UNIVERSITA' DEGLI STUDI DI PAVIA

FACOLTA' DI INGEGNERIA
DIPARTIMENTO DI INGEGNERIA INDUSTRIALE E DELL'INFORMAZIONE

DOTTORATO DI RICERCA IN
TECNOLOGIE PER LA SALUTE, BIOINGEGNERIA E BIOINFORMATICA
XXXIII CICLO - 2020

OPTIMIZATION OF A NETWORK DESIGN TO CONTROL THE EXPRESSION OF ANY TARGET GENE IN BACTERIA

PhD Thesis by

DAVIDE DE MARCHI

Advisors:
Prof. PAOLO MAGNI
Prof. LORENZO PASOTTI

PhD Program Chair:
Prof. SILVANA QUAGLINI

What I cannot create, I do not understand.

Richard P. Feynman,
Professor in theoretical physics at Caltech, 1965

Abstract (English)

Synthetic biology, a discipline born from the interconnection between biology and engineering, is assuming a pivotal role in the world scientific panorama due to its ability to rationally modify existing organisms by improving their biological performances or insert new functions through the rewriting of their genetic program. The sectors in which synthetic biology has brought significant technological developments are disparate, such as: agriculture, cosmetics, therapeutics, energy. Microorganisms can be adopted to face different problems in such areas, via the *ad hoc* design of novel customized biological functions. Nonetheless, the environmental selectivity of the ecological niches belonging to the different fields of application mentioned above generally prevents the model microorganisms used in the laboratory (e.g., *Escherichia coli*, *Bacillus subtilis*) from establishing a symbiotic bond with the environment, leading to the consequent failure of the genetic circuit incorporated in these hosts. The environmental hurdle has stimulated the scientific community to look for new biological *chassis* that are suitable for coming into contact with the ecological niche considered, sometimes selecting them directly *in situ*. A major issue in the use of non-model hosts is that the previously characterized libraries of biological parts (e.g., promoters, Ribosome Binding

Site - RBS, plasmid vectors) used to create gene expression, widely and successfully tested for laboratory strains, are sometimes incompatible with them motivating the need to find new efficient and rational methods to control gene expression in any biological *chassis*. Two approaches can be adopted to achieve this aim: the first is based on the characterization of new regulatory parts in new *host*, which may require huge investments in terms of time and money, while the second is based on the construction of a synthetic circuit architectures that are able to regulate gene expression regardless of the strength of the specific regulatory parts used or from environmental sources of variability. A particular realization of the second approach has been defined as the *holy grail* of synthetic biology, as: "*the realization of a synthetic system which, once transformed, can operate efficiently in any biological chassis*".

This thesis is centered on the development of genetic and computational tools, based on control circuits, to facilitate the engineering of non-standard *chassis*. In particular, the work is focused on a genetic architecture for controlling the expression of a gene of interest that stably and robustly rejects the sources of variability that commonly affect a genetic circuit when inserted into a new bacterial *host* in terms of the variation of its key parameters, such as transcription rate, translation rate and number of copies of the gene. The designed architecture has been investigated *in silico* and *in vivo* to characterize its properties and limitations. Furthermore, computational analyses have been carried out to support the improvement of the control circuit via a different architecture, and to select regulatory parts that can increase the portability of the circuits in different host strains.

In Chapter 1, an overview of the various applications of synthetic biology, with particular attention to the limitations deriving from the use of model microorganisms has been presented. Furthermore, the state of the art of portable control circuits for gene expression in different *hosts* and the biological parts used in them has been described,

focusing on the CRISPRi technology, used in the development of the genetic architectures subsequently reported, and followed finally by the purpose of the thesis.

In Chapter 2, an overview of the mathematical modelling tools commonly used to describe genetic circuits is carried out. The chapter is focused on a case study whose purpose is to show possible limitations of traditional Hill-function models and to expand the descriptive capabilities of models through the addition of new factors such as, variation in the level of regulatory proteins, variation in the number of copies of the gene and the cell load exerted by heterologous gene expression. The descriptive power and the identification of such models is finally discussed, showing that traditional tools are sometimes unable to capture experimental measurements.

In Chapter 3, a new circuit design, based on the *incoherent feedforward-loop (iFFL)* control architecture, called Sad-iFFL, has been presented, whose automatic regulation is based on a new repressor enzyme: the dead Cas9 from *Staphylococcus aureus* (SadCas9). The description of the circuit scheme and its biological implementation is followed by its analysis of the mathematical model, which is used to compare the *in silico* performances of Sad-iFFL with an expression cassette without control architectures (herein named Open loop circuit). Subsequently, the *in vivo* characterization of the SadCas9 enzyme, i.e., the main biological module of the circuit, and the final Sad-iFFL circuit have been reported.

In Chapter 4, a new circuit design, called U-iFFL, based on incoherent feedforward-loop and positive autoregulation has been presented, aimed to improve the portability performances of the circuit developed in the previous chapter. The analysis of the circuit scheme has been carried out analytically and the *in silico* performances with the Sad-iFFL and Open loop circuits have been compared. The characterization of the new biological module included in U-iFFL for the positive autoregulation motif, RNAPT7-P_{T7}, has been presented but,

unfortunately, its inclusion in the final circuit has not been reported due to delays in the assembly of the U-iFFL controller due to to the global pandemic by Sars-CoV-2.

In Chapter 5, a bioinformatics procedure, including two pipelines, has been developed for the *in silico* detection of promoter and RBS sequences in annotated genomes, from genomics and transcriptomics data. Their dual purpose is of 1) providing new libraries of regulatory parts to the scientific world for non-model microorganisms; 2) increasing the portability of the previously analyzed circuits by enhancing the probability that promoters and RBSs are functional in different hosts.

In Chapter 6, the summary and overall conclusions of this work have been reported.

Abstract (Italian)

La biologia sintetica, disciplina sorta dal connubio tra biologia e ingegneria, sta assumendo un ruolo centrale nel panorama scientifico mondiale per la sua capacità di modificare razionalmente organismi esistenti, apportandone migliorie o nuove funzioni tramite la riscrittura del loro programma genetico. I settori nei quali la biologia sintetica ha apportato notevoli sviluppi tecnologici sono molteplici e molto diversi tra loro, quali: agricolo, cosmetico, terapeutico, energetico. La selettività ambientale delle nicchie ecologiche appartenenti ai diversi campi di applicazione sovracitati impedisce, generalmente, a microorganismi modello utilizzati in laboratorio (e.g., *Escherichia coli*, *Bacillus subtilis*) di creare un legame simbionte con l'ambiente, con il conseguente fallimento del circuito genetico in essi inserito. Tale ostacolo ambientale ha stimolato la comunità scientifica nel cercare nuovi *chassis* biologici da ingegnerizzare che risultino adatti a entrare in contatto con la nicchia ecologica considerata, talvolta selezionandoli direttamente *in situ*. Ciò nonostante, le librerie di parti biologiche (e.g., promotori, Ribosome Binding Site - RBS, vettori plasmidici) precedentemente caratterizzate nella letteratura scientifica, usate per creare cassette di espressione genica, risultano essere incompatibili con diversi microorganismi non-modello, motivando la necessità di trovare nuovi metodi

per controllare l'espressione genica in nuovi *chassis* biologici. Due approcci possono essere adottati per raggiungere questo scopo: il primo si basa sulla caratterizzazione di nuove parti regolatrici in nuovi *host*, il quale richiede ingenti investimenti in termini di tempo e denaro, mentre il secondo si basa sulla costruzione di circuiti genetici secondo opportuni schemi di controllo, che riescano a regolare l'espressione genica indipendentemente dalla forza delle parti regolatorie usate o da fonti di variabilità ambientali. La realizzazione del secondo approccio è stata definita come il *santo graal* della biologia sintetica, ovvero: "*la realizzazione di un sistema sintetico che, una volta trasformato, possa operare efficientemente in ogni chassis biologico*".

Il lavoro di questa tesi è centrato nello sviluppo di un' architettura genetica per il controllo dell'espressione di un gene di interesse che sia robusta alle fonti di variabilità che influenzano un circuito genetico quando viene inserito in un nuovo *host* batterico, in termini di variazione dei suoi parametri chiave, come velocità di trascrizione, velocità di traduzione e numero di copie del gene.

Nel capitolo 1 viene esposta una panoramica generale sulle diverse applicazioni della biologia sintetica ponendo particolare attenzione alle limitazioni derivate dall'uso di microorganismi modello. Inoltre, viene descritto lo stato dell'arte dei circuiti di controllo dell'espressione genica in diversi *host* e le parti biologiche in essi utilizzate, soffermandosi sulla tecnologia CRISPRi, impiegata nello sviluppo delle architetture genetiche successivamente discusse, seguito infine dalla descrizione dello scopo della tesi.

Nel capitolo 2 è illustrata una sintesi sui modelli matematici comunemente utilizzati per descrivere il comportamento di circuiti sintetici. In particolare, il capitolo si focalizza su un caso di studio il cui scopo è di mostrare le limitazioni dei tradizionali modelli di Hill e come è possibile aumentare le capacità descrittive di tali modelli tramite l'aggiunta di nuovi fattori in grado di rappresentare la variazione del livello di proteine regolatrici e del numero di copie del gene

oltre al carico metabolico. Infine, viene discussa la capacità descrittiva e l'identificabilità di tali modelli, mostrando come gli strumenti modellistici tradizionali sono talvolta incapaci di descrivere dati sperimentali misurati da circuiti sintetici *in vivo*.

Nel capitolo 3 viene presentato un nuovo *design* circuitale basato sull'architettura di controllo *incoherent feedforward-loop (iFFL)*, chiamato Sad-iFFL, la cui regolazione automatica si basa su un nuovo enzima repressore, la dead Cas9 di *Staphylococcus aureus* (SadCas9). La descrizione dello schema circuitale e della sua rappresentazione biologica è seguita dall'analisi del modello matematico, confrontando le *performance in silico* di Sad-iFFL con una cassetta di espressione priva di architetture di controllo (chiamata circuito Open loop). A seguire, è riportata la caratterizzazione *in vivo* dell'enzima SadCas9, il principale modulo biologico di Sad-iFFL, e del circuito finale Sad-iFFL stesso.

Nel capitolo 4 viene riportato un nuovo *design* circuitale, chiamato U-iFFL, che si propone di migliorare le *performance* del modello sviluppato nel capitolo precedente mediante l'utilizzo di regolazioni di tipo *incoherent feed-forward loop* oltre all'autoregolazione positiva. L'analisi dello schema circuitale è stata effettuata per via analitica e le *performances in silico* sono state confrontate con i circuiti Sad-iFFL e Open loop. La caratterizzazione del nuovo modulo biologico necessario nel circuito U-iFFL per l'autoregolazione positiva, RNAPT7-P_{T7}, viene presentato nel capitolo ma, sfortunatamente, la sua inclusione nel circuito finale non è stata riportata a causa dei rallentamenti nei lavori di assemblaggio del controllore U-iFFL dovuti alla pandemia globale di Sars-CoV-2.

Nel capitolo 5 è riportata una procedura bioinformatica, basata su due *pipeline*, per l'identificazione *in silico* di sequenze promotrici e RBS in genomi annotati, a partire da dati di genomica e trascrittomica, con lo scopo duplice di: fornire nuove librerie di parti regolatrici alla comunità scientifica per microorganismi non modello e aumentare la portabilità dei circuiti proposti nei capitoli precedenti, aumentando

la probabilità di funzionamento di promotori ed RBS in diversi organismi; entrambi punti ad elevato interesse nel mondo della biologia sintetica attuale.

Nel capitolo 6 sono riportate le conclusioni generali di questo lavoro, tenendo in considerazione tutti i risultati ottenuti nei precedenti capitoli.

Contents

List of Figures

List of Tables

Chapter 1

Background

1.1 Synthetic biology and its application: a new interest on non-model microorganisms

If synthetic biology could be expressed in one sentence, it would probably be represented by Richard Feynman's quote: "What I cannot create I do not understand" where precisely, the word creation refers to life. The increasingly knowledge of the manipulation of the information contained in DNA has created different perspectives from which we are able to modify nature. From the creation of the first GMO mouse (Rudolf Jaenisch, 1974) by insertion of foreign DNA taken from a retrovirus [1], we have arrived, nowadays, to design complex genetic circuits that can be chemically synthesized in order to rationally modify living organisms [2]. The idea behind synthetic biology, deriving from the engineering approach, is the rational design for which genetic circuits are seen as the interconnection of elementary biological modules in order to provide one or more new functions to the target

cell [3, 4]. While, at the beginning, the study of genetic circuits and the characterization of their regulatory components (e.g., promoters, RBSs, transcriptional factors, etc.) took place in model microorganisms (e.g., *Escherichia coli*, *Bacillus subtilis*) to understand how to control gene expression for a range of applications [5, 6, 7], synthetic biology is currently going through a transition period. The toolkit of circuit parts that has been developed for model bacteria must be expanded to enable reliable engineering and gene expression control in non-model bacteria to achieve a predictable and stable protein expression in microorganisms better suited to work in the ecological niche required in any application [8, 9, 10]. To understand better this issue, the most representative macro-areas of synthetic biology applications have been illustred in Fig. 1.1 and, for each of them, representative examples and the limitations emerged in the use of model bacteria are reported.

- **Biomaterials**: Windmaier et al. and, later, Azam et al., engineered the type III secretion system of *Salmonella enterica* subsp. *Typhimurium* for the production of high-value polymers as: spider silk protein (first study) [11] and pro-resilin, tropoelastin (second study) [12]. Although protein secretion was significantly higher than control strain, it has been shown that the engineered bacteria is very sensitive to conformational alteration of its natural secretion system complex; furthermore, the performances in terms of secreted protein level could change significantly as new target protein is chosen.

- **Cosmetic**: A *Bacillus subtilis* strain has been engineered for the production of hyaluronic acid by hijacking the endogenous pathway using CRISPRi technology, but it has been demonstrated that the production efficiency is limited by the intracellular concentration of the necessary substrates (e.g., UDP-

Figure 1.1: **Synthetic biology applications.** The engineering and *ad hoc* manipulation of biological systems has created new alternatives to existing technologies in different sectors (e.g., industrial, therapeutical) and generated new ones. The wide set of applications ranges from the production of molecules (e.g., therapeutics treatment, cosmetics, biofuels) to the creation of new biomaterials or the bioconversion of pollutants in harmless products.

GlcUA, UDP-GlcNAc) for high production of high molecular-weight hyaluronic acid [13].

- **Industrial & Biofuels**: The same strain of *Pseudomonas putida* was used by Gonzales et al. and Samuel et al. for the biodegradation of phenol and p-nitrophenol, respectively, present in waste water that derived from industrial processes [14, 15]. The resistance of *P. putida* to these molecules favors its use in the

industrial scenarios but the dependence on the carbon source introduced into the growth media to allow bacterial growth makes the overall process not convenient economically [14, 15, 16]. To overcome this limitation, new soil bacteria have been proposed (e.g., *Cupriavidus pinatubonensis*) which used molecules such as phenol as carbon source for their growth [17]. A further well-known industrial application is the production of biofuels from waste or non-edible products such as ligno-cellulosic biomass (carbon source: xylose, arabinose, galactose, mannose and glucose) and whey permeate (carbon source: lactose) through genetically modified model and non model bacteria in order to maximize the conversion yield between the sugar present and the biomolecule produced (e.g., bioethanol, biopropanol, biodiesel) [18, 19, 20, 21]. In several studies, biological catalysts have been obtained with high yields compared to wild type strains but, nevertheless, it has been estimated that the monetary revenue from the enhancement process is significantly influenced by the chemical reagents costs used to control the environmental parameters (e.g., pH, osmotic pressure, antibiotic for strain selection) within the bioreactor [22, 23].

- **Agriculture**: The plants optimization for agriculture purposes, in terms of growth and productivity, has covered a pivotal role for centuries. The limit of plants on nitrogen fixation processes (differently from carbon) increased the used of chemical nitrogen-enriched fertilizers in the agriculture fields which worsens the CO2 levels in the atmosphere [24, 25]. An alternative solution come through the use of genetic modified nitrogen-fixing bacteria which are able to enstablish a symbiotic relashionship with the plant. In addition to the known model bacterium *B. subtilis*, other microorganisms have been identified for their use (e.g., salt tolerance, drought resistance) in the rhizosphere (narrow

regions of soil influenced directly from plant root secretions), such as *Azospirillum brasilense* or *Burkholdera cepacia*, which lack a rich toolkit of regulatory parts for gene expression and thus no synthetic circuits have been made so far [26, 27, 28].

- **Bioremediation**: Bacteria have been selected to eliminate toxic compounds (e.g., anti-inflammatories, insecticides, antibiotics, etc) from soils, water and surface materials in order to improve environment bioremediation [29, 30]. The main difficulty in using bacteria for this purpose is summarized in the sentences reported from the study of Dvorak et al. "*A critical issue will be the choice and establishment of new chassis organisms better suited to field biotransformations than the currently available laboratory strains in terms of resistance to harsh conditions including extreme pH, temperature, osmotic pressure or fluctuating concentrations of toxic chemical*". Indeed, the harsh growth conditions of these environments do not allows the usage of model bacteria directly in the application fields, highligthing even more the interest from this research field on *new chassis organisms* [31, 10, 29].

- **Diagnosis**: The logic design of the 'sense-logic-actuate' behavior of engineered bacteria has made possible to obtain numerous biosensors which, by interrogating the environment in which they are placed, are able to detect the presence of the disease by altering the logical state of the circuit [32, 33]. An example has been reported in the study of Danino et al. in which the *E. coli Nissle 1917* strain has been modified to detect the presence of tumor inside the liver of a transgenic mouse through a bioluminescent signal present in the urine [34].

- **Therapeutics**: One of the emerging fields in which synthetic biology has devoted much attention is the therapeutic one. The historical symbiotic interaction between Microbes and Man has

meant that many pathologies of the human body are attributable to dysbiosis of the microbiota in certain anatomical areas [35]. In fact, scientific evidence shows a possible correlation between dysbiosis, most of which can be traced back to the intestinal microbiota [36], and diseases such as: immune-mediated diseases, including obesity [37], malnutrition [38], intestinal inflammatory disease [39], as well as to anti-cancer immunity [40, 41]. These assumptions have laid the foundations to engineering bacteria (called probiotics) which, once inserted in the ecological niche for which they were designed, can provide well-being for the human host, in terms of treatment from a dysbiotic state or prevention from a pathological state [42]. Examples are the studies of Yang et al. and Jacouton et al. which, through the engineering of *Salmonella tyhimurium* and *Lactococcus lactis*, respectively, have created two anti-tumor systems: the first silence oncogenes expression with shRNA technology [43], while the second inhibits tumor growth by IL-17A interleukin secretion [44]. Alternative solutions to drug therapies for Inflammatory Bowel Disease (IBD) have been proposed by Palmer et al. through the use of an engineered version of *Escherichia coli Nissle 1917* for the production of a selective bacteriocin against *Salmonella typhimurium*, a bacteria that can cause intestinal infections [45]. In parallel, Praveschotinunt et al. proposed, a system, based on the same bacterial chassis of the previous study, that is able to create an extracellular matrix containing all three trefoil factors to treat inflammation and help re-build the intestinal epithelium [46].

The limitation that emerged in the use of a non-optimal model microorganism related to the context in which it is inserted is a common factor to all the above-reported applications. A representative analysis for all scenarios is given below taking as example one of the

prominent fields for synthetic biology over the last few years: gut microbiota manipulation via engineered probiotics. The highly dynamic environment inside the gut prevents us from predicting with reasonable confidence the effect that an *ad hoc* engineered microorganism can lead to the gut bacteria population system in our body [47]. In fact, the overall contribution of a genetic circuit implementing a function of interest can be weak or null if the final bacterial concentration, determined by the interaction between bacteria and environment, is not sufficiently high to guarantee a physiological response in the host [48]. An additional factor, which limits the use of engineered probiotics, is that the perturbation created by the introduction of a new microorganism can introduce a new state of dysbiosis or aggravate an existing one [42]. In recent years, therefore, research in this field has focused on characterizing, and subsequently engineering, bacteria present in human instestine, so that, once modified and re-introduced into their natural habitat, they have more chances to establish a stable, safe and durable connection with the host. Among them, the focus is on the overrepresented bacteria of the human intestine as belonging to the genus of *Bacteroides* (e.g., *B. fragilis*, *B. melaninogenicus*, *B. thetaiotaomicron*) [49, 50], *Bifidobacteria* (e.g., *B. longum*, *B. bifidum*) [51] and *Enterococci* (e.g., *E. faecium*, *E. faecalis*) [52]. Unfortunately, the toolkit of regulatory parts (e.g., promoters, RBSs, plasmid vectors) to create functional synthetic circuits for the above bacterial strains is reduced or, sometimes, not available. The expression of new proteins can therefore take place in two ways: the first, through the *in vivo* or *in silico* discovery of large libraries of new regulatory elements for the fine tuning of predictable gene expression [53] or, in the second case, through the use of control circuits, usually inspired by the engineering world, for the automatic and adaptive production of the desired proteins using a smaller set of available regulatory parts [54]. Since the first solution requires huge investments in terms of time and money, aggravated by the difficulty to create in laboratory an environ-

ment suitable for bacterial growth of microorganisms such as obligate anaerobic bacteria (e.g., *B. fragilis*), an intriguing option is to focus the research in this sector on the second solution, described in more detail in the next Section.

1.2 Synthetic circuits for the control of gene expression in bacteria

The central dogma of molecular biology describes the one-way flow of genetic information whereby the gene, replicating within the DNA, is initially transcribed into mRNA and subsequently translated into protein (Fig. 1.2). The numerous sources of intra-(e.g., cell burden) and extra-(e.g., temperature, pH, osmotic pressure) cellular variability [55], which significantly affect the gene expression steps, can result in the variation of the three processes described in Fig. 1.2 and their main parameters such as: transcription rate, translation rate and gene copy number. Controlling the variation of these three processes allows to control the expression of a gene in a stable and predictable way [56]. Furthermore, since these are the three sources of variability that affect the behavior of a pre-characterized circuit when inserted into a new microorganism, their fine control would achieve one of the most ambitious goals of synthetic biology: the stable and robust re-usability of genetic circuit across several bacterial species. In synthetic biology, the simplest way to express a gene at a predictable level is through the interconnection of regulatory elements taken from libraries of previously characterized parts, such as: promoters, ribosome binding sites (RBSs) and plasmid vectors. In model organisms, such as the well-characterized *E. coli* bacterium, the libraries of regulatory parts have a size suitable to span a large range of strength values and the expression of the gene is calculated from the combination of the activ-

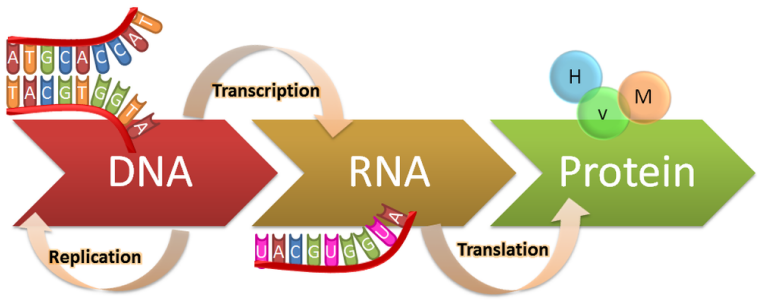

Figure 1.2: **Central dogma of molecular biology.** The flow of genetic information from DNA to proteins is possibile thorughout three main processes: DNA replication, DNA transcription into mRNA and mRNA translation into protein [57].

ities of the chosen elements. As mentioned above, such components are inevitably context-dependent elements, which may show quantitative changes in their activities depending on biological (e.g., circuit or flanking regions), host (e.g., strain) or environmental (e.g., growth media) contexts, and the accurate characterization of such context-dependent variation is one of the main goals in predictable design of synthetic circuits. For such reasons, these 'open loop' regulators have considerable limitations, such as: cell-cell variability [55], orthogonality with the enzymatic host machinery is not always guaranteed [58] and, if the regulatory parts are not chosen appropriately, can generate metabolic load for the cell. These difficulties are even more important in the engineering of non-model bacteria due to the poor, or even unavailable, library of regulatory parts [59]. To overcome these limits, genetic controller schemes have been developed inspired by biological network embedded in life genomes in order to create synthetic circuits which finely control the expression of a target gene. The main genetic networks over-represented in nature are: simple regulation (a gene influences directly the expression of a second gene), Negative and Positive AutoRegulation (NAR and PAR) and Feed-Forward Loop (FFL)

which, in turn can be divided, according to the global effect on the target gene, into coherent and incoherent FFLs [60, 61]. The characterization of biological networks discovered in microorganisms allowed the isolation of different actuators (e.g., LuxR from the Lux-circuitry of *Vibrio fischeri*, LacI from the Lac-circuitry of *E. coli*) and re-use them to create new gene control network.

In the literature, several groups tried to robustly control the expression of one or multiple genes through a synthetic controller that could works in different, also phylogenetically distance, species. Kushwaha et al. created UBER (Universal Bacterial Expression Resource), a combination of positive-autoregulation and incoherent feedforward loop networks which use a bacterial-orthogonal RNA polymerase taken from phage T7 to control the transcriptional process of a target gene in three different bacteria (*E. coli*, *B. subtilis* and *P. putida*). Despite that, the final gene expression was different due to the different rate of translation or circuit copy number among the aforementioned species [56]. The orthogonality of the RNAPT7-T7 system in bacteria leads several groups to use this transcriptional regulator to improve the production of metabolites such as Deoxychromoviridians pigment [62]. Ceroni F. et al. implemented an auto-regulatory feedback logic to avoid intracellular burden that arise from the overexpression of a metabolic route in bacterial genome through CRISPRi technology [63]. A slightly different scheme, based on the theory of quasi-integral controller, has been realized from Huang HH. et al. to tune the expression of a target gene independently from the ribosome demand using RNA interference [64]. Among the various network schemes characterized in the literature, one theoretically achieving the perfect adaptation of a gene in response to context-dependent disturbances is the incoherent-feedforward loop (iFFL) [65, 66]. Segall-Shapiro et al. achieved the perfect adaptation on copy number variation that affect the expression of a target gene by implementing the aforementioned schema using TALE protein and engineered TALE-repressible promoters [65].

The functionality of iFFL has been also tested in mammalian cells by Bleris et al. and Jones et al., reaching robust adaptation to variations of genetic template and cell load [67, 68]. Nonetheless, the robust adaptation of a gene through the three main sources of gene expression variation (transcription rate, translation rate and copy number) has not been achieved yet and new network architecture are needed in order to control gene expression throughout different bacterial species.

1.3 Synthetic biology tools: focus on CRISPRi technology

The biological networks mentioned in the previous Section based their functionality on regulators, generally proteins which modify reversibly the expression of target genes. This regulation can occur in different stages of gene expression, such as: transcriptional, post-transcriptional, translational or post-translational. Usually, the most frequently used actuators in synthetic biology are the regulators of transcriptional processes which, throught DNA binding in a portion close to the promoter sequence, increase or repress the transcriptional activity of the downstream gene. These regulators are called respectively: activators or repressors. One of the most widely used transcription modulators in the literature is the Cas9-engineered dCas9 protein (Fig. 1.3) belonging to the CRISPR-Cas immune system proteins whose original purpose was to recognize exogenous DNA sequences within the bacteria cell and cleave them [70]. The dCas9 protein is a catalytically inactive version of Cas9 ('d' stands for 'dead') obtained by the de-activation of the active sites responsible for DNA cutting, while leaving active its DNA binding capability [69]. The target recognition of dCas9 protein is mediated by single guide RNA (sgRNA). The latter is the engineered version of the system found in nature, originally

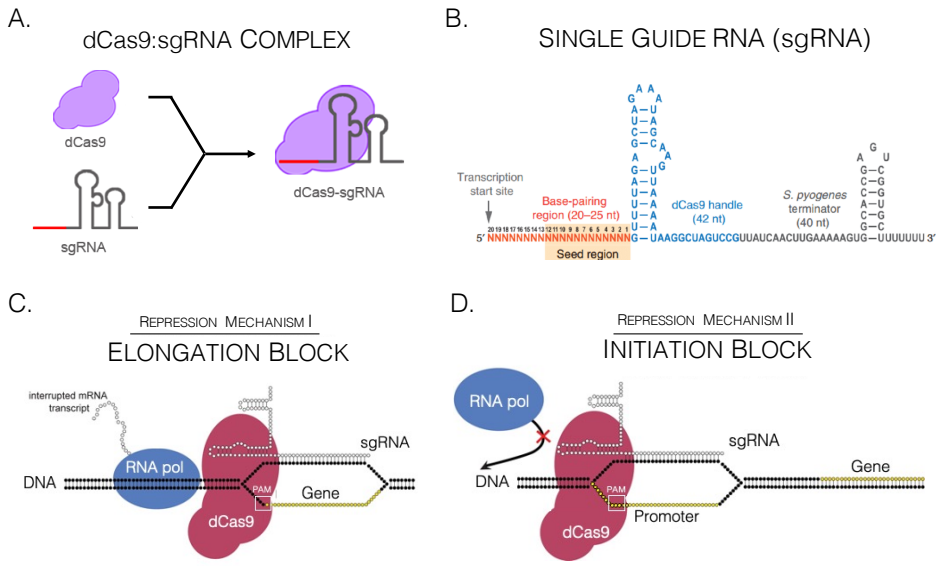

Figure 1.3: **CRISPR interference (CRISPRi) technology.** **A**, The activated repressor (dCas9:sgRNA) is composed by two single biological modules: the nuclease null Cas9, called dCas9 (d stands for 'dead') and the single guide RNA (sgRNA) [69]. **B**, The single guide RNA is an RNA molecule, designed from the natural guide RNA produced in *Streptococcus pyogenes* by Qi et al. [69], composed by three parts with indipendetly function: the *base pairing region* is responsible to DNA targeting through DNA-RNA complementarity, the *dCas9 handle* allows the link with the dCas9 protein to form the dCas9:sgRNA complex and the sgRNA terminator which blocks the RNA polymerase transcription process. **C**, dCas9:sgRNA complex can block transcription process in two ways: obstructing the RNA polymerase in its elongation phase or **D**, hiding the promoter sequence and prevent the transcription initiation.

composed of two RNA molecules expressed separately, in a single RNA molecule so that it can be expressed by a single expression cassette. The single guide RNA is composed of three parts: the base pairing region, the dCas9 handle and the transcriptional terminator where the last two are responsible for binding to dCas9 and for terminating the transcriptional activity of the guide, respectively. The base pairing region is the most important part of the sgRNA since it mediates the recognition of the target region from dCas9 protein; the targeted region can be changed through the modification of an about 20bp se-

quence of sgRNA. The binding between dCas9:sgRNA complex and the DNA is possible only if the PAM (Proto-spacer Adjacent Motif) region is present upstream of the target region, essential to start the binding event of dCas9 (Fig. 1.3 C and D). The nucleotide composition region of PAM sequence is specific to Cas9 protein and the specificity of the latter is intimately linked to it; in fact, from the data reported in the literature, dCas9 proteins with more specific PAM are less likely to generate cellular toxicity due to the lowest number of possible genomic off-targets. The CRISPR technology has been used to develop two types of actuators, repression and activation, called CRISPRi (interference) [69] and CRISPRa (activation) [71], respectively. Differently from activation, the repression based on CRISPR technology has been thoroughly characterized and several circuits developed in the literature base their functionality on it. Mainly, dCas9 can repress the expression of a gene in two ways which are both based on the interference of the normal functioning of bacterial RNA polymerase. The first repression mechanism is based on blocking the RNA polymerase elongation process and truncating the mRNA production of the target gene (Fig. 1.3C), while, the second prevents the binding between the RNA polymerase and the bacterial promoter (Fig. 1.3D) [69]. Although CRISPRi-based technology has found considerable success in the literature, its application in the endogenous metabolic pathways regulation deserves more accurate design and optimization study. In fact, although more specific PAMs give less off-target probabilities and therefore better performances in terms of toxicity and time, the probability to find a target in a specific portion of the genome decreases and, usually, it can be difficult to design a functional regulation system [72]. When the gene regulation pathways are embedded in a plasmid circuits and, therefore, is possible to design *de novo* the target sequence (as it has been done in this work), dCas9 proteins with more selective PAM are desirable in order to increase the aforementioned benefits and the orthogonality with the host, especially if the future

goal of the project is to expand the portability of the optimized circuit to different bacterial species [73]. The most widely used CRISPR (and CRISPRi) systems are based on the Cas9 (and dCas9) protein of *Streptococcus pyogenes*, in which the PAM motif has the 5'-NGG-3' sequence [70]. A more specific PAM motif is found in other CRISPR systems such as the one of *Staphylococcus aureus*, used in this work, and having the 5'-NNGRRT-3' PAM sequence [74].

1.4 Project idea and bigger picture

Synthetic biology is currently in a period of transition determined by the fact that applications demand new methodologies for the rapid and predictable engineering of bacterial chassis suitable for working in selective ecological niches. Although several groups are working to find new solutions in terms of characterization of new bacterial hosts, the *in vivo* and *in silico* search for new regulatory parts (e.g., promoters, RBS, plasmid replication vectors, etc.) and the development of new gene control architectures, a concrete ending has not yet been reached and the forecast investments, in terms of money and time, are huge due to the enormous amount of work required.

The innovative solution that could break this harmful scheme was explained by Adam BL in his work: "*Building genetic toolkits for each member of the next generation SB chassis panel is an extensive and laborious undertaking. For this reason, the holy grail of SB remains a synthetic system that is universal and can be transformed into and operate efficiently within any chassis*" [10].

Based on the concepts illustrated above, the aim of the PhD project is based on the development of a genetic architecture to control the expression of a gene of interest and to robustly reject the sources of variability that affect a genetic circuit when it is moved into a new host in terms of variation of its key parameters, such as transcription

rate, translation rate and circuit copy number. At the same time, this architecture must take into account the knowledge gaps inherent in the engineering of non-model bacteria. In fact, the design phase is based on the hypothesis that at least one set of promoter sequence, RBS and a plasmid vector is known, and the circuit can adapt protein expression to reach a target level regardless of the activity of these three components. The actuators that has been used to implement the automation control of gene expression must be orthogonal (or at least minimally cross-talking) with the cell machinery throughout all bacterial species. The genetic architectures developed in this work are two, called Sad-iFFL and U-iFFL. The first architecture is based on an architecture well known in the literature (Section 1.2), i.e., the incoherent-feedforward loop, for which the regulation system has been optimized according to its portability and orthogonality. The second circuit is a completely new architecture which, combining more subnetworks, aims to overcome the limitations of the first and achieve better performance in terms of output stability, robustness and predictability. Both architectures base their automatic output regulation system on the dCas9 repressor engineered from its wild type Cas9 version, while the second circuit relies, in addition, on a phage transcription system (RNAPT7$-P_{T7}$). The characterization of the two essential biological modules has been carried out *in vivo* to test their suitability in the composition of the designed circuits, based on their regulatory parameters that have been estimated from experimental data. The comparison among three architectures (Sad-iFFL, U-iFFL and Open Loop control) has been analyzed *in silico* (both at the steady$-$state and dynamics), and the performances theoretically investigated based on robustness and stability in different scenarios (e.g., upon variations of transcription rate, translation rate and protein target level). The full theoretical and experimental characterization of Sad-iFFL is reported in Chapter 3. For U-iFFL, only the theoretical characterization of the circuit has been carried out and reported in Chapter 4, while its

assembly has been interrupted due to Sars-CoV-2 worldwide pandemic and the work is still on-going.

The functionality of the automatic controllers explained so far rely on the availability of regulatory parts (e.g., promoters, RBS) that drive the genes in the two circuits. Under a set of assumptions, illustrated in detail in Chapter 3 and 4, the circuits are expected to adjust the target protein level even if transcription, translation and copy number varies, thereby making the circuits portable modules for protein expression. To support the rational choice of promoters and RBSs that may be functional (though with diverse activities) in different bacterial species, in Chapter 5 two bioinformatic pipelines are discussed for parts (promoters and RBSs) selection based on publicly available high-throughput expression data and genome sequences. The same pipelines could also support the generation of new promoter and RBS libraries that research laboratories can refer to engineering new bacteria with a considerable range of strengths.

Chapter 2

Mechanistic models to expand the predictability of inducible systems in synthetic biology[1]

This chapter reports a number of mathematical modelling tools that will be used throughout the project. They include widely used equations, such as empirical Michaelis-Menten and Hill functions, as well as more mechanistic equations that are not commonly used in mathematical analysis of synthetic circuits. Such more complex equations are derived and their identifiability discussed to support their use. Finally, a mathematical model of cell load is illustrated and a full model jointly describing the effects of variations of protein level, circuit copy number and cell load is provided. All the chapter adopts the lux inducible system as a case study, including a regulatory protein (LuxR) and the cognate promoter (P_{Lux}), to demonstrate that all the studied aspects are crucial to quantitatively capture the transfer

[1]The content of this chapter is pubblished in the article "Mechanistic Models of Inducible Synthetic Circuits for Joint Description of DNA Copy Number, Regulatory Protein Level, and Cell Load" [75].

function of a synthetic circuit.

2.1 Mathematical modeling of biological systems

Among the many issues currently limiting model-based approaches in synthetic biology, several unpredictability sources affect the re-use of biological parts in different contexts (i.e., strains, growth media, and even circuits). Cell-to-cell variability, flanking regions-dependent behavior, and cell load are among the major features causing such variability [76, 77]. Many of such effects have often been neglected in mathematical models, thereby limiting their predictive power. In addition, widely used modeling approaches describe recombinant protein regulation via empirical Michaelis–Menten or Hill equations [78]. Empirical models are popular tools and have many advantages (e.g., low number of parameters, overall good descriptive capability and no need to know the biomolecular interactions underlying the described process) [79, 80]. However, they may have poor predictive power on unseen data when one or more circuit elements are changed [81], like copy number of DNA or protein regulators that have been reported to be essential for the biological systems behavior [82]. In specific, transfer functions of inducible devices can be characterized *in vivo* through dose–response experiments, in which a constitutively expressed regulator (activator or repressor) is activated or inhibited by an exogenously added molecule and the complex eventually affects the transcription of a cognate regulated promoter [83]. While empirical models can be easily identified from such experimental data, it is not trivial to generalize them in situations in which molecule copy number changes, also for the empirical nature of model parameters [81]. This weak aspect of empirical models might be crucial in practical bottom-up

design situations, in which circuit parts are interconnected and the behavior of the system is predicted from the functioning of individual parts, by using previously estimated parameters [84]. Mechanistic models are able to overcome some of the issues mentioned above: parameters usually have biological meaning (e.g., dissociation constants, copy numbers, etc.) and predictive power is expected to be higher than empirical models, since circuit changes can be translated into the variation of specific parameters [81]. However, mechanistic models usually have a larger number of parameters to be estimated, thereby raising issues of model identifiability [79, 85]. Since such models require a deep knowledge of occurring biomolecular interactions in the biological system under study, they are less popular than empirical models. Different mechanistic modeling efforts have been undertaken, describing gene regulation using thermodynamics or law of mass action [79, 86]. However, the application of these models in bottom-up design approaches remains low due to the absence of studies on their identification and lack of broad-range advantages demonstrations over the empirical ones. In this work, a novel mechanistic steady-state models of the lux inducible system, used as case study, have been derived and compared with empirical model equations. Mechanistic models rely on different assumptions on regulatory protein (LuxR) and cognate promoter (P_{Lux}) concentrations, inducer-protein complex formation, and resource usage limitation. It has been demonstrated that the change of model assumptions significantly affects the circuit output; preliminary data are in accordance with the activation curves.

2.2 Models definition

2.2.1 Inducible system description

The inducible system studied in this work is described in Fig. 2.1. It includes a constitutively expressed LuxR protein, which acts as transcriptional activator of the cognate P_{Lux} promoter in presence of N-(3-oxohexanoyl)-L-homoserine lactone (HSL). The LuxR-HSL active complex binds to the lux box of the P_{Lux} promoter, triggering its activity in an HSL concentration-dependent fashion, thereby regulating the transcription of the downstream gene, which is then translated into protein. We considered a set of reactions affecting the production of the (reporter) protein-encoded gene regulated by P_{Lux}, similar to the ones described by Carbonell-Ballestero et al. [81] (Fig. 2.1b). Specifically, a LuxR protein dimer (R_{2T}) and HSL (L) form a complex (Q), which binds the free promoter (P) to form a transcriptionally active bound promoter (S). As an alternative reaction scheme, we also considered a similar set of biomolecular interactions, with complex formation occurring in two steps (Fig. 2.1c): binding of two steps (Fig. 2.1c): binding of R_{2T} and L to form Q, and subsequent binding of L and Q to form the hetero-tetramer Q_2. This alternative reaction set was defined to refine the mechanistic model above, in accordance with previous investigations, in which a hetero-tetrameric structure was suggested for the activated complex [87].

In the following part of this section, different models describing the lux system are illustrated. The essential elements of gene expression modeling are described in 2.2.3 together with the description of a basic empirical model for transcription activation. The description of all the models used in this work is provided in Sections 2.2.3 - 2.2.6. Due to the complexity of model structure under some assumptions, the theoretical analysis of the activation function parameters in provided only for the models described in Section 2.2.2 and 2.2.3.

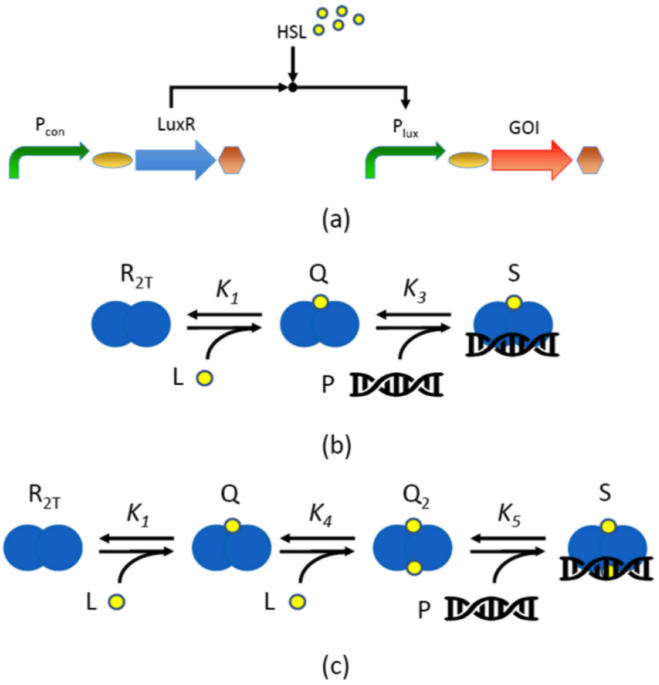

Figure 2.1: **Description of the lux inducible system.** (**a**) The gene encoding the LuxR protein transcriptional regulator is expressed by a constitutive promoter (Pcon); the LuxR regulator becomes activated upon HSL molecule binding to form a complex which can bind the P_{lux} promoter in its single lux box sequence, thereby activating the expression of the gene of interest (GOI), placed downstream of P_{lux}. Curved arrows represent promoters; ovals represent ribosome binding sites (RBSs); straight arrows represent coding sequences; hexagons represent transcriptional terminators; circles represent HSL molecules. (**b**) and (**c**) Biomolecular reactions modeled in this work. The LuxR dimer (R_{2T}, blue overlapping circles) is assumed to be activated upon binding of one (b) or two (c) HSL molecules (L, yellow circles) to form an activated complex (Q or Q_2, respectively), which binds DNA (double helix icon) to enable transcription. In panel (c), Q is the LuxR dimer form bound to one HSL molecule and it is assumed to be unable to bind DNA. The equilibrium constants are reported for each reaction occurring in panels (b) and (c).

2.2.2 Empirical Michaelis-Menten models (M0)

Assumptions

Assuming that no post-transcriptional or post-translational regulations are involved in the circuit, the intracellular dynamics of mRNA (M), immature protein (X) concentration regulated by P_{lux} and the mature form of the protein (Y) can be described via Equations (2.1)−(2.3):

$$\frac{dM}{dt} = H(L) - \gamma_M \cdot M \tag{2.1}$$

$$\frac{dX}{dt} = \rho \cdot M - (\gamma_X + \sigma) \cdot X \tag{2.2}$$

$$\frac{dY}{dt} = \sigma \cdot X - \gamma_X \cdot Y \tag{2.3}$$

In Equation (2.1), γ_M is the mRNA degradation rate and H(L) is an HSL-dependent activation function describing transcription rate, expressed in ($RNA\ time^{-1}$) units. In Equation (2.2), γ_X ($time^{-1}$), includes the protein degradation and dilution rates, and ρ ($time^{-1}$) is the translation rate. If the expressed protein is very stable, is equal to the rate of cell division. When relevant in terms of dynamics, protein maturation or folding is also included: Y represents mature protein, σ ($time^{-1}$) represents protein maturation rate and protein degradation rate is assumed to be the same for immature and mature forms. Assuming the steady-state of all the intracellular processes, the output commonly considered for such system, i.e., the mature protein synthesis rate per cell (y), which is equal to the synthesis term of Equation (2.3), is proportional to H(L) (Equation (2.4))

$$y = \frac{\sigma \cdot \rho \cdot H(L)}{(\gamma_X + \sigma) \cdot \gamma_M} \tag{2.4}$$

Another output commonly found in literature is the pre-cell mature protein (Y), which is proportional to y, thereby enabling to generalize all the modeling work presented in this study. In many works considering a constant LuxR production, not changing throughout the experiments [83, 88], H(L) is modeled via a Hill equation and the circuit output, y, can be written as in Equation (2.5)

$$y = \delta + \frac{\alpha}{1 + \left(\frac{\kappa}{L}\right)^{\eta}} \tag{2.5}$$

Assuming no cooperativity, Equation (2.5) with $\eta = 1$ is equivalent to a Michaelis–Menten equation with three parameters: δ is the basic protein synthesis rate when P_{Lux} is in its off state; α is the activity range in the on state; κ is the HSL concentration corresponding to half of the maximum activation [81]. The δ and α parameters are expressed as intracellular protein concentration per time. The experimental measurements routinely performed in laboratory to characterize synthetic circuits usually exploit fluorescent reporter proteins, which are quantified via *in vivo* assays by means of plate reader or flow-cytometry. In these cases, per-cell arbitrary units of fluorescence (AU) can be adopted to express intracellular protein concentration, assuming their proportionality.

Derivation

When LuxR level is also needed to be described (situation of interest in the present work), another commonly found modeling approach includes the following equations (Equations (2.6) and (2.7)) [87], describing the LuxR-HSL complex formation and the subsequent activa-

tion of protein synthesis by this complex

$$C = \frac{U}{1 + \left(\frac{\kappa_R}{L}\right)^{\beta}} \tag{2.6}$$

$$y = \widehat{\delta} + \frac{\widehat{\alpha}}{1 + \left(\frac{\widehat{\kappa}}{C}\right)^{\widehat{\eta}}} \tag{2.7}$$

In Equation (2.6), C is the intracellular LuxR-HSL complex concentration, U is the total LuxR concentration, κ_R is the concentration of HSL required for half-activation of LuxR, and β is the Hill coefficient. In Equation (2.7), symbols have the same meaning as in Equation (2.5), with the cap denoting that this Hill function has C as input. For this reason, $\widehat{\kappa}$ has the same units as C and U (protein concentration or AU, as explained above). Assuming no cooperativity, as before, the Hill coefficients (β and $\widehat{\eta}$) can be fixed to 1. The expressions in Equations (2.6) and (2.7) can be lumped into a single equation describing y as a function of L (Equation (2.8))

$$y = \widehat{\delta} + \frac{\widehat{\alpha}/(1 + \widehat{\kappa}/U)}{1 + \frac{\kappa_R/(1+U/\widehat{\kappa})}{L}} \tag{2.8}$$

which is equivalent to the Michaelis–Menten function in Equation (2.5) with the following parameters (Equations (2.9)–(2.11))

$$\delta = \widehat{\delta} \tag{2.9}$$

$$\alpha = \frac{\widehat{\alpha}}{1 + \frac{\widehat{\kappa}}{U}} \tag{2.10}$$

$$\kappa = \frac{\kappa_R}{1 + \frac{U}{\widehat{\kappa}}} \tag{2.11}$$

Final Expression

Although the parameters of Equation (2.8) have empirical nature, the copy numbers of P_{Lux} (n_1) and LuxR (n_2) can be included in the model. Specifically, n_1 and n_2 can act as known scale factors of $\widehat{\delta}$, $\widehat{\alpha}$ and U (Equations (2.12) - (2.14)), thereby enabling the description of copy number changes among different situations. Following a bottom-up approach, model parameters ($\widehat{\delta}$ or d, $\widehat{\alpha}$ or v, $\widehat{\kappa}$ and κ_R) can be estimated from experimental data and the parametrized model can be adopted to predict unseen situations with different n_1 and n_2

$$\widehat{\delta} = n_1 \cdot d \tag{2.12}$$

$$\widehat{\alpha} = n_1 \cdot v \tag{2.13}$$

$$U = n_2 \cdot u \tag{2.14}$$

As before, the system output (y) can be expressed in per-time intracellular protein concentration or $AUtime^{-1}$, like $\widehat{\delta}$ and $\widehat{\alpha}$. Analogously, U and $\widehat{\kappa}$ are intracellular protein concentrations that can also be expressed in AU, while $\widehat{\kappa}$ has the same units as HSL concentration. For this reason, the relative copy number changes, instead of the absolute ones, are sufficient to express n_1 and n_2. Specifically, n_1 can be set according to plasmid copy number and n_2 is proportional to the strength of the promoter expressing LuxR. Finally, the u value (Equation (2.14)) can be set to 1 without any loss of generality, since U is always present in ratio with $\widehat{\kappa}$ (Equation (2.8)). The final empirical model, called M0, describing the output of the lux inducible system as a function of HSL and the per- cell copy numbers of P_{Lux} and LuxR, is reported in Equation (2.15)

$$y = n_1 \cdot d + \frac{n_1 \cdot v/(1 + \widehat{\kappa}/(n_2 \cdot u))}{1 + \frac{\widehat{\kappa}/(1+n_2 \cdot u/\widehat{\kappa})}{L}} \tag{2.15}$$

2.2.3 Mechanistic model with LuxR abundance assumption (M1)

Assumptions

A mechanistic model of the biomolecular reactions shown in Fig. 2.1b has been previously reported [81] and is herein briefly illustrated. The main underlying assumptions are that HSL molecules bound to LuxR are negligible compared to the total HSL amount, and that $P << R$. A difference from the work by Carbonell-Ballestero et al. is that no binding is assumed between LuxR dimer and P_{Lux}, and the low-entity promoter activation by this unspecific complex is neglected. Previous experimental work showed that this activity increase is undetectable in the commonly tested conditions [89].

