35 (TTGACA) box is recognized by domain 4, and domain 3 recognizes the extended −10 (TRTG) element, when it exists. Finally, the UP element, not present in all the promoters, is contacted by the carboxy-terminal domain of the α subunit of the RNA polymerase.[43–45]

The other family of sigma factors in *E. coli* is named σ^{54} and has a single member. Unlike the σ^{70} family, σ^{54} recognizes boxes at −12 and −24 bp and normally works together with TFs that bend the DNA strands.[46] The necessity of long DNA fragments for the operation of σ^{54} might conflict with the principle of space-economy in bacteria and is a proposed cause for why this kind of sigma factor is not abundant in bacteria.[46]

In *E. coli*, these seven sigma factors are interchangeable subunits of the RNA polymerase. As expected, most genes are transcribed by the housekeeping σ^{70} (1590 genes, of 2323 with an assigned sigma) whereas σ^{19}, the smallest factor, transcribes just 5 genes of a single operon for iron metabolism (*fecC*, *fecB*, *fecE*, *fecA*, *fecD*). Additionally, in many cases, several sigma factors co-transcribe the same operon. The operons transcribed by the most sigma factors are *clpPX-lon* and *rfaD-waaFCL* each transcribed by four different sigma factors. Multiple promoters could assist the same sigma factor, possibly responding to different combinations with TFs. Following this line of description, 966 operons are transcribed from a unique promoter recognized by a unique sigma factor, 221 operons are transcribed by two sigma factors, 34 by three sigma factors and, as mentioned above, two operons are transcribed by four sigma factors. There are no operons transcribed by more than four sigma factors.

In *E. coli*, 8597 documented promoters exist, but only 3109 are assigned to different sigma factors: 1884 promoters are recognized by σ^{70}, 94 promoters by σ^{54}, 165 by σ^{38}, 307 by σ^{32}, 141 by σ^{28}, 517 by σ^{24}, and 1 by σ^{19}. The consensus of promoters for each sigma factor is shown in Fig. 2, which also shows the occurrence of distances between each promoter-box. In the case of σ^{38}, it is not possible to get a consensus sequence for the −35 box; it has been reported that this sigma factor recognizes the promoters mainly through the −10 and −10 extended elements.[47] Consistently we find that these genetic elements are more clearly present in these promoters. These analyses of promoter consensus reveal differences in the promoters' composition of the σ^{70} family, and sigmas in this family should not be considered as uniform.

The upper part of Fig. 2, in bars, shows the frequencies in terms of nucleotide distances between the boxes of promoters for each kind of sigma factor reported in the literature. In the lower part of the figure, the consensus sequences of promoter boxes for each sigma factor are shown. Note that in the case of σ^{38}, we could not identify a consensus sequence in the −35 box. In the case of σ^{19} the consensus is not shown since there exists just a single promoter for this sigma factor.

The transcription of genes by a limited number of sigma factors divides the universe of promoters into seven sigmulons of different sizes. The transcriptional activity of σ^{70} is associated with the activity of 175 TFs (*i.e.* these promoters have DNA binding sites for σ^{70} and for some of these 175 TFs). On the other hand, the smallest sigmulon corresponds to σ^{19}, which is associated with the fewest number of TFs (4 TFs). It is interesting to note that the most global TF, CRP (catabolic repressor protein), is co-regulating at least in one promoter with all the seven sigma factors. Therefore, this global regulator could master the whole transcriptional activity.

In the case of operons regulated by TFs, there is a wider distribution, with 331 operons having a single DNA-binding site for a TF, and one operon (*csgDEFG*) having up to 32 DNA-binding sites for TFs. When we computed the number of TFs that regulate operons, we get that 425 operons are regulated by a single TF whereas one operon (*gadAXW*) is regulated by up to 17 TFs. Again, we observed that the number of TF DNA-binding sites does not necessarily correlate with the number of different TFs because TFs could have more than one DNA-binding site on one promoter.

Transcription start sites (TSS). The TSS is the nucleotide where the transcription initiates and it is usually numbered +1.[48,49] When the RNA polymerase has been positioned on the promoter region it starts to polymerize the ribonucleotides in

