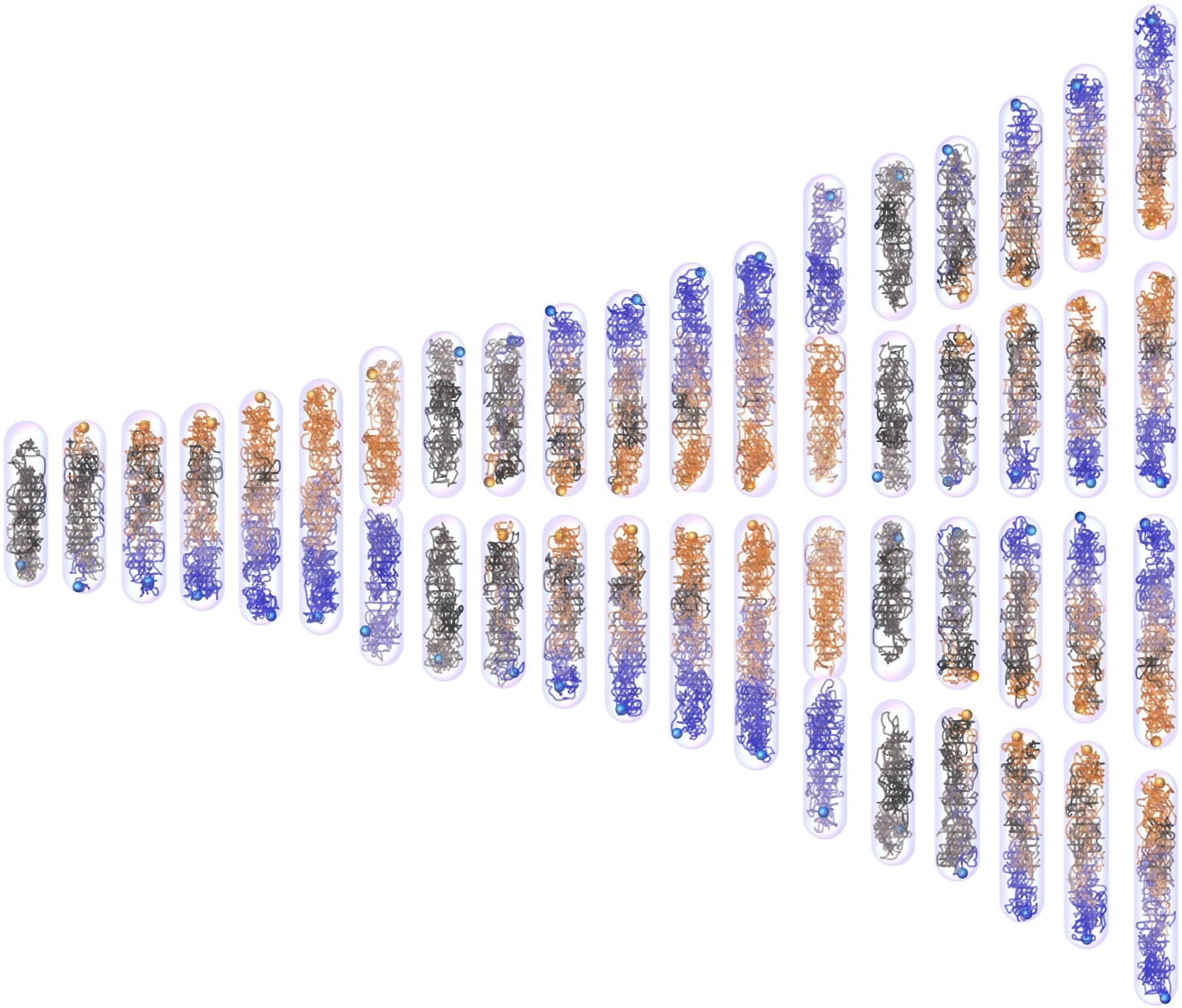
# Bacterial chromosome organization and cell size through space and time

**Joris J.B. Messelink**



# Bacterial chromosome organization and cell size through space and time

**Joris J.B. Messelink**

A dissertation

submitted to the Faculty of Physics

at the Ludwig–Maximilians–Universität München

for the degree of

Doctor rerum naturalium


Munich, 04.02.2022

First referee: Prof. Dr. Chase Broedersz
Second referee: Prof. Dr. Erwin Frey
Day of the oral examination: 16th March 2022

# Zusammenfassung **(German summary)**

Ein universelles Merkmal des Lebens ist die Fähigkeit, Informationen zu speichern und sie an zukünftige Generationen weiterzugeben. Dieser genomische Code, der in der DNS gespeichert ist, enthält alle Informationen, die zur Erzeugung einer lebenden Zelle erforderlich sind. Beim Ablesen des genomischen Codes ist die DNS jedoch kein passiver Informationsträger; ihre Organisation in Form eines gefalteten Chromosoms ist wesentlich für die Regulierung der Gentranskription. Bei der Weitergabe der genetischen Information an die nächsten Generationen muss die Replikation der DNS wiederum eng mit dem Wachstum und der Teilung der Zellen koordiniert werden, um eine getreue Vererbung sicherzustellen. In dieser Arbeit untersuchen wir zwei Aspekte dieser Organisation des Lebens in Bakterien: die räumliche Organisation der Chromosomen und das Zellwachstum. Diese Arbeit ist wie folgt gegliedert:

**Kapitel 1 - Einleitung**
In diesem Kapitel werden grundlegende Konzepte und Methoden für die Untersuchung der Chromosomenorganisation und des Zellwachstums eingeführt. Wir gehen auf die wichtigsten zellulären Komponenten ein, die an der bakteriellen Chromosomenorganisation beteiligt sind, und zeigen, wie Hi-C-Daten detaillierte Informationen über diese Organisation in Form von Zweipunkt-Kontakthäufigkeiten liefern. Anschließend wird das Konzept der maximalen Entropie vorgestellt, und die grundlegenden Gesetze des bakteriellen Einzelzellwachstums diskutiert. Abschließend werden die Ziele und die Bedeutung dieser Arbeit dargelegt.

**Kapitel 2**
*Mit Muriel van Teeseling, Jacqueline Janssen, Martin Thanbichler und Chase Broedersz.*
In diesem Kapitel entwickeln wir ein Maximum-Entropie-Modell (MaxEnt), um die vollständige Verteilung der bakteriellen Chromosomenkonfigurationen in einer Zelle zu inferieren. Der Modellinput besteht aus Hi-C-Daten von *Caulobacter crescentus*, die detaillierte Informationen über die räumliche Chromosomenorganisation in Form von Kontakthäufigkeiten zwischen Paaren genomischer Regionen darstellen. Das Modell wird durch unabhängige Fluoreszenzmikroskopie-Experimente validiert und zeigt die chromosomale Struktur über genomische Skalen hinweg. Auf großen Skalen finden wir ein auffälliges Muster von Positionskorrelationen entlang der langen Zellachse. Dieses Korrelationsmuster wird durch das Vorhandensein großer genomischer Cluster, so genannter Superdomänen, erklärt, die dazu neigen, sich gegenseitig entlang der langen Zellachse räumlich auszuschließen, wenn sie auf gegenüberliegenden Chromosomenarmen liegen. Auf kleineren genomischen Skalen finden wir ein Muster lokaler Streckungen, das teilweise mit der Aktivität der Gentranskription korreliert. Schließlich quantifizieren wir die von jeder genomischen Region kodierte Lokalisierungsinformation, die von der Zelle als zelluläre Straßenkarte für die Positionierung von Proteinen und Proteintröpfchen verwendet werden könnte.

**Kapitel 3**
*Mit Chase P. Broedersz, Grzegorz Gradziuk, Janni Harju, Imesha R. Mudiyanselage, Muriel C.F. van Teeseling und Lucas Tröger.*
Aufbauend auf dem vorangegangenen Kapitel erweitern wir das MaxEnt-Chromosomenmodell, um die räumliche Organisation von sich replizierenden Chromosomen zu beschreiben. Durch

die Kombination von Hi-C-Daten und Fluoreszenzmikroskopie für eine Reihe von Zeitpunkten während des Zellzyklus erhalten wir ein stroboskopisches Modell der Chromosomenorganisation im Laufe der Zeit. Wir validieren die Vorhersagekraft des sich replizierenden MaxEnt-Modells mit Fluoreszenzmikroskopie-Experimenten. Das replizierende MaxEnt-Modell zeigt eine lineare Organisation, die während des gesamten Replikationszyklus über die Chromosomensegmente hinweg bestehen bleibt. Ein Modell, das nur Beschränkungen für die Position der Replikationsursprung (*ori*) enthält, deutet darauf hin, dass die lineare Organisation des replizierten Chromosoms hauptsächlich auf das Ziehen der replizierten *ori* zu den entgegengesetzten Polen zurückzuführen ist. Die lineare Organisation des nicht replizierten Chromosoms wird durch diesen Mechanismus jedoch nicht erklärt. Das replizierende MaxEnt-Modell bietet Zugang zu einer Vielzahl von weiteren organisatorischen Merkmalen während der Replikation, von denen einige am Ende des Kapitels hervorgehoben werden.

**Kapitel 4**

*Mit Fabian Meyer, Marc Bramkamp und Chase P. Broedersz.*

In diesem Kapitel wechseln wir die Perspektive von der chromosomalen Organisation während des Zellwachstums zu den Dimensionen der gesamten Zelle während dieses Prozesses. Das Zellwachstum ist eng mit der Chromosomenreplikation verbunden, und die Regulierung des Wachstums ist erforderlich, um eine stabile Zellgrößenstatistik zu gewährleisten. Hier untersuchen wir das bakterielle Einzelzellwachstum in *Corynebacterium glutamicum*, das aufgrund seiner atypischen Wachstumsmechanismen ein vielversprechender Kandidat ist, um gängige Wachstumsmuster zu hinterfragen. Wir entwickeln ein Inferenzverfahren, um aus detaillierten Mikroskopiedaten trotz des Rauschens und der intrinsischen Variabilität des Einzelzellwachstums durchschnittliche Trajektorien der Einzelzellwachstum zu extrahieren. Wir stellen fest, dass die mittleren Wachstumskurven von dem üblicherweise anzutreffenden exponentiellen Einzelzellwachstum abweichen; vielmehr erhöhen die Zellen anfangs ihre Wachstumsrate, wechseln aber später auf ein lineares Wachstum. Dieses asymptotisch lineare Wachstum steht im Einklang mit einem Modell, bei dem der Mechanismus der apikalen Zellwandbildung der ratenbegrenzende Schritt für das Wachstum ist. Schließlich zeigen wir anhand von Simulationen des Populationswachstums, dass das asymptotisch lineare Wachstum als Regulator der Zellgröße fungiert, was eine evolutionäre Erklärung für das Fehlen vieler gängiger Mechanismen zur Wachstumsregulierung in diesem Bakterium nahelegt.

# Summary

A universal feature of life is the ability to store information and pass it on to next generations. This genomic code, stored in DNA, contains all information needed to generate a living cell. In the process of reading out the genomic code, DNA is not a passive container of information however, but the organization of DNA into a folded chromosome is intrinsic to how gene transcription is regulated. In the passing on of genetic information to next generations, the replication of DNA in turn needs to be tightly coordinated with cellular growth and division to ensure faithful inheritance. In this thesis, we explore two aspects of this organization of life in bacteria: the spatial organization of chromosomes, and cellular growth. This thesis is organized as follows:

**Chapter 1 - Introduction**
Here, we introduce basic concepts and methods for the study of chromosome organization and cellular growth in this thesis. We review the main cellular components involved in bacterial chromosome organization, and see how Hi-C data provides detailed information on this organization in the form of two-point contact frequencies. Next, the concept of maximum entropy inference is introduced, and the laws of bacterial single-cell growth are discussed. Lastly, the goals and significance of this thesis are laid out.

**Chapter 2**
*With Muriel van Teeseling, Jacqueline Janssen, Martin Thanbichler and Chase Broedersz.*
In this chapter we develop a Maximum Entropy (MaxEnt) model to learn the full distribution of single-cell bacterial chromosome configurations. The model input consists of Hi-C data on *Caulobacter crescentus*, which provides detailed information on spatial chromosome organization in the form of contact frequencies between pairs of genomic regions. The model is validated by previous independent fluorescence microscopy experiments, and reveals chromosomal structure across genomic scales. At large scales, we find a striking pattern of positional correlations along the long cell axis. This correlation pattern is explained by the presence of large genomic clusters, termed Super Domains, which tend to spatially exclude each other along the long cell axis if they lie on opposite chromosomal arms. At smaller genomic scales, we find a pattern of local extensions that partially correlates with gene transcription activity. Lastly, we quantify the localization information encoded by each genomic region, which could be used by the cell as a cellular road map for positioning proteins and protein droplets.

**Chapter 3**
*With Chase P. Broedersz, Grzegorz Gradziuk, Janni Harju, Imesha R. Mudiyanselage, Muriel C.F. van Teeseling and Lucas Tröger.*
Building upon the previous chapter, we expand the MaxEnt chromosome model to describe the spatial organization of replicating chromosomes. Combining Hi-C data and fluorescence microscopy for a series of time points throughout the cell cycle, we obtain a stroboscopic model of chromosome organization over time. We validate the predictive power of the replicating MaxEnt model with fluorescence microscopy experiments. The replicating MaxEnt model reveals a linear organization that persists across chromosomal segments throughout the replication cycle. A model containing only constraints on the origin of replication (*ori*)

position suggests that the linear organization of the replicated chromosome is mainly due to the pulling of replicated *ori*'s to opposite poles. The linear organization of the unreplicated chromosome is not explained by this mechanism however. The replicating MaxEnt model provides access to a wealth of further organizational features during replication, a few of which are highlighted at the end of the chapter.

**Chapter 4**

*With Fabian Meyer, Marc Bramkamp and Chase P. Broedersz.*

In this chapter, we shift perspective from chromosomal organization during cellular growth, to the dimensions of the entire cell during this process. Cellular growth is intimately linked to chromosome replication, and regulation of growth is required to ensure stable cell size statistics. Here, we study single-cell bacterial growth in *Corynebacterium glutamicum*, which is a promising candidate to challenge common growth patterns due to its atypical growth mechanisms. We develop an inference procedure to extract average single-cell elongation trajectories from detailed microscopy data, despite noise and intrinsic variability in single-cell growth. We find the mean elongation trajectories to deviate from the commonly found exponential single-cell growth; rather, cells initially increase their elongation rate, but to level off to a linear growth regime for later times. This asymptotically linear growth is found to be consistent with a model of the apical cell wall formation mechanism being the rate-limiting step for growth. Lastly, with population growth simulations we show that asymptotically linear growth acts as a regulator of cell size, suggesting an evolutionary explanation for the absence of many common growth-regulation mechanisms in this bacterium.

# Contents

# Chapter 1

# Introduction

Bacteria form a fascinating subject to investigate all aspects of life in one of its simplest yet essential forms. Ten times shorter than a single human cell, and containing a thousandth of its genetic material, bacteria have a life cycle of only a few hours, continually splitting in two to generate offspring. Despite their relative simplicity, they echo patterns of organization and behaviour found throughout the phylogenetic tree of life. During the bacterial life cycle the internal metabolism is regulated, DNA is replicated, but bacteria also interact with their environment. They search their surroundings for nutrients using active propulsion, communicate with fellow bacteria using chemical signalling, and even exchange genetic material between each other. They also engage in warfare between bacterial colonies, collectively performing coordinated strikes, preemptive attacks, and chemical alarm calling that mirrors animal world behaviour [1]. In some ways, bacteria even outshine their multicellular counterparts. For example, they are extremely versatile in the nutrients they use, and can switch their entire metabolic pathway to process a different food source if nutrient conditions change [2]. Due to this rich phenomenology despite a modest organism size, understanding the principles of bacterial organization might offer a window into principles underlying the organization of life in general.

From a physics point of view, there are interesting problems abound relating to bacterial organization. Many cellular dynamics involve active processes, which defy description in terms of traditional equilibrium statistical physics. As all cellular processes take place at the micrometer scale or below, there is a constant interplay between thermodynamic noise and self-organised order. The bacterium is a highly complex self-regulating system, establishing robustness of its biochemical interaction networks under chemical and thermal fluctuations [3]. These fluctuations also raise questions on the limits of cellular information processing and computation, where cells have been found to function close to optimality in the processing of a few crucial pieces of information [4–6]. Proper cellular function also requires the establishment of spatial patterns [7, 8], for example to ensure accurate division at midcell. The active self-propulsion of bacteria through liquids as well as on surfaces raises questions on how this movement is established [9–11], and which search strategies bacteria follow while using chemical gradients to find nutrients [12]. The competition between bacterial species over nutrients in turn leads to intricate cooperation and predator-prey dynamics, which reveal that competition can be a source as well as a disruption of species diversity [13, 14]. Finally, due to their generation time on the order of hours, bacteria are a prime candidate to study evolutionary processes in real-time [15–17].

In this thesis, we dive into two aspects of bacterial organization: spatial chromosomal

folding and cellular growth. The folding of the bacterial chromosome is intimately connected to gene expression patterns, which ultimately dictate all cellular processes. How chromosome folding influences gene expression, and how the chromosome folding itself is modified by cellular processes is unclear however. In fact, even a coherent description of the order and variability of chromosome configurations is presently lacking. To make progress on this major outstanding problem, in chapters 2 and 3 we employ a top-down Maximum Entropy approach, using ideas from statistical physics and information theory to infer chromosome organization directly from experimental data. In chapter 4, we consider the growth of bacterial cells across generations. Under favourable conditions, bacterial colonies can grow rapidly, doubling in size in under an hour. During this process individual cells must continue to establish stable single-cell growth and internal homeostasis, despite growth fluctuations propagating across generations. We infer the single-cell growth mode of the unusually growing *Corynebacterium glutamicum* from detailed experimental data, and reveal the implications of this growth mode for cellular growth mechanisms and cell size regulation.

In preparation for the following chapters, we will introduce a few major concepts in chromosome organization, Maximum Entropy inference, and bacterial growth, and discuss how the work in this thesis builds upon previous work.

## 1.1 The main actors of bacterial chromosome organization

The bacterial chromosome, which is a circular polymer typically around 1mm long, must be highly compacted to fit inside the approximately 1-2$\mu$m long bacterial cell. How this compaction is achieved, and how chromosomal processes like transcription, replication and segregation are established and regulated under this compaction has been a longstanding subject of research. The earliest investigation into spatial chromosome organization dates back to the 1880's [18], where staining was used to visualize chromosomal localization in the cell. In contrast to human chromosomes, which display a dramatic spatial rearrangement during mitosis, bacterial chromosomes appeared uniform throughout the cell cycle. This observation led to the hypothesis that the bacterial chromosome is unstructured. With major experimental advances in recent years, a wide range of organizational features are however being discovered, and a picture of a much more structured polymer is starting to emerge. We will now give a brief overview of the main actors in bacterial chromosome organization identified so far.

At the small chromosomal length scales, the bacterial chromosome forms twisted loops termed **plectonemes**. These plectonemes, which typically contain about 10kb [19] of chromosomal length, form through underwinding of the DNA double helix, termed negative supercoiling. The degree of negative supercoiling is maintained via antagonizing effecs of gyrase, which windes up the DNA, and topoisomerase, which relaxes strain by cutting the chromosome and subsequently reattaching loose ends [20–22]. Plectonemes are found be topologically insulated; many single-strand knicks are needed to fully relax the chromosome [23], with the total number of such topological domains estimated at around 400 [19]. The locations of topological domains are found to be highly dynamic [24, 25], and the likely separation of topological domains at a central core [18, 26], gives rise to a 'bottle brush' picture of chromosome organization. Topological domains aid in the separation of chromosome copies by pulling non-contiguous DNA strands away from each other, and are proposed to facilitate strand break repair by keeping loose ends spatially close [19] .

Further structure is imposed on the chromosome via interactions with various nucleoid associated proteins (**NAP's**). The protein H-NS binds to the chromosome and itself, facilitating 'daisy-chaining' of multiple DNA-bound H-NS proteins [27], forming chromosomal filaments and loops [28, 29]. Local bending of the DNA is performed by Fis, HU and IHF, further inducing local compaction [18, 30]. Fis is additionally found to preferentially bind at DNA overlaps, potentially stabilizing plectonemes [31–33].

On larger chromosomal length scales, organziational structure is imposed via ring-shaped Structural Maintenence of the Chromosome (**SMC**) proteins, which link two chromosomal regions and extrude loops. Via this process, they are involved in the segregation of newly replicated sister chromosomes [33–35]. Although the mechanism of SMC loop extrusion is still unclear, recent years have seen rapid progress on its experimental characterization, revealing that loop extrusion is an active process [36], that SMC's can traverse one another on the chromosome [37] and can overcome roadblocks significantly larger than their ring size [38]. SMC induces alignment between the two chromosomal arms at each side of its loading site, resulting in a juxtaposed organization between the origin (*ori*) and terminus (*ter*) of replication in *Caulobacter crescentus* and *Bacillus subtilis* [39–41].

During chromosomal replication, the segregation of newly replicated sister chromosomes is induced by active transportation of the newly replicated *ori* via the **Par system**. This system consists of three components: the proteins ParA and ParB, and the chromosomal *ParS* loci. ParB specifically binds to *ParS* sites, which sit close to the *ori* region on de DNA [42–44]. ParA binds nonspecifically to the entire chromosome, after which the ParB$S$ complex is actively moved across the cell by 'surfing' the DNA-bound ParA [45–48]. The transport mechanism hinges on ParA being removed from the chromosome after binding to ParB$S$ complex, creating a ParA gradient for the ParB$S$ complex to follow. Several analytical models elucidate the dynamics of this elegant transportation mechanism [49–53]. The Par system separates replicated *ori*'s in several bacterial species, among which *C. crescentus* [54] and *B. subtilis* [55], however other bacteria such as *Escherichia coli* [18] lack this mechanism.