Derivation

Considering the same equations for transcription, translation and maturation (Equations (2.1) - (2.3)) as in Section 2.2.2, H(L) can be expressed in a more mechanistic fashion (Equation (2.16))

$$H(L) = \kappa_{m0} \cdot P + \kappa_{mL} \cdot S \tag{2.16}$$

where P and S represent the intracellular concentrations of free and complex-bound promoter, while κ_{m0} and κ_{mL} are the transcription rate constants in off and on state, respectively. Based on the law of mass action, the biomolecular interactions in Fig. 2.1b depend on the following equilibrium constants and conservation expressions (Equations

(2.17) - (2.20))

$$K_1 = \frac{Q}{L \cdot R_2} \tag{2.17}$$

$$K_3 = \frac{S}{P \cdot Q} \tag{2.18}$$

$$P_T = P + S \tag{2.19}$$

$$R_{2T} = R_2 + Q + S \tag{2.20}$$

Symbols in Equations (2.17) - (2.20) are described in Fig. 2.1b; briefly, R_{2T} and R_2 are the total and free LuxR dimer concentrations, respectively, where $R_{2T} \approx R_T/2$ (and is the total concentration of LuxR monomer). Under the assumption that $P << R$, the bound promoter S in Equation (2.20) can be neglected: $R_{2T} = R_2 + Q$. Considering Equations (2.1) - (2.3) and Equations (2.16) - (2.20), the following expressions for LuxR-HSL complex and circuit output can be written (Equations (2.21) and (2.22)) [81]

$$Q = \frac{R_{2T}}{1 + \frac{1}{K_1 \cdot L}} \tag{2.21}$$

$$y = \widehat{\kappa}_{m0} \cdot P_T + \frac{(\widehat{\kappa}_{mL} - \widehat{\kappa}_{m0}) \cdot P_T}{1 + \frac{1}{K_3 \cdot Q}} \tag{2.22}$$

where $\widehat{\kappa}_{m0}$ and $\widehat{\kappa}_{mL}$ are κ_{m0} and κ_{mL}, respectively, scaled by $\sigma \cdot \rho / ((\gamma_X + \sigma) \cdot \gamma_M)$ (see Equation (2.4)). These equations show that, like in the empirical model in Equations (2.6) and (2.7), in this mechanistic model the intracellular complex concentration and circuit output can be both described by Michaelis–Menten functions. Analogously, circuit output as a function of HSL is a Michaelis– Menten function (Equation (2.23))

with coefficients described in Equations (2.24) - (2.26)

$$\begin{aligned} y =& P_T \cdot \widehat{\kappa}_{m0} + \frac{P_T \cdot (\widehat{\kappa}_{m0} + \widehat{\kappa}_{mL})}{1 + 1/(K_2 \cdot R_{2T})} \\ & \cdot \frac{1}{1 + 1/(K_1 \cdot (1 + K_3 \cdot R_{2T}) \cdot L)} \end{aligned} \tag{2.23}$$

$$\delta = P_T \cdot \widehat{\kappa}_{m0} \tag{2.24}$$

$$\alpha = \frac{P_T \cdot (\widehat{\kappa}_{m0} + \widehat{\kappa}_{mL})}{1 + 1/(K_3 \cdot R_{2T})} \tag{2.25}$$

$$\kappa = 1/(K_1 \cdot (1 + K_3 \cdot R_{2T})) \tag{2.26}$$

In summary, M0 (Equation (2.15)) and M1 (Equation (2.23)) have the same $y(L)$ function and analogous LuxR-dependent Michaelis – Menten parameters expressions (Equations (2.9) - (2.11) and Equations (2.24) - (2.26)). However, the mechanistic nature of M1 gives the advantages that molecule concentrations are explicitly described and the link between model parameters and biological mechanism is not lost. As a result, M1 offers the opportunity to parametrize the model with biologically meaningful parameters, like intracellular concentrations, equilibrium constants, and copy numbers. Nonetheless, hard-to-measure quantities can still be expressed in AUs instead of concentrations, as in M0. For instance, while promoter concentration and DNA copy number are easy to retrieve (e.g., from assumptions about cell volume and plasmid datasheets or quantitative PCR), protein concentration requires more resource consuming experiments (e.g., Western blot) and depends not only on DNA copy number, but also on transcription, translation and degradation rates [81]. Nevertheless, the relative protein level can still be approximated via the strength of the upstream promoter. For the reasons above, in this work M0 will not be further analyzed, and only serves as reference for M1 parameters expressions.

Final Expression

The final expression of M1, explicitly including n_1 and n_2, is reported in Equation (2.27)

$$y = n_1 \cdot P_U \cdot \widehat{\kappa}_{m0} + \frac{n_1 \cdot P_U \cdot (\widehat{\kappa}_{m0} + \widehat{\kappa}_{mL})/(1 + 1/(K_3 \cdot n_2 \cdot r_T))}{1 + 1/(K_1 \cdot (1 + K_3 \cdot n_2 \cdot r_T) \cdot L)} \quad (2.27)$$

where P_U and r_T are the intracellular concentrations of one DNA copy of promoter and LuxR protein, respectively. The n_1 and n_2 parameters are, respectively, the P_{Lux} copy number, and the scale factor between one protein monomer and the actual dimer concentration.

2.2.4 Mechanistic model without LuxR abundance assumption (M2)

Assumptions and Final Expression

Following the same biomolecular reactions and calculation steps as in M1, it is possible to calculate circuit output without the $P << R$ assumption. In this situation, S cannot be neglected in Equation (2.20). Circuit output can be computed from Equation (2.28), in which R_2 and P are the roots of a second order two-equation system (Equations (2.29) - (2.30)). In this equation system, only one root has a biologically acceptable meaning for any value of L (not shown). Analytical formulas expressing R_2 and P are not reported due to their complexity and, as a result, Equations (2.28) - (2.30) represent the final reported expression for M2

$$y = P \cdot (\widehat{\kappa}_{m0} + \widehat{\kappa}_{mL} \cdot K_1 \cdot K_3 \cdot R_{2T} \cdot L) \quad (2.28)$$

$$(K_1 \cdot K_3 \cdot L) \cdot P^2 + (1 + K_1 \cdot L + K_1 \cdot K_3 \cdot n_2 \cdot r_T \cdot L - K_1 \cdot K_3 \cdot n_1 \cdot P_U \cdot L) \cdot P - (P_T + K_1 \cdot n_1 \cdot P_U \cdot L) = 0 \quad (2.29)$$

$$(K_1 \cdot K_3 \cdot L + K_1^2 \cdot K_3 \cdot L^2) \cdot R_{2T}^2 + (1 + K_1 \cdot L + K_1 \cdot K_3 \cdot n_1 \cdot P_U \cdot L - K_1 \cdot K_3 \cdot n_2 \cdot r_T \cdot L) \cdot R_{2T} - n_2 \cdot r_T = 0 \quad (2.30)$$

2.2.5 Mechanistic model with LuxR-HSL hetero-tetramerization (M1T, M2T)

Assumptions

The biomolecular interactions illustrated in Fig. 2.1c can be used to derive mathematical models in which, differently from Fig. 2.1b, the activated complex is a hetero-tetramer, formed by two LuxR and two HSL molecules. In particular, under the $P << R$ assumption, we also assumed that: i) the LuxR dimer has two binding sites for HSL; ii) the probability of HSL binding to a free site of R_2 and Q is the same (i.e., there is no cooperative behavior); iii) the probability of HSL unbinding from an occupied site of R_{2T} and Q is the same.

Derivation

The output can be expressed as in Equation (2.16), the K_1 equilibrium constant as in Equation (2.17) and the conservation of total, free and bound promoter as in Equation (2.19). The following expressions describe equilibrium constants K_4 and K_5, as well as the LuxR dimer (free, bound with L, bound with $2L$ and bound with $2L$ and the promoter) conservation (Equations (2.31) - (2.33))

$$K_4 = \frac{Q_2}{Q \cdot L} \tag{2.31}$$

$$K_5 = \frac{S}{P \cdot Q_2} \tag{2.32}$$

$$R_{2T} = R_2 + Q + Q_2 + S \tag{2.33}$$

Thanks to the $P << R$ assumption, as in M1 the S concentration becomes negligible in Equation (2.33), thereby having: $R_{2T} = R_2 + Q + Q_2$. The forward and reverse rate constants in Equation (2.34) k_+ and k_-) and Equation (2.35) (k'_+ and k'_-) describing the two HSL binding steps, have the following relations: $k_+ = 2 \cdot k'_+$ and $k'_- = k_- \cdot 2$:

$$R_2 + L \underset{k_-}{\overset{k_+}{\rightleftharpoons}} Q \tag{2.34}$$

$$Q + L \underset{k'_-}{\overset{k'_+}{\rightleftharpoons}} Q_2 \tag{2.35}$$

For the described reasons, a relation between K_1 and K_4 can be written (Equation (2.36)),

$$K_4 = \frac{k'_+}{k'_-} = \frac{k_+}{4 \cdot k_-} = \frac{K_1}{4} \tag{2.36}$$

Final Expression

Based on the relations above and following the same mathematical steps as in M1, the final output for this model (M1T) is reported in Equation (2.37); the M1T expression in Equation (2.37) does not resemble a Michaelis – Menten function (differently from M1).

$$y = n_1 \cdot P_U \cdot \frac{\widehat{k}_{m0} + (\widehat{k}_{m0} \cdot K_1)}{1 + K_1 \cdot L + (K_1^2/4 + K_1^2/4 \cdot K_5 \cdot n_2 \cdot r_T) \cdot L^2} \cdot L + n_1 \cdot P_U \cdot \frac{(\widehat{k}_{m0} \cdot K_1^2/4 + \widehat{k}_{mL} \cdot K_1^2/4 \cdot K_5 \cdot n_2 \cdot r_T)}{1 + K_1 \cdot L + (K_1^2/4 + K_1^2/4 \cdot K_5 \cdot n_2 \cdot r_T) \cdot L^2} \cdot L^2 \tag{2.37}$$

Without the $P << R$ assumption, the expressions of circuit output become more complex. By following the same steps previously done in Section 2.2.4 for M2, the final model (M2T) is described by the following expressions (Equations (2.38)-(2.40)).

$$y = P \cdot \left(\widehat{k}_{m0} + \widehat{k}_{mL} \cdot \frac{K_1^2}{4} \cdot K_5 \cdot R_2 \cdot L \right) \tag{2.38}$$

$$\begin{cases} a_{R_2} \cdot R_2^2 + b_{R_2} \cdot R_2 + c_{R_2} = 0 \\ a_{R_2} = \left(\dfrac{K_1^2}{4} \cdot K_5 \cdot L^2 + \dfrac{K_1^3}{4} \cdot L^3 + \dfrac{K_1^4}{16} \cdot K_5 \cdot L^4 \right) \\ b_{R_2} = \left(1 + K_1 \cdot L + \dfrac{K_1^2}{4} \cdot L^2 + \dfrac{K_1^2}{4} \cdot K_5 \cdot n_1 \cdot P_U \cdot L^2 - \dfrac{K_1^2}{4} \cdot K_5 \cdot n_2 \cdot r_T \cdot L^2 \right) \\ c_{R_2} = -n_2 \cdot r_T \end{cases} \tag{2.39}$$

$$\begin{cases} a_P \cdot P^2 + b_P \cdot P + c_P = 0 \\ a_P = \left(\frac{K_1^2}{4} \cdot K_5 \cdot L^2 \right) \\ b_P = \left(1 + K_1 \cdot L + \frac{K_1^2}{4} \cdot L^2 + \frac{K_1^2}{4} \cdot K_5 \cdot n_2 \cdot r_T \cdot L^2 - \frac{K_1^2}{4} \cdot K_5 \cdot n_1 \cdot P_U \cdot L^2 \right) \\ c_P = -n_1 \cdot P_U \cdot \left(1 + K_1 \cdot L + \frac{K_1^2}{4} \cdot K_5 \cdot L^2 \right) \end{cases} \tag{2.40}$$

2.2.6 Modeling Cell Load (M1L, M2L)

Assumptions

Unless differently indicated, in the two-gene circuit considered in this work (Fig. 2.1a) we assume that only the protein with expression regulated by P_{Lux} is characterized by a relevant resource usage, while LuxR does not cause relevant burden. Since LuxR is constitutively expressed, thereby giving constant load, this assumption will not affect any HSL-dependent function, as previously discussed for genes with constant resource usage acting as background [84].

Derivations

A mechanistic model has been recently proposed to describe the effects of cell load caused by the expression of proteins with high resource demand [90, 91]. In such context of transcriptional/translational resource limitation, the synthesis rates of all the proteins of a synthetic

circuits are globally scaled by a factor D (Equation (2.41)).

$$D = 1 + \sum_{i=1}^{c} J_i \cdot s_i \tag{2.41}$$

where c indicates the number of expressed proteins in the synthetic circuit, s_i represents the synthesis rate of the i-th protein and J_i its resource usage parameter. Model derivation is extensively discussed in the original publications [90, 91], while in this study only the expression with lumped parameters is reported and used (Equation (2.41)).

Final Expression

The final expression of model output in presence of cell load is reported in Equations (2.42)-(2.44)

$$D(L, n_1, n_{2,b}) = 1 + E + J \cdot \left(\frac{\sigma + \gamma_X}{\sigma} \right) \cdot y(L, n_1, n_{2,b}) \tag{2.42}$$

$$n_{2,b}(L, n_1, n_{2,b}) = \frac{n_2}{D(L, n_1, n_{2,b})} \tag{2.43}$$

$$y_b(L, n_1, n_{2,b}) = \frac{y(L, n_1, n_{2,b})}{D(n_1, n_{2,b})} \tag{2.44}$$

where y_b and $n_{2,b}$ indicate the model output and the LuxR scale factor affected by cell load, y is the output of one of the models described in Sections 2.2.2-2.2.5, n_1 and n_2 have the same meaning as before, and J is the resource usage parameter (in min/AU) associated to the output protein. Finally, the contribution of an external constant load (E) was studied analogously, with the exception that the value of E was added to Equation (2.42).

2.3 Model comparisons on different assumptions

The effect of different model assumptions on the transfer functions shape was studied via numerical simulations. The considered assumptions were on LuxR abundance (Section 2.3.1), cell load (Section 2.3.2) and LuxR-HSL binding mechanism (Section 2.3.3). We evaluated if such assumptions exert a relevant contribution to output variation and, in some cases, if their inclusion contributes to the superior descriptive power of preliminary experimental data.

2.3.1 Effect of LuxR abundance assumption: M1 vs. M2

We compared the simulated outputs of M1 and M2, for different DNA/protein copy number situations, to evaluate the effect of the assumption of LuxR concentration abundance over P_{Lux} (Fig. 2.2). As expected from Equations (2.25) and (2.26) and previous works [81], for different values of R_{2T} (equal to $n_2 \cdot r_T$) the output curves generated by M1 showed diverse maximum (α) and switch point (κ) values, increasing and decreasing respectively as a function of R_{2T} (Fig. 2.2a,c,e). On the other hand, for increasing values of plasmid copy number (n_1), the values of α showed a linear increase and κ remained constant (Fig. 2.2b,d,f) as expected from Equation (2.24). In both cases, the Hill coefficient was always equal to 1 as expected (Fig. 2.2g,h). While the M1 model was able to capture the effects of LuxR level variation in some experimentally tested model systems, as previously demonstrated [81], the effect of changes in plasmid copy number are intuitively non-realistic since M1 assumes that the concentration of LuxR is much higher than the one of the promoter, thereby resulting in an unlimited increase of protein synthesis rate for high

DNA copy numbers, with unchanged shape of the curve in terms of κ and η. Since LuxR may be expressed over a wide range of values to tune the sensitivity of the inducible device [78], the removal of its abundance assumption can be of interest and led us to develop M2.

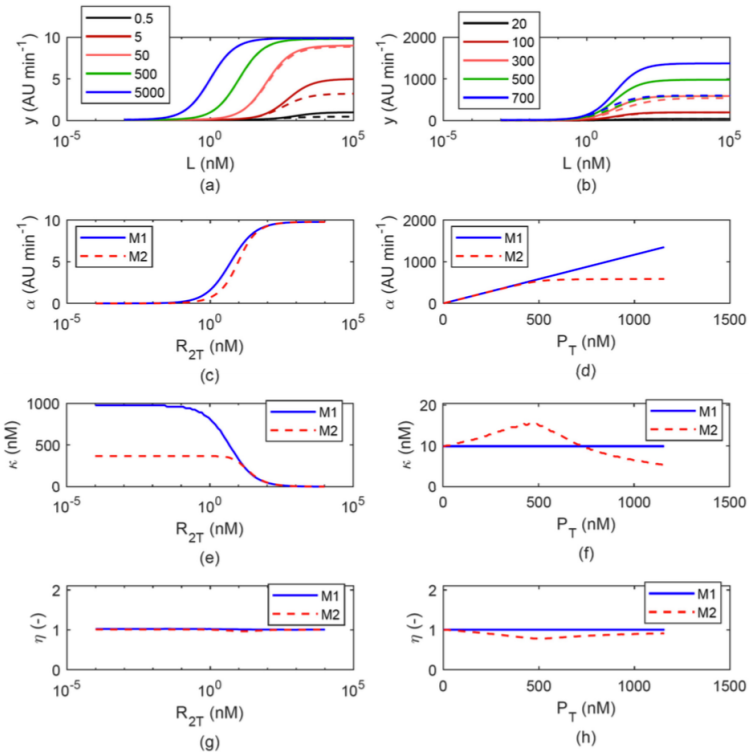

Figure 2.2: **Comparison between M1 and M2.** Panels (**a**) and (**b**) report the output activation curves of M1 (solid lines) and M2 (dashed lines) for different values of R_{2T} (panel (**a**), as indicated in the legend, expressed as nM) and P_T (panel (**b**), as indicated in the legend, expressed as per cell copy number). The (**c**)-(**d**), (**e**)-(**f**), and (**g**)-(**h**) panel pairs report the $R_{2T}-$ and P_T-dependent trend of α, κ and η, respectively, with solid and dashed lines representing M1 and M2, respectively. In panels (**a**),(**c**),(**e**),(**g**) the promoter copy number value was set to 5, while in panels (**b**),(**d**),(**f**),(**h**) the LuxR concentration was set to 500 nM.

The M2 model was investigated in the same situations considered

for M1 (Fig. 2.2). Upon changes of LuxR values, the relationship between R_{2T} and α or κ were qualitatively analogous to M1. However, when R_{2T} decreases below the promoter concentration value (about 8 nM) the output of M2 showed both lower α and κ levels compared to M1, with the decrease of κ showing the highest-entity effect (>2-fold with the used parameters, Fig. 2.2e). When plasmid copy number is varied in a range below a constant LuxR concentration value (500 nM), α still showed a linear increase as observed in M1, but, differently from M1, κ showed a low-entity increase (less than 2-fold). For concentrations of promoter higher than LuxR, the α value showed saturation, intuitively bacause all the P_{Lux} promoters in the cell cannot be occupied by limiting amount of LuxR, and as a result RFP synthesis cannot increase anymore for higher concentrations of promoter. Although in this latter condition RFP maximum expression is constant, an increase in plasmid copy number results in a decrease of κ (about 2-fold). Also, in the M2 model, the Hill coefficient was equal to 1 upon R_{2T} variation. However, interestingly, it decreased to values slightly lower than 1 (up to 0.8) upon variations of plasmid copy number (Fig. 2.2g,h). The illustrated results demonstrate that the removal of the LuxR abundance assumption can affect all the parameters of a Hill function, even when R_{2T} and P_T are tuned over ranges of values not violating this assumption.

2.3.2 Effects of cell load

The M1L model was simulated to investigate the effects of cell load, which was assumed to derive from RFP expression alone, or from both RP and a constant load outside the inducible circuit, caused by the expression of another heterologous protein (Fig. 2.3).

If the load was caused by RFP, its expression affected both LuxR and RFP itself when induction levels became high upon HSL addition. As expected, the maximum level of RFP expression reached by

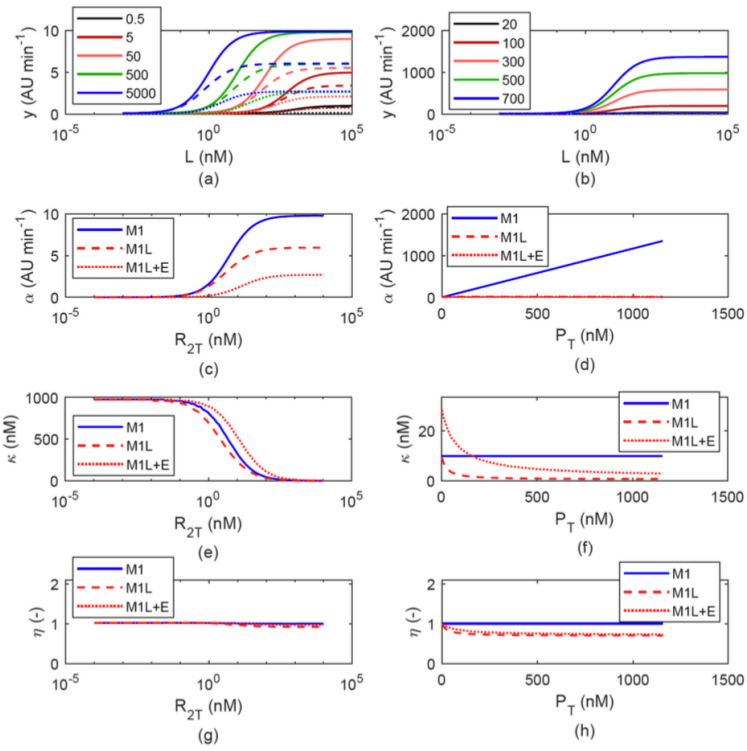

Figure 2.3: **Comparison between M1 and M1L.** Panels (**a**) and (**b**) report the output activation curves of M1 (solid lines) and M1L (dashed lines) and M1L with external load, indicated as M1L+E (dotted lines), for different values of R_{2T} (panel (**a**), as indicated in the legend, expressed as nM) and P_T (panel (**b**), as indicated in the legend, expressed as per cell copy number). The (**c**)-(**d**), (**e**)-(**f**), and (**g**)-(**h**) panel pairs report the $R_{2T}-$ and P_T-dependent trend of α, κ and η, respectively, with solid, dashed, and dotted lines representing M1, M1L, and M1L+E, respectively. In panels (**a**),(**c**),(**e**),(**g**) the promoter copy number value was set to 5, while in panels (**b**),(**d**),(**f**),(**h**) the LuxR concentration was set to 500 nM.

M1L was lower than the respective levels reached by M1 (Fig. 2.3c,d) upon changes of both R_{2T} and plasmid copy number. In particular, the increase of DNA copy number resulted in a saturated maximum RFP expression, which was much lower than in the no-burden model (Fig. 2.3d). The burden effect on κ resulted in a slight decrease com-

pared with M1 upon R_{2T} variation (Fig. 2.3e), while the decreasing effect on κ was more relevant upon plasmid copy number increase (Fig. 2.3f). Assuming that a constant load (E) also affects protein expression, α showed a further decrease, and κ showed an increase compared with M1, that is, activation curves showed a systematically higher switch point than in a situation without load. In all the described cases, the Hill coefficient showed values slightly lower than 1 upon plasmid copy number increase and almost unchanged values (close to 1) in case of R_{2T} variation. We then analyzed the *in vivo* data from a previously published experiment, in which the activation curve of a medium copy plasmid-borne lux inducible device was characterized in absence and presence of a second co-transformed plasmid, considered as an additional constant load for the host [88]. Transfer functions were measured in four conditions (with/without load, low/high expression of IPTG-driven LuxR) and a summary of the resulting Hill function parameters is reported in Table 2.1. As the M1L model predicts, the additional load affects the transfer function parameters upon both low and high LuxR expression level conditions: α showed a decrease in presence of E, while κ increased. It is worth noting that a simpler model including cell load only on RFP expression (and not on LuxR), as previously adopted to analyze *in vivo* data [84], fails to describe the joint increase of α and decrease of κ upon LuxR overexpression, captured by the M1L model and observed in experimental data resembling the modeled situation (data not shown). Analogous results were obtained from the simulations via M2L (data not shown). The reported simulations showed that cell load can quantitatively affect all the Hill parameters of the activation curves, thereby demonstrating the relevance of model assumptions in the analysis of dose–response curves of inducible systems.

Table 2.1: **Parameters estimated from *in vivo* experiments.** Measurements were carried out using MG1655-Z1 as host strain. [a]Percent expression, relative to the maximum expression value obtained in the same study; [b]BioBrick™ construct BBa_J107063 in the pSB3K3 medium-copy vector, IPTG-inducible LuxR expression cassette; [c]Additional load provided by the OL1 low-copy plasmid, described in the same paper; [d]BioBrick™ construct BBa_J107032 in the pSB4C5 low-copy vector, ATc-inducible LuxR expression cassette.

Condition	$\alpha(\%)^a$	$\kappa(nM)$	$\eta(-)$	Reference
Medium copy[b], no IPTG	47	194.01	1.01	[88]
Medium copy + E[c], no IPTG	27	474.1	0.98	[88]
Medium copy, IPTG = 500 μM	100	0.77	1.54	[88]
Medium copy + E, IPTG = 500 μM	68	0.99	1.29	[88]
Low copy[d], no ATc	83	34.29	0.98	This study
Low copy, ATc = 2.5 ng/ml	92	11.32	1	This study
Low copy, ATc = 5 ng/ml	97	3.73	1.15	This study
Low copy, ATc = 50 ng/ml	100	1.52	1.39	This study

2.3.3 Evaluation of LuxR-HSL complex formation assumptions: M(1-2) vs. M(1-2)T

By assuming that the LuxR dimer has two binding sites for HSL molecule, we defined models including a hetero-tetramer formation step (M1T and M2T, depending on the LuxR abundance assumption as above). We also assumed a non-cooperative behavior for the HSL ligand binding, i.e., the two successive HSL binding events occur with the same probability and $K_4 = K_1/4$, as described in Equations (2.34)-(2.36). Considering the Adair equation (Equation (2.45)), describing the fraction (F) of HSL-bound sites of LuxR over the total number of sites [92], the resulting Hill coefficient (computed as in Equation (A.3)) is always 1 under the non-cooperativity assumption, for any parameter and R_{2T} value

$$F = \frac{Q + 2 \cdot Q_2}{2 \cdot R_2 + 2 \cdot Q + 2 \cdot Q_2} = \frac{1}{2} \cdot \frac{K_1 \cdot + 0.5 \cdot K_1^2 \cdot L^2}{1 + K_1 \cdot L + 0.25 \cdot K_1^2 \cdot L^2} \quad (2.45)$$

Differently from Equation (2.45), the expressions of Q_2 (which can be calculated from Equations (2.17), (2.31) and (2.33)) and y show LuxR level-dependent Hill coefficients. For both Q_2 and y, such features can be observed in the closed-form expressions which can be obtained from M1T (Section 2.2.5). In particular, model simulations of M1T showed that the Hill coefficient increases as a function of R_{2T}, which represented the main difference from the respective model without hetero-tetramerization assumption (M1), while the other parameters showed analogous trends as above (Fig. 2.4). The simulation of M2T also showed this feature on the Hill coefficient, together with the same trends described in Section 2.3.1 due to removal of LuxR abundance assumption.

Although a number of previous experimental works on the lux system showed a Hill coefficient around 1 or slightly lower in the tested conditions [81, 93], in some works a higher number was reported [88, 94]. While Hill coefficient values lower than 1 could be due to burden effects and/or violation of the LuxR abundance assumption, as described by the models illustrated in Sections 2.3.1 and 2.3.2, higher values could not be described by those models in any tested case. Using preliminary data from a previous study [88] and a novel *ad hoc* experiment (Table 2.1), we showed that the Hill coefficient increases upon increase of LuxR level, consistently with the simulations of M1T and M2T. By also considering cell load in these more detailed models (obtaining the M1TL and M2TL models), the same trends and conclusions illustrated in Section 2.3.2 could be observed (data not shown). Importantly, the resulting Hill coefficient value, as well as the other parameter values, could be tuned by the joint contribution of different effects, e.g., cell load, LuxR abundance assumption violation, and hetero-tetramerization, with the latter exerting an increase of η for increasing R_{2T} levels, and the other effects causing a decrease of its value.

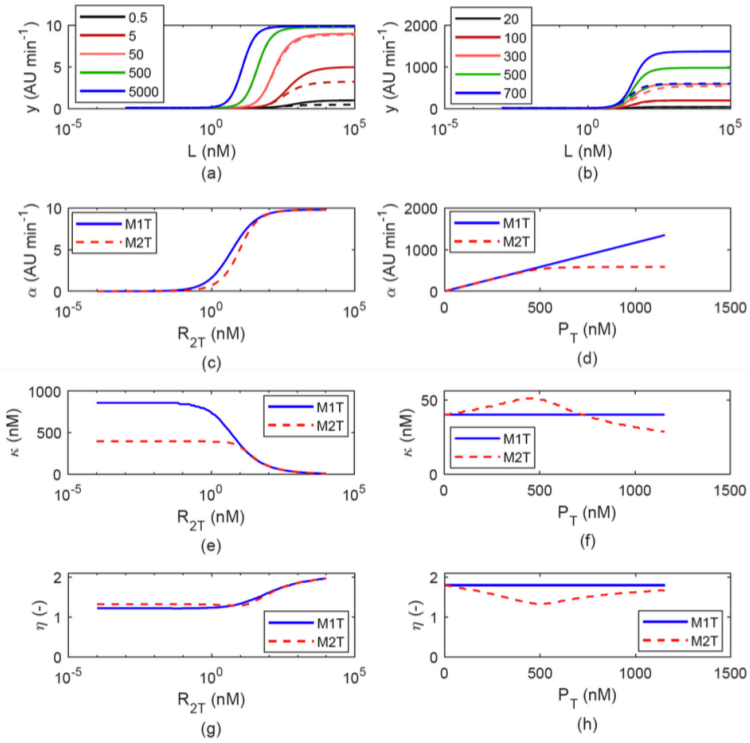

Figure 2.4: **Comparison between M1T and M2T.** Panels (**a**) and (**b**) report the output activation curves of M1T (solid lines) and M2T (dashed lines) for different values of R_{2T} (panel (**a**), as indicated in the legend, expressed as nM) and P_T (panel (**b**), as indicated in the legend, expressed as per cell copy number). The (**c**)-(**d**), (**e**)-(**f**), and (**g**)-(**h**) panel pairs report the $R_{2T}-$ and P_T-dependent trend of α, κ and η, respectively, with solid and dashed lines representing M1T and M2T, respectively. In panels (**a**),(**c**),(**e**),(**g**) the promoter copy number value was set to 5, while in panels (**b**),(**d**),(**f**),(**h**) the LuxR concentration was set to 500 nM.

2.3.4 Model identifiability

The usability of the M1, M2, M1T, and M2T models was evaluated by studying their structural and practical identifiability, to eventually understand if their parameters can be reliably estimated and which experiments are needed for this task.

M1

The M1 model is a priori identifiable: two experiments, in which the transfer function is measured for different LuxR levels (tuned by n_2), are needed to properly estimate model parameters in Equation (2.23). In particular, assuming that the Michaelis – Menten function parameters defined in Equations (2.24)-(2.26) are known from the fitting of experimental data, the following expressions can be written (Equations (2.46)-(2.51))

$$\overline{\delta} = P_U \cdot n_1 \cdot \widehat{k}_{m0} \tag{2.46}$$

$$\overline{\alpha} = \frac{P_U \cdot n_1 \cdot (\widehat{k}_{m0} + \widehat{k}_{mL}}{1 + 1/(\widehat{K}_3 \cdot \overline{n_2})} \tag{2.47}$$

$$\overline{\kappa} = 1/(K_1 \cdot (1 + \widehat{K}_3 \cdot \overline{n_2})) \tag{2.48}$$

$$\dot{\delta} = P_U \cdot n_1 \cdot \widehat{k}_{m0} \tag{2.49}$$

$$\dot{\alpha} = \frac{P_U \cdot n_1 \cdot (\widehat{k}_{m0} + \widehat{k}_{mL})}{1 + 1/(\widehat{K}_3 \cdot \overline{\overline{n_2}})} \tag{2.50}$$

$$\dot{\kappa} = \frac{1}{K_1 \cdot (1 + \widehat{K}_3 \cdot \overline{\overline{n_2}})} \tag{2.51}$$

where the single and double bars on Hill function parameters and copy numbers indicate the parameters of the first and second experiment, respectively. These expressions demonstrate that LuxR must not be very low or very high (compared with $1/\widehat{K}_3$) in both experiments to enable identifiability ($n_2 << 1/\widehat{K}_3$ or $n_2 >> 1/\widehat{K}_3$). These expressions also demonstrate that two or more experiments in which n_1 changes

(while keeping n_2 constant) do not lead to a structurally identifiable model. The a posteriori identifiability was also confirmed by fitting synthetic data generated from parameters in Table 1 (for $n_2 = 50$ and 500 AU, and $r_T = 1$) and a realistic number of data points. As expected, this model is practically identifiable since it led to reasonably low estimation errors and CV for its four unknown parameters (Fig. 2.5).

M2

The a priori identifiability of the M2 model could not be studied due to its complex expression. Only a posteriori identifiability could be addressed. Since M2 has the same parameters as M1 with the addition of r_T (which cannot be estimated separately from K_3 in M1), its proper estimation in M2 is intuitively possible only when LuxR is not abundant compared with promoter concentration, otherwise the M2 expression would become identical to M1. Accordingly, the synthetic experiments were simulated with $n_2 = 0.005, 0.5$ and 5 AU (with $r_T = 100nM/AU$). The simultaneous estimation of its five parameters did not lead to a structurally identifiable model due to the high REE (detailed explanation in A.3) and CV, even by considering more activation curves with different LuxR levels (e.g., $n_2 = 0.05$ which was added to the ones above - data not shown). For this reason, we investigated a two-stage procedure in which synthetic data obtained by setting $n_2 = 0.5$ and 5 AU (corresponding to $R_{2T} = 50$ and $500nM$) were fitted with M1 (first stage). This fitting is expected to provide reliable estimates (as shown in Section 2.3.1) because LuxR is highly abundant and the assumptions of M1 are not violated. Since M1 and M2 share the same $\widehat{k}_{m0}$, $\widehat{k}_{mL}$ and K_1 parameters, their estimated values were fixed in the second stage, in which the M2 model was used to fit the data with $n_2 = 0.005$ (corresponding to $R_{2T} = 0.5nM$) and, as before, 0.5 and $5AU$, to estimate K_3 and r_T, exploiting at least

one condition in which LuxR is not abundant compared with the P_{Lux} promoter. This procedure led to a structurally identifiable condition for the M2 model, since it could estimate its parameters with reasonably low estimation error and CV, even if it had higher REE compared with M1 but much lower CV (Fig. 2.5).

M1T

The M1T model is a priori identifiable from only one experiment. In particular, assuming that the coefficients of the rational function in Equation (2.37) are known (Equation (2.52)), the Equations (2.53)-(2.57) system can be solved, with the only constraint that no solution is obtained for $n_2 << 1/K_5$

$$y = \frac{a' + b' \cdot L + c' \cdot L^2}{1 + d' \cdot L + e' \cdot L^2} \tag{2.52}$$

$$a' = P_U \cdot n_1 \cdot \widehat{k}_{m0} \tag{2.53}$$

$$b' = P_U \cdot n_1 \cdot \widehat{k}_{m0} \cdot K_1 \tag{2.54}$$

$$c' = P_U \cdot n_1 \cdot \widehat{k}_{m0} \cdot K_1^2/4 + P_U \cdot n_1 \cdot \widehat{k}_{mL} \cdot K_1^2/4 \cdot \widehat{K}_5 \cdot n_2 \tag{2.55}$$

$$d' = K_1 \tag{2.56}$$

$$e' = K_1^2/4 + K_1^2/4 \cdot \widehat{K}_5 \cdot n_2 \tag{2.57}$$

The a posteriori identifiability was investigated by using only one acti-

vation curve with $n_2 = 50AU$ (with $r_T = 1$). Despite the identifiability was successfully confirmed, the REE and CV were both higher than in M1 (Fig. 2.5). We also investigated the availability of a second activation curve with $n_2 = 500AU$, as in the case of M1. In this case, REE and CV (Fig. 2.5) were systematically lower than in M1T with one LuxR level and also in M1.

M2T

Since the M2T model structure did not enable the study of a priori identifiability, only the a posteriori one was investigated. Similar consideration to M2 also persist for M2T when different activation curves were fitted with this model to estimate its five parameters, leading to an a posteriori non-identifiability (data not shown). For this reason, the same two-stage identification procedure described in Section 2.3.2 was adopted, with the exception that the first estimation step was carried out via M1T to estimate its four parameters, and then M2T was used to estimate K_5 and r_T from the synthetic data as described in Section 2.3.2 by fixing the $\widehat{k}_{m0}$, $\widehat{k}_{mL}$, and K_1 parameters to the previously estimated values. Results showed that, following the described procedure, the model is a posteriori identifiable with low REE and CV (Fig. 2.5).

2.4 Final considerations

Accurate predictive mathematical models are needed to support the bottom-up design of complex biological systems in synthetic biology. In an effort towards the development of computational tools to overcome this need, different mechanistic models were herein proposed for the lux inducible system. Their mechanistic structure was expected to increase the details of the described system, thereby mak-

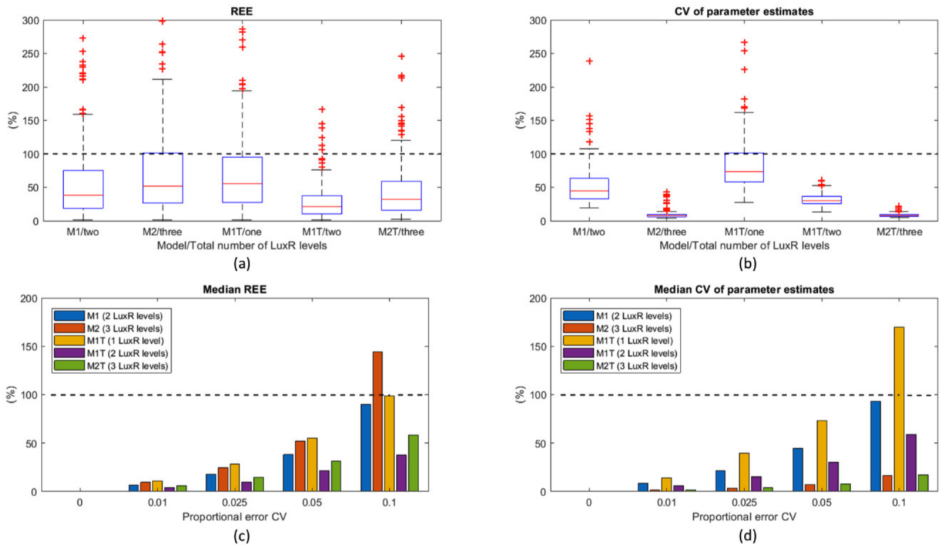

Figure 2.5: **Relative estimation error (REE) and uncertainty of parameter estimates (CV) in a posteriori identifiability.** Panels (**a**) and (**b**) report the distribution of REE (**a**) and CV (**b**) among 200 runs starting from different simulated data, with a 5% proportional error. The model and the number of LuxR levels included in the fitted data are specified for each boxplot. The red line represents the median of the distribution. Panels (**c**) and (**d**) report the median of the REE (**c**) and CV (**d**) distribution as a function of the proportional error entity, from 0 to 10%. The models and specific LuxR levels are described in the legend. In all the panels, the dashed horizontal line indicates the 100% value to facilitate the interpretation of the graphs.

ing the model more generalizable to context variations than traditional empirical equations. Eight different mechanistic models, based on different assumptions on regulatory protein abundance, ligand binding, and resource usage (and their combinations), were studied *in silico* and compared, with the final goal of understanding the impact of the underlying assumptions on the transfer function of the circuit. One of the models was strongly inspired by a previously published work [81], while the others represent novel computational tools. The different assumptions summarized above affected the simulated output of

the models (i.e., dose–response curve of recombinant protein production as a function of inducer concentration), thereby demonstrating that significant diversity in model output could be generated by different features. The initially considered model (M1) assumed that one HSL binding event to the LuxR dimer occurred for the activation of the complex, and that LuxR concentration was much higher than the one of P_{Lux}, with a resulting mathematical expression equivalent to a Michaelis – Menten function model. When the assumption on the LuxR abundance was removed (M2), the model gave different outputs compared with M1 when the assumption was violated. In particular, the increase of P_{Lux} copy number was predicted to result into a linearly increasing expression by M1, while M2 showed a saturating trend, thereby demonstrating that the removal of this simplifying assumption can lead to an intuitively more realistic behavior of the inducible system upon DNA copy number changes. In this situation and with the structural parameter values used in our study, the output showed maximum activity variations up to 2.3-fold between M1 and M2 for biologically plausible values of DNA and protein concentrations. When a limited resource framework was assumed (M1L), the expression of LuxR and output protein were globally affected by cell load due to the expressed output protein (in an HSL-dependent fashion) and/or to a load outside the inducible circuitry. As previously demonstrated *in vivo* and *in silico*, the output of inducible systems is affected by the resource usage of the expressed proteins [84]. In the M1L model, we showed that an external load causes a decrease in maximum output and an increase of the half-maximum HSL concentration. This result could not be predicted by some of the traditionally used models in which the transcriptional regulator is assumed to be constant and is not modeled [84, 95]. In this situation and with the structural parameter values used in our study, the output showed parameter variations up to 2.5-fold between M1 and M1L for biologically plausible values of resource usage parameters, and DNA and protein

concentrations. The predicted effects were consistent with previously published experimental data of our group, in which the same inducible system with and without cell load (due to the presence of an additional co-transformed plasmid) showed the same effect on maximum activity and activation curve sensitivity [88]. In both M2 and M1L models, we also showed that the Hill coefficient of the output curve could decrease (compared to 1, i.e., the one of M1) up to 0.8 with the parameters used in our study. When a different ligand binding reaction was considered (M1T), i.e., two HSL molecule binding to the LuxR dimer non-cooperatively, the output expression introduced a power over the HSL concentration term. As a result, output curves had a Hill coefficient greater than one, even if the described binding mechanism was assumed to be non-cooperative. To our knowledge, this is the first study explicitly highlighting that a Hill coefficient greater than one could occur in absence of cooperative binding in the transcriptional activator. This effect was consistent with previously published experimental data from our group [88] and others [94], as well as novel preliminary experimental data explicitly measured in this work: the Hill coefficient of the output curve increased as a function of LuxR level. In our M1T model, an increase up to 2-fold was observed for the Hill coefficient compared to M1, which relied on a different assumption on HSL binding. The behavior of the remaining models, including combinations of the described assumptions (M2L, M2T, M1TL, M2TL), showed more complex features, but similar conclusions on the effects of the investigated assumptions could be drawn. The M1, M2, M1T, and M2T models were also studied in terms of usability, by investigating their identifiability, to eventually understand if their parameters could be estimated from experimental data and which experimental design is recommended. In fact, in addition to simulation, parameter estimation is a crucial step in model usability that enables the re-use of well-characterized regulatory components in synthetic biology. All the models enabled parameter estimation with reasonably low error

and a few constraints (described in Section 2.3). In general, as expected, by increasing the random noise affecting experimental data REE and CV increase. As for the individual models, M1 required two experiments with different LuxR levels, while M1T required only one experiment with a single LuxR expression value, despite its parameters could be estimated much more reliably by adding a second experiment, like in the M1 case. The M2 and M2T models required an additional experiment, compared to M1 and M1T, respectively, to be properly identified, in which a curve obtained with a LuxR level that is not much higher than P_{Lux} concentration had to be measured. Considering a proportional error model for the generated data (5% CV), all the models enabled the estimation of structural parameters within 2-fold compared with the true value (considering interquartile ranges of REE distribution). The M2 model showed the highest estimation error (>100%, median among 200 runs with different datasets simulated with a 10% proportional error), making it the less robust model among the tested ones in estimation tasks. The M1T model with a single LuxR level showed similar drawbacks, which could be overcome by adding activation curve data with more LuxR levels, while M2 could not be improved following the same procedure. If used to fit the data in real experimental works, a major advantage of the M2 and M2T models is that they potentially enable the estimation of the actual LuxR intracellular concentration, which could not be estimated with the LuxR abundance assumption (M1 and M1T). However, the simultaneous estimation of all the parameters of M2 and M2T failed, thereby leading to the definition of an alternative procedure for the identification of such models: first, M1 (or M1T) was identified by fitting two activation curves data that conformed to the LuxR abundance assumption to estimate all model parameters; then, M2 (or M2T) was identified by fixing three parameters, previously estimated via M1 (or M1T), and estimating the two remaining ones by fitting the two activation curves data used in the first stage, together with an additional

curve obtained for a low LuxR level. Given the same number of experiments, M1T and M2T could be identified with lower estimation error and parameter uncertainty than M1 and M2, respectively. The identifiability of models including cell load was not herein addressed, but their usability with real data was investigated previously [84]; they are not expected to include additional identifiability issues, since resource usage parameters could be estimated separately if required [84, 96] and their identification can rely on a second output of the system, i.e., a cell burden monitor which acts as a proxy of cell load. A number of limitations may affect the usability and predictive performance of the studied models against *in vivo* data. In fact, although mechanistic details have been herein added to traditional models to improve their generalization performance, other assumptions may still be inaccurate in capturing the real behavior of a synthetic circuit. Among the potential crucial aspects, it is worth noting that cell systems are inherently stochastic and when the reacting molecules are present at small intracellular copy numbers stochasticity can result in large fluctuations in the behavior of single cells in a population. In addition, the non-cooperativity of the inducer-regulator binding, herein assumed, should be further investigated. More in general, despite preliminary data have been used to confirm some of the transfer function variations found in this work, a larger-scale experimental validation should be required to investigate and confirm the described effects. Such effort should experimentally validate the impact of the individual assumptions. The different LuxR-HSL binding assumptions will need such validation to select the one best describing experimental measures. However, the validation of the other assumptions (LuxR abundance and cell load) should not lead to the selection of a best-performing model since it is expected to be application-specific, e.g., a no-burden model may have high predictive power if all the genes of a circuit have low resource usage. In summary, we have proposed different usable mechanistic models that had a significant impact on the predicted output of an in-

ducible system, and they were also consistent with a set of preliminary experimental data. In the future, the reported models may support synthetic circuits output prediction in practical situations with unprecedented details, also facilitating the bottom-up design of complex circuits due to their generalization power. In this framework, highly relevant applications of such models in synthetic biology are, e.g., the prediction of circuit output as a function of unseen DNA, protein and inducer concentrations; estimation of protein regulator abundance; investigation of different binding mechanisms for subsequent model selection. The proposed model definition and analytic procedures may be used to study other systems different from the LuxR/P_{Lux} module, considered in this work. The underlying binding reactions could be known, or they might be investigated by testing different model assumptions against experimental data for model parametrization and mechanistic understanding.

Chapter 3

Sad-iFFL: Improved iFFL network based on the CRISPRi system from *Staphylococcus aureus*

The focus of synthetic biology on new microorganisms leads several research groups on the rational design of new DNA-encoded controllers in order to achieve stable and predictable protein expressions, indipendent from the host machinery and thus indipendent from the three main gene expression process parameters that are affected by context-dependent variation: transcription rate [56], translation rate [64] and copy number variations [65]. In order to provide a genetic system able to work in any bacteria with high probability, a new circuit design, herein called Sad-iFFL, has been developed based on the incoherent feedforward-loop (iFFL) network (described in Section 3.1), using a novel repressor enzyme: *Staphylococcus aureus* dCas9, which has different advantages compared with current solutions. The biological circuit description and the mathematical model subsequently developed are discussed in Section 3.2; the latter has also been analyzed to investigate its theoretical working constraints and the rejection ca-

pability of the three aforementioned parameters has been discussed. Subsequently, the dynamic and steady-state analysis of the Sad-iFFL circuitry has been characterized *in silico* and compared with the open-loop scheme (detailed in Appendix B) in terms of output stability due to noise propagation, robustness on parameter variations and settling time. In order to mimic different hosts scenarios, the designed circuit has been tested *in vivo* with different combinations of regulatory parts and its performances has been discussed in Section 3.3.

3.1 Introduction

The gene networks encoded in bacterial DNA exert gene expression regulation in two main ways in open-loop, in which regulatory sequences (e.g., promoter, RBSs, terminators) are solely responsible for the gene expressions process of each sequence or adopting regulatory subnetwork which include ubiquitous network motifs and exert a robust control of the expression of a defined set of genes. The interest by Synthetic Biology to create a stable and robust circuit that could work in the highly variable and dynamic intracellular environment brought the control theory discipline to attention, in order to apply such engineering concepts to the gene regulation process [10]. In the set of network motifs discovered so far, one of the most represented within the bacterial cell is the feedforward loop. They are subject to several engineering constraints, including that (i) they are finely-tuned so that the system returns to the original steady state after a disturbance occurs (adaptation), (ii) they are typically implemented in the combination with negative feedback, and (iii) they can greatly improve the stability and dynamical characteristics of the conjoined negative feedback loop. On the other hand, in biology, these network loops can serve many purposes, one of which may be the implementation of robust control schemes against environmental perturbations or

cell-to-cell variability [97]. Indeed, as reported in Section 1.2, a version of feedforward loop (FFL), called incoherent-feedforward loop (iFFL), has been used to make a gene of interest indipendent from the parameter desirable to reject (e.g., copy number variation [65], [67]). The basic iFFL network is reported in Fig. 3.1A where the signal X regulate positively and negatively, through the entity Y, the output Z. The opposite effect exerted by X on Z gave the suffix 'incoherent' to the FFL network motif. The network presented in this work is based on the iFFL scheme in order to control the expression of a gene of interest (GOI) rejecting the context-dependent variation on the three main parameters of the gene expression process: transcription rate, translation rate and copy number. The logic scheme is shown in Fig. 3.1B.

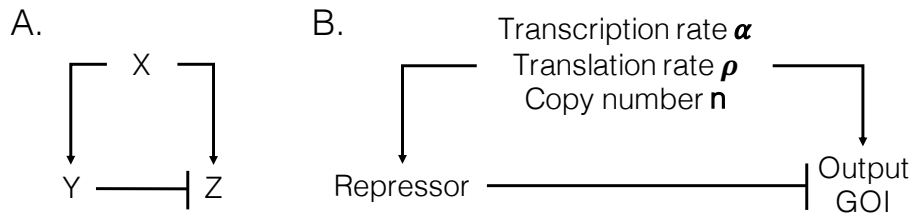

Figure 3.1: **Incoherent feedforward-loop (iFFL) network.** **A**, Incoherent (type I) feedforward network can, under some assumptions on the repressor protein Y, make the output Z independent from the signal X. **B**, Desing of iFFL-based network to obtain the perfect adaptation of a gene of interest (GOI) to: transcription rate α, translation rate ρ and copy number n variations.