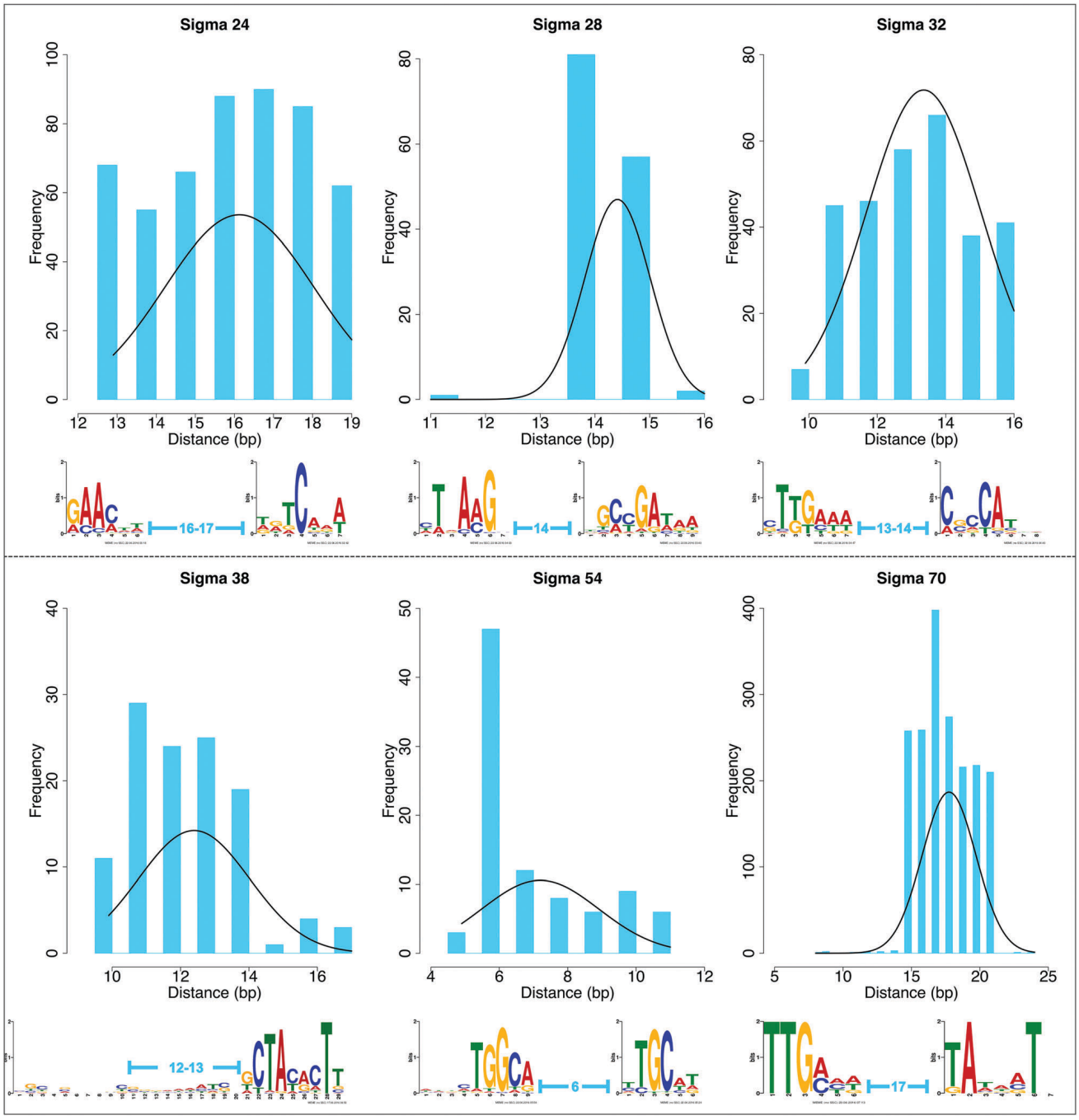

Fig. 2 Promoters for each sigma factor in *E. coli*.

the antiparallel RNA strand at this position.[20] The ribonucleotide reported as the most commonly used template for starting the mRNAs is adenine (43.88%), followed by guanine (24.31%), uracil (18.03%), and cytosine (13.76%). This distribution was obtained from a population of 8500 mapped initiations of transcription. We do not understand how the nucleotide is chosen to initiate the transcription; it seems to be merely based upon whether there is an adequate distance from the promoter to where the RNA polymerase binds to start the transcription.

Ribosome-binding sites. Translation is the process of protein synthesis, and it occurs in a multi-component structure called the ribosome. In bacteria, the ribosome (with relative sedimentation rate of 70S) is composed of two subunits, the large subunit 50S (which includes the 5S and 23S ribosomal RNA) and the small subunit 30S (which includes the 16S ribosomal RNA).[50–53] The ribosome consists of two-thirds of RNA and one-third of proteins, and has three tRNA-binding sites designated as the aminoacyl (A), peptide (P), and exit (E) sites. The translational process can be subdivided into initiation, elongation, and termination. Three initiation factors (IFs) are involved in the interaction with the ribosome: IF1 blocks the site A, which is the entry site of aminoacyl-tRNA; IF3 blocks the site E, which is the exit site of

aminoacyl-tRNA; and IF2 is a GTPase, which binds to the first aminoacyl-tRNA and helps it to engage with the ribosome at the P site.[54,55] During elongation, the polypeptide chain is synthesized from a start to a stop codon. For elongation termination, the A-site recognizes a stop codon and the accessory factors are released. In bacteria, the release factor 1 (RF1) recognizes the UAA and UAG stop codons, whereas RF2 recognizes UAA and UGA, the third release factor (RF3) catalyzes to RF1 and RF2 in the termination process.[56,57]

In this sense, the ribosome binds to a site on the mRNA called the ribosome-binding site (RBS). The RBS is also known as the Shine–Dalgarno sequence based on the last names of the scientists who first described it.[28,29] The ribosomes bind the RBS by the complementary sequence UCCUCC present at the 3′ end of the 16S rRNA.[26,29] From 179 annotated RBSs in *E. coli* we determined that the most conserved nucleotide is the 4 bp sequence AGGG. This, however, is not the best complementary sequence to the 3′ end of 16S rRNA, which should be AGGAGG. This sequence has already been used in synthetic genetic constructions with good results for protein expression.[58]

The codon most used for translation initiation. In prokaryotes, the first codon (AUG) is recognized by a specific tRNA ($tRNA^{Met}$); other codons that can also be used to initiate translation are GUG and UUG. Methionine bound to the $tRNA^{Met}$ at the initiation of translation is *N*-formylated, which selectively distinguishes it from the Met-$tRNA^{Met}$ used during the peptide elongation phase.[59–61]

Like in transcription initiation, where one of the four nucleotides is more used, in translation initiation some codons are more used than others. We compared 4356 translation initiations and found that AUG (methionine, 89.56%) is the codon most used for translation initiation in *E. coli*. Codons less frequently used as start codons are GUG at 8.61% and TTG at 1.69%. This preference in the codons used for initiation of translation is shared with other divisions of bacteria.

Genes. The concept of a gene has changed over time. The term was first used by Wilhelm Johannsen in 1909, when he called genes ''special conditions, foundations and determiners which are present in unique, separate and thereby independent ways many characteristics of the organism are specified''. This was the basis for the concept developed later by Gregor Mendel.[31,62] Classical genetics defines a gene as ''the unit of inheritance that ferried a characteristic from parent to child''.[63] An alternative, and more contemporary, concept should comprise the structure, function, and encoding properties; in this way, genes can be defined as ''the segment of DNA involved in producing a polypeptide chain or stable RNA; it includes regions preceding and following the coding region''.[64]

E. coli contains 4650 annotated genes distributed in 4239 genes that encode for proteins, 86 for tRNAs, 22 for rRNA, 144 pseudo-genes, 38 phantom-genes, and 124 for small RNA (three genes are classified into two categories, phantom and small RNA).