Apart from these major components, there are several other factors that modify bacterial chromosomal structure [18, 33, 56]. Importantly, these components do not act independently, but frequently interact with each other, are influenced by cellular processes such as transcription, and interact with other proteins present in the bacterial cytoplasm [18, 33]. Therefore, elucidating the large-scale organization that emerges via these processes poses a major challenge. In recent years, significant advancement has been made in our understanding of this large-scale organization through the development of the chromosome conformation capture technique Hi-C, which we discuss in the following section.

## 1.2 Probing chromosome organization with Hi-C

In parallel to increasing detail on molecular mechanisms of bacterial chromosome organization being revealed in recent years, breakthroughs in experimental techniques have provided insight into organizational features across chromosomal length scales. The chromosome conformation capture technique Hi-C [57, 58] yields detailed information on spatial chromosome organization in the form of two-point contact frequencies between pairs of genomic regions. The result is an interaction map of genomic region pairs throughout the chromosome, three examples of which are shown in Fig. 1.1. Within this map, each pair of genomic coordinates has a Hi-C score associated with it, which is a measure of how often these genomic regions

are in close proximity, averaged over a population of cells.

From Hi-C data, a wide range of of organizational features can be deduced. In *E. coli* [59] (left panel Fig. 1.1), large contact clusters spanning up to 1.5Mb were identified, termed macrodomains. These macrodomains are associated with an increased fidelity of chromosome segregation [60], and chromosomal loci within a macrodomain display reduced mobility compared to loci in unstructured regions [61]. In *C. crescentus* (middle panel Fig. 1.1), the chromosome was found to organize into Contact Interaction Domains (CIDs), spanning up to a few hundred kb [58]. The boundaries of CID's consist of plectoneme-free regions, and correlate with the positions of highly transcribed genes. Furthermore, the juxtaposed organization of chromosomal arms under the influence of SMC can clearly be observed, with the role of SMC in establishing this organization confirmed via Hi-C measurements on a $\Delta smc$ mutant [58]. In *B. subtilis* (right panel Fig. 1.1), Hi-C revealed that the two chromosomal arms going from *ori* to *ter* are also juxtaposed in this organism, and that SMC is in fact required for this juxtaposition [62]. Subsequent Hi-C experiments performed on cells subject to transpositions of the SMC loading site *ParS* revealed that SMC tethers chromosomal chromosomal arms as it moves from *ori* to *ter*, and an estimate of the SMC progression rate was obtained [39]. In addition to these examples, many further insights on bacterial chromosome organization have been obtained from Hi-C [40, 63–68].

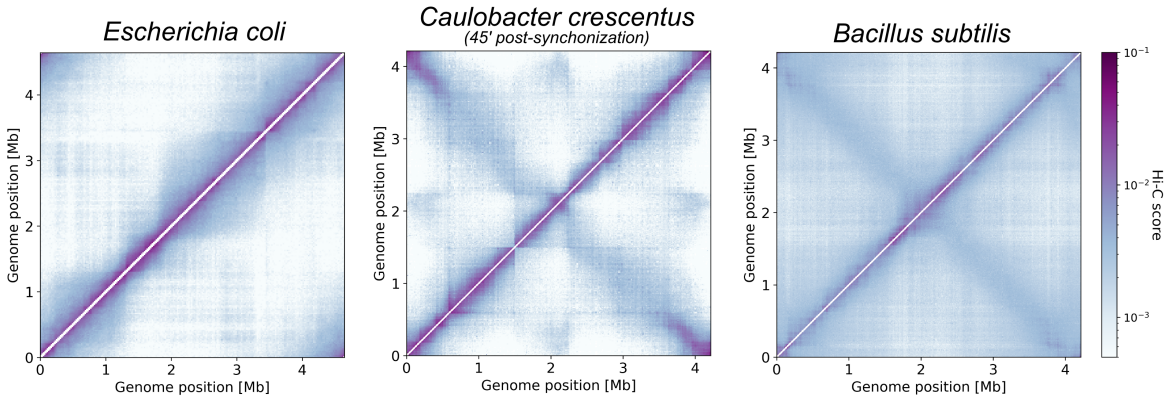



Figure 1.1: Hi-C data sets shown for three bacteria: *E. coli* (left, data from [59]), *C. crescentus* (middle, data from [58]), and *B. subtilis* (right, data from [62]). The Hi-C maps, which are subject to an overall unknown scale factor [69, 70], are rescaled such that the mean score per genomic pair is equal for all three data sets, facilitating direct comparison. The data sets for *E. coli* and *B. subtilis* are for a mixed population of cells distributed over all cell cycle stages. The data set for *C. crescentus* is collected at 45 minutes after the synchronization of newborn swarmer cells, implying that the replication fork has approximately crossed half the chromosome on average (see Chapter 3).

Hi-C data thus provides a wealth of information on many aspects of spatial chromosome organization. The vast amount of information contained in this data however raises the question: is it also possible to infer the full three-dimensional organization of the bacterial chromosome from Hi-C? This is a hard problem, as many possible chromosome models could in principle be consistent with Hi-C data, and a Hi-C map typically contains ∼80000 data points that serve as constraints [69]. In one class of approaches, consensus chromosome structures were obtained by converting Hi-C scores to average distances using an assumed Hi-C score-distance relation [71–73]. Other approaches model the chromosome as an equilibrium polymer with pairwise interactions [74–76], or construct an ensemble of configurations consistent with Hi-C data [77]. Many possible ensembles of configurations could however be consistent with Hi-C data, and an equilibrium polymer might not be suitable to describe a chromosome in a

living cell, which exhibits non-equilibrium fluctuations [78–80].

To overcome the challenge of obtaining a principled model of chromosome organization, in this thesis we construct a Maximum Entropy model to derive the full distribution of chromosome configurations directly from Hi-C data. MaxEnt models constitute a general, unbiased approach to infer a model from experimental data, the construction of which is the subject of the following section.

## 1.3 Maximum entropy inference

Maximum entropy (MaxEnt) inference is a method of constructing statistical models directly from empirical data. The MaxEnt method is the inverse of common physics approaches to model construction; rather than starting from fundamental interactions between system constituents, and deriving system properties from this, we search for the least-structured model that is consistent with our observations. This approach provides a rigorous answer to the question: given a set of observations on a system, which other properties is the system most likely to have?

MaxEnt approaches have been employed in a wide variety of biological contexts to further system understanding, including neuronal firing patterns in the retina [81], bird flocking [82], protein structure prediction [83], antibody diversity [84], and metabolic networks [85]. Compared to traditional bottom-up physics approaches, such an approach differs in what it can teach us about a system. A MaxEnt model does not offer an explanation for *why* a system behaves the way it does; it does not offer an understanding in terms of more fundamental interactions. What it does however do, is provide a way to get all available information about a system out of an experimental data set. Especially for data sets that are hard to interpret in terms of individual system states, this approach can make the difference between uncovering fundamental system properties, or these properties remaining veiled within the data. Furthermore, the MaxEnt model provides constraints for any bottom-up model; the effective behaviour of any bottom-up model should be consistent with MaxEnt predictions. Lastly, a MaxEnt model might still give us insight into system interactions, if we can perform experiments where some interactions are modified. Comparing MaxEnt predictions between differently modified systems can offer insight into how these modifications alter system behaviour, which in turn might offer insight into the system interactions themselves. A concrete example of this will be presented in Chapter 2, where we learn MaxEnt chromosome models for various growth conditions and mutants in *C. crescentus*, and investigate differences in inferred organizational features.

### Quantifying uncertainty

Before developing a MaxEnt inference procedure, we need a measure of the amount of uncertainty contained within a statistical model. A statistical model we will here take as an ensemble of $N$ system microstates $\sigma$, where each microstate has a probability $P(\sigma)$ associated to it. We now imagine a measurement of some system property $\mu$ being performed, with a resulting value $f_\mu$. Via this measurement, we have now gained information about the system, but how can we quantify how much information exactly? This quantified information content lies at the heart of the field of information theory, launched by the seminal 1948 paper by Shannon [86].

To construct a measure of information content [86–88], we note that it should be a function of the probability $P(f_\mu)$ of observing $f_\mu$, given our system description in terms of the $P(\sigma)$. We subsequently impose the following properties on $I(P(f_\mu))$:

1. Information is non-negative: $I(P) \geq 0$.
2. An event with certain occurance contains no information: $I(1) = 0$.
3. For two independent events, the information obtained from the simultaneous observation of these events is equal to the sum of the information gained from each event separately: $I(P_1 P_2) = I(P_1) + I(P_2)$. In other words, the order in which observations are obtained should not matter for their information content if they pertain to independent system properties.
4. The information measure should be a continuous function of the probability $P$.

These properties turn out to be highly constraining for the functional form of $I(P)$. In fact, from these properties we can derive [88] that our information measure must be of the form

$$I(P) = -\log_b(P), \tag{1.1}$$

with the base $b$ a free parameter that determines the units with which we measure information content.

With a definition of information content available to us, we define the model information entropy $S$ as the expectation value of the information gained upon performing a measurement of the system state, $\langle I(P(\sigma))\rangle$. This leads to the Shannon Entropy formula [86]

$$S = -\sum_\sigma P(\sigma) \ln P(\sigma). \tag{1.2}$$

This defines the amount of uncertainty, or information entropy, that we associate to a statistical model.

### Constructing a MaxEnt model

The construction of a MaxEnt model, as first introduced in [87], amounts to finding a system description that maximizes the Shannon entropy (Eq. 1.2), under the constraint that the model should match experimental constraints. This ensures that we find a model that is consistent with experimental data, but otherwise contains as little structure as possible.

The first step in this construction, is the definition of the possible system microstates. This choice defines which states a system could possibly be in at our chosen level of description, and implicitly encodes our prior knowledge on the multiplicities of system states [89]. The emsemble of all such possible system microstates we denote with $\{\sigma\}$. Next, we introduce a set of experimentally measured system properties $f_\mu^{\text{expt}}$, and demand that our model matches these values. This implies

$$\sum_\sigma P(\sigma) f_\mu(\sigma) \stackrel{!}{=} \langle f_\mu^{\text{expt}} \rangle. \tag{1.3}$$

Furthermore, we require that our probability distribution is normalized:

$$\sum_\sigma P(\sigma) \stackrel{!}{=} 1. \tag{1.4}$$

We now maximize 1.2 under the constraints 1.3 and 1.4 via the method of Lagrange multipliers:

$$\tilde{S}[P(\sigma)] = -\sum_{\sigma} P(\sigma)\ln P(\sigma) - \sum_{\mu}\lambda_\mu\left(\sum_{\sigma}(P(\sigma))f_\mu(\sigma) - \langle f_\mu^{\text{expt}}\rangle\right) - \lambda_0\left(\sum_{\sigma}(P(\sigma)) - 1\right), \tag{1.5}$$

where we have one Langrage multiplier $\lambda_\mu$ for each experimental constraint, and an additional $\lambda_0$ ensuring normalization. Setting $\frac{\partial\tilde{S}[P(\sigma)]}{\partial P(\sigma)} \stackrel{!}{=} 0$ and rewriting we obtain

$$P(\sigma) = \frac{1}{Z}\exp\left[-\sum_{\mu=1}^{K}\lambda_\mu f_\mu(\sigma)\right], \tag{1.6}$$

where $Z = \exp[1+\lambda_0]$. This gives us the general form of $P(\sigma)$ for any MaxEnt model, in terms of the set of Lagrange multipliers $\lambda_\mu$. The values of the Lagrange multipliers are obtained by solving Eq. 1.3.

### Relation to the Bolzmann distribution and equilibrium models

The MaxEnt probability distribution (Eq. 1.6) is strongly reminiscent of the familiar Bolzmann distribution. In general, this similarity is strictly an analogy, and the assumptions used to construct the Bolzmann distribution need not apply to our MaxEnt probability distribution. For example, a Bolzmann distribution description requires a system to be in thermodynamic equilibrium, whereas in our MaxEnt derivation no such assumptions are required.

There is however a special case where the MaxEnt probability distribution is identical to the Bolzmann distribution; this occurs when we impose a constraint on the mean system energy $\langle E\rangle$:

$$\langle E\rangle \stackrel{!}{=} \sum_{\sigma} P(\sigma)E(\sigma). \tag{1.7}$$

Solving the corresponding MaxEnt model indeed yields the Boltzmann distribution

$$P(\sigma) = \frac{1}{Z}\exp\left[-\lambda E(\sigma)\right]. \tag{1.8}$$

In this case, the Lagrange multiplier $\lambda$ has a physical interpretation as the inverse temperature. Furthermore, we can now rewrite Eq. 1.5 to

$$-\tilde{S}\lambda^{-1} = \langle E(\sigma)\rangle - S\lambda^{-1} + \text{const.}, \tag{1.9}$$

which is the expression for the free energy $F$ under the identification $F = -\tilde{S}\lambda^{-1}$ and $T = \lambda^{-1}$. This means that the probability distribution that maximizes the entropy functional $\tilde{S}$ is also the distribution that minimizes the free energy.

Constructing a MaxEnt model by constraining the expected system energy is a logical approach if we are dealing with a system in thermodynamic equilibrium. If our system is out of equilibrium however, the energy of a system state can not in general be used to predict its probability, since in a non-equilibrium system active agents continuously consume energy to perform work, and energy is continuously dissipated within the system. Thus, when constructing a MaxEnt model for a biological system, we must carefully consider whether non-equilibrium effects can be ignored before energy constraints are applied.

## 1.4 The laws of bacterial growth

In the fourth chapter of this thesis, we will consider single-cell bacterial growth over time. In preparation for this chapter, we will give a short overview of the bacterial growth laws that have been established so far.

Pioneering work on studying bacterial growth was done in [90–92], where the size of a bacterial colony was measured over many generations. Several distinct growth phases were identified, amongst which the phase of exponential population growth. The appearance of an exponential growth phase is to be expected: in favourable growth conditions, each bacterium grows and divides into two daughter cells. This implies that after $N_{\text{gen}}$ generations, there will be $2^{N_{\text{gen}}}$ bacteria in the population, i.e. an exponential increase. This result remains true regardless of how each individual bacterium grows from its birth to division length.

The characterization of single-cell growth is more challenging than that of population growth, since it requires accurate measurements of size changes over time for individual cells. The presence of measurement noise and intrinsic cell-to-cell variations also implies that statistics over many such single cells must be collected before inferences on mean growth behaviour can be made. In recent years, major advances in automated single-cell measurement techniques have enabled the collection of such data sets [93–95], leading to what has been termed a modern renaissance in microbial physiology [95]. From these measurements, universal properties of bacterial growth have started to emerge.

### The adder principle

A basic quantity to characterize single-cell growth, is the relationship between birth length and division length. Given that a cell is born with at a certain volume $v_b$, at what volume $v_d$ do we expect it to divide? Different strategies imply different possible underlying cellular mechanisms, and lead to different cell size distributions. In [96, 97], a mathematical framework was developed to systematically think about the relation between $v_b$ and $v_d$. Denoting a species-specific growth policy by $v_d = f(v_b)$, the class of affine linear policies is given by

$$f(v_b) = \Delta + cv_b. \tag{1.10}$$

Depending on the values of $\Delta$ and $c$, we have different possible size policies:

- $f(v_b) = \Delta$. This is known as the **sizer** model, where cells grow to a specific final size $\Delta$.
- $f(v_b) = cv_b$. Under exponential single-cell growth, this implies a **timer** model, where cells grow for a fixed time before dividing.
- $f(v_b) = \Delta + v_b$. This corresponds to an **adder** model, where cells add a fixed amount $\Delta$ to their birth volume.

All measurements on single-cell bacterial growth so far have found adder behaviour [98–102], suggesting the universality of this policy. A timer principle combined with exponential single-cell growth can be excluded by stability arguments; in [96] it was shown that this size control strategy leads to divergences in the cell size distribution.

These size policies do not make any statements on the underlying cellular mechanisms, only on the effective behaviour at the population level. In principle, many regulatory mechanisms could give rise to effective adder behaviour. We can however gain insight into possible mechanisms by considering how cell size increase and initiation of chromosome replication are coordinated.

### The General Growth Law

In bacteria, cellular growth has the remarkable property that the time between division events can be significantly shorter than the time needed to complete one round of chromosome replication. This feat is achieved by initiating new rounds of replication before the previous one is completed [103, 104]. This results in multiple replication forks per chromosome, most of which will only be completed in future generations. As all *ori* copies simultaneously fire at initiation of replication, this implies that the number of *ori*'s in the cell is of the form $2^n$, with $n$ an integer.

The multifork replication mechanism raises the question how cell mass increase and chromosomal copy number increase are coordinated. On average, the cell mass doubling time must be equal to the *ori* doubling time, to prevent either continuous accumulation of *ori* copies or indefinite dilution. The cell mass doubling time has been found to be tightly connected to cell size; in nutrient-rich environments that support rapid growth, cell size was found to scale with growth rate regardless of the chemical composition of the medium. This is known as the Growth Law [105], and implies that we could grow cells of a given species in an unknown medium, and by measuring the growth rate infer the mean cell size. Combining this finding with multifork replication led to the hypothesis that replication initiation occurs at a fixed cell size per origin of replication [106].

Initially, it was unclear how the parameters of such an adder-per-origin model would vary across growth conditions and cell masses. Subsequent observations however revealed that the mean mass per origin is constant across a range of cell masses [107–111], and that initiation of replication occurs at a fixed mass for a given growth condition [108, 112, 113]. This lead to the postulation of what is known as the General Growth Law [108]:

$$S = S_0 \times 2^{\tau_{\mathrm{cyc}}/\tau}. \tag{1.11}$$

Here, $S$ is the cell size, $S_0$ is the cell size per origin of replication (*ori*), $\tau$ is the mass doubling time, and $\tau_{\mathrm{cyc}}$ is the cell cycle time, defined as the time needed for one round of chromosome replication and division. Conceptually, the General Growth Law represents a quantitative description of the adder-per-origin model, where a cell adds a fixed amount of volume $S_0$ for each newly replicated origin, and the total number of origins is given by $2^{\tau_{\mathrm{cyc}}/\tau}$. The General Growth Law has been validated for a wide range of growth conditions and mutations affecting individual parameters in Eq. 1.11, suggesting this represents a general principle of bacterial growth [108, 114].

### How can we reconcile the General Growth Law with the adder principle?

In general, the General Growth Law and the adder principle do not necessarily imply each other. A reconciliation of these two growth principles might however be found if a mechanistic origin of the General Growth Law is identified. A long-standing hypothesis for this is the initiator-titration mechanism [106, 115, 116], in which an initiator protein accumulates at a rate proportional to the growth rate, and triggers a new replication round when a critical concentration of this protein has titrated at the *ori* sites. After this start of a new replication round, the initiator protein is degraded again and the cycle begins anew. Modelling work performed in [116–118] demonstrated that such a mechanism can indeed produce adder behaviour at sufficiently slow turnover of the initiator protein. So far, no molecular mechanism for such an initiator-titration protein has been identified [119], although clues may be

found in interactions of the replication-initiation protein DnaA, whose accumulation is a key requirement for origin firing [120, 121].

The replication-initiation centred picture of cell size control has led to speculation that the laws governing cell size are not a direct product of evolutionary selection, but are rather a by-product of selection pressure favouring the multifork replication mechanism [122]. In this scenario, it is the mass-doubling time that is the main determinant of cellular fitness. The General Growth Law is then an emergent property of the requirement that all newborn cells have identical chromosome densities under multifork replication. So far, this hypothesis still awaits experimental validation [119], although results from Lenski's long-term evolution experiment [16, 17] show an evolution of the mean cell size that is consistent with this idea [122].

### Going beyond static growth measures

The regulatory mechanisms discussed so far focus on cell lengths at specific waypoints along the cell cycle. We can however uncover much more details on the cellular growth process if we study the size as a function of time between these points. A wide range of previous measurements of bacterial growth over time suggested exponential single-cell growth [100, 123–126], and this growth mode is often assumed in the development of bacterial growth and cell size regulation models [96, 97, 116–118]. However in the last two years deviations from this trend have started to be revealed from detailed inspection of average single-cell growth [127, 128]. In *Escherichia coli*, growth rate was found to increase at later stages in the cell cycle [128]. In *Bacillus Subtilis*, systemtatic deviations from exponential growth were observed as a function of the cell cycle stage, which were found to compensate for growth-rate disturbances and promote growth-rate homeostasis [127].

Deviations from exponential growth thus have implications for cell size control mechanisms, but they could also reveal features of molecular growth mechanism itself. In chapter 4, we study this growth over time in an unusually growing bacterium, *Corynebacterium glutamicum* which displays a completely different growth mode altogether. From this novel growth mode, we then demonstrate implications for the cellular growth process and cell size regulation.

## 1.5 Goals and significance of this thesis

Many details of bacterial chromosome organization have rapidly been uncovered in recent years, however the overall picture of organization that emerges from this is still unclear. The recently developed chromosome conformation capture technique Hi-C yields a wealth of information on chromosome organization across genomic scales, however deciphering this information in terms of single-cell chromosome configurations is difficult. In chapters 2 and 3, we take a top-down approach to this question by learning a MaxEnt chromosome model directly from experimental constraints. From the MaxEnt chromosome model, organizational features across genomic scales are subsequently inferred. Chapters 2 and 3 of this thesis address the question:

*What are the features of the full distribution of bacterial chromosome organization across length scales, and what do they imply for cellular function?*

The identification of bacterial single-cell growth principles has seen rapid progress over recent years, however much is still left unknown. Single-cell bacterial growth is still sparsely charted across species and growth conditions [119], and emerging methods of dynamical growth analysis have the potential to uncover deeper principles governing bacterial growth. In chapter 4, we develop an inference procedure to extract mean elongation curves from single-cell data, and apply this to the atypically growing *Corynebacterium glutamicum*. The inferred growth behaviour sheds light on cellular growth processes and cell size control mechanisms. Chapter 4 is driven by the question:

*What sets the speed limit for single-cell bacterial growth, and how does this limit affect cell size regulation?*

We make progress on these questions in the following ways.