3.2 Sad-iFFL model-based design

3.2.1 Circuit description

The architecture of the iFFL investigated in this work (Sad-iFFL) includes the GOI and a repressor coding sequence (SadCas9) in dif-

ferent expression cassettes under the control of identical promoters and RBSs, and in the same copy number (Fig. 3.2). The SadCas9 protein forms a complex with a sgRNA that is programmed to bind a DNA region downstream of the GOI promoter, thereby repressing GOI transcription. Differently from the approach usually followed for synthetic circuits reported in the literature in which the repressor complex (dCas9:sgRNA) is programmed to bind the promoter region [69], in this study the target region has been inserted immediatly next to the transcription start site (TSS) of the promoter. Indeed, due to the hypothesis that the promoter sequence is the same for all the components within the circuit (e.g., SadCas9, GOI, sgRNA), it is possible to avoid undesirable network regulations (e.g., negative-feedback loop of SadCas9) that would break the logic of the genetic controller. For a sufficiently high expression level of the repressor, this system is expected to be robust against transcription, translation and copy number variations. For these reasons, this kind of circuit could be able to achieve a predictable target protein level in different host strains and under different perturbed environmental conditions and cellular or genetic contexts. As reported in the literature, the perfect adaptation theoretically occurs only if the cooperativity of the repressor used in the circuit is equal to 1 [65], characteristic of a regulator for which only one binding event can occur. This requirement can be met by design via the dCas9 enzyme that can bind a single DNA sequence in its target.

In order to increase the portability of the circuit scheme through different bacterial hosts, a novel CRISPR-family repressor enzyme (dCas9) from *Staphylococcus aureus*, called SadCas9, has been chosen. Compared with the well-known and characterized counterpart from *Streptococcus pyogenes* dCas9 (SpydCas9), it has promising characteristics, in terms of gene length and specificity. The main important feature for which SadCas9 can be superior for working in different bacterial hosts is the PAM sequence recognition site which, based on

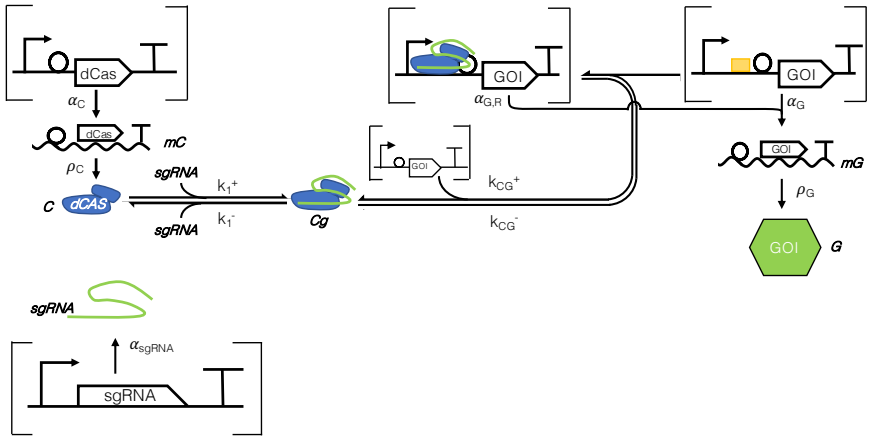

Figure 3.2: ***S. aureus* dCas9 incoherent feedforward loop (Sad-iFFL) biological scheme.** Visual representation of the iFFL gene expression control system. The network functionality is based on the dCas9 (C) repressor, also called SadCas9 in the text due to its presence in the genome of *S. aureus*. Once activated by its binding with the specific sgRNA, the resulting complex can bind the target region (yellow box) and repress the transcription of the downstream gene of interest (GOI) thereby decreasing the final protein level (G). The curved arrow, the circles and the T-shaped lines are respectively the promoter, RBS and terminator parts. The horizontal line inside square brackets represents the DNA while the wavy one the mRNA.The assosiation/dissociation processes (e.g., k_1^+,k_1^-,k_{CG}^+,k_{CG}^-) are represented by bi-directional arrows while the one-way production rates with monodirectional arrow (e.g., transcription rate α_X, translation rate δ_X, where X is reference to the biological entity). The biological scheme lacks of the degradation constants (e.g., mRNA degradation rate, protein degradation rate), which are considered in the associated mathematical model in Section 3.2.1), due to graphical reasons.

its nucleotide composition, is more specific than the one of SpydCas9 (5'-NNGRRT-3' - SadCas9, 5'-NGG-3' - SpydCas9) [74]. The higher specificity reflects a lower probability of off-target binding events in the genome and thus a faster kinetics due to the fact that the enzyme has to 'search-open-check' in a smaller set of targets [72]. The smaller coding sequence (3159 nt compared to 4104 nt of SpyCas9), apart from making DNA-assemby and verification procedures easier, also allows to engineer mammalian cells through Adeno-Associated Viruses (AAV) vector since the latters are able to contain SaCas9 but

not SpyCas9 [98]; this scenario opens the demand for a quantitiative characterization of this so far not widely used enzyme repressor which has been faced in this work by focusing on its use in synthetic circuits.

3.2.2 Mathematical model of Sad-iFFL controller

The biological scheme reported in Fig. 3.2 (Sad-iFFL) has been modeled considering the law of mass action and the law of conservation of mass described in Equations (3.1) and (3.2), respectively, to capture the transcription-, translation- and copy number-dependent effects on the expression of GOI.

$$\begin{cases} C + sgRNA \underset{k_1^-}{\overset{k_1^+}{\rightleftharpoons}} Cg \\ Cg + P_{G,Free} \underset{k_{CG}^-}{\overset{k_{CG}^+}{\rightleftharpoons}} P_{G,Repressed} \end{cases} \tag{3.1}$$

$$\begin{cases} DNA : P_{G,Free} + P_{G,Repressed} = n \\ dCas : C + Cg + P_{G,Repressed} = C_{tot} \\ Single\ guide\ RNA : sgRNA + Cg = g_{tot} \end{cases} \tag{3.2}$$

In Equations (3.1) - (3.2), C, Cg and $P_{G,repressed}$ are the concentrations ($[nM]$) of dCas respectively free, coupled with the single guide RNA and bounded to the target promoter, while $sgRNA$ and P_{Free} are the concentrations ($[nM]$) of free single guide RNA and unbounded promoter inside the cell. The kinetics of repressor complex formation (Cg) and its activity towards the target promoter ($P_{G,Repressed}$) are described respectively with the association rate constants k_1^+ $[nM^{-1}time^{-1}]$, k_{CG}^+ $[nM^{-1}time^{-1}]$ and dissociation rate constants k_1^- $[time^{-1}]$, k_{CG}^- $[time^{-1}]$ constants (Equation (3.1)). The total amount of target promoter, dCas and single guide RNA inside the cell

expressed with the $n[nM]$, $C_{tot}[nM]$ and $g_{tot}[nM]$ parameters in Equation (3.2).

Assuming the level of the target promoter negligible compared to the total dCas9 repressor complex (Hypothesis 1: $Cg >> P_{G,Repressed}$) and the latter negligible compared with the total level of single guide RNA (Hypothesis 2: $sgRNA >> Cg$), the mathematical model of Sad-iFFL circuit can be written as:

$$\frac{dC}{dt} = \frac{n \cdot \rho_C \cdot \alpha_C}{(d_{mC} + \mu)} - (d_C + \mu) \cdot C \tag{3.3}$$

$$\frac{dG}{dt} = \frac{n \cdot \rho_G}{(d_{mG} + \mu)} \cdot \left(\frac{\alpha_G}{(1 + \frac{Cg}{K_{CG}})} \right) - (d_G + \mu) \cdot G \tag{3.4}$$

$$\frac{dsgRNA}{dt} = n \cdot \alpha_{sgRNA} - (d_{sgRNA} + \mu) \cdot sgRNA \tag{3.5}$$

$$Cg = \frac{C}{(1 + \frac{K_1}{sgRNA})} \tag{3.6}$$

No cooperativity was assumed (Hill coefficient $\eta = 1$ − not shown in the model) for the complex formation from sgRNA binding [99] (Equation (3.5)) and for the transcriptional repression from the complex Cg [100] (Equation (3.4)). In Equations (3.3) − (3.5) α_X, ρ_X and δ_X describe the transcription, translation and degradation rate $[time^{-1}]$ of the biological entity X, respectively, while μ is the diluition rate $[time^{-1}]$ due to cell division. In Equation (3.4) and (3.6), $K_1[nM]$ and $K_{CG}[nM]$ are the Michaelis−Menten equilibrium dissociation constants which describe the affinity of the substrates ($sgRNA$, Cg) for their ligands (C, P_{Free}) and have the following relations: $K_1 = k_1^-/k_1^+$ and $K_{CG} = k_{CG}^-/k_{CG}^+$.
Assuming (i) sgRNA overabundance (Hypothesis 3: $sgRNA >> K_1$), (ii) Cg overabundance (Hypothesis 4: $Cg >> K_{CG}$) and describing the parameters as reported in Equations (3.7) - (3.10), is possible to

simplify the model as reported in Equations (3.11) - (3.14).

$$\textbf{Transcription: } \alpha = \alpha_C = \frac{\alpha_G}{f} = \frac{\alpha_{sgRNA}}{s} \tag{3.7}$$

$$\textbf{Translation: } \rho = \rho_C = \frac{\rho_G}{b} \tag{3.8}$$

$$\textbf{Protein Degradation: } \mu >> d_C = d_G \tag{3.9}$$

$$\textbf{RNA Degradation: } \mu << d_{RNA} = d_{sgRNA} = d_{mC} = d_{mG} \tag{3.10}$$

$$\frac{dC}{dt} = \frac{n \cdot \rho \cdot \alpha}{d_{RNA}} - \mu \cdot C \tag{3.11}$$

$$\frac{dG}{dt} = \frac{n \cdot b \cdot \rho \cdot f \cdot \alpha \cdot K_{CG}}{d_{RNA} \cdot Cg} - \mu \cdot G \tag{3.12}$$

$$\frac{dsgRNA}{dt} = n \cdot s \cdot \alpha - d_{RNA} \cdot sgRNA \tag{3.13}$$

$$Cg = C \tag{3.14}$$

Equations (3.11) - (3.14) represent the ordinary differential equation (ode) system used to evaluate the time-continuous behaviour of Sad-iFFL circuit and compared with the open-loop (Section 3.3.4) and U-iFFL (Section 4.2.3) schemes.
The steady-state representation of the model was obtained by evaluating the Equations (3.11) - (3.14) at their equilibrium, $dX/dt = 0$, where $X = [C, G, sgRNA]$. The resulting system (Equations (3.15) - (3.18)) has been used in Section 3.3.4 to evaluate *in silico* the robustness of the Sad-iFFL circuit against parameter variations and to simulate the biological noise throughout the circuit in a stochastic con-

text.

$$C^{SS} = \frac{n \cdot \rho \cdot \alpha}{d_{RNA} \cdot \mu} \tag{3.15}$$

$$G^{SS} = \frac{n \cdot b \cdot \rho \cdot f \cdot \alpha \cdot K_{CG}}{d_{RNA} \cdot \mu \cdot Cg} \tag{3.16}$$

$$sgRNA^{SS} = \frac{n \cdot s \cdot \alpha}{d_{RNA}} \tag{3.17}$$

$$Cg^{SS} = C^{SS} \tag{3.18}$$

Finally, substituting Cg expression(Equation (3.15)) in Equation (3.16) and simplifying the common parameters, the final G protein level is described as follows:

$$G^{SS} = b \cdot f \cdot K_{CG} \tag{3.19}$$

In Equation (3.19), the G protein level at steady-state is equal to the Michaelis−Menten equilibrium dissociation constant of SadCas9 (K_{CG}) multiplied by two proportional-scale factor describing the change of RBS (b) and promoter (f) activities between C and G. In other words, if the transcription rate of the two identical promoters in the circuit was the same and if the translation rate provided by the two identical RBSs upstream of the two genes was the same, the steady state protein level would be K_{CG} (from Equations (3.7) and (3.8) it is possibile to set $b = 1$ and $f = 1$); on the other hand, Equation (3.19) describes a more general situation in which identical components have different activities. This is especially true for RBSs, which are intrinsically highly context-dependent parts and the translation efficiency can significantly change when the same RBS is placed upstream of different coding sequences [101]. In particular, the folding of the transcript RNA can affect the accessibility of RBSs to ribosomes. For these reasons, a biophysical model of RBS efficiency has been studied, including

different contributions that can elucidate if the designed circuits are theoretically capable of rejecting variations of RBS efficiency in different hosts. This topic is discussed in more detail in Section (3.3.3). Promoters, instead, are less prone to change their activities [102, 103], despite significant variations have been reported in the literature [104]. Finally, the equlibrium constant K_{CG} describes the affinity of the repressor protein with its target DNA and its modulation can change the reached output level.

3.3 Sad-iFFL results

In this Section, the results about the theoretical investigation of Sad-iFFL performance (3.3.1-3.3.4) and the *in vivo* characterization results (3.3.5-3.3.6) are provided.

3.3.1 Leakage analysis

In the Sad-iFFL model derivation, the assumption that the dCas9 repressor could stop completely the transcription activity of the target promoter has been made. The scenario in which this assumption is violated is herein considered and quantified the error affecting the steady-state level of G is formally quantified. The G expression is hereby analyzed since it is the only one affected by the dCas9 regulation. The dynamic ordinary equation (Equation (3.4)) describing G expression over time has been changed as follows:

$$\frac{dG}{dt} = \frac{n \cdot \rho_G \cdot \alpha_G}{(d_{RNA} + \mu)} \cdot \left(\delta + \frac{1 - \delta}{(1 + \frac{Cg}{K_{CG}})} \right) - (d_G + \mu) \cdot G \qquad (3.20)$$

In Equation (3.20), δ is the *leakage* parameter, representing the percentage of G synthesis rate in the fully repressed state, occurring due

to non-perfect dCas9 repression and $(1-\delta)$ is the maximum percentage of protein synthesis rate that dCas9 could repress. Assuming the overabundance hypotheses (3)−(4), Equations (3.7) − (3.10) and substituting the Cg expression in the steady-state equation, it follows that:

$$G^{SS} = \underbrace{b \cdot f \cdot K_{CG}}_{\substack{full-repressed \\ steady-state}} + \underbrace{b \cdot f \cdot \left(\frac{\overbrace{n \cdot \rho \cdot \alpha}^{[Cg]^{SS}}}{d_{RNA} \cdot \mu} - K_{CG} \right) \cdot \delta}_{leakage-dependent} \tag{3.21}$$

In Equation (3.21) the protein level G at equilibrium depends on two effects: the steady-state level of G in the full-repressed state (Equations (3.19)) and a leakage-dependent contribution which increase the total protein level G by an additive factor proportional to δ. The assumption for which G is equal to the predicted value in Equation (3.19) is reported in the next Equation (3.22), derived by modeling C (and Cg) as proportional to its half-maximum constant by a multiplicative non-dimensional factor z, representing a fold-increase compared with K_{CG} ($Cg = z \cdot K_{CG}$).

$$\frac{1}{(z-1)} >> \delta \tag{3.22}$$

In Equation (3.22) it is shown that, for the contribution of leakage to be negligible, it is necessary that the inverse of the increase in the protein repressor Cg compared to K_{CG}, has to be much greater compared to the leakage parameter δ, expressed as a percentage of the maximum transcription.

3.3.2 Model constraints

The hypotheses that have been stated in the mathematical model derivation of Sad-iFFL have been collected below and, subsequently,

discussed:

- **Hypothesis 1: overabundance of *Cg* compared to its target.** Since the coding sequences of C and G are assumed to be present at the same copy number, the concentration (n), the production rate of Cg (Equations (3.15) and (3.18)) repressor complex has to be greater than its degradation rate.

$$\rho \cdot \alpha >> d_{RNA} \cdot \mu \tag{3.23}$$

- **Hypothesis 2: overabundance of *sgRNA* compared to free dCas9 (C).** The production rate of $sgRNA$ (Equation (3.17)) has to be greater than the diluted production rate of C (Equation (3.15)). This hypothesis could be violated if a too weak promoter has been chosen for $sgRNA$ expression.

$$s >> \frac{\rho}{\mu} \tag{3.24}$$

- **Hypothesis 3: overabundance of *sgRNA* compared to its half-maximum constant (K_1).** Raper et al. have shown that the K_1 parameter has a value of ≈ 10 $[pM]$ inferring that the linkage between C and $sgRNA$ (Equation (3.17)) is fast enough to be considered instantaneous [105].

$$\frac{n \cdot s \cdot \alpha}{d_{RNA}} >> K_1 \tag{3.25}$$

- **Hypothesis 4: overabundance of *Cg* compared to its half-maximum constant (K_{CG}).** Lower concentration limit of Cg (Equations (3.15) and (3.18)) to achieve transcriptional repression on G.

$$\frac{n \cdot \rho \cdot \alpha}{d_{RNA} \cdot \mu} >> K_{CG} \tag{3.26}$$

- **Hypothesis 5: *Cg* upper limit due to its repression inefficiency.** This represents a notable assumption that must be considered in the design phase of the circuit when the repressor protein is chosen, in fact, the predictable steady-state protein level can be reached and mantained stably if the intracellular repressor concentration is high enough to guarantee its overabundance (Hypothesis 4: $Cg >> K_{CG}$) and lower to a factor dependent on its repression activity and the percentage of leakage due to its repression inefficiency. Equation (3.22) can be written as follows:

$$K_{CG} \cdot \left(\frac{\delta + 1}{\delta}\right) >> Cg \tag{3.27}$$

3.3.3 RBS strength model

The translation efficiency of a coding sequence is strongly dependent on translation initiation rate (TIR), which is the limiting step of protein synthesis. In this section, a mathematical model of TIR is studied to investigate the theoretical working constraints of the designed circuit, which includes two identical RBSs upstream of different coding sequences (SadCas9 and GOI). The portability of a given circuit across different bacterial species will be discussed to elucidate the theoretical host-dependent protein level variation. The total amount of free energy that can be converted into nonmechanical work can be described by Gibbs free energy via the model represented in Equation (3.28). Gibbs free energy model describes the amount of energy present in the mRNA system before and after association of the 30Ss ribosome subunit with ribosome binding site on the mRNA:

$$TIR \propto e^{-\beta \cdot \Delta G_{total}} \tag{3.28}$$

where: β is the Boltzmann constant and ΔG_{total} is the total variation of Gibbs free energy. The total variation of free energy can be written as sum of different energy contributions:

$$\Delta G_{total} = \Delta G_{final} - \Delta G_{initial} =$$

$$(\Delta G_{mRNA:rRNA} + \Delta G_{start} + \Delta G_{spacing} - \Delta G_{standby}) - \Delta G_{mRNA} \tag{3.29}$$

where:

- $\Delta G_{mRNA:rRNA}$ is the energy released when the last nine nucleotides of the *E. coli* 16S rRNA hybridizes and co-folds with the mRNA sub-sequence.
- ΔG_{start} is the energy released when the start codon hybridizes to the initiating tRNA anticodon loop.
- $\Delta G_{spacing}$ is the free energy penalty caused by a nonoptimal physical distance between the 16S rRNA binding site and the start codon.
- $\Delta G_{standby}$ is the work required to unfold any secondary structures sequestering the standby site ($\Delta G_{standby} < 0$), upstream of the RBS, after the 30S complex assembly.
- ΔG_{mRNA} is the work required to unfold the mRNA sub-sequence when it folds according to its most stable secondary structure, called the minimum free energy structure.

The main goal of this part was to study the impact of the ribosome binding site on the efficiency of a genetic circuit in different host organisms, to study the effects of transferring a given iFFL architecture from an organism to another without changing the regulatory

sequences. In specific, since the balance between GOI and repressor level is the key interaction in an iFFL circuit and (as it will be illustrated in the Results section) it affects the steady-state protein level, it is important to investigate if the balance between the TIR of GOI and SadCas9 repressor is maintained constant when the same circuit is transferred in different host strains, for which the translation machinery changes. By design, the RBS sequence is assumed to be identical in both genes, except for the target binding site of Cg which is present only on the GOI gene. For any two genes, this balance in every strain can be expressed as a ratio of their TIR values:

$$\frac{TIR_1}{TIR_2} = \frac{e^{-\beta \cdot \Delta G_{total_1}}}{e^{-\beta \cdot \Delta G_{total_2}}} \tag{3.30}$$

In the case of interest, the two genes are SadCas9 and GOI, and this ratio is the same as the b parameter defined in Equation (3.8). Using the Gibbs free energy model, we can re-write equation (3.30) as:

$$\begin{aligned}\frac{TIR_1}{TIR_2} =& \frac{e^{-\beta \cdot \Delta G_{mRNA:rRNA_1}} \cdot e^{-\beta \cdot \Delta G_{start_1}} \cdot e^{-\beta \cdot \Delta G_{spacing_1}}}{e^{-\beta \cdot \Delta G_{mRNA:rRNA_2}} \cdot e^{-\beta \cdot \Delta G_{start_2}} \cdot e^{-\beta \cdot \Delta G_{spacing_2}}} \cdot \\ & \frac{e^{+\beta \cdot \Delta G_{standby_1}} \cdot e^{+\beta \cdot \Delta G_{mRNA_1}}}{e^{+\beta \cdot \Delta G_{standby_2}} \cdot e^{+\beta \cdot \Delta G_{mRNA_2}}}\end{aligned} \tag{3.31}$$

This model can be simplified as:

$$\frac{TIR_1}{TIR_2} = \frac{e^{-\beta \cdot \Delta G_{mRNA:rRNA_1}}}{e^{-\beta \cdot \Delta G_{mRNA:rRNA_2}}} \cdot P \cdot S \tag{3.32}$$

- ΔG_{start} is always constant in a host species if the start codons are assumed to be the same for the two genes, by design. This leads to complete simplification of the term.
- ΔG_{mRNA} depends on the sequence, which is maintained the same. As figure of merit is made of the difference of energy when we change host the term is simplified.

- P is a constant that remains after simplification of $\Delta G_{spacing}$ terms as theoretically it depends only on the organism optimal spacing and thus the difference remains constant when we change host organism [101, 106, 107, 108].

- S is a constant that remains after simplification of the $\Delta G_{standby}$ terms. Due to the presence of the target binding region in the 5'-UTR of the GOI transcript, SadCas9 and GOI have different standby sites. However, assuming that the mRNA binging region with ribosomes is relatively constant among different organisms for a given gene, the sequence of standby site also remains constant.

- $\Delta G_{mRNA:rRNA}$ significantly change depending on the affinity with rRNA. However, this change is expected to occur equally for the two genes, thereby enabling to treat the ratio between these terms as a constant

Under the assumptions above, a change in the host strain is expected to affect the TIR of each of the two genes. However, their ratio is theoretically constant and equal to the b scale constant, thereby making the theory of Sad-iFFL valid when the circuit is moved through different bacterial species.

3.3.4 *In silico* comparison with open-loop control scheme

Steady-state analysis

Steady state analysis shows how a given system performs at its steady state as a function of perturbation factors like transcription, translation and copy number variation. This type of test was done for

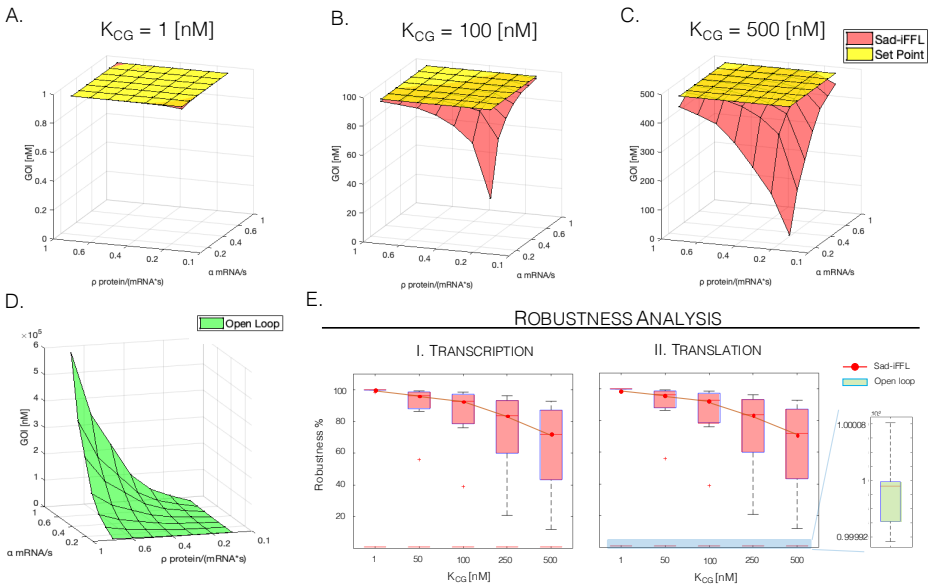

Figure 3.3: ***In silico*** **steady-state and robusteness analysis of the Sad-iFFL network compared with the open-loop scheme on different transcription and translation rates and different theoretical set points (K_{CG}).** The predicted GOI protein concentration due to transcription and translation rate variation in the Sad-iFFL circuit is shown for different K_{CG} (set point) values: 1 [nM] (**A**), 100 [nM] (**B**) and 500 [nM] (**C**). **D**, The predicted GOI protein concentration due to transcription and translation rate variation in the open-loop scheme. **E**, Comparison of the robustness performances of the circuits in terms of rejection of transcription and translation rate variations.

the Sad-iFFL circuit (Fig. 3.3A-C) and compared with the open loop scheme (the latter reported in Fig. 3.3D) as control of the unregulated gene expression. The Equation system (3.15)-(3.18), solved at the steady state, was used for this analysis.Each circuit was investigated for a range of transcription rates (α $[0.05-1]mRNA/s$), translation rates (ρ $[0.05-1]protein/(mRNA*s)$) and different desired GOI concentrations (equal to K_{CG}) $[1-500]nM$.

The results in Fig. 3.3A, Fig. 3.3B, Fig. 3.3C show that Sad-iFFL

has large limitations for specific ranges of parameters, as at low transcription and low translation rate the system is not able to converge to its set point. This is happening because a key working hypothesis of Sad-iFFL (SadCas9 >> K_{CG}) is satisfied only for sufficiently high transcription and translation rates values. Particularly, the Sad-iFFL expression level is dependent on the microenvironment where the gene cassette is placed and thus it is highly dependent on the transcription and translation rates of the host.

Due to the fact that during this test a large range of parameters was set (through variation of desired GOI set point, translation and transcription), clearly it is impossible to meet circuit requirements in every simulation and, for this reason, in Fig. 3.3A, Fig. 3.3B, Fig. 3.3C Sad-iFFL (represented in red) are unable to converge to the set point (shown as a yellow surface) at the lowest values of transcription and/or translation rates, especially for high K_{CG} values, as expected from analytical formulas above. In order to understand how the controller reject the transcription and translation rate variation, the robustness analysis of the three model reported in Fig. 3.3E with different values of the theoretical set point K_{CG}. By fixing a set point value, the robustness to a parameter variation (e.g. transcription rate) is calculated as the median of the percentages fold-change distribution computed by varying the other parameter (e.g., translation rate). The robustness index synthetized the circuit capacity to mantain the curve on a strict plane but does not include how the curve is misplaced compared to the set-point plane (yellow surface); the latter infomation is easily qualitatively estimated from the previous described graphs. In Fig. 3.3E the circuit robustness on transcriptional and translational rate variation is negatively correlated with the set point-value, indeed, an increase of the latter is reflected in a tightening of the repressor overabundance hypothesis; thus SadCas9 has to be more expressed, in terms of production parameter (α and ρ) to work as a gene repressor. In summary, these analyses showed that Sad-iFFL has strong rejection

capabilities against transcription and translation, but its performance dramatically decreased when model assumptions are not met.

Propagation of biological noise

The simulation of propagation of uncertainty is an important step for understanding the behavior of new designed synthetic circuit because it shows how a population of cells bearing the circuits behave in terms of heterogeneity, and how noise in gene expression diffuses throughout the whole system. The source of this noise in biological experiments derives from the stochasticity of cellular processes and heterogeneity of molecule abundance within a population of genetically identical cells [109]. In the analysis of stochastic processes, it is often beneficial to separate contributions arising from fluctuations that are inherent in the reactions occurring in the system of interest (intrinsic noise) from those arising from variability in factors that are considered to be external or specific of the whole cell (extrinsic noise). In the phenomenological model of gene expression, intrinsic noise is defined by the fluctuations generated by stochastic promoter activation, promoter deactivation, and mRNA and protein production and decay. Extrinsic noise sources are defined as fluctuations and population variability in the rate constants associated with these events, usually caused by heterogeneity of cellular resources like polymerases or ribosomes [110].

Moreover, because of differences in gene copy number at different points in the cell cycle, transcription rates inevitably change as cells grow and divide [111]. Propagation of uncertainty simulation was carried out to check how Sad-iFFL perform in the presence of random biological noise belonging to intrinsic and extrinsic components, using the simulated cell-to-cell variability in a recombinant population as output. Open-loop scheme was used as a control as it represents a traditional protein production cassette without effective control on the

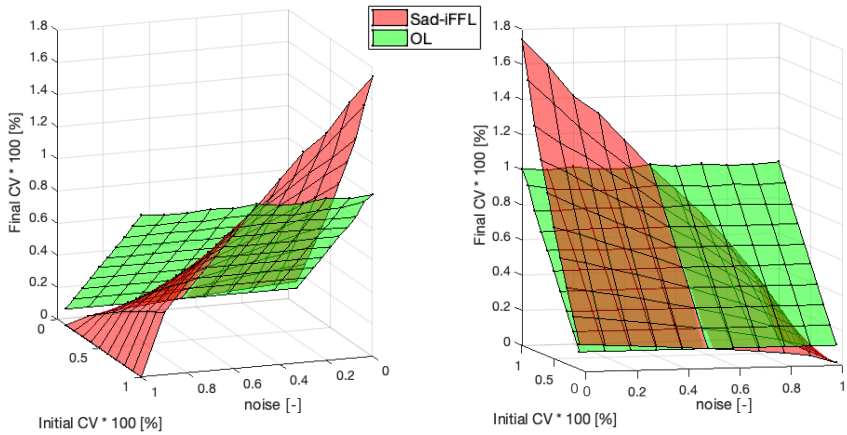

Figure 3.4: **Propagation of biological noise of Sad-iFFL and Open loop circuits.** The graph is represented from two different angular perspectives. The variation in GOI protein concentration is reported in terms of coefficient of variation (final CV) as a results of noise propagation, calculated for different transcription and translation rates; the noise entity is proportional to their coefficient of variation level (initial CV). The Pearson correlation coefficient of cellular noise (noise) is the representation of dominant noise contribution inside the system (0 - all intrinsic, 1 - all extrinsic).

gene expression processes. The noise was considered to be composed of two parts: intrinsic (in cell variation) and extrinsic noise (variation between several cells). The noise modeling and simulations methos has been treated in Appendix D.

In Fig. 3.4 the results of the propagation of uncertainty simulation are represented. Compared with the Open loop configuration, Sad-iFFL shows a lower variability when the main noise contribution is extrinsic (>0.5) and a higher variability when the extrinsic component is <0.5. Biologically plausible values for the two components indicate that the major noise contribution is extrinsic (0.6-0.8), thereby making Sad-iFFL design less noisy than Open loop, confirming the previously characterized advantages of this network motif in terms of cell-to-cell variability [112].

Dynamic analysis

The simulation of dynamic evolution of the system is an important step of synthetic circuit characterization. It describes how does the systems evolve in time and when they reach their steady state. Moreover, it is one of the ways to evaluate the behavior of all the essential parts of the system, and how they react depending on different initial conditions of expressed genes in the cell. The dynamic behavior of the circuits are analyzed and compared based on the settiling time at 95% and the memory effect of the system in response to an induction/de-induction cycle, both illustrated in Fig. 3.5.

Settling time

For this test, Equation system (3.3)-(3.6) was used. The final desired concentration or minimal theoretical limit of GOI was set equal to $500nM$ (represented as black dotted line on Fig. 3.5A). The simulation of the Sad-iFFL circuit shows a converging behavior to the theoretical set point $K_{CG} = 500nM$ (black dotted line) and a settling time of $\approx 3.6h$, which is higher than the one of the Open-loop circuit ($\approx 2.1h$), meaning that the repressor machinery needs more time to reach enough repressor concentration to guarantee its overabundance hypothesis (Hypothesis 4: $Cg >> K_{CG}$) and to reach an equilibrium of all the circuit dynamics.

Induction/de-induction cycle

The basic idea behind this test is to simulate how the Sad-iFFL system evolves in time when transcription is triggered but then it is turned off (e.g., by turning an environmental stimulus off or by washing an inducer molecule out from the growth medium), assuming that the transcriptional activity of SadCas9 and GOI is driven by an ideal inducible promoter with no basic activity in the off state. During this test, the initial conditions of simulations has set to a transcription

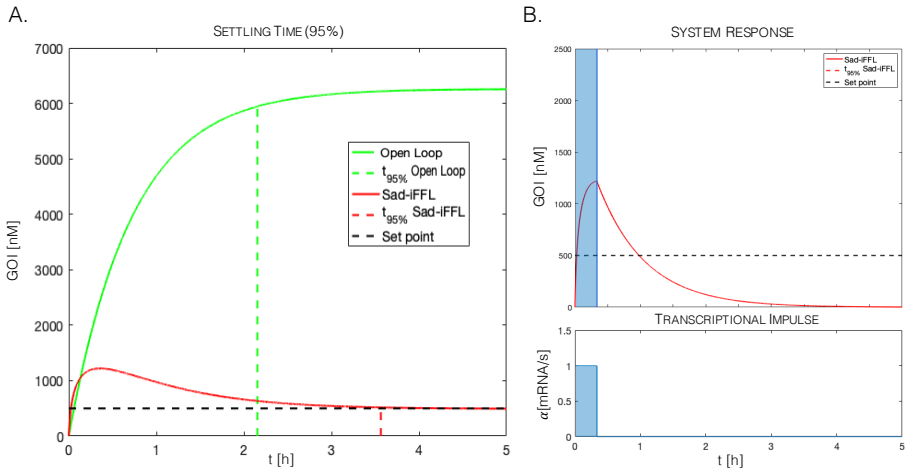

Figure 3.5: **Dynamic analysis of Sad-iFFL circuit and comparison with Open-loop scheme.**(**A**) The simulations of in-time GOI protein concentrations for the two systems are shown combined to the settling time calculated when the protein level reach and remain within an error band of 5%. The settling times are shown through colored (red - Sad-iFFL, green - Open-Loop) dotted vertical lines while the teoretical set point (K_{CG}) is represented by black dotted line. (**B**) The graph represents the system response (above) to a 20-minute impulse (below) described biologically as the transcription activation of the whole set of promoters within the Sad-iFFL circuit. The black dotted line represents the theoretical set point (K_{CG}).

rate value of 1 s^{-1} and turned off after a time $t_{OFF} = 20min$, describing a transcriptional 'pulse' in the specified time window. In the upper graph of Fig. 3.5B the response of the Sad-iFFL model (in red) is reported following a transcriptional 'pulse' highlighted by the blue vertical box. As before, the theoretical set point K_{CG} is reported with black dotted line. Sad-iFFL is not capable to maintain its steady state after turning the transcription off (it goes down to the initial state after the pulse). This behavior was expected since no memory architecture is present in the circuit, and the simulations showed the expected time for a full activation and de-activation cycle of Sad-iFFL.

3.3.5 *In vivo* and *in silico* characterization of individual modules: *S. aureus* dCas9

This section, the SadCas9 enzyme has been characterized to elucidate its suitability as a repressor in synthetic circuits such as Sad-iFFL, a feature that is not studied yet in the literature and depends on the efficiency and host resource usage parameters of the enzyme. In order to fully characterize the repressor protein, it is necessary to understand its quantitative behavior on its target and how it affects the bacterial host. The characterization process is complete when the transfer function between the input (e.g., repressor) and the output (e.g., protein to repress) is known for any relevant range of input values. The process of enzyme characterization is thus reconducible on the knowledge of the input-output relashionship in different conditions such as transcription, translation and copy number contexts. The fastest and easiest technique to estimate protein concentration is using a reporter gene coding for a fluorophore (e.g., Red Fluorescent Protein - RFP), which fluorescence, in terms of wave lenght emission, can be detected and quantified upon light excitation. This process is clearly not available for non-fluorescent protein, like SadCas9, thus its cannot be estimated directly and has to be correlated to another known signal as, in this case, the HSL inducer concentration that drives the LuxR-circuitry. The repression activity of HSL-induced SadCas9 enzyme has been monitored on RFP as target gene as reported on the characterization circuit scheme used to collect experimental (Fig. C.1). The experimental data collected from the circuits listed in Table C.5 are reported in Fig. 3.6. In all experiments, the target-specific sgRNA is expressed by the fully-induced (IPTG=200 $[\mu M]$) P_{Lac} promoter thus guaranteering its high abundance as required by circuit working assumptions. The data have been grouped based on the composition of the target expression cassette, in terms of promoter-RBS pair of the target circuit (from left to right: J119-31, J118-34 and J119-34).

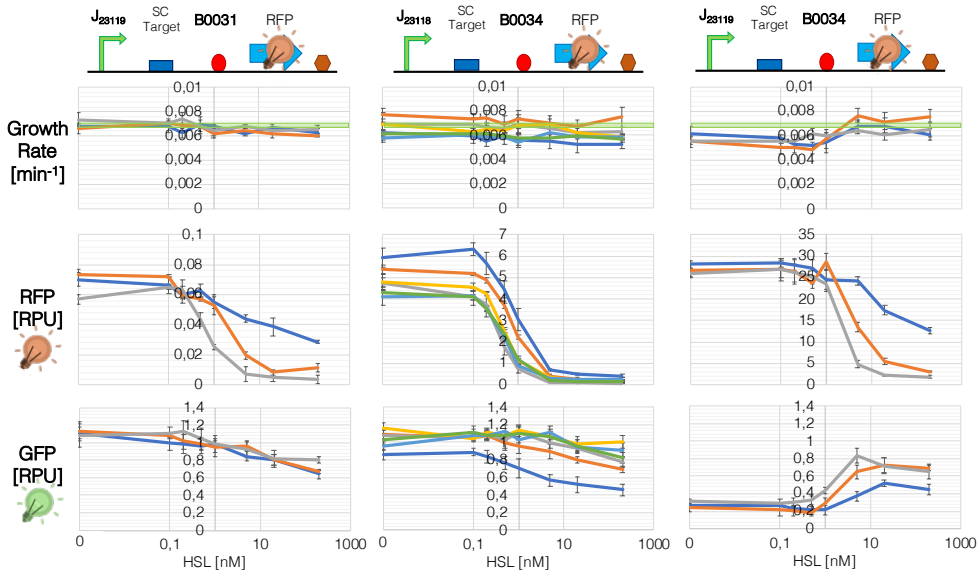

Figure 3.6: **Experimental data of steady-state transfer functions of *S. aureus* dCas9.** The three medium-copy plasmids with different promoter-RBS combinations used in this study are shown in the upper part of the graph (the IPTG-inducible sgRNA expression cassette is not shown for graphical reasons), below each one the experimental data where the error bars representing the standard error of the mean ($n >= 3$ biological replicates for each point) are reported. In all the graphs, each color refers to an RBS upstream of the SadCas9: blue - B0031, orange - B0032, grey - B0034, yellow - CU1, light blue - CU2, green - CA1. The thick green line in the first row of graphs represents the growth rate value of the strains used in this study with no plasmids, used as growth control.

For each of them the RFP and GFP level have been reported that measure the target expression and cell burden, respectively, and the cell growth rate value computed from the OD_{600} (optical density data at 600 nm) time series. The RBSs upstream of the SadCas9 coding sequence that have been used are six (blue - B0031, orange - B0032, grey - B0034, yellow - CU1, light blue - CU2, green - CA1). All of them are used with the target circuit denoted by the J118-34 promoter-RBS combination (second column) since it has a strong output signal (from 1 to 7 RPU) with low cell burden effecting the cell, differently from

the J119-34 combination (third column) in which the non-repressed RFP signal overloads the cell, as shown in the growth rate and, more clearly, in the GFP graphs. Even if the growth rates of all constructs are not highly different from the growth control (green thick line, corresponding to the growth rate of strains without plasmids) cell burden is present in all the promoter-RBS cases, in different ways. In fact, for the J119-34 target, cell burden has been shown to mainly derive from RFP overexpression and the increasing of the SadCas9 concentration is reflected to a benefit to the cell. Differently, for J118-31 and J118-34 targets, where no apparent overload due to RFP is observed, the overexpression of SadCas9 by Lux-circuitry slightly affects the cells by causing a modest load. According with these observations, for the J119-34 target, cell burden is expected to be modulated by RFP - SadCas9 together, despite only the contribution of RFP is easily detectable from experimental data since it is the major burden source for this combination. As expected, the RFP signal shows different transfer functions for different combinations of RBS upstream of SadCas9: an increasing in strength variation on SadCas9 RBS results in a left-shifts of the curve, in fact, more SadCas9 is produced more easily it reaches the critical threshold for repression, and, as a consequence for the Sad-iFFL, the overabundance hypothesis is more easily verified. Full characterization of SadCas9 protein is obtained by the estimation of the parameters describing the system from the experimental data collected. The system could be modeled at its steady state by Hill functions and burden-dependent scale factors, as described in the previous sections of this thesis. Hill equations describe activation or repression functions, and is composed of four parameters that represent different properties of a system: minimum (β_0) and maximum (β_{max}) expression rate values, molecule concentration that is necessary to reach 50% of its effect (K) and the cooperativity of the activation or repression also called Hill coefficient that is linked with the number of binding events between the molecules of interest (η). A deeper de-

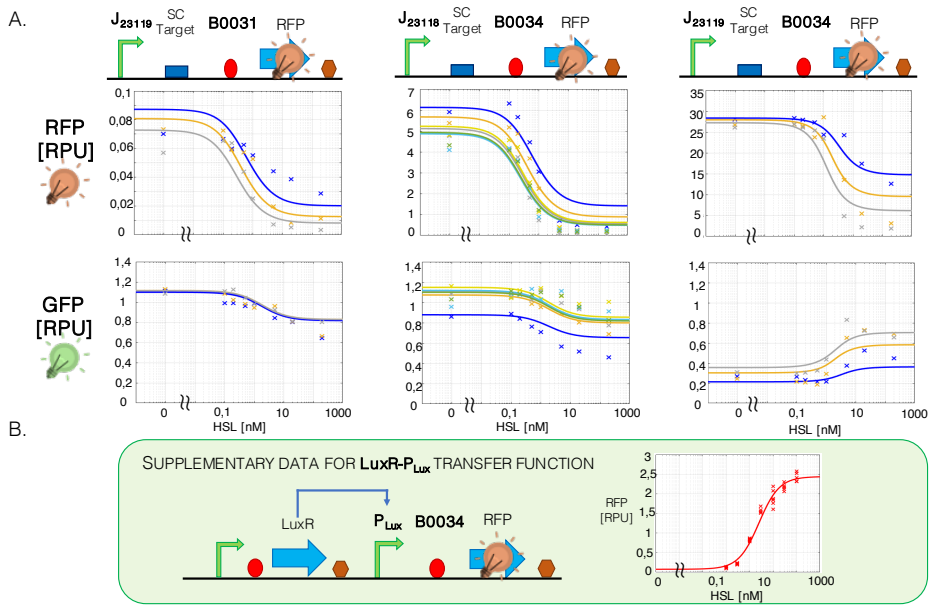

Figure 3.7: **Fitting of steady-state transfer functions to characterize *S. aureus* dCas9 with burden model.** **A**, The three medium-copy plasmids with different promoter-RBS combinations used in this study are shown in upper part of the graph (the IPTG-inducible sgRNA expression cassette is not shown for graphical reasons), below them the model fitting curves (solid lines) and the experimental data (crosses) are reported. In all the graphs, each color refers to an RBS upstream of the SadCas9: blue - B0031, orange - B0032, grey - B0034, yellow - CU1, light blue - CU2, green - CA1. **B**, The data collected from the characterization of LuxR circuitry has been used to increase the information about SadCas9 production by indirectly measuring a signal related with its protein production rate.

scription of the matematical model used for the fitting procedure has been reported and discussed in Section D.2 whereas the mathematical description of cell load has been discussed in Section 2.2.6, and has been considered including the GFP signal in fitting procedure. The fitted data have been reported in Fig. 3.7 and the parameters estimated shown in Table 3.1. The curves has been grouped in the same way as Fig. 3.6 and the cell load modeling procedure has been designed differently among the three constructs, as follows: the J119-31

and J118-34 targets are assumed to be affected only by the cell load due to SadCas9 (not by RFP), while J119-34 only from RFP (not by SadCas9). The fitting procedure has been supported by the introduction of a set of data collected from an independent recombinant strain to carry out the simultaneous estimation of the Lux-circuitry, shown at the bottom of Fig. 3.7 - green box [113]. All the parameters are estimated simoultaneusly on the available experiments and the biological literature-known parameter has been fixed as the Hill coefficients for LuxR ($\eta_{Lux} = 1$) and SadCas9 ($\eta_C = 1$) and the RFP maturation rate [114]. Cell growth rates have been set to the average measured value, depending on the construct (μ_1 for LuxR-circuitry characterization, μ_2 for SadCas9-RFP experiments Fig. 3.6). All the K_C^X parameters estimated values are in [AU]. The results obtained from the fitting showed that a joint model describing target protein expression and cell load was able to capture the main trends in the experimental data. Apart from the circuit in which B0031 was placed upstream of SadCas9, all the other characterized RBSs showed a wide range of SadCas9-dependent repression, from non-repressed values to near zero expression of the target gene. With the used RBSs, the SadCas9 enzyme expression did not show relevant cell load per se, although a high level expression of the target protein could contribute to cell burden. Taken together, these findings suggest that SadCas9 is a suitable component for synthetic circuits in recombinant *E. coli*, in terms of efficiency, tunability and load.

3.3.6 *In vivo* Sad-iFFL performances

The set of characterization circuits, reported in Table 3.2, is based on the biological scheme reported in Fig. 3.8; briefly, the fluorescence protein (RFP) and the repressor complex are expressed by the same HSL-inducible P_{Lux} promoter and the same RBS (B0032, B0034, CU1,

Table 3.1: **Estimated parameters from fitted experimental data of SadCas9-characterization.** [a]Fixed to its literature reported parameter value. [b]Computed from the experimental data collected.

Parameter	Description	Units	Estimated Values
β_0^C	Basal protein level due to leakage activity of P_{Lux} (HSL=0 [nM])	AU	0.131
β_{max}^C	Maximum protein level due to maximum activity of P_{Lux}	AU	4.06
K_{Lux}	Concentration of HSL corresponding to half-maximum induction value of P_{lux}	nM	2.52
η_{Lux}	Hill coefficient of P_{Lux} activation function	-	1^a
μ_1	Cell growth rate	min^{-1}	0.012^b
$\beta_0^{J118-34}$	Residual RFP level (expressed with J23**118** promoter and B00**34** RBS) reached due to the repression of SadCas9 (C) at maximum P_{Lux} activity	AU	0
$\beta_{max}^{J118-34}$	Maximum RFP level (expressed with J23**118** promoter and B00**34** RBS) reached due to the repression of SadCas9 (C) at P_{Lux} leakage activity (HSL=0)	AU	9.38
$\beta_0^{J119-31}$	Residual RFP level (expressed with J23**119** promoter and B00**31** RBS) reached due to the repression of SadCas9 (C) at maximum P_{Lux} activity	AU	0
$\beta_{max}^{J119-31}$	Maximum RFP level (expressed with J23**119** promoter and B00**31** RBS) reached due to the repression of SadCas9 (C) at P_{Lux} leakage activity (HSL=0)	AU	0.139
$\beta_0^{J119-34}$	Residual RFP level (expressed with J23**119** promoter and B00**34** RBS) reached due to the repression of SadCas9 (C) at maximum P_{Lux} activity	AU	0
$\beta_{max}^{J119-34}$	Maximum RFP level (expressed with J23**119** promoter and B00**34** RBS) reached due to the repression of SadCas9 (C) at P_{Lux} leakage activity (HSL=0)	AU	98.96
K_C^{B0031}	Concentration of SadCas9 (C) corresponding to half-maximum enzymatic activity in the construct with **B0031**-SadCas9 combination	AU	199.98
K_C^{B0032}	Concentration of SadCas9 (C) corresponding to half-maximum enzymatic activity in the construct with **B0032**-SadCas9 combination	AU	109.42
K_C^{B0034}	Concentration of SadCas9 (C) corresponding to half-maximum enzymatic activity in the construct with **B0034**-SadCas9 combination	AU	65.60
K_C^{CU1}	Concentration of SadCas9 (C) corresponding to half-maximum enzymatic activity in the construct with **CU1**-SadCas9 combination	AU	71.86
K_C^{CU2}	Concentration of SadCas9 (C) corresponding to half-maximum enzymatic activity in the construct with **CU2**-SadCas9 combination	AU	54.40
K_C^{CA1}	Concentration of SadCas9 (C) corresponding to half-maximum enzymatic activity in the construct with **CA1**-SadCas9 combination	AU	57.06
η_C	Hill coefficient of SadCas9 (C) repression function	-	1^a
μ_2	Cell growth rate	min^{-1}	0.006^b
θ^{RFP}	RFP maturation rate	min^{-1}	0.0167^a
J_C	Resource usage due to SadCas9 (C) expression	$min{\cdot}AU^{-1}$	0.0005
J_R	Resource usage due to RFP (R) expression	$min{\cdot}AU^{-1}$	0.021

CU2, CA1) in order to increase drive the transcription of both genes over different levels. Cell load is monitored from the GFP cassette, and the constitutive expression from the strong promoter J119 in high copy plasmid was used to produce sgRNA thereby meeting the overabundance requirement off sgRNA over free SadCas9. The full description of the biological scheme has been reported in Section C.4.1.