The smallest genes are the pseudo-gene *sokA* (30 bp), *yrb13′* which is a small RNA (40 bp), *pheM* which is a tRNA (45 bp), *trpL* which is a leader peptide (45 bp), and two small RNAs (*FimA34′* and *Spy3′*) with 48 bp each. The larger genes are *lhr* (4617 bp), a member of the ATP-dependent helicase super family II, *yfhM* (4962 bp), which encodes an alpha-macroglobulin, and *yeeJ* that encodes a protein involved in biofilm formation (7104 bp).

The genes with the lowest percent of GC content are *ynbG* (13.64%), *yobI* (19.7%), and *yqcG* (24.11%), while the genes with the largest GC content are *argX* with 67.53% (tRNA for arginine), *yagF* with 66.72% (D-xylonate dehydratase), and *metW* with 66.23% (tRNA for met1). The average GC content is 50.95%, considering all the gene sequences. The GC content in intergenic regions is also variable; in the region between *sfsB* and *murA* it is 42.29%, between *frlB* and *frlC* it is 67.34%, between *yfdP* and *yfdQ* it is 21.53%, and between *yfdp* and *safA* it is 18.91%. The average GC content in intergenic regions is 41.81%.

Translational stop codons. The process of gene translation is terminated by one of three stop codons in *E. coli* (UAG, UAA, and UGA).[54,65,66] The most common stop codon used in *E. coli* is UAA (63.98%), followed by UGA (28.69%), and UAG (7.31%). No natural genes in *E. coli* use double stop codons as are normally implemented in synthetic genetic circuits when assembled with BioBricks.

Terminus of transcription. In bacteria, there are two ways in which transcription can be finished. The first is by Rho-dependent terminators, which involves an interaction between a protein called Rho and a Rho-dependent site in the RNA (*rut* site). Rho is an RNA-binding protein that interacts with RNA polymerase to disrupt transcription.[22,67–69] The second way to terminate translation is Rho-independent; it is managed by stable hairpins followed by U-rich regions in the mRNA. Rho-independent binding causes the pausing and eventual dissociation of the transcription elongation complex from mRNA.[48,70–73]

There are 279 annotated terminators of transcription in *E. coli*, from these just 18 are Rho-dependent without a consensus sequence, whereas the remaining 261 are Rho-independent. There are 11 operons that have double Rho-independent terminators and just two operons with four terminators of this kind (*mtlADR* and *rnpB*). Rho-independent terminators form stem-loop structures, most commonly including 20–32 nucleotides. We used a software program designed for predicting termination efficiency (TTEC, Transcriptional Terminator Efficiency Calculator, http://www.terminatorefficiency.com/). We looked for experimental studies that can validate these predictions for terminator efficiency and found a study by Chen *et al.*, which used two fluorescent proteins separated by different terminators of *E. coli*, the expression of the second protein – downstream – is an indication of the efficiency of the upstream terminator.[74] There is no correlation between the size of the stem-loop structure and the strength of termination activity, for instance, terminators of 15 bp (*rplKAJL-rpoBC* operon) and 58 bp (*mgrP* operon) are the most efficient natural terminators. The largest natural terminator has 152 nucleotides for the operon *ilvLXG_1g_2MEDA*, which encodes for four of the five enzymes required for isoleucine and valine biosynthesis.

Genome distances between contiguous pairs of functional genetics parts

Distances from promoters to the initiation of transcription. The most common way to define the distance from promoters

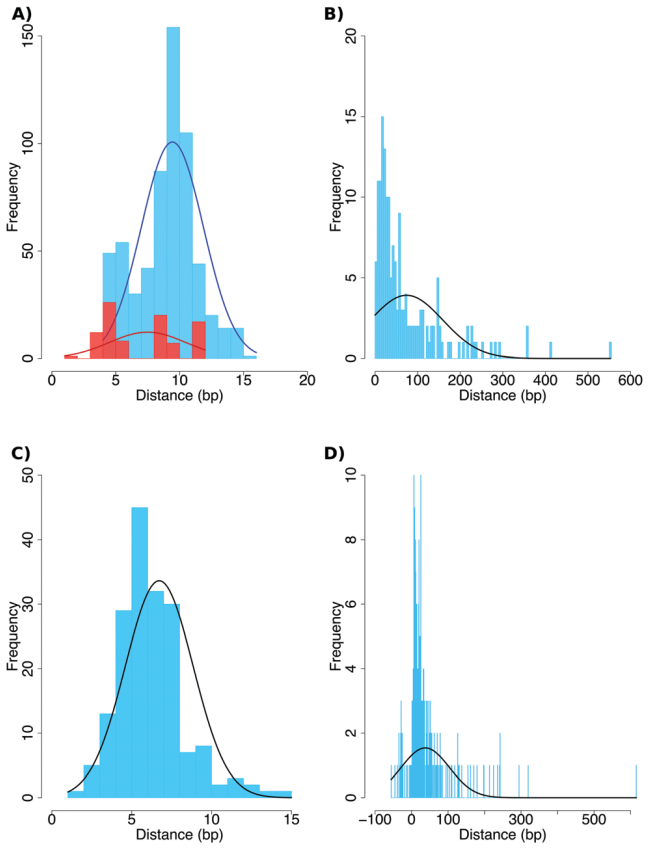

Fig. 3 Distances between genetic elements in transcription units. (A) Frequency of distances between the TSS to the upstream proximal boxes of promoters, in blue color is shown the distance to the −10 box and, in red the distance to the −12 box, which represent the boxes recognized for σ^{70} and σ^{54} respectively. (B) Frequency of distances between the TSS to the RBS. (C) Frequency of distances between the RBS to the initiation of translation. (D) Frequency of distances between the stop codon to the terminus of transcription. These distances are shown in number of nucleotides between each pair of elements. A normal curve was drawn to fit each distribution.

to the initiator of transcription is by considering the position of the middle of the −10 boxes relative to the nucleotide where transcription starts (+1).[11,48,49] In a population of 2907 pairs of documented promoters (for the family of σ^{70}) and initiators of transcription on the same TU, it was found that distances are on average 10 bp. This distance ranges however from as small as 4 bp to 16 bp. There is no consensus on the nucleotide composition of these spacing regions (Fig. 3A). In the case of σ^{54} promoters, this distance corresponds to 7 nucleotides on average; going from 1 to 12 nucleotides in the intergenic regions, as obtained from 91 pairs of σ^{54} promoters and their corresponding +1 positions.