In chapter 2, we derive a Maximum Entropy model for the spatial organization of a bacterial chromosome in *Caulobacter crescentus*, yielding the full distribution of single-cell chromosome configurations directly from Hi-C data. We validate this model with independent chromosomal localization experiments, and show that the MaxEnt model predicts emergent order across genomic scales. We quantify organizational features across length scales, and discuss their implications for cellular organization.

In chapter 3, we expand upon the approach developed in chapter 2 and develop a MaxEnt model for the full distribution of replicating chromosome configurations. We combine constraints from Hi-C and fluorescence microscopy at various stages throughout the cell cycle to obtain a series of replicating chromosome models over time. The replicating MaxEnt model is validated by fluorescence microscopy data across the chromosome and across the cell cycle. From this model we show how chromosome organization changes throughout the replication cycle, and in particular highlight the role of the pulling of replicated origins of replication (*ori*'s) to opposite cell poles. The replicating MaxEnt model provides access to a wealth of further organizational features, a few of which are discussed at the end of the chapter.

In chapter 4, we study the dynamics of single-cell growth in the highly atypically growing *Corynebacterium glutamicum*. From detailed single-cell growth measurements we obtain average elongation rate curves despite noise and cell-to-cell variability, using an inference procedure that carefully avoids inspection bias effects. To understand the obtained elongation rate curves, we develop a model of the apical growth mechanism being the rate-limiting step for growth, and show the results are consistent with observations. Lastly, we show how *C. glutamicum*'s single-cell growth mode acts as a regulator for cell size.

# Chapter 2

# Learning the full distribution of bacterial chromosome conformations

## Chapter Summary

In this chapter, we investigate the spatial organization of bacterial chromosomes. This spatial organization is intimitely connected to the regulation of gene transcription, and facilitates faithful replication and segregation despite the chromosome being highly compacted within the cellular confinement. A state-of-the-art experimental procedure to investigate this spatial organization is Hi-C, which detects contact frequencies between pairs of genomic regions across the chromosome. These two-point contacts reveal several features of chromosomal organization. However, the full distribution of 3D chromosome configurations remains elusive.

In this chapter, we develop a principled Maximum Entropy approach to derive the full distribution of bacterial chromosome configurations directly from Hi-C data. The resulting model is the least-structured model that is consistent with experimental constraints. To test the MaxEnt chromosome model, we compare localizations of genomic regions along the long cell axis to results from independent microscopy experiments. We find a close match on the mean positions as well as their distributions, validating the predictive power of the MaxEnt model.

Next, we investigate organizational features that are yet inaccessible to experiment. To study organization throughout the entire chromosome, we consider two-point correlations in the locations of genomic regions. We find no long-range correlations in the angular or radial directions of the cellular confinement, suggesting an absence of long-range order in these features. By contrast, we find correlations throughout the entire chromosome along the long cell axis, indicating emergent order. The correlation pattern is explained by the presence of large genomic clusters, termed Super Domains, that tend to exclude each other along the long cell axis. Super-resolution experiments measuring chromosome density confirm the clustered nature of the chromosome.

On a smaller length scale, we find a pattern of local chromosomal extensions that partially correlates with transcriptional activity, but only for one chromosomal arm. Finally, we quantify the chromosomal localization information per genomic region. We find that this localization information reaches up to 3 bits at the origin and terminus of replication, which is equivalent to reducing the positional uncertainly to one cellular octant. We hypothesize that this information could be used by the cell as a cellular roadmap to localize proteins and protein droplets. Our MaxEnt method is not organism-specific, and provides a general approach for inferring the full distribution of spatial chromosome organization across genomic scales.

## 2.1 Publication

# Learning the distribution of single-cell chromosome conformations in bacteria reveals emergent order across genomic scales

by

**Joris J.B. Messelink[1], Muriel C.F. van Teeseling[2,6], Jacqueline Janssen[1], Martin Thanbichler[2,3,4] & Chase P. Broedersz[1]**

[1]Arnold Sommerfeld Center for Theoretical Physics and Center for NanoScience, Department of Physics, Ludwig Maximilian University Munich, Munich, Germany.
[2]Department of Biology, University of Marburg, Marburg, Germany.
[3]Max Planck Institute for Terrestrial Microbiology, Marburg, Germany.
[4]Center for Synthetic Microbiology (SYNMIKRO), Marburg, Germany.
[5]Department of Physics and Astronomy, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands.
[6]Present address: Prokaryotic Cell Biology Group, Department of Microbial Interactions, Institute for Microbiology, Friedrich Schiller University Jena, Jena, Germany.

**Reprinted on pages 15 - 72 from**

**Nature Communications 12:196 (2021),**
**doi:10.1038/s41467-021-22189-x.**

nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-021-22189-x OPEN

# Learning the distribution of single-cell chromosome conformations in bacteria reveals emergent order across genomic scales

Joris J. B. Messelink[1], Muriel C. F. van Teeseling[2,6], Jacqueline Janssen[1], Martin Thanbichler[2,3,4] & Chase P. Broedersz[1,5]✉

The order and variability of bacterial chromosome organization, contained within the distribution of chromosome conformations, are unclear. Here, we develop a fully data-driven maximum entropy approach to extract single-cell 3D chromosome conformations from Hi-C experiments on the model organism *Caulobacter crescentus*. The predictive power of our model is validated by independent experiments. We find that on large genomic scales, organizational features are predominantly present along the long cell axis: chromosomal loci exhibit striking long-ranged two-point axial correlations, indicating emergent order. This organization is associated with large genomic clusters we term Super Domains (SuDs), whose existence we support with super-resolution microscopy. On smaller genomic scales, our model reveals chromosome extensions that correlate with transcriptional and loop extrusion activity. Finally, we quantify the information contained in chromosome organization that may guide cellular processes. Our approach can be extended to other species, providing a general strategy to resolve variability in single-cell chromosomal organization.

[1] Arnold Sommerfeld Center for Theoretical Physics and Center for NanoScience, Department of Physics, Ludwig Maximilian University Munich, Munich, Germany. [2] Department of Biology, University of Marburg, Marburg, Germany. [3] Max Planck Institute for Terrestrial Microbiology, Marburg, Germany. [4] Center for Synthetic Microbiology (SYNMIKRO), Marburg, Germany. [5] Department of Physics and Astronomy, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands. [6] Present address: Prokaryotic Cell Biology Group, Department of Microbial Interactions, Institute for Microbiology, Friedrich Schiller University Jena, Jena, Germany. ✉email: c.broedersz@lmu.de

Chromosomes carry all information to generate a living cell. In both prokaryotes and eukaryotes, chromosomal DNA is highly compacted to fit inside its cellular confinement. This implies a major organizational problem: the DNA does not only have to be highly condensed, but its spatial organization also has to facilitate processes such as transcription and replication. In many bacteria, the genetic information is stored on a single chromosome with a contour length three orders of magnitude larger than the cell. Various proteins regulate bacterial chromosome structure[1–5], imposing order on its spatial organization and thereby impacting cellular processes such as transcription[6]. However, this order is opposed by thermal[7] and active chromosomal fluctuations[8], as well as inherent cell-to-cell variability[9]. The resulting degree of organization of the chromosome remains unclear. Resolving this organization requires a characterization of the distribution of single-cell chromosome conformations, posing a key challenge for experiment and theory[10].

The classical picture in which the bacterial chromosome is arranged as an amorphous polymer has become obsolete thanks to recent experimental advances[11–13]. Indeed, fluorescence microscopy experiments revealed that chromosomal loci localize to well-defined cellular addresses in various species[7,14–16], including *Caulobacter crescentus*[17]. This organization helps steer chromosome segregation[18] and cell division[19]. In addition, the level of transcription of several genes depends on their distance to the pole[20]. Further insights were obtained by chromosome conformation capture 5C/Hi–C experiments[21,22], measuring average pair-wise contacts between loci. These experiments revealed Chromosomal Interaction Domains (CIDs) of up to $10^5$ base pairs, comprising loci preferentially interacting within their domain. Various processes[23,24], including transcription[25,26], impact CID organization. On larger genomic scales, locus pairs on opposite chromosomal arms appear to favor a juxtaposed arrangement in several species, induced by the loop extrusion motor SMC (Structural Maintenance of Chromosomes)[23,26–31]. However, it remains challenging to faithfully extract the distribution of 3D chromosome conformations from Hi–C data. Thus, despite these experimental insights, a complete model for the spatial organization of the bacterial chromosome across genomic scales remains elusive.

To exploit advances in Hi–C experiments on various bacteria[23,24,26,29,31,32], a principled data-driven approach is needed that makes an unbiased inference of the distribution of chromosome configurations. However, there are several outstanding challenges that preclude such a fully data-driven model[26,27,33,34]. Several approaches rely on an assumed relation between Hi–C scores and the average spatial distance between locus pairs to obtain a 3D structure [27,33,35]. Other approaches generate an ensemble of configurations consistent with Hi–C data, e.g., using iterative maximum likelihood algorithms[36]. However, Hi–C maps could be consistent with many underlying distributions. For eukaryotes, an equilibrium Maximum Entropy (MaxEnt) distribution selection method was proposed[37–39], as used for protein structure prediction[40]. However, such an approach may be unsuitable for chromosomes in living cells, which exhibit non-equilibrium fluctuations[8,41,42]. Thus, a rigorous approach to derive a distribution of chromosome conformations compatible with non-equilibrium dynamics is still lacking.

Here, we develop a fully data-driven MaxEnt approach for the bacterial chromosome based on Hi–C data. This approach infers the least-structured distribution of chromosome conformations that fits Hi–C experiments, capturing population heterogeneity at the single-cell level. Our MaxEnt model does not rely on equilibrium assumptions, is inferred directly from normalized Hi–C scores, does not require an assumed Hi–C score-distance relation, and we determine the coarse-graining scale of our model using experiments. The MaxEnt model reveals the organization and variability of the bacterial chromosome across genomic scales. Using this model, we quantify the localization information in the cellular location of chromosomal loci that can be used by cellular processes. Our theoretical framework may be generalized to other prokaryotic and eukaryotic species, providing a principled approach to resolve chromosome organization from Hi–C data.

## Results

**Maximum entropy model inferred from chromosomal contact frequencies**. Our goal is to determine the ensemble of single-cell chromosome conformations for a heterogeneous cell population from experimental Hi–C data. To this end, we build on existing MaxEnt methods for analyzing biophysical data[37,38,40,43–49], to develop a principled approach for inferring the statistics of chromosome structure in bacteria from experiments.

The microstates $\{\sigma\}$ of the system are defined as the set of all configurations of the chromosome contained within the cellular confinement. We seek the statistical weights $P(\sigma)$, chosen to be consistent with the experimental Hi–C map. In general, however, a set of experimental constraints does not uniquely determine $P(\sigma)$. The MaxEnt approach is based on selecting $P(\sigma)$ from these possible solutions by choosing the unique distribution with the largest Shannon entropy,

$$S = -\sum_{\sigma} P(\sigma)\ln \mathrm{P}(\sigma), \tag{1}$$

constituting the least-structured distribution consistent with experimental data. Put simply, we require that the only structure present in $P(\sigma)$ is due to experimental constraints from Hi–C scores, rather than assumed features of the underlying polymer model, the interpretation of Hi–C scores, or the ensemble-generating algorithm. A central assumption of our approach is that the experimental Hi–C maps contain sufficient information to constrain the distribution of chromosome conformations.

To apply the MaxEnt method to experimental Hi–C data, we employ a coarse-grained representation of the chromosome: the polymer is represented as a discrete circular chain of length $N$ on a 3D cubic lattice; the chain can self-intersect and is constrained to the cell-shaped confinement. A subset of the $N$ monomers—equally spaced along this chain—represents the centers of the genomic regions, which are defined as the stretch of the DNA associated with an individual bin of the Hi–C map. Thus, the dimensions of the coarse-grained representation are set by the resolution of the available Hi–C data (Supplementary Notes 2, 3.1). This provides an efficient computational framework, while still capturing key organizational features. Specifically, this representation is chosen to preserve experimentally measured distance fluctuations at the coarse-graining scale (see "Methods" section and Supplementary Notes 1–2). At larger scales, the statistics of polymer configurations are only constrained by Hi–C data. Within this representation, a microstate $\sigma = \{\mathbf{r}_1, \mathbf{r}_2, \dots\} = \{\mathbf{r}\}$ is defined by the monomer positions $\mathbf{r}_i$. Two genomic regions have a contact probability $\gamma$ if they occupy the same lattice site, and 0 otherwise.

To obtain the least-structured distribution of microstates consistent with experiments, we seek $P(\{\mathbf{r}\})$ that maximizes $S$ (Eq. (1)) under experimental constraints[45,50]. The two constraints we impose are: 1) the model contact frequencies should match experimental contact frequencies $f_{ij}^{\mathrm{expt}}$ between genomic regions $i$ and $j$ (the correspondence between $f_{ij}^{\mathrm{expt}}$ and Hi–C scores is discussed in the next section), and 2) the distribution should be normalized. To this end, we introduce the functional $\tilde{S}$, with one Lagrange multiplier $\lambda_{ij}$ for each experimental constraint and $\lambda_0$

ensuring normalization:

$$\tilde{S} = -\sum_{\{\mathbf{r}\}} P(\{\mathbf{r}\})\ln \mathrm{P}(\{\mathbf{r}\}) - \sum_{\mathrm{ij}} \lambda_{\mathrm{ij}} \left( \sum_{\{\mathbf{r}\}} \mathrm{P}(\{\mathbf{r}\}) \gamma \delta_{\mathbf{r}_i,\mathbf{r}_j} - f_{ij}^{\mathrm{expt}} \right) - \lambda_0 \left( \sum_{\{\mathbf{r}\}} P(\{\mathbf{r}\}) - 1 \right) \quad (2)$$

Here, $\delta_{\mathbf{r}_i,\mathbf{r}_j}$ is the Kronecker delta. We maximize $\tilde{S}$ under these constraints, setting $\frac{\delta \tilde{S}}{\delta P(\{\mathbf{r}\})} = 0$, yielding

$$P(\{\mathbf{r}\}) = \frac{1}{Z} \exp\left[ -\sum_{ij} \lambda_{ij} \gamma \delta_{\mathbf{r}_i,\mathbf{r}_j} \right], \quad (3)$$

with $Z = \exp[1 + \lambda_0]$. The $\lambda_{ij}$'s parametrizing $P(\{\mathbf{r}\})$ is determined by solving

$$\sum_{\{\mathbf{r}\}} P(\{\mathbf{r}\}) \gamma \delta_{\mathbf{r}_i,\mathbf{r}_j} = f_{ij}^{\mathrm{expt}} \quad (4)$$

for each experimental constraint. For typical Hi–C data on a bacterial chromosome, this amounts to of order $10^5$ constraints[26]. These equations can not be solved directly, as they are highly nonlinear and the state space is very large.

The daunting challenge of finding the Lagrange multipliers can be overcome by noting that the distribution in Eq. (3) can be mapped to a statistical mechanics model: a confined lattice polymer, with a (dimensionless) Hamiltonian

$$H = \frac{1}{2} \sum_{ij} \epsilon_{ij} \delta_{\mathbf{r}_i,\mathbf{r}_j}. \quad (5)$$

The mapping to Eq. (3) is made by setting $\epsilon_{ij} = \gamma \lambda_{ij}$, where $\epsilon_{ij}$ are the effective interaction energies between overlapping loci in the Hamiltonian formulation. Importantly, although a mapping can be made to a statistical mechanics model, our approach does not rely on the chromosome being in thermal equilibrium. This is in contrast to approaches used in refs. [37–39] where a hybrid MaxEnt procedure is employed combining a physical polymer model with Hi–C derived constraints, resulting in an energy landscape description of equilibrium chromosome configurations.

We numerically obtain the inverse solutions of this model using iterative Monte Carlo simulations (Supplementary Note 3). Testing this algorithm on contact frequency maps generated from a set of chosen input $\epsilon_{ij}$, we find that our algorithm precisely and robustly recovers the correct input values (Supplementary Note 4).

**Inferring the MaxEnt model directly from normalized Hi–C scores**. A major hurdle in applying data-driven inference approaches is finding a correspondence between experimental Hi–C scores and the contact frequencies in a coarse-grained polymer model. Published Hi–C maps are typically normalized. This normalization compensates known biases in raw Hi–C data, for instance, due to the proportionality between the number of restriction sites in a genomic region and its Hi–C score[51]. Furthermore, absolute Hi–C scores are hard to interpret because it is difficult to estimate the conversion factor to physical contact frequencies. Importantly, however, even if absolute contact scores could be obtained, a mapping to contact frequencies in a coarse-grained model is challenging.

We address this conversion issue by treating the conversion factor as an unknown parameter $c$ in our MaxEnt procedure. Thus, we write $f_{ij}^{\mathrm{expt}} = c\tilde{f}_{ij}^{\mathrm{expt}}$, with $\tilde{f}_{ij}^{\mathrm{expt}}$ the normalized experimental Hi–C scores. We absorb the contact probability factor $\gamma$ into $c$ (Eq. (2)), setting $\tilde{c} = \frac{c}{\gamma}$, and require that $\tilde{c}$ maximizes the model entropy (Supplementary Note 3.2), yielding the additional constraint

$$\sum_{ij} \epsilon_{ij} \tilde{f}_{ij}^{\mathrm{expt}} = 0. \quad (6)$$

Thus, we infer the least-structured distribution of chromosome conformations from normalized Hi–C data, without assuming a conversion between Hi–C scores and contact frequencies or average distances between loci.

**MaxEnt model of the *C. crescentus* chromosome quantitatively captures measured cellular localization**. We investigate the degree of organization of the bacterial chromosome by considering newborn swarmer cells of the model organism *C. crescentus*. Such newborn swarmer cells contain only a single chromosome, whose replication has not yet initiated[52]. To develop the MaxEnt model for *C. crescentus*, we first experimentally determine the coarse-graining scale, set by the average distance between consecutive 10 kb genomic regions (Supplementary Notes 1–2). Subsequently, we infer the parameters of the MaxEnt model from published experimental Hi–C data (Supplementary Note 5)[26]. Our inverse algorithm robustly converges to an accurate description of the Hi–C map: the modeled and experimental contact maps have an average pair-wise deviation of 6.0% of the total average Hi–C score with a Pearson's correlation coefficient of 0.998 (Fig. 1A, B inset).

Our MaxEnt model quantitatively reproduces essential features of the experimental Hi–C map (Fig. 1A), including the fine structure of the CIDs, as well as the secondary diagonal, which is attributed to the alignment of the two chromosomal arms by SMC[30,53–55]. The inferred $\epsilon_{ij}$'s (Fig. 1B) should not be interpreted as physical interaction energies. Rather, they parametrize the predicted physical distribution of chromosome configurations $P(\{\mathbf{r}_i\})$. We can directly interpret the organizational features implied by $P(\{\mathbf{r}_i\})$ and use it to sample single-cell configurations (Fig. 1C).

We test the predictive power of the MaxEnt model by computing the distribution of axial locations of several loci. Importantly, we do not assume (polar) cell envelope tethering of specific loci, such as the origin of replication (*ori*). We orient cells by setting the *ori* pole in the cell-half containing *ori*. Interestingly, we find a high degree of axial localization of loci: the average axial position of loci is roughly linearly organized, and the predicted positions match previous live-cell microscopy experiments[17] (Fig. 2A). By contrast, simulation results of a confined random polymer—not constrained by Hi–C data—do not exhibit the linear organization, even when *ori* is tethered to the cell pole.

The MaxEnt model also predicts distributions of long-axis positions of chromosomal loci, in remarkable agreement with prior experiments (Fig. 2B). This comparison with independent experimental data constitutes a strong validation of our MaxEnt model. The slight deviation of the position of *ori* compared to the experiments (Fig. 2A, B) can be addressed with an extended MaxEnt model that incorporates the distribution of axial *ori* positions as an additional constraint (Supplementary Note 17). However, other aspects of the predicted chromosomal organization are largely unaffected by this modification, and therefore we will not impose this additional constraint in our analysis.

