Title

Direct and indirect translational control of mRNA concentrations in Escherichia coli

Author
Sheng, Huanjie

Publication Date
2019

Direct and indirect translational control of mRNA concentrations in *Escherichia coli*

By

Huanjie Sheng

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Integrative Biology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Daniela Kaufer, Co-Chair
Doctor Han N. Lim, Co-Chair
Professor Ellen L. Simms
Professor Michiko E. Taga

Summer 2019

Abstract

Direct and indirect translational control of mRNA concentrations in *Escherichia coli*

By

Huanjie Sheng

Doctor of Philosophy in Integrative Biology

University of California, Berkeley

Professor Daniela Kaufer, Co-Chair
Doctor Han N. Lim, Co-Chair

The survival of bacterial cells requires the coordinated regulation of genes and proteins so that cells can quickly and efficiently respond to opportunities for rapid cell growth and to stresses of different environments. Many of the networks and mechanism for the coordinated regulation of genes and proteins involve the control the production, degradation and translation of messenger RNAs (mRNAs). This dissertation focuses on the relationship between translation and the regulation of mRNA production and degradation that occurs without (direct translational control) and with (indirect translational control) the action of small non-coding RNAs (sRNAs). The results shine a light on the designing process of circuits in bioengineering and the mystery of complicated bacterial stress responses and adaptations.

In the first chapter, we developed a new metric named threshold overlap score (TOS) that measures colocalization between biomolecules. TOS describes whether two or three biomolecules occur in the same place more than chance. Specifically, TOS quantifies the percentage of overlapping signals between two or three channels with respect to the uniformly distributed random signals and rescales this value so that it is easily interpretable. The TOS metric was used in the third chapter to quantify sRNA localization in the nucleoid, which provided novel insight into their role in mRNA transcription, degradation and translation.

In the second chapter, we built an open source tool (ImageJ plugin) with an easy-to-use user interface called EzColocalization. EzColocalization gives biologists without programming experience access to imaging segmentation, data visualization, and colocalization analyses including TOS, other classic colocalization metrics and custom colocalization measurements. A form of EzColocalization was used in the first chapter to demonstrate and valid TOS as useful metric and it was used in the third chapter to measure sRNA colocalization and understand the role in mRNA in mRNA transcription, degradation and translation.

In the third chapter, we studied the subcellular localization of sRNAs to elucidate their roles in mRNA production, degradation and translation. Based on our data, we concluded that sRNAs can but mRNAs cannot enter the densely packed nucleoid. This phenomenon is due to the active translation and larger size of mRNAs. These findings indicate that sRNAs have the

potential to regulate nascent mRNAs in the nucleoid prior to the completion of mRNA transcription, which increases their potential impact and efficiency as a regulator.

In the fourth chapter, we constructed mathematical models that capture “real-world” regulation. Unlike the classic central dogma model where mRNA production and degradation and protein production and degradation are all completely independent, the real-world model includes the effects of translation on premature transcription termination and mRNA degradation. The simulations reproduce key experimental observations and show that the coupling of translation to mRNA production and degradation to increase quality control occurs at the cost of efficiency.

Together, the four chapters in this dissertation provide novel insight into the direct and indirect control of mRNA concentrations by sRNA localization and translational feedbacks, and how these sophisticated regulatory processes may benefit bacterial growth and adaptation.

Table of Contents

Acknowledgements

I want to thank Han N. Lim for bringing me to Berkeley and the field of gene regulations in bacteria. Although Han left two years after I got here, he continued to help me with my Ph.D. By the time this dissertation comes out, we have already published three papers together corresponding to the first three chapters.

I also need to acknowledge Prof. Daniela Kaufer for adopting me after Han left. In the last three years, I had a good time in her lab working together with other people doing completely different things. It was a real challenge to work in a very different topic out of the field I had studied before. Nevertheless, I collaborated with a couple of people with their projects on blood-brain barrier and empathy. This actually will not only give me two more papers that are not included in this dissertation, but also a broader view of biological questions.

At the end, I want to thank my other committee members for their help during my five years in Berkeley. Prof Ellen Simms was also the chair of my qualifying exam and we had several helpful conversation on research topics and my interest. Prof. Michiko Taga provided useful guidance when Han left and she has always been responsive and easy to reach.

Introduction: An overview of gene regulation in *Escherichia coli* and the role of small RNAs in controlling gene expression

The production and degradation of mRNAs in bacterial cells are highly regulated and well-organized. As a key intermediate step of the central dogma of molecular biology, mRNAs bridge the gap between the DNA that encodes information to protein products that perform functions in the cell. In this dissertation, I focus on the translational controls of mRNA concentrations in *E.coli*. These controls include regulatory mechanisms like translational coupling as well as indirect factors like localization.

To provide an overview of gene expression systems in bacteria, the first part of this introductory chapter will review some of the fundamental processes in mRNA production and degradation: transcription initiation (I.I.i), transcription elongation (I.I.ii), transcription termination (I.I.iii), 5' end-dependent mRNA degradation (I.II.i), 5' end-independent mRNA degradation (I.II.ii), and mRNA fragment digestion (I.II.iii). The second part of this introductory chapter will examine some of the factors and mechanisms that are used to control the above fundamental processes in mRNA production and degradation: genome structures (II.I), translational coupling (II.II), small non-coding RNA regulation (II.III), and subcellular localization (II.IV). Other factors that are not discussed in the dissertation are briefly touched on in the last section.

I. Mechanisms of mRNA production and degradation

In this part I will briefly described the mechanism of mRNA production (*i.e.* transcription) and mechanism that degrade or remove mRNAs from cells.

I.I. Mechanisms of transcription

Transcription is the first part of central dogma marking the beginning of gene expression. The transcription of mRNAs in bacteria consists of three major phases: initiation, elongation, and termination. In particular, transcription elongation and termination are usually discussed together because stalling RNAP will be rescued by termination. On the other hand, transcription initiation represents the commitment to express the target gene constitutively or conditionally. Here I describe the gene mechanisms of transcription in central dogma.

I.I.i. Mechanism of transcription initiation

Transcription initiation in bacteria is typically regulated by the promoter sequence located between 10 and 35 nucleotides (-10 and -35) upstream of the transcription initiation start site [5, 6]. The promoter is a *cis*-regulatory sequence in the genome that can be recognized by RNAP and sigma factors (σ) [7]. This promoter is usually identified by the RNAP via searching for the correct sequences and bound sigma factors (σ) via one-dimensional or three-dimensional diffusion along the DNA [5, 6, 8, 9]. In *E.coli*, σ70 is the most common sigma factor that transcribes a large number of housekeeping genes [10]. Other sigma factors are present in response to specific stress conditions such as heat shock, osmotic pressure, and transition to stationary growth [11]. Once RNAP and the corresponding sigma factor bind at the promoter, they form a complex known as the RNA polymerase holoenzyme [12]. In this complex, the binding free energy drives the rate-limiting step of breaking the base pairs between the DNA double strands as the closed complex transforms to the open complex [13]. After this transition, RNAP is able to mediate the base pairing between the DNA template and the free single nucleotides to form the 3' of the new growing mRNA.

I.I.ii. Mechanism of transcription elongation

After approximately 10 nucleotides of the nascent mRNA have been created, the sigma factor of the holoenzyme is released marking the start of transcription elongation [7]. The remaining part of elongation is carried out by the subsequently assembled ternary elongation complex (TEC) which is composed of the newly synthesized RNA, the DNA template, and the RNAP [14]. Transcription elongation involves adding new nucleotides that are complementary to the DNA template to the 3' end of the growing mRNA. The nature of elongation requires TEC to be stable enough so that the transcription can continue and also flexible enough so it can disrupt base pairs while moving forward along the DNA template. Studies have shown that the β'-clamp of RNAP is responsible of DNA binding while the β -subunit of RNAP have catalytic and nucleotidyl transferase activity [15, 16]. The sites of RNAP that interacts with DNA and RNA play key roles in the sliding of the TEC while the DNA-RNA hybrid region keeps the TEC stable [17]. A "rachet" model has been proposed that favors the unidirectional movement of the RNAP [18, 19]. While a rachet mechanism can boost the efficiency and fidelity of transcription elongation, RNAP pausing, backtracking, and even disassociation of the TEC in the middle of genes can still occur [20, 21].

I.I.iii. Mechanism of transcription termination

Transcription termination takes place (either in the absence of presence of termination factor Rho) until the RNAP reaches a terminator sequence after the end of a gene or because the RNAP stalls. Termination releases the mRNA from RNAP and dissociates the complex from the DNA template [22-24]. In *E.coli*, there are two types of transcription termination: intrinsic termination and Rho-dependent termination.

Intrinsic termination typically occurs at DNA sequences consisting of a palindromic GC-rich region followed by a series of thymines [24], which forms a hairpin that can disrupt the binding of the TEC [24-27]. Rho-dependent termination relies on the interaction between the termination factor Rho and the NusG protein [28]. The Rho factor recognizes the *rut* sequences on the mRNA, looping the mRNA through its central channel [29], and then pulls the 3' end of the mRNA out of RNAP, thereby leading to transcription termination [30-34]. Both types of termination are frequently found in different parts of the genome [35, 36]. The intrinsic termination sites have some structural constraints like A-region, U-region, and hairpin [37]. In contrast, Rho dependent termination requires an untranslated transcript region with at least 85-90 nucleotides [38, 39].

I.II. Mechanisms of mRNA degradation

To maintain homeostasis, the mRNAs produced by transcription must embrace the ultimate fate of being degraded. There are two main mechanisms of mRNA degradations in *E.coli*: (i) 5' end-dependent mRNA degradation and (ii) 5' end-independent mRNA degradation. These two processes both cleave long mRNA transcript into short mRNA fragments. In addition to these active processes for degrading mRNAs, mRNAs are also in effect removed from cells via the dilution that occurs with cell growth and cell division, which will be briefly discussed in the next section on the regulation of mRNA concentrations.

I.II.i. 5' end-dependent mRNA degradation

mRNA degradation at the 5' end begins with the trimming of the 5' end triphosphate to a monophosphate by RNA pyrophosphohydrolase RppH [40, 41]. Deletion of the gene for RppH in bacteria increases the half-lives of mRNAs between three- and eleven-fold [42]. An mRNA with a monophosphate at the 5' end is more likely to be bound and more efficiently cleaved by RNase E which has endoribonuclease activities [42-46]. The cleavage of the mRNA by RNase E generates mRNA fragment with a 5′ monophosphate which are subject to further 5′ end-dependent degradation [40]. RNase G is a homolog of RNase E with about 50% similarity in the amino acid sequence [47] and it shares the same preference as RNase E for cleaving mRNAs with a 5′ monophosphate and similar cleavage sites [48, 49]. Despite the similarities, the role of RNase G is very limited compared to RNase E, and can often be ignored [50]. Recently, a new study also found that 35%~50% of mRNAs in E.coli are diphosphorylated indicating that removal of phosphates at the 5′ end by RppH is a two-step procedure [51]. However, this does not mean that pyrophosphate removal by RppH is the rate-limiting step. Instead, the decay of many mRNAs are still controlled by the cleavage of the monophosphorylated mRNA fragments generated in the process of degradation [52].

I.II.ii. 5′ end-independent mRNA degradation

An alternative pathway to 5′ end dependent mRNA degradation occurs by RNase E cleavage at AU-rich sequences with little specificity for sequence and secondary structures to generate short mRNA fragments [53-56]. It has been proposed that this process is promoted when RNase E binds to multiple unpaired regions simultaneously [53]. Alternatively, it may occur when RNase E binds to the 5′ end of the mRNA and then the mRNA loops so that the RNase E can simultaneously bind internal distant sequences within the same mRNA and cleave them   [57]. Although the specific mechanism and details of this pathway have not been completely defined, it raises the possibility that under some conditions and for some mRNAs they can be cleaved at a downstream site before upstream sites (*i.e.* cleavage of an mRNA may occur at a site that is 3′ to 5′ sites).

I.II.iii. The digestion of mRNA fragments by exoribonuclease

Exoribonucleases in *E.coli* (including RNase II and polynucleotide phosphorylase (PNPase)) process mRNA fragments and poly (A) tails from the 3′ end of RNAs by successive removal of nucleotides for recycling [58, 59]. In general, exonucleases are thought to be less active without preceding cleavage by endoribonucleases due to the frequent secondary structures at the 3′ end of native mRNAs [60]. PNPase can also degrade uncleaved mRNAs with the assistance of RNA helicase RhlB as part of the *E. coli* degradosome [61, 62]. These exoribonucleases are also referred to as oligoribonuclease when they act on very short mRNA fragments and breaks them into single nucleotides [63]. Even for these short mRNA fragments, direct degradation by exonucleases is believed to be extremely rare without preprocessing by RNase E [64, 65], which may be due to the need for RNase E to disrupt secondary structures. mRNAs with extensive secondary structures, which may have short mRNA fragment that still have extensive secondary structure after RNase E cleavage, may require a different exoribonuclease RNase R to be processed into single nucleotides [66-68]. Note: specific exoribonucleases that are not primarily involved in mRNA degradation, such as those that primarily degrade tRNAs (e.g. RNase T,) are not discussed here [64, 69].

II. Regulation of mRNA concentrations

In the previous section, the key processes and their mechanisms in mRNA production and degradation were discussed. However, mRNAs are not produced and degraded spontaneously without regulatory control. Cells have developed a variety of mechanisms that regulate mRNA concentrations. In this part, four categories of regulatory control are discussed: genome structures (II.I), translational feedbacks (II.II), sRNA regulation (II.III), and change in subcellular localization (II.IV).

II.I. Genome structures

The genome of *E.coli* is circular and is packed into a region of the cell called the nucleoid [7]. Histone-like proteins such as HU and IHF contribute to the structure of the nucleoid [70], and this can impact at a very basic level the chemistry and physics of transcription, degradation and translation of mRNAs and their regulation by factors in the cell (this is examined in Chapter 3). In addition, organizational features of the genome can potentially modulate the control of gene expression, and these include the relative position of a gene in the genome, the clustering and orientation of genes with related functions in the genome, and the organization of genes within the same mRNA (operon). These organizational features are discussed in more detail below.

II.I.i. Gene position

The location of a gene in the genome is an important factor in determining the level of transcription of the corresponding mRNA. According to basic mass action, genes with more copies are expressed at a higher level. In bacteria, the number of copies of a gene before cell division depending on the rate of cell division and the location of the gene on the chromosome. In bacteria it can take 60-90 minutes to replicate the chromosome but cells can double in approximately 20 minutes. That is, bacteria cannot complete chromosome replication when rapidly dividing within a single generation and therefore begin replication of the chromosome for a future generation in advance [71]. As a consequence, there are multiple copies of each gene in a cell during rapid growth. Because chromosome replication starts at the origin of replication (*oriC*), genes that are closer to the origin may have a higher mRNA concentrations due to the greater copy number than those that are distant from the origin [72-74]. Recent studies have also shown that not all chromosomal position effects are due to gene copy number, some are mediated at the level of transcription [75]. That is, different positions on the chromosome can alter access of RNAP, DNA binding proteins including transcription factors, and ribosomes to genes and mRNAs. It has been proposed that transcription initiation is regulated locally and globally by the folding and compaction of the chromosomes [76]. The detailed mechanisms and structure modification of the chromosomes remain unclear and require further studies. In summary, gene expression can be affected by gene position due to gene copy number (which in turn is due to the distance of genes from the origin and chromosome replication) and non-gene copy number effects such as how easily regulatory factors can access a gene.

II.I.ii. Gene clustering and orientation

It has been observed that highly expressed genes and genes in the same pathway tend to occur next to each other on the chromosome and that essential genes can have a preference for the leading strand of the chromosome. While it has been speculated that these observations may be due to effects on gene regulation, studies have shown that gene clustering and orientation have minimal effects on mRNA production and gene noise [72, 77]. It has therefore been proposed that gene clustering may be a consequence of easier acquisition and selection of genes if they are co-acquired with functionally related genes via horizontal gene transfer rather than due to selection for a n effect that gene clustering may have on gene regulation [78, 79]. The reason for the higher proportion of essential genes on the leading strand is not known.

II.I.iii. Operon

An operon is a single mRNA that contains multiple genes with related functions [80]. The expression of genes within one operon has been shown to depend on its relative position [81, 82]. Specifically more proximal genes have higher expression than distal genes. This relationship can be explained by two possible mechanisms. Firstly, RNAP must first transcribe the proximal gene before the distal gene. That is, a distal gene can only be transcribed if the proximal gene is transcribed and the RNAP has not dissociated or terminated transcription, but the converse is not the case. Also, like the copy number effect in DNA replication, the more proximal a gene occurs in an operon the earlier it can be transcribed and therefore it will be more likely to have a greater copy number because it’s simply closer to the site of transcription initiation and has a greater likelihood of completed transcription. Secondly, genes that are earlier in the operon can begin translation before the distal gene have even been transcribed. For short lived mRNAs, this mechanisms can have a substantial impact on protein concentrations.

II.II. Impact of translation

Translation is the process by which ribosomes translate the codons on the mRNA template to amino acids and synthesize the individual amino acids into a peptide. The frequency of ribosome binding and the velocity of ribosomes on the mRNA directly affect the rate of translation. Because translation and transcription occurs simultaneously on the mRNA, increasing the translation rate can also increase the effective mRNA concentration by preventing transcription termination through the interaction between the leading ribosome and RNAP [83, 84]. In addition, mRNA decay also decreases when the translation rate is high because translating ribosomes along the mRNA shelter RNase cleavage sites [85]. The coordination between ribosomes, RNases, termination factors, and RNAP together makes the transcription and translation highly organized and coupled in bacteria cells.

II.III. Regulation by sRNAs

sRNAs are short non-coding transcripts of approximately 100 nucleotides. sRNAs are involved in the regulation of multiple pathways from oxidative stress to sugar and iron metabolism [86-89]. Unlike constitutively expressed mRNAs, sRNA concentrations usually vary in response to environmental stress or growth [90, 91]. The binding of sRNAs to mRNAs to create an mRNA-sRNA duplex can alter translation or degradation of the mRNA to increase protein concentrations (activating sRNAs) or decrease protein concentrations (repressor sRNAs) [92, 93]. Only small proportion of sRNAs seems to be activators [94]. Most sRNAs appear to be repressors and are often co-degraded with the target mRNA after duplex formation [95]. sRNAs can either bind their target mRNAs with perfect complementarity or only partially paired with their target mRNAs [86]. The latter can be highly sequence specific and facilitated by an RNA chaperone, Hfq [96, 97].

sRNAs can regulate mRNA transcription and mRNA degradation [98]. 6S sRNA is able to inhibit transcription initiation by sequestering the most common sigma factor (σ70) and preventing RNAP from forming the transcription initiation complex [99]. Premature transcription termination can also be affected by sRNA binding which can increase target mRNA through the transformation of the secondary structures on the mRNA near the sRNA binding site [100, 101]. sRNAs can also decrease or increase mRNA translation. sRNAs can decrease mRNA translation by several mechanisms, the most common of which is binding near the ribosome binding sequence to prevent the access of ribosomes. The decreased translation can also make the target mRNA more vulnerable to the attack of RNase resulting in notable increase in mRNA degradation [102-105]. Less commonly, sRNAs increase translation primarily by binding to the mRNA to alter secondary structure so that the ribosome binding sequence is more accessible to ribosome. These mechanisms give sRNAs a critical role in the regulation of mRNA concentrations particularly during stress.

II.IV. Impact of subcellular localization

mRNAs in *E.coli* can have diverse localization pattern in the cell. For example, the antiterminator *bglG* and the cytoplasmic chloramphenicol acetyltransferase *cat* mRNAs have been shown to have polar and cytoplasmic localization respectively [106]. Some mRNAs are localize to the site where the protein products are functional [106, 107]. In other cases, mRNAs are transported to specific destinations in the cell before the completion of translation [108, 109]. The *ptsg* membrane localizing mRNA is transported to specific sites in the cell where it is translated, and at this site the translation can be regulated by several factors including by sRNAs [110]. Because mRNA concentrations have been shown to be dependent on the translation, improperly localized mRNAs might be degraded or terminated due to the absence of translating ribosomes.

The localization of mRNAs is also important because the factors that regulate the degradation and translation of mRNA can also localize to different parts of cell. For instance, ribosomes mostly localize in the cytoplasm rather than the nucleoid [111]. Also, RNaseE has been shown to be primarily localize to the cell membrane. Therefore mRNA localization is an important factor in the regulation of mRNA degradation and translation.

III. Summary

In summary, there are many known factors that can modulate gene expression and specifically the production, degradation and translation mRNA concentrations. There are also many factors that are specific and important to individual genes and pathways, and of course there are many biomolecules, pathways and genes that have yet to be discovered and/or understood in terms of their contribution to the regulation of genes. On top of this there is the complexity that arises in cells due to the interaction of all these mechanisms that the cell exploits and benefits from. In my dissertation, I will touch different aspects of these regulatory mechanisms and explain my contribution to our knowledge of regulations of gene expression and mRNA concentrations in *E.coli*.

Chapter 1: Systematic and general method for quantifying localization in microscopy images[1]

1.1 Abstract

Quantifying the localization of molecules with respect to other molecules, cell structures, and intracellular regions is essential to understand their regulation and actions. However, measuring localization from microscopy images is often difficult with existing metrics. Here we evaluate a metric for quantifying localization that is termed the threshold overlap score (TOS), and show it is simple to calculate, easy to interpret, able to be used to systematically characterize localization patterns, and generally applicable. TOS is calculated by: (i) measuring the overlap of pixels that are above the intensity thresholds for two signals; (ii) determining whether the overlap is more, less, or the same as expected by chance (*i.e.* colocalization, anticolocalization, or noncolocalization); and (iii) rescaling to allow comparison at different thresholds. The above is repeated at multiple threshold combinations to generate a TOS matrix to systematically characterize the relationship between localization and signal intensities. TOS matrices were able to identify and distinguish localization patterns of different proteins in various simulations, cell types and organisms with greater specificity and sensitivity than common metrics. For all the above reasons, TOS is an excellent first line metric, particularly for cells with mixed localization patterns.

[1] This chapter has been previously published as "Systematic and general method for quantifying localization in microscopy images" in *Biology open* (2016), 5(12), 1882-1893.

1.2 Introduction

Quantifying the localization of proteins, RNAs, and complexes within cells can help determine their regulation and sites of action [112-114]. Therefore the development and evaluation of metrics to quantify localization is an important and shared goal of many different disciplines. Three common approaches to quantifying localization are: (i) measuring the fraction of two signals that overlap [112, 115]; (ii) measuring the correlation or rank-order correlation of pixel intensities for two signals [112, 116]; and (iii) identifying objects and determining their fractional overlap or the distance separating them [115, 117]. These metrics and less common alternatives [114, 115] have been successfully used in many applications. However, there are also many types of images and samples where the above metrics do not perform well and their results are difficult to interpret [113, 115, 118, 119], inconsistent [113], and/or susceptible to arbitrariness and bias [119].

Metrics often encounter difficulty when images and samples have: a signal of low intensity compared to background and non-specific signals [113, 120], a large proportion of pixels with background or non-specific signals [113, 121], a nonlinear relationship between two signals [115], or mixed patterns of localization. Other important barriers to the use of some metrics include: limited testing (and consequently researchers are uncertain the metric is suitable for their samples and application), underlying assumptions that limit their general application, and the need for customization or simulations that require specialized knowledge and skills. All these issues are common resulting in researchers in many disciplines resorting to qualitative (and often inaccurate) assessments of localization by simply superimposing (or “merging”) images [112, 113]. No single metric or analysis protocol will meet all requirements for all researchers [112, 113], but clearly additional tools to quantify localization are needed.

In this study we evaluated a metric for localization termed the threshold overlap score (TOS), which measures the overlap in signals above threshold intensity values. We use “localization” and “localization pattern” to refer to the measurement of overlap. If the overlap is greater than, less than, or the same as expected by chance then localization is categorized as “colocalization”, “anticolocalization” or “noncolocalization”, respectively. The first part of the study derives TOS and then describes a strategy of using it at many combinations of thresholds to generate a TOS matrix that can identify and distinguish features in mixed patterns of localization. The second part of the study applies TOS analysis to simulated data and experimental data obtained from public image repositories. The latter showed that values from the TOS matrix can distinguish the localization patterns of different proteins for a variety of cell types and organisms, and that they can do so with greater specificity and sensitivity than common metrics (Pearson’s correlation coefficient, Manders’ colocalization coefficients, and Spearman’s rank correlation coefficient).

1.3 Results

1.3.1 Calculating the threshold overlap score (TOS)

The first step in calculating TOS is measuring the observed fraction of pixels that exceed the threshold of one signal that also exceed the threshold of a second signal (**Figure 1.1A**). That is, measuring the "fractional area of overlap" (abbreviated to "AO"). Instead of choosing thresholds by selecting specific values for the intensities, which in turn specify fractions of pixels for signals 1 and 2 ("F_{T1}" and "F_{T2}" respectively), we directly chose these fractions (following rank ordering of the pixels by intensity). This approach of specifying thresholds in terms of selected fractions rather than as values allows observed data from individual cells that have different intensities and total numbers of pixels to be more easily combined. Therefore,

$$\text{observed AO}_1 = \frac{\text{fractional area above signals 1 and 2 thresholds}}{\text{fractional area above signal 1 threshold}} \text{ and}$$
$$\text{observed AO}_2 = \frac{\text{fractional area above signals 1 and 2 thresholds}}{\text{fractional area above signal 2 threshold}}. \qquad \text{Eq. 1.1}$$

The second step is normalizing the observed AO_1 and AO_2 by their expected overlaps (for uniformly distributed random signals), which are F_{T2} and F_{T1} respectively, resulting in the AO ratio. Because it may seem counterintuitive that AO_1 and AO_2 are normalized by the threshold of the other signal we consider the example of a cell with 100 pixels and selected fractions for signal 1 and 2 of 50% ($F_{T1} = 0.5$) and 10% ($F_{T2} = 0.1$) respectively. In this example, 50 and 10 pixels are selected for signals 1 and 2 respectively. If the selected pixels for signal 1 are uniformly distributed throughout the cell, then half of them would be expected to overlay the selected pixels for signal 2 (irrespective of their distribution), which is 5 pixels. For the selected pixels of signal 1, this expected 5 pixel overlap represents 0.1 of them (*i.e.* 5 out of 50), which is equal to the selected fraction for signal 2 (F_{T2}) as stated above. This normalization assumes a null distribution with pixel intensity values uniformly distributed across the cell and independent (note: the point spread function with autocorrelation between pixels does not alter the predicted value but it does affect its variance [113]). From **Eq. 1.1**,

$$\text{AO ratio} = \frac{\text{fractional area above signal 1 and 2 thresholds}}{\text{fractional area above signal 1 threshold} \times \text{fractional area above signal 2 threshold}}. \qquad \text{Eq. 1.2}$$

The AO ratio has the same value when calculated from the observed AO_1 or AO_2. The AO ratio ≈ 1, > 1 or < 1, when the pixels above the threshold of each signal overlap by the same, greater than, or less than the null distribution. The minimum AO ratio depends on $F_{T1} + F_{T2}$ (**Figure 1.1B**). If $F_{T1} + F_{T2} \leq 1$, the minimum AO ratio can be zero because it is possible for the selected pixels for each signal to not overlap (note: the AO ratio is never undefined because both F_{T1} and $F_{T2} > 0$). However, if $F_{T1} + F_{T2} > 1$ the minimum AO ratio cannot be zero because it is impossible for the selected pixels for each signal to overlap less than the sum of $F_{T1} + F_{T2}$ minus 1. The maximum AO ratio also depends on F_{T1} and F_{T2} (**Figure 1.1B**). Specifically, the maximum occurs when the smaller of the selected fractions completely overlaps the larger.

$$\text{Minimum AO ratio} = \begin{cases} \frac{F_{T1} + F_{T2} - 1}{F_{T1} \times F_{T2}}, \text{when } F_{T1} + F_{T2} > 1 \\ 0 \quad\quad\quad , \text{ when } F_{T1} + F_{T2} \leq 1 \end{cases}. \quad \text{Eq. 1.3}$$

$$\text{Maximum AO ratio} = \frac{\text{minimum}\,\{F_{T1}, F_{T2}\}}{F_{T1} \times F_{T2}}. \quad \text{Eq. 1.4}$$

The limits are 0 and 1 for the minimum AO ratio and 1 and $+\infty$ for the maximum AO ratio. When the AO ratio = 1, the observed overlap is the same as expected for the null distribution.

The third step is rescaling AO ratios so they can be compared for different thresholds (**Figure 1.1C**). This is necessary because the AO ratio depends on the product of F_{T1} and F_{T2} (see **Eq. 1.2-1.4**). For example, an AO ratio with 100% overlap will be 2 or 10 depending on whether 50% or 10% of the pixels are selected for both signals. Another reason for rescaling is the inherent asymmetry of ratios. Quadrupling the numerator increases the AO ratio from 1 to 4 while quadrupling the denominator decreases it from 1 to 1/4; the latter is a much smaller absolute change. We found that a simple linear rescaling works well and the results are far easier to interpret.

Linear rescaling generates a metric called the threshold overlap score (TOS). TOS rescales the AO ratios so they have a range from −1 to +1 for all thresholds and have a value of 0 when the observed overlap is exactly the same as expected for the null distribution (**Figure 1.1D**).

$$\text{TOS} = \begin{cases} 0\,, \text{ when AO ratio} = 1 \\ \frac{1 - \text{AO ratio}}{\text{minimum AO ratio} - 1}, \text{when AO ratio} < 1 \\ \frac{\text{AO ratio} - 1}{\text{maximum AO ratio} - 1}, \text{when AO ratio} > 1 \end{cases} \quad \text{Eq. 1.5}$$

The magnitude of TOS reflects how much the overlap lies between the null hypothesis and the maximum or minimum values; for example an absolute value of TOS = 0.9 is nine tenths between the null distribution and the maximum or minimum possible overlap.

Another approach to rescaling is to logarithmically transform the data, which has some advantages but the rescaled values are not easily interpreted. As with linear rescaling, logarithmic rescaling was performed so the minimum and maximum AO ratios after rescaling are −1 and +1 respectively, and an AO ratio of 1 after rescaling is equal to 0 (*i.e.* the null distribution after rescaling = 0). The logarithmically rescaled AO ratios are referred to as “log TOS”. We solved the function

$$\text{log TOS} = \alpha \cdot \ln(\beta \cdot \text{AO ratio} + \gamma)\,, \quad \text{Eq. 1.6}$$

for the above fixed points (note: ln = natural log). That is, we determined the constant coefficients α, β and γ from the following three equations:

$$\begin{cases} -1 = \alpha \cdot \ln(\beta \cdot \text{minimum AO ratio} + \gamma); \\ \quad 0 = a \cdot \ln(b \cdot 1 + c)\,; \\ +1 = \alpha \cdot \ln(\beta \cdot \text{maximum AO ratio} + \gamma). \end{cases} \qquad \text{Eq. 1.7}$$

The coefficients are:

$$\alpha = \frac{1}{\ln\left(\frac{\text{maximum AO ratio - 1}}{\text{1 - minimum AO ratio}}\right)},\ \beta = \frac{\text{maximum AO ratio + minimum AO ratio - 2}}{(\text{maximum AO ratio-1})\cdot(1\text{ - minimum AO ratio})}\ \text{, and } \gamma = 1\text{- }\beta. \qquad \text{Eq. 1.8}$$

Eq. 1.6 can be more simply expressed with the coefficient γ substituted in terms of β:
log TOS = $\alpha \cdot \ln(\beta \cdot (\text{AO ratio - 1}) + 1)$. Eq. 1.9

Coefficients α and β can be rewritten in terms of the selected fractions F_{T1} and F_{T2}, which are defined in the main text as:

$$\text{Minimum AO ratio} = \begin{cases} \frac{F_{T1} + F_{T2} - 1}{F_{T1} \times F_{T2}}, \text{ when } F_{T1} + F_{T2} > 1 \\ 0\text{ , when } F_{T1} + F_{T2} \leq 1 \end{cases} \qquad \text{Eq. 1.10}$$

$$\text{Maximum AO ratio} = \frac{\text{minimum}\left\{F_{T1}, F_{T2}\right\}}{F_{T1} \times F_{T2}}. \qquad \text{Eq. 1.11}$$

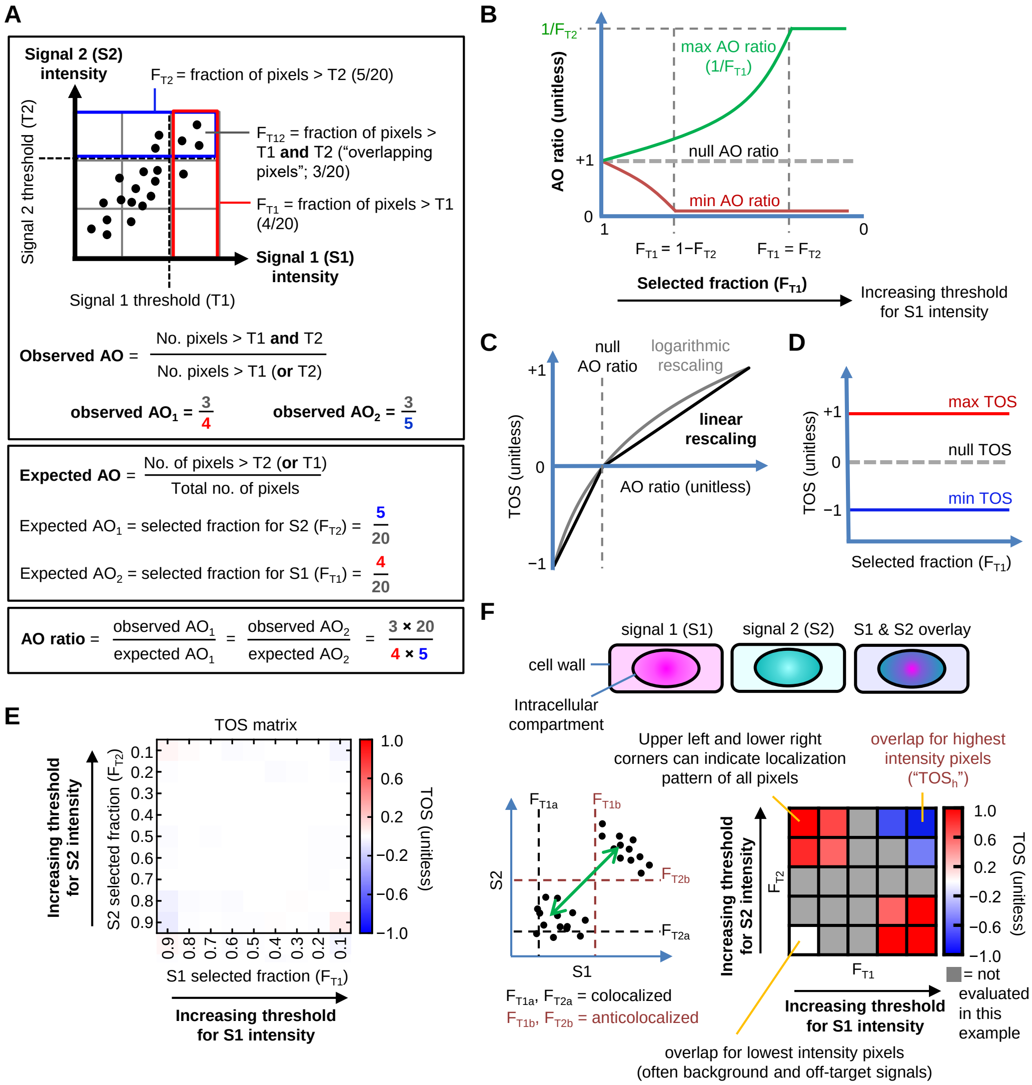
F_{T1} and F_{T2} are fractions and therefore must be greater than or equal to 0, and less than or equal to 1. For simplicity, we let $F_{T1} \leq F_{T2}$ (note: this is not a constraint because the designation of signal 1 and 2 is arbitrary). Substitution of **Eq. 1.10** and **Eq. 1.11** into the expressions for the coefficients α and β in S3 gives:

$$\alpha = \begin{cases} \frac{1}{\ln(\frac{F_{T1}}{1-F_{T1}})}, \text{ when } F_{T1} + F_{T2} > 1 \\ \frac{1}{\ln(\frac{1-F_{T2}}{F_{T2}})}, \text{ when } F_{T1} + F_{T2} \leq 1 \end{cases} \text{ and } \beta = \begin{cases} \frac{F_{T2}\,(2F_{T1} -1)}{(1-F_{T2}\,)(1-F_{T1}\,)}, \text{ when } F_{T1} + F_{T2} > 1 \\ \frac{1-2F_{T2}}{1-F_{T2}}, \text{ when } F_{T1} + F_{T2} \leq 1 \end{cases}. \qquad \text{Eq. 1.12}$$

These expressions show that F_{T1} or F_{T2} must not be = 0.5 otherwise α and β are undefined. If F_{T1} or F_{T2} are equal to 0.5 then a simple linear function is required to rescale the AO ratio to achieve the fixed points. In addition, α and/or β are undefined at the limits of F_{T1} and F_{T2} when they are equal to 1 or 0.

A potential advantage of using log TOS is that there is no discontinuity in the first derivative. However, a disadvantage of log TOS is that values are difficult to interpret unlike the linear TOS where values are the fractions from the null ratio (*i.e.* the observed overlaps is equal to the expected overlap) to the maximum or minimum possible AO ratio. Furthermore, care must be taken to not calculate log TOS at selected fractions for which it is undefined.

It is helpful to divide the spectrum of possible TOS values into categories of “colocalization”, “anticolocalization” and “noncolocalization”. In doing so, it is important to recognize that TOS values may be too small to be biological relevant [122, 123] even if they show statistically significant differences from the null distribution. It is also not useful to define noncolocalization as exactly equal to zero because very few samples would be in this category. For these reasons, we recommend defining noncolocalization as a range of values (*e.g.* TOS between −0.1 and +0.1). A practical advantage of defining noncolocalization as a range is that a “true” noncolocalized pattern can be consistently referred to as such, rather than as “weak colocalization” in one measurement and ‘weak anticolocalization” in another due to measurement error and randomness in biological variation. It must be stressed that these bounds are for the convenience of interpretation and do not affect the analysis itself, and that the definition of noncolocalization should be guided by the design and purpose of the study.



A
Signal 2 (S2) intensity
F_T2 = fraction of pixels > T2 (5/20)
Signal 2 threshold (T2)
F_T12 = fraction of pixels > T1 and T2 (“overlapping pixels”; 3/20)
F_T1 = fraction of pixels > T1 (4/20)
Signal 1 (S1) intensity
Signal 1 threshold (T1)
Observed AO = No. pixels > T1 and T2 / No. pixels > T1 (or T2)
observed AO1 = 3/4
observed AO2 = 3/5
Expected AO = No. of pixels > T2 (or T1) / Total no. of pixels
Expected AO1 = selected fraction for S2 (F_T2) = 5/20
Expected AO2 = selected fraction for S1 (F_T1) = 4/20
AO ratio = observed AO1 / expected AO1 = observed AO2 / expected AO2 = 3 × 20 / 4 × 5
B
1/F_T2
max AO ratio (1/F_T1)
AO ratio (unitless)
+1
null AO ratio
min AO ratio
0
1
0
F_T1 = 1−F_T2
F_T1 = F_T2
Selected fraction (F_T1)
Increasing threshold for S1 intensity
C
+1
null AO ratio
logarithmic rescaling
linear rescaling
TOS (unitless)
0
AO ratio (unitless)
−1
D
max TOS
+1
TOS (unitless)
null TOS
0
min TOS
−1
Selected fraction (F_T1)
E
TOS matrix
Increasing threshold for S2 intensity
S2 selected fraction (F_T2)
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
S1 selected fraction (F_T1)
Increasing threshold for S1 intensity
1.0
0.6
0.2
−0.2
−0.6
−1.0
TOS (unitless)
F
signal 1 (S1)
signal 2 (S2)
S1 & S2 overlay
cell wall
Intracellular compartment
Upper left and lower right corners can indicate localization pattern of all pixels
overlap for highest intensity pixels (“TOS_h”)
F_T1a
F_T1b
S2
F_T2b
F_T2a
S1
F_T1a, F_T2a = colocalized
F_T1b, F_T2b = anticolocalized
Increasing threshold for S2 intensity
F_T2
F_T1
Increasing threshold for S1 intensity
1.0
0.6
0.2
−0.2
−0.6
−1.0
TOS (unitless)
= not evaluated in this example
overlap for lowest intensity pixels (often background and off-target signals)

Figure 1.1 Calculating the threshold overlap score (TOS) and generating TOS matrices. (**A**) Calculation of observed AO, expected AO, and AO ratio. Thresholds are measured by the fraction of pixels with higher intensity in the cell (*i.e.* the "selected fraction", which are F_{T1} and F_{T2} for signal 1 and 2) as explained in the main text. (**B**) Diagram showing maximum, minimum and expected AO ratios as a function of the threshold for signal 1 (*i.e.* F_{T1} is varied and F_{T2} is fixed). Note: expected AO ratio is for the null distribution. (**C, D**) Threshold overlap score (TOS) is obtained from the AO ratio by rescaling linearly (or logarithmically) so the maximum, minimum and null values are +1, −1 and 0 for all selected fractions. (**E**) TOS matrix generated by simulating a uniform distribution for all threshold combinations (n = 500 for each selected fraction). As predicted, the observed AO values for the simulated uniform distribution are close to the expected AO values therefore the AO ratio is ≈ 1 and TOS is ≈ 0 at all threshold combinations. (**F**) Thresholds can affect quantification and characterization of localization. Hypothetical cells with mixed intracellular localization patterns for two signals (S1 and S2). Cells have uncorrelated off-target signals and negatively correlated on-target signal (note: although the off-target signals appear uniform, the signals have variation as shown in the scatterplots). Scatterplot to the left show that the determination of the localization pattern depends on the threshold selected. Thresholds at low signal intensities (F_{T1a} and F_{T2a}) will measure localization of both off-target and on-target signals and together they have a net positive correlation as shown by the green arrow (*i.e.* colocalization). Thresholds at high intensities (F_{T1b} and F_{T2b}) will measure localization of only the on-target signal, which has a negative correlation (*i.e.* anticolocalization).

1.3.2 Generating TOS matrices

One of the most difficult aspects of measuring localization is selecting the thresholds [112]. Thresholds can affect the contribution to the analysis of background signals from the imaging system (if it is not subtracted) and from cells (*e.g.* autofluorescence) and low intensity signals from unbound or non-specifically bound fluorescent, chemiluminescent or colorimetric probes or stains. These low intensity signals, which we refer to as non-specific or "off-target" signals, typically have higher intensity than background signal but lower intensity than "on-target" signals where the probe or stain has localized to the biological target. Separating background and off-target signals is often difficult, and it is typically more important to distinguish both of them from on-target signals. Even if there is no background and off-target signals, threshold selection can affect quantification of localization. Therefore protocols have been developed to make threshold selection less arbitrary [112, 124] but they do not always function well, especially when on-target signals are anticorrelated or the background and off-target signals are high or correlated [113]. Furthermore, cells often have mixed localization patterns making the evaluation of localization at a single set of thresholds, no matter how they are chosen, an inaccurate ensemble description of localization. Based on all the above, we chose to systematically calculate TOS at many different thresholds resulting in a TOS matrix, which can be viewed as a heat map (**Figure 1.1E**). As will be shown below, the TOS matrix can reveal trends between localization and signal intensity, and allow the identification of multiple localization patterns within cells and organisms.

A TOS matrix can in theory be generated by taking any number of combination of thresholds for each channel, ranging from a single set of thresholds to a set of thresholds from every different pixel intensity value in a cell. The former would create a matrix with one element and the latter would create a matrix with up to N^2 elements, where N is the total number of pixels in the cell.

While having the maximum number of combinations will give maximum resolution, creating it for every cell would be problematic for many reasons including: (i) being too computationally intensive; and (ii) resulting in different sized TOS matrices for each cell (because they have different numbers of pixels), which makes it harder to combine them (see below). It must also be kept in mind that if thresholds are taken at the very highest and lowest selected fractions there may be too few overlapping or non-overlapping pixels respectively for statistical significance unless large number of cells are measured (note: *a priori* statistical power can be estimated with standard parametric tests and then increased by up to 15% to account for *post-hoc* non-parametric tests having less power [125, 126]). While bins at the highest thresholds will have a lower number of pixels, these pixels will have the highest numbers of reporter molecules (hence the higher intensity signal) and thus tend to have a lower coefficient of variation.

We chose an intermediate number of threshold combinations (specifically 81 combinations) and found that it gave more than adequate resolution to detect different patterns of localization in our simulations and the experimental data we analyzed (see below). These threshold combinations were 9 selected fractions for signals 1 and 2 (F_{T1} and F_{T2}) from 0.9 to 0.1 in increments of 0.1. Initially, 10% of pixels with the lowest intensity pixels are removed from both signals (leaving a selected fraction of 90% of the pixels; *i.e.* F_{T1} and F_{T2} = 0.9), then 20% of the lowest intensity pixels in the entire cell are removed for one or both signals (leaving a selected fraction of 80% for signal 1 or 2), and so on, until 90% of the lowest intensity pixels in the entire cell are removed for one or both signals (leaving a selected fraction of 10% for signal 1 and 2). Note: F_{T1} and F_{T2} = 1 were not included in the analysis because these selected fractions correspond to 100% of the pixels in the cell therefore all selected pixels must overlap and TOS = 0.

It can be necessary and convenient to extract values from TOS matrices that quantify specific features of mixed localization patterns, and three values that were found to be especially useful were (see below): (i) TOS at the highest thresholds (TOS_h), which correspond to the lowest selected fractions (**Figure 1.1F**); (ii) the maximum TOS in the matrix (TOS_{max}), which if > 0 specifies thresholds with maximum colocalization; and (iii) the minimum TOS in the matrix (TOS_{min}), which if < 0 specifies thresholds with maximum anticolocalization. TOS_h was chosen because many analyses will want to specifically measure the localization pattern of the on-target signal, which will usually be most separated from any background and off-target signals at the highest intensity values (see data below). If the localization pattern is similar at all intensities or the localization pattern of the pixels with the highest signal intensity is not reflective of the biology (including due to noise) then a lower threshold should be selected. Other criteria could also be used to select values from the TOS matrix (see **Discussion**) and their selection should be guided by the experimental system, biological questions, and the heterogeneity of the data.

1.3.3 Interpretation of TOS analysis in samples with mixed localization patterns

We simulated cells to demonstrate how the TOS matrix appears with mixed patterns of localization. The simulated cells had two subpopulations of pixels, which for simplicity had equal counts and uniformly distributed random noise (range 0 to 1 × 10^4 arbitrary units (a.u.)). Population 1 was either positively correlated, uncorrelated, or negatively correlated for signals 1 and 2 (scatterplots, **Figure 1.2A-C**) and population 2 always had uncorrelated signals 1 and 2 (scatterplots, **Figure 1.2A-C**). The two populations initially overlay one another (mean = 6.5 × 10^4 a.u. for both signals; middle scatterplots in **Figure 1.2A-C**). The mean of population 1 or 2 was decreased in 40 equal increments (until the mean = 2.5 × 10^4 a.u.) (left and right scatterplots, **Figure 1.2A-C**). The population with the lower intensity signal can be considered to represent background and off-target signals and the population with the higher intensity signals can represent the on-target signal. Note: the absolute values and units of pixel intensity are not important because the pixels are rank-ordered according to intensity and thresholds are a selected fraction of the pixels rather than values.

For every distribution created we calculated a TOS matrix. But only those distributions where populations 1 and 2 are at their lowest and highest mean intensities (lower left and lower right in **Figure 1.2A-C**), and where populations 1 and 2 maximally overlay each other (low center in **Figure 1.2A-C**) are displayed as heat maps. Visual inspection of the TOS matrices shows the upper right corner accurately reflects whether pixels with the highest intensities are colocalized (positively correlated), noncolocalized (uncorrelated) or anticolocalized (negatively correlated). Similarly, the lower left corner of the TOS matrices reflects whether pixels with the lowest intensities are colocalized, noncolocalized or anticolocalized.

We extracted TOS_h, TOS_{max} and TOS_{min} from the TOS matrices and plotted them as a function of the distance between populations 1 and 2 (with negative and positive shifts indicating the means of signals 1 and 2 for population 2 are less and more respectively than population 1) (**Figure 1.2D-F**). TOS_h correctly shows when the population with the higher means was colocalized, anticolocalized or noncolocalized (*i.e.* $TOS_h > 0.1$, ~ 0 and < -0.1 respectively) (**Figure 1.2D-F**). TOS_{max} identified when the combined population (*i.e.* populations 1 and 2 considered together) displayed colocalization but was insensitive to the localization patterns of each subpopulation. TOS_{min} was sensitive to the presence of anticolocalization in population 1 irrespective of whether it had the lower or higher mean of the two populations (scatterplots in **Figure 1.2C** and black line in **Figure 1.2F**). Values for TOS_h, TOS_{max} and TOS_{min} were found to be relatively insensitive to the mean intensity and distance between the populations; that is, they provide a robust measure of localization.

In summary, the TOS matrix is helpful in distinguishing colocalization, anticolocalization, and noncolocalization within mixed patterns of localization. TOS_h, TOS_{max} and TOS_{min} from the TOS matrices appear to be robust measures of: (i) the localization pattern of the on-target signal, (ii) colocalization for all signals together (background, off-target and on-target), and (iii) the presence of anticolocalization within a mixed localization pattern.

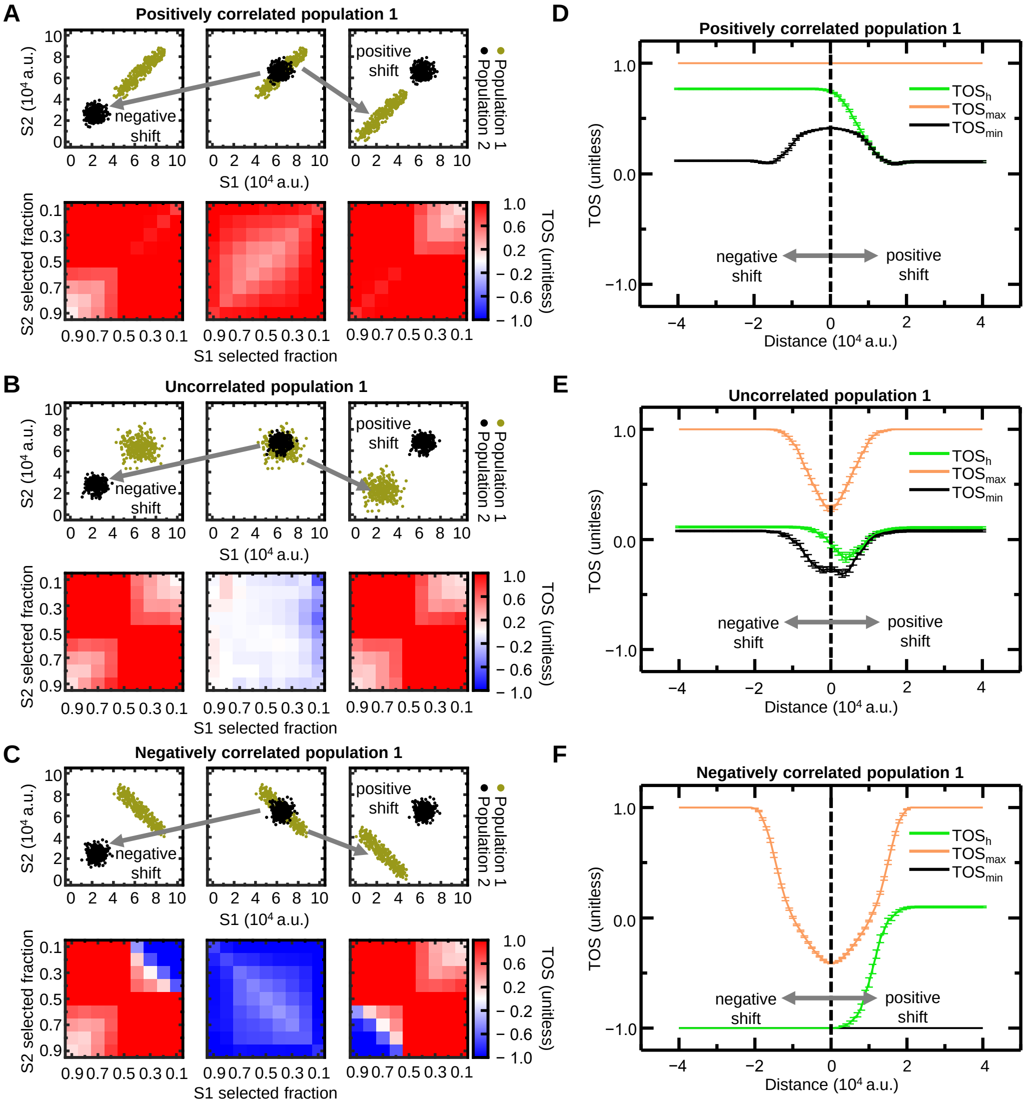


Figure 1.2 Interpretation of TOS analysis in samples with mixed localization patterns. (**A-C**) Scatterplots and their corresponding TOS matrices for images with two populations of pixels (*n* = 300 for each population). Population 1 was positively correlated (**A**), uncorrelated (**B**), or negatively correlated (**C**), and population 2 was always uncorrelated. The means of the two populations were initially the same (center column) and either population 2 was decreased (left column) or population 1 was decreased (right column). Decreasing population 1 and decreasing population 2 corresponds to the positive and negative distances respectively in **panels D-F**. Populations 1 and 2 were decreased in 40 increments. Only scatterplots and TOS matrices for the initial and final positions are shown. (**D-F**) Mean TOS_h, TOS_{max} and TOS_{min} at each distance of the populations (total = 81). **Panels D**, **E** and **F** correspond to distributions where population 1 is positively correlated, uncorrelated, or negatively correlated. Parameters were calculated from 50 TOS matrices simulated for each distance. Error bars are the standard error of the mean (sem).