Distance from the transcription initiation site to the ribosome binding site. A structural component that seems very important for protein expression is the RBS in mRNA. This sequence could be analogous to the promoter on DNA for gene expression. The region between the 5′ mRNA end and the position of the RBS is called the 5′-untranslated region (5′-UTR), and this distance seems to influence the translation initiation efficiency of the mRNA, probably through the formation of secondary structures close to the RBS that might be modulating the ribosome's access to this zone.[75] To calculate the distribution of distances between the TSSs and RBSs, we computed a population of 162 pairs of TSS and RBS annotated over the same mRNA. We found that the distance between these pairs of genetic elements is variable (Fig. 3B). The most frequent distance is 12–40 nucleotides, although the distance can be from −2 to 553 nucleotides as annotated in RegulonDB. Additionally, there is no sequence consensus in these spacing regions. This could indicate that a large distance is enough to promote translation efficiency, possibly due to adequate positioning of the ribosome and accessory proteins for translation initiation.[76]

Distances from the RBS to the initiation of translation. The RBS is the site where the ribosome machinery binds to start the translation from mRNA to a polypeptide. The translation initiation occurs a few nucleotides downstream of the RBS.[77–79] From 179 spacing regions documented for *E. coli*, we determined that the spacing region ranges from 15 to 1 nucleotides, with an average of 6 nucleotides. Coincidentally, the synthetic genetic circuits typically left 6 bp on these regions. Again, there is no consensus among these spacing sequence regions (Fig. 3C).

Distance between genes and genes overlapping in polycistronic TUs. Most TUs encode single genes (2324 TUs) but in some cases, there is more than one gene encoded in a single TU (1225 TUs). For 644 (44%) TUs with more than one gene, at least two genes are overlapping. When genes overlap, it is mostly by 4 bp (264 cases) where the stop codon of the former gene overlaps with the initiation codon of the second gene, in one of the following arrays: ATGA, GTGA, and TTGA (as previously observed in ref. 80). In these cases, the RBS of the second gene falls inside the 3′ sequence of the former gene. The flexibility of sequence composition and distance of the RBS however avoids having conserved amino acids in the region that could correspond to the conserved position of the RBS on the former gene. In the rest of the TUs that have more than one gene, the intergenic distances vary from 0 to 559 nucleotides (in 1247 pairs of genes).

Distance from the stop codon to the terminus of transcription. The marks for transcription and translation both fall at the end of the TUs. The translation stop marks are found before the end of mRNA, and the distances between the translational and transcription finishing marks vary widely. However, the most frequent distances are from 10 to 30 nucleotides, although there are distances as long as 615 bp, if we consider their positions on the DNA (Fig. 3D).

Architecture of synthetic genetic circuits

Design of synthetic TUs. This section is a brief analysis of architectures that currently results in designing synthetic circuits. The iGEM organization (http://parts.igem.org) provides a compendium of standardized genetic parts, including promoters, RBS, coding sequences for proteins and motifs, terminators, reporters, and plasmids, among others (Fig. 4A).

The BioBricks catalogue contains around 709 promoters, including 316 that function in *E. coli*. Of these, only 63 have the

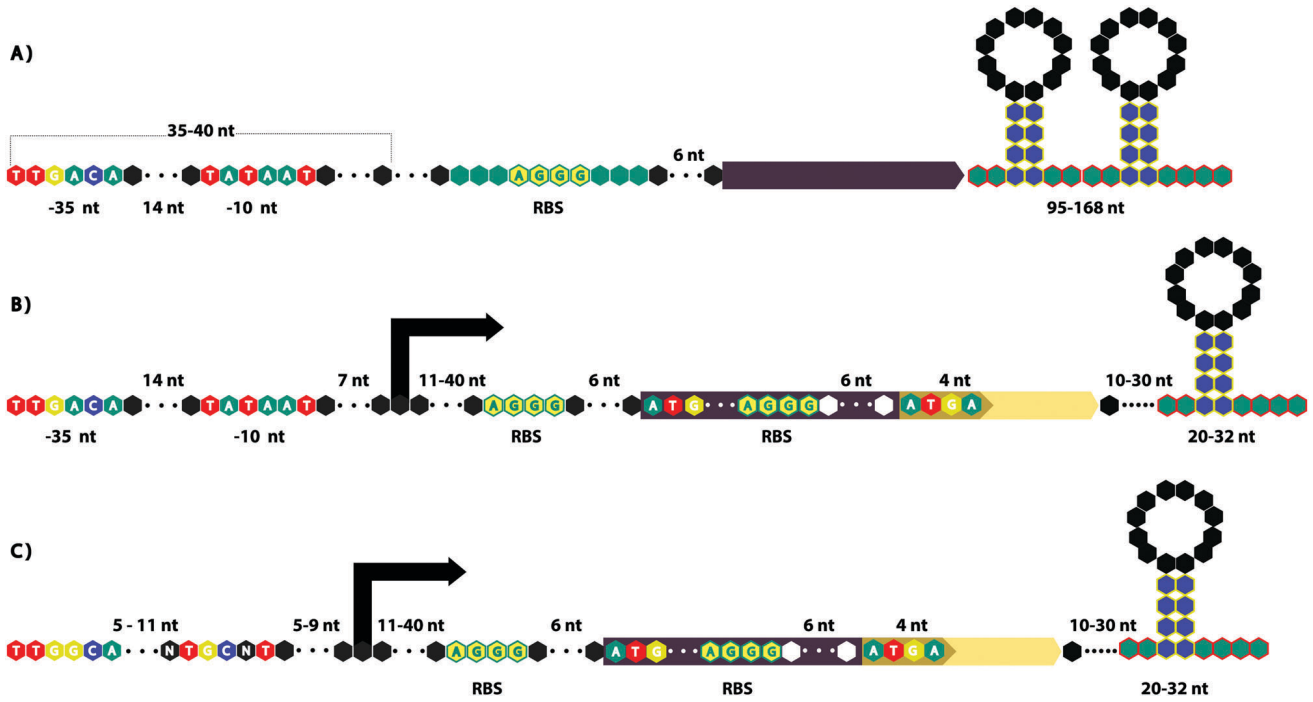

Fig. 4 Synthetic and natural architecture of transcription units. In this figure, we illustrate the genetic elements, their positions, and distances most important in: (A) synthetically designed TUs, (B) the consensus of natural TUs for σ^{70} of *E. coli*, and (C) promoter consensus for σ^{54} of *E. coli*. Hexagons represent nucleotides; with thymine in red, cytosine in blue, adenine in green, guanine in yellow, and non-conserved nucleotides in black. Genes are represented as purple and yellow squares.