**Large-scale chromosome organization primarily characterized by long-axis correlations associated with Super Domains**. Large-scale organizational features of the chromosome can be revealed by measuring various two-point correlation functions. Earlier models suggested a three-dimensional organization in which the two chromosomal arms wind around each other with roughly one helical turn[27,33]. To test if this organization also

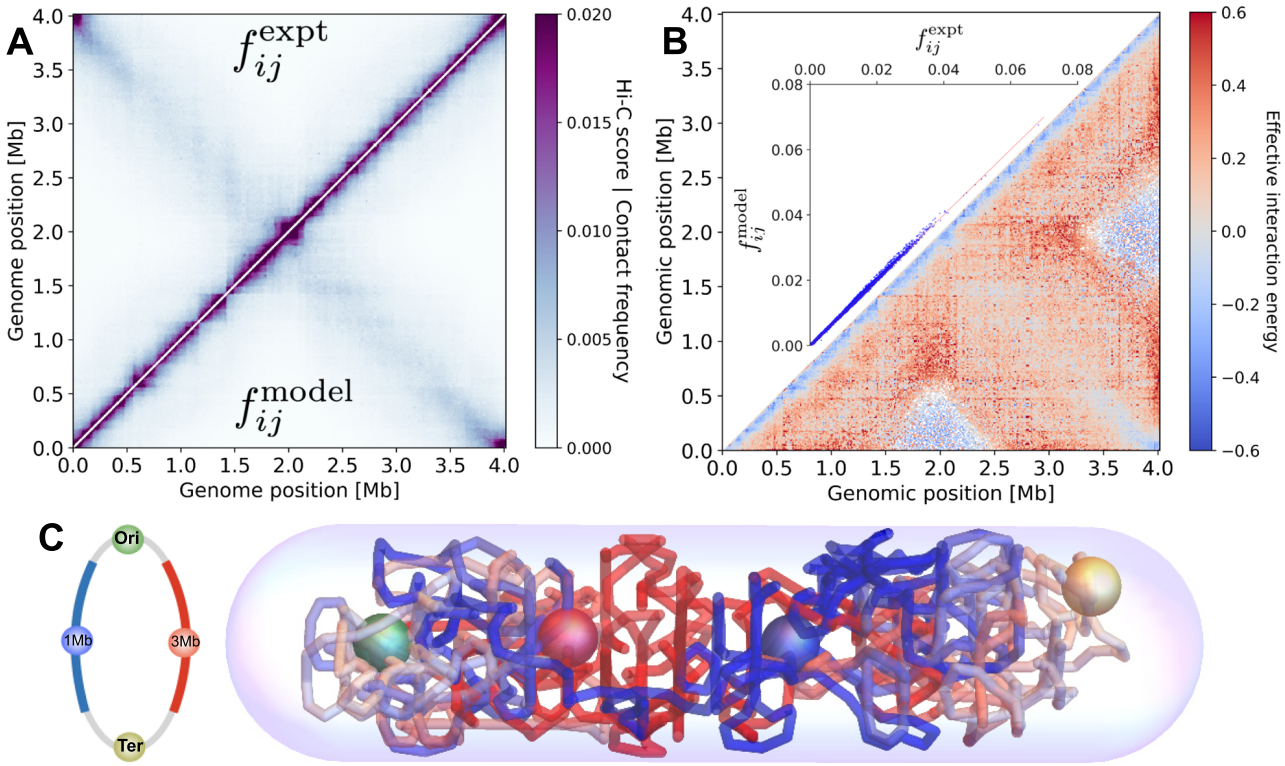



**Fig. 1 Maximum entropy model inferred from Hi–C experiments in *C. crescentus*.** **A** Comparison between experimental contact frequencies $f_{ij}^{\mathrm{expt}}$ (upper left corner, adapted from ref. [26]) and contact frequencies obtained from our inferred MaxEnt model $f_{ij}^{\mathrm{model}}$ (lower right corner). **B** Associated inferred effective interaction energies $\epsilon_{ij}$ (lower right corner, white regions indicate $\epsilon_{ij} \to \infty$) together with a scatter plot of $f_{ij}^{\mathrm{expt}}$ vs. $f_{ij}^{\mathrm{model}}$ (inset). **C** Visualization of a single-cell chromosome configuration predicted by our MaxEnt model; the centers of four distinct chromosome sections are represented in the schematic by colored spheres.

emerges in our MaxEnt model, we compute two-point correlations of angular orientations. For each chromosome segment, we assign an orientation vector in the plane perpendicular to the long axis. We find that angular correlations decay rapidly for genomic distances $\gtrsim$0.2 Mb (Fig. 3A lower right). Large-scale helical order is thus negligible, indicating that a pronounced helical organization is not required to model the experimental Hi–C map.

The two-point correlation function in radial positions decays even more rapidly with genomic distance up to ~0.1 Mb (Fig. 3A upper left), indicating the absence of large-scale order in this direction. By contrast, two-point correlations in the long-axis position exhibit a striking structure: we observe positive long-ranged correlations for pairs of genomic regions on the same chromosomal arm, whereas correlations in axial positions between arms are predominantly negative (Fig. 3B upper left). These long-ranged correlations signify emergent order. Importantly, such organization is absent for a model with a tethered origin not constrained by Hi–C data (Fig. 3B, lower right), as well as for a model with juxtaposed chromosomal arms only constrained by linearly organized average long-axis positions (Supplementary Note 16). Moreover, the structure of the long-axis correlations is inconsistent with global rotational fluctuations (Supplementary Note 12).

We find that these intra-arm anticorrelations are associated with large high-density clusters of subsequent genomic regions, which we term Super Domains (SuDs). SuDs emerge from a clustering analysis of genomic regions in single-cell conformations (Supplementary Note 9). The formation of domain-like structures is revealed by plotting the distance between pairs of loci for a specific chromosome configuration, with single domains spanning up to a quarter of the chromosome length (Fig. 4A, B). On average, 73% of genomic regions are part of a SuD, each chromosomal arm contains ~4 SuDs, and each SuD contains 48 genomic regions (Supplementary Fig. 21). Compared to CIDs, they are typically larger with more variable size and genomic location across chromosome conformations. The variable and delocalized nature of SuDs is apparent from the average distance map between genomic regions, indicating no discrete structure (Fig. 4C). Importantly, SuDs forming on opposing chromosomal arms tend to spatially exclude each other (Fig. 4B, E): the fraction of overlap in axial positions is reduced by 26% compared to randomly paired left and right arm configurations. As a result of this tendency to spatially exclude, chromosomal regions belonging to SuDs on opposing sections of the two arms, are expected to fluctuate in an anti-correlated fashion. (Supplementary Note 9). Thus, this exclusion behavior of opposing SuDs is expected to generate negative intra-arm correlations for pairs of genomic regions with similar average axial positions (Supplementary Note 9).

To experimentally verify signatures of SuDs, we turned towards SIM (structured illumination microscopy) super-resolution microscopy and investigated the intracellular distribution of chromosomal DNA in *C. crescentus* at the single-cell level. These experiments reveal that the chromosome exhibits a highly heterogeneous spatial distribution in the cell, including several dense cluster-like regions (Fig. 4D). We observe that the number, size, and location of these high-density regions are found to vary from cell to cell, consistent with SuD properties derived from our MaxEnt model. To compare these single-cell experimental results with theory, we provide computed density plots of chromosomes based on our MaxEnt model. Specifically, for each chromosome configuration in our model, we compute a chromosome density plot at the experimental resolution (see Methods), as shown in (Fig. 4E). In the computed density plots, we observe high-density

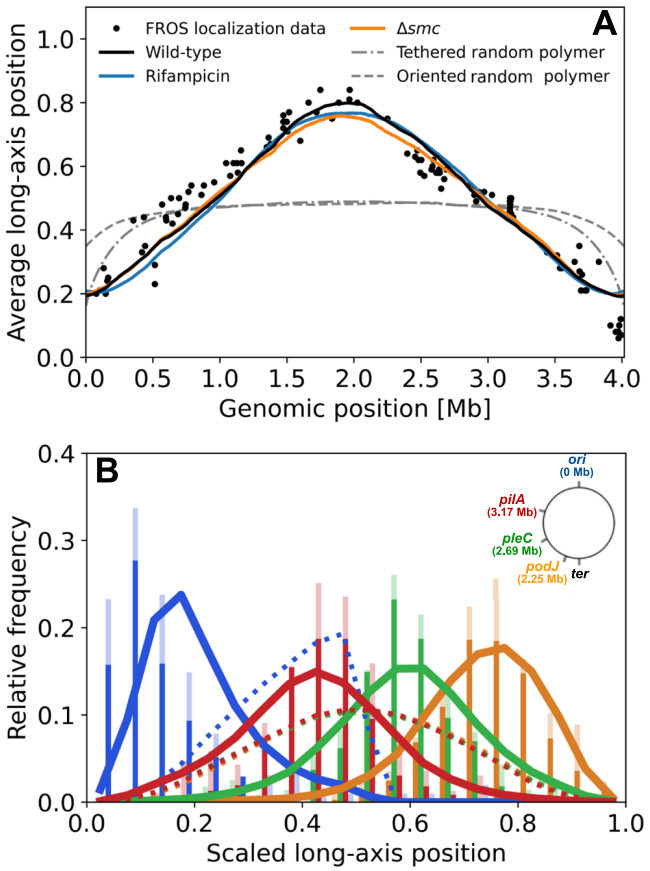



**Fig. 2 Validation of MaxEnt model based on spatial location microscopy data. A** Average scaled long-axis position predicted from MaxEnt models (solid lines) inferred from various Hi-C data sets (from[26]), including wild-type cells (black), rifampicin-treated cells (blue), and Δ*smc* cells (orange), together with results from microscopy experiments (adapted from[17]). Also shown are simulated data for a random polymer with *ori*-pole tether (dash-dotted gray line), and a simulated confined random polymer (dashed gray line), oriented such that *ori* is always on the left cell half. **B** The distribution of single-cell positions (scaled long-axis position) of chromosomal loci (blue: *ori*, red: *pilA*, green: *pleC*, orange: *podJ*), as predicted by the MaxEnt model (solid lines), together with previous experimental data from microscopy experiments (bars, adapted from[17]). To indicate experimental variability, the solid/transparent bars indicate the minimum/maximum measured by two different methods: FROS or FISH. To enable a direct comparison between model and experiment, the model values are distributed over the same number of bins as the experiment. The dotted lines indicate the distribution for a confined oriented random polymer as in **A**.

regions similar to those obtained in our super-resolution experiments. Importantly, the high-density regions in the modeled chromosome density plots correspond to underlying SuD structures (dashed lines in Fig. 4E). Thus, these results allow us to establish a connection between the SuDs predicted by our model and single-cell super-resolution data.

To investigate the influence of cellular processes on long-axis organization, we perform the two-point correlation and SuD structure analysis (Supplementary Note 9) on published Hi–C data of rifampicin-treated cells and a mutant lacking SMC (Δ*smc*)[26] (Supplementary Note 13). Rifampicin treatment inhibits transcription, whereas deletion of SMC abolishes the loop-extrusion activity required to juxtapose the two chromosomal arms[53,56]. For both cases, our models predict an average localization along the long axis similar to those in wild-type cells (Fig. 2A). However, the predicted long-axis correlations exhibit marked differences: for rifampicin-treated cells with inhibited transcription, anticorrelations between chromosomal arms are less pronounced (Fig. 3C upper left). In contrast, Δ*smc* cells display a broad regime with strong anticorrelations between loci on opposite arms (Fig. 3C lower right). These effects are reflected in the statistics of SuDs: upon inhibition of transcription, the SuDs contain 7% more genomic regions per domain than in the wild type. Despite this increased density, the transcription-inhibited cells show a similar overlap of SuDs (29% lower than for randomly paired arms). By contrast, Δ*smc* cells exhibit a similar average SuD density to the wild type (50 genomic regions per cluster on average), but a strong reduction of inter-arm domain overlap (48% lower than for randomly paired arms). Correspondingly, the anticorrelations between long-axis positions of chromosomal arms are much stronger for this mutant (Fig. 3C lower right). Thus, these results suggest that the action of SMC enhances interactions between SuDs, whereas transcription alters their density.

**Local chromosome extension coincides with high transcriptional activity, but only for one chromosomal arm**. The MaxEnt model provides access to local structural features that may be difficult to determine experimentally. Specifically, we consider the local chromosomal extension $\delta_i$, defined as the average spatial distance between two neighboring genomic regions of region $i$ (Supplementary Note 15). Interestingly, the $\delta_i$-profile exhibits an overall trend that is lowest at *ori* and *ter* (Fig. 5A), indicating that these regions are intrinsically more compact (Supplementary Note 15). In addition, pronounced peaks and valleys in the local extension are revealed at a smaller genomic scale similar to that of CIDs. The same structure appears for Δ*smc* cells, although their chromosome appears to be locally more compact than that of the wild type. By contrast, in rifampicin-treated cells, peak amplitudes are significantly suppressed, suggesting a link between local chromosome extension and transcription.

Previous work reported a connection between CID boundaries and highly transcribed genes[26]. Based on this observation and polymer simulations, it was suggested that high transcription creates plectoneme-free regions, physically separating CIDs. To further investigate the impact of gene expression activity on local structure, we compare the locations of local chromosome extension peaks in our MaxEnt model and the 2% most highly transcribed genes. Indeed, we observe a significantly increased overlap between the local chromosome extension peaks and the locations of highly transcribed genes, compared to a random distribution of peaks, but only for genes on the forward strand of the right *ori*-*ter* arm (0–2.0 Mb) (Supplementary Note 10). If the colocalization of local extension peaks by highly transcribed genes would only depend on the relative direction of transcription and replication, this should also occur for highly transcribed genes on backward strands on the left arm, which we do not observe. Thus, while our results indicate a connection between high local chromosome extension and the direction of replication and transcription of highly transcribed genes, the underlying molecular mechanism is still unclear.

**The chromosomal structure provides localization information in the cell**. The inferred structural features of the chromosome not only yield insights into the cellular organization, but they may also have functional significance: organizational features of the chromosome contain spatial information that could guide cellular processes. This spatial information depends on the degree of localization of genomic regions. Put simply, the localization information content of a genomic region increases with the

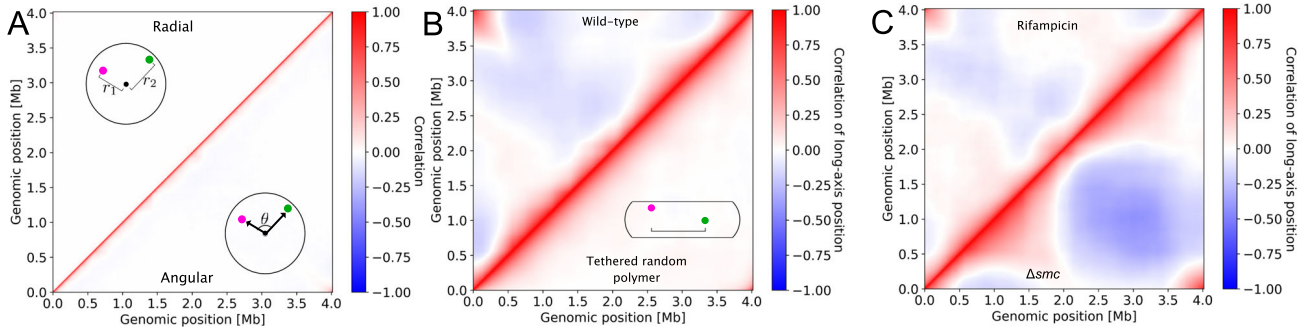



**Fig. 3 MaxEnt model predicts large-scale features of chromosome organization. A** Upper left corner: two-point correlations in the radial positions between genomic regions. Lower right corner: two-point correlations in angular orientations around the long axis. **B** Upper left corner: two-point correlations between long-axis positions for wild-type cells. Lower right corner: the same correlations for a model not constrained by Hi-C data, but with a tethered origin (tethered random polymer). **C** Correlations in the long-axis positions of genomic regions derived from Hi-C data[26] obtained for two modified conditions: cells treated with rifampicin to block transcription (upper left corner), and Δ*smc* cells (lower right corner).

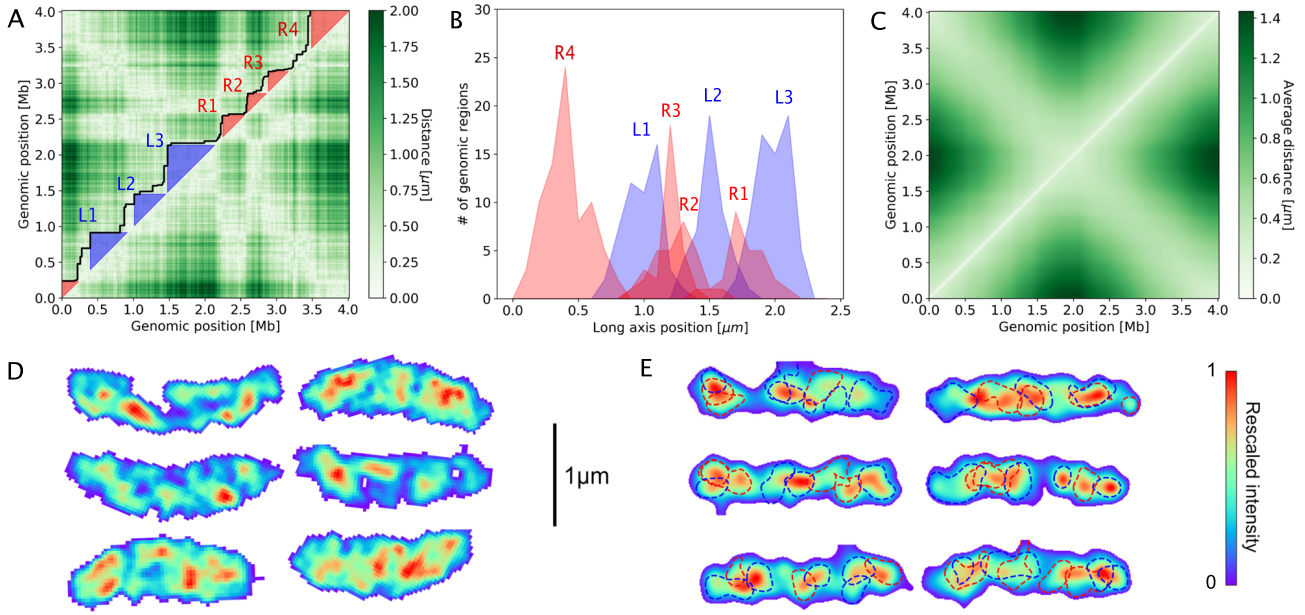



**Fig. 4 Long-axis organization is associated with Super Domain formation. A** Distance map for pairs of genomic regions for one chromosomal configuration. The inferred outlines (Supplementary Note 9) of Super Domains (SuDs) are indicated by a black line, with left/right-arm SuDs shaded blue/red. **B** Long axis distribution of genomic regions in SuDs identified in the configuration depicted in A. **C** Average spatial distances between genomic regions. **D** Super-resolution microscopy images of DAPI-stained DNA inside six synchronized *C. crescentus* swarmer cells. The color code reflects the DAPI fluorescence signal at each pixel, rescaled so that the maximum is at 1 for each cell. **E** Chromosome density plot with the same scaling of several randomly chosen chromosome configurations from our MaxEnt model (with Gaussian blur applied that matches the experimental resolution). Dashed lines indicate the half-maximum density contour of each SuD (identified by the clustering analysis in Supplementary Note 9), with the line color indicating if a SuD predominantly forms on the right (0–2 Mb, blue) or left (2–4 Mb red) chromosomal arm.

precision of its cellular location, i.e., when the spatial distribution of the genomic region is more sharply peaked around a specific point in the cell. This localization information (introduced in the context of developmental patterning[57]) could for example be used to position proteins within the cell: a high relative affinity to a genomic region with high localization information increases the localization of this protein. This mechanism may be exploited to position protein droplets[58], through nucleation on specific chromosomal regions, e.g., droplet-like clusters of DNA-binding chromosome partitioning proteins of the ParB family[3].

Using our MaxEnt model, we can quantify how much localization information (Supplementary Note 14) is encoded by chromosome organization per genomic region (Fig. 5B). This chromosomal localization information is largest near *ori* and *ter*, providing 3 bits of localization information, equivalent to reducing the localization uncertainty to one cellular octant. By contrast, a random polymer provides only 1 bit, enough to reduce localization uncertainty to one cell half. For comparison, with our coarse-grained description, maximal localization information of approximately 9 bits could be achieved. Thus, while this localization information metric indicates that the bacterial chromosome is substantially more ordered than a random polymer, it also highlights that the chromosome is far from having a rigid organization with a precise folded structure.

Comparing these results with those for modified conditions, we find that rifampicin treatment increases chromosomal localization information, whereas information is reduced in Δ*smc* cells, suggesting that SMC action and transcription have opposing effects on localization information. This localization information is just one example of how structural features in the organization

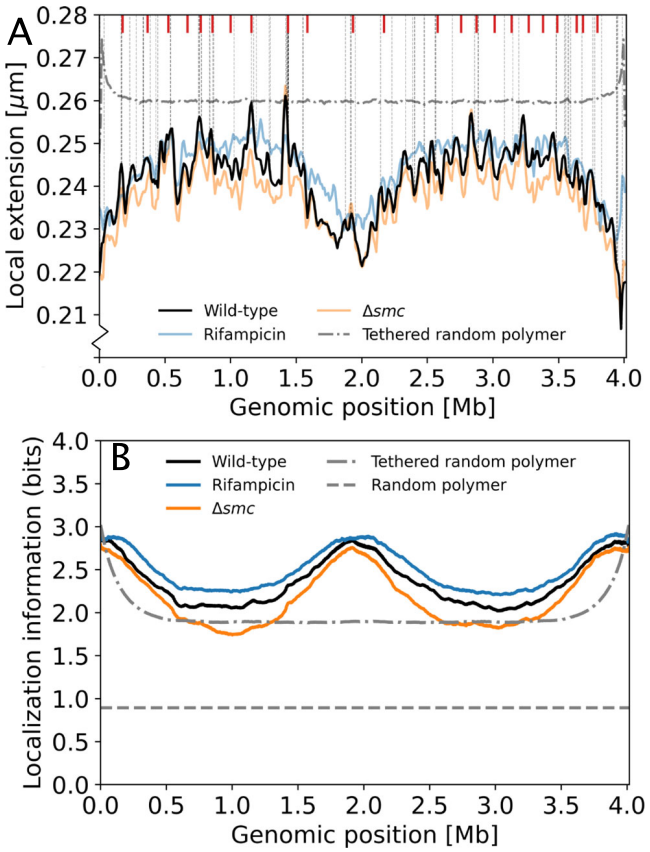



**Fig. 5 The MaxEnt model reveals local features and localization information encoded by chromosome organization. A** The local chromosome extension $\delta_i$ as a function of genomic position. $\delta_i$ is defined as the spatial distance between neighboring genomic regions of site *i* averaged over all chromosome conformations. Model predictions are shown for wild-type cells (black), rifampicin-treated cells (blue), Δ*smc* cells (orange), and a pole-tethered random polymer (gray dash-dotted line). The locations of the top 2% highly transcribed genes are indicated by vertical gray dashed lines, the locations of CIDs determined in ref. [26] are indicated by red markers. **B** Localization information per genomic region in bits for wild-type (black), Δ*smc* (orange), rifampicin-treated cells (blue), a random pole-tethered polymer (dash-dotted line), and a random polymer (dashed line).

of the chromosome can be used to guide cellular processes. The MaxEnt approach provides a scheme to estimate the information available to the cell that is contained in the distribution of chromosome conformations.