1.3.4 Comparison of TOS with other metrics of localization

We next compared TOS_h, TOS_{max} and TOS_{min} to common metrics for evaluating localization using simulated cells with mixed localization patterns. It is not feasible to compare all the alternative metrics to TOS [112-114, 117]. Therefore we selected three metrics that are commonly used by experimentalists, which are Pearson's correlation coefficient (PCC), Manders' colocalization coefficients M1 and M2, and Spearman's rank correlation coefficient (SRCC).

PCC calculates the linear correlation in the intensity of two signals [113]. SRCC evaluates whether the rank order of values for two signals is the same or not and it does not matter whether this monotonic relationship is linear or nonlinear. M1 and M2 are calculated from the sum of the intensities of the pixels that exceed the thresholds for both signals 1 and 2 divided by the sum of the intensities of the pixels that exceed the threshold for signals 1 or 2 respectively [113, 127]. M1 and M2 depend on the fraction of overlapping pixels, the intensities of the pixels, and the thresholds. It has been proposed that the "expected" M1 and M2 (which are equivalent to F_{T2} and F_{T1}) should be subtracted from the observed M1 and M2 respectively, resulting in "M1diff" and "M2diff" [118]. Thresholds for M1 and M2 (and consequently for M1diff and M2diff) are commonly selected using a method (or a variant of it) described by Costes and colleagues [112, 124]. Costes' method evaluates the correlation in pixels below each threshold in the data, and then selects the threshold with the minimum correlation or highest threshold with a non-positive correlation (note: we used the former from JACoP [112]).

We examined how all metrics performed at distinguishing populations of cells with mixed localization patterns for off-target and on-target signals, which are challenging, common and important in localization analysis. The metrics were compared using receiver operating characteristic (ROC) curves, which are commonly used to evaluate image analysis tools and diagnostic tests [128]. To create the ROC curves, we first simulated "condition positive" and "condition negative" populations of cells with mixed localization patterns (**Figure 1.3A**, **B**). Each condition positive cell had an equal combination of pixels with positively correlated off-target signal and positively correlated on-target signal (means = 20,000 and 30,000 a.u. respectively; total pixels per cell = 600). The values were multiplied by a random number from a Gaussian distribution (mean = 1.0 and $\sigma = 0.2$), which was independent for each channel and pixel (**Figure 1.3A**). Each cell in the condition negative population was generated in the same manner except the on-target signal was anticorrelated (**Figure 1.3B**). In the condition positive and negative populations the off-target signal had a slope of $\theta = e^q$, where q had one of 141 equal increments in the range −0.7 to +0.7. Note: the off-target signal was chosen to be positively correlated (rather than uncorrelated or negatively correlated) because this is often harder to threshold and discriminate from on-target signal, and we sought to compare the metrics under challenging conditions. For each slope, 50 cells were simulated resulting in 7050 cells for each condition.

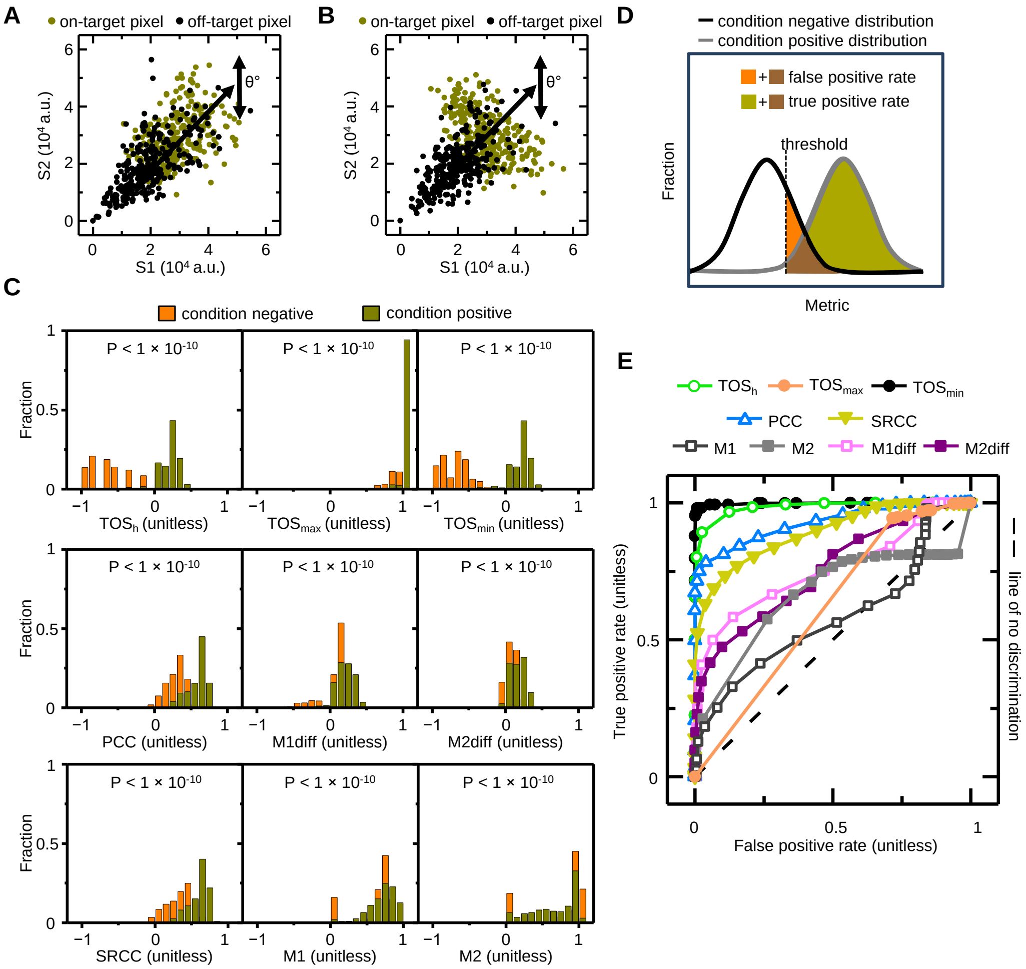


Figure 1.3 Comparison of TOS with other metrics of localization. (**A**, **B**) Representative scatterplots for simulated condition positive (**A**) and condition negative (**B**) cell populations. All cells have positively correlated off-target signal (black symbols) with a slope (θ) that was varied (see main text). Condition positive cells have on-target signal (gold symbols) that is positively correlated and condition negative cells have on-target signal that is negatively correlated. Both conditions have an equal number of off-target and on-target pixels. (**C**) Histograms of TOS_h, TOS_{max}, TOS_{min}, PCC, SRCC, M1, M2, M1diff, and M2diff for the simulated condition positive and negative populations. P values are calculated using a two-tailed Mann-Whitney U test. Note: M2 values appear to exceed 1 because values that are exactly equal to 1 are in the 1 to 1.1 bin. (**D**) Diagram explaining the calculation of the true positive and false negative rates for the metrics in **panel C** (see main text). The fraction of cells in the condition positive population and condition negative population that are above the threshold are the true positive rate and true negative rate respectively. (**E**) Receiver operating characteristic (ROC) curves for each metric (see main text).

We evaluated TOS_h, TOS_{max}, TOS_{min}, PCC, SRCC, M1, M2, M1diff and M2diff for each cell. For every metric, P values were calculated using the two-tailed Mann-Whitney U test which showed all metrics had statistically significant differences in their values for the condition positive and negative populations (displayed in **Figure 1.3C**). Therefore a simple statistical comparison is not helpful in comparing the metrics. Histograms of the values for the condition positive and negative populations were generated and then the fractions of cells in each population above a threshold that slides from highest to lowest value were determined. These fractions for the condition positive and negative populations are the true positive rates (also known as the sensitivity) and false positive rates (which is 1−specificity) respectively (**Figure 1.3D**). The true positive rates were plotted as a function of the false positive rates at each threshold to produce ROC curves for each metric (**Figure 1.3E**). The ROC curve nearest the upper left corner of the plot is closest to an ideal test with perfect classification of localization (*i.e.* 100% sensitivity and 100% specificity).

For the simulations, TOS_{min} was the best classifier followed by TOS_h (**Figure 1.3E**). TOS_{max} did not perform well because there were positive correlations at high selected fractions for both the condition positive and condition negative populations therefore it could not distinguish them (**Figure 1.3A**, **B**). Similarly, PCC was generally positive for both populations, which is why it did not perform as well as TOS_h. M1diff and M2diff did not perform well due to Costes' method for threshold selection (note: also in some cases a threshold could not be identified resulting in undefined M1diff, and M2diff that were not included in the analyses). Consistent with the study mentioned above [118], M1diff and M2diff performed better than M1 and M2, therefore the former two were used in subsequent analyses.

1.3.5 Comparison of TOS with other metrics of localization

We demonstrated the generality of TOS analysis by calculating matrices and extracting TOS_h, TOS_{max}, and TOS_{min} for experiments with different proteins in a variety of cells and organisms, which were obtained from public image repositories (**Materials and Methods**).

The first dataset examined were *Drosophila melanogaster* Kc167 cells (n = 366), which had been probed with fluorescein conjugated phalloidin to identify F-actin in the cytoskeleton and stained with Hoechst 33342 to identify DNA [129, 130] (**Figure 1.4A**). TOS matrices from individual cells were combined and the median TOS value for each threshold combination was presented as a heat map (**Figure 1.4B**). This analysis shows that at most selected fractions the F-actin probe and DNA staining are strongly anticolocalized (*i.e.* TOS << 0), which is expected because they label different parts of the cell (*i.e.* outside the nucleus and in the nucleus respectively). Scatterplots also show anticolocalization with the intensities of F-actin labeling and DNA staining being largely independent (**Figure 1.4C**). TOS_h, TOS_{max}, and TOS_{min} from individual cell TOS matrices were compared to PCC, SRCC, M1diff and M2diff from the same cells. The values of each metric for individual cells were plotted along with the median and 90th and 10th percentile values (note: all metrics have the same range and zero indicates no correlation or noncolocalization) (**Figure 1.4D**). In > 90% of cells, TOS_h and TOS_{min} indicate anticolocalization and their medians are ≈ -1 (*i.e.* maximally anticolocalized). In contrast, TOS_{max}, PCC, SRCC, M1diff and M2diff have many values between the 90th and 10th percentiles that are close to or greater than zero, which indicates these metrics classify many cells as having noncolocalization or colocalization rather than the expected anticolocalization (**Figure 1.4D**).

The second dataset examined were *Saccharomyces cerevisiae* cells (*n* = 38), which have a single copy of the spindle pole body component protein (Spc110) fused to both yellow fluorescent protein (YFP) and cyan fluorescent protein (CFP) (*i.e.* Spc110::YFP::CFP) [131]. We used eight sets of images (YRC PIR ID: 191, 3208, 3559, 3702, 3999, 4722, 5160, and 7396) with YFP, CFP and differential interference contrast (DIC) channels (**Figure 1.4E**). The analysis was performed as described for *D. melanogaster*. Because YFP and CFP are part of the same protein their signals should colocalize, and this was clearly seen at all threshold levels in the TOS matrix and scatterplot (**Figure 1.4F**, **G**). TOS_h, TOS_{max}, TOS_{min}, PCC and SRCC correctly identified that >90% of cells have strong colocalization (**Figure 1.4H**). M1diff and M2diff incorrectly identified most cells as noncolocalized (**Figure 1.4H**) due to Costes' method selecting very low thresholds; and this in turn results in the subtraction of a large "null" value and therefore M1diff and M2diff are small (note: this is one reason the rescaling for TOS is so helpful).

The third dataset examined were *Caenorhabditis elegans* (*n* = 39) which had the production of green fluorescent protein (GFP) regulated by the activity of the *clec-60* promoter (**Figure 1.4I**). In the *pmk-1* deficient mutant that was imaged, GFP production is increased in the anterior intestine next to the pharynx which was identified by a Myo-2::mCherry fluorescent protein fusion [4] (**Figure 1.4I**). The analysis was performed as for the other datasets, and a matrix of TOS median values showed anticolocalization of the GFP and mCherry fluorescence signals at low selected fractions (*i.e.* high threshold) (upper right corner, **Figure 1.4J**). Anticolocalization is both consistent with the biology (because GFP and Myo-2::mCherry label different structures) and a scatterplot of a representative *C. elegans* (**Figure 1.4K**). Values of TOS_h and TOS_{min} indicated anticolocalization in > 90% of worms (both green horizontal bars are below zero in **Figure 1.4L**). Most TOS_{max} values were > 0 therefore it was possible in most cells to identify a set of thresholds where there was a colocalization pattern, which was typically when the selected fraction was high in one channel and low in the other channel. In contrast to TOS_h and TOS_{min}, values for PCC, SRCC, M1diff and M2diff were generally around zero indicating noncolocalization for most worms (**Figure 1.4L**). The latter group of metrics performed poorly because it was difficult to identify a threshold that distinguishes the overlapping off-target and on-target signals and Costes' method tended to choose high selected fractions (*i.e.* low threshold values).

In summary, TOS analysis successfully identified the expected localization patterns of different proteins in various cells and organisms. In these images, TOS_h, TOS_{max}, or TOS_{min} were often able to identify specific features within mixed localization patterns better than PCC, SRCC, M1diff and M2diff.

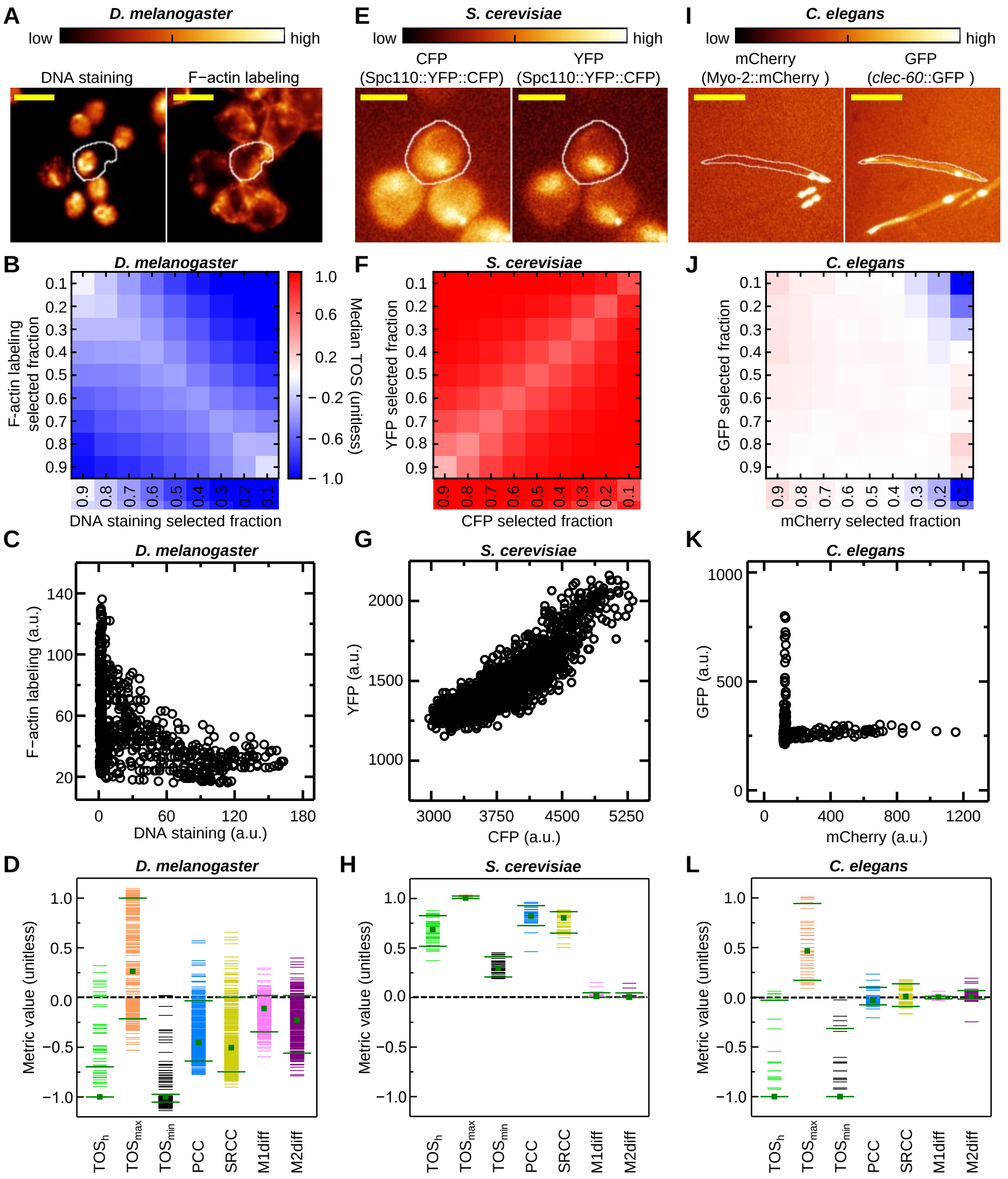

A
D. melanogaster
low
high
DNA staining
F−actin labeling
E
S. cerevisiae
low
high
CFP
(Spc110::YFP::CFP)
YFP
(Spc110::YFP::CFP)
I
C. elegans
low
high
mCherry
(Myo-2::mCherry)
GFP
(clec-60::GFP)
B
D. melanogaster
F−actin labeling selected fraction
0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9
0.9 0.8 0.7 0.6 0.5 0.4 0.3 0.2 0.1
DNA staining selected fraction
1.0 0.6 0.2 − 0.2 − 0.6 − 1.0
Median TOS (unitless)
F
S. cerevisiae
YFP selected fraction
CFP selected fraction
J
C. elegans
GFP selected fraction
mCherry selected fraction
C
D. melanogaster
F−actin labeling (a.u.)
140 100 60 20
0 60 120 180
DNA staining (a.u.)
G
S. cerevisiae
YFP (a.u.)
2000 1500 1000
3000 3750 4500 5250
CFP (a.u.)
K
C. elegans
GFP (a.u.)
1000 500 0
0 400 800 1200
mCherry (a.u.)
D
D. melanogaster
H
S. cerevisiae
L
C. elegans
Metric value (unitless)
1.0 0.5 0.0 −0.5 −1.0
TOS_h TOS_max TOS_min PCC SRCC M1diff M2diff

Figure 1.4 Application of TOS to different types of experimental data. (**A**) Microscopy images of representative *D. melanogaster* cells with DNA staining (Hoechst 33342) and F-actin labeling (fluorescein conjugated phalloidin). Images are pseudocolored and white lines indicate cell boundaries. Scale bar ~ 5 µm [1, 2]. (**B**) TOS matrix analysis in *D. melanogaster* cells (*n* = 366) with selected fractions for DNA staining and F-actin labeling intensity. (**C**) Scatterplot of DNA staining and F-actin labeling in the outlined *D. melanogaster* cell in **panel A**. Note: all intensity values are > 0. (**D**) TOS_h, TOS_{max}, TOS_{min}, PCC, SRCC, M1diff, and M2diff values obtained in individual *D. melanogaster*. Green lines and square indicate the 90th and 10th percentiles and the medians. Horizontal dash line at zero indicates noncolocalization or no correlation. (**E**) Microscopy images of representative *S. cerevisiae* cells with SPC110::YFP::CFP. Images presented as in **panel A**. are pseudocolored. Scale bar = 5 µm (obtained from YRC PIR image). (**F**) TOS matrix analysis in *S. cerevisiae* cells (*n* = 38) with selected fractions for CFP and YFP fluorescence intensity. Heat map scale shown in **panel B**. (**G**) Scatterplot of CFP and YFP fluorescence in outlined *S. cerevisiae* cell in **panel E**. (**H**) TOS_h, TOS_{max}, TOS_{min}, PCC, SRCC, M1diff, and M2diff values obtained in individual *S. cerevisiae*. Data presented as in **panel D**. (**I**) Microscopy images of representative *C. elegans* with mCherry and GFP fluorescence. Scale bar ≈ 500 µm [4]. Images presented as in **panel A** except the white line is a *C. elegans* outline. (**J**) TOS matrix analysis in *C. elegans* (*n* = 42) with selected fractions for mCherry and GFP fluorescence intensity. Heat map scale shown in **panel B**. (**K**) Scatterplot of Myo-2::mCherry and GFP fluorescence signal in the *C. elegans* outlined in **panel I**. (**L**) TOS_h, TOS_{max}, TOS_{min}, PCC, SRCC, M1diff, and M2diff values obtained in individual *C. elegans*. Data presented as in **panel D**.

1.3.6 TOS values can distinguish localization patterns in experimental data with high specificity and high sensitivity

We investigated how well TOS and other metrics can distinguish similar localization patterns. Two types of *Schizosaccharomyces pombe* strains were chosen with fluorescent proteins that were expected to show colocalization. One strain (*n* = 40) had the fusion protein Sid4::YFP::CFP (strain KG4608; ID: 192, 776, 1062 and 1233) (**Figure 1.5A**). Because these fluorescent proteins are fused they should colocalize. A second strain (*n* = 38) had two fusion proteins: Cdc11::CFP and Cdc13::YFP (strain KG3544; ID: 292, 360, 414, and 744) (**Figure 1.5B**). Cdc11::CFP and Cdc13::YFP are known to colocalize to the spindle pole body [132, 133] as well as to other sites.

Cells from each strain were identified and analyzed as described above. TOS matrices for both strains showed colocalization (TOS >> 0) at many threshold combinations (**Figure 1.5C**, **D**). However, at high signal intensities (*i.e.* small F_{T1} and F_{T2}) there are differences in localization between the two strains; cells with Sid4::YFP::CFP have colocalization and cells with Cdc11::CFP and Cdc13::YFP have anticolocalization (**Figure 1.5C**, **D**). The difference in localization can also be seen in the scatterplot of CFP and YFP signal intensity in a representative cell from each strain (**Figure 1.5E, F**). That is, pixels with high intensity CFP and YFP signals tend to occupy the upper right corner for Sid4::YFP::CFP but tend to be at the right or at the top for Cdc11::CFP and Cdc13::YFP. Note: there are many possible causes for why there is more YFP fluorescence for a given amount of CFP fluorescence at high intensity levels compared to lower intensity levels for the YFP::CFP fusion (**Figure 1.5E**) including: increased transcription and translation termination, and decreased CFP fluorescence and/or increased YFP fluorescence due to altered protein folding and aggregate formation.

TOS_h, TOS_{max}, TOS_{min}, PCC, SRCC, M1diff and M2diff were calculated as for the analysis above (**Figure 1.3D**). All metrics except TOS_{max} had statistically significant differences in the distribution of values for the two strains (displayed in **Figure 1.5G**). To compare metrics we generated histograms of each value from individual cells (**Figure 1.5G**) and then ROC curves (**Figure 1.5H**) as described above. Cells with Sid4::YFP::CFP were designated the condition positive population and cells with Cdc11::CFP and Cdc13::YFP were the condition negative population. ROC curves for TOS_h and TOS_{min} demonstrated that they can discriminate the localization patterns of the two cell types with greater specificity and sensitivity than SRCC, PCC, M1diff and M2diff. M1diff and M2diff were not able to distinguish the localization patterns in the two strains because of difficulty with thresholding.

In summary, values from TOS matrices such as TOS_h and TOS_{min} were able to distinguish the localization patterns of the different proteins with greater specificity and greater sensitivity than other common metrics, and they are particularly useful when there are mixed patterns of localization within each cell and thus measures of the localization pattern of entire cells are less meaningful. We stress that this assessment did *not* evaluate whether TOS_h, TOS_{max}, TOS_{min} are better descriptors of an entire population of pixels or whether they identify a specific feature that is reflective of the underlying biology.

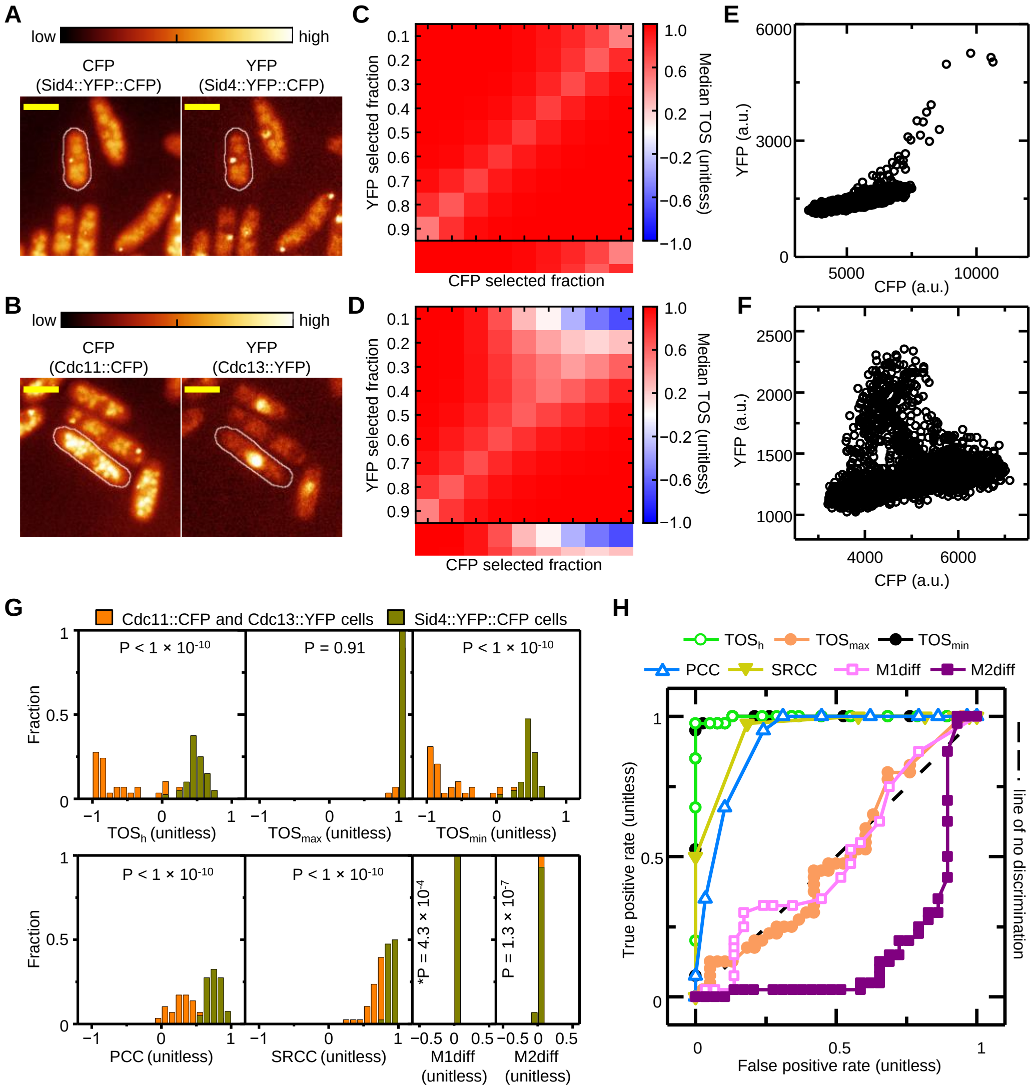


Figure 1.5 TOS values can distinguish localization patterns in experimental data with high specificity and high sensitivity. (**A**) Microscopy images of representative *S. pombe* cells with Sid4::YFP::CFP. Images are pseudocolored. White lines indicate cell boundaries. Scale Bar = 10 µm (obtained from YRC PIR images). (**B**) Microscopy images of representative *S. pombe* cells with Cdc11::CFP and Cdc13::YFP. Images are presented as in **panel A**. (**C, D**) TOS matrices for the strains in **panels A** and **B** respectively (*n* = 40 and 38). Each matrix shows the median values obtained for TOS matrices of individual cells. (**E, F**) Scatterplots of CFP and YFP signal intensity for the outlined cell of each strain in **panel A** and **B** respectively. (**G**) Histograms of TOS$_h$, TOS$_{max}$, TOS$_{min}$, PCC, SRCC, M1diff and M2diff values for individual cells in each strain. * P value is calculated with the raw data, which has more variation than seen in the binned data of the histogram. (**H**) ROC curves for all metrics in **panel G**.

1.4 Discussion

Measuring localization is a basic requirement of cell biology and imaging and yet it is often a challenging task. This study demonstrates the TOS metric and its application, and shows they are valuable tools to help meet the challenge of quantifying localization in a wide range of applications.

TOS has many features that make it suitable for general applications. The first feature is that TOS is simple to interpret because it only quantifies whether signal overlap is the same, more, or less than expected by chance. In contrast, some overlap metrics have a weighting for signal intensities (*e.g.* Mander's coefficients M1 and M2, "overlap coefficient", and "k_1 and k_2 coefficients" [115]), which means that any value is an unknown combination of two factors: overlap and intensities. A second feature is that TOS can be compared at different thresholds, which is not the case with some other metrics (*e.g.* M1 and M2 coefficients) [118]. A third feature is that a single value distinguishes between colocalization, anticolocalization and noncolocalization, whereas some metrics require two values for interpretation, and/or they do not directly distinguish between anticolocalization and noncolocalization [115]. A fourth feature is the null hypothesis for TOS has minimal assumptions and requires no simulations [112, 118], which makes it easier to be implemented. A fifth feature is that TOS is one of the metrics that does not assume a linear correlation in signal intensities [112].

The general applicability of TOS is enhanced by systematically evaluating it at many different threshold combinations. The resulting TOS matrix is particularly useful when there are mixed patterns of localization; and the background and off-target signals are continuous with the on-target signal. TOS matrices are best interpreted holistically with TOS at each selected fraction being evaluated in the context of neighboring TOS (which can detect trends and provide confidence for a specific value for TOS) and in relation to other localization patterns found in the matrix. Within TOS matrices, TOS at the highest thresholds (*i.e.* TOS_h) was particularly helpful in identifying localization patterns in on-target signals when the off-target and background signals were at high levels and/or occupying a large proportion of pixels. We showed that TOS_h, as well as TOS_{max} and TOS_{min}, often had greater specificity and sensitivity than PCC, SRCC, M1diff, and M2diff. Furthermore, TOS was very easy to use with a wide variety of proteins, cell types, and organisms. For all the above reasons, TOS matrices are a good first line of analysis for quantifying intracellular localization. However, we reiterate that there is no best test for all situations [112, 113] and that the selection of a metric must take into account the purpose of the analysis, the underlying biology, and the types of images and samples.

To interpret values of TOS it is important to highlight that in many imaging experiments, including those used in this study, the concentration of reporter is high and single particles cannot be resolved. Therefore the signal in each pixel (or voxel) is the total of many reporter molecules within an area (or volume) of the cell. That is, the signal intensity in each pixel reflects the local concentration of a molecule. Local concentrations may be higher or lower in some cell regions depending on: (i) sites of production and degradation; (ii) diffusion; (iii) kinetics of association and dissociation with cellular structures (*e.g.* nucleus, cell membrane or cytoskeleton); and (iv) attraction to or exclusion from cell regions [134, 135]. With this in mind, colocalization, anticolocalization and noncolocalization should be considered as the relationship in the local concentrations of two types of molecules, which may be due to many factors (and therefore should not be interpreted by itself as revealing as something about the molecules binding [112, 113]). Note: the above mechanisms could potentially generate concentration gradients that contribute more to the spatial autocorrelation of signals in cells than point spread functions [113, 136].

Following from the above, colocalization indicates that higher concentrations of two molecules tend to occur in similar cell regions. This may be due to common sites of production, action, binding, or degradation. Anticolocalization indicates two molecules have high concentrations in different cell regions and thus at least one mechanism is causing the molecules to be recruited to and/or exclude from different regions, one molecule excludes the other from a region [134, 137] or the molecules eliminate each other in the same location (*e.g.* when non-coding RNAs binding to mRNAs both are destroyed [138]). Noncolocalization indicates that molecules have no preference for avoiding or occurring in the same regions. Because the mechanisms responsible for generating anticolocalization and noncolocalization are different, the capacity of metric such as TOS to distinguish these patterns is potentially very useful.

The TOS metric could be adapted for applications that were not examined in this study and to measure localization in different ways. We chose to measure overlap by selecting pixels above thresholds because that approach was most similar to that of Manders' colocalization coefficients. However, AO, AO ratio and TOS could be modified to measure the overlap of pixels below a threshold or within a range (*i.e.* the equivalent of a band-pass filter or low-pass filter instead of a high-pass filter). Another way in which the TOS metric could be altered is to choose selected fractions of pixels by features other than signal intensity such as their distances to the cell poles or membrane. Additionally, TOS analysis could be adapted to examine localization in three dimensional images (*e.g.* images assembled from confocal microscopy) or measure the convergence of more than two signals.

In conclusion, systematic evaluation of the TOS metric at multiple threshold combinations is a valuable addition to the repertoire of tools available for the quantitative analysis of images. TOS analysis is simple to implement and easy to interpret, and it has many features that make suitable for many types of images and samples. Furthermore, values from TOS matrices can distinguish patterns of localization with greater sensitivity and greater specificity than other commonly used metrics. These findings make a strong case for selecting TOS analysis as a first step to evaluating localization in images.

1.5 Materials and Methods

1.5.1 Simulations, calculation of metrics, and statistical analyses

Simulations, calculations and statistical analyses were performed as described in the **Results** using Matlab (R2015a, Mathworks) (code archived at Figshare: https://figshare.com/s/6504f19aef88f1d6cf95). Post-measurement statistical comparisons were performed using the two-tailed Mann-Whitney U test. Note: localization in one set of samples could also be compared to the median of the expected null distribution using the Sign test or by bootstrapping.

1.5.2 Receiver operating characteristic (ROC) curves

Histograms for simulated data were generated using the histcounts function in Matlab. This function, which is based on Scott's rule [139], determined the bin edges in the range of −1.2 to 1.2. For experimental data, bin edges had increments of 0.1 for TOS_h and TOS_{min} (defined in **Results** section), Pearson's correlation coefficient, and Spearman's rank correlation coefficient, and increments of 0.001 for the other metrics. The bin edges were used as thresholds and the fraction of counts in each population above the thresholds were used to create the ROC curves.

1.5.3 Analysis of images

Images of *Drosophila melanogaster* Kc167 cells (BBBC007_v1 (A9)) and whole organism *Caenorhabditis elegans* (BBBC012v1) were obtained from the Broad Bioimage Benchmark Collection [140]. Hand drawn boundaries of *D. melanogaster* cells were downloaded from the same collection and inverted to select cells [141]. *Saccharomyces cerevisiae* (DHY155) and *Schizosaccharomyces pombe* (KG4608 and KG3544) images were obtained from the Yeast Resource Center Public Image Repository (YRC PIR) [142]. Boundaries were traced around *C. elegans* in four different brightfield images, around *S. cerevisiae* cells in eight differential interference contrast (DIC) images, and around *S. pombe* in four DIC images for each strain. Traces were performed in ImageJ [143] and these defined the boundary of a "region of interest (ROI)" (data files at Figshare: https://figshare.com/s/e414b6b45d53f79f7b1f). A "Count Mask" was created in ImageJ to fill each ROI in an image with a unique integer. Count Mask was used to select pixels in the fluorescence images with Matlab that correspond to cells or *C. elegans*. Pixel intensity values within each cell or *C. elegans* were stored in an array, which were used for the analyses.

Some downloaded drawn objects for *D. melanogaster* cells did not identify cell boundaries therefore the Analyze Particle function in ImageJ was used to eliminate small (<400 pixels) and large objects (>5000 pixels) from the analysis. In addition, boundaries that identified areas between cells were eliminated by selecting only ROIs with fluorescence signals greater than the background in non-cell regions. Occasional *S. cerevisiae* cells had binned data so they were removed from the analyses.

Chapter 2: EzColocalization: An ImageJ plugin for visualizing and measuring colocalization in cells and organisms[2]

2.1 Abstract

Insight into the function and regulation of biological molecules can often be obtained by determining which cell structures and other molecules they localize with (i.e. colocalization). Here we describe an open source plugin for ImageJ called EzColocalization to visualize and measure colocalization in microscopy images. EzColocalization is designed to be easy to use and customize for researchers with minimal experience in quantitative microscopy and computer programming. Features of EzColocalization include: (i) tools to select individual cells and organisms from images; (ii) filters to select specific types of cells and organisms based on physical parameters and signal intensity; (iii) heat maps and scatterplots to visualize the localization patterns of reporters; (iv) multiple metrics to measure colocalization for two or three reporters; (v) metric matrices to systematically measure colocalization at multiple combinations of signal intensity thresholds; and (vi) data tables that provide detailed information on each cell in a sample. These features make EzColocalization well-suited for experiments with low reporter signal, complex patterns of localization, and heterogeneous cells and organisms.

[2] This chapter has been previously published as "EzColocalization: An ImageJ plugin for visualizing and measuring colocalization in cells and organisms" in *Scientific reports* (2018), *8*(1), 15764.

2.2 Introduction

Advances in microscopy equipment and labeling techniques make it possible for researchers to image a variety of biological molecules in almost any cell, tissue, or organism [144-150]. However, researchers often find it difficult to rigorously evaluate and interpret the images. In particular, it is often challenging to determine whether the different molecules of interest occur in the same locations, different locations or independent locations (*i.e.* colocalization, anticolocalization and noncolocalization respectively) in cells, tissues or organisms [151].

Several factors limit the use of current software for visualizing the localization of reporters in biological samples and measuring colocalization [112, 113, 152, 153]. One factor is that customization of the software is often required for the equipment, reporters and samples [154, 155], and for automated analyses. A second factor is that the software is often not suited to experiments that push the limits of detection, where the intensity of the intracellular signal is similar to the extracellular signal (*i.e.* “background”) [3], and where there are high levels of non-specific signal in cells [151]. The latter can occur because the probes or reporters are not sufficiently specific [156], are not adequately removed from cells or organisms [157], or have low signal relative to endogenous compounds (*i.e.* “autofluorescence”) [158]. That is, software tools are needed to distinguish intracellular pixels from extracellular pixels, and to select signal intensity thresholds to limit analyses to a subset of intracellular pixels. A third factor is that there are often mixed localization patterns within cells and different localization patterns among cells in a sample [113, 151, 159]. When this heterogeneity is present, software is need to provide measurements for each cell or defined subsets of cells in samples.

It is often possible to address the above challenges by combining multiple existing software programs and customizing them [3, 151]. However, combining and customizing software requires proficiency in programming, experience with quantitative microscopy, comfort with mathematics and statistics, and other support. Many researchers do not have these skills and resources, and this a likely reason that many studies evaluate colocalization by the simple, but often misleading, method of overlaying red and green color images [112, 113]. Therefore there is a pressing need for a single application that provides all the tools for start to finish analysis of colocalization and can be easily customized.

In this study, an open source plugin for ImageJ called EzColocalization was developed so that researchers at all levels of proficiency can visualize the localization of signals and measure colocalization via an easy-to-use graphical user interface (GUI). The first part of the study describes EzColocalization, and the second part demonstrate its use for different sample types and for resolving common issues that prevent rapid and robust quantitative measurements of colocalization. EzColocalization can measure colocalization in cells, tissues, and whole organisms (*e.g. Caenorhabditis elegans* and *Drosophila* embryos); and the software is especially helpful where automation and customization is required, to obtain individual cell measurements in samples with many cells, and for reporters with low signal or low specificity.

2.3 Results

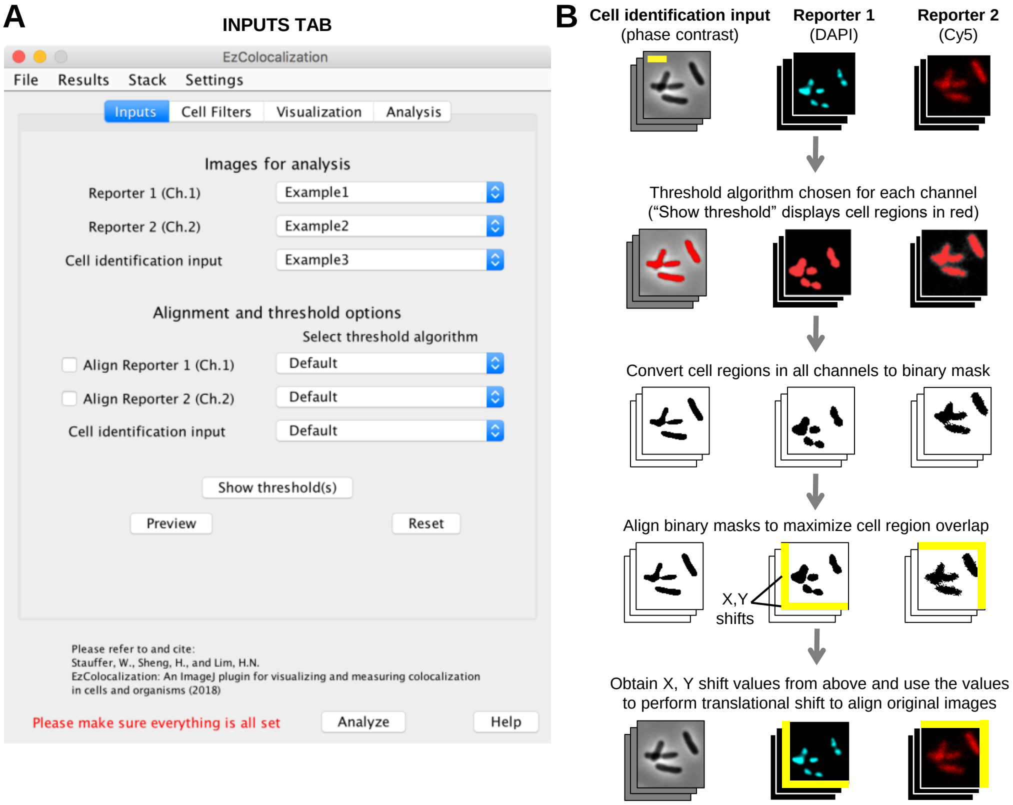
2.3.1 Overview of EzColocalization workflow

The workflow for EzColocalization is divided into four modules each with its own tab on the GUI. The tabs are: (i) “Inputs” where images, masks or regions of interest (ROI) lists are selected and aligned; (ii) “Cell Filters” where cells can be selected based on physical features and signal intensity; (iii) “Visualization” where heat maps, scatterplots, and metric matrices (defined below) are created; and (iv) “Analysis” where the colocalization metrics and outputs are chosen. Not all modules and not all processes within a module have to be used. Some tabs have a “Preview” button to run a specific module instead of the “Analyze” button which runs all selected processes in all modules.

2.3.1.1 Inputs

Image files, which are chosen in the “Input” tab (**Figure 2.1A**), must be: (i) monochromatic (*i.e.* not RGB or CMYK formats); (ii) 8-bit, 16-bit, or 32-bit; and (iii) in a format such as TIFF that retains the original pixel intensity values. Large images may be compressed for file transfer using a lossless format such as ZIP or LZW, and then decompressed for analyses. In addition to images, EzColocalization can accept masks and ROI lists for cell identification (see below). If there are multiple images for each channel, the images should stacked for more efficient analysis in the “Stack” menu (see ImageJ guide for further details [160]). Images in a stack may be different fields of view or a time series, but must have the same magnification and image order for each channel. The input tab also provides options for setting thresholds for signal intensity and aligning misaligned images from different channels (**Figure 2.1B** and **Materials and Methods**). Recommendations for acquiring suitable images for colocalization analysis are provided in the **Materials and Methods**. Note: alignment operates on the assumption that an appropriate threshold for signal intensity can be chosen to distinguish pixels inside and outside of cells; if thresholding includes areas outside the cell or only a limited area within cells, then the alignment may not function properly. For this reason, all alignments should be checked by outputting ROIs and confirming appropriate cell areas are selected.

EzColocalization is primarily designed for one “cell identification” channel and two or three “reporter” channel images. However, it can operate with other input combinations (**Table 2.1**). The cell identification channel is used to identify individual cells, and consequently to distinguish intracellular and extracellular pixels. The cell identification channel can be any image that identifies the cell boundary including: light microscopy images (*e.g.* phase contrast [161, 162] and bright-field), images with a reporter that labels the cell membrane or is throughout the cytoplasm (*e.g.* Cy5, **Figure 2.1B**), and images with an extracellular dye that outlines cells. Differential interference contrast (DIC) images create shadows that make it difficult for automated selection of cells using threshold methods [163]; therefore for DIC images we recommend that ROIs be created using the “selection tools” in ImageJ to manually outline cell areas, and then adding them to a list by choosing “Add to Manager” (in “Selection” submenu of the “Edit” menu). A binary mask can be created using the “Clear Outside” and then “Autothreshold” functions of ImageJ, once there are ROIs for all cells of interest in an image.


A
INPUTS TAB
EzColocalization
File
Results
Stack
Settings
Inputs
Cell Filters
Visualization
Analysis
Show threshold(s)
Preview
Reset
Analyze
Help
B
Cell identification input
(phase contrast)
Reporter 1
(DAPI)
Reporter 2
(Cy5)
Threshold algorithm chosen for each channel
(“Show threshold” displays cell regions in red)
Convert cell regions in all channels to binary mask
Align binary masks to maximize cell region overlap
X,Y
shifts
Obtain X, Y shift values from above and use the values
to perform translational shift to align original images

Images for analysis

Reporter 1 (Ch.1) Example1

Reporter 2 (Ch.2) Example2

Cell identification input Example3

Alignment and threshold options

Select threshold algorithm

☐ Align Reporter 1 (Ch.1) Default

☐ Align Reporter 2 (Ch.2) Default

Cell identification input Default

Please refer to and cite:
Stauffer, W., Sheng, H., and Lim, H.N.
EzColocalization: An ImageJ plugin for visualizing and measuring colocalization
in cells and organisms (2018)

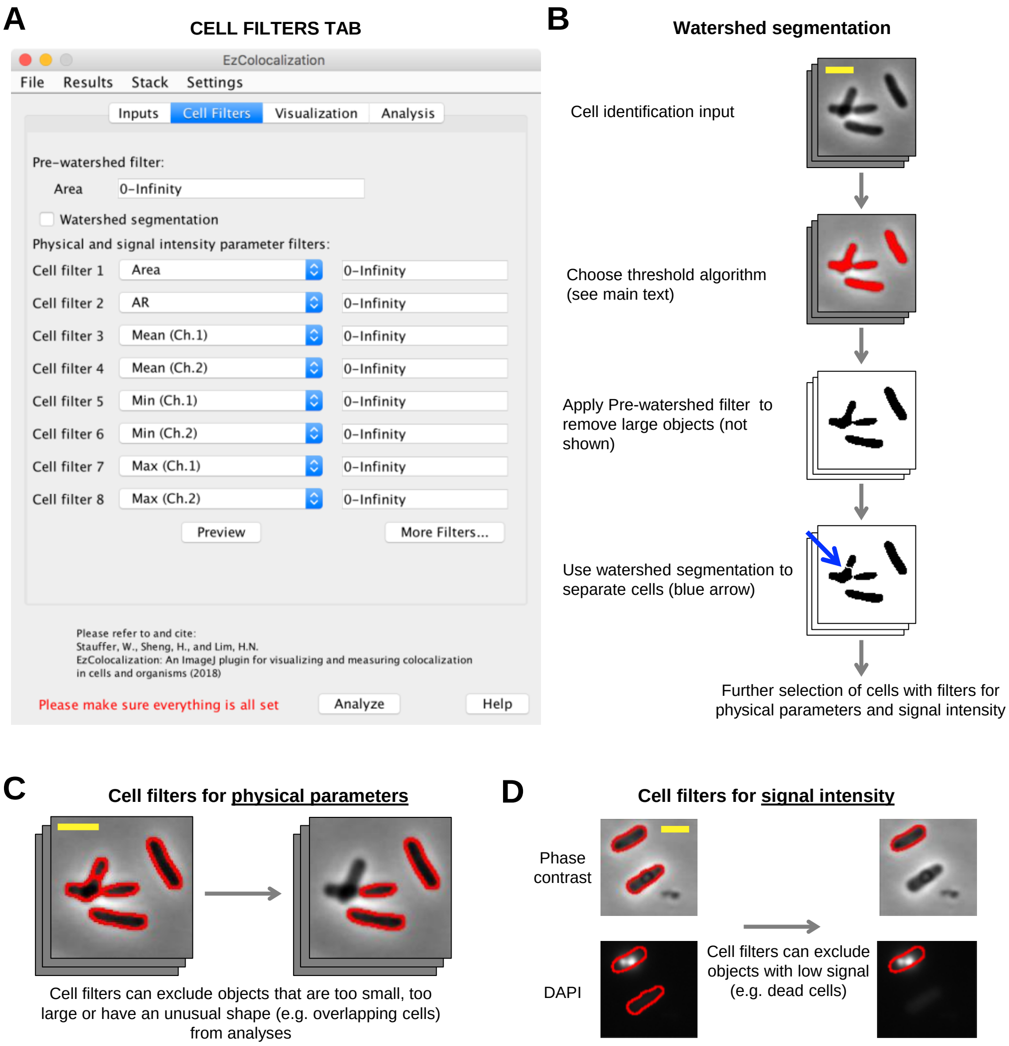
Please make sure everything is all set

Figure 2.1 Input and alignment tab. (**A**) Input tab in the GUI. (**B**) General steps for the alignment of images. The cell identification image stack (phase contrast; left column), reporter 1 image stack (DAPI staining of DNA; center column), and reporter 2 image stack (Cy5; right column) are images of a previously reported bacterial strain (HL6320) [3]. Scale bar is 2 µm. Reporter 1 and 2 images are pseudocolored. Red coloring in the second row of images indicates the objects identified by thresholding of the signal in each channel (“Default” algorithm in ImageJ). Following alignment of the images, some pixels will overhang or need to be filled with pixels (yellow areas) so that all images have the same area in the common aligned region.

2.3.1.2 Cell Filters

The “Cell Filters” tab is used to help select cells in images (**Figure 2.2A**) and distinguish intracellular and extracellular pixels. Cells are identified by: (i) choosing one of the ImageJ threshold algorithms [160], or manually selecting the thresholds (which is done by selecting “*Manual*” from a drop-down list in the Inputs tab and pressing the “Show Threshold(s)” button), to identify regions corresponding to cells in the cell identification channel (**Figure 2.2B**); (ii) using watershed segmentation to separate touching objects in the cell identification channel images (optional) (**Figure 2.2B**); (iii) selecting objects from the cell identification channel images based on physical parameters (**Figure 2.2C**) and signal intensity (**Figure 2.2D**). EzColocalization will attempt to automatically detect whether input images have dark or light background using skewness. Assuming there are more pixels in the background than in the cells, an image with positive skewness indicates a dark background and negative skewness indicates a light background. Users can also manually select whether the input images have dark or light background in the “Parameters…” options of the “Settings” menu. Cells that are only partly within an image, and therefore could provide misleading values, are automatically removed from analyses.