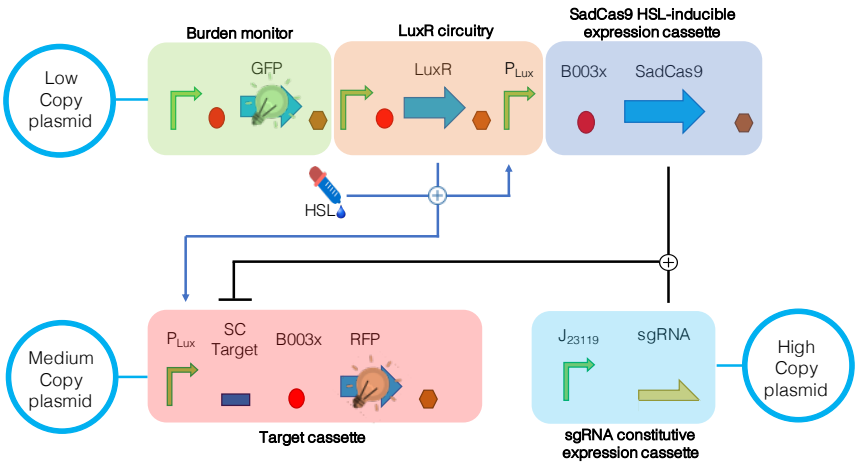

Figure 3.8: **Circuitry scheme for Sad-iFFL *in vivo* characterization.** The synthetic Sad-iFFL circuits tested in this study are based on three-plasmid system. The first plasmid (low copy plasmid) contains two modules responsible, respectively, for monitoring cell load (burden monitor) and for HSL-inducible expression of SadCas9 protein, the second (medium copy plasmid) contains the HSL-inducible expression of RFP protein and the third (high copy plasmid) is responsible for the high constitutive expression of sgRNA which guarantee the overabundance hypothesis for the SadCas9 functionality.

Table 3.2: **Table of synthetic circuits used for Sad-iFFL characterization.** Each circuit reported in the first column is composed of the three plasmids in the adjacent columns. The final construct name is composed as: 'iFFL' - Incoherent Feed-Forward Loop and the codes combination of the RBSs in the upstream regions of SadCas9 and RFP. If the two proteins share the same RBS, only one code is reported near 'iFFL', otherwise both of them are listed. The open-loop control circuits are coupled with 'OL' suffix at the end of the name. The RBS codes are highlighted with bold font within the plasmid name.

Construct name	**Low copy plasmid**	**Medium copy plasmid**	**High copy plasmid**
iFFL32	AE-3A**32**SadCas9	P_{Lux}SCTarget**32**RFP	J119sgRNA
iFFL34	AE-3A**34**SadCas9	P_{Lux}SCTarget**34**RFP	J119sgRNA
iFFLCU1	AE-3A**CU1**SadCas9	P_{Lux}SCTarget**CU1**RFP	J119sgRNA
iFFLCU2	AE-3A**CU2**SadCas9	P_{Lux}SCTarget**CU2**RFP	J119sgRNA
iFFLCA1	AE-3A**CA1**SadCas9	P_{Lux}SCTarget**CA1**RFP	J119sgRNA
iFFL32 OL	AE-3A	P_{Lux}SCTarget**32**RFP	J119sgRNA
iFFL34 OL	AE-3A	P_{Lux}SCTarget**34**RFP	J119sgRNA
iFFLCU1 OL	AE-3A	P_{Lux}SCTarget**CU1**RFP	J119sgRNA
iFFLCU2 OL	AE-3A	P_{Lux}SCTarget**CU2**RFP	J119sgRNA
iFFLCA1 OL	AE-3A	P_{Lux}SCTarget**CA1**RFP	J119sgRNA

The data obtained from the assembled circuits is reported in Fig. 3.9. The GFP data shows how the burden generated within the cell is due to SadCas9 overexpression, in fact, the GFP signal of the open loop circuits (free RFP expression from P_{Lux} promoter, with SadCas9 also expressed but not functional in absence of sgRNA) is not significantly different from the ones that emerged from the iFFL-based controller for which, the RFP expression is low enough not to affect cell metabolism; the only exception is the open loop circuit expressed by the strong strength RBS (B0034) which reaches high RFP values and the GFP value becomes lower than the other constructs. The comparison among all RFP outputs in Fig. 3.9A highlighted the functionality of the Sad-iFFL controller (the comparisons between Sad-iFFL and open loop circuits based on one RBS has been reported in Fig. 3.9D) in terms of key features: all the implemented iFFL circuits reach a steady state value that is lower than the corresponding open loop ones, and the reached value is highly stable ($<$ 2-fold change)

against transcriptional activity variations when HSL concentration is $>= 1nM$. On the other hand, in the same HSL range, the corresponding open loop circuits showed larger fold-changes. For lower HSL concentrations, SadCas9 cannot reach a critical threshold to satisfy model assumptions, also consistent with the repression curves in Fig. 3.8. In disagreement with model prediction, the set point values differs as the RBS sequence change. This probably happens because the set point value of Sad-iFFL depends on K_{CG} but also on the scale parameters for transcription (f) and translation (b) (discussed in Section 3.2.2). Assuming f=1, i.e., transcription is the same between the two copies of the P_{Lux} promoter, the b parameter was investigated to understand if the variability among all the Sad-iFFL circuits curves could be explained by this parameter. For this reason, b has been estimated from the experimental data as follows. Starting from the definition of b formalized in Section 3.2.2, an estimation of the TIRs of both SadCas9 and RFP genes is required:

$$b = \frac{TIR_{RFP}}{TIR_{SadCas9}} \tag{3.33}$$

An estimation of TIR_{RFP} (Equation (3.33)) can be obtained from the open loop data (Fig. 3.9, right) considering that all these circuits are identical except for the RBS upstream of RFP, and thereby assuming that RFP level (RFP_{OL}) is linearly proportional to ρ, as:

$$TIR_{RFP} = Q \cdot \rho_{RFP} \propto RFP_{OL} \tag{3.34}$$

where Q is a constant value relating the TIR with the translation rate. On the other hand, an estimation of $TIR_{SadCas9}$ value can be computed from the $K_C^X, X = [B0032,\ B0034, CU1,\ CU2, CA1]$ values that has been obtained in SadCas9 *in vivo* characterization. In fact, in Equation (D.18) (Appendix D.2) the K parameter has been modeled as $K_C^X = \widehat{K}/\tau_{RBS},\ X = [B0032, B0034, CU1, CU2, CA1]$, where τ_{RBS}

is proportional to SadCas9 translation rate and, therefore, $1/K_C^X$ can be used to approximate $TIR_{SadCas9}$ for every RBS used. Equation (3.34) can be written as follows:

$$b = \frac{TIR_{RFP}}{TIR_{SadCas9}} \propto Q \cdot \frac{\rho_{RFP}}{\frac{1}{K_C^X}} = RFP_{OL} \cdot K_C^X,\ X = RBS \qquad (3.35)$$

The scatter plot between the b estimated parameter and the RFP output level of Sad-iFFL controller is shown (Fig. 3.9C): the variation of Sad-iFFL output values is explained by the b scale parameter with a high correlation. This demonstrates that the steady state value of Sad-iFFL (in the working region in which SadCas9 satisfies model assumptions) is predictable from the knowledge of b. Even if the constructs investigated in this work have different RBSs and the b parameter changes, this scale parameter is expected to be constant for a given RBS used across different strains, thereby enabling the engineering of predictable expression in virtually any strain in which SadCas9 is expressed at a sufficient level, higher than K_{CG}.

3.4 Sad-iFFL overall conclusion

In this chapter, a new iFFL-based architecture has been developed in order to increase the iFFL portability through different bacterial hosts. First, the mathematical model derived from the biological scheme has been analyzed to understand the working constraints and to simulate the steady-state and dynamic characteristics, which were compared with an Open loop model. The steady-state analysis shows that the robustness performances of Sad-iFFL are strongly dependent on the theoretical set-point value K_{CG}, in fact, an increase of the latter leads to raise the SadCas9 biological demand to work as a repressor. Stochastic simulations also showed that, for acceptable biological values of intrinsic and extrinsic noise proportion within a cell system, the

iFFL architecture attenuates the cell-to-cell variability in a better way than Open loop scheme. The functioning of the Sad-iFFL circuit is based on a promising CRISPR-family dCas9 protein which has been *in vivo* and *in silico* characterized in this study and finally tested *in vivo* within the Sad-iFFL circuit. Although the iFFL-regulated output has a good capability to reject the transcription rate when the SadCas9 concentration is high enough to work as repressor, the set-point value of the curves differ as the identical RBS sequence upstream of SadCas9 and GOI changes. This deviation from the expected behavior of a robust controller was further investigated with the estimation of the scale factor parameter between the translation initiation rates of the two genes of the circuits, showing that the variation among all the curves can be predictable from the knowledge of this factor. Despite robust and predictable features were demonstrated *in silico* and *in vivo* for Sad-iFFL, the optimal and promising results obtained needs more refinements to overcome the limits due to hypothesis violation, enhanced by the fact that the set of regulatory parts in new hosts is often poor or even not known. Nonetheless, the advantages provided by the newly characterized dCas9 are expected to be widely beneficial to engineer highly specific regulations in new host strains, for which a validation will be carried out in future works.

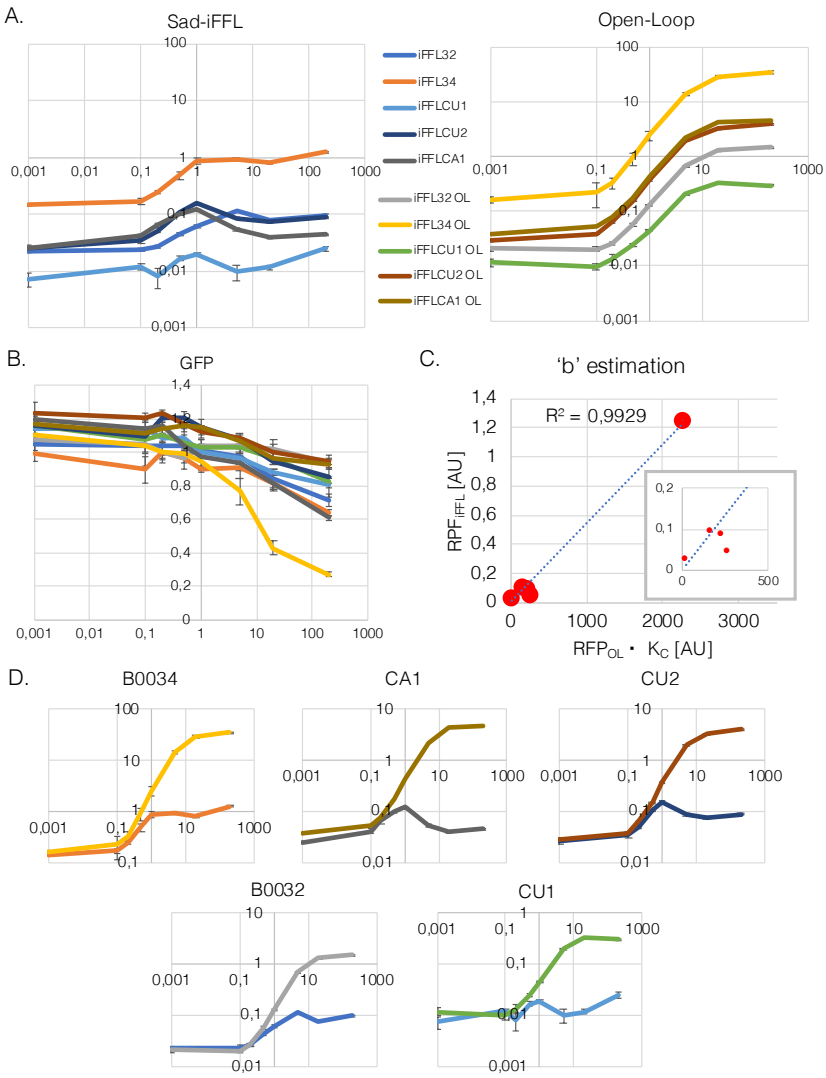

Figure 3.9: **Sad-iFFL *in vivo* performances.**(**A**) Analyzed RFP data for Sad-iFFL and Open loop circuits. Except for (C) all the graphs have HSL [nM] in x-axis and the fluorescence in y-axis has been reported in RPU units. (**B**) Analyzed GFP signal for all the circuits tested. (**C**) Correlation profile of RFP data of Sad-iFFL compared to an estimation of b in the open-loop context in order to verify the predictability of the circuit output as a function of b parameter ($TIR_{RFP}/TIR_{SadCas9}$). (**D**) Each output signal of Sad-iFFL circuit is compared to its relative open loop circuit, paired with the same RBS sequence.

Chapter 4

Universal-iFFL (U-iFFL): Theoretical analysis of an alternative design to improve iFFL network portability through different bacteria

In the previous chapter, a new circuit controller has been described and its *in silico* and *in vivo* performances shown. Although the circuit is able to reject the transcription rate variation and its RBS-dependent effects were predictable, its operability is restricted by several assumptions. In particular, the overabundance assumption of dCas9, compared with the dissociation equilibrium constant with its target DNA, is difficult to achieve when the regulatory parts that drive the repressor are not from a library of well characterized components, that is a typical situation when dealing with a non-model host strain. In this chapter, a new synthetic circuitry (named U-iFFL) has been developed to overcome the above Sad-iFFL limitations, thereby increasing the circuit portability through different bacterial species. The biological schema has been reported in Section 4.1 and subsequently the mathematical model has been studied to analyze, as done previously

in Chapter 3 for Sad-iFFL network, its operability and limitations due to model assumptions. The *in silico* performance metrics of U-iFFL have been compared with the simulations shown and discussed in the previous chapter for Sad-iFFL and open-loop schemes to provide a full comparison among the three configurations. Differently, the *in vivo* experiments has not been performed due to laboratory restriction caused by the Sars-CoV-2 worldwide pandemic. Although it has not been possible to characterize *in vivo* the U-iFFL circuit, its main components, namely the RNAPT7-P_{T7} transcription system, have been studied and tested *in vivo* to evaluate its suitability for being embedded in the final circuit as trancriptional activator. The results of $RNAPT7 - P_{T7}$ system characterization are shown and elucidated in Section 4.2.4, while an overview discussion has been reported in the conclusive Section 4.3.

4.1 U-iFFL model-based design

To overcome the iFFL limitations that has been discussed in Section 3.4, a new iFFL-based architecture has been developed with the aim to achieve higher robustness, stability and to increase the portability of the circuit throughout different bacterial host (Fig. 4.1A). The new design relies on three sub-networks composed by an amplification module (Fig. 4.1B) and two internal incoherent feedforward loop (Fig. 4.1C) and (Fig. 4.1D). The amplification module has been inserted in order to increase the circuit robustness due to the enhancement of the repressor concentrarion, thereby guaranteering its overabundance hypothesis (Hypothesis 4: $Cg >> K_{CG}$, Section 3.3.4); in fact, this assumption has been proven *in silico* (Section 3.3.4) and, subsequently, *in vivo* as well (Section 3.3.6) to be determinant for the Sad-iFFL controller functionality. Since the positive feedback loop on its main components, namely the RNAPT7-P_{T7} transcription system,

couldhave lead to an uncontrollable overproduction of activator and thus provide metabolic overload or unstable behavior, a second iFFL architecture has been embedded, compared with Sad-iFFL, to control the activator concentration and mantain its level in a physiological range for the cell host. This second iFFL motif can simultaneously guarantee the hypothesis of repressor overabundance, under different less restrictive constraints, described in the next Section. As shown in Fig. 4.1, the original iFFL motif regulating the target gene (GOI) has been mantained compared with Sad-iFFL.

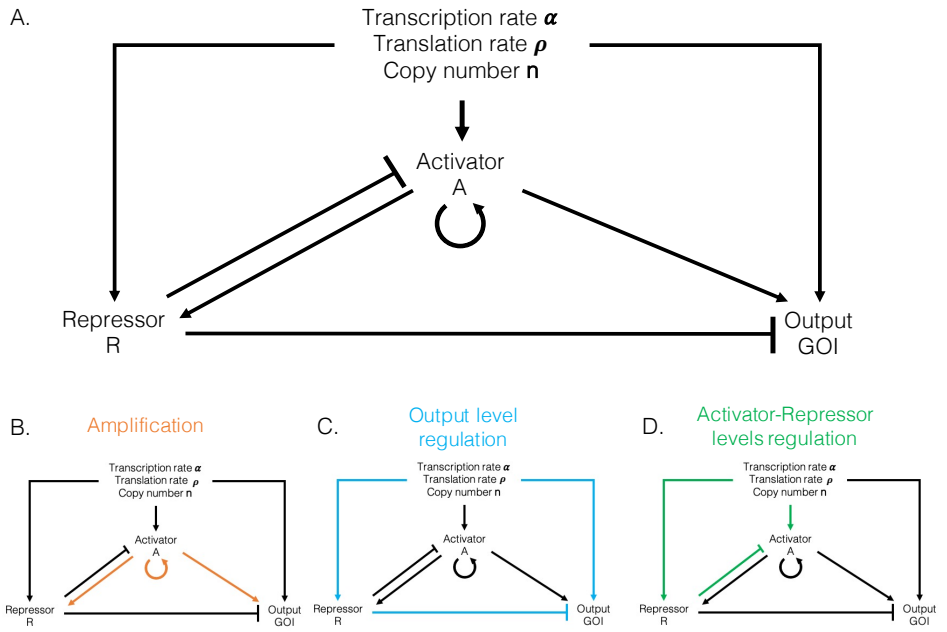

Figure 4.1: **An improved network design (U-iFFL) based on the iFFL scheme.** **A**, The design of the proposed network is based on iFFL scheme with the addition of an activator A and two sub-modules responsible to increase the network robustness and stability (B),(D). **B**, The amplification loop increases the concentration levels of all proteins in the system in order to decrease the probability of violation of actuator over-abundance hypothesis. **C**, The normal functioning of iFFL network is mantained. **D**, A new regulation process based on iFFL scheme is apported to control the amplification loop.

4.1.1 Circuit description

The biological scheme, called universal-iFFL (U-iFFL), of the architecture previously illustrated has been reported in Fig. 4.2. The genetic controller is composed of three proteins: the transcriptional repressor SadCas9 (dCas9), the protein G which is encoded by the expression of the gene of interest (GOI) and the transcriptional activator RNA Polymerase T7 ($T7$). The latter is responsible for the whole transcriptional activity of the circuit including the activator itself, with the exception of the P_C promoter which is placed upstream of the RNAPT7 coding sequence and also upstream of its $T7$ promoter, thereby obtaining a tandem $P_C - P_{T7}$ promoter region. The P_C promoter is important for the circuit activation and will be described at the end of this Section. The RNAPT7 from the T7 bacteriophage system is often referred to as an orthogonal transcriptional system for recombinant bacteria due to the fact that the promoter recognized by the enzyme, called P_{T7} promoter, is not recognized by the bacterial RNA polymerases and viceversa the bacterial promoters are not bound by the phage T7 RNA polymerase; a deeper description will treated in Section 4.2.4. The positive feedback loop on $T7$, shown in Fig. 4.2, enhances the transcription of all the genes in the circuit, decreasing the probability of violation of overabundances hypotheses for Cg repressor, Cg formation and $T7$ activator, which are needed for correct functioning of the circuit; such features are expected to result in robustness and stability enhancement on protein G level control compared with the previous Sad-iFFL design. The two iFFL subnetworks shown in Fig. 4.1C and Fig. 4.1D are regulated by the SadCas9 repressor. The two promoters in the circuit are targeted by the same repressor complex (yellow boxes in Fig. 4.2), thus only one sgRNA has to be expressed, and the affinity for the complex can be modulated by changing the DNA composition of the target region. The latter has been designed immediatly downstream of the P_{T7} promoter region.

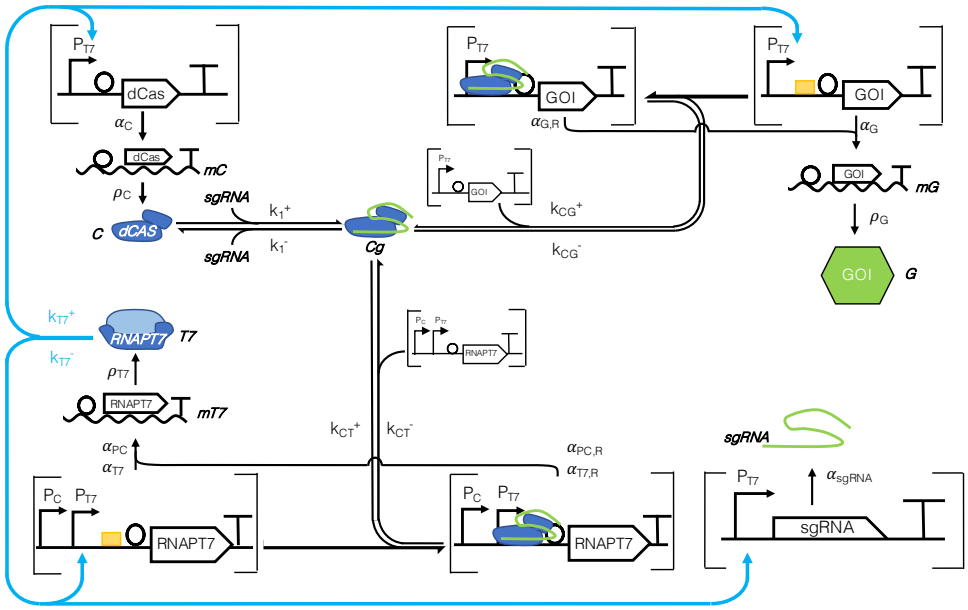

Figure 4.2: **Universal-iFFL (U-iFFL) biological model.** Visual representation of the new iFFL-based design (U-iFFL) for gene expression control system. The network functionality is based on the activities of two proteins: the dCas9 (C) repressor which, once bound with the sgRNA (sgRNA) forming a repressor complex (Cg), it, can bind the target DNA region (yellow boxes) and repress the transcription process of the downstream genes (GOI, RNAPT7); the RNAPT7 (T7) activator, once bound to its cognate promoter (P_{T7}), can start the transcription of the downstream genes (dCas9, GOI, RNAPT7). The curved arrow, the circles and the T-shaped lines are respectively the promoter (P_{T7}, P_C), RBS and terminator parts. The horizontal line inside square brackets represents the DNA while the wavy ones the mRNA. The assosiation/dissociation reactions (k_1^+,k_1^-,k_{CG}^+,k_{CG}^-,k_{CT}^+ and k_{CT}^-) are represented by bi-directional arrows while the production events with monodirectional arrows (e.g., k_{T7}^+, k_{T7}^-, transcription rate α_X, translation rate δ_X, where X is the biological entity).

The binding of the dCas9 complex to the P_{T7} promoter itself would, otherwise, result in the repression of all the copies of this promoter, thereby breaking the circuit logic. A target region is also present downstream of the promoter sequence controlling GOI expression, as in the Sad-iFFL scheme illustrated in Section 3.2.1. In summary, the iFFL loop that has been created beetwen *Cg* and *T7* allows to control

the *T7* trancriptional activation activity, while the iFFL network between C and G, analyzed in the previous chapter, is mantained. The transition from the off- to the on-state of the circuit relies on the transcriptional activity of the P_C promoter: as it will be discussed in the next section, the role of this promoter is to enable RNAPT7 expression at a rate sufficient to trigger the positive autoregulation loop. As it will be discussed in the next section, the strength of the P_C promoter must be sufficient to carry out this task, but component choice is much more flexible than for the selection of promoters in the Sad-iFFL, for which only promoter-RBS pairs providing high expression are needed. In conclusion, U-iFFL is expected to exhibit the same advantages as the Sad-iFFL, but for a wider range of promoter, RBS and copy number conditions, for which Sad-iFFL is expected to fail to show robust control.

4.1.2 Mathematical model of U-iFFL controller

The biological scheme reported in Fig. 4.2 (U-iFFL) has been modeled considering the law of mass action and the law of conservation of mass, described respectively in Equations (4.1) and (4.2).

$$\begin{cases} C + sgRNA \underset{k_1^-}{\overset{k_1^+}{\rightleftharpoons}} Cg \\ Cg + P_{G,Free}^{P_{T7}} \underset{k_{CG}^-}{\overset{k_{CG}^+}{\rightleftharpoons}} P_{G,Repressed}^{P_{T7}} \\ Cg + P_{T7,Free}^{P_{T7}} + P_{T7,Free}^{P_C} \underset{k_{CT}^-}{\overset{k_{CT}^+}{\rightleftharpoons}} P_{T7,Repressed}^{P_{T7}} + P_{T7,Repressed}^{P_C} \\ T7 + P_{Y,Free}^{P_{T7}} \underset{k_{T7}^-}{\overset{k_{T7}^+}{\rightleftharpoons}} T7 + P_{Y,Free}^{P_{T7}} + mY \end{cases} \tag{4.1}$$

$$\begin{cases} DNA : P_{X,Free} + P_{X,Repressed} = n \\ dCas : C + Cg + P_{X,Repressed} = C_{tot} \\ Single\ guide\ RNA : sgRNA + Cg = g_{tot} \end{cases} \tag{4.2}$$

In Equations (4.1) − (4.2), C, Cg and $P_{X,repressed}$ ($X = [G, T7]$) are the concentrations ($[nM]$) of dCas respectively free, coupled with the single guide RNA and bound to the target DNA site (which is also equal to the concentration of promoters in the off-state) promoter, while $sgRNA$ and $P_{Y,Free}$ ($Y = [C, G, T7, sgRNA]$) are the concentrations ($[nM]$) of free single guide RNA and non-repressed promoters inside the cell driving the transcription of C, G, $T7$ and $sgRNA$. The superscript ($T7$ or C) indicates if the promoter is an RNAPT7-regulated P_{T7} promoter or the P_C promoter. The kinetics of the repressor complex formation (Cg) and its activity towards the target DNA ($P_{X,Repressed}$) ($X = [G, T7]$) has been described with the association k_1^+ $[nM^{-1}time^{-1}]$, k_{CG}^+ $[nM^{-1}time^{-1}]$, k_{CT}^+ $[nM^{-1}time^{-1}]$ and dissociation k_1^- $[time^{-1}]$, k_{CG}^- $[time^{-1}]$, k_{CT}^- $[time^{-1}]$ rate constants (Equation (3.1)), respectively. Although the target regions upstream the two coding sequences (C, $T7$) are identical and the Cg complex can be engineered with only one single guide RNA targeting both regions, the rate constants have been defined separately for both genes, based on the subscript: CG refers to the C−G binding while CT to the C−T7 one. The transcriptional activation kinetics of RNAPT7 is determined by the association - dissociation rate constants K_{T7}^+ and K_{T7}^- (Equation (4.1)). The total amounts of target promoter, dCas and single guide RNA inside the cell are defined as $n[nM]$, $C_{tot}[nM]$ and $g_{tot}[nM]$ (Equation (4.2)). The transcription of dCas9 is carried out by the P_{T7} promoter upstream; the transcription of GOI is driven by the non-repressed P_{T7} promoter upstream; the transcription of RNAPT7 is driven by the P_C and the P_{T7} promoters when the target site is not repressed. Assuming that the concentration of target promoters

is negligible compared with the total dCas9 repressor complex (Hypothesis 1: $Cg >> P_{X,Repressed}$, $X = [G, T7]$) and that the latter is negligible as well compared with the total level of single guide RNA (Hypothesis 2: $sgRNA >> Cg$) the mathematical model of U-iFFL circuit can be written as (Equations (4.3) - (4.7)).

$$\frac{dC}{dt} = \frac{n \cdot \rho_C}{(d_{mC} + \mu)} \cdot \left(\frac{\alpha_C}{\left(1 + \frac{K_{T7}}{T7}\right)} \right) - (d_C + \mu) \cdot C \tag{4.3}$$

$$\frac{dG}{dt} = \frac{n \cdot \rho_G}{(d_{mG} + \mu)} \cdot \left(\frac{\alpha_G}{\left(1 + \frac{K_{T7}}{T7}\right)} \cdot \frac{1}{\left(1 + \frac{Cg}{K_{CG}}\right)} \right) - (d_G + \mu) \cdot G \tag{4.4}$$

$$\frac{dT7}{dt} = \frac{n \cdot \rho_{T7}}{(d_{mT7} + \mu)} \cdot \left(\left(\alpha_{T7,P_C} + \frac{\alpha_{T7,P_{T7}}}{\left(1 + \frac{K_{T7}}{T7}\right)} \right) \cdot \frac{1}{\left(1 + \frac{Cg}{K_{CG}}\right)} \right) - (d_{T7} + \mu) \cdot T7 \tag{4.5}$$

$$\frac{dsgRNA}{dt} = n \cdot \left(\frac{\alpha_{sgRNA}}{\left(1 + \frac{K_{T7}}{T7}\right)} \right) - d_{sgRNA} \cdot sgRNA \tag{4.6}$$

$$Cg = \frac{C}{(1 + \frac{K_1}{sgRNA})} \tag{4.7}$$

No cooperativity was assumed (Hill coefficient $\eta = 1$ − not shown in the model) for the complex formation from sgRNA binding [99] (Equation (3.5)) and for the transcriptional repression from the complex Cg [100] (Equation (3.4)).

No cooperativity was assumed ($\eta^{sgRNA} = \eta^{Cg} = \eta^{T7} = 1$ − not shown in the model) for the complex formation from $sgRNA$ binding [99] (Equation (4.7)), the transcriptional repression from the complex Cg (Equations (4.4) − (4.5)) [100] and the transcriptional activity of T7 polymerase (Equations (4.3) − (4.6)) [115]. In Equations (4.3)

– (4.6), α_X, ρ_X and δ_X represent the transcription, translation and degradation rate $[time^{-1}]$ of the biological entity X, respectively, while μ is the diluition rate $[time^{-1}]$ due to cell division. The transcription rate of $T7$ coding sequence has been divided in two contributions, based on its promoters P_C and P_{T7}, resulting in the two transcription-related parameters α_{T7,P_C} and $\alpha_{T7,P_{T7}}$, respectively. In Equation (4.4), (4.5), K_{CG} $[nM]$ and K_{CT} $[nM]$ are the Michaelis – Menten equilibrium dissociation constants for dCas9 with its DNA targets and have the following relations: $K_{CG} = k^-_{CG}/k^+_{CG}$ and $K_{CT} = k^-_{CT}/k^+_{CT}$; in Equation (4.7), the dissociation equilibrium constant is defined as well as $K_1 = k^-_1/k^+_1$. Analogously, the equilibrium dissociation constant for the transcriptional activation of the $T7$ polymerase (Equation (4.3) – (4.6)) has been modeled as the ratio between its dissociation - association rate constants, as: $K_{T7} = k^-_{T7}/k^+_{T7}$.

Assuming (i) $sgRNA$ overabundance (Hypothesis 3: $sgRNA >> K_1$), (ii) Cg overabundance (Hypothesis 4: $Cg >> K_{CG}$, Hypothesis 5: $Cg >> K_{CT}$), (iii) $T7$ overabundance (Hypothesis 6: $T7 >> K_{T7}$) and defining the scale parameters as reported in Equations (4.8) - (4.11), it is possible to simplify the model as in Equations (4.12) - (4.16).

$$\textbf{Transcription: } \alpha^{T7} = \alpha_C = \frac{\alpha_G}{f} = \frac{\alpha_{T7}}{g} = \frac{\alpha_{sgRNA}}{s} \tag{4.8}$$

$$\textbf{Translation: } \rho = \rho_C = \frac{\rho_G}{b} = \frac{\rho_{T7}}{c} \tag{4.9}$$

$$\textbf{Protein Degradation: } \mu >> d_C = d_G = d_{T7} \tag{4.10}$$

$$\textbf{RNA Degradation: } \mu << d_{RNA} = d_{sgRNA} = d_{mC} = d_{mG} = d_{mT7} \tag{4.11}$$

$$\frac{dC}{dt} = \frac{n \cdot \rho \cdot \alpha^{T7}}{d_{RNA}} - \mu \cdot C \tag{4.12}$$

$$\frac{dG}{dt} = \frac{n \cdot b \cdot \rho \cdot f \cdot \alpha^{T7} \cdot K_{CG}}{d_{RNA} \cdot Cg} - \mu \cdot G \tag{4.13}$$

$$\frac{dT7}{dt} = \frac{n \cdot c \cdot \rho \cdot K_{CT} \cdot \left(\alpha_{T7,P_C} + g \cdot \alpha^{T7}\right)}{d_{RNA} \cdot Cg} - \mu \cdot T7 \tag{4.14}$$

$$\frac{dsgRNA}{dt} = n \cdot s \cdot \alpha^{T7} - d_{RNA} \cdot sgRNA \tag{4.15}$$

$$Cg = C \tag{4.16}$$

Equations (4.12) - (4.16) represent the ordinary differential equation (ode) system used in Section 4.2.3 to evaluate the time-dependent behavior of U-iFFL circuit and compared with Sad-iFFL and open-loop schemes.
The steady-state representation of the model has been obtained evaluating the Equations (4.12) - (4.16) at their equilibrium, i.e. for $dX/dt = 0$, where $X = [C, G, T7, sgRNA]$. The system obtained (Equations (4.17) - (4.21)) has been used in Section 4.2.3 to evaluate the model robustness to parameters variations and to analyze the propagation of uncertainty when these parameters are affected by biological noise.

$$C^{SS} = \frac{n \cdot \rho \cdot \alpha^{T7}}{d_{RNA} \cdot \mu} \tag{4.17}$$

$$G^{SS} = \frac{n \cdot b \cdot \rho \cdot f \cdot \alpha^{T7} \cdot K_{CG}}{d_{RNA} \cdot Cg \cdot \mu} \tag{4.18}$$

$$T7^{SS} = \frac{n \cdot c \cdot \rho \cdot K_{CT} \cdot \left(\alpha_{T7,P_C} + g \cdot \alpha^{T7}\right)}{d_{RNA} \cdot Cg \cdot \mu} \tag{4.19}$$

$$sgRNA^{SS} = \frac{n \cdot s \cdot \alpha^{T7}}{d_{RNA}} \tag{4.20}$$

$$Cg^{SS} = C^{SS} \tag{4.21}$$

Finally, substituting Cg expression (Equation (4.17) and (4.21)) in Equations (4.18) − (4.19) and subsequentely to common parameters simplification procedures the final G and $T7$ protein levels can be described as:

$$G^{SS} = b \cdot f \cdot K_{CG} \tag{4.22}$$

$$T7^{SS} = \frac{c}{\alpha} \cdot K_{CT} \cdot \left(\alpha_{T7,P_C} + g \cdot \alpha^{T7}\right) \tag{4.23}$$

In Equation (4.22), the G protein level at steady-state is equal to the Michaelis−Menten constant of SadCas9 K_{CG} multiplied by two scale factors, respectively representing RBS and promoter differential efficiency, between C and G, as shown for the Sad-iFFL model in Section 3.2.2. Analogous to G, the internal iFFL (Fig. 4.1D) between Cg and $T7$, results in a protein level of the $T7$ polymerase proportional to the half-maximum constants of the complex Cg modulated by the scale factor c and by a factor proportional to the total transcription rate of the $T7$ coding sequence ($\alpha_{T7,P_C} + g \cdot \alpha^{T7}$) compared with the reference transcription rate ($\alpha = \alpha_C$). The modulation of steady-state protein levels can be achieved by the modification of the target sequences at DNA-level in order to create a differential steady-state modulator for G and $T7$ with one single guide RNA. The analytical analysis of the U-iFFL steady-state model follows the same hypotheses of Sad-iFFL (described in Section 3.2.2) for the four design scale factors: b, f, c and g. The analysis that has been reported in Section 3.3.3 for the TIR variability through different bacterial hosts are also valid for U-iFFL model.

4.2 U-iFFL Results

4.2.1 Leakage analysis

In the analytical model of U-iFFL no leakage is assumed to occur for promoters with dCas9-repressible transcription. Here, the scenario for which this assumption is restored is modeled and quantified to understand the expected error affecting the steady-state level of G and $T7$. The ode system discussed above (Equations (4.4) - (4.5)) describing the time evolution of the two dCas (C) - regulated genes (G, $T7$) has been modified as follows:

$$\frac{dG}{dt} = \frac{n \cdot \rho_G \cdot \alpha_G}{(d_{mG} + \mu)} \cdot \left(\delta + \frac{1-\delta}{\left(1 + \frac{K_{T7}}{T7}\right)} \cdot \frac{1}{\left(1 + \frac{Cg}{K_{CG}}\right)} \right) - (d_G + \mu) \cdot G \quad (4.24)$$

$$\begin{aligned} \frac{dT7}{dt} = & \frac{n \cdot \rho_{T7} \cdot \alpha_{T7,P_C}}{(d_{mT7} + \mu)} \cdot \left(\delta + \frac{1-\delta}{\left(1 + \frac{Cg}{K_{CG}}\right)} \right) + \\ & \frac{n \cdot \rho_{T7} \cdot \alpha_{T7,P_{T7}}}{(d_{mT7} + \mu)} \cdot \left(\delta + \frac{1-\delta}{\left(1 + \frac{Cg}{K_{CG}}\right)} \right) - (d_{T7} + \mu) \cdot T7 \end{aligned} \quad (4.25)$$

In Equations (4.24) − (4.25), δ is the percent leakage of transcriptional activity that can occur due to the non-zero basic transcriptional activity when dCas9 (C) is bound; the δ parameter has been assumed to be equal for both genes (G, $T7$) regulated by dCas9. Assumed the overabundances hypothesis (3)−(6), Equations (4.8) − (4.11) and by substituting the Cg expression in the steady-state equations of G and

$T7$, it follows that:

$$G^{SS} = \underbrace{b \cdot f \cdot K_{CG}}_{\substack{full-repressed \\ steady-state}} + \underbrace{b \cdot f \cdot \left(\frac{\overbrace{n \cdot \rho \cdot \alpha^{T7}}^{Cg^{SS}}}{d_{RNA} \cdot \mu} - K_{CG} \right) \cdot \delta}_{leakage-dependent} \tag{4.26}$$

$$\begin{aligned} T7^{SS} = & \underbrace{\frac{c}{\alpha^{T7}} \cdot (\alpha_{T7,P_C} + g \cdot \alpha^{T7}) \cdot K_{CT}}_{\substack{full-repressed \\ steady-state}} + \\ & \underbrace{\frac{c}{\alpha^{T7}} \cdot (\alpha_{T7,P_C} + g \cdot \alpha^{T7}) \cdot \left(\frac{\overbrace{n \cdot \rho \cdot \alpha^{T7}}^{Cg^{SS}}}{d_{RNA} \cdot \mu} - K_{CT} \right) \cdot \delta}_{leakage-dependent} \end{aligned} \tag{4.27}$$

In Equations (4.26) − (4.27) the protein levels of G and $T7$, at equilibrium, depends on two effects: the steady-state level in the full-repressed state and a leakage-dependent term which increases the total protein level by an additive factor proportional to δ. The assumptions for which G and $T7$ are equal to their predicted value that has been reported in Equations (4.22) − (4.23) are stated in the next Equations (4.28) − (4.29) derived by modeling Cg as proportional to its half-maximum constant by a multiplicative factor z ($Cg = z \cdot K_{CG}$) and z' ($Cg = z' \cdot K_{CT}$).

$$\frac{1}{(z-1)} >> \delta \tag{4.28}$$

$$\frac{1}{(z'-1)} >> \delta \tag{4.29}$$

In Equation (4.28) − (4.29) it is shown that, for the contribution of leakage to be negligible, it is necessary that the inverse of the increase in the protein repressor Cg compared to $K_{CX}, X = G, T$, has to be greater compared to leakage δ, expressed as a percentage of the maximum transcription. While the first constraint (Equation (4.28)) needs to be met, the second one (Equation (4.29)) is not necessary to achieve a working system, since the level of RNAPT7 can also become greater than its nominal value with no-leakage and hypothesis 6 is still valid.

4.2.2 Model constraints

The hypotheses that have been stated in the mathematical model of U-iFFL have been collected and, subsequently, discussed:

- **Hypothesis 1: overabundance of *Cg* compared to its target.** Since the coding sequences of C and G are assumed to be present at the same copy number, corresponding to the DNA concentration (n), the production rate of the (Equations (4.17) and (4.21)) repressor, and thus the Cg complex, has to be greater than its degradation rate.

$$\rho \cdot \alpha^{T7} >> d_{RNA} \cdot \mu \tag{4.30}$$

- **Hypothesis 2: overabundance of *sgRNA* compared to free dCas9 (*C*).** The production rate of $sgRNA$ (Equation (4.20)) has to be greater than the diluted production rate of C (Equation (4.17)). This hypothesis could be violated if a too weak promoter is chosen for $sgRNA$ expression.

$$s >> \frac{\rho}{\mu} \tag{4.31}$$

- **Hypothesis 3: overabundance of *sgRNA* compared to its half-maximum constant (K_1).** Raper et al. have shown that the K_1 parameter has a value of ≈ 10 $[pM]$ inferring that the linkage between C and $sgRNA$ (Equation (3.17)) is fast enough to be considered instantaneous [105].

$$\frac{n \cdot s \cdot \alpha^{T7}}{d_{RNA}} >> K_1 \tag{4.32}$$

- **Hypothesis 4: overabundance of *Cg* compared to its half-maximum constant (K_{CG}).** Constraint for Cg (Equations (4.17) and (4.21)) to achieve transcriptional repression on G.

$$\frac{n \cdot \rho \cdot \alpha^{T7}}{d_{RNA} \cdot \mu} >> K_{CG} \tag{4.33}$$

- **Hypothesis 5: overabundance of *Cg* compared to its half-maximum constant (K_{CT}).** Constraint for Cg (Equations (4.17) and (4.21)) to achieve transcriptional repression on $T7$.

$$\frac{n \cdot \rho \cdot \alpha^{T7}}{d_{RNA} \cdot \mu} >> K_{CT} \tag{4.34}$$

- **Hypothesis 6: overabundance of *RNAPT7* compared to its half-maximum constant (K_{T7}).** Constraint for $T7$ to achieve transcription activation of P_{T7} promoter. In Equation (4.23), dependently from the strength of the P_C promoter, it is possibile to define three different scenarios, reported below as Case A, B and C. Fixing the scale factor c and K_{CT} value,

it result that, varying the relative promoters strenght α_{T7,P_C} and $\alpha_{T7,P_{T7}}$, the worst case scenario is represented by Case A (Case C: approximating $g \approx 1, (\alpha_{T7,P_C}/\alpha^{T7}) >> 1$); indeed, by ensuring its validity it is possible to guarantee the transcriptional activation hypothesis for all the other scenarios. Case A equation described the trade-off between $T7$-activation and its repression: the relative production of $T7$ polymerase compared to Cg ($c \cdot g$) times K_{CT} has to be greater compared to its dissociation equilibrium constants. This implies that the steady-state value of $T7$ polymerase has to be controlled to be high enought to work properly as an activator but controlled to avoid cell burden due to the overexpression of C and RNAPT7 itself.

$$c \cdot K_{CT} \cdot \left(\frac{g \cdot \alpha^{T7} + \alpha_{T7,P_C}}{\alpha^{T7}} \right) >> K_{T7} \tag{4.35}$$

Case A. $g \cdot \alpha^{T7} >> \alpha_{T7,P_C}$:

$$c \cdot K_{CT} \cdot g >> K_{T7}$$

Case B. $g \cdot \alpha^{T7} = \alpha_{T7,P_C}$:

$$c \cdot K_{CT} \cdot 2 \cdot g >> K_{T7}$$

Case C. $g \cdot \alpha^{T7} << \alpha_{T7,P_C}$:

$$c \cdot K_{CT} \cdot \frac{\alpha_{T7,P_C}}{\alpha^{T7}} >> K_{T7}$$

- **Hypothesis 7: *Cg* upper bound due to its repression inefficiency.** This assumption must be considered in the design phase of the circuit when the repressor protein is chosen. In

fact, the steady-state protein level can be stably mantained if the intracellular repressor concentration is high enough to guarantee its overabundance (Hypothesis 4: $Cg >> K_{CX}, X = [G, T]$) and lower than a factor dependent on its repression activity and by the percentage of leakage due to its repression inefficiency. Equations (4.28)-(4.29) can be written as follows:

$$K_{CX} \cdot \left(\frac{\delta + 1}{\delta} \right) >> Cg, \;\; X = [C, T] \tag{4.36}$$

4.2.3 *In silico* comparisons with the Sad-iFFL and open-loop schemes

Steady-state analysis

The steady state analysis performed previously in Section 3.3.4 is now reproposed with the comparions between U-iFFL circuit scheme under the same perturbation factors: transcription, translation and copy number variations. This analysis is reported in Fig. 4.3; each of them was tested with different ranges of transcription rate (α $[0.05 - 1]$ $mRNA/s$), translation rate (ρ $[0.05 - 1]$ $protein/(mRNA * s)$) and different desired GOI concentrations (set via K_{CG}) $[1 - 500]nM$. Results on Fig. 4.3A, Fig. 4.3B, Fig. 4.3C show that U-iFFL overcomes the major limitation occuring in the Sad-iFFL circuitry: in fact, at low transcription and low translation rate the U-iFFL system is able to converge to the theoretical set point value. The amplification module (Fig. 4.1B) eventually increases the repressor protein level and its overabundance hypothesis can be met more easily than in Sad-iFFL in the same conditions. U-iFFL simulations demonstrated superior performance, as its regulatory network becomes fully active at much

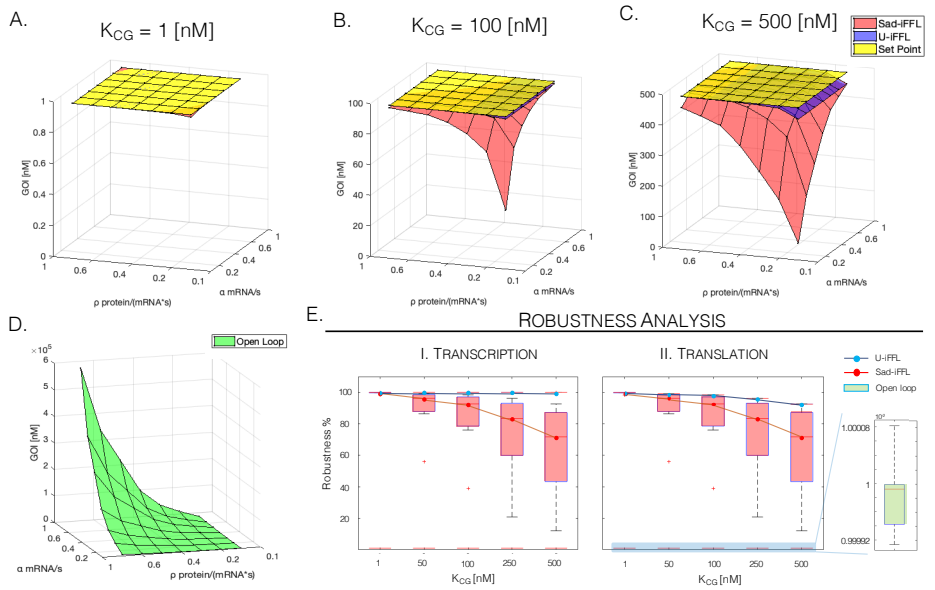

Figure 4.3: ***In silico* steady-state and robusteness analysis of U-iFFL network compared to Sad-iFFL and open-loop scheme on different transcription and translation rates and different theoretic set points (K_{CG}).** The predicted GOI protein concentration due to transcription and translation rate variation in the U-iFFL and Sad-iFFL circuits are shown for different K_{CG} (set point) values: 1 [nM] (**A**), 100 [nM] (**B**) and 500 [nM] (**C**). **D**, Predicted GOI protein concentration due to transcription and translation rate variation in the open-loop scheme. **E**, Comparison of the robustness performances of the circuits in terms of rejection of transcription and translation rate variations.

lower transcription and translation rates than Sad-iFFL, and having convergence issues only with the lowest translation rate used in the simulations.

In order to understand how the controller rejects the transcription and translation rate variations, the robustness analysis of the three circuits are reported in Fig. 4.3E for different theoretical set point tuned via K_{CG} values. Details on the robustness computation have been reported in Section 3.3.4. As shown in Fig. 4.3E, U-iFFL circuit

overperformed in terms of robustness on transcription (lower median $> \approx 98.5\%$) and translation rate (lower median $\approx 91\%$) variation compared to Sad-iFFL controller (lower median of: transcription $\approx 68\%$, translation: $\approx 68\%$). Such results quantify the enormous potential of the newly-designed synthetic circuit (U-iFFL) in comparison with the simpler iFFL composed of one repressor only (Sad-iFFL). These properties open the possibility for new possible applications of gene expression control for synthetic circuits in non-standard hosts with unknown transcription or translation rates, given a promoter and RBS.

Propagation of biological noise

Propagation of biological noise was computationally investigated to check how Sad-iFFL and U-iFFL perform in the presence of random biological noise belonging to intrinsic and extrinsic components, using the simulated cell-to-cell variability in a recombinant population analyzed in Section 3.3.4 for Sad-iFFL as well.

Results for U-iFFL are represented in Fig. 4.4. The most important information from these results is that U-iFFL error propagation is very close to the Sad-iFFL (red surface, not shown due to completely overlapping surfaces).

Such results can be interpreted as U-iFFL has the same features of noise propagation as Sad-iFFL, which underlines its stability and potential. This result was expected since, under conditions of U-iFFL functioning, biological noise affecting RNAPT7 expression does not propagate throughout the circuit (due to its abundance in Hypothesis 6) and the other variable elements (SadCas9 and GOI) are in common between the two iFFL designs. Compared with the Open loop configuration, iFFLs show a lower variability when the main noise contribution is extrinsic (>0.5) and a higher variability when the extrinsic component is <0.5. Biologically plausible values for the two components indicate that the major noise contribution is extrinsic (0.6-0.8),

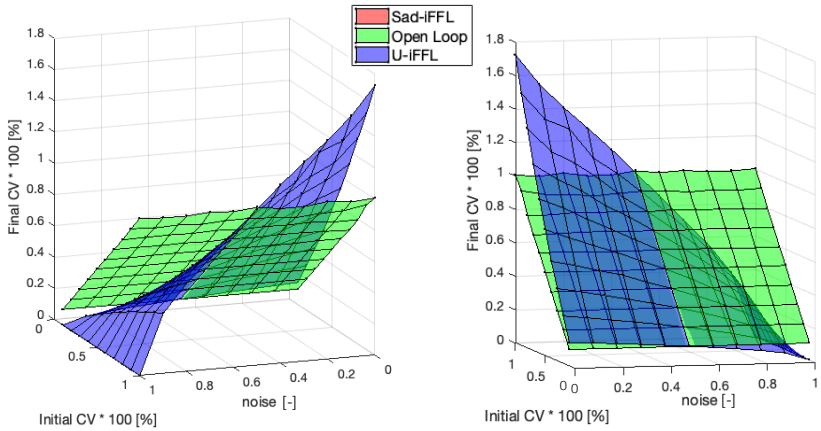

Figure 4.4: **Propagation of biological noise of U-iFFL, Sad-iFFL and Open-loop circuits.** The graph is represented from two different angular perspectives. The variation in GOI protein concentration is reported in terms of coefficient of variation (final CV) as a result of noise propagation, calculated for different transcription and translation rates; the noise entity is proportional to their coefficient of variation level (initial CV). The Pearson correlation coefficient of cellular noise (r) is the representation of dominant noise contribution inside the system (0 - all intrinsic, 1 - all extrinsic). The U-iFFL and Sad-iFFL curves are completly overlapped.

thereby making iFFL design less noisy than Open loop, confirming the previously characterized advantages of feedback and feedforward loops in terms of cell-to-cell variability.