## Discussion

We established a fully data-driven principled approach to infer the spatial organization of the bacterial chromosome at the single-cell level and applied this approach to normalized Hi–C data of the model organism *C. crescentus*. The predictive power of this MaxEnt model is confirmed by prior microscopy experiments[17] showing the distributions of axial positions of chromosomal loci within the cell. Contrary to previous modeling approaches, our MaxEnt model does not rely on an assumed connection between Hi–C scores and average spatial distances[21]. Instead, we can predict how these quantities are related: we recover the approximately linear relation between intra-arm genomic distance and spatial distance used as an input in refs. [21,33] (Supplementary Note 11). However, there are substantial region-to-region deviations in the resulting relation between Hi–C scores and average spatial distances, together with significant correlations in distances between genomic regions. Previous approaches could not account for such deviations and correlations. This may explain differences in model predictions such as the helical chromosomal structure suggested in refs. [27,33], which we do not observe.

By design, the MaxEnt model yields the least-structured distribution of chromosome conformations consistent with experimental constraints, allowing us to investigate the degree of order in the bacterial chromosome. To do this, we considered two-point correlation functions in the cellular positions of genomic regions. We observe negligible correlations in the radial and angular coordinates, indicating an absence of organizational order in these directions. By contrast, there are pronounced long-ranged correlations along the long cell axis, indicating emergent order. This order is related to the observation of variable and delocalized clusters of genomic regions, which we term Super Domains (SuDs). These SuDs manifest in single-cell conformations and are consistent with high-density clusters observed in the *C. crescentus* chromosome by our super-resolution microscopy experiment (Fig. 3E). Similar blob-like structures have previously been observed with (super-resolution) microscopy for the chromosome of *Bacillus subtilis*[23] and *Escherichia coli*[13], suggesting that SuDs are also present in other bacteria. Our MaxEnt model indicates a spatial exclusion of opposing SuDs from different chromosomal arms, which we associate with the long-ranged anticorrelations in axial positions. The interplay between SMC complexes and transcription has been explored in prior work[28,59]. We find that transcription and SMC have opposing effects on SuD properties: inter-arm overlap between domains is reduced by transcription and increased by SMC, consistent with the idea that SMC links chromosomal arms[23,29,30,53].

At the smaller genomic scale of CIDs, we observe a characteristic pattern of local chromosomal extensions, being most compact at *ori* and *ter*. We speculate that the local compaction of the *ori* region may be due to the binding of nucleoid-associated proteins (NAPs)[1,2] such as the ParABS chromosome partitioning system[3,4]. The compaction of the *ter* region might be imposed by the recently discovered NAP ZapT[60], which specifically binds to this region of the chromosome, or by additional as-of-yet undiscovered NAPs. Interestingly, peaks in local extension tend to coincide with highly transcribed genes, but only for the forward strand of the right chromosomal arm (Supplementary Note 10).

From our MaxEnt model, we obtain an estimate of the chromosomal localization information per genomic region. This information reaches up to 3 bits around *ori* and *ter*, equivalent to a localization uncertainty in the cell of one cellular octant. We speculate that such localization information encoded by the organization of the chromosome could be exploited for subcellular positioning of proteins and protein droplets[58] or for the regulation of transcription of genes, as was observed in[20].

Our approach resides in the class of static Maximum Entropy approaches, which make no assumptions or predictions about the underlying dynamics, as opposed to dynamical maximum entropy models or maximum caliber models (see for instance[61,62]). Further model limitations are set by the available input data: organizational features that cannot be faithfully encoded in population-averaged Hi–C data might be absent in the MaxEnt model. The resolution of Hi–C data is limited to 10 kb for the data sets analyzed here, implying that any organizational features below this genomic length scale cannot be explored with our model. However, our approach is not limited to interpreting Hi–C data and can be extended towards an integrated MaxEnt model, simultaneously constrained by both Hi–C and microscopy data (Supplementary Note 17). Furthermore, our approach may be generalized to other prokaryotes, including

systems with replicating chromosomes and multiple replicons, as well as eukaryotes, paving the road for unraveling all information on chromosome conformations at multiple length scales, elucidating single-cell variability and population averages.

## Methods

Here, we consider Hi–C data (replicate 1 of the BglII Hi–C data) on *C. crescentus* newborn swarmer cells published in ref. [26], which have a single, non-replicating chromosome. However, due to imperfect synchronization, a small fraction of cells are included in these experiments in which processes such as chromosome replication and segregation have initiated, which will be reflected in the Hi–C map[27,33]. Before inferring a MaxEnt model, we apply a data-processing scheme to filter out contributions from cells with replicating chromosomes (See Supplementary Notes 5–6). However, we also provide a MaxEnt model inferred directly from the unprocessed Hi–C data (See Supplementary Note 7) and MaxEnt models inferred from Hi–C data sets for replication-arrested cells[25] (See Supplementary Note 8). While there are small differences between the different models, the central behaviors from the MaxEnt model reported in the main text are similar in all cases.

Our algorithm (Supplementary Notes 3,4) requires two length scales: the dimensions of the cellular confinement and the lattice spacing. As cellular confinement, we use a cylinder capped with hemispheres with the dimensions of a newborn swarmer cell minus the cell envelope: 0.63 μm × 2.2 μm (Supplementary Notes 1–2), which is assumed to be the same for all cells. A more detailed representation of the cellular confinement shape does not appear to affect our main results (Supplementary Note 17). To set the coarse-graining scale of our MaxEnt model, we experimentally determined the distribution of spatial distances between subsequent Hi–C bins. Specifically, the lattice spacing, $b$, is set by the average spatial distance between consecutive 10 kb regions (the Hi–C bin size). To determine this parameter, we probed the physical distance of two loci separated by 10 kb in five different regions of the chromosome, using an approach comparable to[63,64]. To this end, we constructed strains whose chromosomes contained two independent arrays of transcription factor binding sites (comprising 10 LacI or TetR binding sites, respectively) inserted at the proper distance (Supplementary Note 1). The sub-cellular positions of these arrays were then determined by producing the respective fluorescently labeled transcription factors (LacI-eCFP and TetR-eYFP) at very low levels, based solely on the basal activity of the inducible promoter driving their expression. Swarmer (G1-phase) cells were imaged immediately after isolation, and the localization of the two arrays was determined with sub-pixel precision by fitting a 2D Gaussian to the acquired images. The Euclidean distances between the two arrays were calculated, taking into account correction factors for a systematic shift produced by the set-up (see Methods for further details) and are shown in (Table S5). The average distance between genomic loci 10 kb apart were found to be 129 ± 7 nm, implying a lattice spacing $b = 88$ nm (Supplementary Note 2). For the selection of cells in Fig. 4D, cells with approximately the average newborn cell length (2.3 ± 0.2 μm (Supplementary Note 2.2)) were chosen. For each cell, out of the z-stack, the plane that corresponded to the mid-cell being in focus was selected. For the calculation of single-cell chromosomal density plots (Fig. 4E), a Gaussian blur was applied, whereby the resolution in the $z$-direction (300 nm) and in the $x$ and $y$ directions (120 nm) were set to match the experimental resolution.

**Reporting summary**. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

## Data availability

Data supporting the findings of this manuscript are available from the corresponding author upon reasonable request. A reporting summary for this article is available as a Supplementary Information file. A sample of chromosome configurations generated by the MaxEnt model is available on GitHub[65].

## Code availability

The code generating the data and implementing the analysis presented in the manuscript is available on GitHub[65].

Received: 31 March 2020; Accepted: 15 February 2021;
Published online: 30 March 2021

## References

1. Dame, R. T., Rashid, F.-Z. M. & Grainger, D. C. Chromosome organization in bacteria: mechanistic insights into genome structure and function. *Nat. Rev. Genet.* **25**, 1–16 (2019).
2. Dillon, S. C. & Dorman, C. J. Bacterial nucleoid-associated proteins, nucleoid structure and gene expression. *Nat. Rev. Microbiol.* **8**, 185 (2010).
3. Broedersz, C. P. et al. Condensation and localization of the partitioning protein ParB on the bacterial chromosome. *Proc. Natl Acad. Sci. USA* **111**, 8809–8814 (2014).
4. Graham, T. G. et al. ParB spreading requires DNA bridging. *Genes Dev.* **28**, 1228–1238 (2014).
5. Brackley, C. A. et al. Nonequilibrium chromosome looping via molecular slip links. *Phys. Rev. Lett.* **119**, 138101 (2017).
6. Dorman, C. J. Function of nucleoid-associated proteins in chromosome structuring and transcriptional regulation. *J. Mol. Microbiol. Biotechnol.* **24**, 316–331 (2014).
7. Wiggins, P. A., Cheveralls, K. C., Martin, J. S., Lintner, R. & Kondev, J. Strong intranucleoid interactions organize the *Escherichia coli* chromosome into a nucleoid filament. *Proc. Natl Acad. Sci. USA* **107**, 4991–4995 (2010).
8. Weber, S. C., Spakowitz, A. J. & Theriot, J. A. Nonthermal ATP-dependent fluctuations contribute to the in vivo motion of chromosomal loci. *Proc. Natl Acad. Sci. USA* **109**, 7338–7343 (2012).
9. Snijder, B. & Pelkmans, L. Origins of regulated cell-to-cell variability. *Nat. Rev. Mol. Cell Biol.* **12**, 119–125 (2011).
10. Imakaev, M. V., Fudenberg, G. & Mirny, L. A. Modeling chromosomes: beyond pretty pictures. *FEBS Lett.* **589**, 3031–3036 (2015).
11. Robinett, C. C. et al. In vivo localization of DNA sequences and visualization of large-scale chromatin organization using *lac* operator/repressor recognition. *J. Cell Biol.* **135**, 1685–1700 (1996).
12. Cattoni, D. I., Valeri, A., Le Gall, A. & Nollmann, M. A matter of scale: how emerging technologies are redefining our view of chromosome architecture. *Trends Genet.* **31**, 454–464 (2015).
13. Wu, F. et al. Direct imaging of the circular chromosome in a live bacterium. *Nat. Commun.* **10**, 2194 (2019).
14. Teleman, A. A., Graumann, P. L., Lin, D. C. H., Grossman, A. D. & Losick, R. Chromosome arrangement within a bacterium. *Curr. Biol.* **8**, 1102–1109 (1998).
15. Bates, D. & Kleckner, N. Chromosome and replisome dynamics in *E. coli*: loss of sister cohesion triggers global chromosome movement and mediates chromosome segregation. *Cell* **121**, 899–911 (2005).
16. Lau, I. F. et al. Spatial and temporal organization of replicating *Escherichia coli* chromosomes. *Mol. Microbiol.* **49**, 731–743 (2004).
17. Viollier, P. H. et al. Rapid and sequential movement of individual chromosomal loci to specific subcellular locations during bacterial DNA replication. *Proc. Natl Acad. Sci. USA* **101**, 9257–9262 (2004).
18. Toro, E., Hong, S.-H., McAdams, H. H. & Shapiro, L. *Caulobacter* requires a dedicated mechanism to initiate chromosome segregation. *Proc. Natl Acad. Sci. USA* **105**, 15435–15440 (2008).
19. Thanbichler, M. & Shapiro, L. MipZ, a spatial regulator coordinating chromosome segregation with cell division in *Caulobacter*. *Cell* **126**, 147–162 (2006).
20. Lasker, K. et al. Selective sequestration of signalling proteins in a membraneless organelle reinforces the spatial regulation of asymmetry in *Caulobacter crescentus*. *Nat. Microbiol.* **5**, 418–429 (2020).
21. Umbarger, M. A. Chromosome conformation capture assays in bacteria. *Methods* **58**, 212–220 (2012).
22. Le, T. B. K. & Laub, M. T. New approaches to understanding the spatial organization of bacterial genomes. *Curr. Opin. Microbiol.* **22**, 15–21 (2014).
23. Marbouty, M. et al. Condensin- and replication-mediated bacterial chromosome folding and origin condensation revealed by Hi-C and super-resolution imaging. *Mol. Cell* **59**, 588–602 (2015).
24. Lioy, V. S. et al. Multiscale structuring of the *E. coli* chromosome by nucleoid-associated and condensin proteins. *Cell* **172**, 771–783 (2018).
25. Le, T. B. K. & Laub, M. T. Transcription rate and transcript length drive formation of chromosomal interaction domain boundaries. *EMBO J.* **35**, 1582–1595 (2016).
26. Le, T. B. K., Imakaev, M. V., Mirny, L. A. & Laub, M. T. High-resolution mapping of the spatial organization of a bacterial chromosome. *Science* **342**, 731–734 (2013).
27. Umbarger, M. A. et al. The three-dimensional architecture of a bacterial genome and its alteration by genetic perturbation. *Mol. Cell* **44**, 252–264 (2011).
28. Tran, N. T., Laub, M. T. & Le, T. B. K. SMC progressively aligns chromosomal arms in *Caulobacter crescentus* but is antagonized by convergent transcription. *Cell* **20**, 2057–2071 (2017).
29. Wang, X. et al. Condensin promotes the juxtaposition of DNA flanking its loading site in *Bacillus subtilis*. *Genes Dev.* **29**, 1661–1675 (2015).
30. Wang, X., Brandão, H. B., Le, T. B. K., Laub, M. T. & Rudner, D. Z. *Bacillus subtilis* SMC complexes juxtapose chromosome arms as they travel from origin to terminus. *Science* **355**, 524–527 (2017).
31. Böhm, K. et al. Chromosome organization by a conserved condensin-ParB system in the actinobacterium *Corynebacterium glutamicum*. *Nat. Commun.* **11**, 1485 (2020).

32. Trussart, M. et al. Defined chromosome structure in the genome-reduced bacterium *Mycoplasma pneumoniae*. *Nat. Commun.* **8**, 14665 (2017).
33. Yildirim, A. & Feig, M. High-resolution 3D models of *Caulobacter crescentus* chromosome reveal genome structural variability and organization. *Nucleic Acids Res.* **46**, 3937–3952 (2018).
34. Imakaev, M. V., Tchourine, K. M., Nechaev, S. K. & Mirny, L. A. Effects of topological constraints on globular polymers. *Soft Matter* **11**, 665–671 (2015).
35. Oluwadare, O., Highsmith, M. & Cheng, J. An overview of methods for reconstructing 3-D chromosome and genome structures from Hi-C data. *Biol. Proced. Online* **21**, 7 (2019).
36. Tjong, H. et al. Population-based 3D genome structure analysis reveals driving forces in spatial genome organization. *Proc. Natl Acad. Sci. USA* **113**, E1663–1667 (2016).
37. Zhang, B. & Wolynes, P. G. Topology, structures, and energy landscapes of human chromosomes. *Proc. Natl Acad. Sci. USA* **112**, 6062–6067 (2015).
38. Di Pierro, M., Zhang, B., Aiden, E. L., Wolynes, P. G. & Onuchic, J. N. Transferable model for chromosome architecture. *Proc. Natl Acad. Sci. USA* **113**, 12168–12173 (2016).
39. Abbas, A. et al. Integrating Hi-C and FISH data for modeling of the 3D organization of chromosomes. *Nat. Commun.* **10**, 2049 (2019).
40. Marks, D. S. et al. Protein 3D structure computed from evolutionary sequence variation. *PLoS ONE* **6**, e28766 (2011).
41. Javer, A. et al. Short-time movement of *E. coli* chromosomal loci depends on coordinate and subcellular localization. *Nat. Commun.* **4**, 3003 (2013).
42. Smith, K., Griffin, B., Byrd, H., MacKintosh, F. C. & Kilfoil, M. L. Nonthermal fluctuations of the mitotic spindle. *Soft Matter* **11**, 4396–4401 (2015).
43. Tkačik, G. et al. The simplest maximum entropy model for collective behavior in a neural network. *J. Stat. Mechan. Exp.* **2013**, P03011 (2013).
44. Mora, T., Walczak, A. M., Bialek, W. & Callan, C. G. Maximum entropy models for antibody diversity. *Proc. Natl Acad. Sci. USA* **107**, 5405–5410 (2010).
45. Bialek, W. et al. Statistical mechanics for natural flocks of birds. *Proc. Natl Acad. Sci. USA* **109**, 4786–4791 (2012).
46. De Martino, D., MC Andersson, A., Bergmiller, T., Guet, C. C. & Tkačik, G. Statistical mechanics for metabolic networks during steady state growth. *Nat. Commun.* **9**, 2988 (2018).
47. Bialek, W. & Ranganathan, R. Rediscovering the power of pairwise interactions. *arXiv preprint arXiv:0712.4397* (2007).
48. Schneidman, E., Berry, M. J., Segev, R. & Bialek, W. Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature* **440**, 1007–1012 (2006).
49. Lapedes, A., Giraud, B. & Jarzynski, C. Using sequence alignments to predict protein structure and stability with high accuracy. *arXiv preprint arXiv:1207.2484* (2012).
50. Pressé, S., Ghosh, K., Lee, J. & Dill, K. A. Principles of maximum entropy and maximum caliber in statistical physics. *Rev. Modern Phys.* **85**, 1115–1141 (2013).
51. Pal, K., Forcato, M. & Ferrari, F. Hi-C analysis: from data generation to integration. *Biophys. Rev.* **11**, 67–78 (2019).
52. Degnen, S. T. & Newton, A. Chromosome replication during development in *Caulobacter crescentus*. *J. Mol. Biol.* **64**, 671–680 (1972).
53. Bürmann, F. & Gruber, S. SMC condensin: Promoting cohesion of replicon arms. *Nat. Struct. Mol. Biol.* **22**, 653–655 (2015).
54. Miermans, C. A. & Broedersz, C. P. Bacterial chromosome organization by collective dynamics of SMC condensins. *J. R. Soc. Interf.* **15**, 20180495 (2018).
55. Ganji, M. et al. Real-time imaging of DNA loop extrusion by condensin. *Science* **360**, 102–105 (2018).
56. Wang, X., Llopis, P. M. & Rudner, D. Z. Organization and segregation of bacterial chromosomes. *Nat. Rev. Genet.* **14**, 191–203 (2013).
57. Dubuis, J. O., Tkacik, G., Wieschaus, E. F., Gregor, T. & Bialek, W. Positional information, in bits. *Proc. Natl Acad. Sci. USA* **110**, 16301–16308 (2013).
58. Shin, Y. & Brangwynne, C. P. Liquid phase condensation in cell physiology and disease. *Science* **357**, eaaf4382 (2017).
59. Brandão, H. B. et al. RNA polymerases as moving barriers to condensin loop extrusion. *Proc. Natl Acad. Sci.* **116**, 20489–20499 (2019).
60. Ozaki, S., Jenal, U. & Katayama, T. Novel divisome-associated protein spatially coupling the Z-ring with the chromosomal replication terminus in *Caulobacter crescentus*. *mBio* **11,** 0487-20 (2020).
61. Cavagna, A. et al. Dynamical maximum entropy approach to flocking. *Phys. Rev. E* **89**, 042707 (2014).
62. Pressé, S., Ghosh, K., Lee, J. & Dill, K. A. Principles of maximum entropy and maximum caliber in statistical physics. *Rev. Modern Phys.* **85**, 1115–1141 (2013).
63. Hensel, Z., Weng, X., Lagda, A. C. & Xiao, J. Transcription-factor-mediated DNA looping probed by high-resolution, single-molecule imaging in live *E. coli* cells. *PLoS Biol.* **11**, e1001591 (2013).
64. Gaal, T. et al. Colocalization of distant chromosomal loci in space in *E. coli*: a bacterial nucleolus. *Genes Dev.* **30**, 2272–2285 (2016).
65. Messelink, J., van Teeseling, M., Janssen, J., Thanbichler, M. & Broedersz, C. Learning the distribution of single-cell chromosome conformations in bacteria reveals emergent order across genomic scales. *GitHub Repository* https://doi.org/10.5281/zenodo.4435038 (2021).


## Acknowledgements

We thank Tung Le for helpful discussions and for generously making experimental data available. In addition, we thank Ben Machta for inspiring discussions, Karsten Miermans and Lucas Tröger for valuable input for the simulations, Gabriele Malengo (Facility for Flow Cytometry and Imaging, MPI Marburg) for help with the super-resolution microscopy, and Maritha Lippmann for excellent technical assistance. This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation, Project 269423233-TRR 174). J.M. is supported by a DFG fellowship within the Graduate School of Quantitative Biosciences Munich (QBM).

## Author contributions

C.P.B. conceived the project; J.J.B.M. performed analyses, simulations, and analysis of microscopy data, J.J. and J.J.B.M. developed the MC algorithm, M.T. and M.C.F.vT. conceived microscopy experiments, M.C.F.vT. performed microscopy experiments and analyzed microscopy data, C.P.B. and J.J.B.M. wrote the paper with input from all authors.