EzColocalization has one optional “Pre-watershed filter” and 8 optional post-watershed filters (with the option to select more). Watershed segmentation can aid the separation of dividing and touching cells [164] but it can also divide large objects such as aggregates of extracellular material into smaller fragments that are the same size as cells. To avoid the latter, the Pre-watershed filter can be used to exclude objects with large areas from the analysis. The Preview button in the Cell Filters tab allows users to see which objects on the current image will be filtered out when the minimum and maximum bounds of the Pre-watershed filter are adjusted. There are two classes of parameters for the post-watershed cell filters (**Table 2.2**): (i) physical parameters based on measurements from the cell identification channel; and (ii) signal intensity parameters from the reporter channels. Physical parameters apply to all channels whereas signal intensity parameters apply only to the reporter channel for which they are selected (because reporters may have very different level of signal). In addition to filtering based on predefined options in ImageJ, EzColocalization has filters for the “MeanBgndRatio” or “MedianBgndRatio”, which are calculated by dividing the mean or median signal intensity of pixels inside an object by the respective mean or median signal intensity of extracellular pixels.



Pre-watershed filter:

Area 0-Infinity

□ Watershed segmentation

Physical and signal intensity parameter filters:

Cell filter 1	Area	0-Infinity
Cell filter 2	AR	0-Infinity
Cell filter 3	Mean (Ch.1)	0-Infinity
Cell filter 4	Mean (Ch.2)	0-Infinity
Cell filter 5	Min (Ch.1)	0-Infinity
Cell filter 6	Min (Ch.2)	0-Infinity
Cell filter 7	Max (Ch.1)	0-Infinity
Cell filter 8	Max (Ch.2)	0-Infinity

Please refer to and cite:
Stauffer, W., Sheng, H., and Lim, H.N.
EzColocalization: An ImageJ plugin for visualizing and measuring colocalization
in cells and organisms (2018)

Please make sure everything is all set

Preview
More Filters...
Analyze
Help
B
Watershed segmentation
Cell identification input
Choose threshold algorithm (see main text)
Apply Pre-watershed filter to remove large objects (not shown)
Use watershed segmentation to separate cells (blue arrow)
Further selection of cells with filters for physical parameters and signal intensity
C
Cell filters for physical parameters
Cell filters can exclude objects that are too small, too large or have an unusual shape (e.g. overlapping cells) from analyses
D
Cell filters for signal intensity
Phase contrast
DAPI
Cell filters can exclude objects with low signal (e.g. dead cells)

Figure 2.2 Cell identification and cell filters tab. (**A**) Cell filter tab in the GUI. (**B**) Cell selection and watershed segmentation. Red coloring in the image in the second row indicates objects identified by thresholding of the signal in the cell identification channel ("Default" algorithm in ImageJ). Cells are the same as in **Figure 2.1**. (**C**) Selection of cells based on physical features using the cell filters. Scale bar is 2 µm. Phase contrast image from **Figure 2.1**. Red outline indicates the objects that were identified by thresholding (**Panel B**), and in the case of the right image, are within the parameter range(s) selected by the filter. (**D**) Selection of cells based on signal intensity using the cell filters. Phase contrast (cell identification image) and DAPI stain (reporter channel) images of bacteria (HL6187). Scale bar is 2 µm. Note: the lower of the two cells in the image (no red border) has been removed from the analysis by the cell filter (that is, it no longer has the red cell outline).

2.3.1.3 Visualization tab

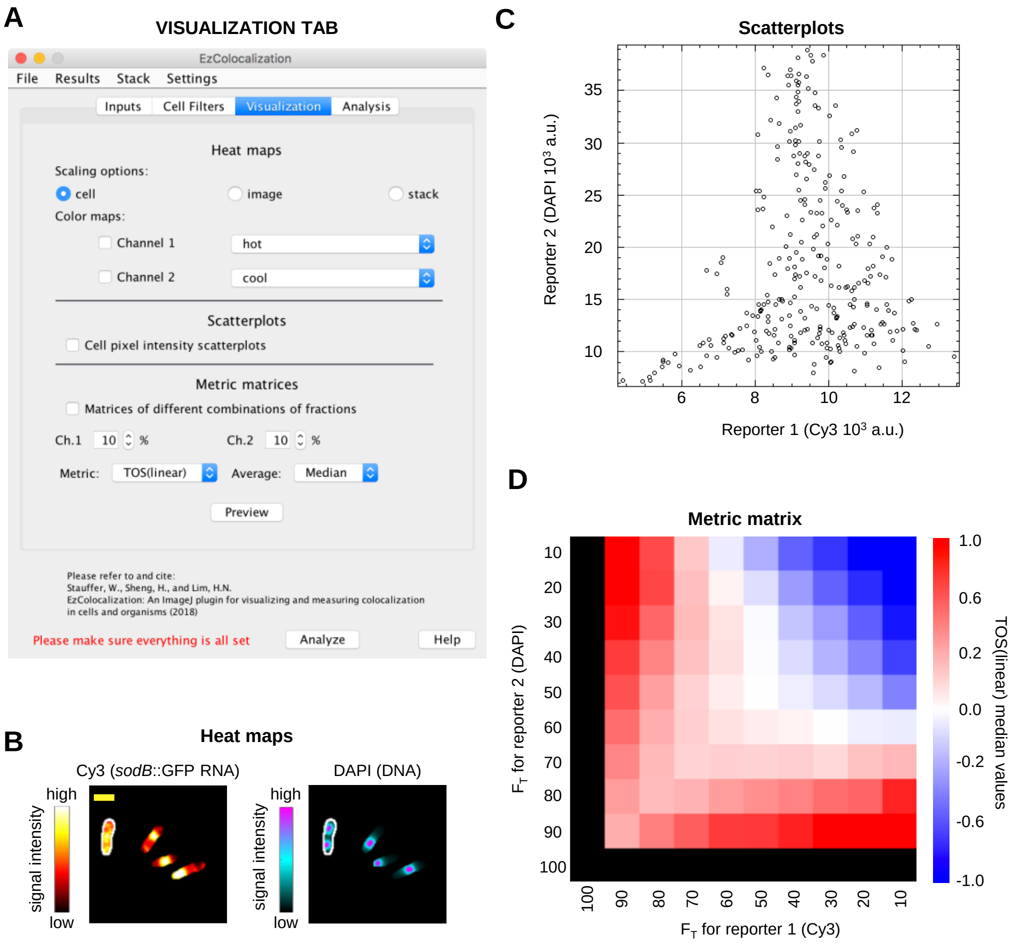
The "Visualization" tab displays signals in cells as: (i) "heat maps"; (ii) scatterplots; and (iii) "metric matrices" (**Figure 2.3A**).

Heat maps are 8-bit color images that show the relative magnitude of reporter signals (**Figure 2.3B**). They are generated by normalization and rescaling so that the minimum and maximum pixel values are 0 and 255 respectively in each cell, image, or stack. There are eight options for coloring the heat maps, and the intensity values for each color are obtained from the "Show LUT" function (within the “Color” submenu of the “Image” menu in ImageJ). Cell heat maps are suited for determining where each reporter occurs with highest intensity in cells. Image heat maps can show if different cells within a field of view have substantially different intensities, which may indicate biological heterogeneity or unevenness in labeling. Stack heat maps can show if cells in different images have substantially different levels of signal intensity, which may indicate unevenness in labeling or measurements across a slide (*e.g.* due to photobleaching) or at different time points (if the stack is a time series). Note: heat map appearance is affected by brightness and contrast settings.

Scatterplots show the relationship between the signal intensity for two or three reporter channels for individual cells, images, and stacks (**Figure 2.3C**). This relationship is important in choosing the appropriate metric (**Materials and Methods**). Scatterplots can also reveal heterogeneity in the localization patterns [151], which may require removal of background pixels or separate analyses for different cell types.

Metric matrices provide an overview of localization patterns by showing the calculated values of a colocalization metric for many threshold combinations. Metric matrices for the threshold overlap score (TOS) have been shown to be useful for the analysis of localization patterns for two reporter channels [3, 151] (**Figure 2.3D**). For completeness, EzColocalization has the option to calculate metric matrices for two reporter channels using five other metrics: threshold overlap score with logarithmic scaling [151], Pearson correlation coefficient (PCC), Manders' colocalization coefficients (M1 and M2), Spearman's rank correlation coefficient (SRCC), and intensity correlation quotient (ICQ) [3, 151]. Colocalization for three channels can also be measured using ICQ, Manders' colocalization coefficients and TOS [127] (**Materials and Methods**).

Thresholds for all metrics are measured as the top percentile (F_T) of pixels for signal intensity [3, 151]. For example, $F_T = 0.1$ is the 10% of pixels with the highest signal. For the metric matrices, F_T is also used to specify the step size for the threshold combinations. That is, $F_T = 0.1$ would be used to set thresholds for the 10%, 20%, …, and 100% of pixels with the highest signal. If F_T does not divide evenly into 100, then the remaining percent is the last step size. For metrics that do not need a threshold (*i.e.* PCC, SRCC, and ICQ) the values are calculated assuming that only the pixels above the thresholds exist. The metric matrix window has options for the values to be saved as text or image, for changing the F_T or type of metric, viewing individual cell metric values as a list, and calculating the mean, median or mode of the metric for each threshold combination. The “Proc” (processed) and “Raw” button determine whether the list of data displayed, copied, or saved with the “List”, “Copy”, or “Save…” buttons respectively is the average value for the sample for each threshold combination (*e.g.* median value) or all values for each cell in the sample for all threshold combinations.



EzColocalization

File Results Stack Settings

Inputs Cell Filters Visualization Analysis

Heat maps

Scaling options:

● cell ○ image ○ stack

Color maps:

□ Channel 1 hot

□ Channel 2 cool

Scatterplots

□ Cell pixel intensity scatterplots

Metric matrices

□ Matrices of different combinations of fractions

Ch.1 10 % Ch.2 10 %

Metric: TOS(linear) Average: Median

Preview

Please refer to and cite:
Stauffer, W., Sheng, H., and Lim, H.N.
EzColocalization: An ImageJ plugin for visualizing and measuring colocalization in cells and organisms (2018)

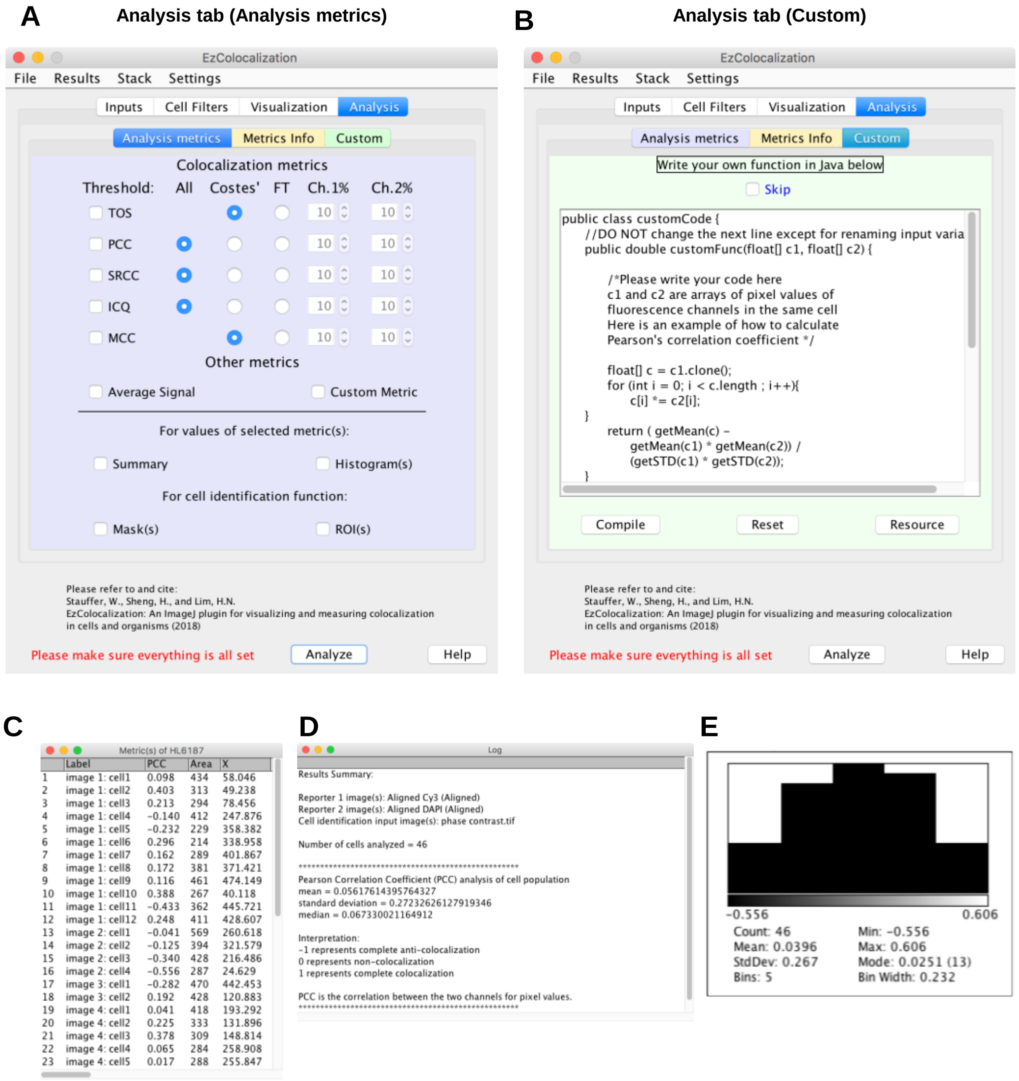
Please make sure everything is all set Analyze Help

Figure 2.3 Visualization tab. Bacterial cells (HL6187) with labeled sodB::gfp RNA (Cy3 channel) and DNA (DAPI). (**A**) Visualization tab in the GUI. (**B**) Heat maps of Cy3 and DAPI signals for bacteria with "cellular scaling" (defined in main text). Scale bar is 2 µm. (**C**) Scatterplot of Cy3 and DAPI for the cell on the left and outlined in white in **Figure 2.3B**. (**D**) Metric matrix for TOS (linear scaling) for the cell on the left and outlined in white in **Figure 2.3B**. F_T is the top percentage of pixels in the channel; for example, if F_T for Cy3 is 80% then it refers to the 80% of pixels with the greatest Cy3 signal. Black color on the left column and bottom row indicate that TOS values are not informative when one threshold is 100%; that is, the overlap of two reporters can only be 100% if 100% of pixels are selected for at least one channel.

2.3.1.4 Analysis

The "Analysis" tab has two subtabs ("Analysis metrics" and "Custom"). The Analysis Metrics subtab has six metrics for measuring colocalization for two reporters (**Figure 2.4A**) and three metrics for three reporters (see previous section). Users may choose a threshold or no threshold for PCC, SRCC and ICQ. TOS and Manders' colocalization coefficients must have a threshold to be calculated. Thresholds can be selected using Costes' method [124] or manually. In the Custom subtab (see **Materials and Methods** for additional information), users can write their own code in Java^{TM} to analyze images (note: the example provided is for calculating PCC) (**Figure 2.4B**). The "Compile" button tests the code and creates a temporary file in the Java temporary directory and displays the outcome of the compiling with a "Succeeded" or "Failed" label. If successful, the compiled custom code is read to the memory again and applied to the selected cells.

The output of every analysis is a table that specifies the image and an identifier number for every cell (**Figure 2.4C**), and for each cell, values are provided for: (i) the selected metric; (ii) physical parameters; and (iii) average signal intensity for each channel (if selected). Note: "NaN" in the output table indicates the failure to calculate a value. Users can also generate summary windows (with the cell number, mean, median and standard deviation for the selected metric) (**Figure 2.4D**), histograms of metric values (**Figure 2.4E**), binary mask images, and ROI lists that specify each cell's position and number on each image in the ROI manager. ROI lists and binary mask images can be saved for re-analysis of the same cells.



Colocalization metrics

Threshold:	All	Costes'	FT	Ch.1%	Ch.2%
□ TOS		●	○	10	10
□ PCC	●	○	○	10	10
□ SRCC	●	○	○	10	10
□ ICQ	●	○	○	10	10
□ MCC		●	○	10	10

Other metrics

□ Average Signal □ Custom Metric

For values of selected metric(s):

□ Summary □ Histogram(s)

For cell identification function:

□ Mask(s) □ ROI(s)

Please refer to and cite:
Stauffer, W., Sheng, H., and Lim, H.N.
EzColocalization: An ImageJ plugin for visualizing and measuring colocalization in cells and organisms (2018)

Please make sure everything is all set

Write your own function in Java below

□ Skip

```
public class customCode {
    //DO NOT change the next line except for renaming input varia
    public double customFunc(float[] c1, float[] c2) {

        /*Please write your code here
        c1 and c2 are arrays of pixel values of
        fluorescence channels in the same cell
        Here is an example of how to calculate
        Pearson's correlation coefficient */

        float[] c = c1.clone();
        for (int i = 0; i < c.length ; i++){
            c[i] *= c2[i];
    }
        return ( getMean(c) -
            getMean(c1) * getMean(c2)) /
            (getSTD(c1) * getSTD(c2));
    }
```

Please refer to and cite:
Stauffer, W., Sheng, H., and Lim, H.N.
EzColocalization: An ImageJ plugin for visualizing and measuring colocalization in cells and organisms (2018)

Please make sure everything is all set

Metric(s) of HL6187

	Label	PCC	Area	X
1	image 1: cell1	0.098	434	58.046
2	image 1: cell2	0.403	313	49.238
3	image 1: cell3	0.213	294	78.456
4	image 1: cell4	-0.140	412	247.876
5	image 1: cell5	-0.232	229	358.382
6	image 1: cell6	0.296	214	338.958
7	image 1: cell7	0.162	289	401.867
8	image 1: cell8	0.172	381	371.421
9	image 1: cell9	0.116	461	474.149
10	image 1: cell10	0.388	267	40.118
11	image 1: cell11	-0.433	362	445.721
12	image 1: cell12	0.248	411	428.607
13	image 2: cell1	-0.041	569	260.618
14	image 2: cell2	-0.125	394	321.579
15	image 2: cell3	-0.340	428	216.486
16	image 2: cell4	-0.556	287	24.629
17	image 3: cell1	-0.282	470	442.453
18	image 3: cell2	0.192	428	120.883
19	image 4: cell1	0.041	418	193.292
20	image 4: cell2	0.225	333	131.896
21	image 4: cell3	0.378	309	148.814
22	image 4: cell4	0.065	284	258.908
23	image 4: cell5	0.017	288	255.847

Log

```
Results Summary:

Reporter 1 image(s): Aligned Cy3 (Aligned)
Reporter 2 image(s): Aligned DAPI (Aligned)
Cell identification input image(s): phase contrast.tif

Number of cells analyzed = 46

*****************************************************
Pearson Correlation Coefficient (PCC) analysis of cell population
mean = 0.05617614395764327
standard deviation = 0.27232626127919346
median = 0.067330021164912

Interpretation:
-1 represents complete anti-colocalization
0 represents non-colocalization
1 represents complete colocalization

PCC is the correlation between the two channels for pixel values.
*****************************************************
```

```
-0.556                          0.606
Count: 46           Min: -0.556
Mean: 0.0396        Max: 0.606
StdDev: 0.267       Mode: 0.0251 (13)
Bins: 5             Bin Width: 0.232
```

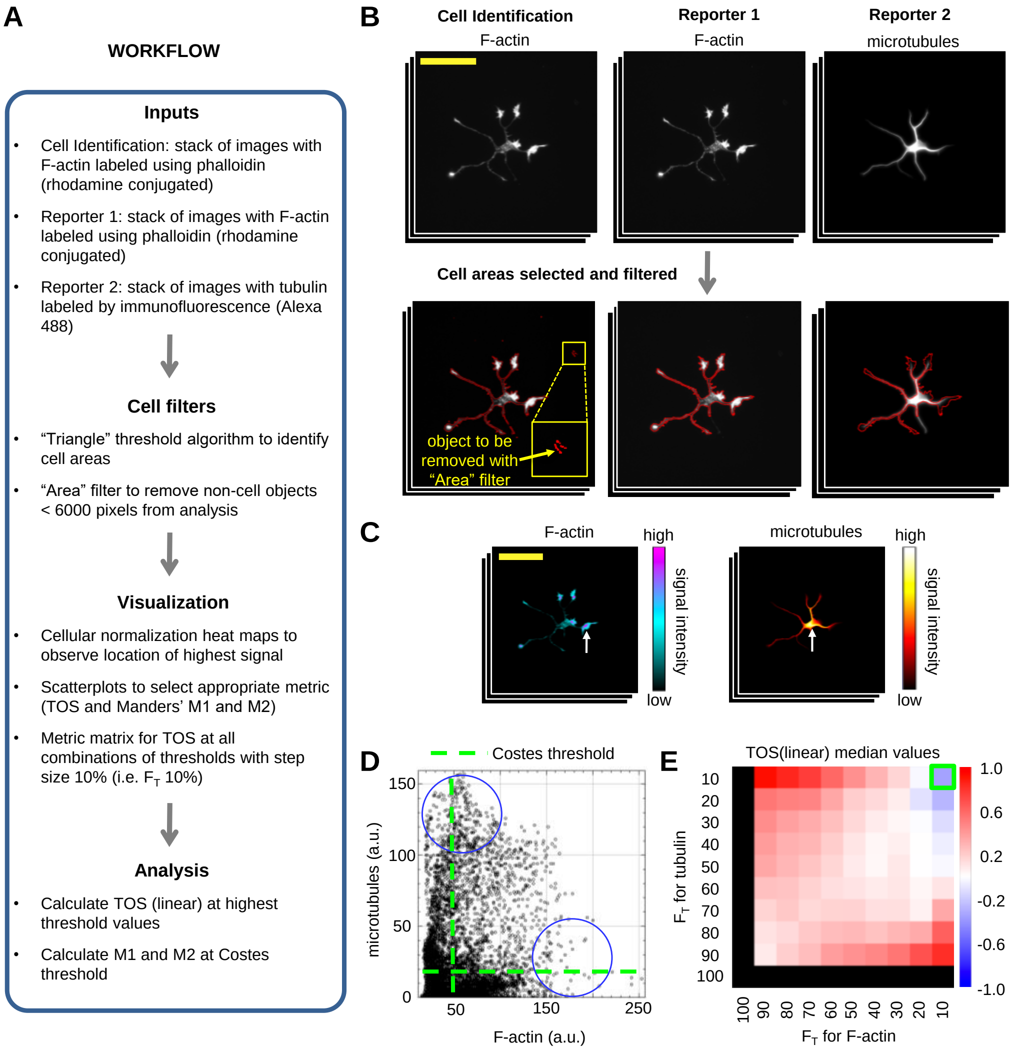
Figure 3.5 Analysis tab. (**A**) Analysis tab in the GUI for selection of default metrics. Note: this example is for two reporter channels (see **Figure 2.8F** for 3 reporter channels). (**B**) Analysis tab in the GUI for users to code custom metrics. The example code provided is for measuring colocalization by Pearson correlation coefficient. (**C**) Example of a data table showing metric values for Pearson correlation coefficient (PCC) and some of the parameter values for each cell in the analysis. Label = the image and unique cell number to identify individual cells; Area = area of each cell in pixels; and X = the average x-value of all pixels in a cell. Data is from the example used in **Figure 2.3**. (**D**) Summary report ("Log") of the results in **Figure 2.4C**. (**E**) Histogram generated from the results in **Figure 2.4C**. The height of each bin is the relative frequency. The Count is the number of cells. Mean is the mean value. StdDev is the standard deviation. Bins is the number of bins. Min and Max are the minimum and maximum values of the lowest and highest bin respectively (which are shown immediately under the histogram). Mode is the mode value. Bin Width is the width of each bin within the histogram.

2.3.2 Applications of EzColocalization

EzColocalization is designed to be used in a modular manner to facilitate customization of analyses for a wide variety of experiments and researcher needs. This section focuses on demonstrating specific tools in EzColocalization to solve real-world problems in diverse image sets.

In the first application of EzColocalization, images of rat hippocampal neurons from the Cell Image Library (CIL:8773, 8775-8788, which are attributed to Dieter Brandner and Ginger Withers) are used to demonstrate: (i) using a reporter channel for cell identification when the experiment does not have separate non-reporter images for cell identification; (ii) cell filters for selecting cells; and (iii) visualization tools for choosing metrics. The workflow of the analysis is outlined in **Figure 2.5A**. In the first step, two reporter image stacks were created: one stack with images where F-actin is labelled (using a phalloidin peptide conjugated to rhodamine); and the second stack with images where tubulin is labelled (using an antibody conjugated to Alexa 488) (**Figure 2.5B**). The interaction of F-actin and tubulin is important for the growth and migration of neurons [165, 166]. We used the F-actin images for cell identification because it is present in all cells and it shows the cell boundaries [151]. Individual cells were selected from the F-actin images by applying a threshold to identify cells [160] and using a cell filter to remove cell debris (note: parameter values in **Figure 2.5A**).

After the cells were selected, the intensity of reporter signals were examined using cellular heat maps and scatterplots. We found the reporters did not colocalize at high signal levels and there was a complex relationship between the signal intensities (**Figure 2.5C, D**). Due to the latter, localization was quantified using Manders' M1 and M2 and TOS (**Materials and Methods**). M1 and M2 were evaluated at thresholds selected by Costes' method, and the values were 0.289 and 0.995 respectively. These values are usually interpreted as indicating that tubulin has high colocalization with F-actin, and F-actin has low colocalization with tubulin. TOS values were evaluated by generating a metric matrix with median TOS values. The matrix showed colocalization, anticolocalization and non-colocalization at different thresholds for the signal intensities of tubulin and F-actin (**Figure 2.5E**). At sites in cells where F-actin and tubulin have the highest intensity signal (top 10% of pixels for each channel), the median TOS value is −0.36 (n = 20). This negative value indicates anticolocalization, which is consistent with the impression obtained from the heat maps and scatterplots, and with other reports [151].


WORKFLOW

Inputs

- Cell Identification: stack of images with F-actin labeled using phalloidin (rhodamine conjugated)
- Reporter 1: stack of images with F-actin labeled using phalloidin (rhodamine conjugated)
- Reporter 2: stack of images with tubulin labeled by immunofluorescence (Alexa 488)

Cell filters

- “Triangle” threshold algorithm to identify cell areas
- “Area” filter to remove non-cell objects < 6000 pixels from analysis

Visualization

- Cellular normalization heat maps to observe location of highest signal
- Scatterplots to select appropriate metric (TOS and Manders’ M1 and M2)
- Metric matrix for TOS at all combinations of thresholds with step size 10% (i.e. F_T 10%)

Analysis

- Calculate TOS (linear) at highest threshold values
- Calculate M1 and M2 at Costes threshold

Figure 2.5 Application 1: Cell selection using reporter images and physical parameters. Images are rat hippocampal neurons labelled with an F-actin probe and anti-tubulin antibody visualized by fluorescence microscopy (see main text). (**A**) Workflow of the analysis. (**B**) Cell identification using the F-actin reporter and filters to remove small non-cell objects (yellow arrow) based on their size (*i.e.* Area option from the cell filters). Large yellow box in left panel is a zoomed in view of the smaller yellow box. Red outline of the neuron indicates it has been identified as an object (*i.e.* a cell) for analysis. Scale bar is 100 µm. (**C**) Heat maps with cellular normalization showing localization regions of signal intensity for the cell shown in **panel B**. Scale bar is the same as **panel B**. (**D**) Scatterplot showing relationship between the signal intensity for two reporter channels for an arbitrary cell in the sample. Pixels with the highest intensity signal for each reporter channel have the lowest intensity signals for the other reporter, which indicates anticolocalization (blue circles). Green dash lines indicate thresholds selected by Costes’ method. (**E**) Metric matrix for the median TOS (linear) value for all cells in the sample (n = 20). Green box indicates the threshold combination where F-actin and tubulin have the highest intensity signal (top 10% of pixels for each channel); the median TOS value is -0.36.

In the second application, images of *Saccharomyces cerevisiae* undergoing mitosis were obtained from the Cell Image Library [167] to demonstrate: (i) cell identification via hand-drawn outlines (for experiments where automated methods of cell identification cannot be applied); and (ii) image alignment. The reporter inputs were an image from a wild type strain ("control"; CIL: 13871) that has the BFA1 protein that loads TEM1 onto the spindle pole body, and an image from a strain without the BFA1 protein (Δ*bfa1* deletion mutant; CIL: 13870). In these reporter images, cells expressed TEM1 protein fused to GFP and the DNA was labelled with DAPI (4', 6-diamidino-2-phenylindole). TEM1 localizes to spindle pole bodies during mitosis and is implicated in triggering exit from mitosis [167]. The workflow is shown in **Figure 2.6A**. In this application, ROIs were manually drawn around cells using the "Freehand" selection tool in ImageJ on DIC images. Binary masks, which were used to select cell areas, were created by selecting the ROIs and using the "Clear Outside" and then "Auto Threshold" functions of ImageJ [160] (**Figure 2.6B**). The cell areas were used for cell identification and to correct alignment between the DIC images and the reporter channels using the "default" threshold algorithm (**Figure 2.6C**). Following this cell identification and image alignment, the images are now ready for visualization and analysis as described in the previous example.

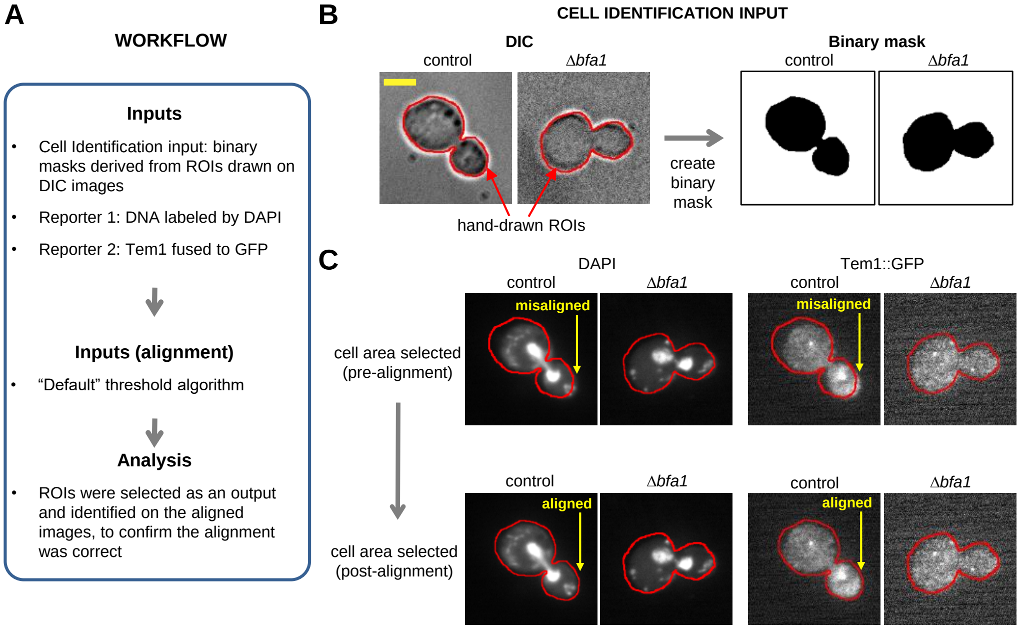
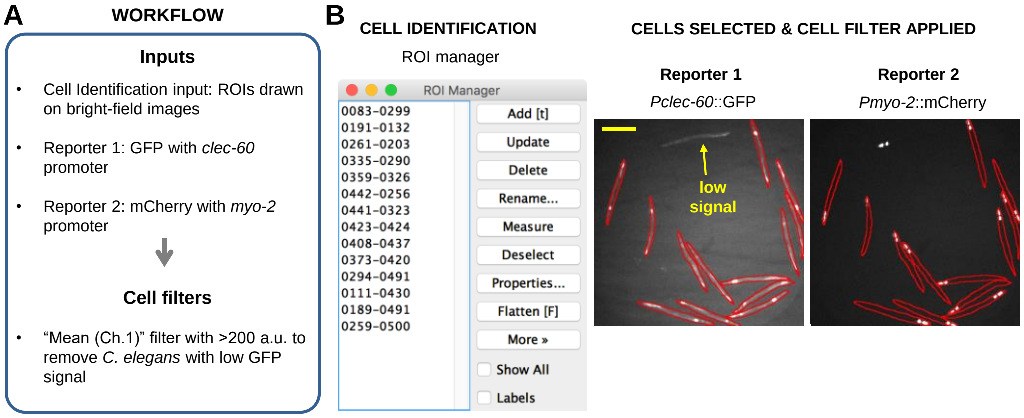

A
WORKFLOW
Inputs
Cell Identification input: binary masks derived from ROIs drawn on DIC images
Reporter 1: DNA labeled by DAPI
Reporter 2: Tem1 fused to GFP
Inputs (alignment)
"Default" threshold algorithm
Analysis
ROIs were selected as an output and identified on the aligned images, to confirm the alignment was correct
B
CELL IDENTIFICATION INPUT
DIC
control
Δbfa1
create binary mask
Binary mask
control
Δbfa1
hand-drawn ROIs
C
DAPI
control
Δbfa1
Tem1::GFP
control
Δbfa1
misaligned
misaligned
cell area selected (pre-alignment)
control
Δbfa1
control
Δbfa1
aligned
aligned
cell area selected (post-alignment)

Figure 2.6 Application 2: Image alignment. Images are *S. cerevisiae* with TEM1 translationally fused to GFP and DAPI staining visualized by DIC microscopy and fluorescence microscopy (see main text). (**A**) Workflow of the analysis. (**B**) Cell identification by hand-drawn ROIs on a DIC image and creation of a binary image mask. Red outline indicates the boundary of the hand-drawn ROI. Scale bar is 3.5 µm. (**C**) Alignment of the reporter images using the binary mask image. Arrows indicate areas of misalignment that are corrected. Red outline is the same as for Panel B.

In the third application, images of whole adult *Caenorhabditis elegans* obtained from the Broad Bioimage Benchmark Collection (BBBC012v1, M14) [140] were used to demonstrate that: (i) EzColocalization can analyze colocalization in whole organisms; and (ii) "cell" filters can select individual organisms based on reporter signal intensity. The images in this example are from the same dataset used in our study describing TOS (but they are not the same images) [151]. The workflow is shown in **Figure 2.7A**. Outlines of individual *C. elegans* were drawn in ImageJ on bright-field images to create ROIs, and the ROIs were added to the ROI manager for “cell” identification. GFP expressed from the *clec-60* promoter in the anterior intestine was reporter 1 and mCherry expressed from the *myo-2* promoter within the pharynx, which is an organ next to the anterior intestine [4], was reporter 2. Cell filters for physical parameters were unnecessary because only those objects considered to be suitable *C. elegans* had outlines drawn around them in the first place. However, cell filters for signal intensity were necessary because some *C. elegans* had low GFP signal, possibly due to transgene silencing [168, 169] (**Figure 2.7B**). Subsequent visualization and analysis was performed as described in the first application.



- Cell Identification input: ROIs drawn on bright-field images
- Reporter 1: GFP with *clec-60* promoter
- Reporter 2: mCherry with *myo-2* promoter

- “Mean (Ch.1)” filter with >200 a.u. to remove *C. elegans* with low GFP signal

```
0083-0299
0191-0132
0261-0203
0335-0290
0359-0326
0442-0256
0441-0323
0423-0424
0408-0437
0373-0420
0294-0491
0111-0430
0189-0491
0259-0500
```

Figure 2.7 Application 3: Cell selection using signal intensity parameters. Images are whole adult *C. elegans* with GFP expressed from the *clec-60* promoter and mCherry expressed from the *myo-2* promoter that are visualized by bright-field microscopy and fluorescence microscopy (see main text). (**A**) Workflow of the analysis. (**B**) Selection of *C. elegans* so that only those individuals with an average intensity for the reporter signal that is above a threshold level are included in analyses. Left image shows the ROI manager with a list of each ROI that was hand-drawn around each *C. elegans*. Right image shows the reporter channel images with red outlines indicating the boundaries of the ROIs. *C. elegans* below the threshold level were excluded (yellow arrow) from the analyses by using the cell filters for signal intensity. Scale bar is 250 µm.

In the fourth application, we demonstrate the analysis of colocalization for three reporter channels. The workflow was the same as for two reporter channels except “3 reporter channels” was first selected in the “Settings” main menu (**Figure 2.8A**). Images were obtained from the Broad Bioimage Benchmark Collection (BBBC025, Version 1, Image set: 37983, image: p23_s9) of U2OS bone cancer cells (*n* = 66) [170]. The three reporter images had DNA, endoplasmic reticulum (ER) and mitochondria respectively stained with Hoechst 33342, concanavalin A/Alexa Fluor488 conjugate, and MitoTracker Deep Red (upper row, **Figure 2.8B**). Cell identification was performed with an image of the plasma membrane labeled with wheat germ agglutinin (WGA)/Alexa Fluor 555 conjugate (upper left, **Figure 2.8B**). Note: the image also had the Golgi apparatus and F-actin network labeled [170]. The plasma membrane was traced using the polygon selection tool in ImageJ to create ROIs for the individual cells, and the ROI manager containing the ROIs was selected for cell identification.

The localization patterns were visualized in the same manner as for two reporters except that: (i) there are three sets of heat maps for the reporters instead of two (lower row, **Figure 2.8B**); and (ii) scatterplots and metric matrices are in three dimensions (**Figure 2.8C-F**). There is the option in the Visualization tab and the Analysis tab (**Figure 2.8G**) to measure colocalization for the three reporters using ICQ, TOS or Manders' M1, M2 and M3 metrics. Of the three metrics, we found that TOS was the easiest to interpret. TOS has a single value for measuring the colocalization of all three reporter signals, and it clearly showed the reporter signals for the nucleus, mitochondria and ER overlapped at low thresholds (i.e. at high F_T values there is colocalization; red color in **Figure 2.8E**) and did not overlap at high thresholds (i.e. at low F_T values there is anticolocalization; blue color in **Figure 2.8E**). These observations are consistent with the nucleus, mitochondria and ER organelles overlapping at their edges (where the signal from their reporters is typically lower) due to known physical interactions, but not at their centers (where the signal from their reporter is typically higher) because they are distinct structures in cells [171-173].

A

WORKFLOW

Settings

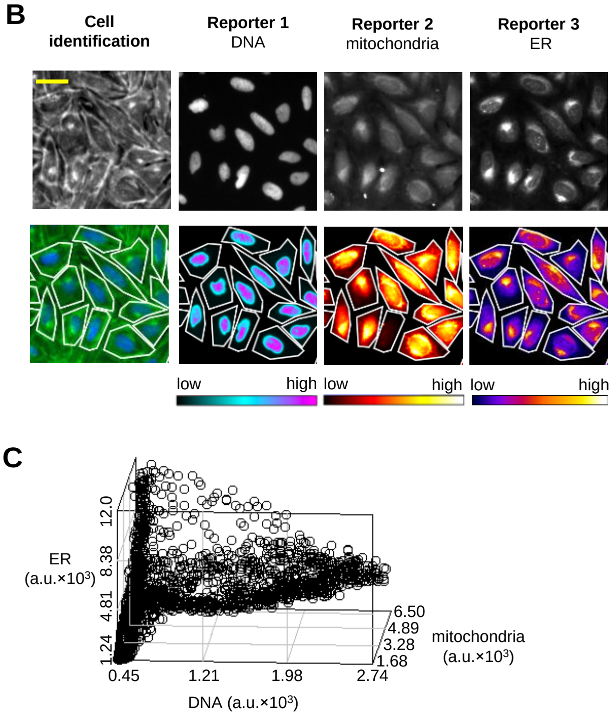
- Select “3 reporter channels”

Inputs

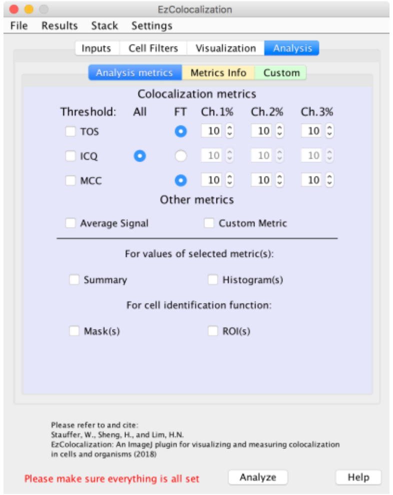
- Cell Identification: Hand drawn ROIs
- Reporter 1: DNA labeled by Hoechst 33342
- Reporter 2: Mitochondria labeled by MitoTracker Deep Red
- Reporter 3: ER labeled by concavalin A/Alexa Fluor 488 conjugate

Visualization

- Cellular normalization heat maps
- Three channel scatterplot
- Three channel metric matrix for TOS, M1, M2, M3, and ICQ at F_T 5%



G **Analysis tab for three reporter channels**



Colocalization metrics

Threshold:	All	FT	Ch.1%	Ch.2%	Ch.3%
□ TOS		●	10	10	10
□ ICQ	●	○	10	10	10
□ MCC		●	10	10	10

Other metrics

□ Average Signal □ Custom Metric

For values of selected metric(s):

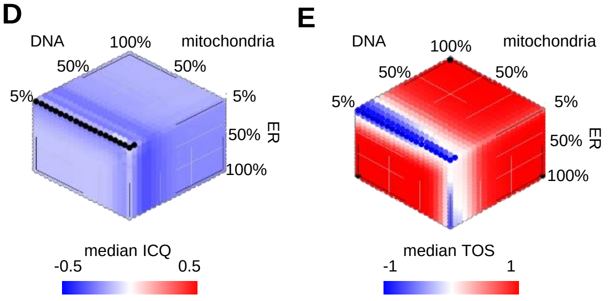
□ Summary □ Histogram(s)

For cell identification function:

□ Mask(s) □ ROI(s)

Please refer to and cite:
Stauffer, W., Sheng, H., and Lim, H.N.
EzColocalization: An ImageJ plugin for visualizing and measuring colocalization in cells and organisms (2018)

Please make sure everything is all set



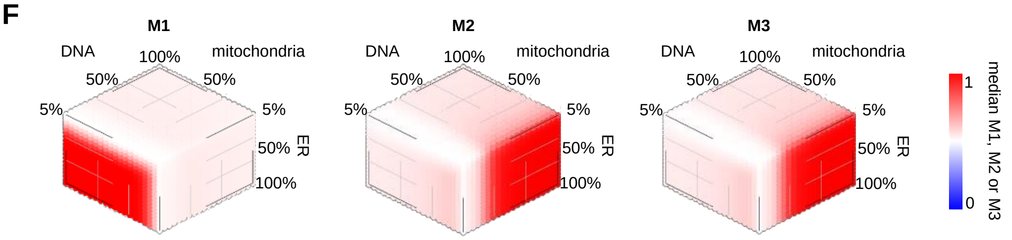


Figure 2.8 Application 4: Measurement of colocalization for three reporter channels. Images are of human bone cancer cells (U2OS) labelled as described in the main text. (**A**) Workflow of the analysis. (**B**) Images of cells in the cell identification and reporter channels. Top row are raw images. Bottom row, left image is the cell identification with pseudocolor (blue is the signal from Hoechst 33342 signal and green is the signal from phalloidin/Alexa Fluor 568 conjugate and wheat germ agglutinin/Alexa Fluor 555 conjugate) and boundaries of the ROIs in white (see main text). Bottom row (except left image) are heat maps for each of the three reporters with the boundaries of the ROIs shown. Signal intensity is indicated by the bar below each reporter image. Scale bar is 20 µm. (**C**) A three channel scatterplot for a single cell is shown for illustrative purposes only. (**D-F**) Metric matrices of median values for ICQ (**D**), TOS (**E**) and Mander's colocalization coefficients M1, M2 and M3 (**F**) for all cells in the analysis (n = 66). Note: black color on metric matrix for ICQ indicates there were no pixels above all three thresholds for some cells, and therefore ICQ could not be calculated. (**G**) Analysis tab for three reporter channels.

2.4 Discussion

EzColocalization was designed to make it easier for researchers to determine where particular types of molecules occur in cells and organisms in relation to other types of molecules. In addition, EzColocalization can provide data on colocalization for each cell or organism in a sample, which is increasingly recognized as being crucial to understanding biological process such as cell differentiation [174], cancer [175], and microbial pathogenesis [176].

Two of the most widely used existing applications for colocalization analysis are JACoP and Coloc2 [112, 153]. JACoP is an ImageJ plugin that can generate pixel intensity scatterplots to visualize localization patterns and measure colocalization with a variety of metrics including PCC (Van Steensel's CCF method or Costes' randomization), Manders' M1 and M2, ICQ, and object based methods [112]. It also permits thresholds to be chosen manually or automatically using Costes' method [112]. Coloc 2 is a plugin for Fiji [153], which builds on the functionality of JACoP by providing options to: analyze selected ROIs within single images, threshold images using a "bisection" algorithm, and measure colocalization with SRCC and Kendall's Tau rank correlation. Unfortunately, JACoP and Coloc 2 do not have built-in options to automate analyses or perform separate colocalization measurements for multiple objects in an image, therefore analyses can be challenging for images with a lot of background pixels or different cell types. The Wright Cell Imaging Facility (WCIF) has helped address these challenge by creating a colocalization plugin that can measure colocalization for individual cells by manually creating individual ROIs [113], but this method cannot be easily automated to analyze many cells across many images.

In addition to the above, software has been reported for measuring colocalization in cells, particularly in cases where the signal is defined to distinct regions or foci. One of these applications is MatCol, which can identify overlapping objects after a threshold is applied, and then calculate if the measured overlap is significantly different to that expected if the same objects were randomly scattered [177]. Another reported script calculates object based colocalization in confocal images [178] from the percent overlap of the objects. A third program measures colocalization for three-dimensional images; it measures the proportion of thresholded objects in one channel that have their center of mass within thresholded objects of another channel [179]. There are practical barriers to the widespread use of these three programs including the need for additional software to identify cell areas and that they are written in Matlab or C++ (therefore users must be familiar with these programming languages to customize them).

To make it easy to optimize analyses, EzColocalization has a simple GUI that requires no programming experience unless a custom metric is created. The GUI template is based on one that is familiar to many microscopists. ImageJ also has a large library of tools that can be used with EzColocalization, and it is open source software [160]. ImageJ has options for creating stacks of images and thresholding images, which were incorporated into EzColocalization for automated analyses. EzColocalization also has tools for the input of images, cell identification, visualizing localization patterns, measuring colocalization, and for displaying and saving results.

EzColocalization can select individual cells from cell identification images using thresholds, ROIs, or mask images. Identification of individual cells allows pixels within cells to be discriminated from pixels in the background and non-cell objects. In addition, cell filters can limit analyses to a subset of cells with certain physical parameters and minimum signal levels. Filters are used to select cells instead of more advanced techniques for cell detection [180] because: (i) they do not require assumptions about cell features (therefore diverse cell types can be analyzed); and (iii) they are intuitive, which makes it easier for researchers to tailor settings for their experiments and identify if patterns of localization are associated with specific cell features.

The visualization tools (heat maps, scatterplots, and metric matrices) can help with choosing the appropriate metrics and thresholds for the analyses. The metric matrices are particularly useful for samples with non-specific binding or localization of probes. These matrices display colocalization values for multiple combinations of thresholds for signal intensity, which facilitates the selection of thresholds so the analysis includes pixels from cellular regions with high signal (due to specific localization) and excludes pixels from regions with low signal (due to non-specific localization).

EzColocalization can not only measure colocalization for two reporters but also for three reporters. The latter is a useful feature that is unavailable for most software applications for measuring colocalization. In addition, custom metrics can be programmed in EzColocalization.

The data table generated by the colocalization analysis is an important feature of EzColocalization. Because the value of the colocalization metric for each cell is provided, and not just the average measurement of colocalization for the sample, it is possible to examine the distribution of metric values, perform statistical analyses, calculate receiver operator characteristic curves, and analyze subsets of cells in heterogeneous samples [151]. The data table also lists the specific image and a unique identifying number for each cell, therefore researchers can examine the images to determine why different cells have different measurements. The data tables can be downloaded and used in any spreadsheet application, which makes the data accessible to researchers without programming experience. Furthermore, the values for the physical parameters, signal intensity, and colocalization metrics can be retrieved from the tables (if the check box is selected) for more sophisticated multivariate analyses, including clustering, classifying and ordination methods.

In conclusion, EzColocalization is an ImageJ plugin with a user-friendly GUI, tools for start-to-finish analysis of colocalization, and many options to customize analyses. The tools are provided to select specific types of cells or organisms, visualize and measure colocalization, and automate analyses. The analysis generates a data table with measurements of colocalization, signal intensity and physical parameters for each cell, which allow users to delve deep into their data. Together these features make EzColocalization ideal for researchers at all levels, and for analyzing heterogeneous samples and complex patterns of localization.

2.5 Materials and Methods

2.5.1 EzColocalization development

The code for EzColocalization was written in Eclipse Java Integrated Development Environment (IDE) release 4.3.0 [181], which is a workspace for writing code and detecting compiling errors in Java^{TM}. EzColocalization incorporates ImageJ Application Program Interfaces (APIs) available from the National Institutes of Health, U.S. Department of Health and Human Services. An environment builder was used so that code written in the IDE ran in an instance of ImageJ as a plugin. This builder was implemented with Java Development Kit 8 [182] and the ImageJ source code within the IDE. The WindowBuilder [183] plugin for the IDE was used to design and generate the code for the GUI, and the code produced was restructured and revised to improve readability, and add listeners, which obtain user inputs from the GUI for running the plugin.

The basic level of organization of the code for EzColocalization are “classes”. Classes are independent blocks of code that represents a set of methods and variables; a class may be devoted to performing calculations which share code or calculations that are most conveniently performed together. Classes with related operations are grouped into a higher level of organization termed “packages”. For example, a class that generates heat maps and a class that displays heat maps may be bundled into the same package. The classes and packages are described in detail in the following section. It should be noted in regard to class functions that many processes within EzColocalization are performed as background computing, and thus the results of some functions, which are intermediates in longer methods, are not displayed and cannot be interacted with via the GUI.

2.5.2 Description of packages and classes for EzColocalization

The first two packages have very basic purposes. The first is the “default” package (by Java^{TM} convention) and its only function is to load the plugin within ImageJ. This package contains a single class, “EzColocalization_”, and outputs from this package are not accessible by other classes in other packages. The second package is “ezcol.files”, which has a single class (“FilesIO”) that loads all emblems and sample images for the GUI.

The third package is “ezcol.main”. It performs shared and general functions, and has six classes (“GUI”, “ImageInfo”, “MacroHandler”, “PluginStatic”, “PluginConstants”, and “AnalysisOperator”). GUI creates the GUI. ImageInfo stores information on the formats of the input images. MacroHandler enables use of the recorder in ImageJ so users can run macros that automatically run commands in batches. For example, the recorder can be used to create a macro to automatically modify and analyze a large set of images with particular settings. PluginStatic contains all static parameters (inputs) and static utility methods (common functions) used in analyses. PluginConstants contains all shared constants that are accessed by other classes. AnalysisOperator coordinates the operation of analyses in response to the inputs.

The fourth package is “ezcol.align”, which performs image alignment and has three classes (“BackgroundProcessor”, “TurboRegMod”, and “ImageAligner”). BackgroundProcessor enhances the contrast of images by: (i) subtracting background signal from pixels using the rolling ball algorithm in the “Subtract Background” function of ImageJ ; (ii) generating binary images of the reporter and cell identification channels with a user chosen algorithm from the “Auto Threshold” function in ImageJ or thresholds manually set by the user (note: only a single manual threshold can be applied for a stack of images, and this is performed by selecting “*Manual*” and then displaying the thresholds by selecting “Show threshold(s)”; if no manual selection is made, the “Default” algorithm is applied); (iii) converting all pixels above the value identified by the Auto Threshold algorithm to a value of 255, and all those below it to 0; (iv) applying the “Fill Holes” function of ImageJ on the binary images to better select the entire area of cells; and (v) calculating the average signal of pixels below the threshold in each reporter channel. TurboRegMod uses the “Translation” alignment algorithm from TurboReg to calculate the required XY coordinate shifts to align the binary images from the output of the BackgroundProcessor class by maximizing the overlap of pixels above the threshold. Note: interpolation of pixel values and other alignment functions that are normally performed by TurboReg are avoided because these functions alter pixel values. ImageAligner performs the image alignment by applying the calculated XY shifts from TurboRegMod to the original images.