Dynamic analysis

The dynamic behavior of the circuits are analyzed and compared based on the settiling time at 95% and the response of the system to an induction/de-induction cycle, both illustrated in Fig. 4.5; the performance indexes computation has been reported in Appendix D.

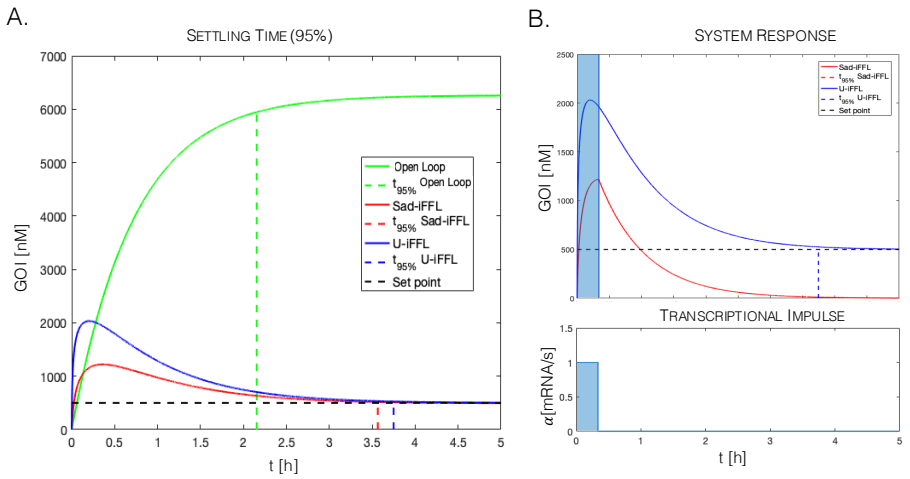

Figure 4.5: **Dynamic analysis of U-iFFL circuit and comparison with Sad-iFFL and Open-loop schemes.** (**A**) Comparison of settling time at $GOI^{SS}{}_{95\%}$ among the U-iFFL, Sad-iFFL and Open-loop circuits. The simulations of time-dependent behavior of GOI protein concentrations for the three systems are shown, together with the settling time calculated when the protein level reaches steady state and remain within a 5% error band. The settling times are shown through colored (blue - U-iFFL, red - Sad-iFFL, green - Open-loop) dotted vertical lines while the teoretical set point (K_{CG}) is represented by black dotted line. (**B**) The graph represents the two system responses (above; blue - U-iFFL, red - Sad-iFFL) to a 20-minute impulse described biologically as the transcription activation of: the P_C promoter in the U-iFFL circuit and the whole set of identical promoters within the Sad-iFFL circuit. The black dotted line represents the theoretical set point (K_{CG}) of iFFLs.

Settling time

For this test the final desired concentration or minimal theoretical limit of GOI was set equal to 500nM (represented as black dotted line on Fig. 4.5). The simulation results of the two circuits showed similar results in terms of converging speed. Even if the new designed U-iFFL simulation is characterized by a slightly higher initial response than Sad-iFFL, it is important to note that the difference in settling time is less than $\approx$ 16 minutes, which is a minor disadvantage in relation to the key advantages shown for the U-iFFL configuration. The predicted time delays of the two model circuit (Sad-iFFL, U-iFFL) are

due to the activation of the regulation machinery: in fact, the cell needs time to produce enough SadCas9 complex, and RNAPT7 in U-iFFL, to meet thir overabundances hypotheses.

Induction/de-induction cycle

During this test, the initial conditions of simulations has set to a transcription rate value of 1 s^{-1} and turned off after a time $t_{OFF} = 20min$, describing a transcriptional 'pulse'. In the upper graph of Fig. 4.5B the behavior of the two controller schemes (red - Sad-iFFL, blue - U-iFFL) in response to a transcriptional 'pulse', reported in the graphs above, highlighted by the blue vertical box. As usual, the theoretical set point K_{CG} is reported with black dotted line in the above graph, set to 500 nM. Sad-iFFL is not capable to maintain its steady state after turning the transcription off (it goes down to the initial state after the pulse) as shown in Section 3.3.4. Differently, the U-iFFL circuit shows the capability to maintain the GOI expression at the desired level after the pulse. This feature makes the system work without necessity of constant presence of the inducer molecule and its behavior is the same as if it was constantly induced. In fact, positive autoregulation is the simplest network motif implementing a genetic "on-memory", as the cell is able to "remember" the active state after the upstream promoter is transiently triggered [116, 60]. In particular, after occurs the first transcription events from the P_C promoter (trigger) and translation by the RBS upstream the *rnapt7* gene, a small quantity of RNAPT7 protein starts to transcibe its own transcript and create the aforementioned 'on-memory' effect. The transition between the off-/on-state depends on the dissociation constant K_{T7} value [117, 60], indeed, if the protein produced by the trigger events (P_C) is not high enought to activate the P_{T7} promoter, the circuit remains in the off-state. To understand if the system composed by RNAPT7-P_{T7} is appropriate for this application, the characterization of the module has been done *in-vivo* and the valueof the dissociation costant K_{T7} has been estimated

(data and discussion have been reported in Section 4.2.4). This feature can represent an advantage of the U-iFFL circuit over Sad-iFFL, depending on its application.

4.2.4 *In vivo* characterization of RNAPT7-P_{T7} transcriptional system from T7 phage

The set of characterizion circuits of RNAPT7-P_{T7} transcriptional system, reported in Table C.6, is based on the biological scheme reported in Fig. C.2; briefly, the fluorescence protein (RFP) is driven by the P_{T7} promoter, which is recognized by RNAPT7, expressed by an HSL-inducible expression cassette; the cell load is monitored from the GFP monitor cassette. The full description of the biological scheme is reported in Section C.4.2. The circuits hereby reported has been obtained after the optimization of RNAPT7 coding sequence: the high toxicity that arose from the wild-type RNAPT7 enzyme was incompatible with cell life, therefore, the enzyme efficiency was decreased through the modification of the RNAPT7 coding sequence. This choice has been made, among different solutions proposed in the literature (e.g., start codon modification [62, 118], mutation of P_{T7} [118], P_{T7} transcriptional repression [119, 120]), for its semplificity and efficacy.

Despite the literature solution of promoter repression is included in the U-iFFL circuit (P_{T7} repressed by SadCas9), it has been demontrated in our laboratory that the transcriptional activation-repression system is not functional with the wild-type RNAPT7 as the system remains constantly in the off-state due to SadCas9 repression and due to cell burden generated from RNAPT7 activity when the SadCas9 concentration falls below a threshold limit (data not shown). The successfull modification carried out in the engineered system is the amino acid substitution R632S of RNAPT7. The nine possible circuit combinations have been designed and constructed with three differ-

ent RBSs (B0031, B0032, B0034), and the transfer functions obtained from the experimental data are shown in Fig. 4.6. The functionality of the system can be outlined in two properties based on the toxicity and activity of the RNAPT7-P_{T7} system. In fact, altough the optimized system reached a non-toxic high target level for HSL concentration values below 0.5 [nM], the behavior changes for HSL above this threshold value and cell burden arose in the cell, as shown in the GFP and growth rate data in Fig. C.2. From these data it emerges the importance to control the steady-state value of RNAP7 when used in synthetic circuit. The transfer functions reported in the middle graph of Fig. 4.6 are clustered in three separated groups accordingly on target RBS strength (from the top-strongest to the bottom-weaker B0034, B0032 and B0031). The independence of each group from the RNAPT7 RBS strength and the high output level reached for HSL = 0 [nM] value indicate that the transcriptional activity of the P_{T7} promoter is saturated for very low concentration of RNAPT7 that arose from the leakage activity of P_{Lux} promoter, thereby suggesting that the half-maximum constant of RNAPT7 is too low to be estimated from the experimental analysis of such circuits, differently from the SadCas9 characterization scenario, for which a wide and tunable repression range was observed. This results lays the foundations for a robust and stable circuit functionality since, for biologically reasonable value of c, the hypothesis assumption (Hypothesis 6: $T7 >> K_{T7}$) of the model (discussed in Section 4.2.2) is easily guaranteed. Moreover, as previously discussed, the K_{T7} is low enough to guarantee the memory effect discussed in Section 4.2.3 (Induction/de-induction cycle). Despite the whole set parts of U-iFFL circuit is characterized, its assembly has been interrupted due to Sars-Cov-2 worldwide pandemic and the work is still on-going.

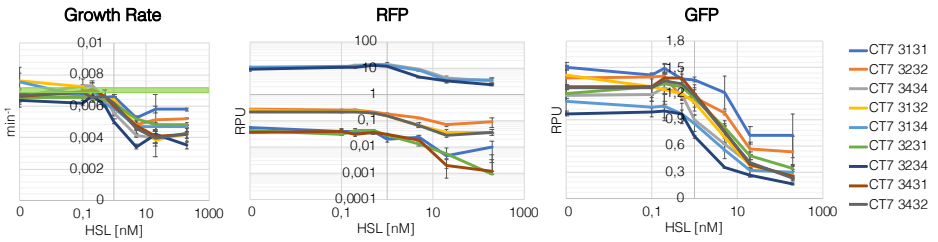

Figure 4.6: **Experimental data of steady-state transfer functions of the RNAPT7-P_{T7} system.** Experimental data in which the error bars represent the standard error of the mean ($n >= 3$ biological replicates for each point). The thick green line in the 'Growth Rate' graph represents the growth rate value of the strain used in this study with no plasmids, used as growth control.

4.3 U-iFFL overall conclusion

In this chapter, a new architecture based on iFFLs and positive feedback sub-networks has been developed in order to overcome the limits that emerged from the Sad-iFFL *in vivo* and *in silico* performances, discussed in the previous chapter. The mathematical model developed from the biological scheme has been deeply analyzed in its constraints and used to simulate the steady-state and dynamic curves which were subsequently compared with the Sad-iFFL ones. The steady-state comparisons between Sad-iFFL and U-iFFL showed that the new architecture is more robust to transcription and translation rate variations in a wide range of theoretical set point values (K_{CG}) and, furthermore, it has been demontrated that the RNAPT7-P_{T7} system does not propagate the noise in a different way to Sad-iFFL. The dynamic analysis demonstrates the importance of the genetic 'on-memory' achieved from the positive autoregulation loop that may be beneficial in applications for which protein expression needs to be triggered and maintained after the removal of inducers. The amplification module composed by the transcriptional activation system of phage T7 has been optimized to limit its toxicity due to a too high

RNAPT7 activity on P_{T7}, and characterized *in vivo*. This analysis validated the overabundances hypothesis of the mathemtical model , thereby indicating that, for biologically measured values of the Sad-Cas9 and T7 systems, the U-iFFL is expected to work and show a more stable behavior than Sad-iFFL due to less restrictive assumptions. Despite the whole set parts of U-iFFL circuit is characterized, its assembly has been stopped due to Sars-Cov-2 worldwide pandemic and the work is still on-going.

Chapter 5

A bioinformatics approach to expand the iFFL portability in different microorganisms

The functionality of the automatic controllers (Sad-iFFL, U-iFFL) explained in Chapter 3 and 4 rely on the availability of regulatory parts (e.g., promoters, RBS) that drive the genes in the two circuits. Under a set of assumptions, illustrated in detail in Chapter 3 and 4, the circuits are expected to adjust the target protein level even if transcription, translation and copy number varies, thereby making the circuits portable modules for protein expression. In this chapter, to support the rational choice of promoters and RBSs, two bioinformatic pipelines based on publicly available high-throughput gene expression data and genome sequences have been developed. After a brief overview on the different strategies adopted in the literature to discover new regulatory parts in bacteria, the transcriptomic and genomic datasets used are illustrated with the thorough exlpanation of the two bioinformatics pipelines. Finally, the results derived from these pipelines and the validation of some of the obtained findings have been reported with

the conclusions and future developments of this work.

5.1 Introduction and project idea

The control of the gene expression can be achieved using previously characterized regulatory parts such as promoters, RBSs and plasmid vectors which in turn determine the transcription and translation rates and the copy number. Fully characterized libraries of regulatory parts, largely known for model bacteria, are not available for non-model bacteria thus limiting the rational design construction of new synthetic circuits for new biological chassis, limitation that decrease the Sad-iFFL (Chapter 3) and U-iFFL (Chapter 4) circuits portability [121]. The interest in the use of non-model organisms is increasing to expand the application range of synthetic biology, as motivated in Chapter 1. Promoters are defined as DNA sequences recognized by the enzyme complex composed of RNA polymerase and sigma factor (σ) responsible for regulating the transcription of the downstream gene into RNA (Fig. 5.1A) [122]. Adaptation to environmental perturbations (e.g., stress, temperature, pH) is essential for the survival of the bacterium, which must rapidly adjust its physiological response. This is possible through the activation of regulatory genes that are turned off or poorly expressed normal conditions, when, e.g., they are not necessary; in fact, usually, the cell does not express all the genes within its genome but modulates the gene expression pattern according to the environmental condition, avoiding to sequester energy from the most important biological function, cell replication [123]. The family of proteins responsible for the modulation of different gene expression patterns is sigma factors which direct the RNA polymerase towards different promoter classes. In bacteria there are several σ-factors but, usually, only one of them promotes the transcription of constitutive genes (e.g., σ^{70} in *E. coli*, σ^{A} in *B. subtilis*) [122]. A promoter is char-

acterized mainly by: the position of the Transcription Start Site (TSS, or +1), the -10 and -35 regions, so called because they are respectively located at about 10 and 35 nucleotides upstream of TSS (Fig. 5.1A). The convention for which these sequences have been marked with the numbers -10 and -35 is linked to the first studies conducted on bacterial promoters in which it was evident that the frequency of *consensus* sequences resided at about -10 and -35 bases from the TSS. In subsequent studies, it has been shown that this is not always respected: in fact, the distance between the TSS and the -10 region can vary from 4 to 12 bases, as well as the length of the core region, the region between the -10 region and the -35 region, which can vary from 15 to 18 bp [124].

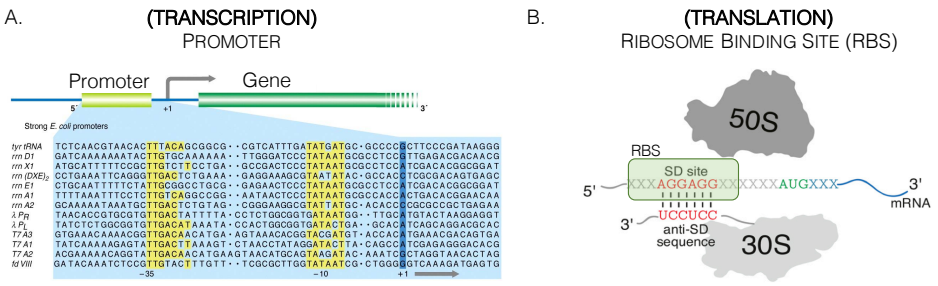

Figure 5.1: **Promoter and Ribosome Binding Site (RBS) sequence.** **A.** The recognition of the promoter sequence (light green box) by RNA polymerase (RNAP) allows the transcriptional process of the downstream gene (dark green box) starting from the Transcription Start Site (TSS, +1). Precisely, RNA polymerase recognizes promoter subsequencies (yellow boxes), called -35 and -10, based on the sigma factor protein that forms a complex with this enzyme. Different *Escherichia coli* promoter sequences recognized by the complex formed between RNAP and σ_{70} have been reported in the graph [125]. **B.** The recognition of the Ribosome Binding Site (RBS - green box) from the ribosome complex (50S + 30S) allows the translational process of the downstream gene (from green AUG) into protein to occur. This process is enabled by the recognition of a sequence within RBS, called SD (Shine-Dalgarno), by the anti-SD (anti Shine-Dalgarno) present in the 16S rRNA gene embedded in the ribosomal subunit 30S [126].

Promoters can be classified into two classes based on their behavior: constitutive or inducible. Constitutive promoters transcribe the

downstream gene regardless of the environmental conditions and their transcription rates depends strictly on the sequence of the promoter [127]. In fact, the variation of a few nucleotides within the promoter sequence can significantly alter its strength. On the other hand, inducible promoters alter their transcription rate as a result of an external stimulus represented by a physical agent, such as temperature, or by a chemical agent such as a specific molecule. Although in some cases of real-world application of engineered organisms the use of inducible promoters has been successful in rationally modulating gene expression, in other cases they are inadequate: the promoter can be hypersensitive to the inducer, its high cost can also make the system unusable in some contexts and, importantly, the levels of expression reached the population may fluctuate significantly within isogenic bacteria leading to a large heterogeneity in gene expression. In light of this, it is generally preferred to use stable and robust constitutive promoters with known strength in the construction of synthetic circuits to be inserted inside the host microorganism [128]. The product of the transcription process is the mRNA which the coding sequence of the gene is present, included within two UnTranslated Regions: 3'UTR and 5'UTR. The 5'UTR region includes the RBS sequence responsible for the recognition of the ribosome and thus the beginning of the translation process (Fig. 5.1B). Precisely, a short *consensus* sequence called Shine-Dalgarno (SD) is contained in the RBS and the mechanism by which the ribosome recognizes it is based on the complementarity between the RNA bases: the SD sequence is complementary to short regions near the 3' end of 16S rRNA, called in turns anti-SD. This sequence was first discovered by Shine and Dalgarno in *E. coli*, however some genes such as for example *rpsA*, they do not present it in their 5' UTR sequence, but the mRNA is still translated efficiently [129]. Many authors use the term Shine-Dalgarno sequence to indicate the AAGGAGG sequence (characteristic of *E. coli* and other bacteria), while others use it to indicate more generally the regions

of the 5' UTR that are recognized by the ribosomes of these organisms. In fact, alternative forms of the AAGGAGG *consensus* sequence have been discovered: many genes of *Cyanobacteria* and *Bacteroides* possess a sequence significantly different from *E. coli* [130]. The traditional approach to discover new promoter or RBS sequences consists in taking portions of the genome for which it is assumed that a regulatory sequence is present, inserting it upstream of a reported gene within an expression vector builted *ad hoc* for that microorganism and quantifying the fluorescence of the downstream gene and compared to a previously characterized part used as a reference. The high amount of time and work behind this process motivated the need of alternative approaches that could identify *in silico* the regulatory sequences and possibly predict their activity. In the literature, several solutions based on a bioinformatics approaches have been conceived and representatives examples have been reported below.

Bioinformatics algorithms for the prediction of promoter sequences based on different modeling approaches have been developed. PePPER [131], BPROM [132], 70ProPred [133] and BacPP [134] are algorithms which based their functionality on the *consensus* sequences recognized by sigma factors. In particular, all the mentioned ones use the information of *E. coli* σ-factors, thus limiting the algorithm portability for species phylogenetically close to it, for which it is assumed that the transcriptional mechanism does not change in a decisive way. Algorithms such as CNNpromoter_b [135] and SAPPHIRE [136], on the other hand, have been developed with a convolutionary neural networks (CNN) modeling to predict promoters based on the knowledge acquired during the training phases on specific datasets of validated promoters. Unfortunately, as is well known in machine learning, their prediction performances strongly depends on the training set used and by its size; in fact, it has been shown that small training sets result in poor accuracy values of the pipeline, thereby limiting its generalization power [135]. A further developed approach is based on bio-

physical models which rely on the calculation of energy variation on moving windows on the whole genome (G4PromFinder and PromPredict) [137, 138]. It has been assumed that the promoter sequences possess, compared with the rest of the genome, binding energy values between adjacent nucleotides significantly lower in order to help RNA polymerase to open the DNA portion and start the transcription process. Similar bioinformatics approaches have also been used for the prediction of RBS sequences within a target genome. Tompa et al. optimized an algorithm that is able to detect short motifs in multiples sequences (such as RBSs in the regions upstream of coding sequences) that overcomes the bias of the current local alignment algorithms for which suboptimal solutions, characterized by short alignments with exact matches, are discarded by the presence of long alignments with the presence of multiple indels. However, the algorithm showed low performances when tested to detect RBS sequences (more complex than a short conserved motif) on different bacterial species due to the high percentages of indels compared with nucleotides on the computed RBS sequence [139].

The state-of-art algorithm for the identification of RBS strength is the biophysical model that takes into account the secondary structure of the mRNA and the hybridization energy between RBS and anti-SD sequences in the rRNA, called RBS Calculator (Salis Lab). In its reverse engineering mode, the method predicts the translation initiation rate for each detected start codon in a given mRNA ($\sim$ 100 bp long) while in its forward engineering mode the method optimizes a synthetic RBS sequence to achieve a targeted translation initiation rate. In particular, the algorithm employs a thermodynamic model of the starting events of bacterial translation to calculate the Gibbs free energy of the ribosomal binding. Using a thermodynamic approach, they related this Gibbs free energy to the translation initiation rate of the ORF [101, 106, 107, 108]. Although the algorithm is widely used to design new RBSs associated with various translation efficiency values,

the accuracy of the prediction of this efficiency is still too low to allow the precise engineering of expression systems in poorly studied hosts. The limitations that emerged from the algorithms previously reported leave the door open for the development of new high-portability bioinformatics algorithms for the identification and possible quantitative prediction of regulatory sequences that can be used in different bacterial strains. In this chapter, to support the rational choice of promoters and RBSs, two bioinformatic data-driven pipelines based on publicly available high-throughput expression data and genome sequences have been developed in order to detect candidate and constitutive promoter sequences with stable activity across different conditions, and design new RBSs that are functional in different bacterial hosts in a wide range of strengths of transcription and translation processes.

5.2 Bioinformatics procedures

Genomics data of promoter sequences were retrieved from public online databas and have been analyzed in MATLAB R2020a (MathWorks, Natick, MA, USA) while transcriptiomics data retrieved from DNA microarray and NGS experiments have been analyzed in R v.3.6.2 software. For NGS data manipulation, where indicated, python v.3.8.5, Linux bash and shell environment have been called through the R software. Clustal-Omega v.1.2.4 software has been embedded in the MATLAB environment and all multiple alignments have been made locally in shell environment in a computer with 2.9 GHz Intel Core i7 processor and 8 GB 1600 MHz DDR3 RAM.

5.2.1 Data resources: genomics and transcriptomics datasets

Genomics datasets

The complete annotated genomes of 120 reference bacterial microorganisms have been obtained in two steps. First, a summary list (in .xls format) has been downloaded from *NCBI Assembly* online database with the following query options: status - 'latest', assembly level - 'complete genome', category - 'representative', exclude - 'partial, derived from surveillance project, anomalous' and annotation status - 'has annotation'. Subsequently, the genomes informations have been retrieved using the *getgenbank* function in MATLAB using the 'Chromosome GenBank' accession code for each bacterium listed.

The complete genomes (Training sets) of *Escherichia coli* str *K-12* substr. *MG1655* (GenBank code: *U00096.3*) and *Bacillus subtilis* subsp. *subtilis* str. *168* (GenBank code: *AL009126.3*) used to test the developed pipeline to detect promoter sequences inside a genomes have been taken from the 120 microorgarnism genomes previously retrieved, while their FASTA genome sequences have been retrieved from *NCBI Assembly* repository. Datasets of promoter sequences (Test sets) of *E. coli* and *B. subtilis* have been retrieved from *EcoCyc* and *DBTBD* databases, respectively. In *EcoCyc*, 1700 promoter sequences with their relative sigma factor(s) and Transcription Start Site (TSS) were present (evidence code - experimental evidence), classified as follows: $\sigma^{70} - 795$, $\sigma^{38} - 181$, $\sigma^{E} - 67$, $\sigma^{32} - 65$, $\sigma^{54} - 39$, $\sigma^{28} - 10$ and $\sigma^{FecI} - 1$, while, in *B. subtilis* 673 promoter sequences with their relative sigma factor(s) were present (evidence code - experimental evidence), classified as follows: $\sigma^{A} - 368$, $\sigma^{B} - 40$, $\sigma^{D} - 13$, $\sigma^{E} - 67$, $\sigma^{F} - 24$, $\sigma^{G} - 52$, $\sigma^{H} - 21$, $\sigma^{K} - 47$, $\sigma^{M} - 3$, $\sigma^{W} - 25$ and $\sigma^{X} - 13$. In the *B. subtilis* test set, the TSS position for each promoter has been computed adding +1 bp to the right absolute position of the promoter sequence if relies on the non-complement strand referred to the genome notation,

otherwise (complement strand) -1 bp has been substracted to the left absolute position of the promoter sequence.

Transcriptomic datasets

Transcriptomics databases have been retrieved from the online database *Gene Expression Omnibus (GEO)* supported by *NCBI* at the *National Library of Medicine (NLM)*. Differently from NGS data, two platforms (GPL - Geo PLatform) have been used to retrieve the transcriptomics experiments of *E. coli* and *B. subtilis* based on DNA-microarray technology: GPL199 (chip manufacturer: *Affymetrix*) and GPL10901 (chip manufacturer: *Agilent*), respectively. In fact, the lack of standardization protocols for data storing of microarray-based experiments on public databases (e.g., *NCBI GEO*) represents a serious limitation in the generalization of the pipeline for which, currently, an adapatation module to integrate data from microarray chips different than those used in this study (e.g., Affymetrix and Agilent) is under development. Nonetheless, these chips brands have been selected since cover a wide subset of microarray experiments ($\sim 44\%$ of GSEs, 3612 GSEs in total). The sets of experiments (GSE - Gene SEries) for each GPL have been listed in Table 5.1. For each experiment, the number of samples (GSM - Geo SaMple), the experimental condition analyzed and its reference have been reported. The experiments (GSE) have been selected by growth condition (mid-exponential phase of the bacterial growth), time-evolution series (static, no dynamic studies have been selected) and if raw data were present. In total, 172 and 96 samples have been collected for *E. coli* and *B. subtilis*, respectively.

Table 5.1: **Experimental datasets used to estimate constitutive and stable genes in *E. coli* and *B. subtilis*.** The transcriptomics datasets used in the first pipeline are reported, divided by technology - Microarray and NGS - for *E. coli* and *B. subtilis* microorganisms. Two platforms have been used for *E. coli* (GPL199) and *B. subtilis* (GPL10901) for microarrays experiments, respectively. The experiments (GSEs) that have been selected are reported with their GSM numerosity, the experimental condition(s) that has(ve) been studied and the reference of the study.

Technology	GPL Code	GSE Code	GSM Number	Experimental condition	Ref.
			Escherichia coli		
Microarray	GPL199	GSE4511	15	pH 5∼8.7, $+O_2$	[140]
		GSE4556	15	pH 5∼8.7, $-O_2$	[141]
		GSE1106	21	$\Delta appY$,$\Delta arcA$,Δfnr,$\Delta oxyR$,$\Delta soxS$, $-O_2$	[142]
		GSE1107	22	$\Delta appY$,$\Delta arcA$,Δfnr,$\Delta oxyR$,$\Delta soxS$, $+O_2$	[143]
		GSE21869	33	Chloramphenicol, Kanamycin	[144]
		GSE22829	6	$\Delta nusA$±nitrofurazone (NFZ)	[145]
NGS	GPL26204	GSE126710	9	Infection from Stx2 phage	[146]
	GPL18956	GSE135706	10	$\Delta rnhA$, $\Delta dnaA$	[147]
	GPL24377	GSE135867	12	$\Delta ubiC$, $\Delta pdhR$	[148]
	GPL16085	GSE65642	12	Δcra, Carbon source variation	[149]
	GPL16085	GSE66481	8	$\Delta gadE$, $\Delta gadW$, $\Delta gadX$	[150]
	GPL21726	GSE80251	9	Erythromycin, Clindamycin	[151]
			Bacillus subtilis		
Microarray	GPL10901	GSE27113	6	DNA damage induced with Argon	[152]
		GSE28872	9	$\Delta ytpQ, \Delta spx$	[153]
		GSE30430	6	Δrny, ±xylose	[154]
		GSE31249	18	$\Delta ywlE, \Delta mcsB$	[155]
		GSE32865	6	DNA damage induced with Argon	[156]
		GSE53333	9	Nutrient limitation and high osmolarity	[157]
NGS	GPL18561	GSE93894	12	$\Delta yacP$	[158]
	GPL18561	GSE94303	6	Δrho	[159]
	GPL18561	GSE120050	6	$\Delta pnpA$	[160]
	GPL23851	GSE121479	6	$\Delta whiA$	[161]
	GPL24109	GSE126233	12	Δefp	[162]

5.2.2 Automatic bioinformatic pipeline to identify promoter sequences with stable activity in bacteria

The pipeline to detect promoters with stable activities in annotated genomes from transcriptomics data is based on two bioinformatics procedures divided as follows. The first has been developed to detect candidate constitutively expressed genes whose expression is stable in transcriptomics data from different experimental conditions, while the second, starting from the genes found with the first pipeline, has been developed to detect the promoter sequences based on the knowledge of the *consensus* sequence related to the sigma factor considered. For promoters that drive the expression of housekeeping genes the sigma factor σ^{70} and σ^{A} for *E. coli* and *B. subtilis* have been used, respectively. The basic assumption is that genes not differentially expressed among different experimental conditions are driven by constitutive promoters whose expression does not vary with external perturbations.

The first part of the pipeline, the analysis of the transcriptomics data, has been divided according to the chip technology, NGS or microarray, and the latter has been divided in turn based on the construction of the DNA-microarray chip (one or two channels - Affymetrix or Agilent, respectively). In fact, data acquired in the two cases are significantly different and their analyses have been adapted accordingly. The overall scheme of the pipeline has been repoted in Fig. 5.2. In the next subparagraphs, will be reported the different variants of the first pipeline that handle microarray and NGS data (numbered points 1-2-3 in Fig. 5.2) in order to detect no differentially expressed (NDE) genes in one experiments, the (common - refers as independent from the type of the initial data) procedure to detect NDE genes on all microarray and NGS experiments (numbered points 4-5-6 in Fig. 5.2)

and, finally, the pipeline that, from the output list of NDE genes of the first pipeline, identify the promoter sequences in the upstream regions of the lattes genes.

Analysis of microarray experiments: Affymetrix chip

The chips produced by Affymetrix (Thermo Fisher Scientific) are used to estimate the level of transcripts inside a cell at a given time by hybridizing the cDNA, obtained from the reverse transcription of the RNA, to probes designed and synthesized *ad hoc* inside the chip. The level of each transcript is estimated from a probe set of about 20 probe pairs of length equal to 25 nucleotides; where a probe pair is composed of two probes called Perfect Match probe (PM), which has been designed to hybridize completely to the sequence of interest, and MisMatch probe (MM), which has been designed to bear a single-point mutation at the center base of the 25-mer probe [163]. The data extracted from the fluorescence image of the chip are used to estimate the intensity of each probe and save the final data in a .CEL file (GSM). These data represent the starting point of the analysis here conducted to identify constant-activity genes through different cell perturbations introduced in several experiments (GSE) made by different research groups (upper part of Table 5.1).

A set of candidate constitutive genes from all the selected GSE experiments has been obtained after a three-step analysis, reported as follows:

1. **Data quality analysis for all GSMs inside each GSE.** The data quality control of each chip is based on the qualitative analysis of the data distribution in each chip and on the evaluation of two indices, called *Relative Log Expression (RLE)* and *Normalized Unscaled Standard Error (NUSE)* calculated both from a

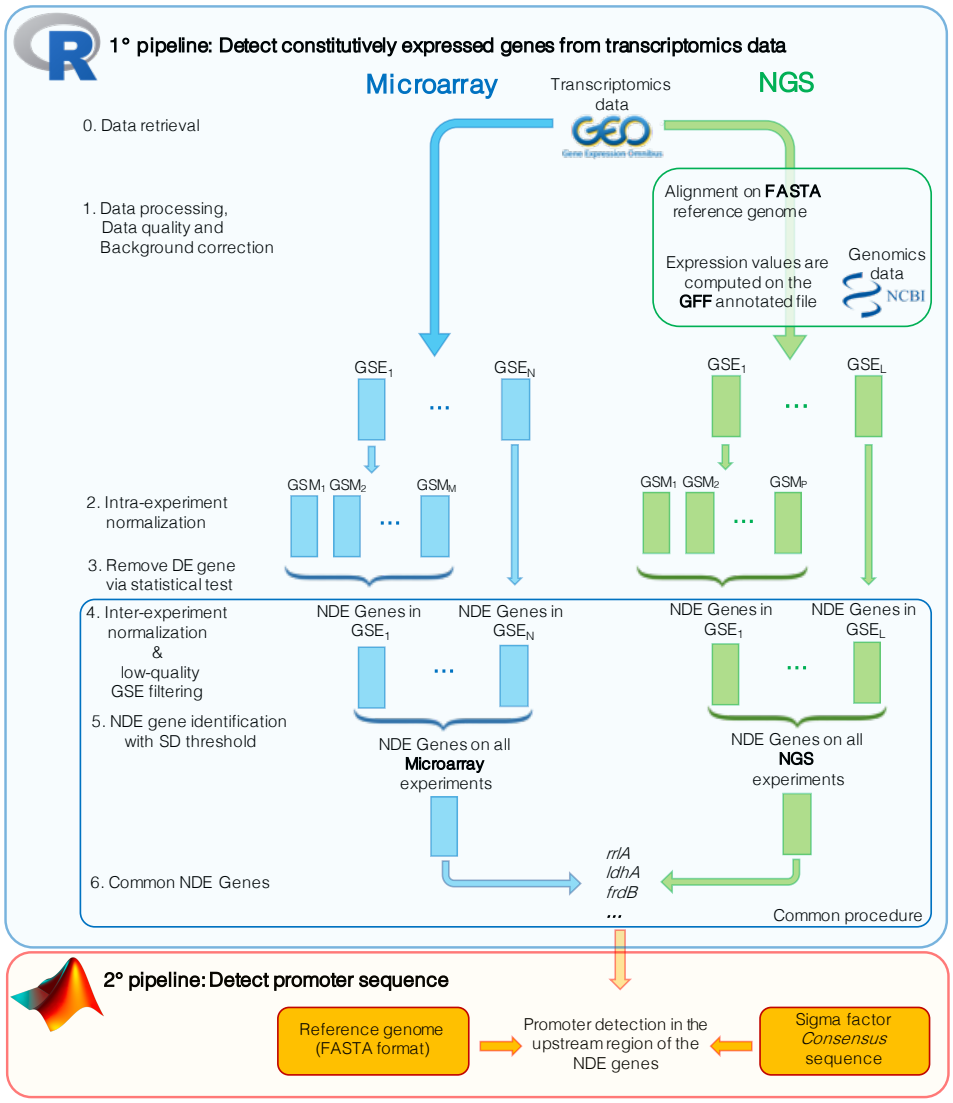

Figure 5.2: **Overall representation of the two bioinformatics pipeline.** The first pipeline - light blue background (R environment) - has been adapted to handle Microarray (blue) and NGS (green) data with different procedures (reported outside the box "common procedure"), due to the different nature of the transcriptomics data, based on a shared rationale reported in the enumerate list on the left part of the graph. The second pipeline - light orange background (MATLAB environment) - starts from the NDE gene list, output of the first pipeline.

data structure called *PLMset* which contains information on the mean, standard deviation and calculated weights of gene expression in each probe set. The PLMset has been obtained through the fitPLM function of the *'affy'* package in R software (in-depth explanation has been reported in [164]).

The RLE index identifies, within a chip, the contribution of the random error due to the variation of environmental conditions (e.g., temperature) among different chips of the same experiment. It is calculated as the distribution of the differences between the expression value of each gene with the median of all values of the same gene in all the chips of the GSE [165]. Since no threshold value has been reported in the literature for which a chip can be considered as low-quality, a *one way ANOVA test* has been applied ($\alpha = 0.05$, H0: all chip have the same mean value) and, if the hypothesis is rejected, *pairwise Tukey test* has been carried out and the chip(s) whose RLE index differs significantly from the others is selected.

The NUSE index represents the distribution of the standard deviations estimates of gene expressions present in a chip. In the literature it has been shown that, in order to consider a good-quality chip, the NUSE index must not exceed the value of 1.05 [165].

The qualitative analysis of data distribution in all chips has been made via the non-parametric *Kruskal-Wallis test* ($\alpha = 0.05$, ANOVA has been not applied due to the violation of normality or homoskedasticity assumptions) which compares the equality of the medians of the distributions. If the null hypothesis is rejected, a *pairwise Wilcoxon signed rank test* has been applied to select the chip(s) which deviates significantly. The chips that do not pass all three of the above criteria are discarded.

2. **Data normalization (RMA - Robust Multi-Array Avarage) inside each GSE.** Data normalization has been made through the RMA method developed by Affymetrix which involves four steps: (i) background correction which it derives from the convolution of the useful signal with signals coming from areas of the chip without probes, (ii) log-data normalization, (iii) Quantile normalization in order to correct the variability of chip data between all chips within the experiment and (iv) data normalization with respect to the probe set between different chips, (identical probe sets of different chips are expected to be related at most linearly). At the end of the RMA procedure, an expression value of a gene is provided as the average of the expression of its associated probe sets [166].

3. **Constitutive genes detection in one experiment (GSE).** The candidate constitutive genes have been selected via the non-parametric *Kruskal-Wallis test* ($\alpha = 0.05$, ANOVA has been not applied due to the violation of normality or homoskedasticity assumptions) based on the mean of the gene expression values on all the GSMs samples. Genes for which the null hypothesis (H0) is not rejected are considered as not differentially expressed. Therefore, for each experiment (GSE) analyzed a set of genes not differentially expressed has been isolated.

Analysis of microarray experiments: Agilent chip

The chips produced by Agilent Technologies are used to estimate the level of transcripts between two samples (usually the same cell culture with and without a perturbation) at a given time by hybridizing the cDNA, obtained from the reverse transcription of the RNA, to probes designed and synthesized *ad hoc* inside the chip. Each set of cDNA obtained from a cell is labeled with a fluorochrome so that the genetic material can be distinguished between the two samples; Cy3

(green) and Cy5 (red) markers are usually adopted. The hybridization between marked cDNA and its relative probe in the chip releases the colored fluorescence [167]. The data on cell transcriptomes are acquired from fluorescence quantification of the image of the chip via an image processing software and saved in .TAR format. For each gene in the latter file the green-red foreground values and green-red background values of fluorescence have been reported. These data represent the starting point of the analysis herein conducted to identify constitutive genes through different cell perturbation conditions, introduced in several experiments (GSE) made by different research groups (lower part of Table 5.1).

A set of candidate constitutively expressed genes from all the GSM samples in one GSE has been obtained after a three-step analysis, reported as follows:

1. **Background correction.** The background signal has been eliminated from the foreground signal using the *backgroundCorrect* function (option set method = 'normexp') implemented in R software which assumes that the background signal has a normal distribution and the noise-free signal an exponential distribution. The background correction has been done independently for the two colors [168].

2. **Data normalization.** Intra- and inter-sample normalizations have been performed to purge the data of systematic effects not associated with biological differences. (i) Intra-sample normalization has been carried out using the *normalizeWithinArrays* function (option set to method = 'loess', where loess = Locally Estimated Scatterplot Smoothing) which considers the signal of a gene as the contribution of several probes which in turns are affected by their relative position in the chip. (ii) Inter-sample normalization has been carried out using the *normalizeBetweenArrays* function (option set to method = 'Aquantile') in order to

equalize the distributions of the two signals. At the end of the normalization process, for each gene, two values have been obtained, M and A, defined in Equation (5.1) and Equation (5.2), respectively [169].

$$M = log_2(R) - log_2(G) \tag{5.1}$$

$$A = \frac{log_2(R) + log_2(G)}{2} \tag{5.2}$$

where M is the fold-change of the expression values of the same gene in the two conditions analyzed; a value of M = 0 indicates that the gene expression is independent of the tested condition, while A represents the mean expression of each gene.

3. **Constitutive genes detection in one experiment (GSE).** The candidate constitutively expressed genes have been selected via *Wilcoxon signed rank test* ($\alpha = 0.05$) on the $\frac{M}{A}$ value of each gene. Therefore, for each experiment (GSE) analyzed a set of genes not differentially expressed has been isolated.

Analysis of NGS experiments

RNA-sequencing is a Next Generation Sequencing (NGS) technology that, like microarray chips, allows to quantify the transcriptome of a biological sample at a given moment in time. In this technology, the chip manufacturing is replaced by the construction of a library of complementary DNA (cDNA) that derives from the total RNA present in the cell where the sequences thus obtained are called reads. Therefore, unlike microarrays, *a priori* knowledge of the gene sequences to be analyzed is not necessary and the signal acquired in the NGS experiements needs an alignment-based reconstruction analysis on the reference genome (FASTA format) of the organism to link them to their genetic *loci*. The NGS data available in the public database

NCBI GEO (downloadable in FASTQ format) are composed by the sequences and their quality score for all the reads sequenced in the experiments. For each sample (GSM) in the experiment, the data (FASTQ format) have been downloaded with the *fastq-dump* function of *SRA Toolkit* in the shell environment, called in turn from R software, while the metadata (e.g., experimental conditions, Taxonomy ID) have been retrieved with the function *metadata* of the python package *pysradb* called from R environment as well.

1. **Data processing and reads quality filtering inside each GSM.** The whole set of reads in each sample (GSM) has been aligned on the reference genome (FASTA format) through the *align* function of the *Rsubread* R package obtaining the BAM file which contains all the alignment informations, such as: genome position and the alignment quality of each reads. In order to discard the reads with a quality score above a threshold limit, the BAM file has been handled in Linux bash environment through the *view* function of the *Samtools* suite; with a quality score of 20, only the aligned reads with a base call accuracy of 99% are mantained. Finally, the gene expression values, called counting value, have been obtained for each gene annotated in the reference genome (GFF format) from the BAM file through the *featureCounts* function of *Rsubread* R package.

2. **Data normalization inside each GSE.** Intra-sample normalization has been made through the *DESeqDataSetFromMatrix* function of *DESeq2* R package. The final value has been obtained from a $\log_2$ transformation biased by a factor of 1 to avoid negative values.

3. **Constitutive genes detection in one experiment (GSE).** The candidate constitutive genes have been selected via the non-parametric *Kruskal-Wallis test* ($\alpha = 0.05$, ANOVA has been

not applied due to the violation of normality or homoskedasticity assumptions) based on the mean of the gene expression values on all the GSEs studies. Genes for which the null hypothesis (H0) is not rejected are considered as not differentially expressed. Therefore, for each experiment (GSE) analyzed a set of genes not differentially expressed has been isolated.

Constitutive genes detection on all GSEs of microarray and NGS data

From the lists of NDE genes isolated within each experiment (GSE), a final list of NDE genes on all the aformentioned experiments have been obtained. This part of the pipeline has been defined as 'common procedure' - light blue box in Fig. 5.2 - since it is independent from the starting data type, differently from the procedures previously explained.

4. **Inter-experiment normalization and low-quality GSE filtering.** The experiments (GSEs) conducted in different laboratories could be affected by non-biological variability that arose from several sources (e.g., DNA extraction quality, different buffers, temperature, operator, different machineries) which could taint the integration of gene expression values when compared together [170]. In order to mitigate this unsolicited variability, a quantile - normalization on all the GSEs (fixed the technology - microarray or NGS) has been performend and the differentially expressed (DE) genes, highlighted in the previous step (3 in Fig. 5.2), were discarded. The experimets which decrease the R^2 coefficient computed among different pairwise experiments by a factor of 50% have been discarded.

5. **Standard deviation (SD) threshold to identify NDE genes on all experiments of microarray or NGS data.** The NDE

genes on all experiments in one technology - microarray or NGS - have been selected filtering the genes which standard deviation values were below a threshold limit.

6. **Identify NDE genes on all experiments.** The final list of NDE genes has been computed by intersecting the two lists obtained in the previous step for the two technologies, microarray and NGS.

Promoter detection pipeline based on sigma factor *consensus* sequence

The detection of promoter sequences inside a genomes of a target microorganisms has been performend via local alignment algorithm (*localalign* function of MATLAB) between the regions 300 bp upstream of each annotated gene and the *consensus* sequence of the *sigma factor*, selected with different core-region length (the region between -35 and -10). The length of the region upstream of the genes has been set to 300 bp as a compromise between the probability to find promoters inside of it (it has been shown that within the 250 bp upstream of the genes there are 89.77 % of bacterial promoters [124]) and the efficiency of the algorithm. In fact, increasing the length of this region the probability of the algorithm to detect false positives (FP) increases (data not shown). The score matrix of the *localalign* function has been modified in order to obtain a linear penalty (from 0 to 1) for uncertainty matches: 1 - match (e.g., A-A), 0.66 - maximum uncertainty of two nucleotides (e.g., A-R where R = [A, T]), 0.33 if maximum uncertainty of three nucleotides (e.g., A-D, where D = [A, G, T]), 0 if uncertainty is maximum (e.g., N, N = [A, G, T, C]) or in case of mismatch. In fact, in the default score matrix, the scores are excessively penalized by the presence of mismatches, limiting the possibility of the recovery of sub-optimal solutions. Indels (INsertions - DELetions) can be inserted to optimize the alignment of the algorithm. For each putative

promoter found in the target genome, a putative TSS has been associated by adding the value of ±8 bp to the distal position of the -10 region, depending on the directionality of the promoter. The comparison with the literature data is based on TSS comparison, in fact, if the true TSS is included inside a window of ± 4 bp on to the position of the putative TSS then the putative promoter has been considered as true positive (TP). The ±4 bp range derives from the evidence that the bp-window between the -10 region and the TSS varies from 4 bp to 12 bp [124]. The test set has been limited to the TSS associated to the promoter sequences recognized by the sigma factor used at the beginning of the pipeline, here σ^{70} and σ^{A}. False negatives (FN) have been considered as the promoter sequences present in the test set that were not detected by the algorithm. Two negative controls, called NC1 and NC2, have been used: (i) NC1 - the test set is composed by the promoter sequences of the sigma factors different from the one that has been used in the pipeline (false positives, FP - promoters found; true negatives, TN - promoters not found), while (ii) NC2 - regions of 300 bp have been taken within the coding sequences with the assumption that no promoters are present (false positives, FP - promoters found; true negatives, TN - promoters not found). For NC2, the median values obtained from 30 independent simulations were analyzed. The performance of the algorithm has been analyzed using three figures of merit:

$$accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{5.3}$$

$$sensitivity = \frac{TP}{TP + FN} \tag{5.4}$$

$$specificity = \frac{TN}{TN + FP} \tag{5.5}$$

Accuracy, sensitivity and specificity are the capabilities to detect a generic istance, promoter sequences within a genetic portion where a

promoters are present and no promoter sequences where no promoters are present, respectively.

5.2.3 Automatic bioinformatic pipeline to estimate RBS *consensus* sequence in bacteria

For each genome of the 120 reference microorganisms, the 18 bp regions upstream of each annotated gene have been taken as follows: 15 bp upstream of the start codon and the 3 bp of the start codon. The length was set at 18 bp as a compromise between the probability of taking an RBS upstream of each gene and the capability of the multiple alignment algorithm (Clustal Omega) to generate a biologically permissible output; in fact, as the length of the regions increases, the proportion of indels (-) become much greater compared with nucleotides (ATGC) - data not shown. The start codon has been kept as a reference to guide the algorithm respect to the starting position of the gene. A *consensus* sequence has been obtained for a list of sequences by multiple alignment via Clustal Omega algorithm and a position-dependent frequency profile of nucleotides (ATGC) or indel (-) has been computed.

5.2.4 Design of new synthetic RBSs based on RBS *consensus* sequence within one or more microorganisms of interest

New synthetic RBSs have been generated from the frequency profile of the *consensus* sequence with the *roulettewheel* algorithm as follows: for each base of the *consensus* sequence, a letter is extracted from the set (ACGT-) proportionally to its frequency of appearance. The sequence thus generated is purified from indels (-) and the start codon. The procedure is repeated for the number of synthetic RBSs to be

generated (N).

5.3 Detection of promoter sequences with stable expression in bacterial genomes

In this section, the results obtained from the single bioinformatics pipelines, illustrated in Section 5.2.2, for the research of genes in bacterial genomes showing stable transcriptional activity in different experimental conditions have been reported and analyzed. At the end of this section, for the constitutively expressed genes obtained from the transcriptomic data of *E. coli* and *B. subtilis*, the upstream regions have been analyzed to detect the promoters sequences recognized by sigma factor σ^{70} and σ^{A} in the two aforementioned species, respectively.

5.3.1 Selection of constitutively expressed genes in public datasets

The scatter plot of the mean gene expression values computed between microarray and NGS chips on all genes (*E. coli* - 4004, *B. subtilis* - 3950) that has been reported in Fig. 5.3A shows a good correlation score for both species infering that, even with the presence of the DE genes, which could decrease the correlation index due to different perturbations on the two technologies, both alternatives - microarray and NGS - are robust sources of transcriptomics data to be used in the pipeline. The validation of the first pipeline has been based on the genes for which, in the literature, have been demonstrated to be constitutively expressed in different perturbations registered. Only the genes with known promoter sequence, sigma factor and experimental evidence code have been used for the validation process. At the end of

the first pipeline, 595 out of 4004 and 223 out of 3950 genes for *E. coli* and *B. subtilis*, respectively, have been identified as non-differentially expressed with a standard deviation threshold value of 0.5, but only 30 and 24 genes have been selected to validate the data based on the aforementioned reasons. The results thus obtained have been reported in the first three columns of Table 5.2 and Table 5.3 for the two microorganisms, where the expression of each genes is explicited in term of mean percentile value in both technologies (Microarray (MA) and NGS); their scatter plot and correlation index have been shown in Fig. 5.3B. For each gene, the information of the upstream promoter(s) has been collected as follows: the sigma factor, the possible transcriptional factors (inside square brackets) and, in order to validate the two pipelines together, the promoter sequence discovered in previous literature studies in which the -35 and -10 regions have been highlighted in capital letters. The data have been collected from *EcoCyc* and *DBTBD* databases and from the literature indicated in references. A diagram tree which summarized the whole data reported in Table 5.2 and Table 5.3 has been represented in Fig. 5.4.

In Fig. 5.4, the transition from the 'Genes' to the 'Promoters' domain consider the nature (e.g., σ^{70}) and the number of promoters found in the upstream sequences of the previously selected stably expressed genes. The sums between nodes of the same level are not exhaustive since it is possible that a gene may have more than one promoter and / or, two genes, if in operon, may share the same promoter.