−10 and −35 boxes that allow constitutive expression. The largest proportion contains a sequence consensus for −10 (TATAAT) and −35 (TTGACA) boxes, and a distance between boxes of 17 bp instead of 14 bp as occurs naturally.

In the case of RBS, 48 are registered in the iGEM repository. These are included in Anderson's RBS library and the Community Collection. The RBSs in Anderson's library have 8 and 6 bp to the initiation codon and could vary on nucleotide compositions. On the other hand, the RBSs in the Community Collection have sizes of 6 bp and from 6 to 9 bp distance from the RBS to the initiation codon.

There are 44 Rho-independent transcriptional terminators as BioBricks; of these only two terminators are natural from *E. coli* (both from the ribosomal operon *rrsB-gltT-rrlB-rrfB*). The terminators have sizes ranging from 33 to 113 bp. Usually, the synthetic genetic circuits are designed with double terminator sequences.

Proposal of consensus architecture for genetic circuits on synthetic biology

Based on the above analyses of the anatomy, consensus, and distances among genetic elements for TUs, we propose an archetypal design for synthetic operons that represents the consensus architecture of TUs in *E. coli*. We hope this architecture will work in the *E. coli* chassis (Fig. 4B and C). We recommend considering consensus sequences as described above for important functional marks. The synthetic genetic constructions should contain the following elements and distances.

(1) For promoters, it depends on the type of regulation you want to modulate, as there are some differences in their consensus. In the case of σ^{70} you can use 14 bp between the −10 and −35 boxes, these boxes can be variable and moldable depending on the desired promoter strength. For a catalog of promoter strengths, you can also consult the bioparts promoter repository at iGEM.org.

(2) Maintain 7 nucleotides for spacing between the −10 box promoter to the start of transcription and seek to use an adenine at the end of the sequence (adenine will indicate the start of transcription).

(3) From the transcription start site to the RBS there should be between 12 to 40 bp. This is the most variable region, and this region is subject to post-transcriptional regulation by riboswitches and attenuators, among others.

(4) Use 6 bp as the distance from the RBS to the initiation of translation.

(5) If more than one gene needs to be encoded, they could be overlapped by 4 bp. These designs should be important for economy of synthetic parts but they are clearly a limitation for biopart recycling in several constructions.

(6) The distance from the stop codon to the Rho-independent terminator should be between 10 and 30 nucleotides. The terminator structure can be of 20–32 bp, and the use of two terminators should be avoided because that could promote homologous recombination in polycistronic operons.[74] You can choose one terminator from Table 4; these terminators efficiently stop transcription and they do not possess sequences susceptible to homologous recombination.

Table 4 The best natural terminators from *E. coli*

Terminator identifier	Sequence	Operon	Efficiency[a]	Size	Experimental[b] average strength
ECK120035133	AAGGCGACTCATCAGTCGCCTT	*rph-pyrE*	100	22	55.17
ECK120017009	AAAAAGGCCGCAGAGCGGCCTTTT	*thrS-infC-rpmI-rplT-pheMST-ihfA*	100	24	67.56
ECK120015170	AAAAAAACCCGCCGAAGCGGGTTTTTT	*rplM-rpsI*	100	27	85.77
ECK120033736	AAAGCCCCCGGAAGATCACCTTCCGGGGGCTTT	*hisLGDCBHAFI*	100	33	164.60
ECK120033737	AAAAAAGCCCGCACCTGACAGTGCGGGCTTTTTTTTT	*thrLABC*[a]	100	37	312.49
ECK120010799	AGGAAAAAAGGCGACAGAGTAATCTGTCGCCTTTTTTCTT	*csrC*	100	40	101.04

[a] Using TTEC software (Transcriptional Terminator Efficiency Calculator). [b] The experimental average strength was reported experimentally by Chen *et al.*[74] The terminator efficiency was measured as the activity of a second fluorescent protein encoded downstream of a first fluorescent protein separated by a tested terminator (GFP and RFP).

Conclusion

In this work, we review the anatomy and natural architecture of TUs and their regulatory signals in *E. coli*. We focus the analysis on this organism, and more specifically on the strain K-12 MG1655, because it is the most widely described and used chassis in synthetic biology. However, much of the natural architecture known in other bacteria reveal that this architecture could be shared.[81] In the case of other *E. coli* strains, *Salmonella typhimurium* and *Pseudomonas aeruginosa*, a preliminary analysis (Tables S2, S3 and Fig. S1, ESI†) with genetic elements characterized to some extent reveals that these bacteria use similar TSS (adenine mainly) as well as the same translation start codon (AUG), translation stop codon (UAA), and ribosomal binding sites, the genes overlap by 4 bp and the average distance between the TSS and the RBS is from 20 to 40 bp. These organisms have much fewer genetic elements annotated in great detail compared to *E. coli*, and it is not possible to do a deep analysis like the one reported in this study but it seems that the main genetic architectures are conserved in *E. coli* strains and even other phylogenetically related bacteria.[82,83] Further analyses are necessary to confirm these preliminary results.

Although many genetic circuits have been made synthetically, the performance of these might be improved if we consider the designs that nature has used for many years. Despite common wisdom about the economy of space in prokaryotes we show that there are spacing regions necessary for the correct operation of the transcription and translational cellular machineries.

The most important difference we note between natural and human-designed genetic circuits, is that the latter are made presently with standardized BioBricks and do not consider the natural distance (−2 to 553 bp) present between transcription initiation and the RBSs. Multiple studies have been done on the engineering of each of these genetic components[6–8] to find an optimal design and facilitate the transcription and translation for the cellular machinery. On the other hand, the transcriptional regulation given mainly by TFs that bind to the operator regions of TUs adjusts the gene expression to endogenous and exogenous conditions. Though, this *trans* regulatory code required for the timing operation of TUs has not yet been described in detail in this organism.