The fifth package is “ezcol.cell”, which identifies cell areas and obtains pixel values. This package has six classes (“ParticleAnalyzerMT”, “CellFinder”, “CellFilterDialog”, “CellDataProcessor”, “DataSorter”, and “CellData”). Note: “cell” refers to any objects being analyzed, including subcellular structures or whole organisms. ParticleAnalyzerMT is a customized multithreading version of the “Analyze Particle” function from ImageJ and it is used to identify cell areas above the thresholds, which are pixels of the objects on the binary images produced by BackgroundProcessor (see previous package)). CellFinder takes inputs from the previous class and converts them into a format for the next class, performs watershed segmentation [184], and removes cells based on user defined cell filters. CellFilterDialog opens the window for additional cell filters. CellDataProcessor obtains the values of pixels identified for each cell. DataSorter and CellData sort the pixel values of cells based on intensity and store them so that these steps do not need to be repeated multiple times for later calculations.

The sixth package is “ezcol.metric”, which performs colocalization analysis in response to inputs from ezcol.cell, and contains six classes (“BasicCalculator”, “MetricCalculator”, “CostesThreshold”, “MatrixCalculator”, “MatrixCalculator3D”, and “StringCompiler”). BasicCalculator is an abstract class containing methods and values shared by the other “calculators” (*i.e.* MetricCalculator, MatrixCalculator, MatrixCalculator3D). MetricCalculator uses previously described algorithms to calculate Li’s ICQ, Manders’ colocalization coefficients M1, M2 and M3, PCC, SRCC, and TOS. “CostesThreshold” uses Costes’ method for determining a threshold and the algorithm was optimized for faster operation using ranked pixel values and dynamic programming as follows. The thresholds start at the maximum pixel values for each channel and PCC is calculated. Then the thresholds are decreased to the next highest pixel value, the values above the new threshold are subtracted from the stored sums, and PCC is calculated again from the new stored sums, and so on. During the entire process when all the pixels have been removed, we compare all the PCC values calculated for all thresholds to find the minimum absolute PCC value. MatrixCalculator calculates metric matrices for two reporter channels. MatrixCalculator3D creates metric matrices for three reporter channels. StringCompiler compiles and performs any custom analysis written by the user.

The seventh and eighth packages are “ezcol.visual.visual2D” and “ezcol.visual.visual3D”, which output plots and data from the analyses. These packages are located in the folder called “visual” and both obtain inputs from ezcol.cell for heat maps and scatterplots and from ezcol.metric for histograms and metric matrices.

The ezcol.visual.visual2D package contains nine classes for visualizing two dimensional data and results (“HeatGenerator”, “HeatChart”, “HistogramGenerator”, “HistogramStackWindow”, “ScatterPlotGenerator”, “PlotStackWindow”, “HeatChartStackWindow”, “OutputWindow”, and “ProgressGlassPane”). HeatGenerator normalizes pixels values so the maximum and minimum values are 0 and 255 (8-bit) or 65535 (16-bit) respectively for each cell, image, or stack. The normalized values are assigned colors from ImageJ lookup tables, or assigned from Matlab (R2015a, Mathworks, Natick, MA, USA) in the case of “hot” and “cool” colors. HeatChart is a modified version of the class JHeatChart (created by Tom Castle) which takes colors from the previous class to generate heat maps as RGB images, and values from MatrixCalculator to generate two dimensional metric matrices. HistogramGenerator and HistogramStackWindow generate histograms by respectively converting cell based data into histogram data starting with ten bins, and generating a stack of histograms for selected metrics. The number of bins can be increased or decreased in increments of one with the “nBin+” or “nBin−” buttons. ScatterPlotGenerator obtains pixel values from two reporter channels for five random cells per image in a stack. If five or less cells are present in an image, then pixel values are obtained for all cells in the image. PlotStackWindow creates and displays a stack of scatterplots, with each plot containing the pixel values for a single cell. HeatChartStackWindow generates the metric matrices window. OutputWindow generates the analysis summary window and its content. ProgressGlassPane generates the progress bar and presents tips in the GUI.

The ezcol.visual.visual3D package has 14 classes for visualization of three reporter channels in dynamic three dimensional scatterplots and metric matrices (“Arrow3D”, “Cube3D”, “Element3D”, “GraphicsB3D”, “Line3D”, “Point3D”, “Polygon3D”, “Rect3D”, “Renderer”, “ScatterPlot3D”, “ScatterPlot3DWindow”, “Spot3D”, “Square3D”, and “Text3D”). All classes are adopted from the jaytools.jar written by Urah Jay. His original classes are modified particularly for three dimensional scatterplots. Element3D is an abstract class (which means it cannot be initialized or constructed) containing methods and values shared by the other classes, including “Arrow3D”, “Cube3D”, “Element3D”, “Line3D”, “Point3D”, “Polygon3D”, “Rect3D”, “Spot3D”, “Square3D”, and “Text3D”. These classes represent the corresponding 3D elements as their names suggest; for example, “Arrow3D” is a class to indicate an arrow on a 3D graph. Some of these 3D elements (“Arrow3D”, “Polygon3D”, “Rect3D”, and “Square3D”) are not used for the purpose of this plugin but are kept for completeness of the package. Renderer is the 3D graphics process of automatically converting 3D elements into 2D image data, the results of which are feed into GraphicsB3D to paint the 2D image data of the projected 3D elements on the image canvas. ScatterPlot3D and ScatterPlot3DWindow generate 3D scatterplots and metric matrices by respectively converting the cell data into a compatible format for plotting and presenting the data in the plot window. 3D scatterplots are created in the same manner as ScatterPlotGenerator except for obtaining pixel values from three reporter channels. Usually the first cell is shown when the image window opens, and users can select the next cell or the previous cell by clicking the forward and back buttons on the window. 3D metric matrices are generated in the same way as HeatChart but all color squares are replaced by 3D spheres to enable visualization of deeper layers in the 3D matrices.

The ninth and final package is “ezcol.debug”. It has two classes, and reports errors and warnings within the plugin. It contains a class, “ExceptionReporter”, which handles and reports errors or warnings, and the class, “Debugger”, which was used during development to debug the plugin.

2.5.3 Testing of EzColocalization

EzColocalization was tested on images from experiments and on modified images created to test specific issues (*e.g.* misalignment). Unpublished images of bacterial cells (HL6187) were used to illustrate the different modules of EzColocalization (**Figure 2.1-2.4**). These bacteria had plasmid pHL1392 in strain HL3338 [185]. pHL1392 has the ampicillin resistance gene, ColE1 origin, and the green fluorescent protein (GFP) fused to part of the *sodB* gene and transcribed from the PLlacO-1 promoter. The sources of the images used for the application experiments (**Figure 2.5-2.8**) are stated in the relevant Results section. Note: images presented in the figures are cropped so that it is easier to see individual cells.

2.5.4 Download and installation

For users without ImageJ, the first step is to download and install the ImageJ application from: https://imagej.nih.gov/ij/download.html. Then the EzColocalization plugin can be downloaded from: http://sites.imagej.net/EzColocalization/plugins/. When saving the file, the user should delete the timestamp at the end of the name of the EzColocalization file. For example, a version named "EzColocalization_.jar-20190501171302" should be renamed as "EzColocalization_.jar". Once the plugin has been renamed "EzColocalization_.jar" it can be moved into the "plugins" folder of ImageJ to install it. Alternatively, users can install it by running ImageJ, selecting "Install..." from the "Plugins" menu of the menu bar, and then selecting the renamed file to install. To use EzColocalization, run the ImageJ application (open “ImageJ.exe” in the ImageJ folder) and choose "EzColocalization" from "Plugins" on the menu bar. For those using Fiji, the EzColocalization update site can be followed according to the instructions at https://imagej.net/Following_an_update_site.

2.5.5 Data acquisition guidelines

Accurate colocalization measurements begin with good experimental data, which depends on the samples, the reporters, the imaging system, data collection methods, controls and replicate measurements. Some guidelines the authors have found useful include the following. Samples should be prepared in a manner that: preserves the native spatial organization, minimizes touching of cells or organisms (which makes the identification of individual cells or organisms easier), and minimizes movement of cells or organisms (especially for live imaging). The reporters should be optimized to specifically label the molecules of interest (which includes minimizing excess or non-specifically bound reporter), and to minimize cross-talk (also known as bleed-through) between the reporter signals and between each reporter signal and non-reporter signals in cells and tissues (*e.g.* autofluorescence). In addition, reporters with the highest specific signal should be preferentially paired with targets that have the lowest concentration. Too much reporter is sometimes more problematic for colocalization measurements than too little because of non-specific labeling or aggregation. A reporter that identifies the cell boundary or entire cell should be considered if the cell boundaries are unclear in bright-field imaging to facilitate automated cell identification so that single cell measurements of colocalization can be easily performed. The imaging system should be set-up with: a high quality monochromatic camera to maximize the signal-to-noise ratio, controls to check the settings and reproducibility of measurements on different days, and adjustments to the light source or neutral density filters to prevent oversaturated pixels with artificially low intensity values. It is important to recognize that misalignment between imaging channels often occurs (and may occur after the initial set-up and alignment) therefore images from different channels should ideally be overlaid in each experiment to evaluate the alignment and to correct any misalignment by adjusting the physical apparatus or the analysis. Differential interference contrast (DIC) is not recommended, and users should instead use phase contrast or another method that does not create shadows for identifying cell boundaries. Generally, it is preferable to maximize the resolution, but the scale of the cells and structures must be considered. For example, measuring the colocalization of reporters in intracellular structures will require a higher level of resolution than measuring colocalization at different tissue structures or organs. Additional guidance on the practical aspects of setting up a system for colocalization measurements is available in several reviews [10-12]. The data should be collected at the highest number of bits to maximize the dynamic range of the signal, and images saved in an appropriate format (see note below). The importance of controls for the proper analysis of the colocalization measurements cannot be overstated. Researchers should not only include appropriate biological controls (*e.g.* deletions strains without the labeled protein) but should also measure some cells with only one of each reporter to quantify and to correct for bleed-through. In addition, independent replicate measurements of controls and samples on different days are important because labeling, microscopy set-up (especially in shared facilities), and any automated settings for image collection can vary dramatically between different experiments and often without the researcher being aware of it until the analysis is performed. As an aside, researchers should only use deconvolution or image corrections that have been proven to provide more accurate representation of localization for their specific reporters, samples, and imaging system.

The format of the images is important. The image file format should be TIFF or another lossless compression format with a single value for pixel intensity. A color camera that records pixel values in RGB can be problematic because it is unknown how the three values contribute to total signal intensity. Pseudocolors can be created for visualization purposes if the pixel intensity values are not changed. RGB and pseudocolor images can be distinguished by looking at the information on top of the image window in ImageJ.

2.5.6 Image alignment

The Inputs tab provides the option for the alignment of images from different channels. The alignment is performed by: (i) subtracting background signal from the cell identification and reporter images to enhance contrast using the rolling ball algorithm of the “Subtract Background” function (note: this step can be turned on or off in the “Parameters…” options of the “Settings” menu); (ii) thresholding the resulting images; (iii) creating a binary mask from the thresholded images; (iv) processing the binary mask with the “Fill Holes” function to ensure cell interiors are selected; (v) aligning the reporter channels and binary mask image using the translation alignment algorithm component of the TurboReg plugin [186]; (vi) obtaining the X and Y coordinate offset values from the alignment and using them to align the original cell identification and reporter images; and (vii) removing overhanging pixels and filling-in pixels (with a value of zero) so all images in the stacks have same size (yellow area in **Figure 2.1B**). Note: TurboReg functions that interpolate pixel values are not used because they change the original values.

2.5.7 Heat maps, scatterplots and metric matrices

Many factors should be considered when performing analyses and selecting a metric for quantifying localization. These should include heterogeneity in the data, the specificity of the reporter, the relative intensity of the intracellular and extracellular background signals, and the relationship between the intensities of the reporter signals. EzColocalization provides tools in the Visualization tab to help users evaluate these considerations.

Heat maps created by EzColocalization can be normalized for each cell, each image, or each stack (“cell heat maps”, “image heat maps” and “stack heat maps” respectively). Cell heat maps can help visually identify the locations in cells where reporters have the highest and lowest intensity, and the localization patterns (*i.e.* colocalization, anticolocalization and noncolocalization of the reporters). Image heat maps can show whether different cells have different average signal intensities within each image. The cell and image heat maps should be carefully inspected for evidence of heterogeneity among cells with respect to: the locations of reporters within cells, the localization pattern (*i.e.* relative positions of the reporters), and the average signal intensity. If there is heterogeneity, then it may be appropriate to limit analysis to a subpopulation of cells by using the cell filters in EzColocalization so that measurements are not an average of multiple populations. Image heat maps should also be examined to determine if the pixels with the highest signal (likely containing reporter) have similar levels of intensity to the pixels with the lowest signal (“background”). If so, then analysis may be improved by selecting individual cells from the image so that the only intracellular pixels are analyzed or by selecting thresholds so that only pixels with signal greater than background levels are analyzed (see metric matrices below) [121, 187] .

Scatterplots reveal the relationship between the signal intensities for different reporters. Evaluating this relationship is important because different assumptions about the relationship of the reporter signals are central to the calculation, interpretation and selection of the metrics for colocalization (see next section). Scatterplots may also reveal if different cells or organisms within a sample have very different intensities or different relationships between pixel intensity. If there is heterogeneity, cell filters may be able to limit analysis to a more homogeneous population. In addition to cell-to-cell heterogeneity, there may be heterogeneity within each cell; that is, different relationships between the signals at different levels of signal intensity. For example, a cell may have pixels with low signal for two reporters that have no correlation and pixels with high signal for the same two reporters that have a positive correlation (due to specific binding to a protein) [112]. In cases where there are different relationships between the pixels at different levels of signal, it may be possible to select thresholds for the reporter signals so that colocalization is only measured for a subset of pixels.

Metric matrices can be calculated for six different colocalization metrics in EzColocalization: TOS with linear or logarithmic scaling [151], PCC [113], SRCC [188], Manders' colocalization coefficients M1 and M2 [113, 127] and ICQ [189]. Each metric matrix calculates the value of the selected metric at every combination of the thresholds chosen (**Figure 2.9**). Metric matrices can quickly determine whether there are general patterns of colocalization, anticolocalization or noncolocalization that depend on signal intensity [3, 151]. A metric matrix can also help to select a threshold that provides a better measure of colocalization for a subset of pixels with different intensities in a cell. That is, the selection of thresholds via the metric matrices can provide more targeted analysis. Because the thresholds in the metric matrix are measured in terms of the percentage of pixels rather than absolute signal level, the metric matrix is well-suited to comparing and aggregating values in a groups of cells where there may be some differences in average signal intensity and cell size.

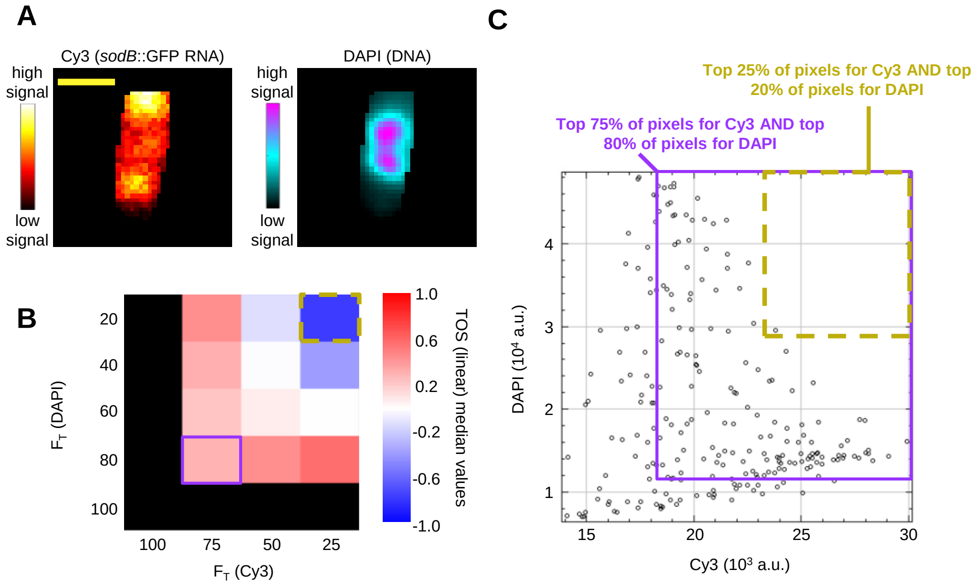


Figure 2.9 Metric matrices and selected fractions. (**A**) Heat maps showing the intensities of Cy3 and DAPI signal for *sodB::gfp* RNA and DNA respectively in a bacterial cell. The *sodB::gfp* RNA was labeled with Cy3 labeled probes by RNA fluorescence in-situ hybridization. Scale bar is 1 µm. (**B**) Metric matrix with TOS values (linear) for the cell in **Panel A**. Each box in the matrix is the TOS value calculated for the pixels that are above the threshold for each channel. The thresholds are measured as the percentage of pixels with the highest signal for each channel (F_T). For this example, the chosen F_T are the top 100%, 75%, 50% and 25% for Cy3 and the top 100%, 80%, 60%, 40% and 20% for DAPI. The calculated value of TOS is shown for every combination of thresholds and the approximate value is displayed in the bar to the right. The box is colored black when at least one threshold is 100% because in such cases TOS values are not informative; that is, when 100% of pixels are selected for at least one reporter then the overlap with the other channel must always be 100%. Threshold combinations indicated by the purple box and gold dash line box are discussed in **Panel C**. (**C**) Scatterplot of the pixels in the cell in **Panel A**. The purple box has pixels that are both in top 75% and the top 80% of values for Cy3 and DAPI respectively, which are used to calculate the TOS value shown in the purple box in the metric matrix (**Panel B**). The gold dash line box has pixels that are both in top 20% and the top 25% of values for Cy3 and DAPI respectively, which are used to calculate the TOS value shown in the gold box in the metric matrix (**Panel B**).

In relation to selecting thresholds, EzColocalization provides two options: Costes' method and manual selection. Costes' method chooses the thresholds automatically [124]. The advantage of automatic selection of the thresholds is that it decreases the potential for user bias. However, the method often does not work well if the signal intensities of the intracellular and background pixels are not clearly distinguishable, the reporter signals do not have similar levels of intensity or a monotonic relationship, or there are outlier pixels with high signal [113]. Manual selection of thresholds by the user is more flexible but it requires care to ensure they are chosen appropriately. The heat map and scatterplots, as well as the metric matrices, can guide the manual selection of the thresholds. Metric matrices can help ensure the thresholds are chosen so that they are representative of broad trends and the results are robust (*i.e.* a small change in the values of the thresholds should not substantially alter the result). Two notes of caution in regard to the selection of thresholds: (i) the metric matrix should not be used to "fish" for a metric and threshold values to give a result that is not broadly consistent with all the data; and (ii) the selection of thresholds must balance the need to eliminate pixels with background or non-specific signal against the need to keep as many pixels as possible so the results of the analysis are broadly representative and not fluctuating due to the noise associated with having a small number of values. Additional guidance on the selection of thresholds is provided in previous publications [3, 151].

2.5.8 Colocalization metrics for two reporter channels

This section provides brief and general guidelines for selecting a colocalization metric. More detailed information on colocalization metrics is published elsewhere [112, 121, 151, 190]. As mentioned in the previous section, it is important to examine the scatterplots to determine the relationship between the signal intensities of the reporters before choosing a colocalization metric.

Pearson's correlation coefficient (PCC) is the covariance of two variables divided by the product of their standard deviations. It is typically used to measure the linear correlation of the signal intensity values for two reporters (**Figure 2.10**) [115, 191]. PCC values can range from −1 which indicates a strong negative correlation between the signals (anticolocalization) to 1 which indicates a strong positive correlation (colocalization). A PCC value of 0 indicates there is no correlation (noncolocalization). Note: PCC could be used to measure nonlinear relationships following a nonlinear transformation of the data, although this is not typically done for measuring colocalization.

Spearman's rank correlation coefficient (SRCC) is calculated by ranking the pixels according to the intensity of signal for each channel and then measuring the correlation in the rankings between two channels [188]. SRCC measures whether the signal intensities of the reporters have a monotonic relationship (**Figure 2.10**), and it is relatively insensitive to outliers because it is based on rankings. Therefore SRCC is suitable for non-linear, monotonic relationships such as power law or logarithmic functions. SRCC values range from −1 (anticolocalization) to 1 (colocalization) [190], and 0 indicates there is no correlation (noncolocalization).

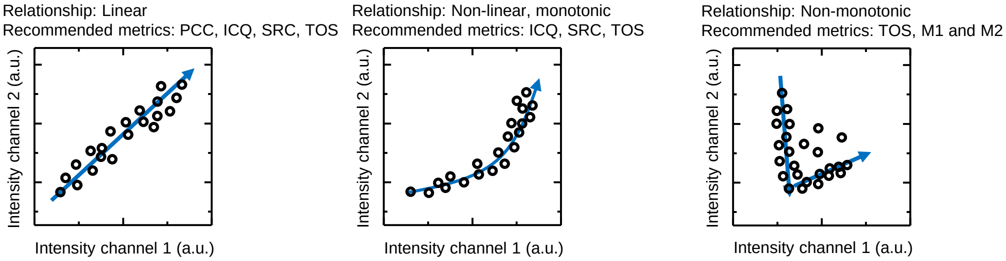


Figure 2.10 Scatterplots identify the relationship between signal intensities. Scatterplots reveal the relationship between the intensities of different reporters, which is important for selecting an appropriate colocalization metric. Three relationships and the recommended metric for measuring colocalization for each are shown (see text of **Materials and Methods**). The blue line and the circles indicate the hypothetical relationship and hypothetical data points respectively.

The intensity correlation quotient (ICQ) is the ratio of the total number of pixels where the signal intensity is above the means for both channels or below the means for both channels (*i.e.* excluding pixels that are above the mean in one channel and below the mean in the second channel), divided by the total number of pixels, minus 0.5 [187, 189]. ICQ ranges from –0.5 to +0.5. ICQ is essentially a sign test with positive or negative values for pixels that are on a positive or negative slope of a function through the mean of both channels. ICQ, like SRCC, is often used to evaluate whether the signal intensities of two reporters have a monotonic relationship (although it could also be used for some types of non-monotonic relationships) (**Figure 2.10**). ICQ is less sensitive to outliers than PCC. ICQ is not an appropriate metric for heterogeneous samples because the mean may not be an appropriate point around which localization should be evaluated.

Manders' colocalization coefficients M1 and M2 are calculated by determining the sum of the intensities of pixels that either exceed thresholds for both signals 1 and 2 divided by the sum of the intensities of the pixels that exceed the threshold for signal 1 or exceed the threshold for signal 2, respectively [113, 127]. The threshold can be determined by several algorithms including Costes' threshold [124]. Disadvantages of M1 and M2 are that both values are needed to determine whether there is colocalization, and the interpretation of these values is complicated by them being dependent on the threshold values [118]. Manders' colocalization coefficients M1 and M2 (and also the threshold overlap score defined below) tend to be better for evaluating colocalization or anticolocalization in cases where there is not a clear localization pattern, there is a mixed pattern of localization, or there is a non-monotonic relationship (**Figure 2.10**).

The threshold overlap score (TOS) is a newer metric that shares some similarity to M1 and M2 in that it calculates the overlap in pixels above a threshold [3, 151]. TOS is calculated by determining the number of pixels that exceed thresholds for both signals 1 and 2 and dividing this number by the number of pixels that exceed the threshold for signals 1 or 2 respectively. Unlike M1 and M2 there is no weighting for signal intensity. In addition, TOS normalizes the observed overlap by the overlap expected to occur simply by chance (which is not done for M1 and M2). A result of this normalization is that TOS measures colocalization as a single value which makes it easier to interpret and compare between experiments than Manders' colocalization coefficients. TOS values are rescaled so that −1 corresponds to the minimum possible overlap (anticolocalization), 0 corresponds to the same overlap as would occur by chance (noncolocalization), and 1 corresponds to the maximum possible overlap (colocalization). The default rescaling option is linear because it is easily interpreted, and its value reflects the fraction between random distribution and the minimum or maximum values (-1 or +1 respectively). For example, a value of 0.5 represents half the maximum possible overlap. EzColocalization also permits a logarithmic rescaling (natural log) for users requiring a metric without a discontinuity in the first derivative, but it is harder to interpret than linear rescaling [151]. As mentioned above, TOS is suitable for the analysis of experiments that have non-monotonic relationships, mixed patterns of localization, or unclear localization patterns (**Figure 2.10**). TOS can also be used for monotonic relationships including linear correlations, although in such cases it may not be as sensitive or specific as other metrics (*e.g.* PCC).

In summary, PCC is often used for datasets where the reporters have an approximately linear relationship between the pixel values. SRCC and ICQ are commonly used to evaluate whether the signal intensities of the two reporters have a monotonic relationship, and are generally considered more robust to outliers than PCC. Manders' M1 and M2 or TOS are often preferred in cases where there is no clear monotonic localization pattern, or mixed patterns of localization.

2.5.9 Colocalization metrics for three reporter channels

The metrics for two channels were: (i) PCC; (ii) SRCC; (iii) ICQ; (iv) Manders' coefficients; and (v) TOS with linear or logarithmic rescaling. Of these metrics, we extended ICQ, Manders' coefficients and TOS (linear or logarithmic rescaling) to measure colocalization for three reporters, and their derivations are below. PCC and SRCC were not extended for three reporters because their meaning and interpretation becomes much more complicated. Specifically, no single value of PCC (or SRCC) can represent the standardized covariance. Instead there are multiple values, each of which reports the extent two channels (independent variables) can predict the signal in the third channel (dependent variable). The first component of principal component analysis (PCA) should be used to measure linearity without assuming dependency of three channels [192]. However, PCA is difficult to interpret in relation to colocalization analysis and therefore was not included [193].

Li's ICQ [189] can be easily expanded to three (or more) channels.

$$\mathrm{ICQ} = \frac{N_{above}+N_{below}}{N_{total}} - 0.5, \qquad \text{Eq. 2.1}$$

where N_{above} is the number of pixels above the means of all three channels, N_{below} is the number of pixels below the means of all channels, and N_{total} is the total number of pixels. For two channels, ICQ is a crude measure of the fraction of pixels that are on the positive diagonal; that is, it can be interpreted as the fraction of pixels that are broadly consistent with a monotonic increasing relationship. For three channels, ICQ provides a crude measure of whether pixel values tend to increase in all three channels. However, the interpretation of the value is more complicated because of the combinatorics; a pixel may have values above or below the mean in 8 possible combinations. A value of −0.25 would be expected if the pixel values have a random distribution, and assuming the median and mean values are approximately equal. In this case, a value >−0.25 may indicate a positive relationship, but it does not exclude the co-presence of a negative relationship. A value of <−0.25 indicates a negative relationship but it does not rule out a positive relationship in a subset of pixels.

The use of Manders' colocalization coefficients for three channels (*i.e.* M1, M2, and M3) has been previously reported [127]. The derivation of Manders' colocalization metrics for more than two channels is straight forward as it simply evaluates the proportion of overlapping signal. However, Manders' colocalization metric are often used with an automated method of threshold selection, such as Costes' method, and these methods typically do not readily extend to three channels [124]. Therefore, EzColocalization users can either select thresholds manually or by using the metric matrix for Manders' colocalization coefficients with three channels. The thresholds are measured as F_T. M_1, M_2, … M_n can be calculated by **Eq. 2.1** where there are at least two reporters:

$$\mathrm{M}_i = \frac{\sum \mathrm{G}_{i,coloc}}{\sum \mathrm{G}_i}, \qquad \text{Eq. 2.2}$$

where $\mathrm{G}_{i,coloc}$ is the value of each pixel in channel i that is above all thresholds and G_i represents the value of each pixel in channel i that is above the threshold for only channel i. The number of Manders' colocalization coefficients is equal to the number of channels, therefore three values need to be interpreted for three channels. Three values can be difficult to interpret collectively and to compare colocalization between samples. Another challenge is that the interpretation of the Manders' colocalization coefficient depends on the selected thresholds [151].

TOS measures the overlap of the signal above the threshold for each channel accounting for the amount of overlap that would be expected to occur by random chance for different thresholds [151]. One of the first steps in calculating TOS is to determine the number of pixels in each cell that exceed the thresholds for all three reporter channels (A_{coloc}) and the number of pixels that exceed the threshold for one of the reporter channels (A_i, where *i* is the *i*th channel). Dividing the former by the latter is the "observed AO". This calculation, is equivalent to calculating the fraction of pixels in the cell that exceed the threshold for all three channels (F_{coloc}) divided by the fraction of pixels that exceed the threshold for the chosen channel *i* (F_{Ti}). That is,

$$\text{observed } \mathrm{AO}_i = \frac{\mathrm{A_{coloc}}}{\mathrm{A}_i} = \frac{\mathrm{A_{coloc}}/\mathrm{A_{total}}}{\mathrm{A}_i\,/\mathrm{A_{total}}} = \frac{\mathrm{F_{coloc}}}{\mathrm{F}_{\mathrm{T}i}}, \text{where } i = 1, 2 \text{ or } 3. \qquad \text{Eq. 2.3}$$

Note: F_{Ti} and F_{coloc} are fractions rather than percentages for all equations in this section, and are defined as greater than zero and less than or equal to one.

The next calculation is the expected AO value assuming uniformly distributed random pixel values. If the pixels above the threshold for the first channel are randomly distributed throughout the cell, then the chance a pixel above the threshold for the second channel overlaps one of the pixels that exceeds the threshold for the first channel, is simply equal to the fraction of pixels above the threshold for the first channel (previously explained elsewhere [151]). Following from this, the chance a pixel that exceeds the threshold for the third channel overlaps a pixel that already exceeds both the first and second channels is simply the product of the fraction of pixels that exceed the first and second reporter channels. Therefore, the

$$\text{expected } \mathrm{AO}_i = \frac{\mathrm{F_{T1}} \times \mathrm{F_{T2}} \times \mathrm{F_{T3}}}{\mathrm{F}_{\mathrm{T}i}}, \text{where } i = 1, 2 \text{ or } 3. \qquad \text{Eq. 2.4}$$

The observed AO is divided by the expected AO to generate the "AO ratio", which accounts for the increase in overlap that occurs with selection of more pixels (*i.e.* greater F_T).

$$\text{AO ratio} = \frac{F_{\text{coloc}}}{F_{T1} \times F_{T2} \times F_{T3}}. \qquad \text{Eq. 2.5}$$

The AO ratio is equal to 1 for cells where the overlap is the same as expected by chance. The value of the AO ratio depends on whether the observed overlap is more or less than expected by chance as well as the selected thresholds. The latter can make interpretation difficult, therefore the AO ratio is rescaled to generate the TOS, which enables easier comparison of analyses with different thresholds.

To rescale the AO ratio, the minimum and maximum value must be determined for the thresholds. The minimum AO ratio can be zero if the sum of the F_T for two channels is less than or equal to 1 (*i.e.* $F_{T1} + F_{T2} \leq 1$). Note: for the above F_{T1} and F_{T2} are greater than zero; if A_1 or A_2 are zero then there is no overlap and F_{T1} and F_{T2} should not be calculated. In the case where the first two channels do not overlap, the threshold for the third channel is inconsequential. If there is no overlap of pixels above the thresholds for two channels then there can be no overlap of all three channels, even if all the pixels are selected for the third channel (*i.e.* $F_{T3} = 1$) and consequently the minimum AO ratio would be zero. That is, if $F_{T1} + F_{T2} + F_{T3} \leq 2$, then it is possible for the minimum AO ratio to be equal to zero. If $F_{T1} + F_{T2} + F_{T3} > 2$ then overlap of all three channels must occur by at least the amount exceeding 2. In summary,

$$\text{minimum AO ratio} = \begin{cases} \frac{F_{T1} + F_{T2} + F_{T3} - 2}{F_{T1} \times F_{T2} \times F_{T3}}, \text{when } F_{T1} + F_{T2} + F_{T3} > 2 \\ 0, \text{ when } F_{T1} + F_{T2} + F_{T3} \leq 2 \end{cases}. \qquad \text{Eq. 2.6}$$

The limits of the minimum AO ratio are 0 and 1.

The maximum AO ratio occurs when all three channels maximally overlap, and the maximum amount of overlap can be no more than the minimum F_T. For example, if two channels both have thresholds that select 80% of pixels and the third channel only selects 5% of pixels in the cell, then the maximal overlap of the selected pixels can be no more than 5% of the pixels in the cell; that is, the minimum of the three F_T values.

$$\text{Maximum AO ratio} = \frac{\text{minimum}\{F_{T1},\ F_{T2},\ F_{T3}\}}{F_{T1} \times F_{T2} \times F_{T3}}. \qquad \text{Eq. 2.7}$$

The last step in calculating TOS is to rescale the AO ratio using the limits for the minimum AO ratio and for the maximum AO ratio as previously reported [151]. A TOS value reflects the fraction of the "distance" between random chance (also known as the null distribution) and the minimum or maximum possible overlap for the thresholds. A positive value indicates colocalization, zero indicates overlap that is no more or less than a random distribution, and a negative value is anticolocalization. For example, 0.5 is halfway between a random distribution and maximum TOS value (half-maximal colocalization for the chosen thresholds) and −0.5 is halfway between a random distribution and the minimum possible TOS value (half-maximal anti-colocalization for the chosen thresholds). It should be noted that the contribution of each channel to the colocalization measurement is not specified in the TOS value. Therefore, anticolocalization may be due to one single channel not overlapping with the other two (as opposed all channels not overlapping).

2.5.10 Custom analysis

The Custom subtab in the Analysis tab allows users to perform custom mathematical analysis for all pixel intensity values in selected cells without having to directly modify the code for EzColocalization. In brief, custom written code inserted into the Custom tab of the plugin uses the same cells or organisms that would be selected by the cell filters (with any alignment used) for non-custom analyses. Each cell's pixel intensity value for each reporter channel are stored in an array, named c1, c2, and c3 for reporter channels 1, 2, and 3 respectively. The order of the pixels within each array is the same; that is, the same index within each array is the same pixel in each channel, and is the intensity value for that channel. The pixel values in the arrays can be analyzed using code written with standard mathematical functions in Java. Selecting the "Resource" button takes the user to a website with a list of operators and functions in Java for mathematical calculations.

Table 2.1 Input images and possible outputs from EzColocalization. * Cell identification channel may be reporter images as discussed in main text. Therefore it is possible to perform all possible analyses with two sets of images (with one set being used as both a cell identification image and a reporter image). # Cell identification images are required to distinguish intracellular and extracellular signal therefore without them any analysis or normalization must be for whole images or image stacks.

Inputs			Outputs						
Cell identification channel*	1 reporter channel	2 or 3 reporter channels	Report of physical features of cells	Visualization: scatterplots, & metric matrices of cell	Visualization: heat maps of cell	Visualization: scatterplots & metric matrices of image or stack	Visualization: heat maps of image or stack	Analysis: measure colocalization in cells	Analysis: measure colocalization in images & stacks #
Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Yes	Yes	No	Yes	No	Yes	No	Yes	No	No
Yes	No	No	Yes	No	No	No	No	No	No
No	Yes	Yes	No	No	No	Yes	Yes	No	Yes
No	Yes	No	No	No	No	No	Yes	No	No
No	No	No	No	No	No	No	No	No	No

Table 2.2. Physical and signal intensity parameters for cell features. *Units for pixel size are arbitrary units unless users set a scale on the images.

Physical (P) or Signal Intensity (S) parameter	Name	Units	Description
P	Area	pixels*	Number of pixels in a cell.
P	X	pixels*	Average x-coordinate of a cell.
P	Y	pixels*	Average y-coordinate of a cell.
P	Perimeter length	pixels*	Length of the outside boundary of a cell.
P	Width	pixels*	Width of a cell in the x-axis.
P	Height	pixels*	Height of a cell in the y-axis.
P	BX	pixels*	Top left x-coordinate of the smallest rectangle enclosing a cell.
P	BY	pixels*	Top left y-coordinate of the smallest rectangle enclosing a cell.
P	Major	pixels*	Primary axis of the best fit ellipse for a cell.
P	Minor	pixels*	Secondary axis of the best fit ellipse for a cell.
P	Circularity	unitless	Circularity of a cell calculated by $4 \pi \times$ area $\div$ perimeter^2. A value of 1 is a perfect circle and <1 is an ellipse.
P	Angle	degrees	Angle between the main axis of an ellipse fit to a cell and x-axis of the entire image containing the cell.
P	Feret’s diameter	pixels*	Longest possible distance between any two points on a cell boundary.
P	FeretX	pixels*	Starting x-coordinate of the Feret’s diameter of a cell.
P	FeretY	pixels*	Starting y-coordinate of the Feret’s diameter of a cell.
P	Feretangle	degrees	Angle between a cell’s Feret’s diameter and its images x-axis.
P	MinFeret	pixels*	Minimum caliper diameter of a cell.
P	AR	unitless	Aspect ratio of a cell calculated by major axis $\div$ minor axis.
P	Round	unitless	Roundness of a cell calculated by $4 \times$ Area $\div \pi \times$ major axis^2.
P	Area fraction	unitless	Percentage of pixels in an image which are included in a cell.
P	Solidity	unitless	Solidity of a cell calculated by its area $\div$ area of its convex hull.
S	Mean (Ch1), (Ch2), or (Ch3)	arbitrary	Mean of pixel values for a cell in reporter channels 1, 2, or 3.
S	Mode (Ch1), (Ch2), or (Ch3)	arbitrary	Mode of pixel values for a cell in reporter channels 1, 2, or 3.
S	Median (Ch1), (Ch2), or (Ch3)	arbitrary	Median of pixel values for a cell in reporter channels 1, 2, or 3.
S	Minimum (Ch1), (Ch2), or (Ch3)	arbitrary	Minimum pixel value for a cell in reporter channels 1, 2, or 3.
S	Maximum (Ch1), (Ch2), or (Ch3)	arbitrary	The maximum pixel value for a cell in reporter channels 1, 2, or 3.
S	StdDev (Ch1), (Ch2), or (Ch3)	arbitrary	Standard deviation of pixel values for a cell in reporter channels 1, 2, or 3.
S	Skew (Ch1), (Ch2), or (Ch3)	unitless	Skewness of pixel values for a cell in reporter channels 1, 2, or 3.
S	Kurt (Ch1), (Ch2), or (Ch3)	unitless	Kurtosis of pixel values for a cell in reporter channels 1, 2, or 3.
S	RawIntDen (Ch1), (Ch2), or (Ch3)	arbitrary	Sum of all pixel values for a cell in reporter channels 1, 2, or 3.
S	IntDen (Ch1), (Ch2), or (Ch3)	arbitrary	Product of pixel number and average pixel value for a cell in reporter channels 1, 2, or 3.
S	Mean BgndRatio (Ch1), (Ch2), or (Ch3)	fold background	Fold change of average pixel value for a cell versus the average pixel value for all pixels outside of cells in reporter channels 1, 2, or 3. For example, “1x-2x” for Ch1 would select cells with a mean pixel value one to two fold the mean pixel value outside cells for reporter channel 1.
S	Median BgndRatio (Ch1), (Ch2), or (Ch3)	fold background	Fold change of median pixel value for a cell versus the median pixel value for all pixels outside of cells in reporter channel 1, 2, or 3. For example, “1x-2x” for a Ch1 filter input would select cells with a median pixel value one to two times the value of the median pixel value outside cells in reporter channel 1.

Chapter 3: Nucleoid and cytoplasmic localization of small RNAs in *Escherichia coli*[3]

3.1 Abstract

Bacterial small RNAs (sRNAs) regulate protein production by binding to mRNAs and altering their translation and degradation. sRNAs are smaller than most mRNAs but larger than many proteins. Therefore it is uncertain whether sRNAs can enter the nucleoid to target nascent mRNAs. Here, we investigate the intracellular localization of sRNAs transcribed from plasmids in *Escherichia coli* using RNA fluorescent *in-situ* hybridization. We found that sRNAs (GlmZ, OxyS, RyhB, and SgrS) have equal preference for the nucleoid and cytoplasm, and no preferential localization at the cell membrane. We show using the *gfp* mRNA (encoding green fluorescent protein) that non-sRNAs can be engineered to have different proportions of nucleoid and cytoplasmic localization by altering their length and/or translation. The same localization as sRNAs was achieved by decreasing *gfp* mRNA length and translation, which suggests that sRNAs and other RNAs may enter the densely packed DNA of the nucleoid if they are sufficiently small. We also found that the Hfq protein, which binds sRNAs, minimally affects sRNA localization. Important implications of our findings for engineering synthetic circuits are: (i) sRNAs can potentially bind nascent mRNAs in the nucleoid, and (ii) localization patterns and distribution volumes of sRNAs can differ from some larger RNAs.

[3] This chapter has been previously published as "Nucleoid and cytoplasmic localization of small RNAs in *Escherichia coli*" in *Nucleic acids research* (2017), *45*(5), 2919-2934.

3.2 Introduction

Bacterial small RNAs (sRNAs) regulate the production of diverse classes of proteins in a wide variety of pathways [194]. Most sRNAs bind to target mRNAs at or near the translation initiation region (TIR) and form sRNA-mRNA duplexes. Duplex formation commonly decreases translation and/or increases mRNA degradation resulting in decreased target protein production [194]. Less often, duplex formation has the opposite effect, causing increased target protein production [194]. Some sRNAs can decrease the production of target proteins and increase the production of others [195-200]. Additionally, some sRNAs regulate gene expression by: binding directly to the σ^{70}-RNA polymerase holoenzyme to alter transcription [201], sequestering proteins [202], and being translated into an active peptide [203].

Most sRNA regulation in *Escherichia coli* and in many other bacteria requires the Hfq protein [204]. Hfq primarily exists as a hexamer that can bind sRNAs and mRNAs to alter their folding and/or facilitate duplex formation. In addition, Hfq can mediate the interaction of proteins and complexes (including RNase E, ribosomes, poly(A) polymerase I and polynucleotide phosphorylase) with sRNAs, mRNAs and/or duplexes [205, 206]. Hfq has been shown by electron microscopy to be present at the inner cell membrane, as well as in the nucleoid and cytoplasm [207]. Many of the proteins that bind to Hfq are also found in the cytoplasm and/or at the cell membrane [208-210].

It has yet to be determined where sRNAs localize to in the cell, which is a barrier to understanding their mechanism of action and the constraints on their activity. It is often assumed that most RNAs are small, and thus they can move anywhere in the cell. However in actuality, RNAs are usually large compared to the proteins they encode due to: (i) each RNA having three nucleotides for each amino acid encoded (in addition there are 5' and 3' untranslated RNA sequences); (ii) the average nucleotide is three times the mass of an amino acid (≈330 Daltons and ≈110 Daltons respectively) [211]; and (iii) RNAs often have less compact structures than globular proteins [211]. Therefore even relatively short sRNAs are large compared to some of the small proteins that act as transcription factors. For example, the diameters of the MicA and DsrA sRNAs are approximately 87.5 Å and 111.5 Å [212] whereas the typical globular protein has a diameter of 50 Å [211]. Therefore while transcription factors can move through the densely packed DNA of the nucleoid to bind near the promoters of target genes, it is possible that many sRNAs and mRNAs are not able to do so because of their larger size.

In general, factors other than size and structure can also affect RNA and protein localization, including: (i) the molecules they form complexes with, which can transport them [213, 214] or restrict them [215] to specific sites in the cell; (ii) covalent modifications [216, 217]; and (iii) net electrostatic charge and charge distribution [218, 219]. In bacteria, a variety of RNA localization patterns and mechanisms have been reported. It has long been recognized that the signal recognition particle (SRP) pathway is an important mechanism for RNA and protein localization. SRP recognizes signal sequences at the N-terminal end of nascent proteins, leading to the transport of a complex containing the partial mRNA, ribosome, and partly synthesized protein [220] to the cell membrane [note: it has also been proposed that these components may be transported separately [221]]. Once at the cell membrane, translation continues in conjunction with translocation of the protein across the cell membrane [220]. Recently, it has been demonstrated that mRNAs can also be transported to the cell membrane without being coupled to translation via a mechanism that has not been fully elucidated involving RNA zip codes [106]. Bacteria also have mechanisms to localize RNAs to other cellular regions including the cytoplasm [106, 222, 223], cell poles [106, 223], and septa of dividing cells [224]. Other studies have shown that some mRNAs primarily localize to their site of transcription [109] [note: it is unclear whether this transcription is taking place at the edge or the center of the nucleoid]. In summary, regulation of RNA localization is important to cells, there are diverse sites and complex patterns of localization, and multiple localization mechanisms, of which most are poorly understood.

Whether or not sRNAs can localize to the nucleoid has important implications for gene regulation. An inability of sRNAs to enter the nucleoid would prevent them binding the TIR on target mRNAs as soon as it is transcribed, and give ribosomes greater opportunity for assembling at the TIR and initiating translation. sRNAs would instead only be able to bind the mRNA after the transcription-translation complex has formed and moved to either the outer edge of the nucleoid or the membrane [96, 110]. At the edge of the nucleoid, transcription and translation occur where there is a high concentration of ribosomes [137, 225-228], which may make it more difficult for the sRNA and Hfq to compete for binding at the TIR. Localization has been reported for one sRNA-mRNA pair (SgrS-*ptsG* mRNA); and in this pair, translation of the transmembrane domain of the *ptsG* mRNA is required for SgrS to mediate degradation of this mRNA [110]. Therefore in this case, it appears the sRNA does not need to enter the nucleoid to mediate its actions. Because *ptsG* requires at least one round of translation for transport to the membrane, SgrS cannot completely silence PtsG production [110]. In theory, complete silencing of target protein production is achievable if sRNAs bind to target mRNAs soon after they are transcribed and before any translation is initiated. The 6S sRNA, which regulates gene transcription by binding to the RNA polymerase holoenzyme with the sigma70 factor [229], indicates that it is possible for at least some sRNAs to move through the nucleoid to sites where transcription is initiated.

Due to the fundamental roles of sRNAs in bacterial survival and pathogenesis, identification of their cellular localization will benefit many areas of basic and medical research. Knowledge of sRNA localization within cells will also aid the rational design and optimization of their use in engineered gene regulatory circuits. sRNAs are useful components in regulatory circuits because of their properties, including rapid signaling [230], programmable specificity [96, 231], and threshold-linear responses [230, 232]. sRNAs have been used as tools to investigate the properties of gene regulation [105, 185, 230, 232] and to construct circuits for metabolic engineering and “knock-down” studies [233, 234].

In this study we investigated whether sRNAs preferentially accumulate in the nucleoid, cytoplasm, or cell membrane using synthetic sRNA systems on plasmids. In the first part, we examined the localization of four sRNAs (GlmZ, OxyS, RyhB and SgrS), and the *gfp* mRNA encoding the green fluorescent protein (GFP) using RNA fluorescent *in-situ* hybridization (FISH). We evaluated localization by measuring the overlap of the sRNA signal with the nucleoid or with pixels at the cell membrane. We found that sRNAs localized in both the nucleoid and cytoplasm. In contrast, the *gfp* mRNA control showed less localization in the nucleoid than in the cytoplasm. Further examination of the localization of the sRNAs found that very few cells had membrane localization compared to a control mRNA (*bglF*), which was fused to *gfp* and was known to have membrane localization [106]. In the second part of the study, we engineered RNAs, with the *gfp* mRNA as the starting point, to determine whether we could alter nucleoid and cytoplasmic localization. We found that decreasing RNA length and decreasing translation increased nucleoid localization, and that these effects can be combined resulting in the same level of nucleoid localization as the sRNAs. Conversely, we demonstrated that increasing RNA length via fusion of native target mRNA sequences to the *gfp* mRNA, increased the preferential localization of RNAs in the cytoplasm. We also demonstrated that Hfq had no effect on sRNA localization in the cytoplasm and nucleoid. Together our results suggest that RNA size is an important factor, but not the only factor, in determining RNA localization, and that because of their small size, sRNAs can enter the nucleoid.

3.3 Results

3.3.1 sRNAs with equal probability in the nucleoid and cytoplasm

We sought to examine the localization of sRNAs in the nucleoid and cytoplasm. The genes for three silencing sRNAs (OxyS, RyhB, and SgrS) and two activating sRNAs (DsrA and GlmZ) were placed on plasmids in *Escherichia coli*. The advantages of having the sRNAs on plasmids are: (i) it is more common to synthetic circuits; (ii) it directly examines whether sRNAs can enter the dense structure of the nucleoid (whereas if sRNAs are transcribed from the chromosome it is unclear if their presence in the nucleoid is due to it being their site of production); and (iii) there are multiple copies of the genes which increases the sRNA concentrations thereby making it easier to detect them. Note: one study found no difference in the localization of mRNAs transcribed from a plasmid or from the chromosome [106]. We selected the *gfp* mRNA to compare with sRNA localization because it is a non-native mRNA (and therefore less likely to be subject to control mechanisms), a common reporter in synthetic biology, and GFP is readily quantified by fluorescence microscopy.

We measured the localization of the sRNAs by RNA fluorescent in-situ hybridization (RNA FISH) because it does not require any modification of the sRNA sequence or structure. Other studies have used RNA FISH to count the number of sRNAs in single cells of *Yersinia pseudotuberculosis* and *Yersinia pestis* [235] to characterize the search kinetics of the SgrS sRNA for *ptsG* [236], and to localize mRNAs [106]. Phase contrast microscopy was used to identify the cell boundary and the DNA stain 4',6-diamidino-2-phenylindole (DAPI) was used to identify the nucleoid.

In our first experiment we examined whether we could detect sRNAs in exponentially growing cells. We compared sRNA signal intensities between strains with and without transcription of the sRNA (the latter was performed in strains without promoters) using “global normalization” for the RNA signal heat maps (**Figure 3.1A**). Global normalization linearly scaled the signals using the highest pixel value in all of the strains (“high”) and the lowest pixel value in all of the strains (“low”) to set the range. The normalized signals in each pixel in individual cells were plotted as heat maps. Strong signal was observed with the sRNA probes only in the strains where GlmZ, OxyS, RyhB and SgrS were transcribed and not in control strains without the promoter (**Figure 3.1A, C**). These results indicate the probes only detect sRNAs and not their DNA sequences or endogenous RNAs. The signal for the transcribed DsrA was very low therefore no further experiments were performed with it (**Figure 3.2**). The gfp probes were also specific for when the *gfp* mRNA was transcribed (**Figure 3.1B, C**).

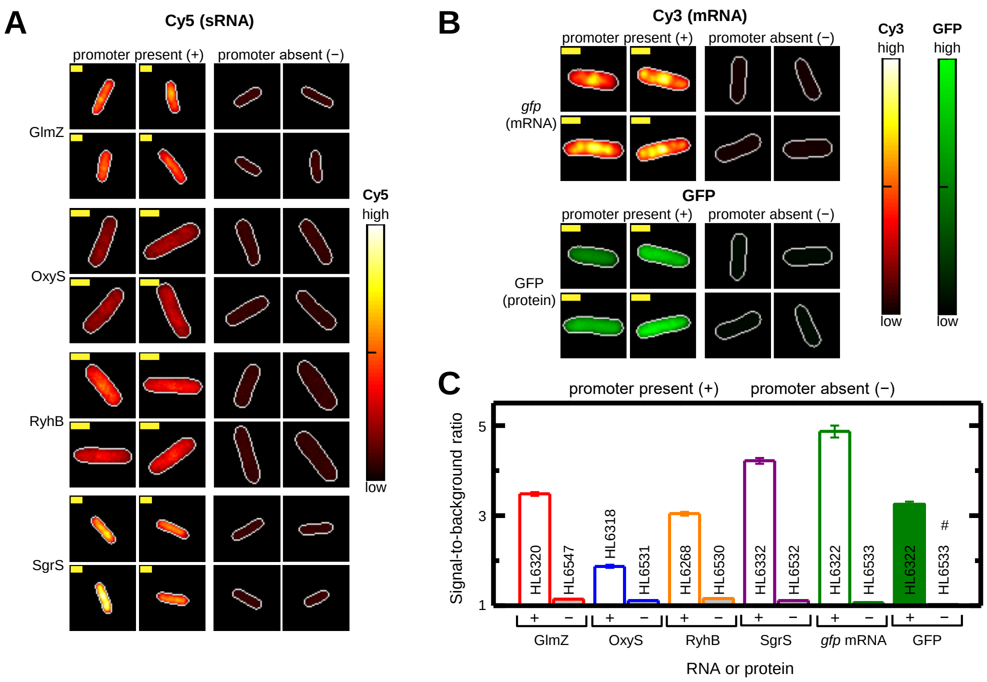


Figure 3.1 RNA FISH specifically detects sRNAs and mRNAs. (**A**, **B**) sRNA and mRNA signal intensities in representative cells with and without a promoter (PCon or PLlacO-1) in the Cy5 (**A**), Cy3 (**B**), or GFP (**B**) channels. Signal intensity in cells is shown as a heat map with "global normalization" (main text). Cell edges (white line) were identified by phase contrast and transferred to Cy5, Cy3 or GFP channels. Yellow scale bar is 1 µm for all images. Strains with promoter: *glmZ* (HL6320; *n* = 375); *oxyS* (HL6318; *n* = 105); *ryhB* (HL6268; *n* = 233); *sgrS* (HL6332; *n* = 499); *gfp* (HL6322; *n* = 133). Strains with no promoter: *glmZ* (HL6547; *n* = 428); *oxyS* (HL6531; *n* = 281); *ryhB* (HL6530; *n* = 199); *sgrS* (HL6532; *n* = 579); *gfp* (HL6533; *n* = 278). All *gfp* measurements were in the presence of 1 mM IPTG. (**C**) Signal-to-background ratios with and without transcription for each RNA or protein. Error bars are the SEM. # bar and error are approximately one and zero and therefore not visible. Statistical comparison of mean signal-to-background ratios between strains with and without a promoter was significant for all pairs (P values for all pairs < 1 × 10^{-58}; Mann-Whitney U two-tailed test and two-tailed t-test).