In Table 5.2, the gene with higher percentile values on both Microarray and NGS is *mdoG*, which has been proposed by Heng et al. to be used as reference gene in *E. coli K-12*. Indeed, in the aforementioned study, it has been shown that *mdoG* has the lowest CV index (= 9.9%) on expression values among different transcriptomics datasets, differently form the 6 housekeeping genes known in the literature (i.e., *recA*, *porC*, *gyrA*, *map*, *rpoC*, *alaS*), tipically used as reference genes for experiment normalization procedure, whose CV indexes

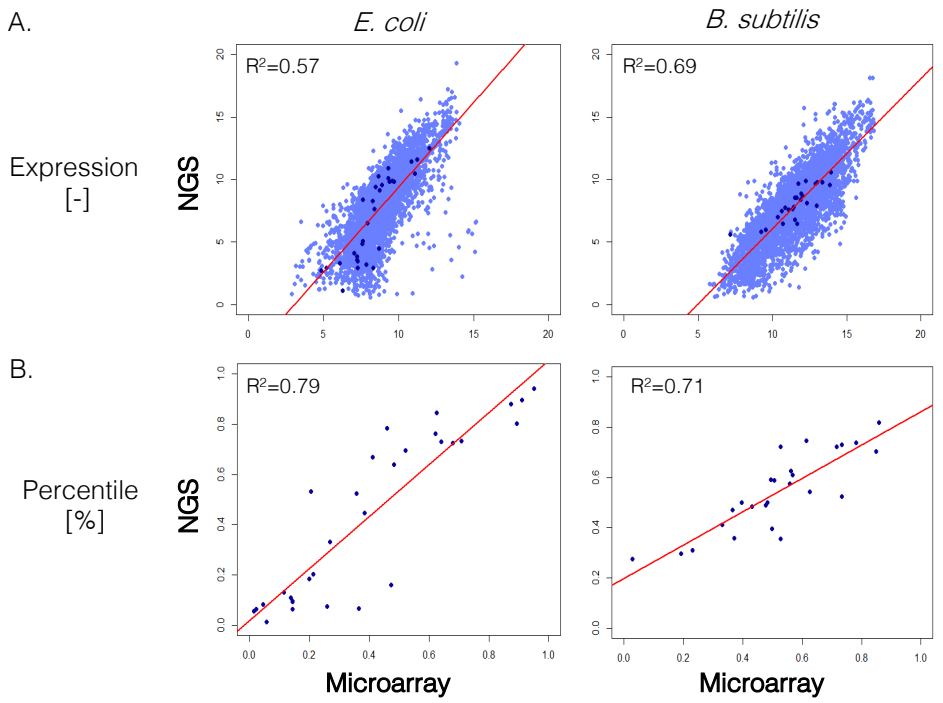

Figure 5.3: **Scatter plot of mean expression values of all genes and mean percentile values of the NDE genes selected from the first pipeline on all experiments between Microarray and NGS data in *E. coli* and *B. subtilis*.** **A**, Scatter plots of mean expression values of all genes (4004 for *E. coli*, 3950 for *B. subtilis*) on all experiments between Microarray and NGS data in *E. coli* and *B. subtilis* are shown - light blue - and the R^2 computed from the linear regression model are reported in the upper left part of each graph. In dark blue are highlighted the NDE genes (40 for *E. coli*, 24 for *B. subtilis*) selected from the first pipeline - and selected as reference set to study since they are annotated in online databases - whose percentile values have been reported in the graphs below (**B**) with the R^2 values described aforementioned as well.

were much higher (16% $\sim$ 53%) [171]. The pipeline detected 5 out of 6 genes within the same operon (*agaS* - *kbaY* - *agaBDI*) despite the promoter is regulated by the transcriptional factor AgaR but, since it

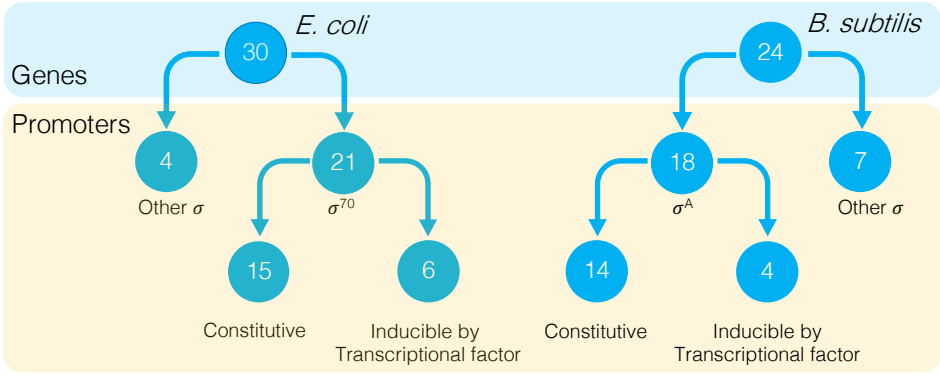

Figure 5.4: **Overview of the genes selected as constitutively expressed.** Tree diagram that summarixes the gene / promoter counts present in Table 5.2 and Table 5.3. The number present in the 'parent node' refers to the number of genes available in the online databases among the set of NDE genes selected at the end of the first pipeline. The transition from the 'Genes' domain to the 'Promoters' domain refers to the nature and the number of promoters found in the sequences upstream of the genes studied. The counts are not exhaustive since it is possible that a gene can has more than one promoter and/or, two genes, if present within an operon, can share the same promoter. The issue with operons has not been considered in the pipeline and will be solved as future work.

becomes active only with the presence of N-acetylgalactosamine and d-galactosamine in the medium and no *E. coli* experiment contains this perturbation, is possibile to infer that the pipeline has identify these genes as NDE correctly. The same results could be translated for *feaB* gene for which, its transcriptional regulator (FeaR) is inactive due to the lackness of its activator (succinate) variation [172]. Differently, 6 genes have been identified wrongly (i.e., *chiA*, *caiT*, *ptrA-recD*, *uvrA*, *rcsD*) since the perturbations (i.e., stress, -O_2) that are responsible to activate the transcriptional factors H-Ns, LexA and ArcA were present in the transcriptomics studies reported in Table 5.1. The genes reported with sigma factor different from σ^{70} have been identified uncorrectly from the pipeline.

5.3. Detection of promoter sequences with stable expression in bacterial genomes

In *B. subtilis* validation set, 3 (i.e., *natB*, *spo0F*, *fabL*) out 17 genes regulated by the housekeeping sigma factor σ^A have been uncorrectly identified as NDE since their regulation by NatR, Spo0A and RsfA are altered by the sodium ion concentration, starvation and stress perturbation, respectively, which are present in the initial transcriptomics experimental set; differently from *fur* gene for which its perturbation (i.e., peroxide stress) is not present. The gene subset composed by *mta*, *cssR*, and *rocR* has been correctly identified as NDE since each of them create a positive or negative feedback loop that regulates its own expression upon different perturbation entities, behavior demonstrated in several studies of network topology [60]. Nonetheless, if a promoter sequence isolated from one of these genes would be used to construct a synthetic expression cassette, it will be self-defeating due to a lackness of ortogonality with the enzymatic host machinery. At the end, the first bioinformatics pipeline detect correctly 20 (66.6 %) and 14 (58.3 %) NDE genes in *E. coli* and *B. subtilis*, respectively.

Table 5.2: **Set of *E. coli* genes filtered by standard deviation (SD) and selected to validate the pipeline.** 30 genes obtained by filtering the list of potential housekeeping genes of *E. coli* with a SD value less than 0.5%. For each gene have been reported: the mean percentile values in Microarray (MA) and NGS data, the sigma factor(s) with the promoter sequence(s) recognized and the reference. Genes within the same operon are reported together mantaining the genome order. If present, on the right of the sigma factor name has been reported the transcriptional factor that regulates the activity of the promoter, in square brackets. Within the promoter sequences have been highlighted in uppercase font the -35 region (left) and the -10 region (right). The promoter sequences have been taken from 3 bp upstream the -35 region to the base before the TSS. The informations have been collected by EcoCyc database and, if a reference is present, in scientific articles.

Gene	Percentile MA $\sim$ NGS[%]	Sigma Factor	Promoter sequence	Ref.
yiaKM	1.5 $\sim$ 5.4	σ^{70}	agaTTGCCTtgttaaaaagtgatcgaTATATTtgaaatca	[173]
yjjQ	2.2 $\sim$ 6.3	σ^{70}	tatTTCACCaggttatcatttgatgcTAACCTtctgac	[174]
agaS-kbaY-agaBDI	14.3 $\sim$ 9.0	σ^{70}[AgaR]	ccaTTGAACtttcagtttcttttcTATAGAttttaatcaa	[175]
rihB	14.5 $\sim$ 6.3	σ^{70}	tatTTAGCAtacgtacgctgttgtTAAGATtcaatc	[124]
ulaA	20.1 $\sim$ 18.3	σ^{70}	atcTTGTTGtgattacttttgaaaatTAGAGTgagtgc	[176]
chiA	20.5 $\sim$ 53.0	σ^{70}[H-Ns]	ataTTGAAGggttctcgtaaaacgtaAATAATtcctgc	[177]
caiT	27.0 $\sim$ 33.0	σ^{70}[H-Ns]	ccaATCACAgaatacagcttattgaataccCATTATgagtta	[178]
paaX	35.7 $\sim$ 52.4	σ^{70}	ccaGGGCCAgaagtcgatacgacctgtgcTATGATtcata	[179]
feaB	38.4 $\sim$ 44.6	σ^{70}[FeaR]	aaaGTGACTtttcttgtcgctgcgTACACTgaaatc	[172]
yoaE	41.1 $\sim$ 66.9	σ^{70}	tctTTGCCAaatcagaggcgtctctgaTATGTTTaact	[124]
hemN	46.0 $\sim$ 78.2	σ^{70}	atgTTATCTggtgtggttattcgtTAAACTacagcg	[180]
alsR	47.4 $\sim$ 15.8	σ^{70}	accAGAAAAacaaataacatcatgttttTAAACTaattaaatg	[181]
ybeD	48.3 $\sim$ 63.8	σ^{70}	cacTTGAAAgtgtaatttccgtccccaTATACTaagc	[182]
yajG	52.2 $\sim$ 69.5	σ^{70}	cacTTCCCCtcgccgttggcaatgtTATGATggcgg	[183]
ptrA-recD	55.9 $\sim$ 60.4	σ^{70}[LexA, ArcA]	aaaTTGCGCaatctatccgcttacttTATGATgcgca	[184]
argR	62.1 $\sim$ 76.2	σ^{70}	tgtTTGCATaaaaattcatcttgtatgcacAATAATgtt	[185]
uvrA	62.5 $\sim$ 84.4	σ^{70}[LexA, ArcA]	tatTTCATTcaggtcaatttgtgtCATAATta	[186]
rcsD	87.4 $\sim$ 88.1	σ^{70}[H-Ns]	agtCTGGAAtttcacactgtaccttTATACTgccct	[187]
mdoG	91.1 $\sim$ 89.7	σ^{70}, σ^{70}	tgcTGAACGataccgggattctgttgtcggAATGGCtgg, ctgTTGTCGgaatggctggttatccatTAAAATagatcgg	[188]
fabB	95.1 $\sim$ 94.0	σ^{70}	gacTTGTTCggcgtacaagtgtacgcTATTGTgcattc	[189]
atoE	13.8 $\sim$ 10.9	σ^{54}		
arrD	21.2 $\sim$ 20.2	σ^{24}		
pldB	70.7 $\sim$ 73.3	σ^{32}		
phoP	89.4 $\sim$ 80.2	σ^{32}		

Table 5.3: **Set of *B. subtilis* genes filtered by standard deviation (SD) and selected to validate the pipeline.** 24 genes obtained by filtering the list of potential housekeeping genes of *E. coli* with a SD value less than 0.5%. For each gene have been reported: the mean percentile values in Microarray (MA) and NGS data, the sigma factor(s) with the promoter sequence(s) recognized and the reference. Genes within the same operon are reported together mantaining the genome order. If present, on the right of the sigma factor name has been reported the transcriptional factor that regulates the activity of the promoter, in square brackets. Within the promoter sequences have been highlighted in uppercase font the -35 region (left) and the -10 region (right). The promoter sequences have been taken from 3 bp upstream the -35 region to the base before the TSS. The informations have been collected by DBTBS database and, if a reference is present, in scientific articles.

Gene	Percentile MA ∼ NGS[%]	Sigma Factor	Promoter sequence	Ref.
ydfJ	2.9 ∼ 27.3	σ^A	tttTTTTTCAtttcattgtcaacTACAATgagaaa	[190]
mta	19.1 ∼ 29.5	σ^A[Mta]	ggaTTGACCctaacgttgcgtgattgttTACGATaaaaaa	[191]
natB	36.5 ∼ 47.0	σ^A[NatR]	ttaTCCAACtaatccagctttttcgtaTATAGTcattact	[192]
bglC	37.1 ∼ 35.7	σ^A	tgaTAGACAatcatgagaaaagattttTACAATgagttc	[193]
yaaJ	39.7 ∼ 50.0	σ^A	caaTCGACAgtctcctttccgtttcagtTATAGTtaatatgtag	[194]
citA	43.1 ∼ 48.3	σ^A	ataTTGATTatttttttaaatattatatatTTTACTataata	[195]
ogt	47.8 ∼ 48.9	σ^A	aacTGGACTtgggcttatggtaagctaTAAAAATtattgaagaac	[196]
cssR	48.2 ∼ 49.9	σ^A[CssR], σ^A	tttTTCTCTCTtcctttcttccctaaccaTATCATaaacaaa, aatTGCCTGccgtatgttaaaactagttTATAATgacgtt	[197]
ywbI	50.5 ∼ 58.8	σ^A	ttaTTGTACataaggtctctctataggTAAAATatat	[198]
rocR	56.3 ∼ 62.6	σ^A[RocR]	cttTTGCATatcctctccgttttttttaTAAAATagaagc	[199]
ytrDEF	56.8 ∼ 60.8	σ^A	ggaTTGACTttgtgagtcaaagtattTATTGTattaag	[200]
albD	62.5 ∼ 54.2	σ^A	gtaTTGAATtagtaatttgatagtttTAAGATaaaagt	[201]
fur	71.6 ∼ 72.2	σ^A[PerR]	tagTTGGAActctgcgcgtatttttgtTATAATgagtc	[202]
spo0F	73.3 ∼ 52.3	σ^A[Spo0A]	gctCAGAAAatgtcgtaaagtagactatTATAATtaaa	[203]
pabB	73.3 ∼ 73.0	σ^A	aaaTTCACTtttttcactaacaacatTGCTTTacaattaaaa	[204]
comA	78.0 ∼ 73.8	σ^A	gacTTGGCAcaggccaagtctttttTATAAAatgga	[205]
fabL	85.8 ∼ 81.9	σ^A[RsfA]	tgaTTGGCCttcttcgttccaaataaggCATAATttgct	[206]
gerKB	23.0 ∼ 30.8	σ^G		
spoVAF	33.0 ∼ 41.1	σ^G		
ylbB	49.7 ∼ 39.6	σ^F		
splA	52.7 ∼ 35.5	σ^G		
ywbD	52.7 ∼ 72.2	σ^B		
ywbF	55.8 ∼ 57.4	σ^B		
yaaR	84.9 ∼ 70.3	σ^E		

5.3.2 Performances of the promoter identification pipeline on the literature-validated test sets

The performances of the pipeline for the detection of promoter sequences within bacterial genomes have been computed as illustrated at the end of Section 5.2.2 on a test set of promoters with numerosity equal to 1700 (σ^{70} = 795) and 673 (σ^{A} = 368) for *E. coli* and *B. subtilis*, respectively. Two negative controls have been used: in the first one, the test set used has been composed by the set of promoter sequences recognized by sigma factors different from the one selected (e.g., σ^{70}, σ^{A}) - NC1, while in second one, sequences of 300 bp within genes of the microorganism have been taken randomly to search for promoter sequences - NC2.

The values of sensitivity, specificity and accuracy obtained are (42 %, [80-98.3] %, [51.5-70.1] %) and (61 %, [72-99] %, [66.1-80.2] %) for *E. coli* and *B. subtilis* respectively, where the specificity and accuracy values were calculated with the negative control, NC1 (first value reported in square brackets) and NC2 (second value). The sensitivity has a unique value because it is independent on the TN and FP values (determined by the negative control sets). The lower values given by the negative control NC1 compared to NC2 indicate that the algorithm is less specific, and therefore less accurate, when tested on the promoter sequences recognized by other sigma factors. This is due to the fact that different sigma factors can recognize similar sequences within the same microorganism (e.g., σ^{70} and σ^{38} of *E. coli* recognize the *consensus* sequence TTGACA(-35) − TATAAT(-10) and TTGACA(-35)−TATACT(-10), respectively).

Regardless of the negative control procedure, the algorithm has a satisfactory capability to discriminate non-promoter sequences (high specificity) and good sensitivity and accuracy values have been obtained for the two microorganisms.

The pipeline performances have been computed to the previously

identificated genes, in Section 5.3.1. From the constitutives genes selected, the pipeline to detect promoter sequence has been applied. On 21 and 18 constitutive promoters with experimental evidence, 12 (57.14%) and 12 (66.66%) promoters have been predicted correctly for *E. coli* and *B. subtilis*, respectively.

5.4 RBS *consensus* sequence estimation and new RBSs design

The *consensus* sequence obtained for *Escherichia coli* str. *K-12* substr. *MG1655* via multiple alignment of the 18 bp sequences upstream of 4357 annotated coding sequences is shown in Fig. 5.5A. The final sequence has been obtained by removal of the gaps introduced by Clustal Omega algorithm. In the *consensus* sequence shown in Fig. 5.5A an AG-rich region has been detected upstream of the start codon, correlated with the *consensus* sequence estimated in the literature (5'-AGGAGG-3') for *Escherichia coli* at a distance of 6∼8 bp upstream of the start codon [130]. This procedure has been repeated for all 120 reference microorganisms thus producing 120 *consensus* sequences (not reported here) which can be used to design 5'-UTR sequences (therefore containing RBS) in the microorganisms studied without any prior knowledge about the their translation control mechanism (e.g. anti-SD, optimal spacing between RBS and start codon, etc.). To support the design of RBS sequences that could work in several different organisms, two approaches can be followed: perform a multiple alignment starting from all the sequences upstream of the genes of all the reference microorganisms, or perform the multiple alignment of the *consensus* sequence previously computed for each microorganism indipendently. Since the first approach requires a large amount of computational time and preliminary results based on ≈

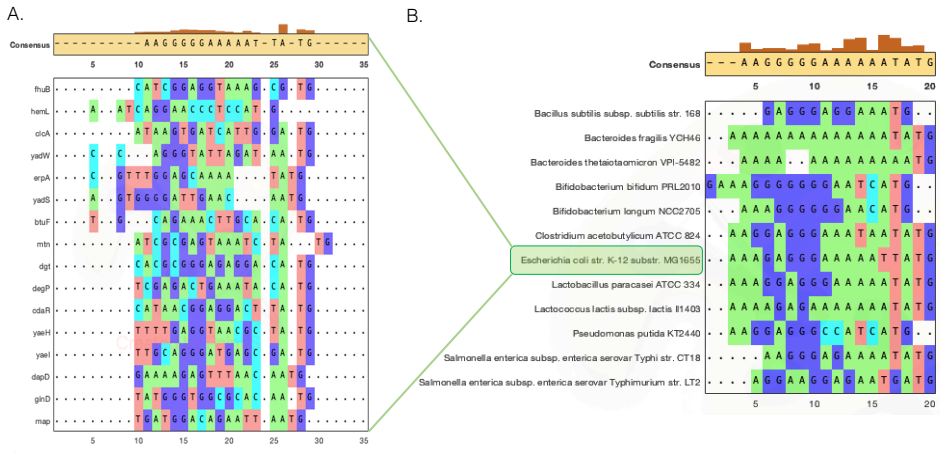

Figure 5.5: ***Consensus* sequence estimation through Clustal Omega multiple alignment.** **A.** Estimation of the *Escherichia coli consensus* sequence from 4357 18bp-region upstream of annotated genes **B.** Twelve previously estimated (example in figure for *E. coli* - green box) *consensus* sequences of therapeutic-notable probiotics have been aligned to obtain a new synthetic *consensus* sequence.

10 microorganisms produces an output sequence with indels (-) only, the second approach has been chosen. The analysis has been carried out on 12 bacteria known for their beneficial properties as probiotics which have been chosen on the basis of scientific interest for their use in the therapeutic field; the results have been reported in Fig. 5.5B. The sequence profile thus obtained has been used to generate new RBS sequences (reported in Table 5.4) that can be used to create new expression systems of translation in the selected bacterial species.

Table 5.4: **New synthetic RBS sequences obtained from *consensus* sequence profile obtained for the twelve probiotics selected.**

***Consensus* sequence** **AAGGGGGAAAAAAT**	
AGGAGGAAGGAAAT	AAGAAAGAAATCTT
AGGAGAGAAAACAT	AAGAAGACAATAAT
AAAGAGGAAAAAAT	AGGAAGGAAAAATT
AGGAGGGAAGACAT	AGGGAGGAAAACAT
AAGAAGGGGAAAAA	AGGGGAACAAAAAT
AAAAGGAACAAAAT	AAGGGAAAAATAAG
AAAAGGGGAAAAAT	AAAAGAGAAAAAAT
AAGAGGGGCAACAT	AGGAAGGAGAAAAT
AAAAAGGGAAGTAT	AAGGGAAGGAAAAT
AAAAGGGAGAACAG	AGAAAGGAAAAAAT
AAGAAGGCAGACAT	AAGAGGGAAAAAAT
AAAGAGGGGAAAAT	AGAGGGAAGTCAAT
AAAGAAGGAAAACT	AGAAGGGAAGTAAT
AGAGGGAAGAACAT	AGGAAGGAAAAATA
AAAAGAACAAAAAT	AAAGAAGGAAACTA

5.5 Discussion

The incresingly demand of expanding the libraries of regulatory parts (e.g., promoters and RBS) for non-model microorganisms motivated the development of novel strategies based on bioinformatics approaches for their *in silico* prediction as an alternative to the high-cost and time-consuming *in vivo* search and characterization.

In this chapter, two new bioinformatics pipelines have been developed with the aim to: detect constitutive endogenous promoters in microorganisms with annotated genomes based on transcriptomics data in different experimental conditions (172 and 96 samples for *E. coli* and *B. subtilis* have been collected, respectively, to test the pipeline), and to design new RBS sequences from the knowledge acquired on the *consensus* sequences computed via multiple alignment algorithm of the 18bp regions upstream each genes, via a process that does not need prior knowledge of the translation mechanism within the considered

microorganism.

The two procedures adopted for promoter sequence detection and for selection of constitutively expressed genes, which compose the first mentioned pipeline, have been tested and the results compared with genomics data (dataset of promoters obtained from the *EcoCyc* and *DBTBS* databases) and scientific literature articles on the two model microorganisms. The performances reached show a promising prediction capability (accuracy of promoters detection in genome = [51.5∼70.1, 66.1∼80.2]%, percentage of putative constitutive genes founded = [40, 85.71]% where ',' separate the two bacterial species, '∼' the values obtained from the two negative controls NC1 and NC2), furthermore the accuracy of the whole pipeline reached 57.14% (*E. coli*) and 66.66% (*B.subtilis*) values. Even if the latter values are high enought to guarantee a good stable and constitutive promoter detection, further optimization steps are needed to achieve better results. Indeed, the optimization to increase the generalization power of this bioinformatics pipeline is in progress. In particular, the process of automatic reading of data from DNA-microarray chips of different manufacturers is still on-going with the aim to increase the experimental data collected throughout multiple perturbations and thus increase the pipeline accuracy: 3612 total GSEs (306 have been used in this study) are present in the *GEO NCBI* database (filtering by Organism - 'Bacteria' and Study type - 'Expression profiling by array') covering 1315 unique GPLs (2 have been analyzed in this study) out of 972 unique microorganisms (526 for NGS) in order to discover good promoter candidates for non-model microorganisms as well. Aforementioned GSE and GPL counts reported are updated to July 2020. Actually, a new pipeline is under development which take into accounts the Transcription Start Sites (TSSs) prediction of the promoter from the NGS data, in fact, the single-base resolution of this technology [207] (gene-resolution for DNA-microarray technology) allows to estimate the TSSs of the promoters thus estimating with greater accuracy

their sequences (generally the promoter relies on the -35 bp upstream its TSS - in this study a 300 bp region has been considered to detect promoter sequences), overcoming also the limitation of the knowledge of *consensus* sequences recognized by sigma factor, generally not available for non-model microorganisms. In parallel, a new branch of the project will be open in order to use the part of information that has been discarded in the first part of this study: the subset of DE genes. In fact, from the latter is possible to infering the DNA operator sites linked to a particular perturbation; usefull in the synthetic circuit design step in order to increase the circuit orthogonality from the enzymatic host machinery. A new approach to design synthetic RBSs within one or more selected microorganisms based on their genomic information has been illustrated. The fully automatic procedure and the lacks of assumptions on the translation regulatory mechanisms increase the scalability and portabilty of the pipeline to more bacterial species. The synthetic RBS sequences obtained in Section 5.4 will be synthesized and tested in laboratory in order to engineer a set of bacteria of interest via *ad hoc* expression vectors. The experimental work will allows an appropriate validation of the developed pipeline. Taken together, the illustrated results and the numerosity of available public datasets pave the way to an expanded bioinformatics workflow to increase the toolkit of regulatory parts in non-model organisms and to design novel regulatory parts that could work in different hosts with high probability, without any prior knowledge on host regulation mechanisms.

Chapter 6

Overall conclusions

The low portability of synthetic circuits characterized in model microorganisms (e.g., *Escherichia coli*, *Bacillus subtilis*) has limited the applicative potential of synthetic biology in many sectors, such as, for example, agricultural, cosmetic, therapeutical, biofuel, opening up the need of designing a gene expression control system that can function in a stable, robust and predictable way in any *chassis*. In this work, two new control architectures, called Sad-iFFL and U-iFFL whose functionality rely in a novel repressor system based on *Staphylococcus aureus* dCas9 (SadCas9), have been developed to mitigate the sources of variability that affect a genetic circuit in the context of different bacterial species, in terms of the variation of transcription rate, translation rate and gene copy number. The bottom-up design of the circuits has been supported by *in silico* analyses of *ad hoc* defined mathematical models, whose accuracy is fundamental to support the design of complex synthetic circuits like those developed in this study.

Chapter 2 has been entirely dedicated to illustrate different modelling tools for synthetic circuits. Eight different mathematical models have been developed, focusing on the Lux-circuitry as a case study, to

show the impact of different model assumptions and deviations from the traditional Hill function models. In particular, regulatory protein abundance, ligand binding, and resource usage were made and the resulting models were compared to analyze the effects on the transfer function of the circuits.

Although it is hard to select a single model as the best tool to describe synthetic circuits, it has been shown that in some cases different assumptions lead to input-output characteristics that significantly deviate from traditional Hill function models and such effects may be important for proper description of synthetic circuit behavior, inference on biological phenomena and prediction of unseen circuits. Importantly, considering the one-step LuxR-P_{Lux} regulatory network, all the considered models have been shown to be identifiable, paving the way to an expanded toolkit of computational methods for circuit analyses. In specific, the simulations carried out on the case study circuit showed that all the relevant parameters of a Hill function may change depending on the underlying assumptions. The selection of the best model will have to be done depending on the experimental situation of interest and some of them have been adopted in the main work of this thesis.

The first designed circuit, Sad-iFFL, relies on the incoherent feed-forward-loop motif enabled by the SadCas9 repressor enzyme. The mathematical treatment has been fundamental to characterize the transfer function of the new SadCas9 repressor, never done in literature so far, to understand the working constraints of the Sad-iFFL circuit and to simulate its steady-state and dynamic characteristics, which were compared with an expression cassette circuit without control, called Open loop. It has been demonstrated that the robustness performances of Sad-iFFL are strongly dependent on the theoretical set-point value K_{CG} (an increase of the latter leads to raise the SadCas9 biological demand to work as a repressor) in agreement with the data collected *in vivo*, for which the Sad-iFFL circuit fails to regulate

the expression of the target gene low transcriptional activities by the input promoter (P_{Lux}), used to drive the circuit over a range of expression values, and the lower bound for HSL concentration was ≈ 1 nM. The *in vivo* results showed consistency with model simulations in terms of expression-dependent behavior and robustness against transcription variation. In addition, by testing five different ribosome binding sites (RBSs) upstream of the SadCas9 and GOI genes, it was possible to confirm circuit robustness in different translation rate contexts. In this case, the steady-state values were different across the used RBSs, since the translation initiation rates (TIRs) of the two genes were different (despite identical RBSs were used for SadCas9 and GOI), and the quantitative output values measured for the Sad-iFFL circuit variants were highly correlated with the expected ratio between SadCas9 and GOI, thereby demonstrating the predictability of the designed architectures.

Despite robust and predictable features have been demonstrated *in silico* and *in vivo* for Sad-iFFL, the perfect adaptation of the target gene has not been achieved for a range of transcriptional activities due to model hypothesis violation. For this reason, a novel architecture, called U-iFFL, has been developed in which a new network motif (positive autoregulation) was added to the incoherent feedforward-loop to implement a more robust circuitry. The transcriptional machinery taken from the phage domain (RNAPT7-P_{T7}) has been used for positive autoregulation and has been inserted to drive all the genes within the circuit, to increase the probability of guaranteering overabundance hypotheses of the circuit actuators. The model-simulated steady-state comparisons between Sad-iFFL and U-iFFL showed that the new architecture is more robust to transcription and translation rate variation in a wide range of theoretical set point values (K_{CG}). Furthermore, it has been demontrated that Sad-iFFL and U-iFFL have theoretically similar noise propagation properties, for both yielding lower cell-to-cell variability than the Open loop circuitry. The dynamic anal-

ysis demonstrates the importance of the genetic 'on-memory' device achieved from the positive autoregulation loop that overcomes the issue of selecting the right promoter sequence in new bacterial hosts. The positive autoregulation module composed by the transcriptional activation system of phage T7 has been optimized and successfully characterized *in vivo*, validating its suitability for U-iFFL in meeting the overabundances hypothesis of the mathematical model.

The promising results achieved by Sad-iFFL and U-iFFL rely on the availability of regulatory parts (e.g., promoter, RBS) that drive the genes inside the circuits in all the target hosts. Such regulatory parts do not need to have a specific quantitative activity (since the circuits are expected to adjust the output by design), but it is important that they are functional and are sufficient to express the circuit proteins at a minimum level, which is especially limiting and crucial the for Sad-iFFL circuit. To support the rational choice of promoters and RBSs, thereby increasing the probability to get a functional circuit, two bioinformatic pipelines based on publicly available high-throughput transcriptomic and genomic data have been developed. In particular, these methods are expected to: (i) enrich the library of regulatory parts in bacteria, thus expanding the range of strenghts available for promoters and RBSs, and (ii) increase the portability of genetic circuits in different non-model microorganisms. The first pipeline, adopted for promoter sequence and strength identification from high-throughput transcriptomics microarray and NGS data selected a set of genes among which $\approx 57.14\%$ and 66.66% are known constitutive promoters, considering two case studies (*Escherichia coli*, *Bacillus subtilis*, respectively). Although the pipeline efficiency is sufficient to help the discovery of new promoter sequences for the aforementioned scenario, further optimization steps are needed to achieve better results. The optimization of the pipeline is in progress in order to expand the availability of microarray data through the optimization of the data retrieval algorithm on different chip manufacturers,

thus including more trascriptomics data on different perturbation conditions on several microorganisms. Furthermore, a new procedure on NGS data is under development in order to increase the accuracy of promoter sequence identification by Transcription Start Site (TSS) estimation, possible with NGS data due to their single base resolution. The second pipeline, for identification of RBS *consensus* sequence in one or more microorganisms, has been provided a new tool to generate new synthetic RBSs that can ben used to create new expression system of translation in the target bacteria, without any knowledge about their translation control mechanism (e.g., anti-SD, optimal spacing between RBS and start codon, etc.). Several synthetic RBSs have been designed and will be *de novo* synthesized and tested in the probiotic organisms considered thus validating the here illustrated pipeline. Taken together, the developed pipelines are expected to support the design of new synthetic parts for non-model bacteria and increase the probability to get functional iFFL circuit architectures, using public data without any prior knowledge on the specific expression machinery of the hosts.

In conclusion, the work of this thesis aimed to increase the portability of new genetic circuit designs in non-model bacteria ensuring their predictability, stability and robustness through the contribution of different approaches, such as: automatic control, bioinformatics, mathematical analyses of dynamic systems and protein engineering. The results of the interconnection of all these disciplines can be summarized in the performances reached for the circuit models, here developed, Sad-iFFL and U-iFFL. In particular, the promising *in silico* behavior, the *in vivo* functioning of the new repression system based on the SadCas9 enzyme and the efficiency of the new optimized transcriptional activation system using RNAPT7-P_{T7} lays the foundations to the use of control circuits capable of adapting to different hosts, thereby expanding the potential of synthetic biology applications.

Appendix A

Supplementary Information for Chapter 2

A.1 Model parametrization

Models were parametrized using plausible values, according to available biological knowledge (e.g., DNA/protein concentrations) and previously published experimental data (e.g., P_{Lux} activation curve, resource usage [84, 88]). A summary of parameters is provided in Table A.1 for each model used in this work. When indicated, structural parameters were fixed to the reported values in simulations and they were assumed to be unknown during parameter estimation tasks (e.g., when studying a posteriori identifiability). *E. coli* cell volume was assumed to be $1\mu m^3$, corresponding to 10^{-15} L. Under this assumption, the concentration of one molecule of promoter DNA or protein corresponds to 1.66 nM [208]. A variation of the average *E. coli* cell volume has been previously reported in the range $0.5 - 2\mu m^3$ for different growth rates and conditions [208]. This variation is expected to affect the absolute values but not the trends of the numerical simulations shown of this work, thereby not affecting the drawn conclusions. Nonetheless, the variation of cell volume will be highly relevant in the

analysis of real data, in which model parameters have to be estimated based on the knowledge of this value.

Table A.1: **Model parameters.** The indicated values were used for simulation and, unless differently indicated, for the study of a posteriori identifiability. [a]During the study of a posteriori identifiability, n_2 was expressed as AU with known value, while r_T was expressed as nM/AU with unknown value; [b]When indicated, a range of copy number values was spanned; [c]Typical value of *E. coli* growth rate, which is used as the rate of intracellular protein dilution; [d]A value of 2 was used to simulate the presence of external load where indicated.

Parameter	Units	Models	Value
P_U	nM	All	1.66
r_T	nM^a	All	1.66
n_1	–	All	5^b
n_2	$-^a$	All	301^b
γ_X	min^{-1}	All	0.01^c
σ	min^{-1}	All	0.0167
$\widehat{k}_{m0}$	$AUnM^{-1}min^{-1}$	All	0.0192
$\widehat{k}_{mL}$	$AUnM^{-1}min^{-1}$	All	1.907
K_1	$nM{-}1$	M1, M1L, M2, M2L	0.001
		M1T, M1TL, M2T, M2TL	0.0055
K_3	nM^{-1}	M1, M1L, M2, M2L	0.2
K_4	nM^{-1}	M1L, M2L, M1TL, M2TL	0.001375
K_5	nM^{-1}	M1L, M2L, M1TL, M2TL	0.2
J_{RFP}	$minAU^{-1}$	M1L, M2L, M1TL, M2TL	0.04
E	–	M1L, M2L, M1TL, M2TL	0^d

As mentioned in Section 2.2.3, to enable the application of the models in popular situations occurring in synthetic biology (i.e., model identification with the data routinely measured in fluorescent reporter protein-based assays) promoter copy number (n_1) was assumed to be available; on the other hand, the actual LuxR protein concentration (dependent from luxR DNA copy number, transcription rate, translation rate, dimerization, and mRNA/protein degradation) is harder to

measure. While in simulation the R_{2T} quantity was spanned to explore the effects of having wide ranges of LuxR values, its value was assumed to be unavailable during model identification. Nonetheless, the relative strength of the promoter expressing the luxR gene is commonly known, thereby enabling to approximate the relative level of LuxR. For model amenability reasons, in estimation steps we parametrized n_2 as the (known) relative strength of this promoter (in AU) and r_T as the (unknown) scale factor between protein concentration (in nM) and AU, in which all the biologically occurring processes, described above, are lumped without any loss of generality and maintaining the mechanistic nature of the model. In M1, the r_T quantity always appears multiplied by K_3 (Equation (2.23)) and, for this reason, only their product is identifiable. By defining $\widehat{K_3} = K_3 \cdot r_T$ (in AU^{-1}), the M1 model can be re-parametrized as Equation (A.1)

$$y = n_1 \cdot P_U \cdot \widehat{k}_{m0} + \frac{n_1 \cdot P_U \cdot (\widehat{k}_{m0} + \widehat{k}_{mL})/(1 + 1/(\widehat{K}_3 \cdot n_2))}{1 + 1/(K_1 \cdot (1 + \widehat{K}_3 \cdot n_2) \cdot L)} \tag{A.1}$$

Analogously, the $K_5 \cdot r_T$ product in M1T (Equation (2.37)) can be re-parametrized as $\widehat{K}_5 = K_5 \cdot R_T$ (in AU^{-1}), thereby yielding the model in Equation (A.2)

$$y = n_1 \cdot P_U \cdot \frac{\widehat{k}_{m0} + (\widehat{k}_{m0} \cdot K_1) \cdot L + (\widehat{k}_{m0} \cdot K_1^2/4 + \widehat{k}_{mL} \cdot K_1^2/4 \cdot \widehat{K}_5 \cdot n_2) \cdot L^2}{1 + K_1 \cdot L + (K : 1^2/4 + K_1^2/4 \cdot \widehat{K}_5 \cdot n_2) \cdot L^2} \tag{A.2}$$

A.2 *A priori* identifiability

Given a circuit output expression with known form (e.g., a Michaelis –Menten or a rational function) with specific coefficients that could be

estimated from data, an expression for each parameter of the model was found as a function of the theoretically known coefficients. If the system can be solved uniquely, the model was considered as structurally (or a priori) identifiable.

A.3 *A posteriori* identifiability

Given a model, simulated data were generated and parameters were estimated. Different experiments were simulated varying n_2 as required, while the other structural parameters were kept constant to the values in Table A.1, unless differently indicated. Proportional Gaussian random noise was added to the simulated data with a coefficient of variation (CV) of 5% as default value; other values were also evaluated (0, 1, 2.5, and 110%). A limited number of data points (12) was assumed to be available, resembling a realistic dose–response experimental setup. Parameter estimation was performed via least squares method with the MATLAB R2017b (MathWorks, Natick, MA, USA) *lsqnonlin* routine. If the estimated parameters are consistent with the ones used to generate the data, the model is considered as practically (or a posteriori) identifiable. For each proportional error entity, 200 simulation and estimation steps were carried out, thereby identifying the model using different synthetic data with random noise. Relative estimation error ($REE = 100 \cdot |p_{est} - p_{true}|/p_{true}$), where p_{est} and p_{true} are the estimated and true parameter values, respectively) was used to express parameter consistency. Uncertainty of parameter estimates, in terms of CV, was also computed as reported previously [208]. For each run, the maximum REE and CV among all the estimated parameters was considered.

A.4 Simulations

The MATLAB roots function was used to find P and R_2 as polynomial roots to solve the M2, M2T, and M2L model equations. Implicit equations, commonly occurring in models including cell load terms, were solved using the fixed-point method, as previously described [84].

A.5 Analysis of activation curves

Hill function parameters, described in Equation (2.5), were calculated for each activation curve: δ and α were computed as the y values at lowest and highest L value, respectively; κ was computed as the value of L corresponding to half-maximum activation; η was computed according to Equation (A.3).

$$\eta = \frac{log(81)}{log(L90/L10)} \tag{A.3}$$

where $L90$ and $L10$ are the L values corresponding to 90% and 10% of the maximum value of y. On the other hand, *in vivo* measured activation curves were fitted with Equation (2.5) to estimate its parameters, as described in Section 2.2.4.

A.6 *In vivo* experiments

Circuit output, i.e., RFP synthesis rate per cell (S_{cell}) at steady-state, was measured for recombinant MG1655-Z1 strain [88] bearing the low-copy plasmid pSB4C5 with X3r as insert [84]. In this construct, previously described in [84] and with sequence available as BBa_J107032 code in the Registry of Standard Biological Parts (http://parts.igem.org), *luxR* is under the control of an anhydrotetracycline (ATc) inducible promoter, which works as gene expression knob

in *Escherichia coli* strains overexpressing TetR, like the one used in this study. The detailed experimental protocol for S_{cell} measurement was previously described [84]. Briefly, 0.5 ml of M9 medium (11.28 g/L M9 salts, 1 mM thiamine hydrochloride, 2 mM MgSO4, 0.1 mM CaCl2, 0.2% casamino acids and 0.4% glycerol) were inoculated with a colony from a freshly streaked LB agar plate in a 2-ml tube and the culture was incubated at 37°C, 220 rpm for at least 16 h. The grown culture was 100-fold diluted in 200 μL of M9 in a 96-well microplate. ATc and HSL (2 μL) were added to the microplate wells to reach the desired concentrations. The microplate was incubated at 37°C in the Infinite F200 (Tecan) microplate reader and an automatic measurement procedure was programmed via the i-control software v.2.0.10 (Tecan, Switzerland): shaking (15 s, 3-mm amplitude), wait (5 s), optical density (600 nm) acquisition, red fluorescence (excitation: 535 nm, emission: 620 nm, gain = 50) acquisition, sampling time = 5 min. Raw data time series were background-subtracted using sterile medium (absorbance) and a non-fluorescent culture (fluorescence), incubated in the same experiment. The resulting data were used to compute S_{cell} as the numeric time derivative of fluorescence, divided by absorbance over time. S_{cell} was averaged in the exponential growth phase, typically occurring at absorbance values between 0.05 and 0.18 [209]. Finally, the average S_{cell} values of the X3r strains at the desired inductions were normalized by the Scell value of a reference culture (MG1655-Z1 bearing the BBa_J107029 constitutive RFP expression cassette in the pSB4C5 plasmid) as internal control to obtain S_{cell} in standardized relative units.

The resulting dose-response curves were fitted using the MATLAB *lsqnonlin* routine to estimate Hill function parameters. For each HSL and ATc condition, at least three independent experiments were carried out.

Appendix B

Supplementary information: Open loop model as control for iFFL-based controllers

In this appendix the Open loop genetic circuit that has been used in this study as a term of comparison for the synthetic circuit architectures, Sad-iFFL and U-iFFL, is reported. It is based on the unregulated expression of a gene of interest (GOI). The biological circuit description and the mathematical model equations are reported in the following Sections, while the *in vivo* circuits realizations are listed in Table 3.2, represented with an OL-suffix at the end of the construct name.

B.1 Circuit Description

The genetic circuit includes no feedback control, and it is composed of a single DNA expression cassette for a gene of interest (GOI) under the control of a constitutive (or inducible) promoter. The main ex-

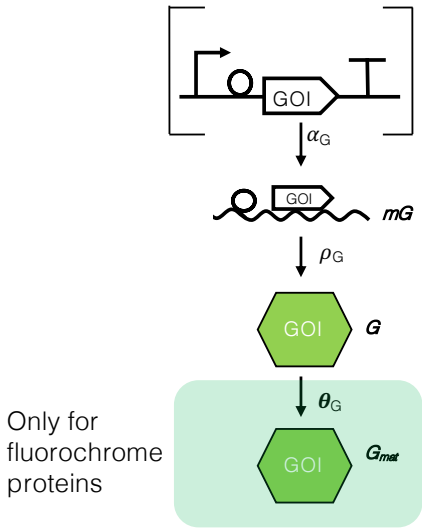

Figure B.1: **Open loop (OL) biological model.** Visual representation of gene expression. The curved arrow, the circles and the T-shaped lines are respective the promoter, RBS and terminator parts. The horizontal line inside square blakets represents the DNA while the wavy one the mRNA. The production rates are represented with monodirectional arrows (e.g., transcription rate α_G, translation rate δ_G, maturation rate θ_G). The biological scheme lacks of the degradation constants (e.g., mRNA degradarion rate, protein degradation rate), which are considered in the associated mathematical model in Section B.1), due to graphical reasons.

pression processes occurring in this circuit are summarized in Fig. B.1: transcription α_G (regulated by the promoter) and translation ρ_G (regulated by the RBS) and, if the protein is a fluorophore (e.g., Red Fluorescent Protein - RFP), protein maturation θ_G. This circuit represents a control configuration that is expected to provide different protein expression levels as a function of promoter, RBS, plasmid copy number and host strain.

B.2 Mathematical model

In Equation (B.1) the ordinary differential equation describing the time-dependent evolution of a generic protein G is reported, while in Equation (B.2) the same equation has been represented but for the case which G is a fluorophore and, therefore, its maturation step has been also modeled.

$$\frac{dG}{dt} = \frac{n \cdot \alpha_G \cdot \rho_G}{(d_{mG} + \mu)} - (d_G + \mu) \cdot G \tag{B.1}$$

$$\frac{dG_{mat}}{dt} = \theta_G \cdot \frac{n \cdot \alpha_G \cdot \rho_G}{(d_{mG} + \mu) \cdot (d_G + \theta_G + \mu)} - (d_G + \mu) \cdot G \tag{B.2}$$

In Equations (B.1) - (B.2), n, α_G, ρ_G are the copy number $[nM]$, transcriptional rate $[time^{-1}]$ and translation rate $[time^{-1}]$, respectively, while d_{mG}, d_G and μ are the mRNA degradation rate $[time^{-1}]$ for the mG transcript, protein degradation rate $[time^{-1}]$ for protein G and the diluition rate $[time^{-1}]$ due to cell division. In Equation (B.2), θ_G $[time^{-1}]$ is the maturation rate from G to its mature form, G_{mat}. The Equations (B.1) represent the ordinary differential equation (ode) used to study the dynamic behavior of the Open loop circuit *in silico* and its performances has been compared with the Sad-iFFL and U-iFFL circuits in Section 3.3.4 and Section 4.2.3.

The steady−state representation of the model has been obtained by evaluating Equation (B.1) and Equation (B.2) at their equilibrium, i.e. for $dG/dt = 0$, and, considering the protein degradation rate negligible compared with cell diluition rate $\mu >> d_G$ and the latter negligible compared with the mRNA degradation rate, $d_{mG} >> \mu$.

$$G^{SS} = \frac{n \cdot \alpha_G \cdot \rho_G}{d_{mG} \cdot \mu} \tag{B.3}$$

$$G_{mat}^{SS} = \frac{\theta_G}{(\theta_G + \mu) \cdot \mu} \cdot \frac{n \cdot \alpha_G \cdot \rho_G}{d_{mG}} \tag{B.4}$$

Equation (B.3) has been used to evaluate the robustness of Open loop model on parameter variations and to analyze the propagation of biological noise on α and ρ parameters in Section 3.3.4 and Scetion 4.2.3. Equation (B.3) has been used to describe mature RFP formation in the mathematical model used for the *in silico* SadCas9 characterization in Section D.2.

Appendix C

Supplementary information: wet lab protocols and data analysis

C.1 Materials and reagents

C.1.1 Inducers

Inducers are small molecules that regulate gene expression in two ways: repressing or activating the transcriptional activity of the promoter upstream the target gene. In this study two inducer molecules, HSL and IPTG, have been used that indirectly activate the Lux- and Lac-circuitries, respectively.

- **N-(3-oxohexanoyl)-L-homoserine lactone (HSL)**: the binding between HSL molecule and LuxR protein forms a complex responsible for the activation of the transcriptional activity of P_{Lux} promoter. The HSL molecule has been purchased from Sigma Aldrich (K3007), dissolved in deionized water at final concentration of $200mM$ and conserved at -20°C. The small size of this molecole allows to freely permeate through bacterial membrane.

- **Isopropyl β-D-thiogalactopyranoside (IPTG)**: the binding between the IPTG molecule and the LacI protein inactivates LacI, which is not able to bind DNA anymore. IPTG molecule has been bought from Sigma Aldrich (I1284) in a ready made solution at a concentration of $2mM$ and conserved at -20°C. IPTG molecule is a synthetic, structural analogue, of allolactose (also inducing the of Lac-circuitry) which, differently from the latter, is not metabolized thus providing a constant concentration of IPTG within the cell.

C.1.2 Antibiotics

Antibiotics are molecules with antimicrobial properties which are used in synthetic biology to select a bacterial population bearing a plasmid which the resistance gene is encoded, in order to selectively kill all the non-engineered bacteria. In this study, high-, medium- and low-copy plasmids have the ampicillin, kanamycin and chloramphenicol resistance, respectively:

- **Chloramphenicol (Cm)**: molecule that inhibits the protein synthesis, blocking the translation process of mRNAs. Chloramphenicol has been used as a marker for the selection of bacteria bearing the low-copy number plasmid vector pSB4C5. This molecule is conserved at -20°C at a concentration of $34mg/ml$; the working concentration is $12.5\mu g/ml$.

- **Kanamycin (Kan)**: molecule that inhibits the protein sythesis. Kanamycin has been used as a marker for the selection of bacteria bearing the medium-copy number plasmid vector pSB3K3. This molecule is conserved at -20°C in a concentration of $50mg/ml$; the working concentration is $25\mu g/ml$.

- **Ampicillin (Amp)**: molecule that prevents the formation of

plasma membrane of new forming bacteria. Ampicillin has been used as a marker for the selection of bacteria bearing the high-copy number plasmid vector pSB1A2. This molecule is conserved at -20°C at a concentration of $100mg/ml$; the working concentration is $100\mu g/ml$.