## Funding

Open Access funding enabled and organized by Projekt DEAL.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1038/s41467-021-22189-x.

**Correspondence** and requests for materials should be addressed to C.P.B.

**Peer review information** *Nature Communications* thanks the anonymous reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

**Reprints and permission information** is available at http://www.nature.com/reprints

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Supplementary Information**

# Learning the distribution of single-cell chromosome conformations in bacteria reveals emergent order across genomic scales

J.J.B. Messelink et al.

# Contents

# 1 Experimental procedures on *C. crescentus* cells

**Bacterial strains and growth conditions**

All *C. crescentus* strains used in this study were derived from the synchronizable wild-type CB15N (NA1000). Cells were grown in peptone-yeast extract (PYE) medium (Pointdexter, 1964) at $28°C$ under aerobic conditions (shaking at 210 rpm). When appropriate, the medium was supplemented with antibiotics at the following concentrations ($\mu$g ml$^{-1}$ in liquid/solid medium): kanamycin (30/50), gentamicin (15/20), and spectinomycin (50/100).

## 1.1 Experimental determination of distances between loci 10 kb apart

**Plasmid and strain construction**

To measure the distances between chromosomal loci that are located 10 kb apart, *C. crescentus* strains were constructed whose chromosomes contained binding sites for fluorescently tagged DNA binding proteins. The bacterial strains, plasmids, and oligonucleotides used in this study are listed in Tables S1-S4. *Escherichia coli* TOP10 (Invitrogen) was used as host for cloning purposes. All plasmids were verified by DNA sequencing. Plasmids carrying 10 copies of either *lac*O (PCR-amplified from plasmid pLAU43 [1]) or *tetO* (PCR-amplified from plasmid pLAU44 [1]) were transferred to *C. crescentus* by electroporation [2] and integrated at various chromosomal loci by single-homologous recombination. Subsequently, a two-gene operon encoding LacI-eCFP and TetR-eYFP was integrated at the *xylX* locus by phiCr30-mediated phage transduction [2], using a lysate of a strain transformed with plasmid pHPV472 [3]. Proper chromosomal integration was verified by colony PCR.

**Measurement of distance between pairs of loci 10 kb apart**

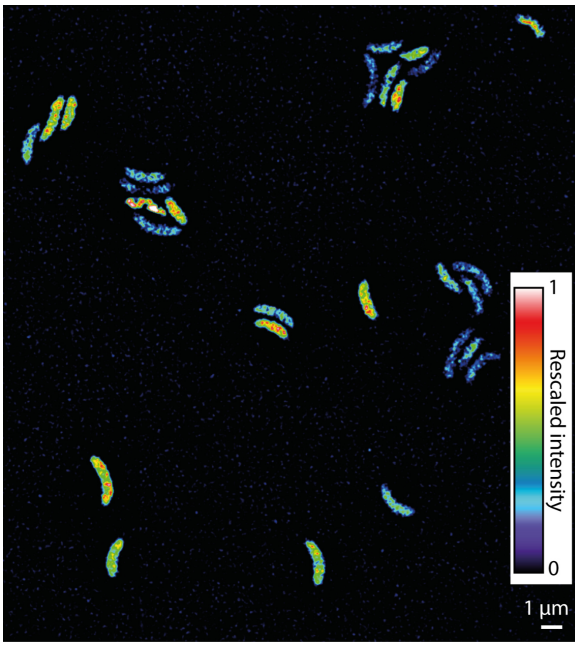
All microscopy analyses to determine the distance between chromosomal loci were performed on cells grown in PYE medium containing kanamycin and gentimicin to the mid-exponential phase (OD 0.4), and subsequently synchronized [4]. Immediately after synchronization, swarmer cells were immobilized on pads made of 1% agarose in PYE medium. Cells were observed with a Zeiss Axio Observer.Z1 microscope equipped with an alpha Plan-Apochromat 100x/1.46 Oil Ph3 objective (Zeiss, Germany). An X-Cite 120PC metal halide light source (EXFO, Canada), combined with ET-CFP and ET-YFP filter cubes (Chroma, USA), was used for the detection of fluorescent foci. Images were taken with a pco.edge sCMOS camera (pco, Germany) and recorded with VisiView 2.1.4 (Visitron, Germany). To identify the subpixel localization of the fluorescent foci, a 2D Gaussian was fitted to each fluorescent focus using the GDSC SMLM plugin [5] [1] for ImageJ2 [6]. In order to correct for systematic shifts between the YFP and CFP channels, fiducials (Tetraspeck microspheres, 0.5 $\mu$m, Invitrogen/Thermo Fischer Scientific, USA) were imaged in the YFP and CFP channels and analyzed with the same set-up and pipeline.

## 1.2 Determination of chromosome density via SIM microscopy

In order to investigate the intracellular distribution of the chromosome, *C. crescentus* wild-type cells were grown and synchronized as described above. Immediately after synchronization, the cells

[1] http://www.sussex.ac.uk/gdsc/intranet/microscopy/UserSupport/AnalysisProtocol/imagej/smlm_plugins/

were incubated with 0.5 $\mu$g/ml of the DNA-stain DAPI (4',6-diamino-2-phenylindole) for 5 min at 28°C. Cells were then washed (5 min at 4000 g), resuspended in M2 salts buffer [2] and applied to pads made of 1% agarose in water, before they were imaged with a Zeiss Elyra 7 Lattice SIM microscope equipped with an alpha Plan-Apochromat 100x/1.46 Oil Objective (Zeiss, Germany). DAPI was excited with a 405 nm laser and its emission was recorded in the 420-480 nm range.




**Supplementary Figure 1: SIM microscopy image example** SIM microscopy image of a single focal plane out of a z-stack shows the DAPI-stained DNA inside multiple *C. crescentus* cells immediately after synchronization. The DNA is organized in a heterogeneous fashion, with several regions of high-density chromosome packing per cell, and shows a clear cell-to-cell variation. The intensity is rescaled for the entire image, such that the highest measured intensity is 1, and the lowest is 0. Scale bar: 1 $\mu$m. Shown is a representative image of one of the two biological replicates, which both showed similar results.

**Supplementary Table 1:** Strains used in this study.

| Strain | Genotype/description | Construction/Reference |
|---|---|---|
| ***E. coli* strains** | | |
| TOP10 | Cloning strain | Invitrogen |
| ***C. crescentus* strains** | | |
| CB15N | Synchronizable wild-type strain | Evinger & Agabian (1977) [7] |
| MvT151 | CB15N $P_{xyl}$::$P_{xyl}$-*lacI-ecfp-tetR-eyfp* 10x *tetO* and 10x *lacO* spaced 10.0 kb apart at 196° | Consecutive integration of pMvT149, pMvT150 and $P_{xyl}$-*lacI-ecfp-tetR-eyfp* into CB15N |
| MvT152 | CB15N $P_{xyl}$::$P_{xyl}$-*lacI-ecfp-tetR-eyfp* 10x *tetO* and 10x *lacO* spaced 10.1 kb apart at 212° | Consecutive integration of pMvT151, pMvT152 and $P_{xyl}$-*lacI-ecfp-tetR-eyfp* into CB15N |
| MvT170 | CB15N $P_{xyl}$::$P_{xyl}$-*lacI-ecfp-tetR-eyfp* 10x *tetO* and 10x *lacO* spaced 10.1 kb apart at 21° | Consecutive integration of pMvT161, pMvT162 and $P_{xyl}$-*lacI-ecfp-tetR-eyfp* into CB15N |
| MvT171 | CB15N $P_{xyl}$::$P_{xyl}$-*lacI-ecfp-tetR-eyfp* 10x *tetO* and 10x *lacO* spaced 10.0 kb apart at 108° | Consecutive integration of pMvT163, pMvT164 and $P_{xyl}$-*lacI-ecfp-tetR-eyfp* into CB15N |
| MvT172 | CB15N $P_{xyl}$::$P_{xyl}$-*lacI-ecfp-tetR-eyfp* 10x *lacO* and 10x *tetO* spaced 10.0 kb apart at 108° | Consecutive integration of pMvT165, pMvT166 and $P_{xyl}$-*lacI-ecfp-tetR-eyfp* into CB15N |
| MvT179 | CB15N $P_{xyl}$::$P_{xyl}$-*lacI-ecfp-tetR-eyfp* 10x *tetO* and 10x *lacO* spaced 10.1 kb apart at 311° | Consecutive integration of pMvT159, pMvT160 and $P_{xyl}$-*lacI-ecfp-tetR-eyfp* into CB15N |

**Supplementary Table 2:** Plasmids used in this study.

| Plasmid | Description | Reference |
|---|---|---|
| **Basic vectors** | | |
| pLAU43 | Plasmid carrying 240 LacI binding sites (*lacO*), $Kan^R$ | Lau et al., 2003 [1] |
| pLAU44 | Plasmid carrying 240 TetO binding sites (*tetO*), $Gen^R$ | Lau et al., 2003 [1] |
| pHPV472 | Plasmid carrying $P_{xyl}$-*lacI-ecfp tetR-eyfp*, $Spc^R$ $Str^R$ | Viollier et al., 2004 [3] |
| pMCS-2 | Integrating plasmid containing multiple cloning site, $Kan^R$ | Thanbichler et al., 2007 [8] |
| pMCS-4 | Integrating plasmid containing multiple cloning site, $Gen^R$ | Thanbichler et al., 2007 [8] |
| **Plasmids constructed in this work** | | |
| pMvT149 | pMCS-2 including 10x *tetO* and part of CCNA_02049, $Kan^R$ | This study |
| pMvT150 | pMCS-4 including 10x *lacO* and part of a chromosomal fragment close to CCNA_02054, $Gen^R$ | This study |
| pMvT151 | pMCS-2 including 10x *tetO* and part of a chromosomal fragment close to CCNA_02228, $Kan^R$ | This study |
| pMvT152 | pMCS-4 including 10x *lacO* and part of a chromosomal fragment close to CCNA_02233, $Gen^R$ | This study |
| pMvT159 | pMCS-2 including 10x *tetO* and part of CCNA_03310, $Kan^R$ | This study |
| pMvT160 | pMCS-4 including 10x *lacO* and part of a chromosomal fragment close to CCNA_03317, $Gen^R$ | This study |
| pMvT161 | pMCS-2 including 10x *tetO* and part of CCNA_00217, $Kan^R$ | This study |
| pMvT162 | pMCS-4 including 10x *lacO* and part of a chromosomal fragment close to CCNA_00226, $Gen^R$ | This study |
| pMvT163 | pMCS-2 including 10x *tetO* and part of CCNA_01105, $Kan^R$ | This study |
| pMvT164 | pMCS-4 including 10x *lacO* and part of a chromosomal fragment close to CCNA_01112, $Gen^R$ | This study |
| pMvT165 | pMCS-2 including 10x *lacO* and part of CCNA_01105, $Kan^R$ | This study |
| pMvT166 | pMCS-4 including 10x *tetO* and part of a chromosomal fragment close to CCNA_01112, $Gen^R$ | This study |

**Supplementary Table 3:** Construction of plasmids.

| Plasmid | Description |
|---|---|
| pMvT149 | a) amplification of 10 *tetO* motifs from pLAU44 using oMvT789 & oMvT790 (product 433 bp) and 800 bp from NA1000 gDNA using oMvT791 & oMvT792 (product 848 bp) |
| | b) fusion of two inserts with pMCS-2/NdeI+NheI via Gibson Assembly |
| pMvT150 | a) amplification of 10 *lacO* motifs from pLAU43 using oMvT796 & oMvT797 (product 547 bp) and 800 bp from NA1000 gDNA using oMvT798 & oMvT799 (product 845 bp) |
| | b) fusion of two inserts with pMCS-2/NdeI+NheI via Gibson Assembly |
| pMvT151 | a) amplification of 10 *tetO* motifs from pLAU44 using oMvT803 & oMvT804 (product 435 bp) and 800 bp from NA1000 gDNA using oMvT805 & oMvT806 (product 843 bp) |
| | b) fusion of two inserts with pMCS-2/NdeI+NheI via Gibson Assembly |
| pMvT152 | a) amplification of 10 *lacO* motifs from pLAU43 using oMvT808 & oMvT809 (product 548 bp) and 800 bp from NA1000 gDNA using oMvT810 & oMvT811 (product 845 bp) |
| | b) fusion of two inserts with pMCS-2/NdeI+NheI via Gibson Assembly |
| pMvT159 | a) amplification of 10 *tetO* motifs from pLAU44 using oMvT789 & oMvT839 (product 435 bp) and 800 bp from NA1000 gDNA using oMvT840 & oMvT841 (product 843 bp) |
| | b) fusion of two inserts with pMCS-2/NdeI+NheI via Gibson Assembly |
| pMvT160 | a) amplification of 10 *lacO* motifs from pLAU43 using oMvT819 & oMvT842 (product 549 bp) and 800 bp from NA1000 gDNA using oMvT843 & oMvT844 (product 851 bp) |
| | b) fusion of two inserts with pMCS-2/NdeI+NheI via Gibson Assembly |
| pMvT161 | a) amplification of 10 *tetO* motifs from pLAU44 using oMvT789 & oMvT849 (product 436 bp) and 800 bp from NA1000 gDNA using oMvT850 & oMvT851 (product 840 bp) |
| | b) fusion of two inserts with pMCS-2/NdeI+NheI via Gibson Assembly |
| pMvT162 | a) amplification of 10 *lacO* motifs from pLAU43 using oMvT819 & oMvT854 (product 549 bp) and 800 bp from NA1000 gDNA using oMvT855 & oMvT856 (product 844 bp) |
| | b) fusion of two inserts with pMCS-2/NdeI+NheI via Gibson Assembly |
| pMvT163 | a) amplifation of 10 *tetO* motifs from pLAU44 using oMvT789 & oMvT859 (product 436 bp) and 800 bp from NA1000 gDNA using oMvT860 & oMvT861 (product 848 bp) |
| | b) fusion of two inserts with pMCS-2/NdeI+NheI via Gibson Assembly |
| pMvT164 | a) amplification of 10 *lacO* motifs from pLAU43 using oMvT819 & oMvT863 (product 547 bp) and 800 bp from NA1000 gDNA using oMvT864 & oMvT865 (product 851 bp) |
| | b) fusion of two inserts with pMCS-2/NdeI+NheI via Gibson Assembly |
| pMvT165 | a) amplification of 10 *lacO* motifs from pLAU43 using oMvT819 & oMvT867 (product 548 bp) and 800 bp from NA1000 gDNA using oMvT868 & oMvT861 (product 845 bp) |
| | b) fusion of two inserts with pMCS-2/NdeI+NheI via Gibson Assembly |
| pMvT166 | a) amplification of 10 *tetO* motifs from pLAU44 using oMvT789 & oMvT869 (product 435 bp) and 800 bp from NA1000 gDNA using oMvT870 & oMvT865 (product 848 bp) |
| | b) fusion of two inserts with pMCS-2/NdeI+NheI via Gibson Assembly |

**Supplementary Table 4:** Oligonucliotides used in this study.

| ID | Name | Sequence (5' to 3') |
|---|---|---|
| oMvT789 | tetO _CCNA_02049_p1 | cgagacgtccaattgcatatgtccctatcagtgatagagaggggaaagg |
| oMvT790 | tetO _CCNA_02049_p2 | cgccgctggccaccggatctctatcactgatagggaccttcccttctg |
| oMvT791 | tetO _CCNA_02049_p3 | gggaaggtccctatcagtgatagagatccggtggccagcggcgaac |
| oMvT792 | tetO _CCNA_02049_p4 | gatcccccgggctgcagctagcgcgcactgaggccgatggcg |
| oMvT796 | lacO _CCNA_02049_p1 | gcgagacgtccaattgcatatgttgtgagcggataacaattggagcaag |
| oMvT797 | lacO _CCNA_02049_p2 | cttcgaccgctgggacttcttgttatccgctcacaatttgccttttgc |
| oMvT798 | lacO _CCNA_02049_p3 | ggcaaattgtgagcggataacaagaagtcccagcggtcgaagaggacg |
| oMvT799 | lacO _CCNA_02049_p4 | gatccccgggctgcagctagcgcctatgacgtgatgagctccaagcac |
| oMvT803 | tetO _CCNA_02228_p1 | cgagacgtccaattgcatatgtccctatcagtgatagagaggggaaagg |
| oMvT804 | tetO _CCNA_02228_p2 | gacgacccctactggtcctctctatcactgatagggaccttccc |
| oMvT805 | tetO _CCNA_02228_p3 | ggtccctatcagtgatagagaggaccagtaggggtcgtcgaacg |
| oMvT806 | tetO _CCNA_02228_p4 | gatccccgggctgcagctagcccagccccgccgccgacatcg |
| oMvT808 | lacO _CCNA_02228_p1 | gcgagacgtccaattgcatatgttgtgagcggataacaattggagcaag |
| oMvT809 | lacO _CCNA_02228_p2 | cccaggcaacttgtctttcgttgttatccgctcacaatttgccttttgc |
| oMvT810 | lacO _CCNA_02228_p3 | ggcaaattgtgagcggataacaacgaaagacaagttgcctgggc |
| oMvT811 | lacO _CCNA_02228_p4 | gatccccgggctgcagctagcctagcggatcgggcgcgcgaag |
| oMvT819 | lacO _CCNA_01737_p1 | gcgagacgtccaattgcatatgttgtgagcggataacaattggagcaag |
| oMvT839 | tetO _nusG_p2 | ggtcgaaaagatcgcctgatctctatcactgatagggaccttcccttc |
| oMvT840 | tetO _nusG_p3 | ggtccctatcagtgatagagatcaggcgatcttttcgacctgattg |
| oMvT841 | tetO _nusG_p4 | gatccccgggctgcagctagccgcgacagccgccgccgctcc |
| oMvT842 | lacO _CC-3211_p2 | gcagccgcgatttccattgagttgttatccgctcacaatttgccttttg |
| oMvT843 | lacO _CC-3211_p3 | ggcaaattgtgagcggataacaactcaatggaaatcgcggctgcgg |
| oMvT844 | lacO _CC-3211_p4 | ctagtggatcccccgggctgcagctagcctgccaggagacgcggcc |
| oMvT849 | tetO _CC-0217_p2 | cagcgcatagcccagcgcgctctctatcactgatagggaccttcccttc |
| oMvT850 | tetO _CC-0217_p3 | ggtccctatcagtgatagagagcgcgctgggctatgcgctgac |
| oMvT851 | tetO _CC-0217_p4 | cccccgggctgcagctagcctagctccccgccctctcgatcg |
| oMvT854 | lacO _CC-0226_p2 | caactatgtcgatgacgagcattgttatccgctcacaatttgccttttg |
| oMvT855 | lacO _CC-0226_p3 | caaattgtgagcggataacaatgctcgtcatcgacatagttgctgcg |
| oMvT856 | lacO _CC-0226_p4 | ggatcccccgggctgcagctagcgtgatgaccaagaccatgcttctggc |
| oMvT859 | tetO _CC-1053_p2 | gcccagatgccggcgcaatctctctatcactgatagggaccttcccttc |
| oMvT860 | tetO _CC-1053_p3 | gggaaggtccctatcagtgatagagagattgcgccggcatctgggcc |
| oMvT861 | tetO _CC-1053_p4 | gatcccccgggctgcagctagcggcaggatcgaccaccgcgc |
| oMvT863 | lacO _CC-1059_p2 | ccagttcgcagagccggcgttgttatccgctcacaatttgccttttgc |
| oMvT864 | lacO _CC-1059_p3 | caaaaggcaaattgtgagcggataacaacgccggctctgcgaactggag |
| oMvT865 | lacO _CC-1059_p4 | ggatcccccgggctgcagctagctcatgccatccggtagtgtcgggc |
| oMvT867 | lacO _CC-1053_p2 | gcccagatgccggcgcaatcttgttatccgctcacaatttgccttttgc |
| oMvT868 | lacO _CC-1053_p3 | ggcaaattgtgagcggataacaagattgcgccggcatctgggc |
| oMvT869 | tetO _CC-1059_p2 | ccagttcgcagagccggcgtctctatcactgatagggaccttcccttc |
| oMvT870 | tetO _CC-1059_p3 | ggaaggtccctatcagtgatagagacgccggctctgcgaactggag |

## 2 Data analysis: using experimental distance distributions to set the coarse-grained representation of the lattice polymer

We require a coarse-grained representation of the bacterial chromosome that is consistent with experimentally determined statistics beyond the coarse-graining length scale. Furthermore, our coarse-grained representation should allow for efficient computation. The resolution of the Hi-C data set (10 kb) sets a natural coarse-graining scale for the polymer, but we require additional experiments for the statistics at this length-scale: the distribution of spatial distances between pairs of loci at a 10 kb genomic distance. Here we demonstrate that a lattice polymer representation of the chromosome captures the statistics at this length scale. In this representation, the measured average spatial distance between a pair of loci sets the lattice spacing of our representation of the bacterial chromosome.

### 2.1 Analysis of experimental distance distributions of pairs of loci in *C. crescentus*

From the experimental procedure described in Note 1, a data set of 100 2D distance vectors are obtained in *C. crescentus* for five pairs of loci separated by 10 kb. Note, microscopy data only gives us the projected 2D distances, while the actual distance vectors are in 3D. From the 2D data set, however, we can infer the underlying distribution of 3D distances. To make this inference, two effects are considered:

1. Measurement errors. This has two sources: finite localization precision and drift between the two consecutive images, taken to determine the positions of the two fluorescently (YFP and CFP) labeled loci using two different fluorescence channels.

The measurement noise due to finite localization precision depends on the intensity of the fluorescent probe and the brightness of its direct surroundings. We calculated this precision using the GDSC SMLM plugin to have a standard error of 32.63 nm, with an average variation between measurements of 0.02 nm.