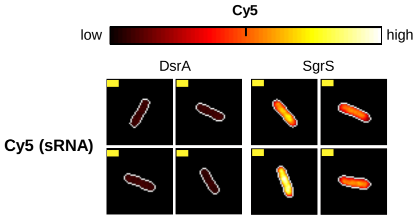


Figure 3.2 RNA FISH for DsrA and SgrS. Representative cells transcribing DsrA and SgrS with Cy5 labeled probes. White line is the cell boundary identified by phase contrast, which after alignment was transferred to the Cy5 channels. Yellow scale bar indicates 1 µm. The intensity of Cy5 signal in cells is shown as heat maps with "global normalization" (described in main text). Strains: *dsrA* (HL6269) with PCon promoter; *sgrS* with PCon promoter (HL6332; data from **Figure 3.1A** is reshown for comparison).

To visualize intracellular localization of sRNAs and mRNAs in individual cells we performed "cellular normalization" for the RNA signal heat maps rather than "global normalization". That is, we linearly scaled the pixel values in each cell using the highest ("high") and lowest ("low") pixel value for that particular cell (**Figure 3.3A**). Inspection of representative cells indicated that pixels with the highest intensity signal for each sRNA appear to occur in regions with DAPI (*i.e.*

in the nucleoid) as well as regions outside it (*i.e.* in the cytoplasm). The *gfp* mRNA displayed a very different pattern of localization with high signal predominantly in the cytoplasm (**Figure 3.3A**). The negative control strains for Cy3 and Cy5 (described in figure legend) had diffuse localization of signal as expected for background signal (**Figure 3.3B**). The signal-to-background ratio was <1.2 and <1.3 for Cy3 and Cy5 respectively in > 95% of the negative control cells; these values were used as cut-offs to identify cells in our experiments with no signal and therefore not included in the analyses.

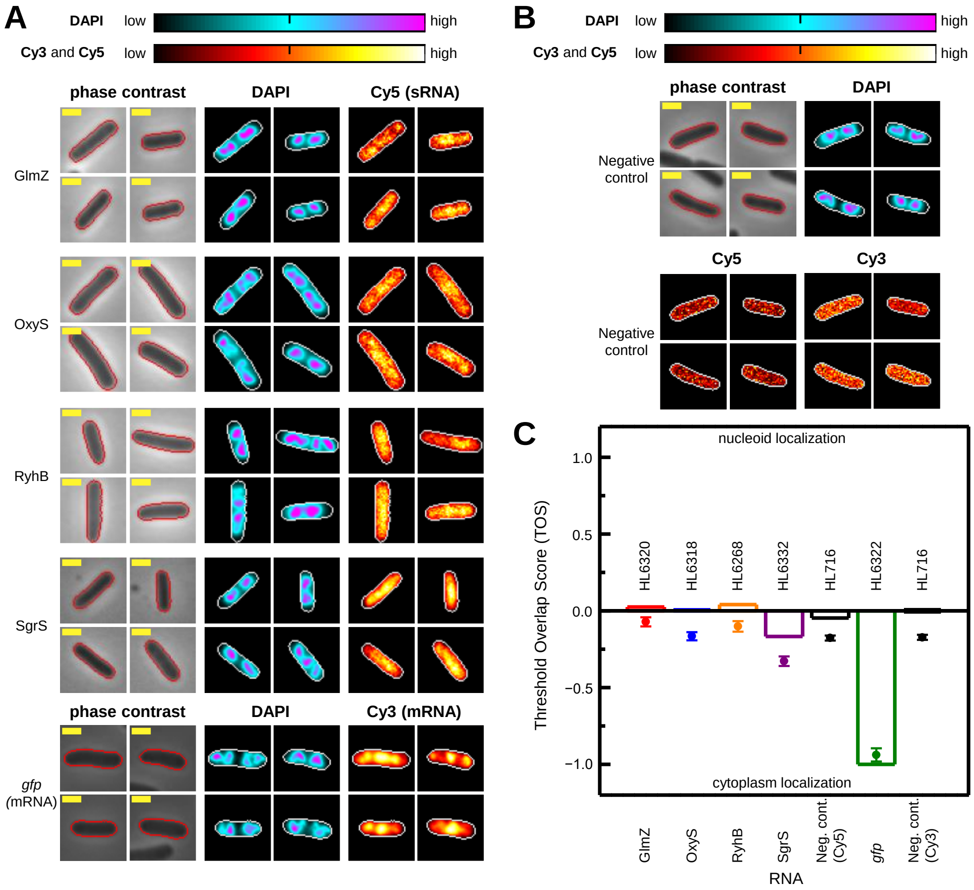


Figure 3.3 sRNAs occur with equal probability in the nucleoid and cytoplasm. (**A**) sRNA and mRNA localization in representative cells transcribing an sRNA or mRNA in phase contrast, DAPI, and Cy5 or Cy3 channels. Signal intensities within cells are shown as heat maps with "cellular normalization" (main text). Cell edges were identified by phase contrast (red line) and transferred to the DAPI, Cy5, and/or Cy3 channels (white line) following alignment. Yellow scale bar indicates 1 µm. Strains with promoter: *glmZ* (HL6320; *n* = 81); *oxyS* (HL6318; *n* = 126); *ryhB* (HL6268; *n* = 93); *sgrS* (HL6332; *n* = 209); *gfp* (HL6322; *n* = 34). Data for *gfp* are from the same experiment as in **Figure 3.1** but here they are shown with cellular normalization instead of global normalization (note: cells with saturated pixels in the DAPI channel were not included in this analysis). (**B**) Representative cells in the negative control (HL716; *n* = 355) without Cy3 or Cy5 probes in phase contrast, DAPI, Cy5 and Cy3 channels. The negative control is the host strain without any plasmid (HL716) and without probes. (**C**) Threshold overlap score (TOS) for each sRNA, the mRNA, and negative control. Bars are the medians, circle symbols are the means, and error bars are the SEMs. TOS is a normalized measure of the overlap of the top 10% of the DAPI and Cy5 (or Cy3) signals (see main text for further details). Median TOS of each sRNA to the *gfp* mRNA was significantly different (P = 2.0×10^{-14}, 6.0×10^{-16}; 9.0×10^{-15} and 2.5×10^{-11} for GlmZ, OxyS, RyhB and SgrS respectively; Mann-Whitney U two-tailed test).

The nucleoid does not have a distinct boundary but instead has parts that extend into the cytoplasm. Therefore to evaluate whether sRNAs can enter the nucleoid we need to focus on their localization in the center of the nucleoid where the DNA is densest and the DAPI signal is highest. Specifically, we needed to set a threshold to select the pixels with highest DAPI signal. However, we did not want to set the threshold too high so that there were too few pixels to evaluate the statistical significance of the measured overlap of the sRNA signal with the center of the nucleoid. In other words, the threshold needs to select a fraction of pixels with the highest intensity signal (F_T) that is neither too small nor too large. We performed a power calculation (**Materials and Methods**) and determined $F_T = 0.1$ to be the threshold for our experiments. That is, we set thresholds for the sRNA and DAPI signals (and also for the mRNA and GFP signals) so that the 10% of pixels with the highest intensities in each cell were selected for our analyses.

To quantify localization we first determined the observed overlap of our selected pixels for the sRNA signal (or *gfp* mRNA signal) with our selected pixels for the DAPI signal (*i.e.* the center of the nucleoid). This was calculated by counting the number of overlapping pixels and dividing it by the number of selected pixels for the DAPI channel. Because we selected the pixels with the highest signal, the effect of background signal was minimal. We then divided the observed overlap by the expected overlap for a uniform distribution of random signal intensities across the cell. We rescaled this ratio resulting in a threshold overlap score (TOS) with −1 and +1 as the minimum and maximum respectively [151]. That is,

$$\text{TOS}=\begin{cases} 0, & \text{when observed overlap = expected overlap,} \\ \dfrac{\text{observed overlap}}{\text{expected overlap}} -1, & \text{when observed overlap < expected overlap, and} \\ \dfrac{\text{observed overlap - expected overlap}}{\text{1 - expected overlap}}, & \text{when observed overlap > expected overlap.} \end{cases}$$

The observed and expected overlaps are both fractions with a value between 0 and 1; and in this case the expected overlap is 0.1. TOS > 0, ≈ 0, and < 0 indicate the sRNA or mRNA occurs in the nucleoid more, the same, or less than a signal that is uniformly distributed throughout the cell (*i.e.* colocalization, noncolocalization and anticolocalization with the nucleoid respectively). It is important to note that TOS is designed to evaluate the fractional overlap of the sRNA signal or mRNA signal with DAPI independent of the level of clustering of selected pixels (see summary below). Otherwise, any change in localization measured by the TOS metric would be due to an unknown combination of changes in signal overlap and/or clustering.

Our analysis revealed that three sRNAs (GlmZ, OxyS, and RyhB) had median TOS values of 0.02, 0.01, and 0.04 respectively (**Figure 3.3C**). That is, the top 10% of pixels with the highest intensity signal for each sRNA and the "center" of the nucleoid (where the top 10% of pixels with the highest intensity signal for DAPI occur) have TOS values that are close to zero. This indicates the sRNA signals overlap as much as would be expected for a signal that was uniformly distributed in the cell (*i.e.* noncolocalization). For SgrS, there was weak anticolocalization of the signal and the nucleoid (median TOS ≈ −0.17). In contrast to the sRNAs, *gfp* mRNA has strong anticolocalization with a median TOS = −1.00 (**Figure 3.3C**). Therefore there is essentially no overlap between the top 10% of pixels with highest intensity *gfp* mRNA signal and the top 10% of pixels with highest intensity DAPI signal. To better highlight this difference between sRNAs and mRNA localization we show only the 10% of pixels with highest signal intensity for the sRNA or mRNA (magenta color), and only the 10% of pixels with highest signal intensity for the DNA (*i.e.* DAPI; cyan color) in **Figure 3.4**. Selected pixels for the sRNAs or mRNAs that overlap the selected pixels for the DNA have a yellow color. **Figure 3.4** shows that relatively few of the selected pixels for the mRNAs overlap the selected pixels for the DNA compared to the sRNAs (note: this can be most clearly seen in "zoomed" views of the digital images).

In summary, sRNAs tend to occur in both the nucleoid and cytoplasm and the *gfp* mRNA occurs predominantly in the cytoplasm. The mechanistic basis for this difference is examined in later experiments. Our finding that the fractional overlap of the 10% of pixels with highest intensity signals for the sRNAs and the 10% of pixels with the highest intensity signal for the DNA, is approximately the same as expected by chance for a uniform distribution, does not necessarily mean the pixels for *both* signals are actually uniformly distributed throughout the cell. First, it is sufficient for the pixels of only one of the signals to be uniformly distributed. Second, TOS evaluates overlap independent of the level of clustering of the selected pixels. In the second case, the selected pixels for the RNA signal and the DAPI signal may be clustered but if these clusters are randomly distributed in the cell then the overlap may be same as expected for a uniformly distributed signal.

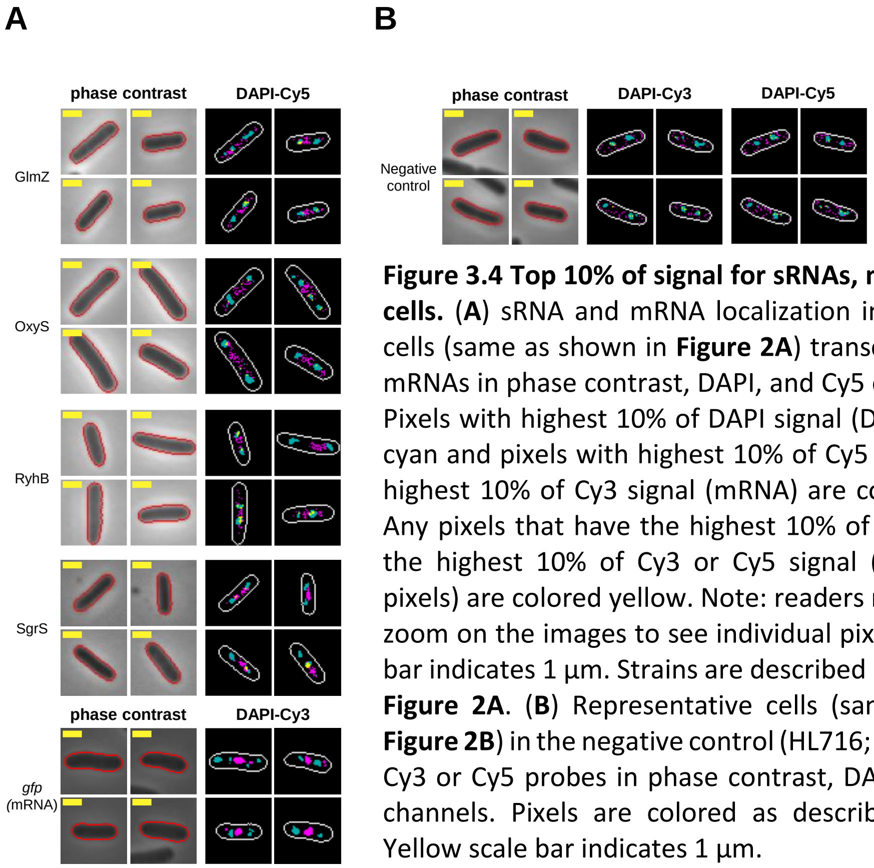


Figure 3.4 Top 10% of signal for sRNAs, mRNA or DNA in cells. (**A**) sRNA and mRNA localization in representative cells (same as shown in **Figure 2A**) transcribing sRNAs or mRNAs in phase contrast, DAPI, and Cy5 or Cy3 channels. Pixels with highest 10% of DAPI signal (DNA) are colored cyan and pixels with highest 10% of Cy5 signal (sRNA) or highest 10% of Cy3 signal (mRNA) are colored magenta. Any pixels that have the highest 10% of DAPI signal and the highest 10% of Cy3 or Cy5 signal (*i.e.* overlapping pixels) are colored yellow. Note: readers may need to use zoom on the images to see individual pixels. Yellow scale bar indicates 1 µm. Strains are described in the legend for **Figure 2A**. (**B**) Representative cells (same as shown in **Figure 2B**) in the negative control (HL716; *n* = 355) without Cy3 or Cy5 probes in phase contrast, DAPI, Cy3 and Cy5 channels. Pixels are colored as described in **panel A**. Yellow scale bar indicates 1 µm.

3.3.2 sRNAs display no preferential membrane localization

We analyzed images from the above experiments to determine if there was increased localization of sRNAs at the cell membrane. We included the *gfp* mRNA as a negative control because it does not localize at the membrane [222, 237]. In addition, we created a positive control by fusing the *bglF* mRNA, which encodes β-glucoside phosphotransferase permease (BglF), to the *gfp* mRNA. BglF has eight domains that span the inner membrane [238], and the *bglF* mRNA has previously been shown to localize to the membrane by a translation independent mechanism [106]. RNA FISH was performed as above and GFP was measured by fluorescence microscopy (**Materials and Methods**).

The BglF::GFP protein was observed in regions near the membrane as expected (**Figure 3.5A**). The *bglF*::*gfp* mRNA was not obviously at the membrane from visual inspection of cell images but it was detectable by quantitative analysis. Quantitative analysis was performed by identifying the cell boundaries in the phase contrast images and removing the outermost layer of pixels using the "erode" function in ImageJ. The outermost layer of pixels were removed because they include areas outside the cell and have less signal from the point spread function of neighboring pixels. Together these factors create an "edge effect" with lower signal in the outermost layer of pixels (**Figure 3.5B**). The next outermost layer was termed the "membrane" and we determined by TOS whether the top 10% of intensity values in the whole cell overlap with this membrane layer more, less, or the same as a uniform distribution. We created histograms of this "membrane TOS" obtained from each cell and found that the mean and median values often did not capture differences between samples because of heterogeneity in the cell populations (**Figure 3.6**). Therefore we measured the fraction of cells in each sample that had colocalization of the sRNA or mRNA signal with the membrane region (*i.e.* membrane TOS > 0) (**Figure 3.5C**).

BglF::GFP and *bglF*::*gfp* mRNA had 93.8% and 57.1% of cells with membrane TOS > 0 respectively (**Figure 3.5C**). That is, the 10% of pixels with the highest BglF::GFP and *bglF*::*gfp* mRNA signal in the cell overlapped with the membrane region more than expected by random chance in the majority of cells. These results in the positive controls are consistent with membrane localization. In contrast, all the sRNAs and the *gfp* mRNA had very low percentages of cells with TOS > 0 (**Figure 3.5C**), and these percentages were significantly lower than the *bglF::gfp* mRNA positive control (P values < 1×10^{-6}; Fisher's exact test). The control with uniform randomly distributed background signal (HL716) is not affected by the edge effect and was expected to have a 50:50 random split of cells with TOS > 0, and this was observed (**Figure 3.5C**). In summary, none of the sRNAs displayed evidence of membrane localization.

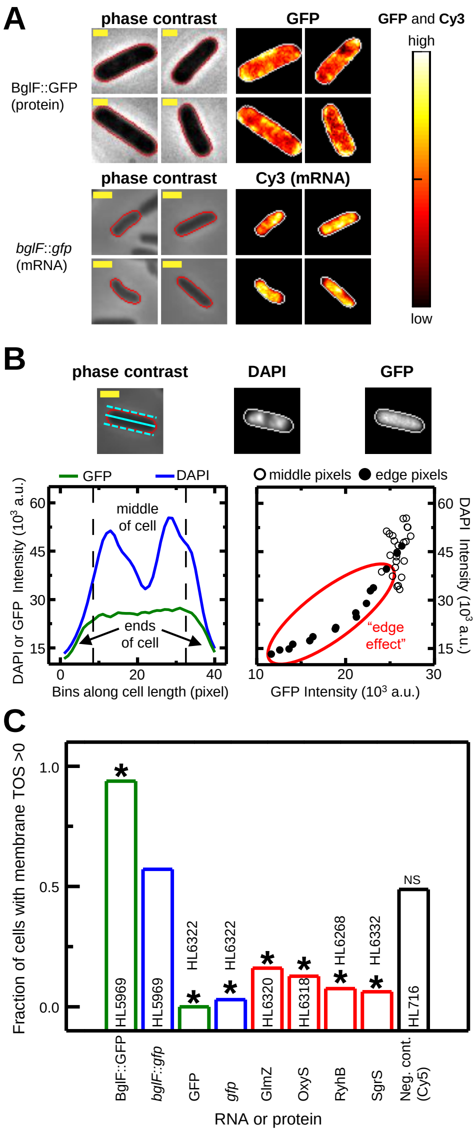


Figure 3.5 sRNAs display no preferential membrane localization. Cell edges were identified as in **Figure 3.3A**. Yellow scale bar indicates 1 µm. (**A**) mRNA and protein localization of *bglF*::*gfp* (HL5969) in representative cells shown for the phase contrast, GFP (n = 304) or Cy3 (n = 84) channels. Signal intensities of Cy3 and GFP are shown as heat maps with cellular normalization. Measurements were made at 1 mM IPTG. (**B**) "Edge effect" shown in a cell with GFP expression (HL6322). Average GFP fluorescence and DAPI signals along the longitudinal axis (upper cell images); both signals decrease at the cell ends (left lower plot) resulting in a positive correlation for binned pixel values (right lower plot; circled with a red line). Solid and dash cyan lines in the phase contrast image show the center and edges of the longitudinal axis. (**C**) Fraction of cells with membrane localization ("membrane TOS" > 0) in strains probed for sRNAs and mRNAs, or expressing GFP. Analysis was performed on measurements collected for **Figure 3.3** (HL6322, HL6320, HL6318, HL6268, and HL716). The number of cells with TOS > 0 was compared between the positive control (*bglF*::*gfp*) and each of the sRNAs and mRNAs using Fisher's exact test. * indicates statistical significance ($P < 1 \times 10^{-6}$). NS indicates no significance ($P = 1.8 \times 10^{-1}$) for the negative control (Cy5).

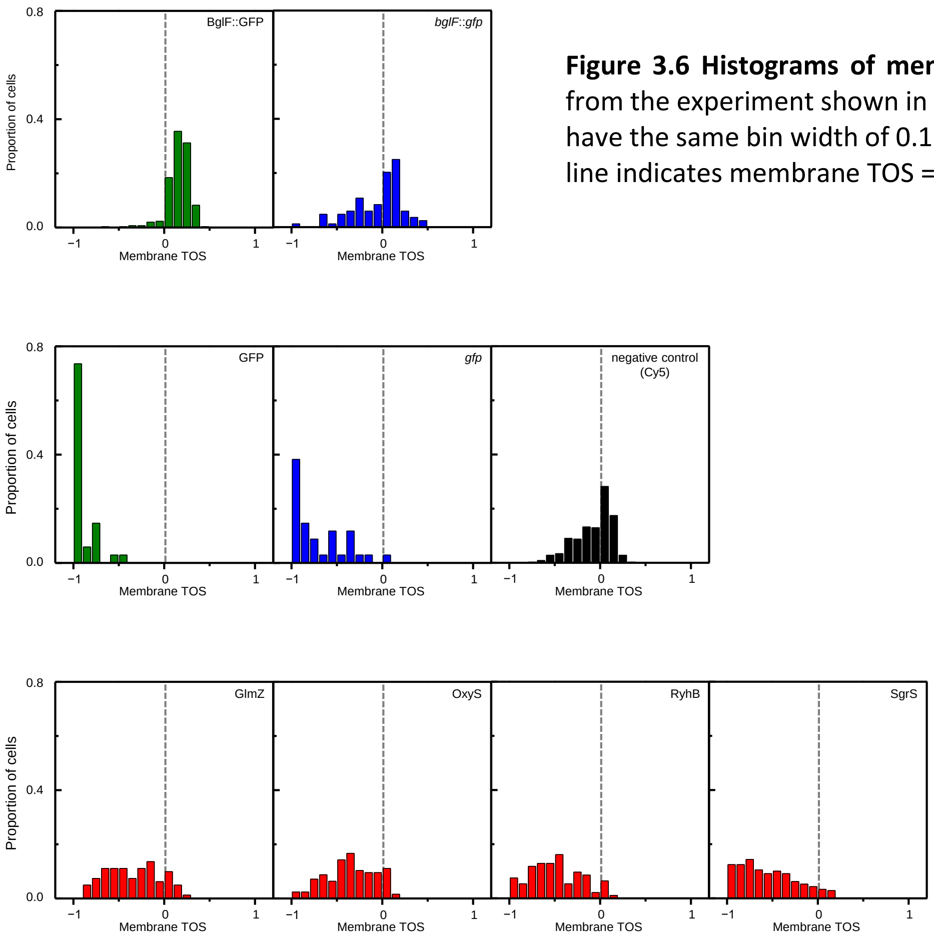


Figure 3.6 Histograms of membrane TOS. The data is from the experiment shown in **Figure 3.5**. All histograms have the same bin width of 0.1 TOS (unitless). Grey dash line indicates membrane TOS = 0.

3.3.3 RNA length and translation affect nucleoid localization

We attempted to engineer RNAs with the same level of nucleoid localization as sRNAs. Creating nucleoid localization for the full length *gfp* mRNA was judged to be more likely to be informative than simply disrupting nucleoid localization of sRNAs. Moreover, sRNAs have very important structure-function relationships that when altered could have unexpected and unexplainable effects. The first factor that was considered was RNA length. RyhB, OxyS, GlmZ and SgrS have lengths of 102, 121, 207, and 238 nucleotides respectively whereas the *gfp* mRNA has a total length of 994 nucleotides (717 nucleotides for the coding sequence and ~277 nucleotides for the 5’ untranslated region and T1 terminator). Therefore the *gfp* mRNA is expected to have a larger size resulting in more difficulty diffusing through the compact chromosomal DNA of the nucleoid. The second factor that was considered were polysomes, which are complexes comprised of an mRNA, 70S ribosomes, the translated peptide and other factors [239]. The presence of polysomes increases mRNA size and therefore could impede mRNA movement through the nucleoid.

To evaluate the effects of RNA length and polysome formation on localization we compared the full length *gfp* mRNA to a partial length mRNA, and both lengths with and without translation.

Translation was prevented by deleting the ribosome binding sequence (RBS) and the start codon, which for the full length mRNA abolished GFP fluorescence. To summarize, there were four sets of samples in this experiment (**Figure 3.7A**): (i) full length *gfp* mRNA with translation (≈ 994 nucleotides); (ii) full length *gfp* mRNA without translation (≈ 976 nucleotides); (iii) first quarter of the *gfp* mRNA with an introduced stop codon and translation (≈ 460 nucleotides); and (iv) first quarter of the *gfp* mRNA with an introduced stop codon and no translation (≈ 442 nucleotides). Localization of these mRNAs was measured by RNA FISH using the Cy3 probes for the *gfp* sequence (representative cells in **Figure 3.7B**). TOS was calculated as described above to measure the overlap of pixels with the top 10% of the mRNA signal and pixels with the top 10% of the DAPI signal (which are primarily in the center of the nucleoid). We found that in most samples, the median and mean TOS were not the same due to a long-tailed distribution. This distribution required a rank-order test for statistical significance (Mann-Whitney U two-tailed test) and therefore we primarily compared the median TOS between samples.

We examined the effect of translation on localization by comparing median TOS for full length mRNA with and without the RBS (−1.00 and −0.70 respectively) (**Figure 3.7C**). These differences were statistically significant (P = 2.6 × 10^{-4}; total *n* = 299). Therefore decreasing translation to reduce the number of polysomes along the mRNA increased the overlap of pixels with the highest (*i.e.* top 10%) *gfp* mRNA and DAPI signals; that is, polysomes appear to prevent nucleoid localization of the mRNA. We next examined the effect of mRNA length on localization by comparing median TOS for the full length *gfp* without the RBS and the first quarter of *gfp* without the RBS, which were −0.70 and 0.05 respectively (**Figure 3.7C**). The difference in median TOS was statistically significant (P = 1.4 × 10^{-27}; total *n* = 395) indicating that decreasing mRNA length enabled greater localization of the top 10% of pixels with the *gfp* mRNA signal in the nucleoid. We did not compare full length *gfp* with the RBS to the quarter length *gfp* with the RBS because altering mRNA length can also potentially decrease the number of polysomes along the mRNA; therefore observed differences may reflect the combined effects of length and translation. It is notable that the partial length *gfp* mRNA without an RBS has a median and mean TOS ≈ 0; that is, it has the same fractional overlap with the nucleoid as the sRNAs.

The partial length *gfp* mRNAs with and without an RBS had similar median TOS (0.01 and 0.05 respectively) but large differences in mean TOS (−0.23 ± 0.03 and −0.07 ± 0.02 respectively) (**Figure 3.7C**). This difference in the means was due to the *gfp* mRNA without the RBS having more cells in the tail of the distribution with greater localization of mRNA in the cytoplasm. The data indicate the effects of shortening RNA length and decreasing translation can be combined, which is expected if polysomes and length contribute at least partly independently to RNA size, and if RNA size affects nucleoid localization.

We replaced the first quarter of the *gfp* coding sequence without the RBS and start codon with the sequence from the last quarter of the *gfp* coding sequence and repeated the experiments and analysis (representative cells in **Figure 3.7D**). Despite the different sequences, these partial length *gfp* mRNAs had very similar median and mean TOS (**Figure 3.7E**), which suggests the degree of nucleoid localization is primarily determined by length and not sequence.

To determine how RNA lengths relate to their size, we plotted the radius of gyration (R_g) for sRNAs, ribozymes, transfer RNAs, ribosomal RNAs and mRNAs (**Figure 3.7F** and **Table 3.6**). The radius of gyration is a way of describing the distribution of mass of an RNA or protein around its axes of rotation. If the shape of a RNA or protein is approximated by a solid sphere then the diameter is roughly equal to the radius of gyration multiplied by 2√(5/3) [240]. The values were obtained from the literature or by searching the Nucleic Acid Database Project for bacterial RNA structures with > 30 nucleotides and without any protein binding [241]. We fitted the measurements to the power law relationship that exists between R_g and the number of bond segments for a polymer (N). Specifically, we used the function $R_g = a \cdot N^{v}$, where N is the number of nucleotides, *a* is a pre-factor, and *v* is an exponent that specifies the compactness of the RNA in a solvent [242-246].

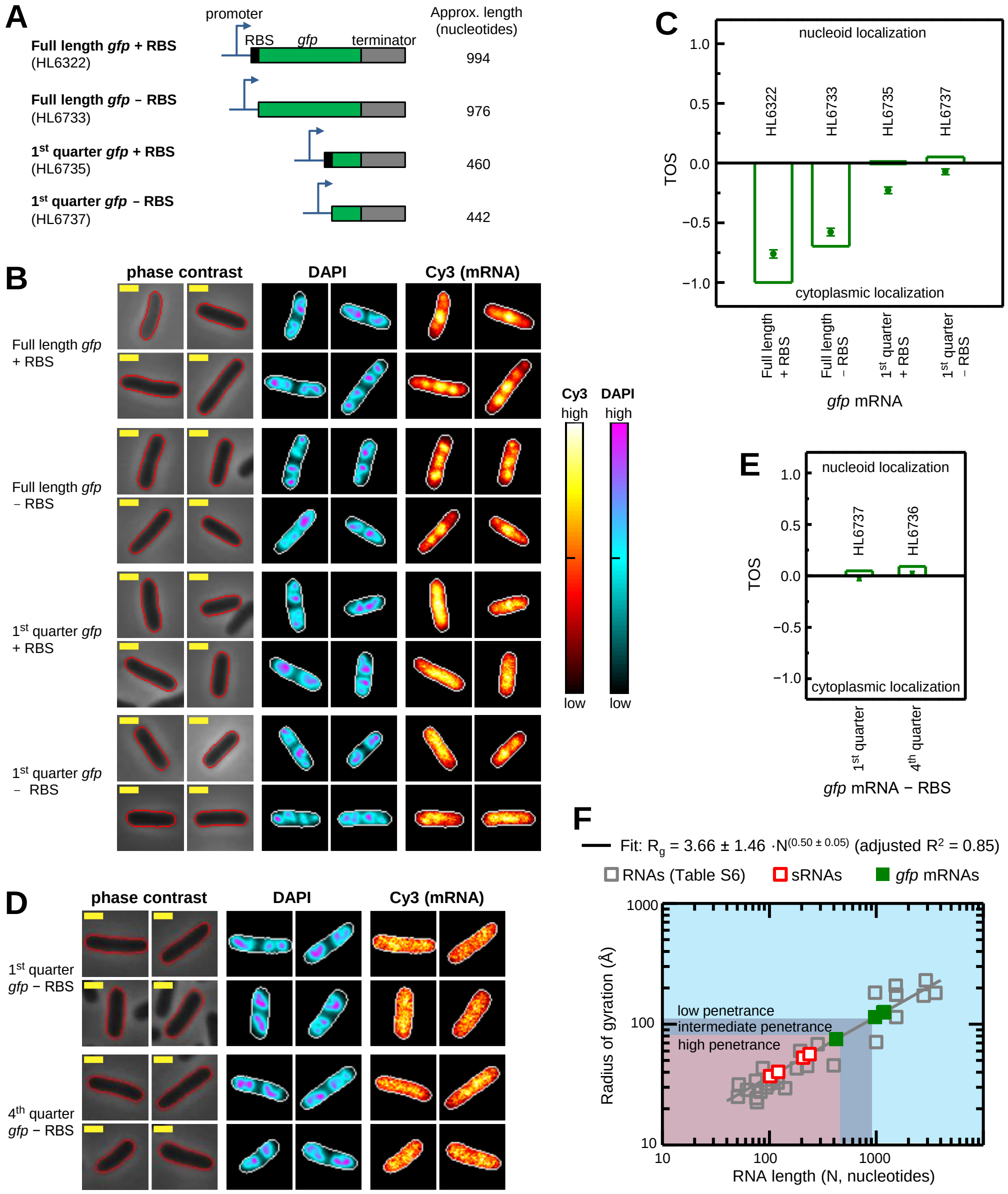


Figure 3.7 RNA length and translation affect nucleoid localization. Cell edges were identified as in **Figure 3.3A**. Yellow scale bar indicates 1 µm for all images. Measurements were made at 1 mM IPTG. (**A**) Full and partial length *gfp* mRNA with and without the RBS (st7) and start codon. (**B**) Localization of *gfp* mRNA in representative cells with each of the genes in **panel A**. DAPI and Cy3 signal intensities (cellular normalization) represented as heat maps. Sample sizes: HL6322 (*n* = 113), HL6733 (*n* = 186), HL6735 (*n* = 229), and HL6737 (*n* = 209). (**C**) TOS for strains with each of the genes shown in **panel A**. Bars are the medians, circle symbols are the means, and error bars are the SEMs. Median TOS for the following pairwise combinations were statistically significant (Mann-Whitney U two-tailed test): full length *gfp* mRNA ± RBS (P = 2.6 × 10^{-4}), 1st quarter *gfp* mRNA ± RBS (P = 3.8 × 10^{-4}), and full length *gfp* mRNA – RBS versus 1st quarter *gfp* mRNA – RBS (P = 1.4 × 10^{-27}). (**D**) Comparison of *gfp* mRNA localization in cells with 1st quarter *gfp* mRNA – RBS (HL6737; *n* = 231) and 4th quarter *gfp* mRNA – RBS (HL6736; *n* = 205). (**E**) TOS for strains with the genes in **panel D**. Plot is presented as in **panel C**. The difference in median TOS was small and barely significant (P = 1.2 × 10^{-2}, Mann-Whitney U two-tailed test). (**F**) Radius of gyration (R_g) as a function of RNA length. R_g values from the literature, which are provided in **Table 3.6**, were fitted to a power function as defined in main text (grey line). Parameter values from the fit were then used to calculate R_g for the sRNA and *gfp* mRNAs without RBS. Two target mRNA::gfp mRNA fusions from **Figure 3.5** are included in the plot for comparison. Because these fusions (*rpoS::gfp* and *fhlA::gfp*) have similar lengths their symbols overlap (see **Figure 3.8** and main text for more details). Parameter errors are the standard deviations. Shading shows RNA size ranges that may have potentially high, intermediate and low nucleoid penetrance.

Our fit using the Levenberg-Marquardt algorithm yielded *v* = 0.50 ± 0.05 and *a* = 3.66 ± 1.46 Å (adjusted R^2 = 0.85; *n* = 28). The exponent (*v*) is consistent with an ideal polymer chain with a simple random walk in a θ solvent [242]. Studies of relatively short RNAs, which tend to be tRNAs, riboswitches and ribozymes with high levels of self-annealing and more compact structures, often have exponents ≈1/3 to ≈2/5 [244, 247]. We reanalyzed the data for only the tRNAs, riboswitches and ribozymes (51-400 nucleotides), and the fit yielded an exponent *v* = 0.36 ± 0.10 (*a* = 6.18 ± 3.25 Å; adjusted R^2 = 0.53; *n* = 13), which was similar to that reported in the other studies [244, 247]. Conversely, our fit to experimentally measured sRNAs, mRNAs, and random sequences (75 to 1523 nucleotides) yielded a high exponent value indicating self-avoiding interactions and a larger volume (*v* = 0.60 ± 0.11 and *a* = 2.68 ± 2.04 Å; adjusted R^2 = 0.98; *n* = 5). Note: all errors for the fits are the standard deviations.

We estimated the R_g for our sRNAs using the parameters from the first fit because it had the lowest relative errors and was the most general fit. The calculated R_g for the sRNAs are: 37.0 Å (RyhB, 102 nucleotides); 40.3 Å (OxyS, 121 nucleotides), 52.7 Å (GlmZ, 207 nucleotides), and 56.5 Å (SgrS, 238 nucleotides). The predicted R_g for the partial length *gfp* mRNA without the RBS is 75.2 Å. This value is only slightly larger than for the sRNAs and similar to the 30S and 50S ribosomes which are approximately 70-80 Å [248-250] and can enter the nucleoid [226].

The predicted R_g for the full length *gfp* mRNA without the RBS is 114.3 Å, which is approximately the same radius as the 70S ribosome [251, 252]. That is, the R_g for the full length *gfp* mRNA, which does not enter the nucleoid, is 1.5-fold larger than the R_g of the partial length *gfp* mRNA that does enter the nucleoid. While this fold difference is relatively small, the absolute difference in diameter is large (≈ 100 Å) (note: diameters of the partial and full length *gfp* mRNAs are 194.1 Å and 295.2 Å respectively, assuming they are spherical).

We also measured the localization of mRNAs that are longer than the full length *gfp* mRNA. These mRNAs had the non-translated region and partial coding sequence of two native mRNAs, *rpoS* and *fhlA*, translationally fused to *gfp* (**Figure 3.8A**). These native mRNAs are known targets for sRNA regulation, and both fusions were previously described and shown to have relatively low translation, particularly for *fhlA::gfp* [185]. Because of this low level of translation, it was more appropriate to compare the localization of these target mRNA::*gfp* fusions to the localization of the full length, non-fusion *gfp* mRNA without the RBS rather than with the RBS. The lengths of the *rpoS::gfp* and *fhlA::gfp* fusion mRNAs were ≈1161 and ≈1185 nucleotides respectively. We measured localization by the same method as used for the other *gfp* mRNAs, and determined the mean TOS to be −0.79 ± 0.03 and −0.71 ± 0.04 for *rpoS::gfp* and *fhlA::gfp* respectively, and the median TOS to be −1.00 for both mRNAs (**Figure 3.8B, C**). The *rpoS::gfp* (≈1161 nucleotides) and *fhlA::gfp* (≈1185 nucleotides) fusion mRNAs, which are longer than the full length *gfp* mRNAs and have even less overlap with the nucleoid, have predicted R_g of 124.7 Å and 126.0 Å respectively, and diameters of 322.0 Å and 325.3 Å respectively (**Figure 3.7F**). These results show that mRNAs longer than the full length *gfp* mRNA without the RBS have even greater localization in the cytoplasm, consistent with their longer length further reducing nucleoid localization.

Together the data indicate that RNAs can be engineered to increase or decrease their localization by altering their length and/or level of translation to increase or decrease their size. For RNAs to have the same level of nucleoid localization as sRNAs, they appear to require an R_g $< \approx 80$ Å or diameter $< \approx 200$ Å, with all other factors being equal. Larger RNAs appear to have more difficulty moving into and through the nucleoid, and thus tend to localize outside the nucleoid.

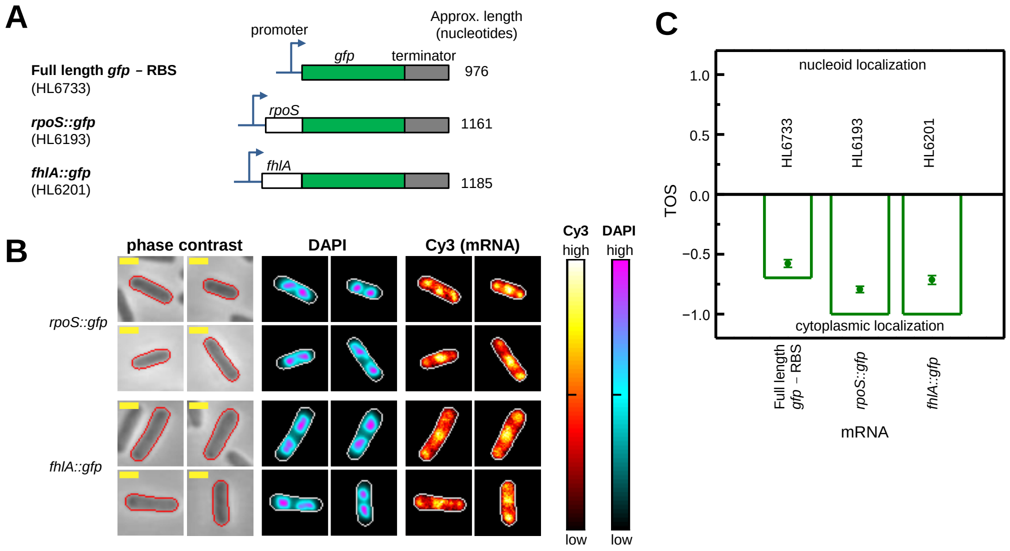


Figure 3.8 Localization of target mRNAs fused to *gfp*. Cell edges were identified as in **Figure 3.3A**. Yellow scale bar indicates 1 µm for all images. Measurements were made at 1 mM IPTG. Note: HL6733 data are the same as that shown in **Figure 3.7** and are reshown to enable convenient comparison. (**A**) Full length *gfp* mRNA with and without additional untranslated and coding sequences. (**B**) Localization of *gfp* mRNA in representative cells with each of the genes in **panel A**. DAPI and Cy3 signal intensities (cellular normalization) represented as heat maps. Sample sizes: HL6733 (*n* = 186), HL6193 (*n* = 177), and HL6201 (*n* = 98). (**C**) TOS for strains with each of the genes shown in **panel A**. Bars are the medians, circle symbols are the means, and error bars are the SEMs. Differences in the TOS values for the following pairwise combinations of samples were statistically significant (Mann-Whitney U two-tailed test): full length *gfp* mRNA – RBS versus *rpoS::gfp* mRNA (P = 2.3 × 10^{-7}) and full length *gfp* mRNA – RBS versus *fhlA::gfp* mRNA (P = 1.2 × 10^{-2}).

3.3.4 Hfq has minimal effect on sRNA localization

Given the prominent role of Hfq in regulating sRNA activity in *E. coli* it is important to establish whether Hfq affects sRNA localization. Hfq could potentially affect sRNA localization in several ways. The Hfq hexamer has a diameter of 62-65 Å [239, 253] therefore its binding to sRNAs could potentially add to their size and limit their movement in the nucleoid to regions with the densest DNA. Alternatively, the binding of Hfq to sRNAs may decrease their size as occurs with the *rpoS* target mRNA, which has a smaller size in the Hfq::*rpoS* complex (R_g = 58.0 ± 1.0 Å) than alone (R_g = 68.1 ± 1.6 Å) [254]. In addition, Hfq can bind to DNA [255] therefore it could potentially sequester sRNAs in the nucleoid.

We examined sRNA localization in strains without Hfq (Δ*hfq*) by RNA FISH (**Figure 3.9A**). These measurements were performed in parallel in strains with Hfq (**Figure 3.3**). We calculated TOS and found little or no difference in nucleoid localization with and without Hfq (**Figure 3.9B**). The exception was SgrS, which lost its slight preference for the cytoplasm with the deletion of *hfq* resulting in equal preference for the cytoplasm and the nucleoid (median TOS = − 0.17 and 0.02 respectively; P = 5.23 × 10^{-4}; total *n* = 535). The deletion of *hfq* prevents SgrS from forming duplexes with its target mRNA (*ptsG*) [256], therefore the latter finding of decreased anticolocalization with the nucleoid is probably due to less SgrS binding to its target mRNA outside the nucleoid at the cell poles [236]. It should be noted that under our growth conditions it has been established that SgrS translation does not occur [256, 257].

The minimal effects of Hfq on sRNA localization are unlikely to be due to the sRNAs being in such excess that there is insufficient Hfq to bind most of the sRNA molecules. One reason it is unlikely that there is a large pool of unbound sRNAs, is that many studies have shown that sRNAs are rapidly degraded in the absence of Hfq [see references and data in [185, 230]]. Furthermore, Hfq is clearly interacting with at least GlmZ, RyhB, and SgrS because the deletion of *hfq* altered their mean signals (and thus their concentrations) (**Figure 3.9C**). The halving of the GlmZ and SgrS concentrations (mean signal above background) when *hfq* was deleted suggests that at least 50% of the GlmZ and SgrS sRNAs are bound to Hfq (see calculations and model in the **Materials and Methods**). If Hfq was only binding to a small fraction of these sRNAs then the deletion of *hfq* should have had a correspondingly small effect on their concentrations.

There are many reasons OxyS and RyhB may have shown no difference in concentration and an increase in concentration with the deletion of *hfq*. These include similar degradation rates for the unbound and Hfq bound forms in our experiments, and decreased duplex formation in the *hfq* deletion strains causing increases in the sRNA concentrations by amounts that offset (OxyS), or more than offset (RyhB), the decreases in sRNA concentrations caused by the loss of Hfq protection of the sRNAs. We tested duplex formation of RyhB with a fusion of its target sequence (*sodB*) to *gfp* and found that SodB::GFP levels were increased in the Δ*hfq* strain (signal-to-background ratio in wild-type and Δ*hfq* strains were 3.28 ± 0.18 and 6.54 ± 0.46; HL6284 and HL6285). Therefore RyhB activity was decreased in the Δ*hfq* strain, and consequently Hfq concentrations in the wild-type background are sufficiently high to be a major contributor to duplex formation.

Together our experiments indicate the similarity in sRNA localization in the wild-type and Δ*hfq* strains is not explained by insufficient Hfq, and is most likely due to Hfq not having much effect on sRNA localization.

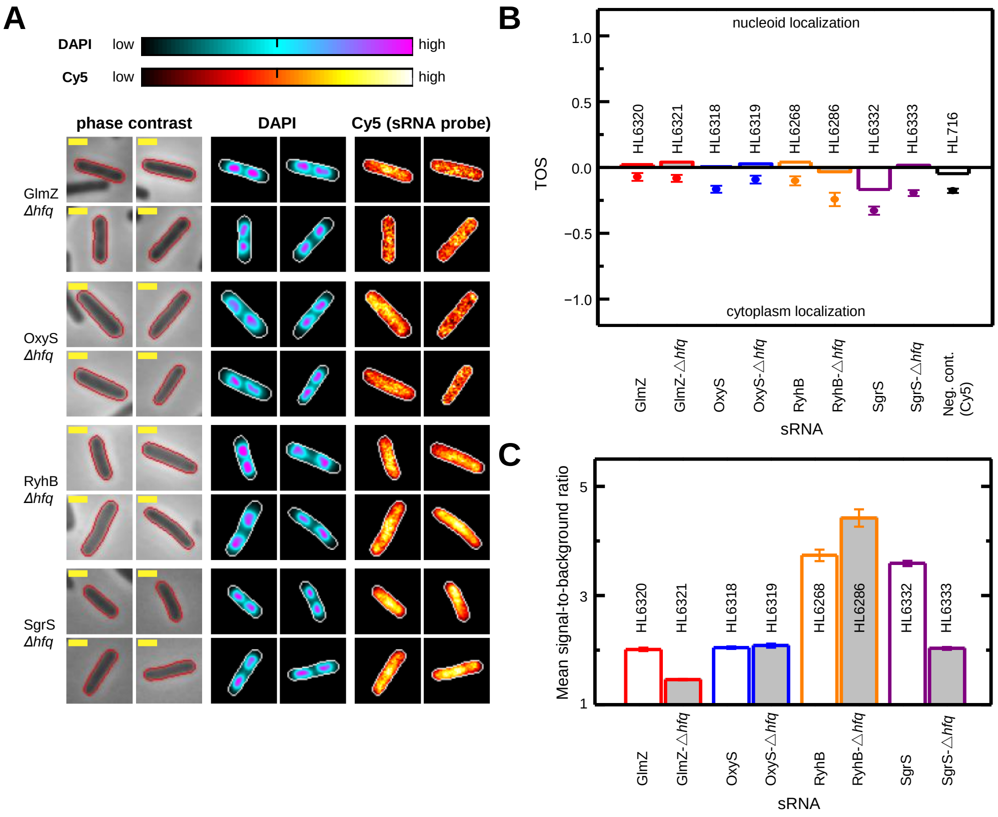


Figure 3.9 Hfq has minimal effect on sRNA localization. Cell edges were identified as in **Figure 3.3A**. Yellow scale bar indicates 1 µm for all images. Measurements for the negative control strain (HL716) without any plasmid and probes were made at 1 mM IPTG. (**A**) sRNA localization in strains without *hfq* (△*hfq*). Signal intensities of DAPI and Cy5 in individual cells are shown as heat maps with cellular normalization. All strains have the sRNA transcribed from the PCon promoter. Strains: *glmZ* (HL6321; *n* = 144); *oxyS* (HL6319; *n* = 97); *ryhB* (HL6286; *n* = 62); and *sgrS* (HL6333; *n* = 326). (**B**) TOS for each sRNA with *hfq* (data from the experiments in **Figure 3.3A**) and without *hfq* (**panel A**). Bars are the medians, circle symbols are the means, and error bars are the SEMs. Median TOS for pairs of strains with or without *hfq* were very similar for GlmZ, OxyS, and RyhB (P = 4.4 × 10^{-1}, 1.3 × 10^{-2}, and 1.7 × 10^{-2} respectively; Mann-Whitney U two-tailed test). Median TOS was significantly different for SgrS with or without *hfq* (P = 5.2 × 10^{-4}). (**C**) Signal-to-background ratios for each sRNA with and without *hfq*. Error bars are the SEMs. Mean signal-to-background ratio in strains with and without *hfq* were statistically significant for GlmZ, RyhB, and SgrS but not for OxyS (P = 1.4 × 10^{-50}, 2.6 × 10^{-4}, 4.0 × 10^{-125} and 3.7 × 10^{-1} respectively; two-tailed t-test).

3.4 Discussion

In this study we showed that sRNAs occur throughout the nucleoid and cytoplasm. The four sRNAs that were measured are diverse: (i) they ranged in size from 102 to 238 nucleotides; (ii) one (GlmZ) acts to increase target protein production and three (OxyS, RyhB, and SgrS) act to decrease target protein production; and (iii) they are involved in regulating different classes of proteins in different pathways including iron storage (RyhB), oxidative stress (OxyS), and carbohydrate metabolism (GlmZ and SgrS) [194]. Given the variety of the sRNAs studied, the findings of this study are likely to be general. In contrast to the sRNAs, the full length *gfp* mRNA almost exclusively occurs in the cytoplasm, which is consistent with nucleoid exclusion and the observation in another study that diffusion of *gfp* mRNA appeared to avoid the nucleoid [222]. We hypothesized that the difference between the sRNAs and the *gfp* mRNA was due to the larger size of the latter because of its greater length and polysomes. Decreasing each of these factors reduced nucleoid exclusion of the full length *gfp* mRNA, and decreasing both of these factors completely eliminated the nucleoid exclusion. Together the results indicate that sRNAs are able to enter the nucleoid due to their smaller size, and our observation that there is no preferential localization in the cytoplasm or nucleoid suggests that sRNAs probably move into and out of the nucleoid at similar rates (see below).