C.2 Cloning

The *E. coli* TOP10 strain (Invitrogen) has been used for the *in vivo* amplification of plasmids. TOP10 was transformed by heat-shock according manufacturer's instructions and transformed bacteria were grown in L-broth (LB: sodium chloride 10 g/l, tryptone 10 g/l, yeast extract 5g/l) at 37°C. Glycerol stocks have been prepared with 750 μl of saturated culture with proper antibiotics and 250 μl of glycerol 80%, and stored at -80°C. All the circuits used in this study were assembled from existing plasmids avaliable in the MIT Registry of Standard Biological Parts (reported in Table C.1) according to the BioBrick™ Standard Assembly procedure or from standardized parts taken from another plasmid (*sadcas9* - Section C.4.1) or bacterial genome (*rnapt7* - Section C.4.2) and a number of standard molecular biology methods: plasmids were extracted from saturated 5 ml cultures (grown in LB at 37°C, 220rpm) through the NucleoSpin Plasmid kit (Macherey-Nagel); DNA was digested as appropriate, with the EcoRI/XbaI/SpeI/PstI enzymes, and the fragments of interest were extracted from 1% agarose gel by NucleoSpin PCR cleanup and gel extraction kit (Macherey-Nagel) before proceeding with ligation. As a result, each part used in this work is compliant to the BioBrick™ Standard. Consequently, every junction between assembled parts has the TACTAG sequence if the downstream part is a coding sequence (only exception for synthetic the new RBSs designed in this work and reported in Section C.6), otherwise the sequence is TACTAGAG. All the DNA-modifying

enzymes were purchased from Thermo Fisher Scientific. The DNA of all the constructed parts was screened via diagnostic restriction digest/electrophoresis, and was sequence-verified via the Eurofins Genomics Germany GmbH DNA analysis service (Ebersberg, Germany). All single-plasmid construct that have been obtained in this study are reported in Table C.2.

Table C.2: **Synthetic constructs obtained in this study.** For each single-plasmid synthetic constructs is reported the description of the single parts used to assemble it is reported. The description of the final circuits is outlined.

Name	**Construct**	**Purpose**
SadCas9	new part	Dead-endonuclease Cas9 from *S. aureus* mutagenized to allow BioBrick^{TM} assembly taken from Addgene plasmid #113718.
sgRNA	new part	single-guide RNA with terminator (taken from Addgene plasmid #44251) designed for SadCas9 compatibility.
AE-3A31SadCas9	AE-3A + BBa_B0031 + SadCas9 + BBa_B0015 + pSB4C5	LuxR-Inducible SadCas9 expression under B0031 RBS
AE-3A32SadCas9	AE-3A + BBa_B0031 + SadCas9 + BBa_B0015 + pSB4C5	LuxR-Inducible SadCas9 expression under B0032 RBS
AE-3A34SadCas9	AE-3A + BBa_B0034 + SadCas9 + BBa_B0015 + pSB4C5	LuxR-Inducible SadCas9 expression under B0034 RBS
AE-3ACU1SadCas9	AE-3A + CU1 + SadCas9 + BBa_B0015 + pSB4C5	LuxR-Inducible SadCas9 expression under CU1 RBS
AE-3ACU2SadCas9	AE-3A + CU2 + SadCas9 + BBa_B0015 + pSB4C5	LuxR-Inducible SadCas9 expression under CU2 RBS
AE-3ACA1SadCas9	AE-3A + CA1 + SadCas9 + BBa_B0015 + pSB4C5	LuxR-Inducible SadCas9 expression under CA1 RBS
Plac_sgRNA J119SCTarget34RFP	BBa_R0011 + sgRNA + BBa_J23119 + SCTarget + BBa_B0034 + BBa_E1010 + BBa_0015 + pSB3K3	LacI-inducible sgRNA expression and constitutive RFP cassette expression bearing the target region for SadCas9 repression
Plac_sgRNA J119SCTarget31RFP	BBa_R0011 + sgRNA + BBa_J23119 + SCTarget + BBa_B0031 + BBa_E1010 + BBa_0015 + pSB3K3	LacI-inducible sgRNA expression and constitutive RFP expression cassette bearing the target region for SadCas9 repression

Continues on next page...

Table C.2 – *...continued from previous page*

Name	Construct	Purpose
Plac_sgRNA J118SCTarget34RFP	BBa_R0011 + sgRNA + BBa_J23118 + SCTarget + BBa_B0034 + BBa_E1010 + BBa_B0015 + pSB3K3	LacI-inducible sgRNA expression and constitutive RFP cassette expression bearing the target region for SadCas9 repression
P_{Lux}SCTarget31RFP	BBa_R0062 + SCTarget + BBa_B0031 + BBa_E1010 + BBa_B0015 + pSB3K3	LuxR-inducible RFP expression cassette under B0031 RBS bearing the target region for SadCas9 repression
P_{Lux}SCTarget32RFP	BBa_R0062 + SCTarget + BBa_B0032 + BBa_E1010 + BBa_B0015 + pSB3K3	LuxR-inducible RFP expression cassette under B0032 RBS bearing the target region for SadCas9 repression
P_{Lux}SCTarget34RFP	BBa_R0062 + SCTarget + BBa_B0034 + BBa_E1010 + BBa_B0015 + pSB3K3	LuxR-inducible RFP expression cassette under B0034 RBS bearing the target region for SadCas9 repression
P_{Lux}SCTargetCU1RFP	BBa_R0062 + SCTarget + CU1 + BBa_E1010 + BBa_B0015 + pSB3K3	LuxR-inducible RFP expression cassette under CU1 RBS bearing the target region for SadCas9 repression
P_{Lux}SCTargetCU2RFP	BBa_R0062 + SCTarget + CU2 + BBa_E1010 + BBa_B0015 + pSB3K3	LuxR-inducible RFP expression cassette under CU2 RBS bearing the target region for SadCas9 repression
P_{Lux}SCTargetCA1RFP	BBa_R0062 + SCTarget + CA1 + BBa_E1010 + BBa_B0015 + pSB3K3	LuxR-inducible RFP expression cassette under CA1 RBS bearing the target region for SadCas9 repression
J119sgRNA	BBa_J23119 + sgRNA + pSB1A2	Constitutive expression of sgRNA cassette under J119 promoter
AE-3A31T7(R632S)	AE-3A + BBa_B0031 + RNAPT7(R632S) + BBa_B0015 + pSB4C5	LuxR-Inducible RNAPT7 expression under B0031 RBS
AE-3A32T7(R632S)	AE-3A + BBa_B0032 + RNAPT7(R632S) + BBa_B0015 + pSB4C5	LuxR-Inducible RNAPT7 expression under B0031 RBS
AE-3A34T7(R632S)	AE-3A + BBa_B0034 + RNAPT7(R632S) + BBa_B0015 + pSB4C5	LuxR-Inducible RNAPT7 expression under B0034 RBS
P_{T7}SCTarget31RFP	BBa_I719005 + SCTarget + BBa_B0031 + BBa_E1010 + BBa_B0015 + pSB3K3	RNAPT7-Inducible RFP expression cassette under B0031 RBS
P_{T7}SCTarget32RFP	BBa_I719005 + SCTarget + BBa_B0032 + BBa_E1010 + BBa_B0015 + pSB3K3	RNAPT7-Inducible RFP expression cassette under B0032 RBS

Continues on next page...

Table C.2 – *...continued from previous page*

Name	Construct	Purpose
P_{T7}SCTarget34RFP	BBa_I719005 + SCTarget + BBa_B0034 + BBa_E1010 + BBa_B0015 + pSB3K3	RNAPT7-Inducible RFP expression cassette under B0034 RBS

C.2.1 Mutagenesis

Synthetic circuits modifications have been performed through mutagenesis of two types: convergent or divergent. Divergent primers have been used to carry out single point-mutations or to delete DNA portion (e.g., TACTAGAG-scar between the promoter transcription start site and the single guide RNA after assembly procedure with BioBrick^{TM} enzymes, delete a whole promoter sequence). Convergent primers have been used to amplify DNA sequences and, when specified, add a DNA portion (e.g., RBS) to them. The convergent primers have, on their tail, a complete (EcoRI, XbaI) or partial (XbaI) BioBrick^{TM} prefix on the forward primer and a complete (SpeI, PstI) or null suffix (if the DNA template includes a BioBrick-standardized - VR reverse primer to enable the assembly procedure of the PCR-amplified DNA sequence. The experimental protocol includes:

- DNA amplification: template plasmid DNA has been purified through the NucleoSpin Plasmid kit (Macherey-Nagel) in case it was a plasmid in a bacterial culture, otherwise, in case of amplification from genomic template, a bacterial colony has been suspended in 100 μl of deionyzed water and 1 μl has been used for the PCR protocol. Phusion Hot Start Flex II (ThermoFisher Scientific) has been used according to manufacturer protocol and using primer pairs added (reported in Table C.3), for which annealing temperatures have been estimated on the free online tool T_m Calculator (ThermoFisher Scientific) with parameters set to 'Phusion or Phire DNA polymerase'.

Table C.1: **BioBrick™ parts and constructs used in this study for circuits assembly.**

Name	BioBrick Code	Description
P_{Lux}	BBa_R0062	LuxR positive-inducible promoter
P_R	BBa_R0051	Constitutive promoter from lamba bacteriophage for LuxR expression
P_{Lac}	BBa_R0011	IPTG-inducible promoter
J23100	BBa_J23100	Constitutive promoter of the Monitor cassette
J23101	BBa_J23101	Reference constitutive promoter
J23118	BBa_J23119	Medium-strength constitutive promoter
J23119	BBa_J23119	Strong constitutive promoter
P_{T7}	BBa_I719005	Wild type promoter of T7 bacteriophage
B0030	BBa_B0030	Strong RBS of LuxR
B0031	BBa_B0031	Weak RBS
B0032	BBa_B0032	Medium-strength RBS
B0034	BBa_B0034	Strong RBS
B0015	BBa_B0015	Double transcriptional terminator
LuxR	BBa_C0062	LuxR coding sequence
mRFP1	BBa_E1010	RFP coding sequence
GFPmut3b	BBa_E1040	GFP coding sequence
AE-3A	BBa_J23100 + B0032 + E0040 + B0015 + BBa_R0051 + BBa_B0030 + BBa_C0062 + BBa_B1006 + BBa_J107202 + pSB4C5	Optimized LuxR-inducible ex- pression cassette characterization
sadcas9	Addgene plasmid #113718	Bacterial vector for expression of Snap-tagged *Staphylococcus aureus* dCas9

- Template DNA digestion: after PCR cycles, the DpnI (Roche) enzyme has been added in order to digest the methylated template DNA by cutting a commonly occurring sequence.
- DNA ligation: PCR product has been separated in 1% agarose gel and the selected DNA band extracted by NucleoSpin Extract II kit (Macherey-Nagel). The blunt-end DNA fragment obtained after PCR has been phosphorylated by T4 Polynucleotide Kinase (PNK - ThermoFisher Scientific) and ligated with T4 ligase (Roche).

Table C.3: **Primers used in this study.** For each template reported in the first coloumn the primer pairs used to obtain the final product are shown; the purpose of each DNA manipulation is outlined, together with the final product obtained.

Template	Primer Pairs	Sequence (5' - 3')	Aim	Product
Addgene plasmid #113718	FW_BB_SadCas9	gtgcttctagagcggttaccacgttgtaaaggaacaacagaatggcttcctccgaagacgt	Conversion of SadCas9 in BioBrick™ format	SadCas9wEXsites
	RV_BB_SadCas9	gtgcttctagagcggttaccacgttgtaaaggaacaacagaatggcttcctccgaagacgt		
SadCas9wEXsites	FW_DeleteEcoRI	aactctaaagatgcacaaaaaatg	Delete EcoRI site in SadCas9 CDS	SadCas9wXsite
	RV_DeleteEcoRI	cttttctctcgcaagttcaa		
SadCas9wXsite	FW_DeleteXbaI	ctcgaagatttacttaatatccatt	Delete XbaI site in SadCas9 CDS	SadCas9
	RV_DeleteXbaI	agggattgcttctaacgagt		
SadCas9	FW_B0031SadCas9	ccgcttctagagtcacacaggaaacctactagatgaaaaggaattatatc	Insertion of BBa_B0031 RBS in the upstream region of SadCas9	31SadCas9
	VR	attaccgcctttgagtgagc		
SadCas9	FW_B0032SadCas9	cgcttctagagtcacacaggaaagtactagatgaaaaggaattatatcttaggattagc	Insertion of BBa_B0032 RBS in the upstream region of SadCas9	32SadCas9
	VR	attaccgcctttgagtgagc		
SadCas9	FW_B0034SadCas9	cgcttctagagaaagaggagaaatactagatgaaaaggaattatatcttaggattagc	Insertion of BBa_B0034 RBS in the upstream region of SadCas9	34SadCas9
	VR	attaccgcctttgagtgagc		
SadCas9	FW_CU1SadCas9	gtgcttctagagccataaaaacttgacactagggtcaaaatatgaaaaggaattatatcttaggattagc	Insertion of CU1 RBS in the upstream region of SadCas9	CU1SadCas9

Continues on next page...

Table C.3 – *...continued from previous page*

Template	Primer Pairs	Sequence (5' - 3')	Aim	Product
	VR	attaccgcctttgagtgagc		
SadCas9	FW_CU2SadCas9	gtgcttctagagtatttaaaaggaaaacatcaaaggcactatgaaaaggaattatatcttaggattagc	Insertion of CU2 RBS in the upstream region of SadCas9	CU2SadCas9
	VR	attaccgcctttgagtgagc		
SadCas9	FW_CA1SadCas9	gtgcttctagagcggttaccacgttgtaaaggaacaacagaatgaaaaggaattatatcttaggattagc	Insertion of CA1 RBS in the upstream region of SadCas9	CA1SadCas9
	VR	attaccgcctttgagtgagc		
J11934RFP	FW_SCTarget11934	caatatggtcttgatcctactagagaaagaggagaaat	Insertion of the Target region for the complex SadCas9:sgRNA between J119 and B0034	J119SCTarget34RFP
	RV_SCTargetJ11934	aaccatgagttagctagcattatacctaggactg		
J119SCTarget34RFP	FW_Suffix	tactagtagcggccgctgcag	Deletion of B0034 RFP part	J119SCTarget
	RV_SCTarget	ggatcaagaccatattgaaccatgagtta		
118119SCTarget34RFP	FW_SCTarget	taactcatggttcaatatggtcttg	Deletion of B0034 RFP part	J118SCTarget
	RV_J118	gctagcacaatacctaggactgag		
P_{Lux}119SCTarget34RFP	FW_SCTarget	taactcatggttcaatatggtcttg	Deletion of B0034 RFP part	P_{Lux}SCTarget
	RV_Plux-3A	attcgactataacaaaccattttcttgcgtaaacctgtac		
J119scarsgRNA	FW_sgRNA	ggatcaagaccatattgaaccgtt	Deletion of the BioBrick™ assembly-derivate scar between J119 and sgRNA	J119sgRNA
	RV_J119	gctagcattatacctaggactgagctagct		
P_{Lac}scarsgRNA	FW_sgRNA	ggatcaagaccatattgaaccgtt	Deletion of the BioBrick™ assembly-derivate scar between P_{Lac} and sgRNA	P_{Lac}sgRNA
	RV_Plac	gtgctcagtatcttgttatccgc		
BBa_E1010	FW_CU1RFP	gtgcttctagagccataaaaacttgacactagggtcaaaaatatggcttcctccgaagacgt	Insertion of CU1 RBS in the upstream region of RFP	CU1RFP
	VR	attaccgcctttgagtgagc		
BBa_E1010	FW_CU2RFP	gtgcttctagagtatttaaaaggaaaacatcaaaggcactatggcttcctccgaagacgt	Insertion of CU2 RBS in the upstream region of RFP	CU2RFP
	VR	attaccgcctttgagtgagc		

Continues on next page...

Table C.3 – *...continued from previous page*

Template	Primer Pairs	Sequence (5' - 3')	Aim	Product
BBa_E1010	FW_CA1RFP	gtgcttctagagcggttac cacgttgtaaaggaacaa cagaatggcttcctccga agacgt	Insertion of CA1 RBS in the upstream region of RFP	CA1RFP
	VR	attaccgcctttgagtgag c		
BL21(DE3)	FW_B0034T7	gtgcttctagagaaagag gagaaatactagatgaac acgattaacatcgctaag	Amplification of RNAPT7 from *E. coli BL21(DE3)* genome in BioBrick^{TM} format under B0034 RBS	34T7
	RV_BB_RNAPT7	gtttcttcctgcagcggcc gctactagtttattacgcg aacgcgaagtcc		
34T7	FW_B0031T7	gcttctagagtcacacag gaaacctactagatgaac acgattaacatcgc	RBS changing from B0034 to B0031	31T7
	VR	attaccgcctttgagtgag c		
34T7	FW_B0032T7	gcttctagagtcacacag gaaagtactagatgaaca cgattaacatcgc	RBS changing from B0034 to B0032	32T7
	VR	attaccgcctttgagtgag c		
31T7	FW_T7(R632S)	gactaagagttcagtcat gacgctg	R632S mutation in 31T7 CDS	31T7(R632S)
	RV_T7(R632S)	acactgcgagtaacaccg taagcc		
32T7	FW_T7(R632S)	gactaagagttcagtcat gacgctg	R632S mutation in 32T7 CDS	32T7(R632S)
	RV_T7(R632S)	acactgcgagtaacaccg taagcc		
34T7	FW_T7(R632S)	gactaagagttcagtcat gacgctg	R632S mutation in 34T7 CDS	34T7(R632S)
	RV_T7(R632S)	acactgcgagtaacaccg taagcc		
P_{T7}119SCTarget34RFP	FW_SCTarget	taactcatggttcaatatg gtcttg	Deletion of B0034 RFP part	P_{T7}SCTarget
	RV_PT7	tctccctatagtgagtcgt attactctag		

C.2.2 Amplification and BioBrick^{TM}-standardization of biological elements

Biological elements not available as Registry parts have been PCR-amplified to convert them into BioBrick^{TM} standard format.

- *sadcas9* gene has been taken from Addgene plasmid #113718 used in the study of Savic et al. (Schwank Lab) which has the two single-point mutations (D10A and N580A) that convert the

Cas9 protein to its dead version [210]. The *sadcas9* gene has been inserted in pSB1A2 to sequently cloning procedures.

- *rnapt7* gene has been taken from *E. coli* BL21(DE3) strain. BL21(DE3) is an optimized bacterial strain used in labotatory for high level protein expression using IPTG-inducible T7 RNA polymerase system [211]. The *rnapt7* gene amplification has been coupled with insertion in its upstream region of BBa_B0034 RBS and sequently integrated into pSB1A2 vector.

The primer that have been used are reported in Table C.3.

C.3 sgRNA design

The single guide RNA (sgRNA) has been designed in the free online tool Benchling (https://benchling.com) and *de novo* synthesized (GenScript Biotech). The whole sgRNA sequence is composed of three parts: base pairing region, dCas9 handle and terminator sequence. The base pairing region has been designed manually by avoiding sequences that could interfere with synthetic circuits funcionality or assemby procedures (e.g., BioBrick^{TM} restriction sites, hairpins, -35 or -10 promoter *consensus* sequence, or sequences resembling the RBS *consensus* of *E. coli*). The target region, which has been derived by base pairing complementarity, has been coupled with the PAM sequence recognized from *Staphylococcus aureus* dCas9 protein NNGRRTN (in this study the TTGAGTA sequence was used) [74]. The dCas9 handle sequence has been designed based on the study of Ran et al. [74] and the terminator sequence has been taken from *Streptococcus pyogenes*, already used in our laboratory to terminate the transcription of sgRNA compatible with SpydCas9. All sequences have been reported in Table C.4.

Table C.4: **sgRNA components and their relative sequence.**

Component	Sequence
Base pairing region	GGATCAAGACCATATTGAACC
	GTTTTAGTACTCTGGAAACAGAATCTACTAA
S. aureus dCas9 handle	AACAAGGCAAAATGCCGTGTTTATCTCGTCA
	ACTTGTTGGCGAGATTTTTT
	GAAGCTTGGGCCCGAACAAAAACTCATCTCA
	GAAGAGGATCTGAATAGCGCCGTCGACCAT
	CATCATCATCATCATTGAGTTTAAACGGTCT
	CCAGCTTGGCTGTTTTGGCGGATGAGAGAA
	GATTTTCAGCCTGATACAGATTAAATCAGAA
	CGCAGAAGCGGTCTGATAAAACAGAATTTG
S. pyogenes terminator	CCTGGCGGCAGTAGCGCGGTGGTCCCACCT
	GACCCCATGCCGAACTCAGAAGTGAAACGCC
	GTAGCGCCGATGGTAGTGTGGGGTCTCCCC
	ATGCGAGAGTAGGGAACTGCCAGGCATCAA
	ATAAAACGAAAGGCTCAGTCGAAAGACTGG
	GCCTTTCGTTTTATCTGTTGTTTGTCGGTGA
	ACT
PAM sequence	TTGAGTA
Target region	GGTTCAATATGGTCTTGATCC

C.4 *In vivo* enzyme characterization

In this Section the *ad hoc* designed biological circuits used for the characterization of the transcriptional repression (SadCas9) and activation (RNAPT7-P_{T7}) systems have been reported. The rationale behind both characterization circuits is to analyze how the fluorescent output level (RFP) changes as a function of the repressor (SadCas9) or activator enzyme (RNAPT7) concentration variation. Their expressions have been modulated with different HSL inducer concentrations via the P_{Lux} promoter activated from Lux-circuitry activity and different RBS sequences upstream of the activator or repressor gene. All the assembled biological circuits have been transformed in the test strain

E. coli TOP10 F' since it carries the $lacI^q$ repressor for inducible expression from lac promoter using the IPTG molecule, which is used to trigger sgRNA expression (see below).

C.4.1 *Staphylococcus aureus* dCas9 (SadCas9) transcriptional repression system

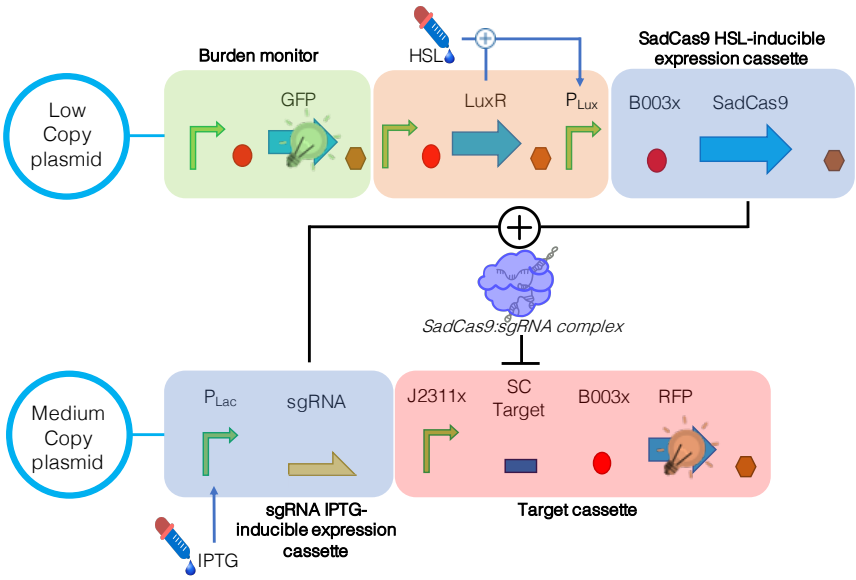

Figure C.1: **Circuitry scheme for *S. aureus* dCas9 characterization.** The synthetic circuits used for SadCas9 characterization are based on a two-plasmid system. The first plasmid (low-copy plasmid) contains two modules responsible, respectively, for monitoring cell load (burden monitor) and for HSL-inducible expression of the SadCas9 protein. The second (medium-copy plasmid) contains the IPTG-inducible expression of sgRNA module and the target cassette characterized by the fluorescent protein RFP.

Originally, the coding sequence of SadCas9 in the purchased Addgene plasmid #113718 presented two BioBrick™ restriction sites (EcoRI, XbaI) that interfere with the standard assembly procedures.

Two single-point mutations (c.1449T>C and c.1629A>C) have been carried out to delete them without changing the amino acid composition of the protein, via the protocol that has been reported in Section C.2.1.

The characterization circuits scheme, based on two-plasmid system, for the SadCas9 repressor enzyme has been reported in Fig. C.1. The expression of the active SadCas9 repressor complex (SadCas9:sgRNA) is modulated by two inducers: HSL that drives the P_{Lux} transcriptional activity (and thus the expression of free SadCas9 protein) and IPTG that drives the single guide RNA expression through the P_{Lac} promoter. The target DNA region of SadCas9 has been placed downstream of the Transcription Start Site (TSS) of the constitutive promoter responsible for the expression of the fluorescent protein (RFP). Constitutive GFP expression cassette (in the low-copy plasmid) has been used as burden monitor. The concentration of the IPTG inducer has been set to $200\mu M$ (fully inducing the P_{Lac} promoter in the used strain) in order to meet the overabundance hypothesis of sgRNA compared with SadCas9 level. Several circuit combinations have been constructed by changing the RBS sequence upstream of SadCas9 (B0031, B0032, B0034, CU1, CU2, CA1) and RFP (B0031, B0034) coding sequences and the target promoter sequence (J118, J119), as reported in Table C.5. The design of CU1, CU2 and CA1 RBSs has been described in Section C.6.

Table C.5: **Synthetic circuits used for *S. aureus* dCas9 (SadCas9) characterization.** Each circuit reported in the first column is composed of the two plasmids in the adjacent columns. The final construct name, CSxyz, is composed as: 'CS' - Characterization SadCas9, 'x' - SadCas9 RBS code, 'y' - RFP promoter code, 'z' - RFP RBS code. The 'x', 'y' and 'z' codes are highlighted with bold font within plasmid name.

Construct name	**Low copy plasmid**	**Medium copy plasmid**
CS191	AE-3A3**1**SadCas9	Plac_sgRNA J11**9**SCTarget**31**RFP
CS291	AE-3A3**2**SadCas9	Plac_sgRNA J11**9**SCTarget**31**RFP

Continues on next page...

Table C.5 – *...continued from previous page*

Construct name	Low copy plasmid	Medium copy plasmid
CS491	AE-3A3**4**SadCas9	Plac_sgRNA J11**9**SCTarget3**1**RFP
CS184	AE-3A3**1**SadCas9	Plac_sgRNA J11**8**SCTarget3**4**RFP
CS284	AE-3A3**2**SadCas9	Plac_sgRNA J11**8**SCTarget3**4**RFP
CS484	AE-3A3**4**SadCas9	Plac_sgRNA J11**8**SCTarget3**4**RFP
CSU184	AE-3A3C**U1**SadCas9	Plac_sgRNA J11**8**SCTarget3**4**RFP
CSU284	AE-3A3C**U2**SadCas9	Plac_sgRNA J11**8**SCTarget3**4**RFP
CSA184	AE-3A3C**A1**SadCas9	Plac_sgRNA J11**8**SCTarget3**4**RFP
CS194	AE-3A3**1**SadCas9	Plac_sgRNA J11**9**SCTarget3**4**RFP
CS294	AE-3A3**2**SadCas9	Plac_sgRNA J11**9**SCTarget3**4**RFP
CS494	AE-3A3**4**SadCas9	Plac_sgRNA J11**9**SCTarget3**4**RFP

C.4.2 RNAPT7-P_{T7} transcriptional activation system

The high toxicity that arose from the expression of the wild-type RNAPT7-P_{T7} system in the test strain *E. coli* TOP10 F' in preliminary experiments (data not shown) indicated the need to reduce the transcriptional system efficiency to enable the use of an activation motif compatible with cell life. As reported in Section 4.2.4, a single-point mutation in the RNAPT7 coding sequence (c.1894C>A) has been carried out that results in an amino acid alteration in the protein sequence (p.Arg632Ser), inspired by previous works in which this (and other) mutations were used to decrease the efficiency of RNAPT7. The characterization scheme, based on a two-plasmid system, for the RNAPT7 activator enzyme, has been reported in Fig. C.2. The RNAPT7 activator is expressed usign the P_{Lux} promoter, whose transcriptional activity is triggered by using the HSL inducer. The recognition of the cognate promoter P_{T7} by the RNAPT7 enzyme enhances the expression of the fluorescent protein (RFP). Several circuit combinations have been constructed by changing the RBS sequence (B0031, B0032,

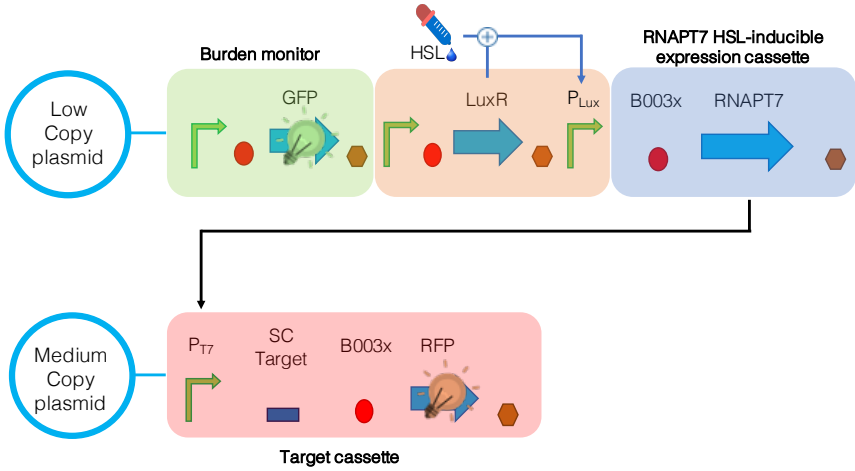

Figure C.2: **Circuitry scheme for RNAPT7-P_{T7} *in vivo* characterization.** The synthetic circuits used for RNAPT7-P_{T7} characterization are based on two-plasmid system. The first plasmid (low copy plasmid) contains two modules responsible, respectively, for monitoring cell load (burden monitor) and for HSL-inducible expression of RNAPT7 protein. The second (medium copy plasmid) contains the RNAPT7-inducible expression of RFP cassette.

B0034) upstream of the RNAPT7 and RFP coding sequences, reported in Table C.6.

Table C.6: **Synthetic circuits used for the characterization of the RNAPT7-P_{T7} transcriptional system.** Each circuit reported in the first column is composed of the two plasmids in the adjacent columns. The final construct name, CT7xy, is composed as: 'CT7' - Characterization RNAPT7(R632S), 'x' - RNAPT7 RBS code, 'y' - RFP RBS code. The 'x' and 'y' codes are highlighted with bold font within plasmid name.

Construct name	Low copy plasmid	Medium copy plasmid
CT7 3131	AE-3A**31**T7(R632S)	P_{T7}SCTarget**31**RFP
CT7 3231	AE-3A**32**T7(R632S)	P_{T7}SCTarget**31**RFP
CT7 3431	AE-3A**34**T7(R632S)	P_{T7}SCTarget**31**RFP
CT7 3132	AE-3A**31**T7(R632S)	P_{T7}SCTarget**32**RFP
CT7 3232	AE-3A**32**T7(R632S)	P_{T7}SCTarget**32**RFP
CT7 3432	AE-3A**34**T7(R632S)	P_{T7}SCTarget**32**RFP

Continues on next page...

Table C.6 – *...continued from previous page*

Construct name	Low copy plasmid	Medium copy plasmid
CT7 3134	AE-3A**31**T7(R632S)	P_{T7}SCTarget**34**RFP
CT7 3234	AE-3A**32**T7(R632S)	P_{T7}SCTarget**34**RFP
CT7 3434	AE-3A**34**T7(R632S)	P_{T7}SCTarget**34**RFP

C.5 Quantitative assays and data analysis

C.5.1 Fluorescence and growth assays

Bacteria from long-term glycerol stocks were streaked on LB agar plates supplemented with the proper antibiotic(s), with the purpose of isolating single colonies. These colonies were considered as biological replicates. Plates have been incubated overnight at 37°C, then 0.5 ml of selective M9 (M9 salts - #M6030, Sigma Aldrich - 11.28g/l, thiamine hydrochloride 1mM, MgSO4 2mM, CaCl2 0.1mM, casamino acids 0.2%, glycerol 0.4%) have been inoculated with single colonies and incubated overnight in 2−ml tubes at 37°C, 220rpm (when explicitly indicated, IPTG inducer has been also added at the final concentration of 200 μM). Cultures were 100-fold diluted in the same medium (with IPTG when indicated) in 200 μl in a 96-well microplate. The HSL inducer was also added to the microplate wells as indicated to trigger the expression of the P_{Lux} promoter. Cultures were assayed via the Infinite F200Pro microplate reader (Tecan), programmed with the i-control (Tecan) software to perform a kinetic cycle as follows: linear shaking (3 mm amplitude, 15s), wait (5s), absorbance measurement (600nm), red fluorescence measurement (excitation at 535nm, emission at 620nm, gain=50), repeat cycle every 5min. In every microplate experiment, 200μl of M9 and a non-fluorescent TOP10 F' *E.*

coli culture were included in triplicate to enable the estimation of absorbance and fluorescence backgrounds, respectively. Finally, the ConstRFP and ConstGFP strains, including a constitutive RFP and GFP expression cassette, respectively, under the control of the J23101 constitutive promoter (with B0034 and B0032 RBS, respectively), herein named REF cultures, where also included in triplicate as controls to enable the computation of fluorescence outputs in normalized units. Absorbance and fluorescence background signals were subtracted from raw absorbance and fluorescence over time (t), to obtain cell density (X, expressed as optical density - OD - proportional to the per-well cell count), and RFP (R, expressed as arbitrary units of raw RFP - AUR - proportional to the per-well number of fluorescent proteins) and GFP (G, expressed as arbitrary units of raw GFP - AUG - proportional to the per-well number of fluorescent proteins). Since a significant cell density-dependent autofluorescence was previously reported for GFP measurements with the adopted experimental setup, green fluorescence has been blanked via a different procedure, described in [212]: the raw green (auto)fluorescence (GFPauto) as a function of OD600 has been computed as:

$$GFP_{auto}(t) = e^{q+m \cdot OD_{600}(t)} \tag{C.1}$$

and it has been subtracted from the raw fluorescence value of each GFP-expressing strain to obtain a signal proportional to the GFP level in the whole culture. This exponential function was previously parametrized with the growth rate-dependent coefficients q and m, measured from different exponentially growing cultures used for calibration. A signal proportional to RFP or GFP synthesis rate per cell S^{raw}_{cell}, expressed as $AUR \cdot OD^{-1} \cdot min^{-1}$ has been computer over time in the exponential growth phase (EGP, identified via visual inspection, typically $0.02 < OD_{600} < 0.14$ in microplate experiments) for each culture as the numeric time derivative of fluorescence, divided by the

mean cell density.

$$S^{raw}_{cell,i} = \frac{F_i - F_{i-1}}{t_i + t_{i-1}} \cdot \frac{2}{X_i + X_{i-1}}, \forall i \in EGP, F = [R, G]; \tag{C.2}$$

The mean value ($\overline{S^{raw}_{cell}}$) of the S^{raw}_{cell} time series of the REF strain replicates $S^{raw}_{cell,REF}$ has been computer over the EGP and it has been divided by ($\overline{S^{raw}_{cell,REF}}$) for normalization, obtaining the RPU (Relative Promoter Units) value, expressed as $AU \cdot min^{-1}$:

$$RPU = \frac{S^{raw}_{cell,i}}{\overline{S^{raw}_{cell,REF,i}}} \tag{C.3}$$

RPU is proportional to the synthesis rate of the fluorescence protein (RFP or GFP), and it is also is proportional to the intracellular level of RFP or GFP produced, for a given growth rate. Growth rate (μ) was computed as the slope of the regression line of the $log(X(t))$ time series in the EGP.

C.6 New RBS design for SadCas9 expression

New synthetic RBSs for SadCas9 expression have been designed using the free online tool RBS Calculator v2.1 (Salis Lab) with the aim of expanding the SadCas9 expression range and, at the same time, ensuring expression of the target RFP protein in order to test new combinations of Sad-iFFL in Chapter 3. New synthetic RBSs have been generated based on the information of: (i) host organism (*Escherichia coli str. K-12 substr MG1655*, NCBI accession: NC_000913), (ii) the coding sequence of the target protein (*sadcas9*) and (iii) the desidered Translation Initiation Rate (TIR) value. The latter has been set by considering values higher than the predicted TIR values for

B0032 upstream of the SadCas9 coding sequence, and around B0034 upstream of the same gene. From a candidate RBS, with the same algorithm, the TIR of RFP (with upstream the target sequence for SadCas9:sgRNA binding) has been predicted in order to guarantee the functionality of the RBS designed also for the expression of the reporter protein. Three synthetic RBSs have been designed: CU1 and CU2 with $TIR_{B0032}^{SadCas9} < TIR_{X}^{SadCas9} < TIR_{B0034}^{SadCas9}$ where X=[CU1, CU2], and CA1 with $TIR_{CA1}^{SadCas9} > TIR_{B0034}^{SadCas9}$. $TIR_{B0032}^{SadCas9}$ has been used as lower constraint for TIR since the RFP signal produced is low but well detectable (when driven by Lux circuitry at full induction) usgin the Infinite F200Pro microplate reader (Tecan) and good results have been obtained in characterization of SadCas9 with this RBS. The three synthetic RBS candidates obtained have been reported in Table C.7 and their predicted TIRs are reported as percentages of B0034 for SadCas9 ad RFP.

Table C.7: **Synthetic RBS designed with RBS Calculator**

RBS	**Sequence**	[%] **of** $TIR_{B0034}^{SadCas9}$	[%] **of** TIR_{B0034}^{RFP}
B0032	Reference CCATAAAAACTT	79.94	35.72
CU1	GACACTAGGGTC AAAAT TATTTAAAAGGA	86.31	77.83
CU2	AAACATCAAAGG GCACT CGGTTACCACGT	91.59	65.94
CA1	TGTAAAGGAACA ACAGA	162.14	92.8

Appendix D

Supplementary information: in silico models simulation and data fitting

In this appendix the model parameters that have been used for the *in silico* study of Open Loop, Sad-iFFL and U-iFFL models are reported. The methods to compute the model performance at the steady state (steady state analysis, biological noise propagation) and the dynamic evolution (induction/de-induction cycle, settling time) are also illustrated. In the last Section, the procedure on the experimental data of SadCas9 characterization has been reported.

D.1 Simulations

The model parameters that have been used to simulate the models (Open Loop, Sad-iFFL and U-iFFL) are reported in Table D.1. P_U is the intracellular concentrations of one DNA copy of promoter and it is used to convert the number of DNA molecules into its relative concentration value ([nM]), while the remaining parameters have been explained in Section 3.2.2 and Section 4.1.2. All simulations and data

fitting steps have been made with MATLAB R2018a (MathWorks, Natick, MA, USA); parameter estimation procedure has been performed through the ordinary least squares algorithm via the *lsqnonlin* function on all experimental data simultaneously.

Steady state analysis

Steady state analysis shows how a given system performs at its steady state as a function of perturbation factors like transcription, translation and target protein levels variations. This analysis has been made for the three designed circuits (Open Loop, Sad-iFFL and U-iFFL). Each of them was tested with different range of transcription rate (α $[s^{-1}]$), translation rate (ρ $[s^{-1}]$) and different desired GOI concentration, defined via the corresponding K_{CG} $[nM]$ value. The robustness index on transcription rate α and translation rate ρ has been calculated as the median of the fold-change distribution on the other parameter (translation rate ρ, transcription rate α, respectively) variation (Equation (D.2)), as follows:

$$Robustness(\alpha) = median([FC_{\rho_1}, ...FC_{\rho_i}, ...FC_{\rho_N}]) \tag{D.1}$$

$$Robustness(\rho) = median([FC_{\alpha_1}, ...FC_{\alpha_i}, ...FC_{\alpha_N}]) \tag{D.2}$$

where FC_{ρ_i} and FC_{α_i} are the fold-change values computed by fixing the i-th transcription rate or the i-th translation rate value, respectively, as reported in Equations (D.3) and (D.4):

$$FC_{\delta_i} = \begin{cases} \frac{GOI(\alpha_{min},\rho_i)}{GOI(\alpha_{max},\rho_i)} \; if \; GOI(\alpha_{max},\rho_i) > GOI(\alpha_{min},\rho_i) \\ \frac{GOI(\alpha_{max},\rho_i)}{GOI(\alpha_{min},\rho_i)} \; if \; GOI(\alpha_{min},\rho_i) > GOI(\alpha_{max},\rho_i) \end{cases} \tag{D.3}$$

$$FC_{\alpha_i} = \begin{cases} \frac{GOI(\alpha_i,\rho_{min})}{GOI(\alpha_i,\rho_{max})} \; if \; GOI(\alpha_i,\rho_{max}) > GOI(\alpha_i,\rho_{min}) \\ \frac{GOI(\alpha_i,\rho_{max})}{GOI(\alpha_i,\rho_{min})} \; if \; GOI(\alpha_i,\rho_{min}) > GOI(\alpha_i,\rho_{max}) \end{cases} \tag{D.4}$$

Propagation of biological noise

Noise has been assumed to affect transcription and translation rate ($\alpha \cdot \rho$) on the steady state model of Open Loop (Equation (B.3)), Sad-iFFL (Equations (3.15)−(3.18)) and U-iFFL (Equations (4.17)−(4.21)) as follows:

$$P = p \cdot v \ , \ p = \alpha \cdot \rho \tag{D.5}$$

$$v \sim LogN(0, \sigma^2) \tag{D.6}$$

where the logarithm of the lognormal distribution (v) is a Gaussian distribution with mean $\mu = 0$ and variance σ^2; this noise representation has been widely used in literature to describe the fluorescence distribution of reporter proteins in cell populations engineered with synthetic circuits [111]. The statistical proprieties of P ($AVE(P) = e^{\frac{\sigma^2}{2}}$ and $VAR(P) = e^{\sigma^2} \cdot (e^{\sigma^2} - 1)$) show that the mean is biased by a factor of $e^{\frac{\sigma^2}{2}}$ due to lognormal distribution. A correction term has been introduced to clear variable P from this undesirable contribution, as follows:

$$p = p_{pop} \cdot e^{-\frac{\sigma^2}{2}} \tag{D.7}$$

The final noise model has been derived on constant coefficient of variation (CV) assumption, in which, a value of variance σ^2 has been obtained as:

$$\sigma^2 = ln(1 + CV^2) \tag{D.8}$$

Cellular noise can be split in two components: extrinsic and intrinsic noise. The first (extrinsic) refers to the variation that affect equally the whole set of regulatory parts (as reported in Equation (D.5), the single part is considered here as the sequence composed by promoter and RBS pair) within a system, while, the second (intrinsic) refers to the variation in identically-regulated quantities that arises from the stochasticity of biochemical reactions in the same cell. A correlation coefficient ϕ ($0 \leq \phi \leq 1$) has been introduced to model the proportion between intrinsic ($\phi = 0$) and extrinsic ($\phi = 1$) factors that compose

the total noise. Particularly, if noise is composed only by the intrinsic component, the behaviors of the regulatory parts in an isogenic recombinant cell population are indipendent. The parameters that have been used for simulate the propagation of biological noise in the Open loop, Sad-iFFL and U-iFFL models have been reported in Table D.1. The CV values for multiplicative lognormal noise range between 0% and 100% and the final CV on model outputs (GOI $[nM]$) has been computed from 10,000 independent generated samples.

Induction/de-induction cycle
To test the designed systems dynamic behavior, transcriptional activity was assumed to be null and to be triggered at t=0. The models were analyzed computationally with MATLAB using a differential equation solver by low order method (*ode23s*) and the solution of each model (Open Loop, Sad-iFFL and U-iFFL) has been computed separately. The transcriptional activity has been described through a transcriptional impulse of time t_{set} for which the target promoters is active with transcription rate of α $[s^{-1}]$ that has been setted opportunatly as reported in Section 3.3.4 and Section 4.2.4.

Settling time
The settling time is the time required for the response curve to reach and stay within a range around the final value of size specified by absolute percentage of the final value (usually 2% or 5%). The settling time was here recorded at 5% of the difference from the steady state, which was considered equal to the final estimated value from the differential equation solver.

Table D.1: **Model parameters for *in silico* simulations of Open Loop, Sad-iFFL and U-iFFL models.** [a]Intracellular concentration of one DNA copy of promoter; [b]When indicated, a range of values was spanned.

Parameter	Units	Models	Value
P_U	nM	All	1.66^a
d_{RNA}	s^{-1}	All	0.0042
$d_{Protein}$	s^{-1}	All	3.83e-4
μ	s^{-1}	All	0.023
n	–	All	1^b
α	s^{-1}	All	1^b
ρ	s^{-1}	All	0.4^b
b	–	Sad-iFFL, U-iFFL	1
f	–	Sad-iFFL, U-iFFL	1
K_{CG}	nM	Sad-iFFL, U-iFFL	500^b
c	–	U-iFFL	1
g	–	U-iFFL	1
K_{CT}	nM	U-iFFL	500
K_{T7}	nM	U-iFFL	10

D.2 Models implementation for the characterization of SadCas9 repressor

The repression and activation of protein expression in the circuit for SadCas9 characterization (Fig. C.1) have been represented with Hill functions modulated by their degradation rate assuming the steady-state of all the molecular species within the cell in the exponential growth phase, as reported in Equation (D.9).

$$S_j = \left(\beta_0^j + \frac{\beta_{max}^j}{1 + \left(\frac{K_j}{I_j}\right)^{\pm \eta_j}} \right) \cdot \frac{1}{\mu} \tag{D.9}$$

where β_0, β_{max}, K and η are the Hill parameters that characterize the transfer function between the Input molecule (I) and the output protein (S); in more details, β_0 is the minimum synthesis rate, $\beta_0 + \beta_{max}$ is the maximum synthesis rate, K is the input (I) level corresponding to the half-maximum concentration of the output (S) and η is the Hill coefficient which describes the cooperativity and the effect of the input (I) to the output (S) (no cooperativity - $\eta = 1$, (I) repressor - $\eta < 0$, (I) activator - $\eta > 0$); μ is the dilution rate due to cell division (approximation due to negligible protein degradation rate compared with growth rate). The Hill parameters reported hereby lumped the information on the biologically-meaningful parameters used in this work (e.g., transcription rate α, translation rate ρ) in order to work with an identifiable model from the experimental data of SadCas9 characterization and thus enabling the fitting procedure. For fluorescent proteins, as RFP, the maturation process has been modeled as follows:

$$R_{immature} = \frac{1}{\mu + \theta^{RFP}} \cdot S_j \tag{D.10}$$

$$R_{mature} = \theta^{RFP} \cdot R_{immature} \tag{D.11}$$

where θ^{RFP} is the maturation rate for RFP. In Equation (D.11), R_{mature} (called R from now) is expressed in $AU \cdot cell^{-1} \cdot min^{-1}$ and refers to the normalized RFP value of S_{cell} computed from the *in vivo* experimental data of SadCas9 characterization that has been illustrated in Section 3.3.5 (Fig. 3.6). The cell load modeling (described in Section 2.2.6) has been used here to better describe the data affected by metabolic burden caused by the incorporation of the circuit in the bacterial host. As mentioned in Section 2.2.6, the cell load effect is modeled through the parameter D as denominator on all the protein synthesis rates, reported as follows:

$$S_{j,B} = \frac{S_j}{D} \tag{D.12}$$

$$D = 1 + \sum_{i=1}^{c} J_i \cdot S_{i,B} = 1 + J_R \cdot R + J_C \cdot C \tag{D.13}$$

where $S_{j,B}$ is the j-th protein synthesis rate affect from burden. The parameter D has been readapted (from Equation (2.41)) within this context and reported in Equation (D.13) assuming that the cell load is generated by RFP and SadCas9 proteins. Describing RFP and SadCas9 production rates as Equation (D.9), considering the maturation rate of RFP (Equations (D.10) − (D.11)) and modeling cell burden as reported in Equations (D.12)−(D.13), the whole model used to fit the experimental data of SadCas9 characterization has been derived as follows:

$$C = \left(\beta_0^C + \frac{\beta_{max}^C}{1 + (\frac{K_{Lux} \cdot D}{HSL})^{\eta_{Lux}}} \right) \cdot \frac{1}{\mu_1 \cdot D} \tag{D.14}$$

$$R = \frac{\theta^{RFP}}{(\mu_2 + \theta^{RFP}) \cdot \mu_2 \cdot D} \cdot \left(\beta_0^X + \frac{\beta_{max}^X}{1 + (\frac{C}{K_C^Y})^{\eta_C}} \right) \tag{D.15}$$

$$G = \frac{max(G_{Data})}{D} \tag{D.16}$$

$$\begin{aligned} D = 1 + J_R \cdot \left(\beta_0^X + \frac{\beta_{max}^X}{1 + (\frac{C}{K_C^Y})^{\eta_C}} \right) \cdot \frac{\theta^{RFP}}{(\mu_2 + \theta^{RFP}) \cdot \mu_2 \cdot D} + \\ J_C \cdot \left(\beta_0^C + \frac{\beta_{max}^C}{1 + (\frac{K_{Lux} \cdot D}{HSL})^{\eta_{Lux}}} \right) \cdot \frac{1}{\mu_1 \cdot D} \end{aligned} \tag{D.17}$$

where C, R, G and D are the SadCas9, RFP, GFP protein level and burden parameter, respectively. $\beta_0^C[AU]$ and $\beta_{max}^C[AU]$ are the minimum synthesis rate and the maximum excursion between off- and on- state rate for SadCas9 (C) expressed from Lux-circuitry (HSL); $\beta_0^X[AU]$ and $\beta_{max}^X[AU]$ have the same meaning but for RFP protein expressed by the promotor-RBS combination X ($X = [J118-34, J119-$

31, $J119-34$]). η_C and η_{Lux} are the Hill parameters which describe the cooperativity of Lux-activation and dCas9-repression. The only equation based on experimental data by design is Equation (D.16), in which $max(G_{Data})$ is the maximum value of GFP on different HSL concentration value. The resource usage terms, J_R and J_C for RFP (R) and SadCas9 (C), respectevely, represent a weight coefficient related with of the burden caused by the associate protein; both are expressed in $[AU^{-1} \cdot min]$. $K_C^Y[AU]$ are the SadCas9 level from RBS Y ($Y = B0032, B0034, CU1, CU2, CA1$) for which the half-maximum concentration of the output (RFP) is reached and K_{Lux} is the HSL concentration for which half-maximum level of SadCas9 is achieved. In biological terms, the K_C^Y value is unique for each dCas9 protein and does not change with the expression level of the repressor protein. As mentioned at the beginning of this Section, in order to obtain an identifiable model from the available data, the parameters hereby used lump the information of two or more biological descriptors; here, K_C^Y describes the variation of SadCas9 due to different RBS strength and the real half-maximum constant $\widehat{K_C}$. Precisely, the K_C^Y can be written as

$$K_C^Y = \widehat{K_C}/\tau_{RBS} \tag{D.18}$$

where τ_{RBS} is the RBS strength, which modulates the resulting repressor protein level and a lumped K parameter was estimated for any given RBS combinations in the tested strains.

Bibliography

[1] R Jaenisch and B Mintz. Simian virus 40 dna sequences in dna of healthy adult mice derived from preimplantation blastocysts injected with viral dna. *Proceedings of the national academy of sciences*, 71(4):1250–1254, 1974.

[2] F Meng and T Ellis. The second decade of synthetic biology: 2010–2020. *Nature Communications*, 11(1):1–4, 2020.