Acknowledgements

The authors thank Cei Abreu-Goodyear and Luis Jose Delaye-Arredondo for useful suggestions for these analyses, and Jose Ramón Mendoza for his assistance with Fig. 4 and Fig. S1 (ESI†). CR-Ch held a Conacyt master fellow (324427) during this study. The authors thank Collen Beard for editing the manuscript.

Notes and references

1 F. R. Blattner, G. Plunkett 3rd and C. A. Bloch, *et al.*, The complete genome sequence of *Escherichia coli K-12*, *Science*, 1997, **277**(5331), 1453–1462.
2 S. Gama-Castro, H. Salgado, A. Santos-Zavaleta and D. Ledezma-Tejeida, *et al.*, RegulonDB version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond, *Nucleic Acids Res.*, 2016, **44**(D1), D133–D143.
3 T. Ellis, T. Adie and G. S. Baldwin, DNA assembly for synthetic biology: from parts to pathways and beyond, *Integr. Biol.*, 2011, **3**(2), 109–118.
4 A. A. Cheng and T. K. Lu, Synthetic biology: an emerging engineering discipline, *Annu. Rev. Biomed. Eng.*, 2012, **14**, 155–178.
5 J. Bonnet, P. Subsoontorn and D. Endy, Rewritable digital data storage in live cells *via* engineered control of recombination directionality, *Proc. Natl. Acad. Sci. U. S. A.*, 2012, **109**(23), 8884–8889.
6 R. P. Patwardhan, C. Lee and O. Litvin, *et al.*, High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis, *Nat. Biotechnol.*, 2009, **27**(12), 1173–1175.
7 H. M. Salis, E. A. Mirsky and C. A. Voigt, Automated Design of Synthetic Ribosome Binding Sites to Precisely Control Protein Expression, *Nat. Biotechnol.*, 2009, **27**(10), 946–950.
8 J. B. Kinney, A. Murugan and C. G. Callan Jr., *et al.*, Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence, *Proc. Natl. Acad. Sci. U. S. A.*, 2010, **107**(20), 9158–9163.
9 A. Martinez-Antonio, *Escherichia coli* transcriptional regulatory network, *Network Biol.*, 2011, **1**(1), 21–33.

10 H. Weber, T. Polen and J. Heuveling, *et al.*, Genome-wide analysis of the general stress response network in *Escherichia coli*: sigmaS-dependent genes, promoters, and sigma factor selectivity, *J. Bacteriol.*, 2005, **187**(5), 1591–1603.

11 A. Mendoza-Vargas, L. Olvera and M. Olvera, *et al.*, Genome-Wide Identification of Transcription Start Sites, Promoters and Transcription Factor Binding Sites in *E. coli*, *PLoS One*, 2009, **4**(10), e7526.

12 J. P. Richardson, Rho-dependent Termination of Transcription Is Governed Primarily by the Upstream Rho Utilization (rut) Sequences of a Terminator, *J. Biol. Chem.*, 1996, **271**(35), 21597–21603.

13 A. Ray-Soni, M. J. Bellecourt and R. Landick, Mechanisms of bacterial transcription termination: all good things must end, *Annu. Rev. Biochem.*, 2016, **85**, 319–347.

14 H. Salgado, G. Moreno-Hagelsieb and T. F. Smith, *et al.*, Operons in *Escherichia coli*: genomic analyses and predictions, *Proc. Natl. Acad. Sci. U. S. A.*, 2000, **97**(12), 6652–6657.

15 M. Thomas-Chollier, C. Herrmann and M. Defrance, *et al.*, RSAT peak-motifs: motif analysis in full-size ChIP-seq datasets, *Nucleic Acids Res.*, 2011, **40**(4), 1–9.

16 M. Touchon, C. Hoede, O. Tenaillon, V. Barbe, S. Baeriswyl, P. Bidet and A. Calteau, *et al.*, Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths, *PLoS Genet.*, 2009, **5**(1), e1000344.

17 J. Tamames, Evolution of gene order conservation in prokaryotes, *Genome Biol.*, 2001, **2**(6), research0020-1.

18 M. Chamberlin and P. Berg, Deoxyribonucleic acid-directed synthesis of ribonucleic acid by an enzyme from *Escherichia coli*, *Proc. Natl. Acad. Sci. U. S. A.*, 1962, **48**(1), 81–94.

19 D. G. Vassylyev, S. I. Sekine, O. Laptenko, J. Lee, M. N. Vassylyeva, S. Borukhov and S. Yokoyama, Crystal structure of a bacterial RNA polymerase holoenzyme at 2.6 Å resolution, *Nature*, 2002, **417**(6890), 712–719.

20 D. F. Browning and S. J. Busby, The regulation of bacterial transcription initiation, *Nat. Rev. Microbiol.*, 2004, **2**(1), 57–65.

21 G. Zhang, E. A. Campbell and L. Minakhin, *et al.*, Crystal Structure of Thermus aquaticus Core RNA Polymerase at 3.3 Å Resolution, *Cell*, 1999, **98**(6), 811–824.

22 J. W. Roberts, Termination factor for RNA synthesis, *Nature*, 1969, **224**, 1168–1174.

23 E. Skordalakes and J. M. Berger, Structure of the Rho transcription terminator: mechanism of mRNA recognition and helicase loading, *Cell*, 2003, **114**(1), 135–146.

24 R. C. Huang, N. Maheshwari and J. Bonner, Enzymatic Synthesis of RNA, *Biochem. Biophys. Res. Commun.*, 1960, **3**(1), 689–694.

25 J. Hurwitz, A. Bresler and R. Diringer, The enzymatic incorporation of ribonucleotides into polyribonucleotides and the effect of DNA, *Biochem. Biophys. Res. Commun.*, 1960, **3**(1), 15–19.

26 B. S. Laursen, H. P. Sørensen, K. K. Mortensen and H. U. Sperling-Petersen, Initiation of Protein Synthesis in Bacteria, *Microbiol. Mol. Biol. Rev.*, 2005, **69**(1), 101–123, DOI: 10.1128/MMBR.69.1.101-123.2005.