We also found that the deletion of *hfq* had minimal effect on the localization of sRNAs. This result includes sRNAs that bound to Hfq in our experimental systems in sufficient amounts that the deletion of *hfq* affected their concentrations and/or activities. To explain our findings we consider three plausible scenarios that take into account Hfq stabilizes sRNAs [185, 194, 204]. In scenario 1, Hfq binds sRNAs in the nucleoid or the cytoplasm, and unbound sRNA movement into and out of the nucleoid is slow or limited. In this scenario we would expect the concentration to decrease at the site of Hfq binding resulting in decreased or increased nucleoid localization, which is not consistent with our observations. In scenario 2, Hfq binds sRNAs in the nucleoid or the cytoplasm, and unbound sRNA movement into and out of the nucleoid is fast and unlimited. In this scenario we would expect the deletion of *hfq* to decrease the total cellular sRNA concentration but there would be minimal effect on localization because of rapid movement of sRNA. In scenario 3, Hfq binds sRNAs in both the cytoplasm and nucleoid. For the sRNAs to localize with equal probability in the nucleoid and cytoplasm as observed, Hfq must also localize with equal probability in the nucleoid and cytoplasm (or less likely, there is a difference in Hfq activity in the nucleoid and cytoplasm that is exactly counterbalanced by a difference in Hfq concentration in the nucleoid and cytoplasm so the sRNA localization appears to occur with equal probability in the nucleoid and cytoplasm). In this scenario, deletion of *hfq* decreases the concentration of sRNAs in both the nucleoid and the cytoplasm. However, for sRNA localization not to change with the deletion of *hfq* (as was observed), then the sRNAs must also be able to move without Hfq equally to the nucleoid and cytoplasm. Scenarios 2 and 3 are compatible with our observations, and both are consistent with sRNAs being able to move in and out of the nucleoid with minimal bias. To be clear, our findings do not specify where in the cell sRNAs bind and act on mRNAs.

Our observation that sRNAs can readily move into and through the nucleoid indicates they have the potential to bind mRNAs at the earliest stages of transcription and therefore can compete with ribosomal subunits for binding at the TIR. As mentioned above, the advantage of sRNAs binding to the TIR before the ribosomal subunits (instead of waiting until afterwards), is that it can potentially prevent the first round of translation initiated within the nucleoid and therefore prevent any protein at all being produced. This advantage is important for target proteins that exert their actions at low concentrations [258] and in systems that have high cooperativity, ultrasensitive switches, or positive feedback [259, 260]. As an example, proteins such as the outer membrane proteins, OmpA, OmpC and OmpF, which are regulated by the MicA, MicC and MicF sRNAs, where the expression of even a single protein could provide a route for bacteriophage to enter the cell and cause cell death [261]. In addition, if sRNAs bind to the TIR immediately after its transcription they can potentially prevent the leading ribosome from being in close proximity to the RNA polymerase, which may increase the probability of transcription termination for some genes [262-264]. The increased transcription termination would further enhance gene silencing by sRNAs. For sRNAs that increase target gene expression via opening up hairpins at the TIR [195] to facilitate ribosome binding, the capacity to enter the nucleoid and bind during the early stages of transcription would be expected to further enhance their activity by preventing transcription termination. Our demonstration that short lengths of mRNA can move through the nucleoid suggests that partial length mRNAs that are generated during transcription termination [23, 264, 265] can easily diffuse out of the nucleoid. This movement of partial length mRNAs will reduce entropic forces acting to expand the nucleoid and allow mRNA fragments to be quickly broken down and recycled by the RNA degradosome at the inner membrane [209, 266].

During stress conditions and slow growth rates, the nucleoid can become more compact resulting in less space between the folded DNA, and consequently greater resistance to the diffusion of large molecules through the nucleoid [227]. Therefore, while we found that RNAs of 442 nucleotides or less in length were able to localize in the nucleoid, this may not be the case during stress, which may explain why sRNAs are much shorter (50-250 nucleotides) [267]. The effect of stress on nucleoid localization of sRNAs and mRNAs needs to be further investigated and should be kept in mind when designing synthetic circuits.

Our findings are relevant in many ways for the design of synthetic gene circuits incorporating sRNAs and other non-coding RNAs. They directly demonstrate that the construction of synthetic gene circuits with sRNAs on plasmids will not impair these sRNAs from accessing the nucleoid and regulating target genes on the chromosome. In addition, we found little difference in the nucleoid localization of RNAs over a wide range of sizes from 102 nucleotides (RyhB) to 442 nucleotides (partial length *gfp* mRNA without RBS). Therefore synthetic sRNAs should be designed to be less than 442 nucleotides (or an $R_g < \approx 80$ Å or a diameter $< \approx 200$ Å), and probably shorter if they need to function during stress conditions for the reasons mentioned above. This constraint on size may limit the use of long non-coding RNAs, which are typically cis-acting and bind to complementary target mRNAs [203], particularly for applications where they need to act within the nucleoid to be efficient. Within the range of 442 to 1185 nucleotides it appears that as the RNA becomes larger it has more difficulty entering the nucleoid; this relationship between size and nucleoid localization needs to be further characterized. It must be stressed that size is not the only factor that may affect RNA localization. As we showed with the *bglF::gfp* mRNA, and others have shown for other RNAs [215], specific sequences can affect RNA localization, which could conceivably affect nucleoid localization of sRNAs.

Another point that is relevant to synthetic biology is the effect of localization on local RNA concentrations. Because sRNAs we investigated do not appear to sequester or concentrate in any specific regions of the cell their concentrations are simply determined by the whole volume of the cell. In contrast, mRNAs such as *gfp* (as well as *ptsG* and *bglF*) occupy a smaller volume because of exclusion from the nucleoid and therefore have higher local concentrations. Estimates of the volume of the nucleoid range from ~50-75% of the cell volume [226, 268], which means that with the same number of sRNA and mRNA molecules, the effective cytoplasmic mRNA concentration (if the mRNA is excluded from 75% of the cell volume) may be four times higher than the sRNA concentration. This difference is important in quantitative models of gene regulation, particularly for sRNAs due to their stoichiometric action [232, 269], and for mRNAs encoding cooperative proteins and other proteins with steep response curves [259, 260]. Another consequence for the modeling of sRNAs that enter the nucleoid and therefore have greater potential for silencing, is that this is expected to alter several aspects of their threshold-linear response [230, 232] (*i.e.* target protein concentration as a function of sRNA production) (**Figure 3.10**). Specifically, it is expected that: (i) the linear graded response will be "steeper" because each sRNA prevents more target proteins from being produced; (ii) the transition at the threshold will be sharper [232]; and (iii) the minimum target protein concentration will be lower (**Figure 3.10**).

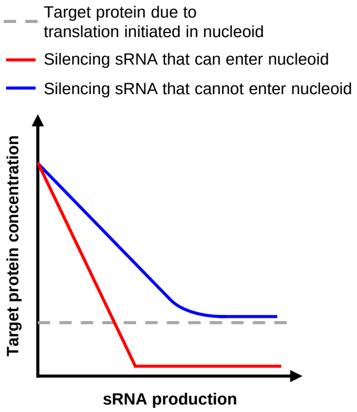


Figure 3.10 Expected threshold-linear responses for sRNAs that can and cannot enter the nucleoid with all other factors being equal. The sRNAs are decreasing translation of the target mRNA.

In conclusion, this study reveals that sRNAs can move into the nucleoid and because of this they have the potential to regulate mRNAs deep within the nucleoid, soon after mRNA transcription is initiated and the TIR is synthesized, and before the transcription-translation complex moves to the edge of nucleoid. Furthermore, sRNAs appear to occur with equal probability in the nucleoid and cytoplasm which suggests there is no bias or sequestration of sRNAs in either region. This information provides a deeper understanding of the potential roles for sRNAs in gene regulation and of the potential constraints on the evolution of sRNAs, and allows the construction of more accurate and more detailed models to optimize the engineering of synthetic circuits incorporating sRNAs.

3.5 Materials and Methods

3.5.1 Bacterial plasmids and strains

Strains, plasmids and oligonucleotide sequences are in **Table 3.1-3.3**. Plasmids were assembled using components of the pZ system [270] including the ColE1 origin, terminator sequences and promoters. Plasmid maps are in **Figure 3.11**. sRNA sequences were amplified from MG1655 and cloned downstream of the synthetic PCon [185] promoter (or PConshort promoter for GlmZ [271]). The *gfp* sequence was obtained from pTAK102 [272] and cloned downstream of the PLlacO-1 promoter [270]. The *sodB* (−56 to +141) [105], *rpoS* (−149 to +30), and *fhlA* (−107 to +96) mRNA target sequences, and the full length *bglF* sequence (without stop codon) were amplified from MG1655 and translationally fused to the *gfp* mRNA (note: numbering relative to start codon). Chromosomal sRNA genes and *hfq* were deleted using the lambda Red method [273] with the oligonucleotides in **Table 3.3**.

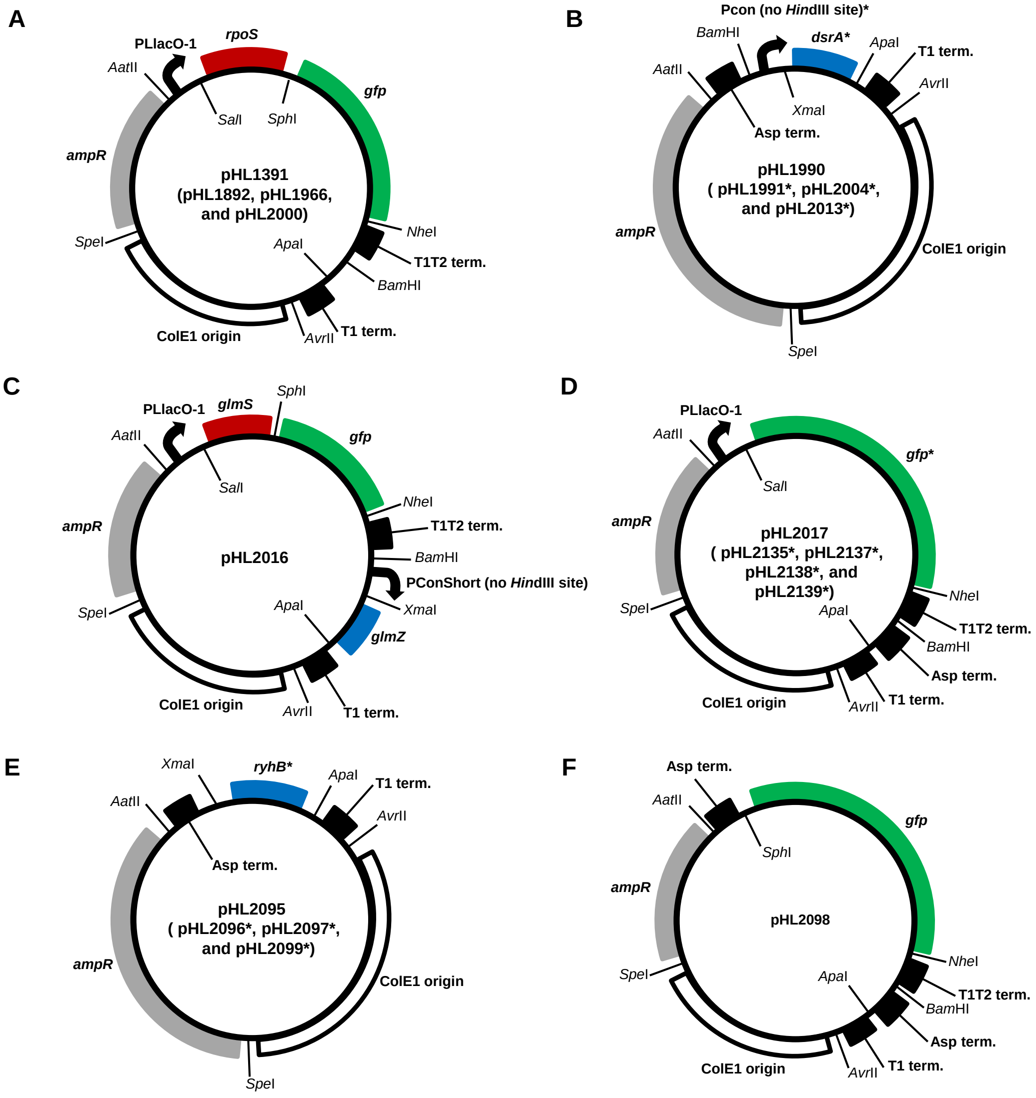


Figure 3.11 Plasmid maps. (**A**) pHL1391 and its derivatives (pHL1892, pHL1966, and pHL2000). pHL1391 was modified by replacing *rpoS* with either RBS(st7)::*bglF* or *fhlA*, and inserting an Asp terminator between the *Bam*HI and *Apa*I restriction sites to generate pHL1892 or pHL1966 respectively. pHL2000 is the same as pHL1391 except the partial *rpoS* sequences is replaced by the partial *sodB* sequence, and PCon (no *Hin*dIII site)::*ryhB* is inserted between *Bam*HI and *Apa*I in the same clockwise direction as *gfp*. (**B**) pHL1990 and its derivatives (pHL1991, pHL2004 and pHL2013). *pHL1990 plasmid was modified by replacing PCon (no *Hin*dIII site)::*dsrA* with: (i) PCon (no *Hin*dIII site)::*ryhB* generating pHL1991; (ii) PCon::*oxyS* generating pHL2004; (iii) PCon (no *Hin*dIII site)::*sgrS* generating pHL2013. (**C**) pHL2016. (**D**) pHL2017 and its derivatives (pHL2135, pHL2137, pHL2138 and pHL2139). *pHL2017 was modified by replacing the RBS(st7)::*gfp* with: (i) *gfp* (no ATG or RBS(st7)) generating pHL2135; (ii) first quarter RBS(st7) *gfp* generating pHL2137; (iii) last quarter *gfp* (no ATG or RBS(st7)) generating pHL2138; (iiii) first quarter *gfp* (no ATG or RBS(st7)) generating pHL2139. (**E**) pHL2095 and its derivatives (pHL2096, pHL2097 and pHL2099). *pHL2095 was modified by replacing *ryhB* with: (i) *oxyS* to generate pHL2096; (ii) *sgrS* generating pHL2097; (iii) *glmZ* generating pHL2099. (**F**) pHL2098.

3.5.2 RNA fluorescent in situ hybridization (RNA FISH)

RNA FISH was performed on cells inoculated from overnight culture into fresh lysogeny broth (LB) media with 100 µg/mL of ampicillin and grown at 37°C and 200 revolutions per minute (rpm) for 3.5 - 5 hours until they reached an OD_{600nm} ~ 0.3-0.5, and then they were harvested. Isopropyl β-D-1-thiogalactopyranoside (IPTG; Fisher Scientific, Fair Lawn, NJ, USA) was added as specified for individual experiments in the figure legends and protocols below. The RNA FISH protocol was the same as reported [274] except for the following modifications: (i) the volume of the cell culture and the amount of fluorescent probe were halved; (ii) probes for DsrA, RyhB, SgrS and OxyS sRNAs were in a single mix of DNA probes (labeled with Quasar Cy5); and (iii) GLOX was added after hybridization and washing (final concentration of 0.4% glucose, 10mM Tris HCL, 2 x SSC, 1% glucose oxidase and 2% of 21.6 mg/mL of catalase from bovine liver) to increase signal and prevent bleaching of Cy3 and Cy5 in accordance with the manufacturer's recommendations (Biosearch Technologies, Novato, CA, USA).

The first part of the process of RNA FISH was growing the cells and fixing them. Cells from an overnight culture were inoculated into fresh lysogeny broth (LB) media with 100 µg/mL of ampicillin (and 1 mM IPTG for the *gfp* mRNAs). The cells were grown at 37°C and 200 revolutions per minute (rpm) on a shaker for 3.5 - 5 hours to an OD_{600nm} ~ 0.3-0.5. Then 7.5 mL of the culture was removed and centrifuged for 10 minutes at 4°C and 3650 g, the supernatant was removed and the cell pellet was resuspended in 0.5 mL of fixation solution [3.7% formaldehyde (Mallinckrodt Chemicals, Phillipsburg, NJ, USA) in 1 x PBS (phosphate buffered saline)] and incubated at 25^{o}C for 30 minutes at 30 rpm. The fixed cells were centrifuged at 25^{o}C for 8 minutes at 400 g, the supernatant was removed, and the cell pellet resuspended in 1 x PBS. The cells were then centrifuged at 25^{o}C for 3.5 min at 600g and resuspended in 1 x PBS and this was repeated. The cells were resuspended in 70% ethanol and incubated at 25^{o}C for 1 h to permeabilize them, and then centrifuged at 25^{o}C for 7 minutes at 600 g, and the cell pellet resuspended in wash solution [35.43% formamide (Fisher Scientific, Fair Lawn, NJ, USA) in 2 x saline sodium citrate (SSC)] [274].

The second part of the process of RNA FISH was probe hybridization and DNA staining. These probes were designed using the manufacturer's proprietary software and labeled with Quasar Cy5 for the sRNAs and Quasar Cy3 for the *gfp* mRNA (**Table 3.4**). Fixed cells in wash solution were centrifuged at 25^{o}C for 7 minutes at 600 g and resuspended in 25 µl hybridization solution [10% w/v dextran sulfate (Pharmacia; now part of GE Healthcare Life Sciences), 2x SSC, 10% formamide, 2 mM ribonucleoside vanadyl complex (New England Biolabs, Ipswich, MA, USA), 200 µg/mL of bovine serum albumin (Sigma, St. Louis, MO, US), and 1 mg/mL of tRNA from *E. coli* MRE600 (Roche Diagnostics, Indianapolis, IN, USA)]. Note: we found the hybridization solution became unstable after multiple uses and therefore it was stored as single-use aliquots. Two µl of each fluorescently labeled probe at a concentration of 25 µM was added to each 25 µl hybridization solution (final volume is 27 µl) at 30^{o}C overnight at 30 rpm on a shaker. The next morning the probe was washed away as previously described [274]. In the final wash step, 4', 6-diamidino-2-phenylindole (DAPI; Invitrogen, Grand Island, NY, USA) was added to a final concentration of 10 µg/mL to stain the DNA and incubated at 30^{o}C for 30 minutes and then centrifuged at 25^{o}C for 3.5 minutes at 600 g. After the final wash, the cells were resuspended in 100 µl of GLOX buffer (0.4% glucose, 10mM Tris-HCL and 2 x SSC) and incubated at 25^{o}C for 5 minutes. The cells were then centrifuged for 3.5 minutes at 600 g and resuspended in 10 µl of GLOX buffer with a final concentration of 0.43 mg/ml of bovine catalase (Sigma) and 1% glucose oxidase from *Aspergillus niger* (Sigma), placed on slides with coverslips and examined immediately by fluorescence microscopy.

Cells were visualized with a Zeiss AxioObserver Z1 inverted microscope with Plan-Neofluar 100x/1.3 oil Ph3 objective and with or without the 1.6x optovar. Images were captured with a Hamamatsu EM-CCD digital camera (Model C9100-13) and iVision-Mac software (Biovision Technologies, Exton, PA, USA). The filter sets are: Cy3 (560/40 nm exciter, 660 nm longpass beamsplitter and 630/75 nm emitter); Cy5 (620/60 nm exciter, 660 nm longpass beamsplitter and 700/75 nm emitter); DAPI (350/50 nm exciter, 400 nm longpass beamsplitter and 420 nm longpass emitter); and GFP (470/40 nm exciter, 495 nm longpass beamsplitter and 525/50 nm emitter). The light source was an X-cite 120Q lamp or X-cite 120LED (Lumen Dynamics, Mississauga, Canada). Power settings, exposure times, and gain of the photomultiplier tube detector were adjusted for individual experiments to maximize the signal-to-background ratio.

3.5.3 Analysis of RNA FISH images

Images were processed in ImageJ [143]. The first step (except for the negative control without GFP, Cy3 and Cy5) was alignment of phase-contrast and fluorescence images. This was performed by subtracting background signal ("Subtract Background" function), thresholding (default algorithm), aligning thresholded images [customized "StackReg" plugin [186]], extracting offset values from this alignment, and applying the offset values to align the original phase-contrast and fluorescence images. Note: background signal is still present in these original images and consequently in the localization analyses. The second step was identification of cells in the phase-contrast images. This was done by thresholding the images (default algorithm) and converting them to binary. On the binary images, cells were initially selected based on size ("Analyze Particle" function) and then watershed segmentation [164] was used to separate dividing and touching cells. A second more stringent selection was performed to select cells: (a) with a narrow range of sizes; (b) that were rod shaped with a major axis to minor axis > 2.01 (AR filter); (c) that were below a threshold width (MinFeret filter); and (d) that did not have saturated pixels (Max measurement). Cells with an average signal-to-background of less than 1.2-fold for Cy3 and 1.3 -fold for Cy5 were not included in the analyses (see main text) except in **Figure 3.1** and for the negative control (HL716). see **Table 3.5** for analysis parameters. The cell boundaries were "regions of interest (ROI)". A "Count Mask" was created in ImageJ which filled each ROI within an image with a unique integer. The Count Mask was then used to select pixels in the fluorescence images that correspond to cells using Matlab (R2015a, Mathworks, Natick, MA, USA). Pixel intensity values within each cell were stored in an array with a unique location identifier for each cell.

3.5.4 Measurements of GFP fluorescence for membrane proteins

Bacteria with BglF::GFP were prepared as follows. An overnight culture was inoculated into fresh LB media with 100 µg/mL of ampicillin and grown at 37°C and 200 revolutions per minute (rpm) for 2-2.75 hours to an OD_{600nm} ~ 0.1. Cells were then induced at 1 mM IPTG for one hour, grown to an OD_{600nm} ~ 0.4, and placed on ice for 20 minutes. One mL of iced culture was centrifuged at 1610 g, the supernatant removed, and the pellet resuspended in 7 µl of iced LB. Three microliters of resuspended cells were mounted on glass slides with a cover slip. Fluorescence microscopy was performed using a Nikon TE2000 inverted microscope with 100x objective, 1.5 × optovar, with Ph3 annulus, X-cite 120PC lamp (Exfo, Waltham, MA, USA) and an excitation filter/dichroic mirror/emission filter set for GFP (470 ± 20 nm/495 nm/525 ± 25 nm respectively). Images were acquired using a Pixus 1024 pixel CCD camera (Princeton Instruments, Trenton, NJ, USA) and Metamorph 7.0 software (Molecular Devices, Sunnyvale, CA, USA).

3.5.5 Power calculation for determining selected fraction (F_T)

To determine the selected fraction of pixels with the highest intensity (F_T) that were needed to measure overlap with the nucleoid and cell membrane, and thus the threshold, we performed a power calculation assuming: (i) equal numbers of selected pixels for the signal of interest and the center of the nucleoid; (ii) at least 30 cells will be measured and each cell has 300 pixels (*i.e.* total "population" size = 9000 pixels); (iii) a type I error (α) = 0.05; (iv) power = 1 – type 2 error = 0.8; and (v) the observed overlap of the signal of interest with the center of the nucleoid will be at least 30% of the maximum possible overlap or 30% of the minimum possible overlap (after taking into account the expected overlap of the null distribution which is equal to F_T). There is no analytical solution so we approximated a function using the Matlab "sampsizepwd" to calculate F_T according to the above criteria. We calculated an F_T = 0.0886 and rounded to 0.1 for our analyses. Note: with a greater number of cells this selected fraction can detect statistically significant differences in the overlap from the null hypothesis of < 30% of the maximum or minimum possible overlap.

3.5.6 Calculating the fraction of GlmZ and SgrS sRNAs bound to Hfq

We observed that the deletion of *hfq* decreased the signal for GlmZ and SgrS by approximately 50% (main text). This finding is consistent with previous reports that the binding of Hfq to these sRNAs decreases their degradation rates [230, 269], and therefore the deletion of *hfq* will decrease their concentrations. From our finding we estimated the fraction of GlmZ and SgrS that are bound to Hfq using sets of equations to create a model of sRNA production and degradation (both in the presence and absence of Hfq). These equations are simplified from our previously reported models [230, 275].

The model has a constant production rate *P* (units: $M \cdot s^{-1}$) because the same promoter and plasmid is used in strains with and without *hfq*. The degradation of the unbound sRNAs is proportional to their concentration *U* (unit: M) and is specified by the rate constant γ_U (units: s^{-1}). Unbound sRNA can bind Hfq to form a bound form with concentration *H* (unit: M). The rate constants for the sRNA and Hfq binding and unbinding reactions are k_U (units: $M^{-1} \cdot s^{-1}$) and k_H (unit: s^{-1}) respectively. The degradation of sRNAs bound to Hfq is also proportional to their concentration and is specified by the rate constant γ_H (unit: s^{-1}).

In the presence of Hfq, the equations for the system are:

$$\frac{dU}{dt} = P + k_H \cdot H \cdot Hfq - k_U \cdot U - \gamma_U \cdot U, \text{ and} \qquad \text{Eq. 3.1}$$

$$\frac{dH}{dt} = - k_H \cdot H \cdot Hfq + k_U \cdot U - \gamma_H \cdot H. \qquad \text{Eq. 3.2}$$

In our experiments we measured the total sRNA concentration by RNA FISH, which is U + H, therefore we combine the above equations to give

$$\frac{d(U+H)}{dt} = P + -\gamma_U \cdot U - \gamma_H \cdot H. \qquad \text{Eq. 3.3}$$

At steady state,

$$P = \gamma_U \cdot U + \gamma_H \cdot H. \qquad \text{Eq. 3.4}$$

In the absence of Hfq, bound sRNA does not occur therefore the equation for the system is:

$$\frac{dU}{dt} = P - \gamma_U \cdot U. \qquad \text{Eq. 3.5}$$

At steady state,

$$P = \gamma_U \cdot U^*. \qquad \text{Eq. 3.6}$$

* indicates the steady state concentration of unbound sRNA is not necessarily the same in the systems with Hfq and without Hfq (**Eq. 3.4**).

For both GlmZ and SgrS, the total sRNA concentration in the *hfq* deletion mutant, which has only unbound sRNA, is approximately half that of the wild-type with *hfq*. That is,

$$U^* = \frac{1}{2}(U + H). \qquad \text{Eq. 3.7}$$

The substitution of **Eq. 3.7** into **Eq. 3.6**, and using the equality of the right hand sides of **Eq. 3.4** and **Eq. 3.6** gives

$$\gamma_U \cdot \frac{1}{2}(U + H) = \gamma_U \cdot U + \gamma_H \cdot H. \qquad \text{Eq. 3.8}$$

Rearranging **Eq. 3.8** specifies the ratio of U and H in terms of the degradation constants and incorporates the constraint obtained from the experimental observations (defined in **Eq. 3.7**). That is,

$$\frac{U}{H} = 1 - \frac{2\gamma_H}{\gamma_U}, \text{ where } H > 0 \text{ and } \gamma_U > 0. \qquad \text{Eq. 3.9}$$

Biologically U and H must be greater than or equal to zero. We set aside the case where H is equal to zero for the purpose of interpreting **Eq. 3.9** to avoid the ratio of U/H being undefined and also because it is already considered in **Eq. 3.5** and **Eq. 3.6** for the *hfq* deletion strain. Consequently,

$$1 - \frac{2\gamma_H}{\gamma_U} \geq 0, \text{ where } \gamma_U > 0. \qquad \text{Eq. 3.10}$$

Therefore,

$$\frac{\gamma_H}{\gamma_U} \leq \frac{1}{2}, \text{ where } \gamma_U > 0. \qquad \text{Eq. 3.11}$$

Because $\gamma_H \geq 0$ and $\gamma_U > 0$ then

$$0 \leq \frac{\gamma_H}{\gamma_U} \leq \frac{1}{2}. \qquad \text{Eq. 3.12}$$

We now consider the limits of the $\frac{\gamma_H}{\gamma_U}$ ratio. The lower limit, $\frac{\gamma_H}{\gamma_U} = 0$, occurs when degradation of bound sRNA is zero and only unbound sRNAs are degraded. Under these conditions, **Eq. 3.9** indicates the U/H ratio must be 1/1 to satisfy our experimental observations for GlmZ and SgrS. That is, approximately 50% of the sRNA are bound to Hfq when there is no degradation of Hfq bound sRNAs. Therefore the deletion of *hfq* and elimination of bound sRNAs would decrease the total sRNA concentration by 50%, as observed for GlmZ and SgrS. The upper limit, $\frac{\gamma_H}{\gamma_U} = \frac{1}{2}$, occurs when the value for the degradation rate constant for bound sRNAs is one half that of unbound sRNAs. In this scenario, the wild-type strain has a concentration of unbound sRNA that is essentially zero because sRNAs rapidly and stably bind to Hfq. Bound sRNAs have a concentration that is twice that of unbound sRNAs (the latter occur when *hfq* is deleted), because the degradation rate for bound sRNAs is half that of unbound sRNAs. The $\frac{\gamma_H}{\gamma_U}$ ratio cannot exceed 1:2 (*e.g.* 1:1) otherwise the total sRNA concentration in the wild-type strain is less than two-fold the concentration in the *hfq* deletion strain, and this would not be consistent with our measurements.

In summary, the 50% decrease in total sRNA concentration of GlmZ and SgrS with the deletion of *hfq* indicates that 50%-100% of these sRNAs are bound to Hfq.

Table 3.1 Strains.

Strain	Description	Source	Antibiotic Resistance
HL1	MG1655 + pKD46	[185]	amp
HL713	HL1 + integration at *intS* of PCR product *kanR::lacIq* amplified from pHL67 with intspkd1f and laciqints	This study	kan
HL716	HL713 + pCP20 and cured	[185]	none
HL744	HL716 + pKD46	[276]	amp
HL751	HL744 + Δ*hfq* using pKD13 and oligonucleotides hfqpkd1f and hfqpkd4r	This study	kan
HL752	HL744 + Δ*sgrS* using pKD13 and oligonucleotides sgrsko1pkd1f and sgrsko2pkd4r	[276]	kan
HL756	HL752 + pCP20 and cured	This study	none
HL770	HL751 + pCP20 and cured	This study	none
HL772	HL770 + pKD46	This study	amp
HL852	HL744 + Δ*dsrA* using pKD13 and oligonucleotides dsrako1pkd1f and dsrako2pkd4r	This study	kan
HL865	HL852 + pCP20 and cured	[185]	none
HL2729	HL744 + Δ*ryhB* using pKD13 and oligonucleotides rybpkd1f and rybpkd4r	This study	kan
HL2752	HL2729 + pCP20 and cured	[185]	none
HL3221	HL744 + Δ*oxyS* using pKD13 and oligonucleotides oxys1pkd1f and oxys2pkd4r	This study	kan
HL3262	HL3221 + pCP20 and cured	[185]	none
HL3325	HL772 + Δ*ryhB* using pKD13 and oligonucleotides rybpkd1f and rybpkd4r	This study	kan
HL3338	HL3325 + pCP20 and cured	[185]	none
HL3387	HL772 + Δ*oxyS* using pKD13 and oligonucleotides oxys1pkd1f and oxys2pkd4r	This study	kan
HL3425	HL3387 + pCP20 and cured	[185]	none
HL5212	HL744 + Δ*glmZ* using pKD13 and oligonucleotides glmzkopkd1f and glmzkopkd4r	This study	kan
HL5226	HL5212 + pCP20 and cured	This study	none
HL5316	HL5226 + pKD46	This study	amp
HL5378	HL5316 + Δ*glmY* using pKD13 and oligonucleotides glmykopkd1f and glmykopkd4r	This study	kan
HL5390	HL5378 + pCP20 and cured	This study	none
HL5969	HL716 + pHL1892	This study	amp
HL6040	HL744 + Δ*yhbJ* using pKD13 and oligonucleotides yhbjshortpkd1f and yhbjshortpkd4r	This study	kan
HL6128	HL5390 + Δ*hfq* via transduction from HL6040	This study	kan
HL6190	HL6128 + pCP20 and cured	This study	amp
HL6193	HL716 + pHL1391	[185]	amp
HL6201	HL716 + pHL1966	This study	amp
HL6268	HL2752 + pHL1991	This study	amp
HL6269	HL865 + pHL1990	This study	amp

HL6284	HL2752 + pHL2000	This study	amp
HL6285	HL3338 + pHL2000	This study	amp
HL6286	HL3338 + pHL1991	This study	amp
HL6317	HL756 + Δ*hfq* via transduction from HL751	This study	kan
HL6318	HL3262 + pHL2004	This study	amp
HL6319	HL3425 + pHL2004	This study	amp
HL6320	HL5390 + pHL2016	This study	amp
HL6321	HL6190 + pHL2016	This study	amp
HL6322	HL716 + pHL2017	This study	amp
HL6332	HL752 + pHL2013	This study	(kan) amp
HL6333	HL6317 + pHL2013	This study	(kan) amp
HL6530	HL2752 + pHL2095	This study	amp
HL6531	HL3262 + pHL2096	This study	amp
HL6532	HL752 + pHL2097	This study	(kan) amp
HL6533	HL716 + pHL2098	This study	amp
HL6547	HL5390 + pHL2099	This study	amp
HL6733	HL716 + pHL2135	This study	amp
HL6735	HL716 + pHL2137	This study	amp
HL6736	HL716 + pHL2138	This study	amp
HL6737	HL716 + pHL2139	This study	amp

(kan) = kanamycin resistance but kanamycin not used for selection in the experiment

Table 3.2 Plasmids.

Plasmid	Description	Source	Antibiotic Resistance
pHL67	*lacIq* from pTrc99a + ColE1 from pZE21 + *kanR* cassette from pKD13 (including P1 and P4 oligonucleotide sites); template for lacIq insertion into the genome	[185]	kan
pHL1391	*ampR* + PLlacO-1::*rpoS*::*gfp*::T1T2 terminator + T1 terminator + ColE1	[185]	amp
pHL1966	*ampR* + PLlacO-1::*fhlA*::*gfp*::T1T2 terminator + T1 terminator + ColE1	This study	amp
pHL1892	*ampR* + PLlacO-1::RBS (st7)::*bglF*::*gfp*::T1T2 terminator + Asp terminator + T1 terminator + ColE1	This study	amp
pHL1990	*ampR* + Asp terminator + PConNoHind::*dsrA*::T1 terminator + ColE1	This study	amp
pHL1991	*ampR* + Asp terminator + PConNoHind::*ryhB*::T1 terminator + ColE1	This study	amp
pHL2000	*ampR* + PLlacO-1::*sodB*::*gfp*::T1T2 terminator + PConNoHind::*ryhB*::T1 terminator + ColE1	This study	amp
pHL2004	*ampR* + Asp terminator + PCon::*oxyS*::T1 terminator + ColE1	This study	amp
pHL2013	*ampR* + Asp terminator + PConNoHind::*sgrS*::T1 terminator + ColE1	This study	amp
pHL2016	*ampR* + PLlacO-1::*glmS*::*gfp*::T1T2 terminator + PConShortNoHind::*glmZ*::T1 terminator + ColE1	This study	amp
pHL2017	*ampR* + PLlacO-1::RBS (st7) *gfp*::T1T2 terminator + Asp terminator + T1 terminator + ColE1	This study	amp
pHL2095	*ampR* + Asp terminator::*ryhB*::T1 terminator + ColE1	This study	amp
pHL2096	*ampR* + Asp terminator::*oxyS*::T1 terminator + ColE1	This study	amp
pHL2097	*ampR* + Asp terminator::*sgrS*::T1 terminator + ColE1	This study	amp
pHL2098	*ampR* + Asp terminator::*gfp*::T1T2 terminator + Asp terminator + T1 terminator + ColE1	This study	amp
pHL2099	*ampR* + Asp terminator::*glmZ*::T1 terminator + ColE1	This study	amp
pHL2135	*ampR* + PLlacO-1::*gfp* (no RBS, no ATG)::T1T2 terminator + Asp terminator + T1 terminator + ColE1	This study	amp
pHL2137	*ampR* + PLlacO-1::RBS (st7) *gfp* (first quarter)::T1T2 terminator + Asp terminator + T1 terminator + ColE1	This study	amp
pHL2138	*ampR* + PLlacO-1::*gfp* (last quarter, no RBS, no ATG)::T1T2 terminator + Asp terminator + T1 terminator + ColE1	This study	amp
pHL2139	*ampR* + PLlacO-1::*gfp* (first quarter, no RBS, no ATG)::T1T2 terminator + Asp terminator + T1 terminator + ColE1	This study	amp

Table 3.3 Oligonucleotides.

Forward	Reverse	Sequence	Function (strain created)
dsrako1pkd1f	dsrako2pkd4r	atatggcgaatattttcttgtcagcgaaaaaaattgcggataaggtgatggtgtaggctggagctgcttc	delete *dsrA* using pKD13 as template in HL852
dsrako2pkd4r	dsrako1pkd1f	tattcatgacttcagcgtctctgaagtgaatcgttgaatgcacaataaaaattccggggatccgtcgacc	delete *dsrA* using pKD13 as template in HL852
glmykopkd1f	glmykopkd4r	agttcagatacaacaaagccgggaattacccggctttgttatggaataaggtgtaggctggagctgcttc	delete *glmY* using pKD13 as template in HL5378
glmykopkd4r	glmykopkd1f	cgttaccaaactattttctttattggcacagttactgcataatagtaaccattccggggatccgtcgacc	delete *glmY* using pKD13 as template in HL5378
glmzkopkd1f	glmzkopkd4r	tagttccttctcacccggaggcaagcacctccggggccttcctgatacatgtgtaggctggagctgcttc	delete *glmZ* using pKD13 as template in HL5212
glmzkopkd4r	glmzkopkd1f	acaagtgttaagggatgttatttcccgattctctgtggcataataaacgaattccggggatccgtcgacc	delete *glmZ* using pKD13 as template in HL5212
hfqpkd1f	hfqpkd4r	tcagaatcgaaaggttcaaagtacaaataagcatataaggaaaagagagagtgtaggctggagctgcttc	delete *hfq* using pKD13 as template in HL751, HL1120, and HL1179
hfqpkd4r	hfqpkd1f	ggaacgcaggatcgctggctccccgtgtaaaaaaacagcccgaaaccttaattccggggatccgtcgacc	delete *hfq* using pKD13 as template in HL751, HL1120, and HL1179
intspkd1f	laciqints	ccgtagatttacagttcgtcatggttcgcttcagatcgttgacagccgcagtgtaggctggagctgcttc	PCR amplify *kanR::lacIq* using pHL67 as template to integrate at *intS*
laciqints	intspkd1f	atagttgttaaggtcgctcactccaccttctcatcaagccagtccgcccagctaactcacattaattgcgttgc	PCR amplify *kanR::lacIq* using pHL67 as template to integrate at *intS*
oxys1pkd1f	oxys2pkd4r	agcaatgaacgattatccctatcaagcattctgactgataattgctcacagtgtaggctggagctgcttc	delete *oxyS* using pKD13 as template in HL3221 and HL3387

oxys2pkd4r	oxys1pkd1f	atttatatgtataaatttgagcctggcttatcgccgggcttttttatggcattccggggatccgtcgacc	delete *oxyS* using pKD13 as template in HL3221 and HL3387
rybpkd1f	rybpkd4r	gattttgaggatggttgagagggttgcagggtagtagataagttttagatgtgtaggctggagctgcttc	delete *ryhB* using pKD13 as template in HL2729 and HL3325
rybpkd4r	rybpkd1f	tttgcacaaccgcagaacttttccgcagggcatcagtcttaattagtgccattccggggatccgtcgacc	delete *ryhB* using pKD13 as template in HL2729 and HL3325
sgrsko1pkd1f	sgrsko2pkd4r	gcaaaagacagcaattttattttccctatattaagtcaataattcctaacgtgtaggctggagctgcttc	delete *sgrS* using pKD13 as template in HL752
sgrsko2pkd4r	sgrsko1pkd1f	gccatcgtcattatccagatcatacgttccctttttagcgcggcgagaatattccggggatccgtcgacc	delete *sgrS* using pKD13 as template in HL752
yhbjshortpkd1f	yhbjshortpkd4r	atgcccagcttgtttgtgatttcaacagtttgcttgacgggtgtaggctggagctgcttc	delete *yhbJ* using pKD13 as template in HL6040
yhbjshortpkd4r	yhbjshortpkd1f	cggtaatgtctcttttagacgttgtgaggagaaacagtacattccggggatccgtcgacc	delete *yhbJ* using pKD13 as template in HL6040

Table 3.4 FISH probe sequences.

Oligonucleotide	Fluorescence label	Sequences	Function
gfp	Quasar Cy3	aagttcttctcctttacgca gaattgggacaactccagtg acatcgccatctaattcaac gacagagaatttttgcccat catcaccttcaccctctcca agggtaagttttccgtatgt tcccagtagtgcaaataaat gttggccatggaacaggtag ataaccgaaagtagtgacaa atctcgcaaagcattgaaca tgctgtttcatatgatctgg catggcactcttgaaaaagt tttcctgtacataaccttcg gttcccgtcatctttgtaaa tgacttcagcacgtgtcttg acaagggtatcaccttcaaa accttttaactcgattctat ttccatcttctttaaaatca tccattttgtgtccaagaat attatgtgagttatagttgt gtttgtctgccatgatgtat ttaactttgattccattctt aatgttgtgtctaattttga ctaattgaacgcttccatct gtattttgttgataatggtc gacagggccatcgccaattg ggtaatggttgtctggtaaa gaaagggcagattgtgtgga tctcttttcgttgggatctt actcaagaaggatcatgtga gtaatcccagcagctgttac tgtatagttcatccatgcca	probing for *gfp* mRNA
DsrA	Quasar Cy5	caccaggaaatctgatgtgt gcttaagcaagaagcactta tgagggggtcgggatgaaac	probing for DsrA sRNA
GlmZ	Quasar Cy5	gagatggaatgagcatctac tgaggcactaaggcgaacat ctctgcgtcattccggagtt ggacgataagcaccgtaaac ggcataagcgacatctgtca ttgtgtccatggtgtctgat caagtgggtgcttcactcaa	probing for GlmZ sRNA

		gcgttaaaacaggtctgtat gcctgctcttattacggagc	
OxyS	Quasar Cy5	aagaggtgccgctccgtttc gggcagtgacttcaagggtt cgagttgagaaactctcgaa gttcacgttggctttagtta cggatcctggagatccgcaa	probing for OxyS sRNA
RyhB	Quasar Cy5	gcgagggtcttcctgatcgc atgtcgtgctttcaggttct aatactggaagcaatgtgag gccagcacccggctggctaa	probing for RyhB sRNA
SgrS	Quasar Cy5	gggcaccccccttgcttcatc gtgctgataaaactgacgca acttcgctgtcgcggtaaaa cttaaccaacgcaaccagca catggttaatcgttgtggga tcccactgcatcagtccttc tcaactttcagaattgcggt agtcacacatgatgcaggca gggtgattttacaccaatac ccagcaggtataatctgctg	probing for SgrS sRNA

Table 3.5 Parameter values for experimental data analyses.

Microscope type	Initial size filter (pixel^2)	Second size filter (pixel^2)	AR filter	MinFeret filter (pixels)	Radius 1 (pixels)	Radius 2 (pixels)
Nikon TE2000E	[300,1100]	[300,900]	(2.01, ∞)	[0,30]	25	50
Zeiss AxioObserver Z1 (optovar 1.0 ×)	[100,400]	[100,300]	(2.01, ∞)	[0,10]	25	50
Zeiss AxioObserver Z1 (optovar 1.6 ×)	[200,800]	[200,600]	(2.01, ∞)	[0,15]	25	50

"Initial size filter" displays the minimum and maximum particle areas selected before watershed segmentation. "Second size filter" displays the minimum and maximum particle areas selected after watershed segmentation. "AR filter" is the lower and upper bounds of the major to minor axis used to select shape. "MinFeret" is the lower and upper bounds of the minimum caliper diameter for cells. "Radius 1" and "Radius 2" are the rolling ball radii used in the "Subtract Background" function for phase-contrast and fluorescence images respectively.

Table 3.6 Radii of gyration for bacterial single stranded RNAs.

RNA	Length (nucleotides)	Radius of gyration# (Å)	Source
S-adenosyl methionine riboswitch (unbound)	51	25	[277]
S-adenosyl methionine riboswitch (unbound)	52	31.7	[277]
fragment from 5S ribosomal RNA	62	28.5, 28.7 [28.6]	PDB: 357D, 364D
MicA small RNA	75	33.9	[212]
tRNA (valine)	76	24.4	PDB: 2K4C
tRNA (fMet)	77	22.6	PDB: 3CW5
thiamine pyrophosphate riboswitch (unbound)	83	27.5	[277]
DsrA small RNA	87	43.2	[212]
S-adenosyl methionine riboswitch (unbound)	94	29.9	[277]
cyclic diguanylate riboswitch (unbound)	98	32	[277]
yybP-ykoY Mn riboswitch	107	34.6	PDB: 4Y1M
5S ribosomal RNA	120	32.7	[278]
5S ribosomal RNA	120	36.1	[279]
flavin mononucleotide riboswitch (unbound)	141	29.4	[277]
lysine riboswitch (unbound)	181	43	[277]
Ribozyme (*Azoarcus*)	195	~60	[280]
glycine riboswitch (unbound)	226	45	[281]
rpoS mRNA (partial sequence)	284	68.1	[254]
ribonuclease P ribozyme	400	43.9, 44.3, 48 [45.4]	[282]
random intergenic sequence transcribed	975	182	[283]
12S (partial 16S)	~1000	71	[284]
random intergenic sequence transcribed	1523	208	[283]
16S ribosomal RNA	1541	176	[285]
16S ribosomal RNA	1541	189, 161 [175]	[286]
16S ribosomal RNA	1541	114	[250]
mRNA sequence (cowpea chlorotic mottle virus)	2777	172	[283]
23S ribosomal RNA	2904	230	[286]
MS2 RNA (bacteriophage)	3569	181	[287]

§ Radius of gyration was calculated from structures deposited in The Nucleic Acid Database Project at Rutgers, The State University of New Jersey. # The hydrodynamic radius is assumed to be approximately the same for RNAs [288]. * Classified as non-riboswitch and non-ribozyme RNA structures. Values in square brackets [ ] are the average.

Chapter 4: Translational control of mRNA concentrations optimizes both quality control and launch control in bacteria

4.1 Abstract

Bacteria have evolved to be extremely efficient to maximize their growth rate and survival under stress. And yet bacteria spend valuable energy and resources when the translation rate is low on initiating the transcription of mRNAs and further investing in terminating their transcription and degrading the mRNAs ("translational mRNA control") rather than turning off transcription initiation. Here we investigated the trade-offs associated with translational mRNA control using mathematical models and analyses of experimental data. We show that translational mRNA control can function as (1) a quality control mechanism with the leading ribosome acting as the primary evaluator of mRNA quality thereby preventing the creation of potentially deleterious proteins, and (2) a "launch control" system with a high baseline rate of transcription initiation acting as an accelerator and transcriptional termination and mRNA degradation simultaneously acting as a brake on the protein production resulting in a faster response at times of need. These advantages are at the cost of less efficient resource use, including high levels of partially transcribed mRNAs that sequester valuable resources. These findings explain previously unexplained experimental observations and support translational control as a mechanism that expends resources for other potential benefits to gene regulations.

4.2 Introduction

The central dogma of molecular biology is a two-step process whereby genetic information encoded by DNA is first transcribed to create messenger RNA (mRNA) and then translated into proteins that perform most of the functions in the cell (**Figure 4.1A**). Classic models of gene regulation based on the central dogma usually consider four independent rates can be used to control protein concentrations [289]: (i) mRNA production (*i.e.* transcription); (ii) mRNA degradation; (iii) protein production (*i.e.* translation); and (iv) protein degradation. Independently regulating these four different process would seem more than adequate to control protein concentrations, and yet many biological systems have evolved additional interactions and complexity.

In bacteria, mRNA transcription and degradation are usually dependent of mRNA translation. Specifically, low translation rates can lead to less transcription due to increased transcription termination by Rho-dependent and Rho-independent mechanisms [28, 290, 291], which is generally referred to as transcription-translation coupling. A low translation rate also often increases mRNA degradation due to less ribosomes preventing RNases accessing the mRNA [292, 293] (note: we refer to this as "mRNA degradation-translation coupling"). Both mechanisms lower concentration of full length mRNA at low translation rates, therefore we will refer collectively to transcription-translation coupling and mRNA degradation-translation coupling as "translational mRNA control".

The existence of mRNA-translation coupling has long been known and some of the mechanisms have been elucidated in some detail (although there is still much to learn). But the reasons these mechanisms have evolved are unclear. Translational mRNA control may exist to: (i) provide quality control so that mutant proteins, partial proteins or misfolded proteins (that occur due to genetic, transcription or translation errors) that can be toxic to the cell are not created [294]; (ii) prevent the formation of heteroduplexes of RNA and one strand of double stranded DNA (R loops) that are deleterious to the cell [295]; (iii) conserve resources by preventing the synthesis of mRNAs and sequestration of ribosomes by mRNAs that are not being translated [293, 296]; (iv) a mechanism for limiting the use of resources at times of competition for ribosomes, RNA polymerase, ATP, amino acids, and other factors [297-299] (in contrast to point (iii) it is not the case that the encoded protein is not required).

The idea that translation termination is a mechanism that limits the use of energy and resources (point (iii) above) builds on the idea from early research on the *trp* operon, which encodes proteins necessary for the production of tryptophan. In this system, transcription of the operon terminates when there are sufficient levels of tryptophan in the cell, and thus further production of tryptophan would be a waste of resources. However, there are features of this operon that indicate that it may not be a representative example of translational mRNA control. The *trp* operon is a very specific example where the protein product of the gene is directly involved in translation. Furthermore, it is an example of regulation where the cell has sufficient resources (*i.e.* adequate tryptophan) rather than too little resources, which occurs with most genes. Therefore, it is unclear to what extent this operon reflects a general principle for the evolution of transcription termination and translational mRNA control.

While it is clear that terminating transcription part way through translation or rapidly degrading mRNAs that are not being translated could save energy and conserve the use of resources such as nucleotides and amino acids, this approach uses more resources than decreasing transcription initiation in the first place. The case for translational mRNA control being the best available mechanism due to barriers in evolving general mechanisms that couple translation and transcription initiation is not compelling, as such mechanism are already known to exist, including sigma factors and (p)ppGpp [11, 300].

A recent study quantified the effect of mRNA translational control on mRNA and protein concentrations [262]. The study demonstrated a power-law relationship between protein and mRNA concentrations, and further evidence for this was observed in an independent study [262, 301]. Specifically, the protein concentration (P) increase as the mRNA concentration (M) to the power of 3.6; that is, $P \approx M^{3.6}$ [262]. The basis for this empirical observation was shown to be due to the action of both transcription-translation coupling and mRNA degradation-translation coupling. The precise reason that a power-law (or similar relationship) emerges and the implications for gene regulation were not investigated. However, it is clear that translational mRNA control is a mechanism that holds across a large range of translation rates, and therefore the close relationship between transcription, degradation and translation of mRNAs may have regulatory properties that have favored its selection.

To understand the potential regulatory properties of translational control mechanisms it would be helpful to understand how it affects mRNA transcription, and degradation and protein concentrations, then it would be possible to evaluate the possible advantages and disadvantages. Mathematical models can be a powerful tool to do this and provide a framework for a mechanistic understanding gene regulatory interactions [230, 269, 271, 272, 275, 302, 303]. Therefore in this study, mathematical models were created to specifically examine the general regulatory effects of translational mRNA control on mRNA and proteins concentration.

4.3 Results

4.3.1 Power-law relationship and negative cooperativity in a simple translational mRNA control model

In classic models of central dogma ("central dogma model") (**Figure 4.1A, B**), increasing the transcription rate (k_m; [mRNA]·s^{-1}) increases the mRNA concentration and increasing the translation initiation rate (k_p; [protein]·[mRNA]$^{-1}$·s^{-1}) increases the amount of protein made from each mRNA. The steady state mRNA concentration [mRNA]$_{ss}$ is simply the transcription rate divided by mRNA degradation rate (δ_m; s^{-1}). Therefore increasing k_p increases the protein concentration but has no effect on the mRNA concentrations (**Figure 4.1C**); and consequently on a plot of protein concentration as a function of the mRNA concentration, varying k_p simply results in a vertical line (**Figure 4.1C**). Multiplication of the mRNA concentration and the translation rate determines the production rate of proteins. The protein production rate divided by the degradation rate of the protein (δ_p) determines the steady state protein concentration [protein]$_{ss}$. That is,

$$[\mathrm{mRNA}]_{ss} = \frac{k_m}{\delta_m} \qquad \text{Eq. 4.1}$$

and

$$[\mathrm{protein}]_{ss} = \frac{k_p}{\delta_p}[\mathrm{mRNA}]_{ss}. \qquad \text{Eq. 4.2}$$

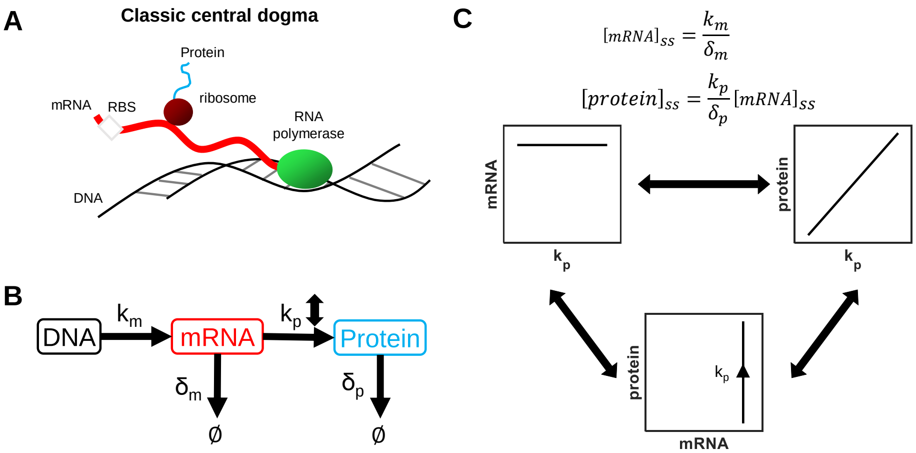


Figure 4.1. The "classic central dogma" model. (**A**) The simplest classic system includes transcription, mRNA degradation, translation, and protein degradation. The system does not contain active mRNA degradation (but dilution of mRNA due to cell growth is included) and it does not contain any transcription termination. RBS: ribosome binding site. (**B**) Kinetic scheme of the classic central dogma. (**C**) The effect of altering the translation rate (k_p) on full length mRNA and protein concentrations in the classic central dogma model. Formulas provided applies to all three plots. Black arrow on the protein-mRNA plot indicates the direction of the change when k_p increases.