To account for drift between two consecutive images, we decompose the distance vector within each pair of foci into an $x$ and $y$ component, and sum these two components separately for all cells. As the orientations of cells are isotropically distributed, both the $x$ and $y$ component sums should go to 0 for increasing sample size. However, we find significant deviations from 0, larger than expected with our finite sampling, indicating a systematic drift estimated to be $35 \pm 4$nm in the $x$-direction, and $52 \pm 5$nm in the $y$-direction (error on the mean). We correct for these deviations by subtracting the systematic drift in the $x$ and $y$ directions from each of the experimentally measured distance vectors, from which a model for the 3D distance distribution is inferred. This correction will, however, be an overestimate: for a finite sample size, the $x$ and $y$ component sums will likely deviate from 0, even in the absence of drift. To account for this bias in the drift estimator, we simulate finite sampling of 2D distance vectors (using the same number of data points as in the experiments) from the inferred model for the 3D distance distribution. Note, we require a self-consistent iterative procedure: the bias in the drift estimator that we correct for, when inferring the 3D model from measured 2D distances, must be consistent with the bias we determine when performing a finite sampling of 2D distances from this model.

2. We consider intrinsic variations in 3D distances between the loci, for instance due to thermal fluctuations of the DNA. We assume that the underlying distribution of relative positions is de-

**Supplementary Table 5:** Inferred average distances for the measured pairs of loci. The data sets MvT171 and MvT172 are for the same loci, just with their markers switched (see Note 1). The determined distances for each of these pairs are within two standard deviations of each other.

| Data set | Average 3D distance (nm) | Inferred $\sigma$ (nm) |
|---|---|---|
| MvT151 | 106 ± 7 | 67 |
| MvT170 | 134 ± 8 | 84 |
| MvT171 | 121 ± 8 | 76 |
| MvT152 | 158 ± 9 | 99 |
| MvT172 | 132 ± 8 | 83 |
| MvT179 | 124 ± 7 | 78 |
| **Inferred average for entire chromosome** | Average 3D distance (nm) | Variance (nm) |
| | 129 ± 7 | 17 |

scribed by a 3D Gaussian with a standard deviation and a mean equal to 0. This results in one fit parameter ($\sigma$) for the underlying distribution.

To determine the value of $\sigma$ for each of the pairs of loci, we also use an iterative procedure: we start by choosing an initial value of $\sigma$, and then simulate the sampling of a large number of 3D distance vectors from this distribution. We then take a 2D projection of these samples and add the random measurement error of 32.63 nm (see point 1). Next, we compute the average 2D distance and compare with the experimentally determined 2D average distance. If these values are not equal, the value of $\sigma$ is updated accordingly, and a new round of the iteration begins. This procedure is repeated until convergence is reached (the average 2D distance is equal to the experimentally determined 2D average distance).

Once convergence is reached, the mean 3D distance for each pair of loci is calculated through a forward simulation of random points being drawn from a 3D Gaussian. The error on the mean inferred 3D distance for a specific pair of loci on the chromosome is determined by bootstrapping (see Table 5). The average distance for the entire chromosome is taken as the average over the means of the 5 pairs of loci we studied experimentally, and is determined to be $129 \pm 7$ nm (standard error of the mean).

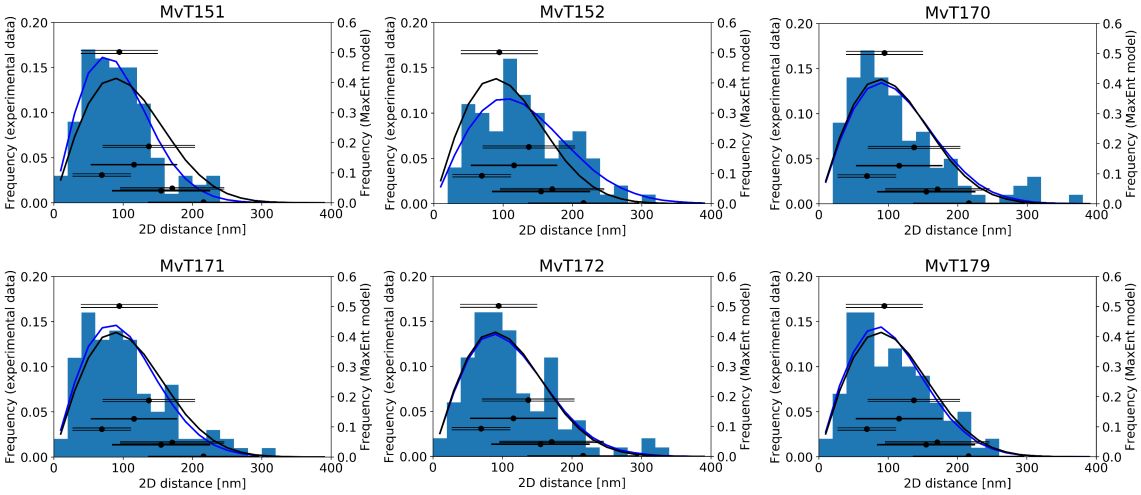
Once the average distances are matched between model and experiment, the distributions of measured distances can also be compared. This distribution matches well between model and experiment (Supplementary Fig. 2), supporting the assumption of a 3D gaussian as an underlying distribution of relative positions between the loci. Once we set the lattice constant of our lattice polymer to match this average 3D distance, our lattice polymer model approximately captures the correct statistics for the distance between neighboring chromosomal regions. This validates the use of a lattice polymer to connect consecutive monomers representing neighboring chromosomal regions.

## 2.2 Setting the dimensions of the lattice spacing and the cellular confinement in the model

We employ a polymer model on a cubic lattice. In this representation, the position of each fourth monomer indicates the unit cell occupied by the center of a Hi-C chromosomal region. The polymer model is allowed to intersect, since multiple centers of genomic regions could reside in the same unit cell volume. In fact, two monomers are assigned a contact probability $\gamma$ only if they simultaneously occupy the same lattice site. This assumes that the dominant contributions to contacts between two chromosomal regions are from configurations where their respective centers occupy the same unit cell. Any excluded volume effects reducing the number of self-overlaps of the coarse-grained polymer manifest through imposed Hi-C score constraints.

To set the scale of the lattice spacing $b$ in the model, we use the average spatial distance between consecutive Hi-C chromosomal regions determined in Note 2.1). If we consider distances between subsequent chromosomal regions, however, coarse-graining effects need to be taken into account: only seven distances between these regions are possible in the lattice representation (Supplementary Fig. 4 B): (0, $\sqrt{2}b$, $2b$, $\sqrt{6}b$, $\sqrt{8}b$, $\sqrt{10}b$, $4b$), which occur with respective relative occurrence frequencies $(f_1, \cdots, f_7)$. In our MaxEnt model, we robustly observe ($f_1 \approx 0.092$, $f_2 \approx 0.50$, $f_3 \approx 0.13$, $f_4 \approx 0.19$, $f_5 \approx 0.041$, $f_6 \approx 0.048$, $f_7 \approx 0.0022$). This coarse-graining effect implies a cut-off of the tail of the underlying Gaussian distribution of 3D distances. To account for this cut-off, we first sample real-space configurations of consecutive chromosomal regions according to the experimentally determined 3D Gaussian distribution of continuous distances (see Note 2.1), and infer the statistics in the corresponding lattice model. For each of the seven possible (discretized) distances in the coarse-grained lattice representation, we thus obtain associated conditional distribution of real-space distances. The sum of the seven conditional real-space distance distributions, weighted by their respective relative occurrence frequencies $(f_i)$, defines the full distribution of distances between neighbouring chromosomal regions in the MaxEnt model. We determine the lattice spacing $b = 88$ nm, such that the average distance between chromosomal regions in our MaxEnt model matches the experimentally determined average distance (Note 2.1)). Note, for this lattice spacing, the distribution of distances between neighbouring chromosomal regions in the MaxEnt model are also in accordance with our experimentally determined distributions (Supplementary Fig. 2).

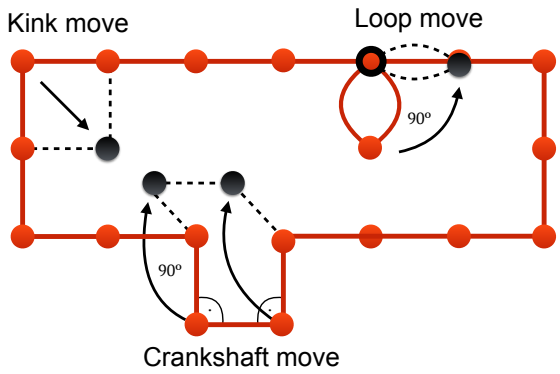
The phase space of chromosome states is restricted to those that fit inside a cell, the sampling thus explores a constrained space (see also [9]). We introduce a confinement formed by a cylinder capped by two hemispheres. The dimensions of the confinement are chosen to match typical dimensions of a newborn swarmer cell. These dimensions are determined by taking a sample of 267 cells from the MvT151 data set, which yields an average length of $2.3 \pm 0.2, \mu m$ and width of $0.75 \pm 0.04 \mu m$, as determined by using the BacStalk software [10]. Subtracting the estimated width of the cell envelope of 61 nm (based on figure 2 of [11]), we arrive at typical chromosome confinement dimensions of $2.2 \times 0.63 \mu m$. With the inferred lattice spacing, this translates to a confinement of 470 unit cells (25 lattice spacings long and 7 wide). This representation of the cell could be refined further to include the crescent shape, but we find that such corrections do not appear to significantly affect the results of our model (see Note 17).




**Supplementary Figure 2: Distributions of 2D projected distances from experiment and MaxEnt model.** Bars: experimentally measured 2D distances (after bias correction, see Note 2.1). Blue lines: distributions of 2D projected distances from the inferred 3D Gaussian distribution. For each data set there is one fit parameter $\sigma$, chosen such that the average distances of measured and inferred distributions match. Black markers: Relative frequencies ($f_i$) of each of the seven possible configurations of two neighboring chromosomal regions of the MaxEnt model with associated average distances determined from coarse-graining. The pairs of horizontal black lines at each dot indicate the mean variance of the MaxEnt configuration frequency for all neighboring pairs of chromosomal regions. The error bar indicates the standard deviation of the underlying distance distribution for each coarse-grained configuration. Black curve: Inferred 2D distance distribution between consecutive genomic regions for the entire chromosome for the MaxEnt model. This distribution is obtained by weighing the inferred distance distribution for each coarse-grained configuration with the associated relative occupancy frequency within the MaxEnt model. To enable a direct comparison with experimental data, the inferred measurement noise is applied over the MaxEnt distance distribution. Note that all MaxEnt data sets are the same in each panel.

# 3 Inverse Monte Carlo algorithm for MaxEnt chromosome model

We solve the inverse problem and obtain the Lagrange multipliers $\epsilon_{ij}$'s by an iterative procedure: we perform a Monte Carlo (MC) simulation (forward algorithm) to sample equilibrium states from the lattice polymer model with an initial guess for $\epsilon_{ij}$. Subsequently, we compare the estimated contact map, $f_{ij}^{\text{sim}}$, obtained from this MC simulation, with the target experimental map $f_{ij}^{\text{expt}}$. When the modeled and experimental contacts deviate, the $\epsilon_{ij}$'s are updated (inverse algorithm). This procedure converges when the modelled normalized contact frequency map matches the Hi-C data set within a tolerance level, yielding the complete set of parameters $\epsilon_{ij}$ that defines the MaxEnt model. The forward and inverse algorithm are described below.




**Supplementary Figure 3: Illustration of the three polymer moves employed in the Monte Carlo simulation.** The simulation employs a kink move, a crankshaft move and a loop move.

## 3.1 Forward algorithm

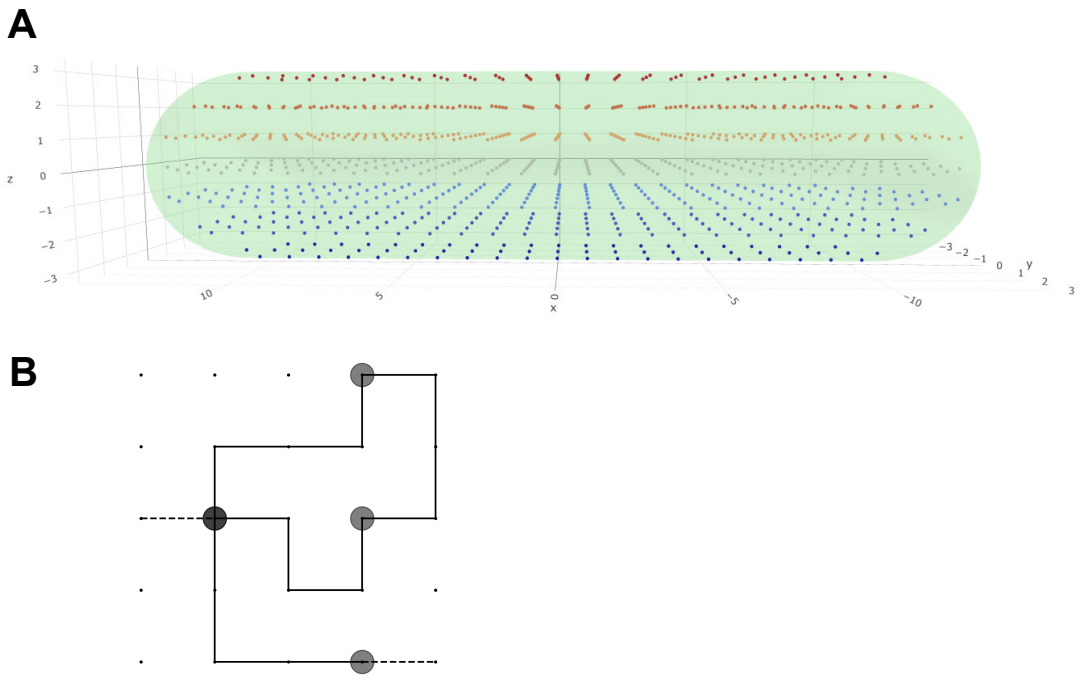
In our coarse-grained model, the bacterial chromosome of *C. crescentus* is represented by a circular lattice polymer with a length of 1620 monomers. Each $4^{\text{th}}$ monomer represents the location of the center of a genomic region, with three monomers in between to ensure Gaussian statistics between subsequent centers of genomic regions (see Note S2). The level of coarse-graining can be adapted to accommodate the resolution of the data on which the model is trained.

The algorithm is initiated with the circular polymer randomly arranged within the confinement. This starting state is obtained by first 'winding up' the polymer in a square that fits in the confinement. Subsequently, a simulation with no interaction energies is run for $10^7$ Monte Carlo moves. The resulting configuration is used as the starting configuration. We simulate the Boltzmann distribution of polymer configurations in the MaxEnt model using Monte Carlo simulations. To sample configurations in the Monte Carlo algorithm, we employ three different polymer moves: the *kink move*, the *Crankshaft move* and the *loop move* (Supplementary Fig. 3). This move set preserves circularity and allows an ergodic sampling of the space of polymer configurations, which is demonstrated in Note 3.3. Moves which would place a monomer outside of the confinement are forbidden.

A potential move $\{\mathbf{r}\} \rightarrow \{\mathbf{r}'\}$ is randomly chosen (based on the move set in Supplementary Fig. 3), and then accepted with a probability $P_{\text{acc}}(\{\mathbf{r}'\},\{\mathbf{r}\})$ according to the Metropolis criterion: $P_{\text{acc}}(\{\mathbf{r}'\},\{\mathbf{r}\}) = \min(1, \exp(E(\{\mathbf{r}\}) - E(\{\mathbf{r}'\})))$, provided the configuration stays within the confinement. Here, $E(\{\mathbf{r}'\})$ and $E(\{\mathbf{r}\})$ are the energies of the proposed configuration $\{\mathbf{r}'\}$ and current configuration $\{\mathbf{r}\}$, respectively. The energies are computed according to the Hamiltonian (Eq. (5) in main text)

$$H(\{\mathbf{r}\}) = \frac{1}{2}\sum_{ij} \epsilon_{ij}\delta_{\mathbf{r}_i,\mathbf{r}_j}. \tag{S1}$$

For pairs of genomic regions $i, j$ for which $\tilde{f}_{ij}^{\text{expt}} = 0$, the corresponding $\epsilon_{ij}$ is set to a high value at the start of the simulation, typically 10, which may further increase during iterations of the inverse algorithm. Note, this initial value is high enough to ensure these contacts do not form in practice. At the start of the forward simulation, we apply a burn in time of $2 \times 10^7$ MC moves before contact




**Supplementary Figure 4: Illustration of the model confinement and chromosome representation A** The cellular confinement used in the simulations. Each dot represents a lattice point. **B** Illustration of the coarse-grained representation of the chromosome, which is shown here in 2D for simplicity. The chromosome is represented by a lattice polymer, where each fourth monomer describes the position of the center of a genomic region. The three monomers in between centers of genomic regions serve to ensure correct distance statistics between subsequent genomic regions. When two centers of genomic regions overlap, they have a probability $\gamma$ of forming a contact that contributes to the Hi-C map.

frequency statistics are calculated. During the inverse algorithm, this burn in time is only applied to the first forward simulation. For subsequent forward simulations, the final configuration of the previous forward simulation is used as a starting state.

## 3.2 Inverse algorithm

As noted in the main text, we learn the MaxEnt model directly from the normalized experimental Hi-C map. During a forward simulation of the polymer, the contact frequency $f_{ij}^{\text{model}}$ of each pair of monomers is counted. After one round of forward simulation, the simulated contact frequencies are normalized and compared to the experimental ones. The pairwise interaction energies are then updated according to

$$\Delta\epsilon_{ij} = \alpha(\tilde{f}_{ij}^{\text{model}} - \tilde{f}_{ij}^{\text{exp}}) \times \frac{1}{\sqrt{\tilde{f}_{ij}^{\text{exp}}}}. \tag{S2}$$

Here, $\alpha$ is the learning rate (which we typically set to 0.2), and the last factor is included to speed up conversion for pairs with a low contact frequency. Note, $\tilde{f}_{ij}^{\text{model}}$ and $\tilde{f}_{ij}^{\text{exp}}$ are the normalized model and experimental contact frequencies, respectively.

Importantly, to impose that the normalized contact frequencies match between model and experiment, we need to determine one remaining parameter: the absolute scale of the model contact frequencies. This is fixed by main text Eq. 6, which is derived as follows. Writing $f_{ij}^{\text{expt}} = c\tilde{f}_{ij}^{\text{expt}}$ and $\tilde{c} = \frac{c}{\gamma}$, the entropy functional becomes

$$\tilde{S} = -\sum_{\{\mathbf{r}\}} P(\{\mathbf{r}\}) \ln P(\{\mathbf{r}\}) - \sum_{ij} \lambda_{ij} \Bigg( \sum_{\{\mathbf{r}\}} P(\{\mathbf{r}\}) \delta_{\mathbf{r}_i, \mathbf{r}_j} - \tilde{c}\tilde{f}_{ij}^{\text{expt}} \Bigg) - \lambda_0 \Bigg( \sum_{\{\mathbf{r}\}} P(\{\mathbf{r}\}) - 1 \Bigg) \tag{S3}$$

we require that $\tilde{c}$ maximizes the model entropy, setting $\frac{\delta \tilde{S}}{\delta \tilde{c}} = 0$. This yields main text Eq. 6:

$$\sum_{ij} \lambda_{ij} \tilde{f}_{ij}^{\text{expt}} = 0. \tag{S4}$$

Ensuring that this condition is satisfied in each iteration step fixes the overall scale of contact frequencies. In the simulation, this is done by applying an overall shift in the interaction energies after the update step in Eq. (S2). This overall shift can be derived as follows: we start from Eq. 6, which imposes $\sum_{ij} \epsilon_{ij} \tilde{f}_{ij}^{\text{expt}} = 0$. In general, a set of $\epsilon_{ij}$ obtained after the update step in Eq. (S2) will not satisfy this constraint. We can, however, introduce a shift $\Delta\epsilon$ of all $\epsilon_{ij}$ such that this condition is satisfied:

$$\sum_{ij} (\epsilon'_{ij} - \Delta\epsilon) \tilde{f}_{ij}^{\text{expt}} = 0. \tag{S5}$$

Rewriting, and making use of $\sum_{ij} \tilde{f}_{ij}^{\text{expt}} = N_{\text{bin}}$ with $N_{\text{bin}}$ is the number of Hi-C bins, yields

$$\Delta\epsilon = -\frac{\sum_{kl} \epsilon'_{kl} f_{kl}^{\text{exp}}}{N_{\text{bin}}}. \tag{S6}$$

Performing this shift after each update step ensures that the condition in main text Eq. 6 is satisfied at each iteration of the inverse algorithm.

We iterate the inverse algorithm until the Pearson's correlation coefficient between the simulated normalized contact frequencies and the experimental data is above 0.98. This is the correlation coefficient of contact frequencies between repeat experiments reported in [12]. In practice, we can obtain even higher correlation coefficients of 0.998, as stated in the main text. With each subsequent forward simulation, the number of Monte Carlo steps is multiplied by $\sqrt{n}$, with $n$ the iteration step. The inverse algorithm is typically started with $\sim 360$ million steps, and run for $\sim 100$ iterations.