27 B. K. Cho, K. Zengler and Y. Qiu, *et al.*, The transcription unit architecture of the *Escherichia coli* genome, *Nat. Biotechnol.*, 2009, **27**(11), 1043–1049.

28 J. Shine and L. Dalgarno, The 3′-terminal sequence of *Escherichia coli* 16S ribosomal RNA: complementarity to nonsense triplets and ribosome binding sites, *Proc. Natl. Acad. Sci. U. S. A.*, 1974, **71**, 1342–1346.

29 J. Shine and L. Dalgarno, Determinant of cistron specificity in bacterial ribosomes, *Nature*, 1975, **254**, 34–38.

30 D. Omotajo, T. Tate, H. Cho and M. Choudhary, Distribution and diversity of ribosome binding sites in prokaryotic genomes, *BMC Genomics*, 2015, **16**(1), 604, DOI: 10.1186/s12864-015-1808-6.

31 M. B. Gerstein, C. Bruce and J. S. Rozowsky, *et al.*, What is a gene, post-ENCODE? History and updated definition, *Genome Res.*, 2007, **17**, 669–681.

32 J. Ma, A. Campbell and S. Karlin, Correlations between Shine–Dalgarno Sequences and Gene Features Such as Predicted Expression Levels and Operon Structures, *J. Bacteriol.*, 2002, **184**(20), 5733–5745.

33 E. S. Poole, C. M. Brown and W. P. Tate, The identity of the base following the stop codon determines the efficiency of *in vivo* translational termination in *Escherichia coli*, *EMBO J.*, 1995, **14**(1), 151–158.

34 M. Kozak, Initiation of translation in prokaryotes and eukaryotes, *Gene*, 1999, **234**(2), 187–208.

35 A. Stevens, Incorporation of the adenine ribonucleotide into RNA by cell fractions from *E. coli*, *Biochem. Biophys. Res. Commun.*, 1960, **3**(1), 93–96.

36 C. A. Gross, C. Chan and A. Dombroski, *et al.*, The Functional and Regulatory Roles of Sigma Factors in Transcription, *Cold Spring Harbor Symp. Quant. Biol.*, 1998, **63**, 141–156.

37 S. L. Dove, S. A. Darst and A. Hochschild, Region 4 of sigma as a target for transcription regulation, *Mol. Microbiol.*, 2003, **48**(4), 863–874.

38 R. R. Burgess, A. A. Travers and J. J. Dunn, *et al.*, Factor stimulating transcription by RNA polymerase, *Nature*, 1969, **221**, 43–46.

39 A. Kumar, R. A. Malloch, N. Fujita, D. A. Smillie, A. Ishihama and R. S. Hayward, The minus 35-recognition region of *Escherichia coli* sigma 70 is inessential for initiation of transcription at an, *J. Mol. Biol.*, 1993, **232**(2), 406–418.

40 G. Zhang and S. A. Darst, Structure of the *Escherichia coli* RNA polymerase α subunit amino-terminal domain, *Science*, 1998, **281**(5374), 262–266.

41 B. Bae, E. Davis, D. Brown, E. A. Campbell, S. Wigneshweraraj and S. A. Darst, Phage T7 Gp2 inhibition of *Escherichia coli* RNA polymerase involves misappropriation of σ70 domain 1.1, *Proc. Natl. Acad. Sci. U. S. A.*, 2013, **110**(49), 19772–19777.

42 R. K. Shultzaberger, Z. Chen and K. A. Lewis, *et al.*, Anatomy of *Escherichia coli* sigma70 promoters, *Nucleic Acids Res.*, 2007, **35**(3), 771–788.

43 D. Pribnow, Bacteriophage T7 early promoters: nucleotide sequences of two RNA polymerase binding sites, *J. Mol. Biol.*, 1975, **99**(3), 419–443.

44 M. Takanami, K. Sugimoto and H. Sugisaki, *et al*., Sequence of promoter for coat protein gene of bacteriophage fd, *Nature*, 1976, **260**, 297–302.

45 K. A. Barne, J. A. Bown and S. J. W. Busby, *et al*., Region 2.5 of the *Escherichia coli* RNA polymerase sigma 70 subunit is responsible for the recognition of the 'extended-10' motif at promoters, *EMBO J.*, 1997, **16**(13), 4034–4040.

46 J. D. Helmann and M. J. Chamberlin, Structure and function of bacterial sigma factors, *Annu. Rev. Biochem.*, 1988, **57**, 839–872.

47 C. Peano, J. Wolf and J. Demol, *et al*., Characterization of the *Escherichia coli* σ(S) core regulon by Chromatin Immunoprecipitation-sequencing (ChIP-seq) analysis, *Sci. Rep.*, 2015, **5**, 10469.

48 M. Rosenberg and D. Court, Regulatory sequences involved in the promotion and termination of RNA transcription, *Annu. Rev. Genet.*, 1979, **13**, 319–359.

49 C. B. Harley and R. P. Reynolds, Analysis of *E. coli* promoter sequences, *Nucleic Acids Res.*, 1987, **15**(5), 2343–2361.

50 R. H. Pain and J. A. Butler, The preparation and properties of ribonucleic acids from rat liver, *Biochem. J.*, 1956, **66**, 299–302.

51 M. Nomura and C. V. Lowry, Phage F2 RNA-directed binding of formylmethionyl-tRNA to ribosome and the role of 30S ribosomal subunits in initiation of protein synthesis, *Proc. Natl. Acad. Sci. U. S. A.*, 1967, **58**(3), 946–953.

52 M. Kozak, Comparison of initiation of protein synthesis in procaryotes, eucaryotes and organelles, *Microbiol. Rev.*, 1983, **47**(1), 1–45.

53 A. Simonetti, S. Marzi, L. Jenner, A. Myasnikov, P. Romby, G. Yusupova and M. Yusupov, *et al*., A structural view of translation initiation in bacteria, *Cell. Mol. Life Sci.*, 2009, **66**(3), 423–436.

54 D. Petrelli, A. LaTeana and C. Garofalo, *et al*., Translation initiation factor IF3: two domains, five functions, one mechanism? *EMBO J.*, 2001, **20**(16), 4560–4569.