In models of bacterial gene regulation with translational mRNA control ("real-world models") (**Figure 4.2A, B**), the rate of translation affects both the mRNA production rate (via its effect on transcription termination) and the mRNA degradation rate (via the effect of ribosomes on degradation). Since translation affects mRNA production and degradation, they both act to increase the full length mRNA concentration or both act to decrease the full length mRNA concentration, one of the simplest ways to represent their combined effect (*i.e.* translational mRNA control) on k_p is by using a Hill function [289] (**Figure 4.2C**). That is,

$$[\mathrm{mRNA}]_{ss} = \frac{k_m}{\delta_m}\left(\epsilon + \frac{k_p^n}{k_p^n + K^n}\right), \qquad \text{Eq. 4.3}$$

where K is the half-maximal effect of translational mRNA control, ϵ is the amount of mRNA that is generated independent of translation, and n is the Hill coefficient that describes the cooperativity. In most systems, the amount of mRNA generated in the absence of translation is negligible therefore ϵ is approximately zero, and the equation can be simplified to

$$[\mathrm{mRNA}]_{ss} = \frac{k_m}{\delta_m}\frac{k_p^n}{k_p^n + K^n} \qquad \text{Eq. 4.4}$$

In this system, when < 1 , $n = 1$ and $n > 1$, it is analogous to negative cooperativity, no cooperativity and positive cooperativity respectively. Negative cooperativity means that initially increasing the translation rate has maximal impact on the mRNA concentration and there is a diminishing return as the translation rate is further increased. Positive cooperativity means that initially increasing the translation rate has minimal impact on the mRNA concentration and there is increasing impact as the translation rate is increased. No cooperativity means that increasing the translation rate has the same relative impact at all levels. Due to the incorporation of translational mRNA control in this model, the mRNA concentration is no longer independent of the translation rate (k_p).

The simple model described above for the translational control model can be compared quantitatively with experimental measurements (**Figure 4.2D**). Although it is difficult to measure the translation initial rate (k_p) directly, it is relatively straight-forward to compare a plot of the protein concentration as a function of the mRNA concentration as k_p is increased (**Figure 4.2E**). It has been observed that there is a power-law relationship with a power of approximately 3.4 when the steady state protein concentration ($[\mathrm{protein}]_{ss}$) is plotted as a function of the mRNA concentration ($[\mathrm{mRNA}]_{ss}$). That is,

$$[\mathrm{protein}]_{ss} = 0.923[\mathrm{mRNA}]_{ss}^{3.405}, \qquad \text{Eq. 4.5}$$

(**Figure 4.2D**) [262].

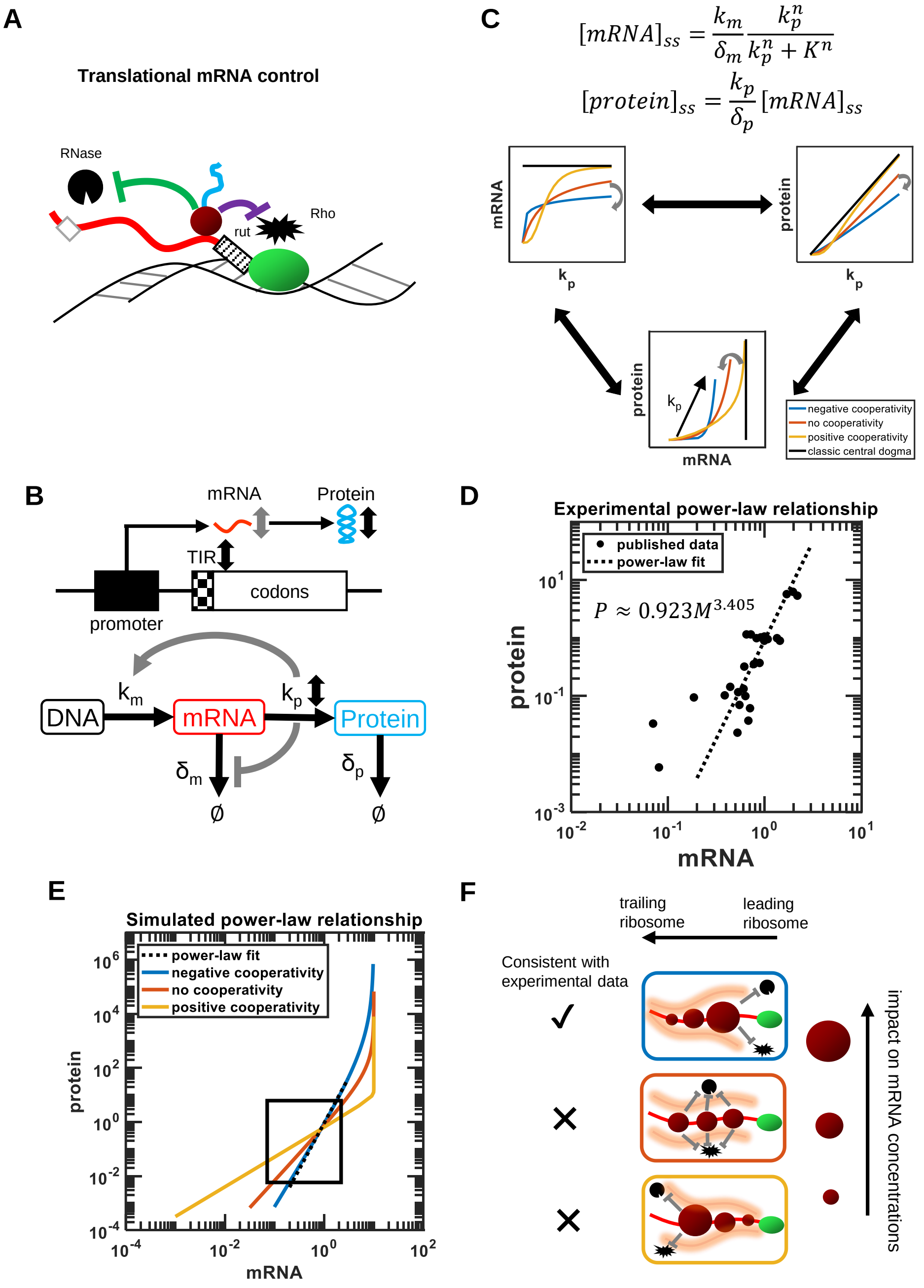

A
Translational mRNA control
RNase
rut
Rho
C
$[mRNA]_{ss} = \frac{k_m}{\delta_m} \frac{k_p^n}{k_p^n + K^n}$
$[protein]_{ss} = \frac{k_p}{\delta_p} [mRNA]_{ss}$
mRNA
k$_p$
protein
k$_p$
protein
k$_p$
mRNA
negative cooperativity
no cooperativity
positive cooperativity
classic central dogma
B
mRNA
Protein
TIR
codons
promoter
DNA
k$_m$
mRNA
k$_p$
Protein
δ$_m$
δ$_p$
∅
∅
D
Experimental power-law relationship
published data
power-law fit
$P \approx 0.923M^{3.405}$
protein
10^1
10^{-1}
10^{-3}
10^{-2}
10^{-1}
10^0
10^1
mRNA
E
Simulated power-law relationship
power-law fit
negative cooperativity
no cooperativity
positive cooperativity
protein
10^6
10^4
10^2
10^0
10^{-2}
10^{-4}
10^{-4}
10^{-2}
10^0
10^2
mRNA
F
trailing ribosome
leading ribosome
Consistent with experimental data
✓
✕
✕
impact on mRNA concentrations

Figure 4.2. The translational mRNA control model. (**A**) Translating ribosomes prevent RNases from accessing the mRNA and degradating it and by preventing transcription termination (the combination of mechanisms is "translational mRNA control"). (**B**) Kinetic scheme for the translational mRNA model. Increasing the translation rate increases full length mRNA concentration by increasing the transcription rate and decreasing the degradation rate for mRNAs. (**C**) Mathematical model as described in the main text. The effect of altering the translation rate (k_p) on full length mRNA and protein concentrations in the translational mRNA model. Grey arrows indicate the direction the translational mRNA control shifts the curve compared to the classic central dogma model. (**D**) Previously reported data between the full length mRNA and protein concentrations, and a fit of a power-law function to the data (Hussein et al., 2015). (**E, F**) Translational mRNA control with negative cooperativity (blue line) is consistent with the experimentally observed power-law relationship (black dash line) between full length mRNA concentrations and protein concentrations. The black box indicates the range of full length mRNA and protein concentrations of the published data in (**C**). Negative cooperativity indicates the leading ribosome has the most the impact. The size of ribosome represents its relative impact on the inhibition of both transcription termination and mRNA degradation.

To understand the implication of this power-law relationship between mRNA and protein concentration, we simplified the translational mRNA control model. We started with the **Eq. 4.4** which describes the overall translational mRNA control with both transcription-translation coupling and mRNA degradation-translation coupling. The exponent of the power-law relationship in **Eq. 4.5** is the slope on the log-log scale. We can obtain this exponent of the function by obtaining the derivative of protein concentration over mRNA concentration on the log-log scale; that is,

$$\text{exponent} = \frac{d\log[\text{protein}]_{ss}}{d\log[\text{mRNA}]_{ss}} = 3.405 \qquad \text{Eq. 4.6}$$

(See more detail in **Materials and Methods** about the exponent definition). Using **Eq. 4.6** to obtain the exponent and the derivatives of **Eq. 4.2** and **4.4** with respect to k_p we obtained

$$\text{exponent} = \frac{1+\left(\frac{k_p}{K}\right)^n}{n} + 1\cdot \qquad \text{Eq. 4.7}$$

The exponent increases as k_p increases, and the upper and lower bounds are

$$\text{exponent} \in [\frac{1}{n} + 1, \infty). \qquad \text{Eq. 4.8}$$

The exponent was observed to be relatively constant in the previously published experimental data as the translation initiation rate (k_p) was broad range, which indicates that the exponent of 3.405 holds when $k_p \ll K$ (and $k_p/K \approx 0$). Therefore **Eq. 4.7** can be simplified and made equal to the experimental measured value for the exponent (**Eq. 4.6**) to solve for *n*:

$$\frac{1}{n} + 1 \geq 3.405. \qquad \text{Eq. 4.9}$$

The upper bound of the Hill coefficient is therefore $n \leq 0.4158$. This bound, however, might be too stringent because experimental observed exponent might be noisy and the previously reported exponent in a different paper is only 3.2 [82]. Therefore, we altered this bound by only requiring the exponent to be greater than three. As a consequence, the Hill coefficient has a new bound of $n < 0.5$. In many biological systems, Hill functions operated near the half maximal point (K) where $k_p/K \approx 1$, and if were the case the Hill coefficient would still be less than 1. Therefore negative cooperativity (*i.e.* n < 1) of translational mRNA control seems to be a robust prediction of the model. Note: a more general solution with any abstract function (not just the Hill function) is derived in the **Materials and Methods** section.

To confirm the results of the above mathematical analysis, a simulation of the translational mRNA control model was performed (**Figure 4.2E**). The simulation was performed in Matlab (R2016b, Mathworks) and parameter values were set manually or calculated based on the conditions (see **Materials and Methods** for further details). Again, the results only aligned with the experimental data when there was negative cooperativity (**Figure 4.2E**).

The negative cooperativity indicates that the leading ribosome has an outsized role in translational mRNA control. This finding is consistent with previous experimental studies that have shown that the leading ribosome has been shown to prevent RNAP from backtracking and unexpected stalling [28, 304] and the leading ribosome has important roles in translation initiation that do trailing ribosomes do not have [305]. In addition, there is some evidence that the leading ribosome has the most impact on mRNA degradation [108, 306].

The importance of the leading ribosome may be explained by at least two possible roles: (i) it is a "test-pilot" checking for errors in the target mRNA sequence (*i.e.* a quality control mechanism); and (ii) it prepares the cell for maximal protein production when the cell decides to switch on the gene (*i.e.* a “launch control” mechanism). As a quality control mechanism, it is the leading ribosome that will first identify issues in the mRNA sequence and translation (such as mutations, secondary structures that cause pausing, and limited availability of tRNAs, amino acids, ribosomes and other factors needed for translation). Launch control is the mechanism used in vehicles to maximize acceleration that works by simultaneously applying an accelerator and brake so the engine is at optimal revolutions per minute (rpm) for acceleration before the brake is removed, which is in contrast to traditional acceleration where the brake is first removed and then the accelerator is applied resulting in a substantial lag in acceleration. Launch control can also apply to gene regulation to maximize the speed of response to changes in the environment. In this case, a high transcription initiation rate is the analogous to applying the accelerator and transcription termination and mRNA degradation are analogous to applying brakes. Negative cooperativity in this system means minimal translation would be needed to remove the brakes thereby allowing protein production to occur with minimal delay.

4.3.2 A detailed kinetic model of translational mRNA control

To understand the role that translational mRNA control may play in quality control and launch control we created a detailed kinetic model to investigate features not included in the simple model above including: (i) the separate contributions of transcription-translation coupling and mRNA degradation-translation coupling to quality control and launch control; (ii) the partial length mRNAs as well as full length mRNAs in quality control; and (iii) the dynamics of mRNA and protein concentration changes for launch control; and (iv) as a separate validation of the findings from the simple model. Because it is critical to anchor models with experimental observations and measurements, the model will focus on the specific system and the quantitative data was previously reported [262].

The kinetic model has three major classes of RNAs: (i) nascent mRNAs that occur because some mRNA synthesis occurs after transcription initiation but before transcription elongation is initiated (see Introduction), (ii) partial length mRNAs that occur after transcription elongation is initiated but before completion of transcription of the full length mRNA, and (iii) full length mRNAs (**Figure 4.3A, B**). At the beginning of transcription, RNA polymerase (RNAP) binds to the promoter and initiates transcription of the nascent mRNA before going onto transcription elongation. During this interval or just after the formation of the nascent mRNA, ribosomes and Rho factor can compete for binding at sequence in the 5’ end of the mRNA. If Rho factor binds, transcription will terminate and partial length mRNA will be generated. On the other hand, if ribosomes binding takes place first, then transcription will continue and a protein can be produced. With a very small probability, neither ribosome nor Rho factor binds during transcription. In this case, a naked full length mRNA is formed.

In contrast to Rho factor which must bind at transcription initiation or soon thereafter, ribosomes may bind at any time to the naked full length or partial length mRNA that is not bound by Rho factor and this will prevent RNases from binding and decrease the mRNA degradation rate. If Rho factor is bound no ribosome can bind to the mRNA even if mRNA transcription is completed [262]. Similarly, RNases may bind at any time to a nascent, partial length mRNA or full length mRNA that is not already bound by a ribosome.

The model assumes: (i) that RNase degradation is an active process and that in its absence mRNAs are simply removed by the dilution that accompanies cell growth or by slower mechanisms of RNA degradation (which is consistent with experimental observations in the system being modeled [262]); (ii) the leading ribosome primarily determines whether the mRNA is translated, terminated or rapidly degraded (which is based on experimental observations as described above) thereby the unmanageable complexity of a system with mRNAs having different number of ribosomes; (iii) that RNases are acting during transcription rather than after mRNA release (which is supported by experimental evidence [307]); and (iv) that Rho factor is the primary cause of transcription termination (which again was experimentally confirmed for this system [262]).

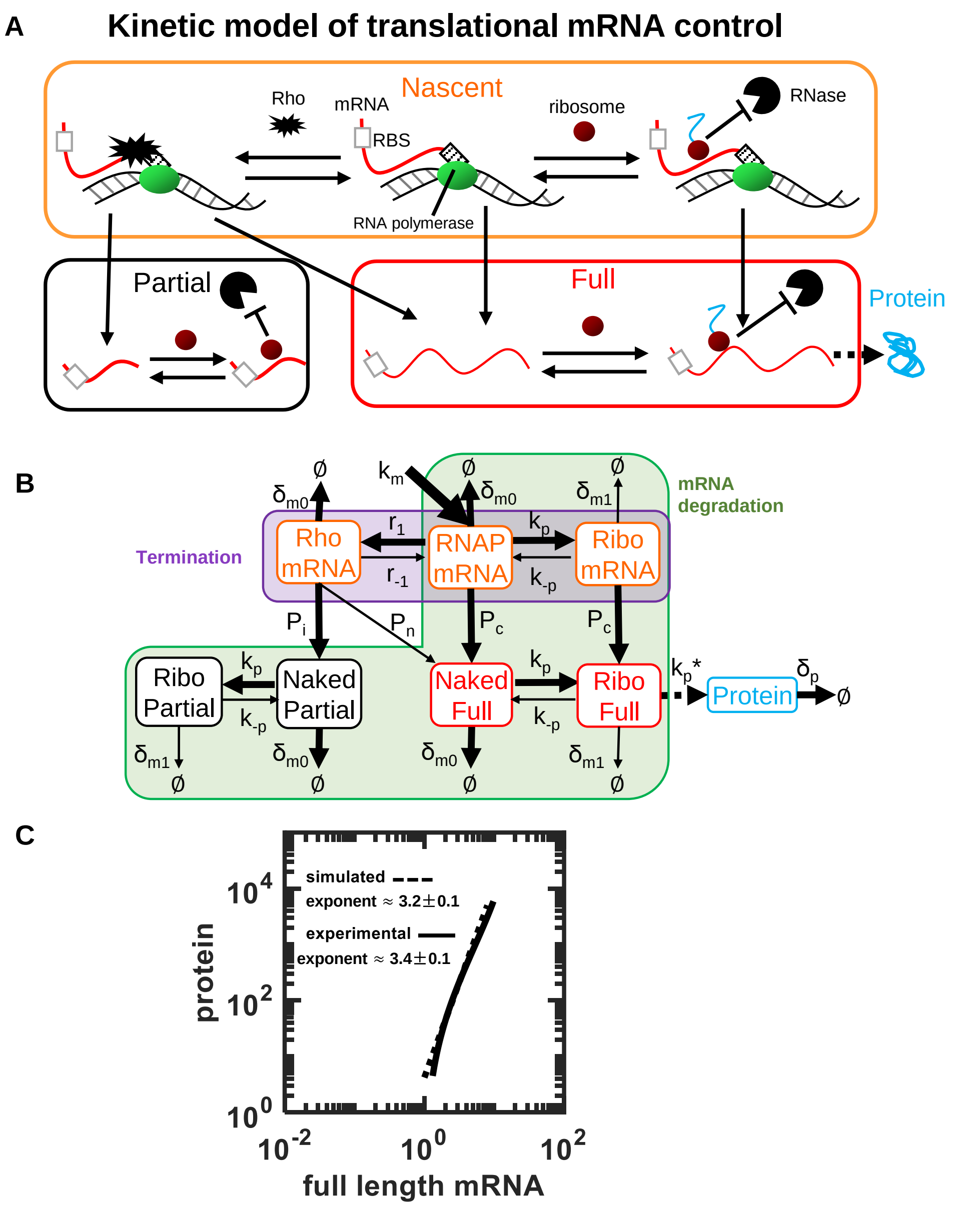


Figure 4.3. Kinetic model of translational mRNA control. (**A**) Reaction steps and complexes in the kinetic model of translational mRNA control. Nascent = nascent mRNA, Partial = partial length mRNA, and Full = full length mRNA. See main text for details. (**B**) Reaction schemer of the model in (**A**) with rate constants. Reaction species in the purple boxes indicate the competing actions between ribosome binding resulting in transcription and Rho factor binding resulting in transcription termination. Reaction species in the green boxes indicate the competing actions between ribosome binding resulting in transcription and RNase binding resulting in mRNA degradation. It should be noted that * indicates that k_p* is proportional to k_p. RNAP mRNA = naked nascent mRNA; Rho mRNA and Rho Partial= Rho factor bound nascent and partial length mRNAs; Ribo mRNA and Ribo Full = ribosome bound nascent mRNA and full length mRNA; Naked Full = neither Rho factor nor ribosome bound full length mRNA. Rate constants are defined in the **Materials and Methods**. (**C**) Simulation of the kinetic model of protein concentration as a function of full length mRNA at varying the translation initiation rates.

The kinetic model was simulated using parameters values reported for this system and in other published experiments. We first simulated the relationship between protein concentrations and full length mRNA concentration and the results were consistent with the previously observed power-law relationship (**Figure 4.2D**). The exponent of the simulated power-law was 3.2 (95% confidence interval ± 0.1), which is close to the experimentally measured exponent of 3.4 (95% confidence interval ± 0.1) (**Figure 4.3C**). Therefore the kinetic model also predicts the experimentally observed relationship between protein and full length mRNA concentrations.

4.3.3 The collaboration and counteraction of transcription-translation coupling and mRNA degradation-translation coupling

Having established that the kinetic model generates similar results to that of the simple model in the first part of the study and a key experimental observation (*i.e.* the power-law relationship with exponent 3.2), we next investigated features that can only be addressed by this model. In particular, whether the detailed kinetic model can also predict the effect of translational mRNA control on partial length mRNA concentrations; this would provide strong validation of the model as the relationship between the partial length mRNA concentration and translational mRNA control, which is not intuitive and not previously explained.

When the translation initiation rate is increased it causes both a decrease in transcription termination and a decrease in mRNA degradation which have complementary effects on full length mRNA; that is both mechanisms act to increase it (**Figure 4.4A**). In contrast, increasing translation has opposing effects on partial length mRNA. A greater number of ribosomes bound to the nascent length mRNAs could prevent transcription termination leading to a decrease in partial length mRNA concentration while this will also increase the number of ribosomes bound to the partial length mRNAs preventing degradation by RNases resulting in an increase in partial length mRNA concentration. This leads to a tug of war between two processes and the strongest mechanism will affect the net change (**Figure 4.4B**). That is, the partial length mRNA can provide indicator of the relative contributions of the two mechanisms (note: it has already been reported that RNase E acts on partial length mRNAs and full length mRNAs with similar efficacy [262]).

Our simulation result showed that in contrast to the full length mRNA which increased with translation the partial length mRNA concentration stayed relatively constant over a large range of translation rate k_p (**Figure 4.4C, D**). That is, it appeared that the effect of increasing translation on transcription termination and on mRNA degradation contributed equally to the change in partial length mRNA. These findings are consistent with the experimental data from the previous study (**Figure 4.4E, F**). This indicates a balance between anti-degradation and anti-termination in the generation of partial length mRNAs.

The above result indicates that the translational mRNA control does not conserve resources. A resource conservation mechanism would enable the effect of transcription termination to dominate the effect of mRNA degradation, especially in regard to partial length mRNAs.

However, as we have shown the system is not configured in that way and therefore a substantial amount of nucleotides are sequestered in these relatively stable mRNA fragments.

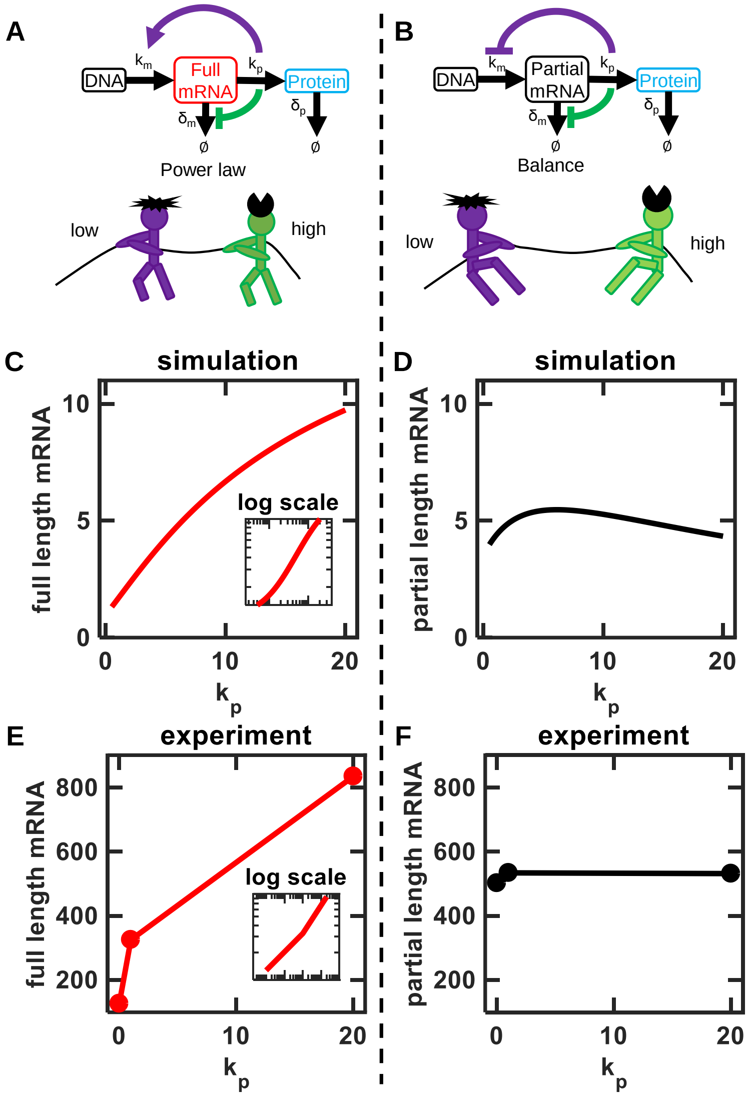


Figure 4.4. Simulation of the concentration of partial length mRNAs using the kinetic model of translational control. (**A, B**) Increasing translation increases the transcription of full length mRNA (by decreasing transcription termination) and decreases mRNA degradation leading to increased full length mRNA. Increasing translation decreases the production of partial length mRNA (by decreasing transcription termination) and decreases mRNA degradation leading to constant partial length mRNA. (**C, D**) The results of the simulation showing full length mRNA and partial length mRNA as a function of the translation initiation rate (k_p). The inset shows the log-log scale plot of the same curves. (**E, F**) Experimental reported in Hussein et al., 2015 for comparison with the simulation results in panels C and D.

4.4 Discussion

The first part of this study describes a simple mathematical model based on first principles and minimal assumptions that reproduces the experimentally observed power-law relationship, and shows that for an exponent to be consistent with experimental measurements therefore must be negative cooperativity; that is, the first ribosome has greater impact of translational mRNA control than trailing ribosomes. The greater impact of the leading ribosome on translational mRNA control is consistent with reported experimental observations [28, 108, 304-306]. This negative cooperativity (which was also observed in the kinetic model) is precisely the behavior needed for translational mRNA control mechanisms to act as a quality control mechanism [294]. Negative cooperativity means the leading ribosome can act as a “pilot” that evaluates the quality of the mRNA and translation and makes the decision as to whether the specific transcription event in which it is involved will generate mRNA structures or proteins that could be detrimental to the cell and therefore whether to terminate transcription and degrade the mRNA.

The second part describes a kinetic model and simulates the behavior of the system with parameters from reported experimental measurements. The kinetic model also generated a power-law relationship over a wide range of values and an exponent value that supports the presence of negative cooperativity for translation. It predicts that a consequence of the combination of having both mechanism of translation mRNA control (transcription-translation coupling and mRNA degradation-translation coupling) is that they can have opposing actions resulting in high levels of partial length mRNA fragments that are more or less constant at all levels of translation. The stability of high concentrations of partial mRNA, which are due to the very mechanism that created them, does not seem consistent with the explanation that translational mRNA control as a mechanism that is primarily for the conservation of resources. As mentioned earlier, cells have alternative mechanisms available to them that can prevent transcription initiation taking place rather than needing to terminate transcription after transcription has been initiated and degrading the mRNA.

The findings reveal that translational mRNA control can give rise to important regulatory properties and that these properties, rather than energy and resource conservation, may have contributed to its widespread and fundamental involvement in gene regulation.

4.5 Materials and Methods

4.5.1 The definition of exponent in a general function

The power-law relationship is characterized by a standard formula like $y = ax^b$ where x is the independent variable (*i.e.* mRNA), y is the dependent variable (*i.e.* protein), a is a scaling constant, and b is defined as the exponent. This definition of exponent, however, only applies to this particular form of power-law function. Therefore, we need to generalize the definition of exponent into a more generic form. To achieve this, we started by taking the logarithmic transformation on both side of the equation. This gives us:

$$\log y \quad = \log a + b\log x. \qquad \text{Eq. 4.10}$$

We noticed that the exponent b appears as the coefficient for the term $\log x$. To get rid of the unwanted constant $\log a$, we took the derivative with respect to $\log x$ on both sides of **Eq. 4.10** converting it to

$$\frac{d\log y}{d\log x} \quad = \frac{d\log a}{d\log x} + \frac{bd\log x}{d\log x}. \qquad \text{Eq. 4.11}$$

Here $\log a$ is independent of $\log x$ so the first term on the right hand side equals zero. The $\log x$ in the numerator and the denominator of the second term cancel each other. Therefore, we have:

$$\frac{d\log y}{d\log x} \quad = 0 + b = b. \qquad \text{Eq. 4.12}$$

Unlike the original definition of power, this formula doesn't rely on specific form of the equation. Thus, it can be applied to equations with more biological meaning including Hill functions. It's simply the slope of the function on the log-log scale. Following this definition, we can write the exponent of protein and mRNA (b) to be the slope of the protein as a function of mRNA on the log-log scale as

$$\text{exponent } (b) = \frac{d\log[\text{protein}]}{d\log[\text{mRNA}]}. \qquad \text{Eq. 4.13}$$

It should be noted that this measure of exponent varies with the parameters of the model. For example, due to the coupling of translation with transcription, the exponent might be a function of translation rate k_p.

4.5.2 The derivation of negative cooperativity from experimental data

Due to the limitation and lack of accurate quantitative measurements, the parameters of the Hill function in **Eq. 4.4** are not clear. To avoid overfitting with specific parameter values or functions, we summarized transcription-translation coupling and mRNA degradation-translation coupling together as one abstract function $f(k_p)$. For further analysis, we assumed that this function is a continuous and differentiable function of translation rate k_p. It's important to note that for a given translation rate k_p, there is only one corresponding mRNA concentration. Using this fact and substituting $f(k_p)$ into **Eq. 4.4**, we obtain a concise model of translational mRNA control as

$$\begin{aligned} [\mathrm{mRNA}]_{ss} &= \frac{k_m}{\delta_m} f(k_p) \\ [\mathrm{protein}]_{ss} &= \frac{k_p}{\delta_p}[\mathrm{mRNA}]_{ss} = \frac{k_p}{\delta_p}\frac{k_m}{\delta_m} f\big(k_p\big). \end{aligned} \qquad \text{Eq. 4.14}$$

Although there is no specific form of translational mRNA control, we could still derive the exponent of $f(k_p)$ from the exponent b of protein-mRNA relationship as in **Eq. 4.13**. To do this, we substituted **Eq. 4.14** into **Eq. 4.13** and simplified the equation, which leads to a formula directly linking k_p to b as:

$$b = \frac{d\log\Big(\frac{k_p k_m}{\delta_p \delta_m} f(k_p)\Big)}{d\log\Big(\frac{k_m}{\delta_m} f(k_p)\Big)} = \frac{1}{\frac{d\log(f(k_p))}{d\log(k_p)}} + 1. \qquad \text{Eq. 4.15}$$

In the last step of **Eq. 4.15**, we immediately noticed that the denominator of the first term has the same form as **Eq. 4.13**. This means the exponent of $f(k_p)$ (n_f) with respect to k_p can be expressed as

$$n_f = \frac{d\log(f(k_p))}{d\log(k_p)}. \qquad \text{Eq. 4.16}$$

Plugging **Eq. 4.16** into **Eq. 4.15** we arrived at a simple formula:

$$b = \frac{1}{n_f} + 1. \qquad \text{Eq. 4.17}$$

This formula draws a connection between translational control of mRNA concentration as a general function with the observed power-law relationship between mRNA and protein. From the published data, we know translation has a positive impact on mRNA concentration and the experimental data shows a value of exponent b around 3.4. This allows us to calculate the power of the translational mRNA control as $n_f = 0.4167$. In a more general form, because the measured power is not accurate, we can also get a looser bound for n_f based on the fact that the observed power is greater than 3. That is, $n_f < 0.5$.

To show the connection between this conclusion of a general function to that of a real biological system, we would like to bring back the Hill functions in **Figure 4.2C**. We modeled the translational mRNA control as in **Figure 4.2C** where we had

$$f(k_p) = k_m \frac{k_p^n}{k_p^n + K^n}. \qquad \text{Eq. 4.18}$$

It's straight-forward to compute the exponent (n_f) of **Eq. 4.18** as

$$n_f = n \frac{K^n}{k_p^n + K^n} \qquad \text{Eq. 4.19}$$

Eq. 4.19 connects the Hill coefficient k and the empirical exponent n_f. If n and K are constants, then n_f is actually a function of k_p. When $k_p = 0$, $n_f = n$; when $k_p = K_k$, $n_f = 0.5n$; when $k_p \to \infty$, $n_f \to 0$. Therefore, $n_f \in (0, n]$. This bound need to be smaller than the upper bound for n_f from above for the exponent n to hold for a broad range of translation rate k_p. That is, $n \in (0,0.5]$. This is the same as the bound we got with Hill function.

4.5.3 Comparing the outcome of different cooperativity

To compare translational mRNA control with different cooperativity, we plotted different cooperativity together as comparisons with the experimentally observed power-law relationship in **Figure 4.2D**. Fitting the data with a regular regression with a power-law function or Hill function will always give a negative cooperativity. Therefore, instead of fitting the model to data, we adjusted the parameters so different functions with different cooperativities always cross the same fixed point. This is done by first calculating the average mRNA level and the corresponding predicted protein level by the fitted power-law function in **Eq. 4.5**. The result gives a fixed point that is on all curves. Next, we set the parameters and the Hill coefficient (cooperativity) leaving only the degradation rate of the protein δ_p unset. In the last step, we computed the required δ_p for each model so that they all cross the fixed point of average mRNA level in the experimental data and plotted the final curves. It should be noted that we didn't specifically choose δ_p to be the last variable to be adjusted. Tuning other parameters will give similar results. The parameter values of **Eq. 4.4** we used are: $k_m = 1$ a.u.·s^{-1}, $\delta_m = 0.1$ s^{-1}, $K = 100$ s^{-1}, $k_p \in [10^{-2}, 10^5]$ s^{-1}, $\delta_p = 14.8164, 1.3812, 125.3703$ s^{-1} for negative, no, and positive cooperativity respectively. Here δ_p values were solved for different cooperativities with the procedure described above by solve function in Matlab (R2016b, Mathworks). These values are not realistic because the unit used in the experimental data is arbitrary.

The procedure described above is also implemented in **Figure 4.3C** where we compared the simulated power-law relationship to the experimentally observed power. In this case, a power law function like $P = aM^{3.4}$ is used to represent the power-law relationship from experiment. Again, the average mRNA and its corresponding protein level on the simulated curve is used as a fixed point to calculate the value of a. We did not use the original fitted power-law function in **Figure 4.2D** because experimental measurements have very different scale in the values and they are not comparable to the simulated values.

4.5.4 The detailed kinetic model of the translational mRNA control

We developed a mathematical model of central dogma including six key processes: transcription initiation, transcription termination, mRNA degradation, translation initiation, translation completion, and protein degradation. In this model, we focused on the status of mRNAs which fall into three big categories: nascent mRNA, full length mRNA, and partial length mRNA.

Transcription initiation occurs with a rate of k_m ([mRNA]·s^{-1}) and gives rise to the initial stage of mRNA marked as $[\mathrm{RNAP} \cdot \mathrm{mRNA}]$. In the following step, Rho dependent transcription termination competes with translation initiation with a rate of r_1 (s^{-1}) and k_p (s^{-1}) respectively [308]. If the ribosome succeed, Rho factor cannot be activated so no partial length mRNA can be generated; if Rho factor reaches the site first without hindering ribosomes on the way, mRNA will be terminated [309]. In addition, it is possible that transcription might finish via intrinsic termination before ribosome or Rho factor binding occurs. Therefore, we built in a step where $[\mathrm{RNAP} \cdot \mathrm{mRNA}]$ is converted to naked full length mRNA ($[\mathrm{Naked} \cdot \mathrm{Full}]$) at a rate of P_c (s^{-1}). The naked full length mRNA will not stay long before ribosomes bind to it forming polysomes ($[\mathrm{Ribo} \cdot \mathrm{Full}]$).

On the other hand, once Rho factor binds to the nascent mRNA, it quickly stops the transcription turning the new transcript into partial length mRNA ($[\mathrm{Naked} \cdot \mathrm{Partial}]$) at a rate of P_i (s^{-1}) preventing the unused transcription from being made [310]. This phenomenon is also known as premature termination [311]. Meanwhile, the Rho bound mRNA has a chance, although very unlikely, to reach the end of the gene without premature termination so we modeled this process with a small rate of P_n (s^{-1}). This creates a narrow corridor for Rho bound mRNA to be correctly translated. After premature termination, sometimes the ribosome binding site (RBS) on the partial length mRNA is still available. As a consequence, ribosome can still bind to the naked partial length mRNA producing ribosome bound partial length mRNA. However, no protein can be produced in this process and ribosomes on the partial length mRNA need to be rescued [312].

In the presence of ribosomes, mRNAs are covered and become inaccessible for RNases which degrades mRNAs [313]. This means, three ribosome bound states of mRNA in our model ($[\mathrm{Ribo} \cdot \mathrm{mRNA}]$, $[\mathrm{Ribo} \cdot \mathrm{Full}]$, $[\mathrm{Ribo} \cdot \mathrm{Partial}]$) should have a lower degradation rate δ_{m1} than the normal degradation rate δ_{m0} of the naked mRNAs ($[\mathrm{RNAP} \cdot \mathrm{mRNA}]$, $[\mathrm{Rho} \cdot \mathrm{mRNA}]$, $[\mathrm{Naked} \cdot \mathrm{Partial}]$, $[\mathrm{Naked} \cdot \mathrm{Full}]$). Moreover, this protection by ribosomes applies to both nascent mRNA, partial length mRNA, and full length mRNA. That is, both $[\mathrm{Ribo} \cdot \mathrm{Full}]$ and $[\mathrm{Ribo} \cdot \mathrm{Partial}]$ have lower degradation rates.

Finally, proteins are produced from the translating mRNA (polysomes) at a rate of k_p^* ([protein] · [mRNA]$^{-1}$·s^{-1}). The rate constant of this process is assumed to be related to the RBS as well. This is because the rate of protein production is determined by the spacing between ribosomes on the translating mRNAs and this spacing is predetermined by the wait time of ribosomes binding at the RBS. In our simplified model, multiple ribosome binding events are compressed into a single step. Thus, the direct influence of RBS on protein production is required to be factored in independently. In this paper, we assume k_p^* has the same value as k_p but a more general form expressing k_p^* as a function of k_p can also be used.

At the last step, we have a constant rate of protein degradation at δ_p. Two notable quantities that can be extracted from the model. The concentration of full length mRNA $[\text{Full} \cdot \text{mRNA}]$ is given by $[\text{Naked} \cdot \text{Full}] + [\text{Ribo} \cdot \text{Full}]$ and the concentration of partial length mRNA $[\text{Partial} \cdot \text{mRNA}]$ is given by $[\text{Naked} \cdot \text{Partial}] + [\text{Ribo} \cdot \text{Partial}]$. The equations of the model are listed below:

$$
\begin{aligned}
\frac{d[\text{RNAP}\cdot\text{mRNA}]}{dt} = \quad & k_m - k_p[\text{RNAP} \cdot \text{mRNA}] + k_{-p}[\text{Ribo} \cdot \text{mRNA}] \\
& -r_1[\text{RNAP} \cdot \text{mRNA}] + r_{-1}[\text{Rho} \cdot \text{mRNA}] \\
& -\delta_{m0}[\text{RNAP} \cdot \text{mRNA}] - P_c[\text{RNAP} \cdot \text{mRNA}] \\
\frac{d[\text{Ribo}\cdot\text{mRNA}]}{dt} = \quad & k_p[\text{RNAP} \cdot \text{mRNA}] - k_{-p}[\text{Ribo} \cdot \text{mRNA}] \\
& -\delta_{m1}[\text{Ribo} \cdot \text{mRNA}] - P_c[\text{Ribo} \cdot \text{mRNA}] \\
\frac{d[\text{Rho}\cdot\text{mRNA}]}{dt} = \quad & r_1[\text{RNAP} \cdot \text{mRNA}] - r_{-1}[\text{Rho} \cdot \text{mRNA}] \\
& -\delta_{m0}[\text{Rho} \cdot \text{mRNA}] - (P_i + P_n)[\text{Rho} \cdot \text{mRNA}] \\
\frac{d[\text{Naked}\cdot\text{Partial}]}{dt} = \quad & P_i[\text{Rho} \cdot \text{mRNA}] - \delta_{m0}[\text{Naked} \cdot \text{Partial}] \\
& -k_p[\text{Naked} \cdot \text{Partial}] + k_{-p}[\text{Ribo} \cdot \text{Partial}] \qquad . \qquad \text{Eq. 4.20} \\
\frac{d[\text{Ribo}\cdot\text{Partial}]}{dt} = \quad & P_i[\text{Rho} \cdot \text{mRNA}] - \delta_{m0}[\text{Naked} \cdot \text{Partial}] \\
& -k_p[\text{Naked} \cdot \text{Partial}] + k_{-p}[\text{Ribo} \cdot \text{Partial}] \\
\frac{d[\text{Naked}\cdot\text{Full}]}{dt} = \quad & P_n[\text{Rho} \cdot \text{mRNA}] + P_c[\text{RNAP} \cdot \text{mRNA}] - \delta_{m0}[\text{Naked} \cdot \text{Partial}] \\
& -k_p[\text{Naked} \cdot \text{Full}] + k_{-p}[\text{Ribo} \cdot \text{Full}] \\
\frac{d[\text{Ribo}\cdot\text{Full}]}{dt} = \quad & P_c[\text{Ribo} \cdot \text{mRNA}] - \delta_{m1}[\text{Ribo} \cdot \text{Full}] \\
& +k_p[\text{Naked} \cdot \text{Full}] - k_{-p}[\text{Ribo} \cdot \text{Full}] \\
\frac{d[Protein]}{dt} = \quad & k_p{}'[\text{Ribo} \cdot \text{Full}] - \delta_p[Protein]
\end{aligned}
$$

The kinetic parameters used in the models are: $k_m = 1\text{nM} \cdot \text{min}^{-1}$, $\delta_{m0} = 0.16\text{min}^{-1}$, $\delta_{m1} = 0.04\text{min}^{-1}$, $r_1 = 2.5\text{min}^{-1}$, $r_{-1} = 4\text{min}^{-1}$, $k_p, k_p{}' \in (0,20]\text{min}^{-1}$, $k_{-p} = 2\text{min}^{-1}$, $P_i = 1\text{min}^{-1}$, $P_n = 0.07\text{min}^{-1}$, $P_c = 0.1\text{min}^{-1}$, $\delta_p = 0.03\text{min}^{-1}$. Among these parameters, k_m and δ_p were previously reported. δ_{m0} is derived from the literature with an additional dilution rate of 0.02min^{-1} added to the original value [230]. Other parameters were set in this study. The steady state solutions were solved by the solve function in Matlab (R2016b, Mathworks).

Conclusions

“The Central Dogma. This states that once 'information' has passed into protein it cannot get out again. In more detail, the transfer of information from nucleic acid to nucleic acid, or from nucleic acid to protein may be possible, but transfer from protein to protein, or from protein to nucleic acid is impossible. Information means here the precise determination of sequence, either of bases in the nucleic acid or of amino acid residues in the protein.” Francis Crick (1957) [314].

In the more than 60 years since the central dogma was described we have learned an immense amount about the control of the flow of information in the cell, including roles for DNA, RNA and protein that were unimaginable when the central dogma was first proposed. And there is much that is still unknown.

In this thesis I have investigated some of the most fundamental questions in biology. Where do mRNAs and the small non-coding RNAs that regulate them occur in cells (Chapter 3)? How does translation alter mRNA transcription and degradation (Chapter 4)? In the process of pursuing answers to these questions we also developed general tools that will benefit other researchers, and allow them to not only build on our findings but also to pursue an unlimited number of other questions. We have created new metrics for quantifying whether biological molecules (such as RNAs and proteins) that are involved in the regulation of cellular processes colocalize in the same parts the cell, and therefore may have regulatory interactions (Chapter 1). In addition, we have developed software tools that will enable researchers from diverse areas of biology to easily visualize and measure the contribution of spatial organization of molecules in cells to better understand their contribution to the regulation of translation and other processes (Chapter 2).

This dissertation has revealed further complexity to translational regulation and that it is more pervasive and more important than we had thought. Translation should never simply be considered as a process that begins after the mRNA is generated and a process for simply converting the mRNA code into the amino code of proteins. We have shown that translation regulates the production and the degradation of its own template (*i.e.* mRNA). We have shown that the specialized non-coding RNAs that have evolved to regulate translation have the capacity to enter into the nucleoid to regulate transcription and translation at its very earliest stages, as well as localizing in the cytoplasm and at the membrane to participate in the translation and degradation of mRNAs at those locations.

The findings in this dissertation not only contribute to our understanding of some of the most basic principles of gene regulation, but they also have broad potential applications for the analysis of the regulation of specific genes and biochemical pathways and for the design of engineered genetic circuits.

References

1. Velichkova, M., et al., *Drosophila Mtm and class II PI3K coregulate a PI(3)P pool with cortical and endolysosomal functions.* Journal of Cell Biology, 2010. **190**(3): p. 407-425.
2. George, C.M., et al., *Condensins are Required for Maintenance of Nuclear Architecture.* Cells, 2014. **3**(3): p. 865-82.
3. Sheng, H., et al., *Nucleoid and cytoplasmic localization of small RNAs in Escherichia coli.* Nucleic Acids Res, 2017.
4. Wahlby, C., et al., *An image analysis toolbox for high-throughput C. elegans assays.* Nat Methods, 2012. **9**(7): p. 714-6.
5. Campbell, E.A., et al., *Structure of the bacterial RNA polymerase promoter specificity sigma subunit.* Mol Cell, 2002. **9**(3): p. 527-39.
6. Wosten, M.M., *Eubacterial sigma-factors.* FEMS Microbiol Rev, 1998. **22**(3): p. 127-50.
7. Lodish, H., et al., *Molecular cell biology*. 2008: Macmillan.
8. Wang, F., et al., *The promoter-search mechanism of Escherichia coli RNA polymerase is dominated by three-dimensional diffusion.* Nature structural & molecular biology, 2013. **20**(2): p. 174.
9. Ricchetti, M., W. Metzger, and H. Heumann, *One-dimensional diffusion of Escherichia coli DNA-dependent RNA polymerase: a mechanism to facilitate promoter location.* Proceedings of the National Academy of Sciences, 1988. **85**(13): p. 4610-4614.
10. Maeda, H., N. Fujita, and A. Ishihama, *Competition among seven Escherichia coli sigma subunits: relative binding affinities to the core RNA polymerase.* Nucleic Acids Res, 2000. **28**(18): p. 3497-503.
11. Paget, M., *Bacterial sigma factors and anti-sigma factors: structure, function and distribution.* Biomolecules, 2015. **5**(3): p. 1245-1265.
12. Murakami, K.S., S. Masuda, and S.A. Darst, *Structural basis of transcription initiation: RNA polymerase holoenzyme at 4 A resolution.* Science, 2002. **296**(5571): p. 1280-4.
13. Ruff, E., M. Record, and I. Artsimovitch, *Initial events in bacterial transcription initiation.* Biomolecules, 2015. **5**(2): p. 1035-1062.
14. Nudler, E., *Transcription elongation: structural basis and mechanisms.* Journal of molecular biology, 1999. **288**(1): p. 1-12.
15. Chakraborty, A., et al., *Opening and closing of the bacterial RNA polymerase clamp.* Science, 2012. **337**(6094): p. 591-595.
16. Murakami, K., *Structural biology of bacterial RNA polymerase.* Biomolecules, 2015. **5**(2): p. 848-864.
17. Nudler, E., *Transcription elongation: structural basis and mechanisms.* J Mol Biol, 1999. **288**(1): p. 1-12.
18. Gusarov, I. and E. Nudler, *The mechanism of intrinsic transcription termination.* Mol Cell, 1999. **3**(4): p. 495-504.
19. Nudler, E., *Spatial Organization of Transcription Elongation Complex in Escherichia coli.* Science, 1998. **281**(5375): p. 424-428.
20. Bar-Nahum, G., et al., *A ratchet mechanism of transcription elongation and its control.* Cell, 2005. **120**(2): p. 183-93.

21. Nudler, E., *RNA polymerase backtracking in gene regulation and genome instability.* Cell, 2012. **149**(7): p. 1438-45.
22. Chamberlin, K.M.A.a.M.J., *Transcription Termination in Escherichia coli: Measurement of the Rate of Enzyme Release From Rho-independent Terminators.* J. Mol. Biol., 1988. **202**: p. 271-285.
23. Ciampi, M.S., *Rho-dependent terminators and transcription termination.* Microbiology, 2006. **152**(Pt 9): p. 2515-28.
24. Nudler, E. and M.E. Gottesman, *Transcription termination and anti-termination in E. coli.* Genes Cells, 2002. **7**(8): p. 755-68.
25. Chamberlin, K.M.A.a.M.J., *RNA Chain Elongation by Escherichia coli RNA Polymerase.* J. Mol. Biol., 1990. **213**: p. 79-108.
26. Erie, D.A., *The many conformational states of RNA polymerase elongation complexes and their roles in the regulation of transcription.* Biochim Biophys Acta, 2002. **1577**(2): p. 224-39.
27. Tagami, S., S.I. Sekine, and S. Yokoyama, *A novel conformation of RNA polymerase sheds light on the mechanism of transcription.* Transcription, 2011. **2**(4): p. 162-167.
28. Proshkin, S., et al., *Cooperation between translating ribosomes and RNA polymerase in transcription elongation.* Science, 2010. **328**(5977): p. 504-8.
29. Koslover, D.J., et al., *Binding and translocation of termination factor rho studied at the single-molecule level.* J Mol Biol, 2012. **423**(5): p. 664-76.
30. Hart, C.M. and J.W. Roberts, *Rho-dependent transcription termination. Characterization of the requirement for cytidine in the nascent transcript.* J Biol Chem, 1991. **266**(35): p. 24140-8.
31. Richardson, J.P., *Rho-dependent Termination of Transcription Is Governed Primarily by the Upstream Rho Utilization (rut) Sequences of a Terminator.* Journal of Biological Chemistry, 1996. **271**(35): p. 21597-21603.
32. Richardson, J.P., *Rho-dependent termination and ATPases in transcript termination.* Biochim Biophys Acta, 2002. **1577**(2): p. 251-260.
33. Roberts, J.W., *Termination factor for RNA synthesis.* Nature, 1969. **224**(5225): p. 1168-74.
34. Xiong Yu, T.H., Katsuya Shigesada, and Edward H. Egelman, *Three-dimensional Reconstruction of Transcription Termination factor Rho: Orientation of the N-terminal Domain and Visualization of an RNA-binding site.* JMB, 2000. **299**: p. 1279-1287.
35. Gardner, P.P., et al., *RNIE: genome-wide prediction of bacterial intrinsic terminators.* Nucleic acids research, 2011. **39**(14): p. 5845-5852.
36. Leela, J.K., et al., *Rho-dependent transcription termination is essential to prevent excessive genome-wide R-loops in Escherichia coli.* Proceedings of the National Academy of Sciences, 2013. **110**(1): p. 258-263.
37. Lesnik, E.A., et al., *Prediction of rho-independent transcriptional terminators in Escherichia coli.* Nucleic acids research, 2001. **29**(17): p. 3583-3594.
38. Hollands, K., et al., *Riboswitch control of Rho-dependent transcription termination.* Proceedings of the National Academy of Sciences, 2012. **109**(14): p. 5376-5381.
39. Ciampi, M.S., *Rho-dependent terminators and transcription termination.* Microbiology, 2006. **152**(9): p. 2515-2528.