[3] T Lewens. From bricolage to biobricks™: Synthetic biology and rational design. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, 44(4):641–648, 2013.

[4] V de Lorenzo. Evolutionary tinkering vs. rational engineering in the times of synthetic biology. *Life sciences and society and policy*, 14(1):1–16, 2018.

[5] S Mukherji and A Van Oudenaarden. Synthetic biology: understanding biological design from synthetic circuits. *Nature Reviews Genetics*, 10(12):859–871, 2009.

[6] MB Elowitz and S Leibler. A synthetic oscillatory network of transcriptional regulators. *Nature*, 403(6767):335–338, 2000.

[7] TS Gardner, CR Cantor, and JJ Collins. Construction of a genetic toggle switch in escherichia coli. *Nature*, 403(6767):339–342, 2000.

[8] E Freed J Fenster SL Smolinski J Walker CA Henard R Gill, and CA Eckert. Building a genome engineering toolbox in non-model prokaryotic microbes. *Biotechnology and bioengineering*, 115(9):2120–2138, 2018.

[9] H Chi, X Wang, Y Shao, Y Qin, Z Deng, L Wang, and S Chen. Engineering and modification of microbial chassis for systems and synthetic biology. *Synthetic and systems biotechnology*, 4(1):25–33, 2019.

[10] BL Adams. The next generation of synthetic biology chassis: moving synthetic biology from the laboratory to the field, 2016.

[11] DM Widmaier and CA Voigt. Quantification of the physiochemical constraints on the export of spider silk proteins by salmonella type iii secretion. *Microbial Cell Factories*, 9(1):78, 2010.

[12] A Azam, C Li, KJ Metcalf, and D Tullman-Ercek. Type iii secretion as a generalizable strategy for the production of full-length biopolymer-forming proteins. *Biotechnology and Bioengineering*, 113(11):2313–2320, 2016.

[13] AW Westbrook, X Ren, J Oh, M Moo-Young, and CP Chou. Metabolic engineering to enhance heterologous production of hyaluronic acid in bacillus subtilis. *Metabolic engineering*, 47:401–413, 2018.

[14] G Gonzalez, G Herrera, MT Garcia, and M Pena. Biodegradation of phenolic industrial wastewater in a fluidized bed bioreactor with immobilized cells of pseudomonas putida. *Bioresource technology*, 80(2):137–142, 2001.

[15] MS Samuel, A Sivaramakrishna, and A Mehta. Bioremediation of p-nitrophenol by pseudomonas putida 1274 strain. *Journal of Environmental Health Science and Engineering*, 12(1):53, 2014.

[16] PI Nikel, M Chavarria, A Danchin, and V de Lorenzo. From dirt to industrial applications: Pseudomonas putida as a synthetic biology chassis for hosting harsh biochemical reactions. *Current opinion in chemical biology*, 34:20–29, 2016.

[17] FA Millacura, F Cárdenas, V Mendez, M Seeger, and LA Rojas. Degradation of benzene by the heavy-metal resistant bacterium cupriavidus metallidurans ch34 reveals its catabolic potential for aromatic compounds. *bioRxiv*, page 164517, 2017.

[18] BS Dien, MA Cotta, and TW Jeffries. Bacteria engineered for fuel ethanol production: current status. *Applied microbiology and biotechnology*, 63(3):258–266, 2003.

[19] Y Lin and S Tanaka. Ethanol fermentation from biomass resources: current state and prospects. *Applied microbiology and biotechnology*, 69(6):627–642, 2006.

[20] MR Connor and S Atsumi. Synthetic biology guides biofuel production. *BioMed Research International*, 2010, 2010.

[21] L Pasotti, S Zucca, M Casanova, G Micoli, MG Cusella De Angelis, and P Magni. Fermentation of lactose to ethanol in cheese whey permeate and concentrated permeate by engineered escherichia coli. *BMC biotechnology*, 17(1):48, 2017.

[22] LO Ingram, PF Gomez, X Lai, M Moniruzzaman, BE Wood, LP Yomano, and SW York. Metabolic engineering of bacteria for ethanol production. *Biotechnology and bioengineering*, 58(2-3):204–214, 1998.

[23] L Pasotti, D De Marchi, M Casanova, I Massaiu, M Bellato, MG Cusella De Angelis, C Calvio, and P Magni. Engineering endogenous fermentative routes in ethanologenic escherichia coli w for bioethanol production from concentrated whey permeate. *New Biotechnology*, 57:55–66, 2020.

[24] MS Roell and MD Zurbriggen. The impact of synthetic biology for future agriculture and nutrition. *Current Opinion in Biotechnology*, 61:102–109, 2020.

[25] C Rogers and GED Oldroyd. Synthetic biology approaches to engineering the nitrogen symbiosis in cereals. *Journal of experimental botany*, 65(8):1939–1946, 2014.

[26] A Coniglio, V Mora, M Puente, and F Cassán. Azospirillum as biofertilizer for sustainable agriculture: Azospirillum brasilense az39 as a model of pgpr and field traceability. In *Microbial Probiotics for Agricultural Systems*, pages 45–70. Springer, 2019.

[27] WO Draghi, J Degrossi, M Bialer, G Brelles-Marino, P Abdian, L Wall A Soler-Bistué, and A Zorreguieta. Biodiversity of cultivable burkholderia species in argentinean soils under no-till agricultural practices. *PloS one*, 13(7):e0200651, 2018.

[28] JF Salles, JD Van Elsas, and JA Van Veen. Effect of agricultural management regime on burkholderia community structure in soil. *Microbial Ecology*, 52(2):267–279, 2006.

[29] S O'Brien and A Buckling. The sociality of bioremediation: hijacking the social lives of microbial populations to clean up heavy metal contamination. *EMBO reports*, 16(10):1241–1245, 2015.

[30] PKR Tay, PQ Nguyen, and NS Joshi. A synthetic circuit for mercury bioremediation using self-assembling functional amyloids. *ACS synthetic biology*, 6(10):1841–1850, 2017.

[31] P Dvořák, PI Nikel, J Damborskỳ, and V de Lorenzo. Bioremediation 3.0: engineering pollutant-removing bacteria in the times of systemic biology. *Biotechnology advances*, 35(7):845–866, 2017.

[32] S Slomovic, K Pardee, and JJ Collins. Synthetic biology devices for in vitro and in vivo diagnostics. *Proceedings of the National Academy of Sciences*, 112(47):14429–14435, 2015.

[33] AS Khalil and JJ Collins. Synthetic biology: applications come of age. *Nature Reviews Genetics*, 11(5):367–379, 2010.

[34] T Danino, A Prindle, GA Kwong, M Skalak, H Li, K Allen, J Hasty, and SN Bhatia. Programmable probiotics for detection of cancer in urine. *Science translational medicine*, 7(289):289ra84–289ra84, 2015.

[35] Y Fan and O Pedersen. Gut microbiota in human metabolic health and disease. *Nature Reviews Microbiology*, pages 1–17, 2020.

[36] MR Charbonneau, VM Isabella, N Li, and CB Kurtz. Developing a new class of engineered live bacterial therapeutics to treat human diseases. *Nature Communications*, 11(1):1–11, 2020.

[37] PJ Turnbaugh, PE Ley, MA Mahowald, V Magrini, ER Mardis, and JI Gordon. An obesity-associated gut microbiome with increased capacity for energy harvest. *nature*, 444(7122):1027, 2006.

[38] LV Blanton, MJ Barratt, MR Charbonneau, T Ahmed, and JI Gordon. Childhood undernutrition and the gut microbiota and microbiota-directed therapeutics. *Science*, 352(6293):1533–1533, 2016.

[39] SK Mazmanian, JL Round, and DL Kasper. A microbial symbiosis factor prevents intestinal inflammatory disease. *Nature*, 453(7195):620–625, 2008.

[40] VR Matson, J Fessler, R Bao, T Chongsuwat, Y Zha, ML Alegre, JJ Luke, and TF Gajewski. The commensal microbiome is associated with anti–pd-1 efficacy in metastatic melanoma patients. *Science*, 359(6371):104–108, 2018.

[41] V Gopalakrishnan, CN Spencer, L Nezi, A Reuben, MC Andrews, TV Karpinets, PA Prieto, D Vicente, K Hoffman, SC Wei, et al. Gut microbiome modulates response to anti–pd-1 immunotherapy in melanoma patients. *Science*, 359(6371):97–103, 2018.

[42] JR Bober, CL Beisel, and NU Nair. Synthetic biology approaches to engineer probiotics and members of the human microbiota for biomedical applications. *Annual review of biomedical engineering*, 20:277–300, 2018.

[43] N Yang, X Zhu, L Chen, S Li, and D Ren. Oral administration of attenuated s. typhimurium carrying shrna-expressing vectors as a cancer therapeutic. *Cancer biology & therapy*, 7(1):145–151, 2008.

[44] E Jacouton, E Torres Maravilla, AS Boucard, N Pouderous, AP Pessoa Vilela, I Naas, F Chain, V Azevedo, P Langella, , and LG Bermúdez-Humarán. Anti-tumoral effects of recombinant lactococcus lactis strain secreting il-17a cytokine. *Frontiers in microbiology*, 9:3355, 2019.

[45] JD Palmer, E Piattelli, BA McCormick, MW Silby, CJ Brigham, J Christopher, and V Bucci. Engineered probiotic for the inhibition of salmonella via tetrathionate-induced production of microcin h47. *ACS infectious diseases*, 4(1):39–45, 2018.

[46] P Praveschotinunt, AM Duraj-Thatte, I Gelfat, F Bahl, DB Chou, and NS Joshi. Engineered e. coli nissle 1917 for the delivery of matrix-tethered therapeutic domains to the gut. *Nature communications*, 10(1):1–14, 2019.

[47] AC Wong and M Levy. New approaches to microbiome-based therapies. *MSystems*, 4(3), 2019.

[48] Z Zhou, X Chen, H Sheng, X Shen, X Sun, Y Yan, J Wang, and Q Yuan. Engineering probiotics as living diagnostics and therapeutics for improving human health. *Microbial Cell Factories*, 19(1):1–12, 2020.

[49] B Lim, M Zimmermann, NA Barry, and AL Goodman. Engineered regulatory systems modulate gene expression of human commensals in the gut. *Cell*, 169(3):547–558, 2017.

[50] M Mimee, AC Tucker, CA Voigt, and TK Lu. Programming a human commensal bacterium and bacteroides thetaiotaomicron and to sense and respond to stimuli in the murine gut microbiota. *Cell systems*, 1(1):62–71, 2015.

[51] A Mauras, F Chain, A Faucheux, S Gontier P Ruffié, B Ryffel, MJ Butel, P Langella, LG Bermúdez-Humarán, and AJ Waligora-Dupriet. A new bifidobacteria expression system (best) to produce and deliver interleukin-10 in bifidobacterium bifidum. *Frontiers in microbiology*, 9:3075, 2018.

[52] H Hanchi, W Mottawea, K Sebei, and R Hammami. The genus enterococcus: Between probiotic potential and safety concerns—an update. *Frontiers in microbiology*, 9:1791, 2018.

[53] L Pasotti and S Zucca. Advances and computational tools towards predictable design in biological engineering. *Computational and mathematical methods in medicine*, 2014, 2014.

[54] D Del Vecchio, AJ Dy, and Y Qian. Control theory meets synthetic biology. *Journal of The Royal Society Interface*, 13(120):20160380, 2016.

[55] S Mitchell and A Hoffmann. Identifying noise sources governing cell-to-cell variability. *Current opinion in systems biology*, 8:39–45, 2018.

[56] M Kushwaha and HM Salis. A portable expression resource for engineering cross-species genetic circuits and pathways. *Nature communications*, 6(1):1–11, 2015.

[57] A Costello and AH Badran. Synthetic biological circuits within an orthogonal central dogma. *Trends in Biotechnology*, 2020.

[58] CC Liu, MC Jewett, JW Chin, and CA Voigt. Toward an orthogonal central dogma. *Nature chemical biology*, 14(2):103–106, 2018.

[59] EM Lammens, PI Nikel, and R Lavigne. Exploring the synthetic biology potential of bacteriophages for engineering non-model bacteria. *Nature communications*, 11(1):1–14, 2020.

[60] U Alon. Network motifs: theory and experimental approaches. *Nature Reviews Genetics*, 8(6):450–461, 2007.

[61] S Mangan and U Alon. Structure and function of the feed-forward loop network motif. *Proceedings of the National Academy of Sciences*, 100(21):11980–11985, 2003.

[62] K Temme, R Hill, TH Segall-Shapiro, F Moser, and CA Voigt. Modular control of multiple pathways using engineered orthogonal t7 polymerases. *Nucleic acids research*, 40(17):8773–8781, 2012.

[63] F Ceroni, A Boo, S Furini, TE Gorochowski, O Borkowski, YN Ladak, AR Awan, C Gilbert, GB Stan, and T Ellis. Burden-driven feedback control of gene expression. *Nature methods*, 15(5):387–393, 2018.

[64] HH Huang, Y Qian, and D Del Vecchio. A quasi-integral controller for adaptation of genetic modules to variable ribosome demand. *Nature communications*, 9(1):1–12, 2018.

[65] TH Segall-Shapiro, ED Sontag, and CA Voigt. Engineered promoters enable constant gene expression at any copy number in bacteria. *Nature biotechnology*, 36(4):352, 2018.

[66] SK Aoki, G Lillacci, A Gupta, A Baumschlager, D Schweingruber, and M Khammash. A universal biomolecular integral feedback controller for robust perfect adaptation. *Nature*, 570(7762):533–537, 2019.

[67] L Bleris, Z Xie, D Glass, A Adadey, E Sontag, and Y Benenson. Synthetic incoherent feedforward circuits show adaptation to the amount of their genetic template. *Molecular systems biology*, 7(1):519, 2011.

[68] RD Jones, Y Qian, V Siciliano, B DiAndreth, J Huh, R Weiss, and D Del Vecchio. An endoribonuclease-based feedforward controller for decoupling resource-limited genetic modules in mammalian cells. *bioRxiv*, page 867028, 2020.

[69] LS Qi, MH Larson, LA Gilbert, JA Doudna, JS Weissman, AP Arkin, and WA Lim. Repurposing crispr as an rna-guided platform for sequence-specific control of gene expression. *Cell*, 152(5):1173–1183, 2013.

[70] M Jinek, K Chylinski, I Fonfara, M Hauer, JA Doudna, and E Charpentier. A programmable dual-rna–guided dna endonuclease in adaptive bacterial immunity. *science*, 337(6096):816–821, 2012.

[71] AA Dominguez, WA Lim, and LS Qi. Beyond editing: repurposing crispr–cas9 for precision genome regulation and interrogation. *Nature reviews Molecular cell biology*, 17(1):5, 2016.

[72] SE Clamons and RM Murray. Modeling dynamic transcriptional circuits with crispri. *BioRxiv*, page 225318, 2017.

[73] JH Hu, SM Miller, M Geurts, W Tang, L Chen, N Sun, CM Zeina, X Gao, HA Rees, Z Lin, et al. Evolved cas9 variants with broad pam compatibility and high dna specificity. *Nature*, 556(7699):57–63, 2018.

[74] FA Ran, L Cong, WX Yan, DA Scott, JS Gootenberg, AJ Kriz, B Zetsche, O Shalem, X Wu, KS Makarova, et al. In vivo

genome editing using staphylococcus aureus cas9. *Nature*, 520(7546):186–191, 2015.

[75] L Pasotti, M Bellato, D De Marchi, and P Magni. Mechanistic models of inducible synthetic circuits for joint description of dna copy number and regulatory protein level and cell load. *Processes*, 7(3):119, 2019.

[76] SB Carr, J Beal, and DM Densmore. Reducing dna context dependence in bacterial promoters. *PloS one*, 12(4):e0176013, 2017.

[77] AP Arkin. A wise consistency: engineering biology for conformity and reliability and predictability. *Current opinion in chemical biology*, 17(6):893–901, 2013.

[78] J Ang, E Harris, BJ Hussey, R Kil, and DR McMillen. Tuning response curves for synthetic biology. *ACS synthetic biology*, 2(10):547–567, 2013.

[79] H Trabelsi, M Koch, and JL Faulon. Building a minimal and generalizable model of transcription factor–based biosensors: Showcasing flavonoids. *Biotechnology and bioengineering*, 115(9):2292–2304, 2018.

[80] L Endler, N Rodriguez, N Juty, V Chelliah, C Laibe, C Li, and N Le Novere. Designing and encoding models for synthetic biology. *Journal of The Royal Society Interface*, 6(suppl_4):S405–S417, 2009.

[81] M Carbonell-Ballestero, S Duran-Nebreda, R Montañez, J Macía R Solé, and C Rodríguez-Caso. A bottom-up characterization of transfer functions for synthetic biology designs: lessons from enzymology. *Nucleic acids research*, 42(22):14060–14069, 2014.

[82] Y Mileyko, RI Joh, and JS Weitz. Small-scale copy number variation and large-scale changes in gene expression. *Proceedings of the National Academy of Sciences*, 105(43):16659–16664, 2008.

[83] B Canton, A Labno, and D Endy. Refinement and standardization of synthetic biological parts and devices. *Nature biotechnology*, 26(7):787–793, 2008.

[84] L Pasotti, M Bellato, M Casanova, S Zucca, MGC De Angelis, and P Magni. Re-using biological devices: a model-aided analysis of interconnected transcriptional cascades designed from the bottom-up. *Journal of biological engineering*, 11(1):50, 2017.

[85] A Gábor, AF Villaverde, and JR Banga. Parameter identifiability analysis and visualization in large-scale kinetic models of biosystems. *BMC systems biology*, 11(1):1–16, 2017.

[86] Y Berset, D Merulla, A Joublin, V Hatzimanikatis, and JR Van Der Meer. Mechanistic modeling of genetic circuits for arsr arsenic regulation. *ACS synthetic biology*, 6(5):862–874, 2017.

[87] H Song FK Balagaddé, J Ozaki, CH Collins, M Barnet, FH Arnold, SR Quake, and L You. A synthetic escherichia coli predator–prey ecosystem. *Molecular systems biology*, 4(1):187, 2008.

[88] L Pasotti, M Bellato, N Politi, M Casanova, S Zucca, MG Cusella De Angelis, and P Magni. A synthetic close-loop controller circuit for the regulation of an extracellular molecule by engineered bacteria. *IEEE transactions on biomedical circuits and systems*, 13(1):248–258, 2018.

[89] L Pasotti, M Quattrocelli, D Galli, MG Cusella De Angelis, and P Magni. Multiplexing and demultiplexing logic functions for computing signal processing tasks in synthetic biology. *Biotechnology Journal*, 6(7):784–795, 2011.

[90] Y Qian, HH Huang, JI Jiménez, and D Del Vecchio. Resource competition shapes the response of genetic circuits. *ACS synthetic biology*, 6(7):1263–1272, 2017.

[91] Y Qian and D Del Vecchio. Effective interaction graphs arising from resource limitations in gene networks. In *2015 American Control Conference (ACC)*, pages 4417–4423. IEEE, 2015.

[92] MI Stefan and N Le Novère. Cooperative binding. *PLoS Comput Biol*, 9(6):e1003106, 2013.

[93] S Zucca, L Pasotti, G Mazzini, MG Cusella De Angelis, and P Magni. Characterization of an inducible promoter in different dna copy number conditions. *BMC bioinformatics*, 13(S4):S11, 2012.

[94] DM Colton, EV Stabb, and SJ Hagen. Modeling analysis of signal sensitivity and specificity by vibrio fischeri luxr variants. *PLoS One*, 10(5):e0126474, 2015.

[95] S Jayanthi, KS Nilgiriwala, and D Del Vecchio. Retroactivity controls the temporal dynamics of gene transcription. *ACS synthetic biology*, 2(8):431–441, 2013.

[96] O Borkowski, C Bricio, M Murgiano, B Rothschild-Mancinelli, GB Stan, and T Ellis. Cell-free prediction of protein expression costs for growing cells. *Nature communications*, 9(1):1–11, 2018.

[97] GT Reeves. The engineering principles of combining a transcriptional incoherent feedforward loop with negative feedback. *Journal of biological engineering*, 13(1):62, 2019.

[98] AE Friedland, R Baral, P Singhal, K Loveluck, S Shen, M Sanchez, E Marco, GM Gotta, ML Maeder, EM Kennedy, et al. Characterization of staphylococcus aureus cas9: a smaller cas9 for all-in-one adeno-associated virus delivery and paired nickase applications. *Genome biology*, 16(1):1–10, 2015.

[99] AAK Nielsen and CA Voigt. Multi-input crispr/c as genetic circuits that interface host regulatory networks. *Molecular systems biology*, 10(11):763, 2014.

[100] S Zhang and CA Voigt. Engineered dcas9 with reduced toxicity in bacteria: implications for genetic circuit design. *Nucleic acids research*, 46(20):11115–11125, 2018.

[101] HM Salis, EA Mirsky, and CA Voigt. Automated design of synthetic ribosome binding sites to control protein expression. *Nature biotechnology*, 27(10):946–950, 2009.

[102] S Kosuri, DB Goodman, G Cambray, VK Mutalik, Y Gao, AP Arkin, D Endy, and GM Church. Composability of regulatory sequences controlling transcription and translation in escherichia coli. *Proceedings of the National Academy of Sciences*, 110(34):14024–14029, 2013.

[103] L Pasotti, N Politi, S Zucca, MG Cusella De Angelis, and P Magni. Bottom-up engineering of biological systems through standard bricks: a modularity study on basic parts and devices. *PloS one*, 7(7):e39407, 2012.

[104] C Lou, B Stanton, YJ Chen, B Munsky, and CA Voigt. Ribozyme-based insulator parts buffer synthetic circuits from genetic context. *Nature biotechnology*, 30(11):1137–1142, 2012.

[105] AT Raper, AA Stephenson, and Z Suo. Functional insights revealed by the kinetic mechanism of crispr/cas9. *Journal of the American Chemical Society*, 140(8):2971–2984, 2018.

[106] A Espah Borujeni, AS Channarasappa, and HM Salis. Translation rate is controlled by coupled trade-offs between site accessibility and selective rna unfolding and sliding at upstream standby sites. *Nucleic acids research*, 42(4):2646–2659, 2014.

[107] A Espah Borujeni and HM Salis. Translation initiation is controlled by rna folding kinetics via a ribosome drafting mechanism. *Journal of the American Chemical Society*, 138(22):7016–7023, 2016.

[108] A Espah Borujeni, D Cetnar, I Farasat, A Smith, N Lundgren, and HM Salis. Precise quantification of translation inhibition by mrna structures that overlap with the ribosomal footprint in n-terminal coding sequences. *Nucleic acids research*, 45(9):5437–5448, 2017.

[109] G Hornung and N Barkai. Noise propagation and signaling sensitivity in biological networks: a role for positive feedback. *PLoS Comput Biol*, 4(1):e8, 2008.

[110] M Kaern, TC Elston, WJ Blake, and JJ Collins. Stochasticity in gene expression: from theories to phenotypes. *Nature Reviews Genetics*, 6(6):451–464, 2005.

[111] N Rosenfeld, JW Young, U Alon, PS Swain, and MB Elowitz. Gene regulation at the single-cell level. *science*, 307(5717):1962–1965, 2005.

[112] A Becskei and L Serrano. Engineering stability in gene networks by autoregulation. *Nature*, 405(6786):590–593, 2000.

[113] M Bellato, AF Chiacchiera, E Salibi, M Casanova, D De Marchi, MG Cusella De Angelis, L Pasotti, and P Magni. Crispr interference as low burden logic inverters in synthetic circuits: characterization and tuning. *bioRxiv*, 2020.

[114] VK Mutalik, JC Guimaraes, G Cambray, C Lam, MJ Christoffersen, QA Mai, AB Tran, M Paull, JD Keasling, AP Arkin, et al. Precise and reliable gene expression via standard transcription and translation initiation elements. *Nature methods*, 10(4):354–360, 2013.

[115] S Ohuchi, Y Mori, and Y Nakamura. Evolution of an inhibitory rna aptamer against t7 rna polymerase. *FEBS open bio*, 2:203–207, 2012.

[116] Z Cheng, F Liu, XP Zhang, and W Wang. Robustness analysis of cellular memory in an autoactivating positive feedback system. *FEBS letters*, 582(27):3776–3782, 2008.

[117] DR Burrill and PA Silver. Making cellular memories. *Cell*, 140(1):13–18, 2010.

[118] H Zhao, HM Zhang, X Chen, T Li, Q Wu, Q Ouyang, and GQ Chen. Novel t7-like expression systems used for halomonas. *Metabolic engineering*, 39:128–140, 2017.

[119] X Liang, C Li, W Wang, and Q Li. Integrating t7 rna polymerase and its cognate transcriptional units for a host-independent and stable expression system in single plasmid. *ACS synthetic biology*, 7(5):1424–1435, 2018.

[120] SI Tan and IS Ng. New insight into plasmid-driven t7 rna polymerase in escherichia coli and use as a genetic amplifier for a biosensor. *ACS Synthetic Biology*, 9(3):613–622, 2020.

[121] Q Yan and SS Fong. Challenges and advances for genetic engineering of non-model bacteria and uses in consolidated bioprocessing. *Frontiers in microbiology*, 8:2060, 2017.

[122] KS Murakami and SA Darst. Bacterial rna polymerases: the wholo story. *Current opinion in structural biology*, 13(1):31–39, 2003.

[123] L López-Maury, S Marguerat, and J Bähler. Tuning gene expression to changing environments: from rapid responses to evolutionary adaptation. *Nature Reviews Genetics*, 9(8):583–593, 2008.

[124] AM Huerta and J Collado-Vides. Sigma70 promoters in escherichia coli: specific transcription in dense regions of overlapping promoter-like signals. *Journal of molecular biology*, 333(2):261–278, 2003.

[125] JE Mitchell, D Zheng, SW Busby, and SD Minchin. Identification and analysis of 'extended–10'promoters in escherichia coli. *Nucleic acids research*, 31(16):4689–4695, 2003.

[126] C Yang, AJ Hockenberry, MC Jewett, and LAN Amaral. Depletion of shine-dalgarno sequences within bacterial coding regions is expression dependent. *G3: Genes and Genomes and Genetics*, 6(11):3467–3474, 2016.

[127] A Feklistov, BD Sharon, SA Darst, and CA Gross. Bacterial sigma factors: a historical and structural and genomic perspective. *Annual review of microbiology*, 68:357–376, 2014.

[128] MD Engstrom and BF Pfleger. Transcription control engineering and applications in synthetic biology. *Synthetic and systems biotechnology*, 2(3):176–191, 2017.

[129] MR Amin, A Yurovsky, Y Chen, S Skiena, and B Futcher. Re-annotation of 12 and495 prokaryotic 16s rrna 3'ends and analysis of shine-dalgarno and anti-shine-dalgarno sequences. *PloS one*, 13(8):e0202767, 2018.

[130] D Omotajo, T Tate, H Cho, and M Choudhary. Distribution and diversity of ribosome binding sites in prokaryotic genomes. *BMC genomics*, 16(1):604, 2015.

[131] A de Jong, H Pietersma, M Cordes, OP Kuipers, J Kok, and Jan. Pepper: a webserver for prediction of prokaryote promoter elements and regulons. *BMC genomics*, 13(1):299, 2012.

[132] VSA Salamov and A Solovyevand. Automatic annotation of microbial genomes and metagenomic sequences. *Metagenomics and its applications in agriculture and biomedicine and environmental studies. Hauppauge: Nova Science Publishers*, pages 61–78, 2011.

[133] W He, C Jia, Y Duan, and Q Zou. 70propred: a predictor for discovering sigma70 promoters based on combining multiple features. *BMC systems biology*, 12(4):44, 2018.

[134] SDA e Silva, S Echeverrigaray, and GJL Gerhardt. Bacpp: bacterial promoter prediction—a tool for accurate sigma-factor specific assignment in enterobacteria. *Journal of theoretical biology*, 287:92–99, 2011.

[135] RK Umarov and VV Solovyev. Recognition of prokaryotic and eukaryotic promoters using convolutional deep learning neural networks. *PloS one*, 12(2):e0171410, 2017.

[136] L Coppens and R Lavigne. Sapphire: a neural network based classifier for σ70 promoter prediction in pseudomonas. *BMC bioinformatics*, 21(1):1–7, 2020.

[137] M Di Salvo, E Pinatel, M Fondi A Talà, C Peano, and A Alifano. G4promfinder: an algorithm for predicting transcription promoters in gc-rich bacterial genomes based on at-rich elements and g-quadruplex motifs. *BMC bioinformatics*, 19(1):36, 2018.

[138] V Rangannan and M Bansal. Relative stability of dna as a generic criterion for promoter prediction: whole genome annotation of microbial genomes with varying nucleotide base composition. *Molecular bioSystems*, 5(12):1758–1769, 2009.

[139] M Tompa. An exact method for finding short motifs in sequences and with application to the ribosome binding site problem. In *ISMB*, volume 99, pages 262–271, 1999.

[140] LM Maurer, E Yohannes, SS Bondurant, M Radmacher, and JL Slonczewski. ph regulates genes for flagellar motility and catabolism and oxidative stress in escherichia coli k-12. *Journal of bacteriology*, 187(1):304–319, 2005.

[141] ET Hayes, JC Wilks, P Sanfilippo, E Yohannes, DP Tate, BD Jones, MD Radmacher, SS BonDurant, and JL Slonczewski. Oxygen limitation modulates ph regulation of catabolism and hydrogenases and multidrug transporters and envelope composition in escherichia coli k-12. *BMC microbiology*, 6(1):89, 2006.

[142] MW Covert, EM Knight, JL Reed, MJ Herrgard, and BO Palsson. Integrating high-throughput and computational data elucidates bacterial networks. *Nature*, 429(6987):92–96, 2004.

[143] S Saito, S Aburatani, and K Horimoto. Network evaluation from the consistency of the graph structure with the measured data. *BMC systems biology*, 2(1):84, 2008.

[144] Y Asakura, H Kojima, and I Kobayashi. Evolutionary genome engineering using a restriction–modification system. *Nucleic acids research*, 39(20):9034–9046, 2011.

[145] SE Cohen, CA Lewis, RA Mooney, MA Kohanski, JJ Collins, R Landick, and GC Walker. Roles for the transcription elongation factor nusa in both dna repair and damage tolerance pathways in escherichia coli. *Proceedings of the National Academy of Sciences*, 107(35):15517–15522, 2010.

[146] P Berger, IU Kouzel, M Berger, N Haarmann, U Dobrindt, GB Koudelka, and A Mellmann. Carriage of shiga toxin phage profoundly affects escherichia coli gene expression and carbon source utilization. *BMC genomics*, 20(1):1–14, 2019.

[147] RT Veetil, N Malhotra, A Dubey, and ASN Seshasayee. Laboratory evolution experiments help identify a predominant region of constitutive stable dna replication initiation. *Msphere*, 5(1), 2020.

[148] A Anand, K Chen, L Yang, AV Sastry, CA Olson, S Poudel, Y Seif, Y Hefner, PV Phaneuf, S Xu, et al. Adaptive evolution reveals a tradeoff between growth rate and oxidative stress during naphthoquinone-based aerobic respiration. *Proceedings of the National Academy of Sciences*, 116(50):25287–25292, 2019.

[149] D Kim, SW Seo, Y Gao, H Nam, GI Guzman, BK Cho, and BO Palsson. Systems assessment of transcriptional regulation on central carbon metabolism by cra and crp. *Nucleic acids research*, 46(6):2901–2917, 2018.

[150] SW Seo, D Kim, EJ O'Brien, R Szubin, and BO Palsson. Decoding genome-wide gadewx-transcriptional regulatory networks reveals multifaceted cellular responses to acid stress in escherichia coli. *Nature communications*, 6(1):1–8, 2015.

[151] E Dzyubak and MNF Yap. The expression of antibiotic resistance methyltransferase correlates with mrna stability independently of ribosome stalling. *Antimicrobial agents and chemotherapy*, 60(12):7178–7188, 2016.

[152] T Winter, J Winter, M Polak, K Kusch, U Mäder, R Sietmann, J Ehlbeck, S van Hijum, KD Weltmann, M Hecker, et al. Characterization of the global impact of low temperature gas plasma on vegetative microorganisms. *Proteomics*, 11(17):3518–3530, 2011.

[153] P Zuber, S Chauhan, P Pilaka, MM Nakano, S Gurumoorthy, AA Lin, SM Barendt, BK Chi, H Antelmann, and U Mäder. Phenotype enhancement screen of a regulatory spx mutant unveils a role for the ytpq gene in the control of iron homeostasis. *PLoS One*, 6(9):e25066, 2011.

[154] M Lehnik-Habrink, M Schaffer, U Mäder, C Diethmaier, C Herzberg, and J Stülke. Rna processing in bacillus subtilis: identification of targets of the essential rnase y. *Molecular microbiology*, 81(6):1459–1473, 2011.

[155] AKW Elsholz, K Turgay, S Michalik, B Hessling, K Gronau, D Oertel, U Mäder, J Bernhardt, D Becher, M Hecker, et al. Global impact of protein arginine phosphorylation on the physiology of bacillus subtilis. *Proceedings of the National Academy of Sciences*, 109(19):7451–7456, 2012.

[156] T Winter, J Bernhardt, J Winter, U Mäder, R Schlüter, KD Weltmann, M Hecker, and H Kusch. Common versus noble b acillus subtilis differentially responds to air and argon gas plasma. *Proteomics*, 13(17):2608–2621, 2013.

[157] M Kohlstedt, PK Sappa, H Meyer, A Zaprasis S Maaß, T Hoffmann, J Becker, L Steil, M Hecker, M van Dijl, et al. Adaptation of b acillus subtilis carbon core metabolism to simultaneous nutrient limitation and osmotic challenge: a multi-omics perspective. *Environmental microbiology*, 16(6):1898–1917, 2014.

[158] M Leroy, J Piton, L Gilet, O Pellegrini, C Proux, JY Coppée, S Figaro, and C Condon. Rae1/yacp, a new endoribonuclease involved in ribosome-dependent mrna decay in bacillus subtilis. *The EMBO journal*, 36(9):1167–1181, 2017.

[159] P Nicolas, U Mäder, E Dervyn, T Rochat, A Leduc, N Pigeonneau, E Bidnenko, E Marchadier, M Hoebeke, S Aymerich, et al. Condition-dependent transcriptome reveals high-level regulatory architecture in bacillus subtilis. *Science*, 335(6072):1103–1106, 2012.

[160] B Liu, G Deikus, A Bree, S Durand, DB Kearns, and DH Bechhofer. Global analysis of mrna decay intermediates in b acillus subtilis wild-type and polynucleotide phosphorylase-deletion strains. *Molecular microbiology*, 94(1):41–55, 2014.

[161] K Surdova, P Gamba, D Claessen, T Siersma, MJ Jonker, J Errington, and LW Hamoen. The conserved dna-binding protein whia is involved in cell division in bacillus subtilis. *Journal of bacteriology*, 195(24):5450–5460, 2013.

[162] KR Hummels and DB Kearns. Suppressor mutations in ribosomal proteins and fliy restore bacillus subtilis swarming motility in the absence of ef-p. *PLoS genetics*, 15(6):e1008179, 2019.

[163] MD Robinson and TP Speed. A comparison of affymetrix gene expression arrays. *BMC bioinformatics*, 8(1):449, 2007.

[164] L Gautier, L Cope, BM Bolstad, and RA Irizarry. affy—analysis of affymetrix genechip data at the probe level. *Bioinformatics*, 20(3):307–315, 2004.

[165] MN McCall, PN Murakami, M Lukk, W Huber, and RA Irizarry. Assessing affymetrix genechip microarray quality. *BMC bioinformatics*, 12(1):137, 2011.

[166] RA Irizarry, B Hobbs, F Collin, YD Beazer-Barclay, KJ Antonellis, U Scherf, and TP Speed. Exploration and normalization and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249–264, 2003.

[167] M Zahurak, G Parmigiani, W Yu, RB Scharpf, D Berman, E Schaeffer, S Shabbeer, and L Cope. Pre-processing agilent microarray data. *BMC bioinformatics*, 8(1):142, 2007.

[168] GK Smyth, K Gordon, and T Speed. Normalization of cdna microarray data. *Methods*, 31(4):265–273, 2003.

[169] A Oshlack, D Emslie, LM Corcoran, and GK Smyth. Normalization of boutique two-color microarrays with a high proportion of differentially expressed probes. *Genome biology*, 8(1):1–8, 2007.

[170] Y Pan, L Bodrossy, P Frenzel, AG Hestnes, S Krause, C Lüke, M Meima-Franke, H Siljanen, MM Svenning, and PLE Bodelier. Impacts of inter-and intralaboratory variations on the re-

producibility of microbial community analyses. *Applied and environmental microbiology*, 76(22):7451–7458, 2010.

[171] SSJ Heng, OYW Chan, BMH Keng, and MHT Ling. Glucan biosynthesis protein g is a suitable reference gene in escherichia coli k-12. *ISRN microbiology*, 2011, 2011.

[172] J Zeng and S Spiro. Finely tuned regulation of the aromatic amine degradation pathway in escherichia coli. *Journal of bacteriology*, 195(22):5141–5150, 2013.

[173] E Ibañez, R Gimenez, T Pedraza, L Baldoma, J Aguilar, and J Badia. Role of the yiar and yiasgenes of escherichia coli in metabolism of endogenously formed l-xylulose. *Journal of Bacteriology*, 182(16):4625–4627, 2000.

[174] T Stratmann, S Madhusudan, and K Schnetz. Regulation of the yjjq-bglj operon, encoding luxr-type transcription factors, and the divergent yjjp gene by h-ns and leuo. *Journal of bacteriology*, 190(3):926–935, 2008.

[175] WK Ray and TJ Larson. Application of agar repressor and dominant repressor variants for verification of a gene cluster involved in n-acetylgalactosamine metabolism in escherichia coli k-12. *Molecular microbiology*, 51(3):813–826, 2004.

[176] E Campos, L Baldoma, J Aguilar, and J Badia. Regulation of expression of the divergent ulag and ulaabcdef operons involved in l-ascorbate dissimilation in escherichia coli. *Journal of bacteriology*, 186(6):1720–1728, 2004.

[177] O Francetic, C Badaut, S Rimsky, and AP Pugsley. The chia (yheb) protein of escherichia coli k-12 is an endochitinase whose

gene is negatively controlled by the nucleoid-structuring protein h-ns. *Molecular microbiology*, 35(6):1506–1517, 2000.

[178] K Eichler, F Bourgis, A Buchet, HP Kleber, and MA Mandrand-Berthelot. Molecular characterization of the cai operon necessary for carnitine metabolism in escherichia coii. *Molecular microbiology*, 13(5):775–786, 1994.

[179] A Ferrandez, B Minambres, B Garcia, ER Olivera, JM Luengo, JL Garcia, and E Diaz. Catabolism of phenylacetic acid in escherichia coli characterization of a new aerobic hybrid pathway. *Journal of Biological Chemistry*, 273(40):25974–25986, 1998.

[180] B Troup, C Hungerer, and D Jahn. Cloning and characterization of the escherichia coli hemn gene encoding the oxygen-independent coproporphyrinogen iii oxidase. *Journal of bacteriology*, 177(11):3326–3331, 1995.

[181] KI Sorensen and B Hove-Jensen. Ribose catabolism of escherichia coli: characterization of the rpib gene encoding ribose phosphate isomerase b and of the rpir gene, which is involved in regulation of rpib expression. *Journal of bacteriology*, 178(4):1003–1011, 1996.

[182] JT Wade, DC Roa, DC Grainger, D Hurd, SJW Busby, K Struhl, and E Nudler. Extensive functional overlap between σ factors in escherichia coli. *Nature structural & molecular biology*, 13(9):806–814, 2006.

[183] S Lindquist, K Weston-Hafer, H Schmidt, C Pul, G Korfmann, J Erickson, C Sanders, HH Martin, and S Normark. Ampg, a signal transducer in chromosomal β-lactamase induction. *Molecular microbiology*, 9(4):703–715, 1993.

[184] F Claverie-Martin, MR Diaz-Torres, and SR Kushner. Analysis of the regulatory region of the protease iii (ptr) gene of escherichia coli k-12. *Gene*, 54(2-3):185–195, 1987.

[185] DB Lim, JD Oppenheim, T Eckhardt, and WK Maas. Nucleotide sequence of the argr gene of escherichia coli k-12 and isolation of its product, the arginine repressor. *Proceedings of the National Academy of Sciences*, 84(19):6697–6701, 1987.

[186] H Ogasawara, J Teramoto, S Yamamoto, K Hirao, K Yamamoto, A Ishihama, and R Utsumi. Negative regulation of dna repair gene (uvra) expression by arca/arcb two-component system in escherichia coli. *FEMS microbiology letters*, 251(2):243–249, 2005.

[187] E Krin, A Danchin, and O Soutourina. Rcsb plays a central role in h-ns-dependent regulation of motility and acid stress resistance in escherichia coli. *Research in microbiology*, 161(5):363–371, 2010.

[188] I Loubens, L Debarbieux, A Bohin, JM Lacroix, and JP Bohin. Homology between a genetic locus (mdoa) involved in the osmoregulated biosynthesis of periplasmic glucans in escherichia coli and a genetic locus (hrpm) controlling pathogenicity of pseudomonas syringae. *Molecular microbiology*, 10(2):329–340, 1993.

[189] S Kauppinen, M Siggaard-Andersen, and P von Wettstein-Knowles. β-ketoacyl-acp synthase i of escherichia coli: Nucleotide sequence of the fabb gene and identification of the cerulenin binding residue. *Carlsberg research communications*, 53(6):357–370, 1988.

[190] M Serizawa and J Sekiguchi. The bacillus subtilis ydfhi two-component system regulates the transcription of ydfj, a member of the rnd superfamily. *Microbiology*, 151(6):1769–1778, 2005.

[191] NN Baranova, A Danchin, and AA Neyfakh. Mta, a global merr-type regulator of the bacillus subtilis multidrug-efflux transporters. *Molecular microbiology*, 31(5):1549–1559, 1999.

[192] M Ogura, K Tsukahara, K Hayashi, and T Tanaka. The bacillus subtilis natk–natr two-component system regulates expression of the natab operon encoding an abc transporter for sodium ion extrusion. *Microbiology*, 153(3):667–675, 2007.

[193] LM Robson and GH Chambliss. Endo-beta-1, 4-glucanase gene of bacillus subtilis dlg. *Journal of bacteriology*, 169(5):2017–2025, 1987.

[194] JCR Struck, RKK Far, W Schröder, F Hucho, HY Toschka, and VA Erdmann. Characterization of a 17 kda protein gene upstream from the small cytoplasmic rna gene of bacillus subtilis. *Biochimica et Biophysica Acta (BBA)-Gene Structure and Expression*, 1050(1-3):80–83, 1990.

[195] S Jin and AL Sonenshein. Transcriptional regulation of bacillus subtilis citrate synthase genes. *Journal of bacteriology*, 176(15):4680–4690, 1994.

[196] F Morohoshi, K Hayashi, and N Munakata. Bacillus subtilis gene coding for constitutive o 6-methylguanine-dna alkyttransferase. *Nucleic acids research*, 17(16):6531–6543, 1989.

[197] E Darmon, D Noone, A Masson, S Bron, OP Kuipers, KM Devine, and JM van Dijl. A novel class of heat and secretion stress-responsive genes is controlled by the autoregulated cssrs

two-component system of bacillus subtilis. *Journal of Bacteriology*, 184(20):5661–5671, 2002.

[198] YI Zhang, SV Taylor, HJ Chiu, and TP Begley. Characterization of the bacillus subtilis thic operon involved in thiamine biosynthesis. *Journal of bacteriology*, 179(9):3030–3035, 1997.

[199] R Gardan, G Rapoport, and M Débarbouillé. Expression of therocdefoperon involved in arginine catabolism inbacillus subtilis. *Journal of molecular biology*, 249(5):843–856, 1995.

[200] KI Yoshida, Y Fujita, and SD Ehrlich. An operon for a putative atp-binding cassette transport system involved in acetoin utilization ofbacillus subtilis. *Journal of bacteriology*, 182(19):5454–5461, 2000.

[201] G Zheng, LZ Yan, JC Vederas, and P Zuber. Genes of the sbo-alb locus ofbacillus subtilis are required for production of the antilisterial bacteriocin subtilosin. *Journal of Bacteriology*, 181(23):7346–7355, 1999.

[202] M Fuangthong, AF Herbig, N Bsat, and JD Helmann. Regulation of the bacillus subtilis fur and perr genes by perr: not all members of the perr regulon are peroxide inducible. *Journal of bacteriology*, 184(12):3276–3286, 2002.

[203] K Trach, JW Chapman, P Piggot, D LeCoq, and JA Hoch. Complete sequence and transcriptional analysis of the spo0f region of the bacillus subtilis chromosome. *Journal of bacteriology*, 170(9):4194–4208, 1988.

[204] A de Saizieu, P Vankan, C Vockler, and APGM van Loon. The trp rna-binding attenuation protein (trap) regulates the steady-state levels of transcripts of the bacillus subtilis folate operon. *Microbiology*, 143(3):979–989, 1997.

[205] Y Weinrauch, N Guillen, and DA Dubnau. Sequence and transcription mapping of bacillus subtilis competence genes comb and coma, one of which is related to a family of bacterial regulatory determinants. *Journal of bacteriology*, 171(10):5362–5375, 1989.

[206] H Yamamoto, M Mori, and J Sekiguchi. Transcription of genes near the sspe locus of the bacillus subtilis genome. *Microbiology*, 145(8):2171–2180, 1999.

[207] S Goodwin, JD McPherson, and WR McCombie. Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, 17(6):333, 2016.

[208] R Milo. What is the total number of protein molecules per cell volume? a call to rethink some published values. *Bioessays*, 35(12):1050–1055, 2013.

[209] S Zucca, L Pasotti, N Politi, M Casanova, G Mazzini, MG Cusella De Angelis, and P Magni. Multi-faceted characterization of a novel luxr-repressible promoter library for escherichia coli. *PLoS One*, 10(5):e0126264, 2015.

[210] N Savic, FCAS Ringnalda, H Lindsay, C Berk, K Bargsten, Y Li, D Neri, MD Robinson, C Ciaudo, J Hall, et al. Covalent linkage of the dna repair template to the crispr-cas9 nuclease enhances homology-directed repair. *Elife*, 7:e33761, 2018.

[211] H Jeong, V Barbe, CH Lee, D Vallenet, DS Yu, SH Choi, A Couloux, SW Lee, SH Yoon, L Cattolico, et al. Genome sequences of escherichia coli b strains rel606 and bl21 (de3). *Journal of molecular biology*, 394(4):644–652, 2009.

[212] I Massaiu, L Pasotti, M Casanova, N Politi, S Zucca, MG Cusella De Angelis, and P Magni. Quantification of the gene silencing performances of rationally-designed synthetic small rnas. *Systems and synthetic biology*, 9(3):107–123, 2015.

List of publications

Articles in peer reviewed journals

- L Pasotti, **D De Marchi**, M Casanova, I Massaiu, M Bellato, MG Cusella De Angelis, C Calvio, P Magni. *Engineering endogenous fermentative routes in ethanologenic Escherichia coli W for bioethanol production from concentrated whey permeate.* New Biotechnology, 57: 55-66, 2020.

- M Bellato, **D De Marchi**, C Gualtieri, E Sauta, P Magni, A Macovei, L Pasotti. *A Bioinformatics Approach to Explore MicroRNAs as Tools to Bridge Pathways Between Plants and Animals. Is DNA Damage Response (DDR) a Potential Target Process?* Frontiers in Plant Science, 10: 1535, 2019.

- L Pasotti, M Bellato, **D De Marchi**, P Magni. *Mechanistic models of inducible synthetic circuits for the joint description of DNA copy number, regulatory protein level and cell load.* Processes, 7(3):119, 2019.

Contributions to conference proceedings

- **D De Marchi**, L Pasotti, M Casanova, I Massaiu, MG Cusella De Angelis, P Magni. *A metabolic engineering approach to optimize ethanol production from dairy waste in Escherichia coli.* Congresso Gruppo Nazionale di Bioingegneria - Atti. Milan, Italy, 2018.

- **D De Marchi**, L Pasotti, M Casanova, I Massaiu, MG Cusella De Angelis, P Magni. *A metabolic engineering approach to improve ethanol production from dairy waste by optimized Escherichia coli.* Proceeding of the 5th International Synthetic and System Biology Summer School (SSBSS). Siena, Italy, 2018.

- **D De Marchi**, L Pasotti, M Bellato, P Magni. *Beyond Hill Equations: Mechanistic Modeling of Inducible Systems to Expand the Predictability of Synthetic Circuits.* Proceeding of SEED 2019. New York, USA, 2018.

- M Bellato, **D De Marchi**, L Pasotti, A Macovei, C Gualtieri, A Balestrazzi, P Magni. *A bioinformatic approach to predict the cross-kingdom potential of plant miRNA in humans.* Proceeding of BITS 2018. Turin, Italy, 2018.

- C Gualtieri, **D De Marchi**, M Bellato, E. Sauta, L Pasotti, P Magni, A Balestrazzi, A Macovei. "In silico evidence of microRNA-mediated cross-kingdom regulation of genes involved in the cellular response to viruses in Medicago truncatula and humans". Proceeding of 3rd iPLANTA Conference. Lisbona, Portugal, 2019.

- I Massaiu, L Pasotti, **D De Marchi**, P Magni. *Evaluation of mathematical methods to in-silico study the metabolic phenotype*

of E. coli. Congresso Gruppo Nazionale di Bioingegneria - Atti. Milan, Italy, 2018.

- L. Pasotti, I. Massaiu, E. Rama, **D. De Marchi**, M. Cavalletti, M. Casanova, M. Bellato, A. Zebre, G. Mazzini, M.G. Cusella De Angelis, C. Calvio and P. Magni. *In-silico and in-vivo metabolic engineering strategied to optimize waste bioconversion pathways in microbial strains: two case studies on ethanol and poly-gamma-glutamic acid biosynthesis.* Proceeding of SEED 2018. Scottsdale, Arizona, 2018.

- A Macovei, C Gualtieri, M Bellato, **D De Marchi**, E Sauta, L Pasotti, P Magni, A Balestrazzi. *MicroRNA as 'communication molecules' between plants and animals.* Proceeding of BIOCONTROL2019. Viterbo, Italy, 2019.

- A Frusteri Chiacchera, L Pasotti, M Bellato, E Salibi, M Casanova, **D De Marchi**, MG Cusella De Angelis, P Magni. *Approaches to optimize CRISPRi based gene regulation in E. coli.* Proceeding of the 6th International Synthetic and System Biology Summer School (SSBSS). Pisa, Italy, 2019.