55 A. Marintchev and G. Warner, Translation initiation: structures, mechanisms and evolution, *Q. Rev. Biophys.*, 2004, **37**(3–4), 197–284.

56 S. Brenner, A. O. W. Stretton and S. Kaplan, Genetic code: The 'nonsense' triplets for chain termination and their suppression, *Nature*, 1965, **206**, 994–998.

57 V. Ramakrishnan, Ribosome structure and the mechanism of translation, *Cell*, 2002, **108**, 557–5572.

58 I. Farasat, M. Kushwaha and J. Collens, *et al*., Efficient search, mapping, and optimization of multi-protein genetic systems in diverse bacteria, *Mol. Syst. Biol.*, 2014, **10**, 731.

59 K. Marcker and F. Sanger, N-Formyl-methionyl-S-RNA, *J. Mol. Biol.*, 1964, **8**, 835–840.

60 B. F. C. Clark and K. Marcker, The role of N-Formyl-methionyl-sRNA in protein biosynthesis, *J. Mol. Biol.*, 1966, **17**, 294–406.

61 C. O. Gualerzi and C. L. Pon, Initiation of mRNA translation in bacteria: structural and dynamic aspects, *Cell. Mol. Life Sci.*, 2015, **72**, 4341–4367.

62 P. Portin, Historical development of the concept of the gene, *J. Med. Philos.*, 2002, **27**(3), 257–286.

63 H. Pearson, Genetics: What is a gene? *Nature*, 2006, **441**, 398–401.

64 K. Scherrer and J. Jost, The gene and the genon concept: a functional and information-theoretic analysis, *Mol. Syst. Biol.*, 2007, **3**, 1–11.

65 S. Brenner, E. R. Barnett and E. R. Katz, Crick FHC. UGA: a third nonsense triplet in genetic code, *Nature*, 1967, **213**, 449–450.

66 G. Korkmaz, M. Holm, T. Wiens and S. Sanyal, Comprehensive Analysis of Stop Codon Usage in Bacteria and Its Correlation with Release Factor Abundance, *J. Biol. Chem.*, 2014, **289**(44), 30334–30342, DOI: 10.1074/jbc.M114.606632.

67 J. L. Pinkham and T. Platt, The nucleotide sequence of the rho gene of *E. coli* K-12, *Nucleic Acids Res.*, 1983, **11**(11), 3531–3545.

68 V. Epshtein, D. Dutta and J. Wate, *et al*., An allosteric mechanism of Rho-dependent transcription termination, *Nature*, 2010, **463**, 245–249.

69 T. M. Henkin, Transcription termination control in bacteria, *Curr. Opin. Microbiol.*, 2000, **3**(2), 149–153.

70 M. Rosenberg, S. Weissman and B. deCrombrugghe, Termination of transcription in bacteriophage λ, *J. Biol. Chem.*, 1975, **250**(12), 4755–4764.

71 S. Adhya and M. Gottersman, Control of transcription termination, *Annu. Rev. Biochem.*, 1978, **47**, 967–996.

72 P. J. Farnham and T. Platt, Rho-independent termination: dyad symmetry in DNA causes RNA polymerase to pause during transcription *in vitro*, *Nucleic Acids Res.*, 1981, **9**(3), 563–577.

73 K. Datta and P. H. von Hippel, Direct spectroscopic study of reconstituted transcription complexes reveals that intrinsic termination is driven primarily by thermodynamic destabilization of the nucleic acid framework, *J. Biol. Chem.*, 2008, **283**(6), 3537–3549.

74 Y. J. Chen, P. Liu and A. A. Nielsen, *et al*., Characterization of 582 natural and synthetic terminators and quantification of their design constraints, *Nat. Methods*, 2013, **10**, 659–664.

75 S. Kosuri, D. B. Goodman and G. Cambray, *et al*., Composability of regulatory sequences controlling transcription and translation in *Escherichia coli*, *Proc. Natl. Acad. Sci. U. S. A.*, 2013, **110**(34), 14024–14029.

76 V. Vimberg, A. Tats and M. Remm, *et al*., Translation initiation region sequence preferences in *Escherichia coli*, *BMC Mol. Biol.*, 2007, **8**, 100.

77 S. Ringquist, S. Shinedling and D. Barrick, *et al*., Translation initiation in *Escherichia coli*: sequences within the ribosome-binding site, *Mol. Microbiol.*, 1992, **6**(9), 1219–1229.

78 H. Chen, M. Bjerknes and R. Kumar, *et al*., Determination of the optimal aligned spacing between the Shine–Dalgarno sequence and the translation initiation codon of *Escherichia coli* mRNAs, *Nucleic Acids Res.*, 1994, **22**(23), 4953–4957.

79 R. K. Shultzaberger, R. E. Bucheimer and K. E. Rudd, *et al*., Anatomy of *Escherichia coli* ribosome binding sites, *J. Mol. Biol.*, 2001, **313**(1), 215–228.

80 G. Moreno-Hagelsieb and J. Collado-Vides, A powerful non-homology method for the prediction of operons in prokaryotes, *Bioinformatics*, 2002, **1**, S329–S336.

81 O. Wurtzel, D. R. Yoder-Himes, K. Han, A. A. Dandekar, S. Edelheit, E. P. Greenberg and S. Lory, *et al.*, The single-nucleotide resolution transcriptome of *Pseudomonas aeruginosa* grown in body temperature, *PLoS Pathog.*, 2012, **8**(9), e1002945.
82 I. D'Arrigo, K. Bojanovič, X. Yang, M. Holm Rau and K. S. Long, Genome-wide mapping of transcription start sites yields novel insights into the primary transcriptome of Pseudomonas putida, *Environ. Microbiol.*, 2016, **18**(10), 3466–3481.
83 C. Kröger, S. C. Dillon, A. D. Cameron, K. Papenfort, S. K. Sivasankaran, K. Hokamp and A. Colgan, *et al.*, The transcriptional landscape and small RNAs of Salmonella enterica serovar Typhimurium, *Proc. Natl. Acad. Sci. U. S. A.*, 2012, **109**(20), E1277–E1286.