40. Hui, M.P., P.L. Foley, and J.G. Belasco, *Messenger RNA degradation in bacterial cells.* Annual review of genetics, 2014. **48**: p. 537-559.

41. Deana, A., H. Celesnik, and J.G. Belasco, *The bacterial enzyme RppH triggers messenger RNA degradation by 5′ pyrophosphate removal.* Nature, 2008. **451**(7176): p. 355.

42. Deana, A., H. Celesnik, and J.G. Belasco, *The bacterial enzyme RppH triggers messenger RNA degradation by 5' pyrophosphate removal.* Nature, 2008. **451**(7176): p. 355-8.

43. Mackie, G.A., *Ribonuclease E is a 5′-end-dependent endonuclease.* Nature, 1998. **395**(6703): p. 720.

44. Callaghan, A.J., et al., *Structure of Escherichia coli RNase E catalytic domain and implications for RNA turnover.* Nature, 2005. **437**(7062): p. 1187-91.

45. Celesnik, H., A. Deana, and J.G. Belasco, *Initiation of RNA decay in Escherichia coli by 5' pyrophosphate removal.* Mol Cell, 2007. **27**(1): p. 79-90.

46. Mackie, G.A., *Ribonuclease E is a 5'-end-dependent endonuclease.* Nature, 1998. **395**(6703): p. 720-3.

47. Jiang, X., A. Diwa, and J.G. Belasco, *Regions of RNase E important for 5′-end-dependent RNA cleavage and autoregulated synthesis.* Journal of bacteriology, 2000. **182**(9): p. 2468-2475.

48. McDOWALL, K.J., et al., *The ams-1 and rne-3071 temperature-sensitive mutations in the ams gene are in close proximity to each other and cause substitutions within a domain that resembles a product of the Escherichia coli mre locus.* Journal of bacteriology, 1993. **175**(13): p. 4245-4249.

49. Tock, M.R., et al., *The CafA protein required for the 5′-maturation of 16 S rRNA is a 5′-end-dependent ribonuclease that has context-dependent broad sequence specificity.* Journal of Biological Chemistry, 2000. **275**(12): p. 8726-8732.

50. Deana, A. and J.G. Belasco, *The function of RNase G in Escherichia coli is constrained by its amino and carboxyl termini.* Molecular microbiology, 2004. **51**(4): p. 1205-1217.

51. Luciano, D.J., et al., *A novel RNA phosphorylation state enables 5′ end-dependent degradation in Escherichia coli.* Molecular cell, 2017. **67**(1): p. 44-54. e6.

52. Luciano, D.J., et al., *Differential control of the rate of 5′-end-dependent mRNA degradation in Escherichia coli.* Journal of bacteriology, 2012. **194**(22): p. 6233-6239.

53. Clarke, J.E., et al., *Direct entry by RNase E is a major pathway for the degradation and processing of RNA in Escherichia coli.* Nucleic acids research, 2014. **42**(18): p. 11733-11751.

54. McDowall, K.J., S. Lin-Chao, and S.N. Cohen, *A+ U content rather than a particular nucleotide order determines the specificity of RNase E cleavage.* Journal of Biological Chemistry, 1994. **269**(14): p. 10790-10796.

55. Mackie, G.A. and J.L. Genereaux, *The role of RNA structure in determining RNase E-dependent cleavage sites in the mRNA for ribosomal protein S20 in vitro.* J Mol Biol, 1993. **234**(4): p. 998-1012.

56. McDowall, K.J., et al., *Site-specific RNase E cleavage of oligonucleotides and inhibition by stem-loops.* Nature, 1995. **374**(6519): p. 287-90.

57. Baker, K.E. and G.A. Mackie, *Ectopic RNase E sites promote bypass of 5'-end-dependent mRNA decay in Escherichia coli.* Mol Microbiol, 2003. **47**(1): p. 75-88.

58. Mohanty, B.K. and S.R. Kushner, *Polynucleotide phosphorylase functions both as a 3' right-arrow 5' exonuclease and a poly(A) polymerase in Escherichia coli.* Proc Natl Acad Sci U S A, 2000. **97**(22): p. 11966-71.
59. Zilhao, R., et al., *PNPase modulates RNase II expression in Escherichia coli: implications for mRNA decay and cell metabolism.* Mol Microbiol, 1996. **20**(5): p. 1033-42.
60. Coburn, G.A. and G.A. Mackie, *Reconstitution of the degradation of the mRNA for ribosomal protein S20 with purified enzymes.* J Mol Biol, 1998. **279**(5): p. 1061-74.
61. Liou, G.G., et al., *DEAD box RhlB RNA helicase physically associates with exoribonuclease PNPase to degrade double-stranded RNA independent of the degradosome-assembling region of RNase E.* J Biol Chem, 2002. **277**(43): p. 41157-62.
62. Steege, D.A., *Emerging features of mRNA decay in bacteria.* RNA, 2000. **6**(8): p. 1079-90.
63. Ghosh, S. and M.P. Deutscher, *Oligoribonuclease is an essential component of the mRNA decay pathway.* Proc Natl Acad Sci U S A, 1999. **96**(8): p. 4372-7.
64. Kushner, S.R., *mRNA Decay in Escherichia coli Comes of Age.* Journal of Bacteriology, 2002. **184**(17): p. 4658-4665.
65. Mackie, G.A., *Specific endonucleolytic cleavage of the mRNA for ribosomal protein S20 of Escherichia coli requires the product of the ams gene in vivo and in vitro.* J Bacteriol, 1991. **173**(8): p. 2488-97.
66. Cheng, Z.F. and M.P. Deutscher, *An important role for RNase R in mRNA decay.* Mol Cell, 2005. **17**(2): p. 313-8.
67. Cheng, Z.-F. and M.P. Deutscher, *An important role for RNase R in mRNA decay.* Molecular cell, 2005. **17**(2): p. 313-318.
68. Venkataraman, K., M. Garcia-Diaz, and W. Karzai, *Non-stop mRNA decay: a special attribute of trans-translation mediated ribosome rescue.* Frontiers in microbiology, 2014. **5**: p. 93.
69. Li, Z. and M.P. Deutscher, *Maturation pathways for E. coli tRNA precursors: a random multienzyme process in vivo.* Cell, 1996. **86**(3): p. 503-12.
70. Drlica, K. and J. Rouviere-Yaniv, *Histonelike proteins of bacteria.* Microbiological reviews, 1987. **51**(3): p. 301.
71. Cooper, S. and C.E. Helmstetter, *Chromosome replication and the division cycle of Escherichia coli B/r.* J Mol Biol, 1968. **31**(3): p. 519-40.
72. Block, D.H., et al., *Regulatory consequences of gene translocation in bacteria.* Nucleic Acids Res, 2012. **40**(18): p. 8979-92.
73. Chandler, M.G. and R.H. Pritchard, *The effect of gene concentration and relative gene dosage on gene output in Escherichia coli.* Mol Gen Genet, 1975. **138**(2): p. 127-41.
74. Sousa, C., V. de Lorenzo, and A. Cebolla, *Modulation of gene expression through chromosomal positioning in Escherichia coli.* Microbiology, 1997. **143 (Pt 6)**: p. 2071-8.
75. Bryant, J.A., et al., *Chromosome position effects on gene expression in Escherichia coli K-12.* Nucleic acids research, 2014. **42**(18): p. 11383-11392.
76. Browning, D.F. and S.J. Busby, *Local and global regulation of transcription initiation in bacteria.* Nature Reviews Microbiology, 2016. **14**(10): p. 638.
77. Liang, L.W., et al., *Minimal effect of gene clustering on expression in Escherichia coli.* Genetics, 2013. **193**(2): p. 453-65.

78. Achaz, G., et al., *Associations between inverted repeats and the structural evolution of bacterial genomes.* Genetics, 2003. **164**(4): p. 1279-1289.

79. Ohno, S. *So much'junk'DNA in our genome*. in *Evolution of Genetic Systems, Brookhaven Symp. Biol.* 1972.

80. Salgado, H., et al., *Operons in Escherichia coli: genomic analyses and predictions.* Proceedings of the National Academy of Sciences, 2000. **97**(12): p. 6652-6657.

81. Hussein, R., T.Y. Lee, and H.N. Lim, *Quantitative characterization of gene regulation by Rho dependent transcription termination.* Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms, 2015. **1849**(8): p. 940-954.

82. Levin-Karp, A., et al., *Quantifying translational coupling in E. coli synthetic operons using RBS modulation and fluorescent reporters.* ACS synthetic biology, 2013. **2**(6): p. 327-336.

83. Abe, H., T. Abo, and H. Aiba, *Regulation of intrinsic terminator by translation in Escherichia coli: transcription termination at a distance downstream.* Genes to Cells, 1999. **4**(2): p. 87-97.

84. Wright, J.J. and R.S. Hayward, *Transcriptional termination at a fully rho-independent site in Escherichia coli is prevented by uninterrupted translation of the nascent RNA.* The EMBO journal, 1987. **6**(4): p. 1115-1119.

85. Deana, A. and J.G. Belasco, *Lost in translation: the influence of ribosomes on bacterial mRNA decay.* Genes & development, 2005. **19**(21): p. 2526-2533.

86. Gottesman, S. and G. Storz, *Bacterial small RNA regulators: versatile roles and rapidly evolving variations.* Cold Spring Harb Perspect Biol, 2011. **3**(12).

87. Gottesman, S., et al., *Small RNA regulators and the bacterial response to stress.* Cold Spring Harb Symp Quant Biol, 2006. **71**: p. 1-11.

88. Vanderpool, C.K., *Physiological consequences of small RNA-mediated regulation of glucose-phosphate stress.* Curr Opin Microbiol, 2007. **10**(2): p. 146-51.

89. Masse, E., et al., *Small RNAs controlling iron metabolism.* Curr Opin Microbiol, 2007. **10**(2): p. 140-5.

90. Wassarman, K.M., *Small RNAs in bacteria: diverse regulators of gene expression in response to environmental changes.* Cell, 2002. **109**(2): p. 141-4.

91. Wassarman, K.M., et al., *Identification of novel small RNAs using comparative genomics and microarrays.* Genes Dev, 2001. **15**(13): p. 1637-51.

92. Gottesman, S., *The small RNA regulators of Escherichia coli: roles and mechanisms*.* Annu Rev Microbiol, 2004. **58**: p. 303-28.

93. Storz, G., J. Vogel, and K.M. Wassarman, *Regulation by small RNAs in bacteria: expanding frontiers.* Mol Cell, 2011. **43**(6): p. 880-91.

94. Fröhlich, K.S. and J. Vogel, *Activation of gene expression by small RNA.* Current opinion in microbiology, 2009. **12**(6): p. 674-682.

95. Massé, E., F.E. Escorcia, and S. Gottesman, *Coupled degradation of a small regulatory RNA and its mRNA targets in Escherichia coli.* Genes & development, 2003. **17**(19): p. 2374-2383.

96. Kawamoto, H., et al., *Base-pairing requirement for RNA silencing by a bacterial small RNA and acceleration of duplex formation by Hfq.* Mol Microbiol, 2006. **61**(4): p. 1013-22.

97. Møller, T., et al., *Hfq: a bacterial Sm-like protein that mediates RNA-RNA interaction.* Molecular cell, 2002. **9**(1): p. 23-30.

98. Storz, G., J.A. Opdyke, and K.M. Wassarman, *Regulating bacterial transcription with small RNAs.* Cold Spring Harb Symp Quant Biol, 2006. **71**: p. 269-73.

99. Barrick, J.E., et al., *6S RNA is a widespread regulator of eubacterial RNA polymerase that resembles an open promoter.* RNA, 2005. **11**(5): p. 774-84.

100. Sedlyarova, N., et al., *sRNA-mediated control of transcription termination in E. coli.* Cell, 2016. **167**(1): p. 111-121. e13.

101. Brantl, S. and E.G. Wagner, *Antisense RNA-mediated transcriptional attenuation: an in vitro study of plasmid pT181.* Mol Microbiol, 2000. **35**(6): p. 1469-82.

102. Aiba, H., *Mechanism of RNA silencing by Hfq-binding small RNAs.* Curr Opin Microbiol, 2007. **10**(2): p. 134-9.

103. Repoila, F. and F. Darfeuille, *Small regulatory non-coding RNAs in bacteria: physiology and mechanistic aspects.* Biol Cell, 2009. **101**(2): p. 117-31.

104. Storz, G., J.A. Opdyke, and A. Zhang, *Controlling mRNA stability and translation with small, noncoding RNAs.* Curr Opin Microbiol, 2004. **7**(2): p. 140-4.

105. Urban, J.H. and J. Vogel, *Translational control and target recognition by Escherichia coli small RNAs in vivo.* Nucleic Acids Res, 2007. **35**(3): p. 1018-37.

106. Nevo-Dinur, K., et al., *Translation-independent localization of mRNA in E. coli.* Science, 2011. **331**(6020): p. 1081-4.

107. St Johnston, D., *Moving messages: the intracellular localization of mRNAs.* Nat Rev Mol Cell Biol, 2005. **6**(5): p. 363-75.

108. Iost, I. and M. Dreyfus, *The stability of Escherichia coli lacZ mRNA depends upon the simultaneity of its synthesis and translation.* EMBO J, 1995. **14**(13): p. 3252-61.

109. Montero Llopis, P., et al., *Spatial organization of the flow of genetic information in bacteria.* Nature, 2010. **466**(7302): p. 77-81.

110. Kawamoto, H., et al., *Implication of membrane localization of target mRNA in the action of a small RNA: mechanism of post-transcriptional regulation of glucose transporter in Escherichia coli.* Genes Dev, 2005. **19**(3): p. 328-38.

111. Bakshi, S., H. Choi, and J.C. Weisshaar, *The spatial biology of transcription and translation in rapidly growing Escherichia coli.* Frontiers in microbiology, 2015. **6**: p. 636.

112. Bolte, S. and F.P. Cordelieres, *A guided tour into subcellular colocalization analysis in light microscopy.* J Microsc, 2006. **224**(Pt 3): p. 213-32.

113. Dunn, K.W., M.M. Kamocka, and J.H. McDonald, *A practical guide to evaluating colocalization in biological microscopy.* Am J Physiol Cell Physiol, 2011. **300**(4): p. C723-42.

114. Zinchuk, V., O. Zinchuk, and T. Okada, *Quantitative colocalization analysis of multicolor confocal immunofluorescence microscopy images: pushing pixels to explore biological phenomena.* Acta Histochem Cytochem, 2007. **40**(4): p. 101-11.

115. Cordelieres, F.P. and S. Bolte, *Experimenters' guide to colocalization studies: finding a way through indicators and quantifiers, in practice.* Methods Cell Biol, 2014. **123**: p. 395-408.

116. French, A.P., et al., *Colocalization of fluorescent markers in confocal microscope images of plant cells.* Nat Protoc, 2008. **3**(4): p. 619-28.

117. Lagache, T., et al., *Statistical analysis of molecule colocalization in bioimaging.* Cytometry A, 2015. **87**(6): p. 568-79.

118. McDonald, J.H. and K.W. Dunn, *Statistical tests for measures of colocalization in biological microscopy.* J Microsc, 2013. **252**(3): p. 295-302.

119. Wu, Y., et al., *Quantitative determination of spatial protein-protein correlations in fluorescence confocal microscopy.* Biophys J, 2010. **98**(3): p. 493-504.

120. Landmann, L., *Deconvolution improves colocalization analysis of multiple fluorochromes in 3D confocal data sets more than filtering techniques.* J Microsc, 2002. **208**(Pt 2): p. 134-47.

121. Barlow, A.L., et al., *Colocalization analysis in fluorescence micrographs: verification of a more accurate calculation of pearson's correlation coefficient.* Microsc Microanal, 2010. **16**(6): p. 710-24.

122. Martinez-Abrain, A., *Statistical significance and biological relevance: A call for a more cautious interpretation of results in ecology.* acta oecologica, 2008. **34**: p. 9-11.

123. Lovell, D.P., *Biological importance and statistical significance.* J Agric Food Chem, 2013. **61**(35): p. 8340-8.

124. Costes, S.V., et al., *Automatic and quantitative measurement of protein-protein colocalization in live cells.* Biophys J, 2004. **86**(6): p. 3993-4003.

125. Hodges, J.L. and L. E.L., *The efficiency of somc nonparametric competitors of the t-test.* Annals of Mathematical Statistics, 1956. **27**: p. 324-335.

126. Clifford Blair, R., J.J. Higgins, and D.S. Smitley, *On the relative power of the U and t tests.* British Journal of Mathematical and Statistical Psychology, 1980. **33**: p. 114-120.

127. Manders, E.M.M., F.J. Verbeek, and J.A. Aten, *Measurement of colocalization of objects in dual-colour confocal images.* J Microsc, 1993. **169**(3): p. 375-382.

128. Metz, C.E., *Receiver operating characteristic analysis: a tool for the quantitative evaluation of observer performance and imaging systems.* J Am Coll Radiol, 2006. **3**(6): p. 413-22.

129. Wheeler, D.B., et al., *RNAi living-cell microarrays for loss-of-function screens in Drosophila melanogaster cells.* Nat Methods, 2004. **1**(2): p. 127-32.

130. Carpenter, A.E., et al., *CellProfiler: image analysis software for identifying and quantifying cell phenotypes.* Genome Biol, 2006. **7**(10): p. R100.

131. Muller, E.G., et al., *The organization of the core proteins of the yeast spindle pole body.* Mol Biol Cell, 2005. **16**(7): p. 3341-52.

132. Krapp, A., et al., *S. pombe cdc11p, together with sid4p, provides an anchor for septation initiation network proteins on the spindle pole body.* Curr Biol, 2001. **11**(20): p. 1559-68.

133. Decottignies, A., P. Zarzov, and P. Nurse, *In vivo localisation of fission yeast cyclin-dependent kinase cdc2p and cyclin B cdc13p during mitosis and meiosis.* J Cell Sci, 2001. **114**(Pt 14): p. 2627-40.

134. Neeli-Venkata, R., et al., *Robustness of the Process of Nucleoid Exclusion of Protein Aggregates in Escherichia coli.* J Bacteriol, 2016. **198**(6): p. 898-906.

135. Mondal, J., et al., *Entropy-based mechanism of ribosome-nucleoid segregation in E. coli cells.* Biophys J, 2011. **100**(11): p. 2605-13.

136. Xu, L., et al., *Resolution, target density and labeling effects in colocalization studies - suppression of false positives by nanoscopy and modified algorithms.* Febs J, 2016. **283**(5): p. 882-98.
137. Bakshi, S., H. Choi, and J.C. Weisshaar, *The spatial biology of transcription and translation in rapidly growing Escherichia coli.* Front Microbiol, 2015. **6**: p. 636.
138. De Lay, N., D.J. Schu, and S. Gottesman, *Bacterial Small RNA-based Negative Regulation: Hfq and Its Accomplices.* Journal of Biological Chemistry, 2013. **288**(12): p. 7996-8003.
139. Scott, D.W., *Optimal and Data-Based Histograms.* Biometrika, 1979. **66**(3): p. 605-610.
140. Ljosa, V., K.L. Sokolnicki, and A.E. Carpenter, *Annotated high-throughput microscopy image sets for validation.* Nat Methods, 2012. **9**(7): p. 637.
141. Jones, T.R., A.E. Carpenter, and P. Golland, *Voronoi-based segmentation of cells on image manifolds. Proceedings of the Workshop on Computer Vision for Biomedical Image Applications (CVBIA)*. Lecture Notes in Computer Science 3765, ed. L. Yanxi, J. Tianzi, and Z. Changshui. 2005: Springer-Verlag, Berlin. 535-543.
142. Riffle, M. and T.N. Davis, *The Yeast Resource Center Public Image Repository: A large database of fluorescence microscopy images.* BMC Bioinformatics, 2010. **11**: p. 263.
143. Schneider, C.A., W.S. Rasband, and K.W. Eliceiri, *NIH Image to ImageJ: 25 years of image analysis.* Nat Methods, 2012. **9**(7): p. 671-5.
144. Kannaiah, S. and O. Amster-Choder, *Methods for studying RNA localization in bacteria.* Methods, 2015.
145. Kocaoglu, O. and E.E. Carlson, *Progress and prospects for small-molecule probes of bacterial imaging.* Nat Chem Biol, 2016. **12**(7): p. 472-8.
146. Xue, L., et al., *Imaging and manipulating proteins in live cells through covalent labeling.* Nat Chem Biol, 2015. **11**(12): p. 917-23.
147. Gautam, S., et al., *Exterior design: strategies for redecorating the bacterial surface with small molecules.* Trends Biotechnol, 2013. **31**(4): p. 258-67.
148. Gruskos, J.J., G. Zhang, and D. Buccella, *Visualizing Compartmentalized Cellular Mg2+ on Demand with Small-Molecule Fluorescent Sensors.* J Am Chem Soc, 2016. **138**(44): p. 14639-14649.
149. Perry, J.L., et al., *Use of genetically-encoded calcium indicators for live cell calcium imaging and localization in virus-infected cells.* Methods, 2015. **90**: p. 28-38.
150. Kervrann, C., et al., *A guided tour of selected image processing and analysis methods for fluorescence and electron microscopy.* IEEE Journal of Selected Topics in Signal Processing, 2016. **10**(1): p. 6-30.
151. Sheng, H., W. Stauffer, and H.N. Lim, *Systematic and general method for quantifying localization in microscopy images.* Biol Open, 2016. **5**(12): p. 1882-1893.
152. Adler, J. and I. Parmryd, *Quantifying colocalization: thresholding, void voxels and the H(coef).* PLoS One, 2014. **9**(11): p. e111983.
153. Schindelin, J., et al., *Fiji: an open-source platform for biological-image analysis.* Nat Methods, 2012. **9**(7): p. 676-82.
154. Yao, Z. and R. Carballido-López, *Fluorescence imaging for bacterial cell biology: from localization to dynamics, from ensembles to single molecules.* Annual review of microbiology, 2014. **68**: p. 459-476.

155. Haas, B.L., et al., *Imaging live cells at the nanometer-scale with single-molecule microscopy: obstacles and achievements in experiment optimization for microbiology.* Molecules, 2014. **19**(8): p. 12116-49.
156. Snapp, E., *Design and use of fluorescent fusion proteins in cell biology.* Curr Protoc Cell Biol, 2005. **Chapter 21**: p. Unit 21 4.
157. Wallner, G., R. Amann, and W. Beisker, *Optimizing fluorescent in situ hybridization with rRNA-targeted oligonucleotide probes for flow cytometric identification of microorganisms.* Cytometry, 1993. **14**(2): p. 136-43.
158. Patterson, G.H., et al., *Use of the green fluorescent protein and its mutants in quantitative fluorescence microscopy.* Biophys J, 1997. **73**(5): p. 2782-90.
159. Li, B. and L. You, *Predictive power of cell-to-cell variability.* Quantitative Biology, 2013. **1**(2): p. 131-139.
160. Ferreira, T. and W. Rasband. *ImageJ User Guide — IJ 1.46*. 2012; Available from: https://imagej.nih.gov/ij/docs/guide/user-guide.pdf.
161. Zernike, F., *Phase contrast, a new method for the microscopic observation of transparent objects.* Physica, 1942. **9**(7): p. 686-698.
162. Zernike, F., *Phase contrast, a new method for the microscopic observation of transparent objects part II.* Physica, 1942. **9**(10): p. 937-1019.
163. Obara, B., et al., *Bacterial cell identification in differential interference contrast microscopy images.* BMC Bioinformatics, 2013. **14**: p. 134.
164. Vincent, L. and P. Soille, *Watersheds in Digital Spaces - an Efficient Algorithm Based on Immersion Simulations.* Ieee Transactions on Pattern Analysis and Machine Intelligence, 1991. **13**(6): p. 583-598.
165. Rodriguez, O.C., et al., *Conserved microtubule-actin interactions in cell movement and morphogenesis.* Nat Cell Biol, 2003. **5**(7): p. 599-609.
166. Coles, C.H. and F. Bradke, *Coordinating neuronal actin-microtubule dynamics.* Curr Biol, 2015. **25**(15): p. R677-91.
167. Valerio-Santiago, M. and F. Monje-Casas, *Tem1 localization to the spindle pole bodies is essential for mitotic exit and impairs spindle checkpoint function.* J Cell Biol, 2011. **192**(4): p. 599-614.
168. Kelly, W.G., et al., *Distinct requirements for somatic and germline expression of a generally expressed Caernorhabditis elegans gene.* Genetics, 1997. **146**(1): p. 227-38.
169. Hsieh, J., et al., *The RING finger/B-box factor TAM-1 and a retinoblastoma-like protein LIN-35 modulate context-dependent gene silencing in Caenorhabditis elegans.* Genes Dev, 1999. **13**(22): p. 2958-70.
170. Bray, M.-A., et al., *Cell Painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes.* Nature protocols, 2016. **11**(9): p. 1757-1774.
171. English, A.R. and G.K. Voeltz, *Endoplasmic Reticulum Structure and Interconnections with Other Organelles.* Cold Spring Harbor Perspectives in Biology, 2013. **5**(4): p. a013227.
172. Marchi, S., S. Patergnani, and P. Pinton, *The endoplasmic reticulum–mitochondria connection: One touch, multiple functions.* Biochimica et Biophysica Acta (BBA) - Bioenergetics, 2014. **1837**(4): p. 461-469.

173. Prachar, J., *Intimate contacts of mitochondria with nuclear envelope as a potential energy gateway for nucleo-cytoplasmic mRNA transport.* General physiology and biophysics, 2003. **22**(4): p. 525-534.
174. Potente, M. and T. Makinen, *Vascular heterogeneity and specialization in development and disease.* Nat Rev Mol Cell Biol, 2017. **18**(8): p. 477-494.
175. Meacham, C.E. and S.J. Morrison, *Tumour heterogeneity and cancer cell plasticity.* Nature, 2013. **501**(7467): p. 328-37.
176. Ackermann, M., *A functional perspective on phenotypic heterogeneity in microorganisms.* Nat Rev Microbiol, 2015. **13**(8): p. 497-508.
177. Khushi, M., et al., *MatCol: a tool to measure fluorescence signal colocalisation in biological systems.* Sci Rep, 2017. **7**(1): p. 8879.
178. Kreft, M., et al., *Automated high through-put colocalization analysis of multichannel confocal images.* Comput Methods Programs Biomed, 2004. **74**(1): p. 63-7.
179. Fletcher, P.A., et al., *Multi-image colocalization and its statistical significance.* Biophys J, 2010. **99**(6): p. 1996-2005.
180. Buggenthin, F., et al., *An automatic method for robust and fast cell detection in bright field images from high-throughput microscopy.* BMC Bioinformatics, 2013. **14**: p. 297.
181. Wiegand, W., *Eclipse: A platform for integrating development tools.* IBM Systems Journal, 2004. **43**(2): p. 371-383.
182. Réveillac, J.-M., *Modeling and Simulation of Logistics Flows 2: Dashboards, Traffic Planning and Management*. 2017, Hoboken, NJ, USA: John Wiley & Sons, Inc.
183. Clayberg, E. and D. Rubel, *Eclipse Plug-ins*. Third ed. The Eclipse Series, ed. E. Gamma, L. Nackman, and J. Wiegand. 2008, Boston, MA.: Addison-Wesley. 928.
184. Ferreira, T. and W. Rasband *ImageJ User Guide — IJ 1.46*. 2012.
185. Hussein, R. and H.N. Lim, *Disruption of small RNA signaling caused by competition for Hfq.* Proc Natl Acad Sci U S A, 2011. **108**(3): p. 1110-5.
186. Thevenaz, P., U.E. Ruttimann, and M. Unser, *A pyramid approach to subpixel registration based on intensity.* IEEE Trans Image Process, 1998. **7**(1): p. 27-41.
187. Adler, J. and I. Parmryd, *Quantifying colocalization by correlation: the Pearson correlation coefficient is superior to the Mander's overlap coefficient.* Cytometry A, 2010. **77**(8): p. 733-42.
188. Adler, J., S.N. Pagakis, and I. Parmryd, *Replicate-based noise corrected correlation for accurate measurements of colocalization.* J Microsc, 2008. **230**(Pt 1): p. 121-33.
189. Li, Q., et al., *A syntaxin 1, Galpha(o), and N-type calcium channel complex at a presynaptic nerve terminal: analysis by quantitative immunocolocalization.* J Neurosci, 2004. **24**(16): p. 4070-81.
190. Adler, J. and I. Parmryd, *Colocalization analysis in fluorescence microscopy.* Methods Mol Biol, 2013. **931**: p. 97-109.
191. Manders, E.M., et al., *Dynamics of three-dimensional replication patterns during the S-phase, analysed by double labelling of DNA and confocal microscopy.* J Cell Sci, 1992. **103 (Pt 3)**: p. 857-62.
192. Andrews, D.T., et al., *Comments on the relationship between principal components analysis and weighted linear regression for bivariate data sets.* Chemometrics and Intelligent Laboratory Systems, 1996. **34**(2): p. 231-244.

193. Suhr, D.D., *Principal Component Analysis vs. Exploratory Factor Analysis*, in *SUGI 30 proceedings*. 2005. p. 203-230.

194. Wagner, E.G. and P. Romby, *Small RNAs in bacteria and archaea: who they are, what they do, and how they do it.* Adv Genet, 2015. **90**: p. 133-208.

195. Soper, T., et al., *Positive regulation by small RNAs and the role of Hfq.* Proc Natl Acad Sci U S A, 2010. **107**(21): p. 9602-7.

196. Sledjeski, D.D., A. Gupta, and S. Gottesman, *The small RNA, DsrA, is essential for the low temperature expression of RpoS during exponential growth in Escherichia coli.* EMBO J, 1996. **15**(15): p. 3993-4000.

197. Prevost, K., et al., *The small RNA RyhB activates the translation of shiA mRNA encoding a permease of shikimate, a compound involved in siderophore synthesis.* Mol Microbiol, 2007. **64**(5): p. 1260-73.

198. Moon, K., et al., *Complex transcriptional and post-transcriptional regulation of an enzyme for lipopolysaccharide modification.* Mol Microbiol, 2013. **89**(1): p. 52-64.

199. Masse, E. and S. Gottesman, *A small RNA regulates the expression of genes involved in iron metabolism in Escherichia coli.* Proc Natl Acad Sci U S A, 2002. **99**(7): p. 4620-5.

200. Lease, R.A., M.E. Cusick, and M. Belfort, *Riboregulation in Escherichia coli: DsrA RNA acts by RNA:RNA interactions at multiple loci.* Proc Natl Acad Sci U S A, 1998. **95**(21): p. 12456-61.

201. Wassarman, K.M., *6S RNA: a small RNA regulator of transcription.* Curr Opin Microbiol, 2007. **10**(2): p. 164-8.

202. Babitzke, P. and T. Romeo, *CsrB sRNA family: sequestration of RNA-binding regulatory proteins.* Curr Opin Microbiol, 2007. **10**(2): p. 156-63.

203. Caldelari, I., et al., *RNA-mediated regulation in pathogenic bacteria.* Cold Spring Harb Perspect Med, 2013. **3**(9): p. a010298.

204. Vogel, J. and B.F. Luisi, *Hfq and its constellation of RNA.* Nat Rev Microbiol, 2011. **9**(8): p. 578-89.

205. De Lay, N., D.J. Schu, and S. Gottesman, *Bacterial small RNA-based negative regulation: Hfq and its accomplices.* J Biol Chem, 2013. **288**(12): p. 7996-8003.

206. De Lay, N. and S. Gottesman, *Role of polynucleotide phosphorylase in sRNA function in Escherichia coli.* RNA, 2011. **17**(6): p. 1172-89.

207. Diestra, E., et al., *Cellular electron microscopy imaging reveals the localization of the Hfq protein close to the bacterial membrane.* PLoS One, 2009. **4**(12): p. e8301.

208. Carpousis, A.J., et al., *Copurification of E. coli RNAase E and PNPase: evidence for a specific association between two enzymes important in RNA processing and degradation.* Cell, 1994. **76**(5): p. 889-900.

209. Liou, G.G., et al., *RNA degradosomes exist in vivo in Escherichia coli as multicomponent complexes associated with the cytoplasmic membrane via the N-terminal region of ribonuclease E.* Proc Natl Acad Sci U S A, 2001. **98**(1): p. 63-8.

210. Jasiecki, J. and G. Wegrzyn, *Localization of Escherichia coli poly(A) polymerase I in cellular membrane.* Biochem Biophys Res Commun, 2005. **329**(2): p. 598-602.

211. Milo, R. and R. Philips, *Cell Biology by the Numbers*. 2015, New York, NY: Garland Science, Taylor and Francis Group. 400.

212. Henderson, C.A., et al., *Characterization of MicA interactions suggests a potential novel means of gene regulation by small non-coding RNAs.* Nucleic Acids Res, 2013. **41**(5): p. 3386-97.

213. Buxbaum, A.R., G. Haimovich, and R.H. Singer, *In the right place at the right time: visualizing and understanding mRNA localization.* Nat Rev Mol Cell Biol, 2015. **16**(2): p. 95-109.

214. Christie, M., et al., *Structural Biology and Regulation of Protein Import into the Nucleus.* J Mol Biol, 2016. **428**(10 Pt A): p. 2060-90.

215. Buskila, A.A., S. Kannaiah, and O. Amster-Choder, *RNA localization in bacteria.* RNA Biol, 2014. **11**(8): p. 1051-60.

216. Moremen, K.W., M. Tiemeyer, and A.V. Nairn, *Vertebrate protein glycosylation: diversity, synthesis and function.* Nat Rev Mol Cell Biol, 2012. **13**(7): p. 448-62.

217. Wang, X. and C. He, *Dynamic RNA modifications in posttranscriptional regulation.* Mol Cell, 2014. **56**(1): p. 5-12.

218. Garg, S.G. and S.B. Gould, *The Role of Charge in Protein Targeting Evolution.* Trends Cell Biol, 2016.

219. Goldenberg, N.M. and B.E. Steinberg, *Surface charge: a key determinant of protein localization and function.* Cancer Res, 2010. **70**(4): p. 1277-80.

220. Saraogi, I. and S.O. Shan, *Co-translational protein targeting to the bacterial membrane.* Biochim Biophys Acta, 2014. **1843**(8): p. 1433-41.

221. Bibi, E., *Is there a twist in the Escherichia coli signal recognition particle pathway?* Trends Biochem Sci, 2012. **37**(1): p. 1-6.

222. Golding, I. and E.C. Cox, *Physical nature of bacterial cytoplasm.* Phys Rev Lett, 2006. **96**(9): p. 098102.

223. Russell, J.H. and K.C. Keiler, *Subcellular localization of a bacterial regulatory RNA.* Proc Natl Acad Sci U S A, 2009. **106**(38): p. 16405-9.

224. dos Santos, V.T., A.W. Bisson-Filho, and F.J. Gueiros-Filho, *DivIVA-mediated polar localization of ComN, a posttranscriptional regulator of Bacillus subtilis.* J Bacteriol, 2012. **194**(14): p. 3661-9.

225. Stracy, M., et al., *Live-cell superresolution microscopy reveals the organization of RNA polymerase in the bacterial nucleoid.* Proc Natl Acad Sci U S A, 2015. **112**(32): p. E4390-9.

226. Sanamrad, A., et al., *Single-particle tracking reveals that free ribosomal subunits are not excluded from the Escherichia coli nucleoid.* Proc Natl Acad Sci U S A, 2014. **111**(31): p. 11413-8.

227. Jin, D.J., C. Cagliero, and Y.N. Zhou, *Role of RNA polymerase and transcription in the organization of the bacterial nucleoid.* Chem Rev, 2013. **113**(11): p. 8662-82.

228. Roggiani, M. and M. Goulian, *Chromosome-Membrane Interactions in Bacteria.* Annu Rev Genet, 2015. **49**: p. 115-29.

229. Steuten, B., et al., *Regulation of transcription by 6S RNAs: insights from the Escherichia coli and Bacillus subtilis model systems.* RNA Biol, 2014. **11**(5): p. 508-21.

230. Hussein, R. and H.N. Lim, *Direct comparison of small RNA and transcription factor signaling.* Nucleic Acids Res, 2012. **40**(15): p. 7269-79.

231. Urban, J.H. and J. Vogel, *Two seemingly homologous noncoding RNAs act hierarchically to activate glmS mRNA translation.* PLoS Biol, 2008. **6**(3): p. e64.

232. Levine, E., et al., *Quantitative characteristics of gene regulation by small RNA.* PLoS Biol, 2007. **5**(9): p. e229.
233. Kang, Z., et al., *Small RNA regulators in bacteria: powerful tools for metabolic engineering and synthetic biology.* Appl Microbiol Biotechnol, 2014. **98**(8): p. 3413-24.
234. Yoo, S.M., D. Na, and S.Y. Lee, *Design and use of synthetic regulatory small RNAs to control gene expression in Escherichia coli.* Nat Protoc, 2013. **8**(9): p. 1694-707.
235. Shepherd, D.P., et al., *Counting small RNA in pathogenic bacteria.* Anal Chem, 2013. **85**(10): p. 4938-43.
236. Fei, J., et al., *RNA biochemistry. Determination of in vivo target search kinetics of regulatory noncoding RNA.* Science, 2015. **347**(6228): p. 1371-4.
237. Golding, I., et al., *Real-time kinetics of gene activity in individual bacteria.* Cell, 2005. **123**(6): p. 1025-36.
238. Yagur-Kroll, S., A. Ido, and O. Amster-Choder, *Spatial arrangement of the beta-glucoside transporter from Escherichia coli.* J Bacteriol, 2009. **191**(9): p. 3086-94.
239. Wingreen, N.S. and K.C. Huang, *Physics of Intracellular Organization in Bacteria.* Annu Rev Microbiol, 2015. **69**: p. 361-79.
240. Atkins, P. and J. de Paula, *Physical Chemistry*. 2009, Oxford, U.K.: Oxford University Press. 1060.
241. Coimbatore Narayanan, B., et al., *The Nucleic Acid Database: new features and capabilities.* Nucleic Acids Res, 2014. **42**(Database issue): p. D114-22.
242. Thirumalai, D. and C. Hyeon, *Theory of RNA folding: From hairpins to ribozymes*, in *Non-protein coding RNAs*, N.G. Walter, S.A. Woodson, and R.T. Batey, Editors. 2009, Springer: Berlin. p. 1 online resource (xi, 398 pages).
243. Yoffe, A.M., et al., *Predicting the sizes of large RNA molecules.* Proc Natl Acad Sci U S A, 2008. **105**(42): p. 16153-8.
244. Hyeon, C., R.I. Dima, and D. Thirumalai, *Size, shape, and flexibility of RNA structures.* J Chem Phys, 2006. **125**(19): p. 194905.
245. Fang, L.T., W.M. Gelbart, and A. Ben-Shaul, *The size of RNA as an ideal branched polymer.* J Chem Phys, 2011. **135**(15): p. 155105.
246. Rawat, N. and P. Biswas, *Shape, flexibility and packing of proteins and nucleic acids in complexes.* Phys Chem Chem Phys, 2011. **13**(20): p. 9632-43.
247. Hajdin, C.E., et al., *On the significance of an RNA tertiary structure prediction.* RNA, 2010. **16**(7): p. 1340-9.
248. Moore, P.B., D.M. Engelman, and B.P. Schoenborn, *Asymmetry in the 50S ribosomal subunit of Escherichia coli.* Proc Natl Acad Sci U S A, 1974. **71**(1): p. 172-6.
249. Serdyuk, I.N. and A.K. Grenader, *Joint use of light, X-ray and neutron scattering for investigation of RNA and protein mutual distribution within the 50S subparticle of E. coli ribosomes.* FEBS Lett, 1975. **59**(1): p. 133-6.
250. Mandiyan, V., et al., *Assembly of the Escherichia coli 30S ribosomal subunit reveals protein-dependent folding of the 16S rRNA domains.* Proc Natl Acad Sci U S A, 1991. **88**(18): p. 8174-8.
251. Schuwirth, B.S., et al., *Structures of the bacterial ribosome at 3.5 A resolution.* Science, 2005. **310**(5749): p. 827-34.

252. Zhu, J., et al., *Three-dimensional reconstruction with contrast transfer function correction from energy-filtered cryoelectron micrographs: procedure and application to the 70S Escherichia coli ribosome.* J Struct Biol, 1997. **118**(3): p. 197-219.

253. Nielsen, J.S., et al., *An Hfq-like protein in archaea: crystal structure and functional characterization of the Sm protein from Methanococcus jannaschii.* RNA, 2007. **13**(12): p. 2213-23.

254. Peng, Y., et al., *Structural model of an mRNA in complex with the bacterial chaperone Hfq.* Proc Natl Acad Sci U S A, 2014. **111**(48): p. 17134-9.

255. Updegrove, T.B., et al., *E. coli DNA associated with isolated Hfq interacts with Hfq's distal surface and C-terminal domain.* Biochim Biophys Acta, 2010. **1799**(8): p. 588-96.

256. Balasubramanian, D. and C.K. Vanderpool, *Deciphering the interplay between two independent functions of the small RNA regulator SgrS in Salmonella.* J Bacteriol, 2013. **195**(20): p. 4620-30.

257. Wadler, C.S. and C.K. Vanderpool, *A dual function for a bacterial small RNA: SgrS performs base pairing-dependent regulation and encodes a functional polypeptide.* Proc Natl Acad Sci U S A, 2007. **104**(51): p. 20454-9.

258. Taniguchi, Y., et al., *Quantifying E. coli proteome and transcriptome with single-molecule sensitivity in single cells.* Science, 2010. **329**(5991): p. 533-8.

259. Ferrell, J.E., Jr. and S.H. Ha, *Ultrasensitivity part I: Michaelian responses and zero-order ultrasensitivity.* Trends Biochem Sci, 2014. **39**(10): p. 496-503.

260. Ferrell, J.E., Jr. and S.H. Ha, *Ultrasensitivity part III: cascades, bistable switches, and oscillators.* Trends Biochem Sci, 2014. **39**(12): p. 612-8.

261. Guillier, M., S. Gottesman, and G. Storz, *Modulating the outer membrane with small RNAs.* Genes Dev, 2006. **20**(17): p. 2338-48.

262. Hussein, R., T.Y. Lee, and H.N. Lim, *Quantitative characterization of gene regulation by Rho dependent transcription termination.* Biochim Biophys Acta, 2015. **1849**(8): p. 940-54.

263. McGary, K. and E. Nudler, *RNA polymerase and the ribosome: the close relationship.* Curr Opin Microbiol, 2013. **16**(2): p. 112-7.

264. Boudvillain, M., N. Figueroa-Bossi, and L. Bossi, *Terminator still moving forward: expanding roles for Rho factor.* Curr Opin Microbiol, 2013. **16**(2): p. 118-24.

265. Santangelo, T.J. and I. Artsimovitch, *Termination and antitermination: RNA polymerase runs a stop sign.* Nat Rev Microbiol, 2011. **9**(5): p. 319-29.

266. Khemici, V., et al., *The RNase E of Escherichia coli is a membrane-binding protein.* Mol Microbiol, 2008. **70**(4): p. 799-813.

267. Michaux, C., et al., *Physiological roles of small RNA molecules.* Microbiology, 2014. **160**(Pt 6): p. 1007-19.

268. Fisher, J.K., et al., *Four-dimensional imaging of E. coli nucleoid organization and dynamics in living cells.* Cell, 2013. **153**(4): p. 882-95.

269. Adamson, D.N. and H.N. Lim, *Essential requirements for robust signaling in Hfq dependent small RNA networks.* PLoS Comput Biol, 2011. **7**(8): p. e1002138.

270. Lutz, R. and H. Bujard, *Independent and tight regulation of transcriptional units in Escherichia coli via the LacR/O, the TetR/O and AraC/I1-I2 regulatory elements.* Nucleic Acids Res, 1997. **25**(6): p. 1203-10.

271. Sagawa, S., et al., *Paradoxical suppression of small RNA activity at high Hfq concentrations due to random-order binding.* Nucleic Acids Res, 2015. **43**(17): p. 8502-15.

272. Gardner, T.S., C.R. Cantor, and J.J. Collins, *Construction of a genetic toggle switch in Escherichia coli.* Nature, 2000. **403**(6767): p. 339-42.

273. Datsenko, K.A. and B.L. Wanner, *One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products.* Proc Natl Acad Sci U S A, 2000. **97**(12): p. 6640-5.

274. Zong, C., et al., *Lysogen stability is determined by the frequency of activity bursts from the fate-determining gene.* Mol Syst Biol, 2010. **6**: p. 440.

275. Adamson, D.N. and H.N. Lim, *Rapid and robust signaling in the CsrA cascade via RNA-protein interactions and feedback regulation.* Proc Natl Acad Sci U S A, 2013. **110**(32): p. 13120-5.

276. Shin, J.E., C. Lin, and H.N. Lim, *Horizontal transfer of DNA methylation patterns into bacterial chromosomes.* Nucleic Acids Res, 2016. **44**(9): p. 4460-71.

277. Zhang, J., C.P. Jones, and A.R. Ferre-D'Amare, *Global analysis of riboswitches by small-angle X-ray scattering and calorimetry.* Biochim Biophys Acta, 2014. **1839**(10): p. 1020-1029.

278. Muller, J.J., et al., *Comparison of the structure of ribosomal 5S RNA from E. coli and from rat liver using X-ray scattering and dynamic light scattering.* Eur Biophys J, 1986. **13**(5): p. 301-7.

279. Noriega, T.R., et al., *Real-time observation of signal recognition particle binding to actively translating ribosomes.* Elife, 2014. **3**.

280. Kilburn, D., et al., *Crowders perturb the entropy of RNA energy landscapes to favor folding.* J Am Chem Soc, 2013. **135**(27): p. 10055-63.

281. Lipfert, J., et al., *Structural transitions and thermodynamics of a glycine-dependent riboswitch from Vibrio cholerae.* J Mol Biol, 2007. **365**(5): p. 1393-406.

282. Kazantsev, A.V., et al., *Solution structure of RNase P RNA.* RNA, 2011. **17**(6): p. 1159-71.

283. Gopal, A., et al., *Visualizing large RNA molecules in solution.* RNA, 2012. **18**(2): p. 284-99.

284. Osterberg, R., et al., *The conformation of a large RNA fragment from the E.coli ribosomal 16S-RNA. An X-ray and neutron small-angle scattering study.* Nucleic Acids Res, 1980. **8**(24): p. 6221-31.

285. Folkhard, W., et al., *Small-angle x-ray studies on the structure of 16-S ribosomal RNA and of a complex of ribosomal protein S4 and 16-S ribosomal RNA from Escherichia coli.* Eur J Biochem, 1975. **59**(1): p. 63-71.

286. Stanley, W.M., Jr. and R.M. Bock, *Isolation and physical properties of the ribosomal ribonucleic acid of Escherichia coli.* Biochemistry, 1965. **4**(7): p. 1302-11.

287. Zipper, P. and W. Folkhard, *A small-angle x-ray scattering investigation on the structure of the RNA from bacteriophage MS2.* FEBS Lett, 1975. **56**(2): p. 283-7.

288. Werner, A., *Predicting translational diffusion of evolutionary conserved RNA structures by the nucleotide number.* Nucleic Acids Res, 2011. **39**(3): p. e17.

289. Alon, U., *An introduction to systems biology: design principles of biological circuits*. 2007: Chapman & Hall.

290. Richardson, J.P., *Preventing the synthesis of unused transcripts by Rho factor.* Cell, 1991. **64**(6): p. 1047-9.

291. Abe, H., T. Abo, and H. Aiba, *Regulation of intrinsic terminator by translation in Escherichia coli: transcription termination at a distance downstream.* Genes Cells, 1999. **4**(2): p. 87-97.

292. Deana, A. and J.G. Belasco, *Lost in translation: the influence of ribosomes on bacterial mRNA decay.* Genes Dev, 2005. **19**(21): p. 2526-33.

293. Simms, C.L., E.N. Thomas, and H.S. Zaher, *Ribosome-based quality control of mRNA and nascent peptides.* Wiley Interdiscip Rev RNA, 2017. **8**(1).

294. Laalami, S., L. Zig, and H. Putzer, *Initiation of mRNA decay in bacteria.* Cell Mol Life Sci, 2014. **71**(10): p. 1799-828.

295. Gowrishankar, J. and R. Harinarayanan, *Why is transcription coupled to translation in bacteria?* Mol Microbiol, 2004. **54**(3): p. 598-603.

296. Yanofsky, C., *Attenuation in the control of expression of bacterial operons.* Nature, 1981. **289**(5800): p. 751-8.

297. Russell, J.B. and G.M. Cook, *Energetics of bacterial growth: balance of anabolic and catabolic reactions.* Microbiol Rev, 1995. **59**(1): p. 48-62.

298. Scott, M., et al., *Interdependence of cell growth and gene expression: origins and consequences.* Science, 2010. **330**(6007): p. 1099-102.

299. Li, G.W., et al., *Quantifying absolute protein synthesis rates reveals principles underlying allocation of cellular resources.* Cell, 2014. **157**(3): p. 624-35.

300. Liu, K., A.N. Bittner, and J.D. Wang, *Diversity in (p)ppGpp metabolism and effectors.* Curr Opin Microbiol, 2015. **24**: p. 72-9.

301. Levin-Karp, A., et al., *Quantifying translational coupling in E. coli synthetic operons using RBS modulation and fluorescent reporters.* ACS Synth Biol, 2013. **2**(6): p. 327-36.

302. Dekel, E. and U. Alon, *Optimality and evolutionary tuning of the expression level of a protein.* Nature, 2005. **436**(7050): p. 588-92.

303. Lim, H.N., Y. Lee, and R. Hussein, *Fundamental relationship between operon organization and gene expression.* Proc Natl Acad Sci U S A, 2011. **108**(26): p. 10626-31.

304. Ikeuchi, K., T. Izawa, and T. Inada, *Recent Progress on the Molecular Mechanism of Quality Controls Induced by Ribosome Stalling.* Front Genet, 2018. **9**: p. 743.

305. Andreeva, I., R. Belardinelli, and M.V. Rodnina, *Translation initiation in bacterial polysomes through ribosome loading on a standby site on a highly translated mRNA.* Proc Natl Acad Sci U S A, 2018. **115**(17): p. 4411-4416.

306. Pedersen, S., et al., *Fast Translation within the First 45 Codons Decreases mRNA Stability and Increases Premature Transcription Termination in E. coli.* J Mol Biol, 2019. **431**(6): p. 1088-1097.

307. Chen, H., et al., *Genome-wide study of mRNA degradation and transcript elongation in Escherichia coli.* Mol Syst Biol, 2015. **11**(1): p. 781.

308. Kriner, M.A., A. Sevostyanova, and E.A. Groisman, *Learning from the leaders: Gene regulation by the transcription termination factor Rho.* Trends in biochemical sciences, 2016. **41**(8): p. 690-699.

309. Fan, H., et al., *Transcription–translation coupling: direct interactions of RNA polymerase with ribosomes and ribosomal subunits.* Nucleic acids research, 2017. **45**(19): p. 11043-11055.

310. Richardson, J.P., *Preventing the synthesis of unused transcripts by Rho factor.* Cell, 1991. **64**(6): p. 1047-1049.

311. Stanssens, P., E. Remaut, and W. Fiers, *Inefficient translation initiation causes premature transcription termination in the IacZ gene.* Cell, 1986. **44**(5): p. 711-718.

312. Keiler, K.C., *Mechanisms of ribosome rescue in bacteria.* Nature Reviews Microbiology, 2015. **13**(5): p. 285.

313. Roy, B. and A. Jacobson, *The intimate relationships of mRNA decay and translation.* Trends in Genetics, 2013. **29**(12): p. 691-699.

314. Cobb, M., *60 years ago, Francis Crick changed the logic of biology.* PLoS Biol, 2017. **15**(9): p. e2003243.