### 3.3 Ergodicity of forward algorithm

Next, we demonstrate that the algorithm is ergodic. A circular path of the polymer can be represented as a sequence of $N$ steps along the lattice, where each step is either up ($U$), down ($\bar{U}$), right ($R$), left ($\bar{R}$), in ($I$) or out ($\bar{I}$). We denote the total number of steps of type $x$ by $N(x)$. Circularity

of the path implies that $N(U) = N(\bar{U})$, $N(R) = N(\bar{R})$ and $N(I) = N(\bar{I})$. Furthermore, we will divide the steps in *types*, where $(U)$ and $(\bar{U})$ are type 1, $(R)$ and $(\bar{R})$ are type 2, and $(I)$ and $(\bar{I})$ are type 3. An individual path can then be described as a sequence of steps, for example

$$[\mathrm{U}, \bar{R}, \mathrm{I}, \mathrm{R}, \bar{U}, \cdots]. \tag{S7}$$

Here, each of the steps is colored by type. In the following we will also consider the sequence within each type. For our example, the sequences for the three types are:

- Type 1: $[\mathrm{U}, \bar{U}, \cdots]$
- Type 2: $[\bar{R}, \mathrm{R}, \cdots]$
- Type 3: $[\mathrm{I}, \cdots]$

We now consider the action of each of the polymer moves on a sequence of steps.

- The **kink move** interchanges two subsequent steps of a different type. Using only this move, any sequence of type 1, type 2 and type 3 steps can be created from a starting sequence that doesn't change the number of each type. Put differently, using the representation in (S7), any sequence of red, green and blue can be created that conserves the original counts of each color. Within each type, the sequence of the possible steps (e.g. $U$ and $\bar{U}$), however, cannot be changed with this move.

- The **crankshaft move** takes a motif of the form $[A, B, \bar{A}]$ and alters this to one of three possible motifs: (i) $[\bar{A}, B, A]$, or (ii) $[C, B, \bar{C}]$, or (iii) $[\bar{C}, B, C]$. The first alteration changes the sequence of steps within a type. Combining this alteration with the kink move, any sequence of steps within each type can be created, provided that there is at least one set of steps of a different type.

  Alteration (ii) and (iii) change the number of steps of each type: $N(A) + N(\bar{A})$ is reduced by 2, and $N(C) + N(\bar{C})$ is increased by 2. Combining this with the kink move, any set of counts of each of the types can be created, provided that polymer length and circularity are preserved, and that in the initial state not all steps are of the same type.

  Combining all three alterations with the kink move, from any starting sequence any final sequence can be created that conserves polymer length and circularity, as long as the starting and final sequence have moves of at least two different types.

- The **loop move** takes a motif of the form $[A, \bar{A}]$ and alters it to either (i) $[\bar{A}, A]$ or (ii)$[B, \bar{B}]$ or (iii) $[\bar{B}, B]$. Alteration (i) enables any change of the sequence within a type when the entire initial sequence is of the same type. Alterations (ii) and (iii) allow the conversion from a state of only one type to a state of two types.

  Combining the loop move with the kink and crankshaft moves, from any starting sequence any final sequence can be created that conserves polymer length and circularity. Thus, an ergodic sampling of the space of polymer configurations is ensured.

Note I: The presence of a confinement introduces a parity on the lattice sites: sites that can be occupied by an even monomer through these 3 moves cannot be occupied by an uneven monomer,

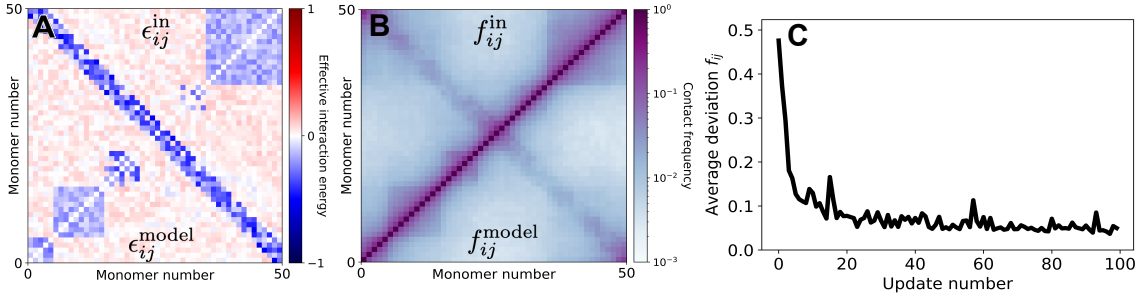
and vice versa. Either choice of parity can be seen as a separate coarse-grained model, as the unit cell locations shift depending on this choice.

Note II: A confinement could be chosen that 'traps' a portion of the polymer in place, making the phase space reachable using the three moves dependent on the initial state. For our confinement consisting of a cylinder with rounded edges such a trapping is not present, thus ergodicity is still preserved.

Note III: ergodicity is already ensured if only the loop and kink moves are used: the crankshaft move can be constructed as a combination of the two. However, the crankshaft move allows for a faster exploration of phase space and is thus also included.

# 4 Testing the inverse Monte Carlo algorithm

To test the performance of our inverse algorithm, we generated trial data sets by running a forward simulation for a chosen set of input effective interaction energies $\epsilon_{ij}^{\text{in}}$ (upper left Supplementary Fig. 5A). The resulting simulated contact map, $f_{ij}^{\text{in}}$, exhibits intricate features, including domain-like structures along the main diagonal and a fainter second diagonal (upper left Supplementary Fig. 5B). Subsequently, we treat this contact map as an experimental data set, which we use as an input to our iterative inverse scheme. We find that our inverse scheme rapidly and accurately retrieves the correct energies, $\epsilon_{ij}^{\text{model}} \approx \epsilon_{ij}^{\text{in}}$, and contact frequencies, $f_{ij}^{\text{model}} \approx f_{ij}^{\text{in}}$, demonstrating that this scheme adequately solves the inverse problem (Supplementary Fig. 5A-C).




**Supplementary Figure 5: Demonstration of numerical inverse algorithm for MaxEnt chromosome model. A** Upper left: input effective interaction energies $\epsilon_{ij}^{\text{in}}$. Lower right: effective interaction energies retrieved by the MaxEnt model. **B** Upper left: simulated contact frequencies $f_{ij}^{\text{in}}$ using $\epsilon_{ij}^{\text{in}}$. Lower right: contact frequencies of the MaxEnt model, using $f_{ij}^{\text{in}}$ as an input. **C** The average relative contact frequency deviation: $\langle f_{ij}^{\text{in}} - f_{ij}^{\text{model}} \rangle / \langle f_{ij}^{\text{model}} \rangle$ vs. iteration number of inverse algorithm.

# 5 Hi-C data filtering

Before the Hi-C data from Ref. [12] can be used to train our single-chromosome MaxEnt model, we need to account for the presence of a small fraction of replicating cells due to imperfect synchronization. Most notably, there is a local increase in Hi-C scores between the *ori* and *ter* genomic

regions, which is attributed to a small fraction of cells that have partially replicated their chromosome and segregated their newly formed *ori* regions towards the other pole, where the *ter* region of the initial chromosome is located [13, 14]. Although this increase is not readily visible on a linearly scaled Hi-C map (Supplementary Fig. 6A, upper left), it is clearly visible on a logarithmic scale (Supplementary Fig. 6A, lower right). Importantly, in experiments where replication is inhibited prior to synchronization, such an increase in contacts between the *ori* and *ter* genomic regions is not observed [15] (Supplementary Fig. 6B). In Ref. [15], two Hi-C experiments were performed on swarmer cells that could not undergo replication or cell division: on cells depleted of *dnaA*, and cells overexpressing *ctrA(D51E)*$\Delta\mathfrak{Z}\Omega$.

## 5.1 DNA replication inhibited Hi-C datasets

In the cells depleted of DnaA, the only copy of *dnaA*, whose product activates the initiation of replication, is driven by an IPTG-regulated promoter. Growth in medium lacking IPTG produced a population of cells that contained only a single, unreplicated copy of the chromosome. Cells were suspended in PYE medium without IPTG to deplete DnaA for 90 min before synchronization [15]. The data set analyzed here is for cells that were formadehyde fixed immediately after synchronization (90 min after IPTG withdrawal).

In the cells overexpressing the hyperactive and non-degradable CtrA variant *ctrA(D51E)*$\Delta\mathfrak{Z}\Omega$, chromosome replication is inhibited by constitutive binding of CtrA close to the origin of replication. *ctrA(D51E)*$\Delta\mathfrak{Z}\Omega$ is expressed from an xylose-inducible promoter on the high copy number pJS14 plasmid in the presence of the chromosomal copy of wild-type *ctrA*. Cells were suspended in PYE medium plus xylose for 60 min before synchronization [15]. The data set analyzed here is for cells that were formadehyde fixed at 0 hr post synchronization (60 min after xylose addition).

For both the DnaA-depleted cells and the cells overexpressing *ctrA(D51E)*$\Delta\mathfrak{Z}\Omega$, average Hi-C scores are found to monotonicaly decrease with inter-arm genomic distance until a noise floor is reached, and to exhibit three distinct scaling regimes (Supplementary Fig. 6C). By contrast, for the wild-type synchronized swarmer cells from Ref. [12], an increase in average Hi-C scores for the largest inter-arm genomic distances is observed (Supplementary Fig. 6F). If we train a MaxEnt model directly on this data, this single-chromosome model will interpet these *ori*-*ter* contacts as inter-chromosomal contacts, resulting in a weaker localization of the *ter* region (Supplementary Note 7). Here, we propose a filtering procedure to process the wild-type data such that we can infer a reliable single-chromosome MaxEnt model, even in the presence of a small fraction of non-synchronized cells.

## 5.2 Filter procedure

The goal of our data processing procedure is to filter out the contribution of the newly replicated *ori* from the wild-type data set, using the two data sets for replication-inhibited cells as a benchmark. The advantages of filtering the wild-type dataset, rather than applying the analysis to the replication-inhibited cells, are two-fold: First, the experimental procedure to inhibit replication might affect features of chromosome organization. Second, a filter method allows for the analysis of data sets for mutants and cells in atypical growth conditions but without replication inhibition, such as the $\Delta smc$ mutant and the rifampicin-treated cells in Ref. [12], using a single chromosome model. For completeness, the results of applying the MaxEnt method directly to the unfiltered wild-type data, as well as to the replication-inhibited cell data, are presented in Supplementary

Notes 7 and 8. Importantly, we find that all the central conclusions drawn in the Main Text based on our MaxEnt model trained on the filtered WT data, can also be drawn for a MaxEnt model on the unprocessed Hi-C data from the replication-inhibited cells.

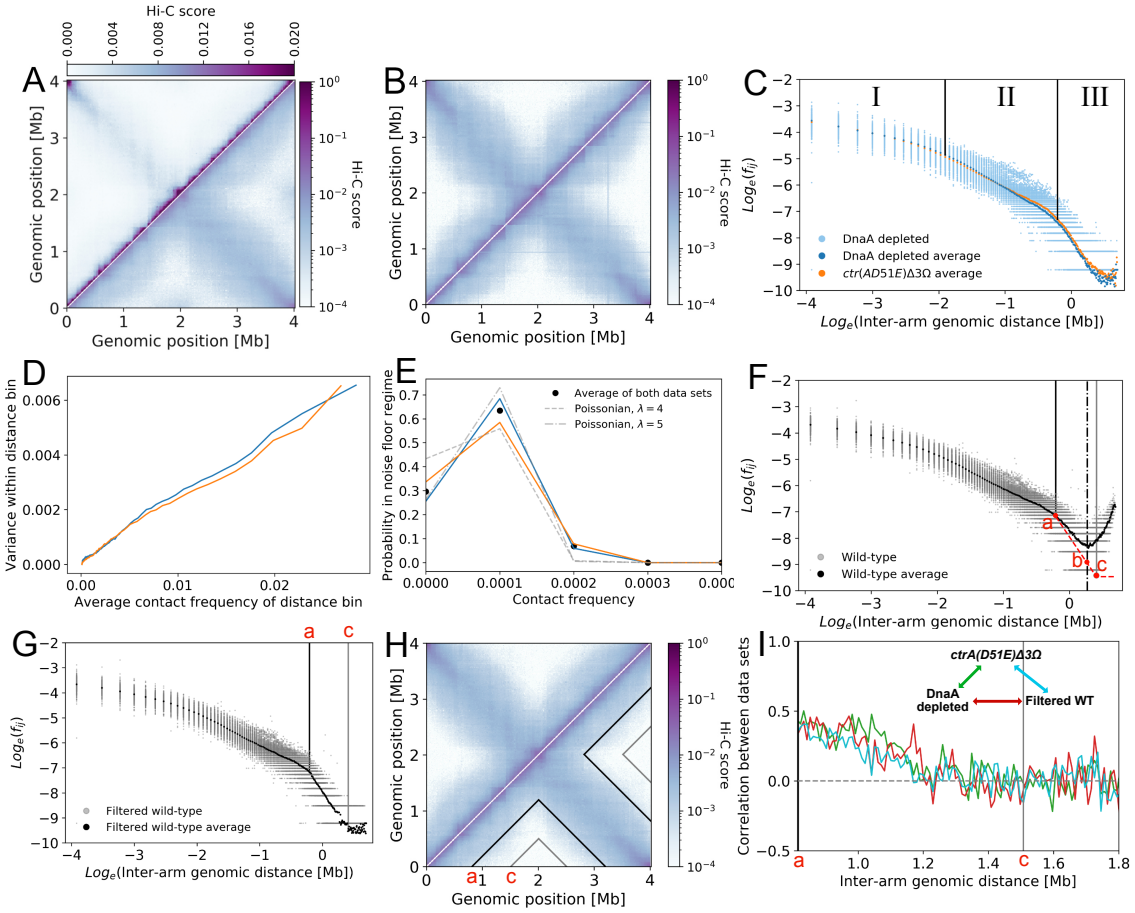
The procedure to filter out the contribution of the newly replicated *ori* aims to reproduce three features observed for replication-inhibited cells: (1) a power law scaling of the average contact frequencies in regime III (Supplementary Fig. 6C), (2) a proportionality between the mean and variance of Hi-C scores across inter-arm genomic distance bins (Supplementary Fig. 6D), and (3) a transition to a noise floor regime for the largest inter-arm genomic distances (Supplementary Fig. 6C,E). The filtering procedure is as follows. First, the estimated average Hi-C scores for the single, unreplicated chromosome $f_{\rm av}^{\rm single}(d)$ are constructed for each inter-arm genomic distance bin, $d$, in regime III (Supplementary Fig. 6F, red dashed line) for the wild-type data set (the construction procedure is detailed in the next paragraph). A rescaling factor $\mu(d) = \frac{f_{\rm av}^{\rm single}(d)}{\langle f_{ij}^{\rm WT}\rangle_d}$ is then obtained between $f_{\rm av}^{\rm single}(d)$ and the unfiltered wild-type data averages $\langle f_{ij}^{\rm WT}\rangle_d$ for a given distance bin $d$. This factor $\mu(d)$ is subsequently used to rescale individual Hi-C scores of the wild type data set at each inter-arm genomic distance bin $d$ within regime III. By construction, this rescaling procedure ensures that the filtered Hi-C scores will not only have the correct estimated average value, but also the correct estimated variance, preserving the proportionality between average Hi-C scores and the asscociated variance observed for replication-inhibited cells. Finally, when the average rescaled contact frequencies fall below the noise floor observed for replication-inhibited cells, Hi-C scores are determined from the observed noise floor distribution (Supplementary Fig. 6E).

To construct $f_{\rm av}^{\rm single}(d)$, we need two points $a$ and $b$ on the log-log plot to define the power law relation associated to regime III. The vertical position of point $a$ is set at $\langle f_{ij}^{\rm WT}\rangle_d$ at the onset of regime III (Supplementary Fig. 6F), beyond which contributions from the newly replicated *ori* are assumed to become significant. To position point $b$, we assume the contributions to $\langle f_{ij}^{\rm WT}\rangle_d$ from inter-chromosomal contacts and the newly replicated *ori* regions to be equal at the minimum of $\langle f_{ij}^{\rm WT}\rangle_d$ (Supplementary Fig. 6F, dash-dotted line), since this marks the distance beyond which contributions from the newly replicated *ori* become dominant. We thus set the vertical position of point $b$ equal to $\langle f_{ij}^{\rm WT}\rangle_d/2$. Hence, $f_{\rm av}^{\rm single}(d)$ follows the power law relation consistent with the line from $a$ to $b$, extending till point $c$, where the noise floor level is reached; this noise floor is found to be at an average Hi-C score of 0.000078 for the replication-inhibited cells(Supplementary Fig. 6F, point $c$). Using this procedure, we now also obtain $\mu(d)$ between point $a$ and $c$.

The filter procedure rescales the Hi-C data by $\mu(d)$ between points $a$ and $c$ in regime III. Beyond point $c$, the noise level is reached, and Hi-C scores are randomly drawn from the observed noise floor distribution. These noise-floor distributions are constructed by counting all Hi-C scores of the two replication-inhibited cells with an inter-arm genomic distance above 1.78 Mb, and are consistent with an underlying Poissonian process (Supplementary Fig. 6E). Importantly, this construction leaves all Hi-C scores in scaling regimes I and II (Supplementary Fig. 6C) unchanged, and filters wild-type Hi-C scores in regime III. The resulting filtered Hi-C scores are shown in Supplementary Fig. 6G and Supplementary Fig. 6H. Finally, we applied the same data processing procedure to two other replicas of the WT experiments, as shown in Supplementary Fig. 7.

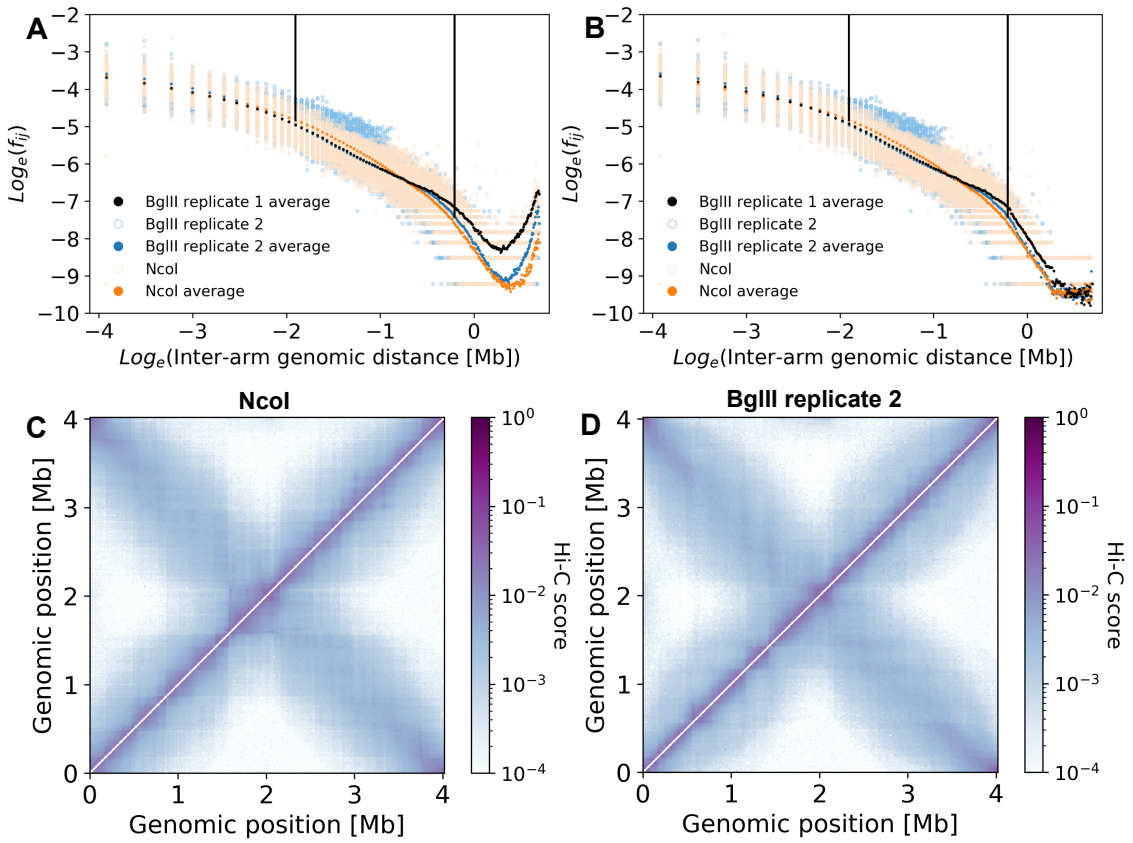
Our data processing procedure ensures that the averages and variances of the contact frequencies per inter-arm genomic distance bin behave as observed for replication-inhibited cells. However, it is possible that additional structure is present in the replication-inhibited data, which is lost in the filtered wild-type data during this data processing procedure. To test this, we compute the correlation between Hi-C contact scores within each inter-arm genomic distance bin, between (1)

the filtered wild-type data set, (2) the DnaA-depleted cells and (3) *ctrA(D51E)*Δ3Ω overexpressing cells. These correlations are a measure for the similarities between the data sets for each inter-arm genomic distance bin. A correlation of 1 corresponds to two data sets being identical up to a proportionality constant, whereas a correlation of 0 corresponds to the variations within a genomic distance bin being linearly independent between data sets. At the onset of regime III (point $a$), we find significant correlations between the three datasets (Supplementary Fig. 6I). Importantly, these correlations do not significantly differ between each of the three pairs of data sets, indicating the presence of similar structure in the filtered wild-type data set and the replication-inhibited data sets. For larger inter-arm genomic distances, the correlations between data sets go to zero, as would be expected at the onset of the noise floor regime (point $c$).



**Supplementary Figure 6** ***(previous page)*****: Hi-C data processing procedure A** Wild-type contact frequencies from [12] on a linear scale (upper left triangle) and a logarithmic scale (bottom right triangle). **B** Contact frequencies for replication-inhibited swarmer cells directly after synchronization. Upper left triangle: DnaA-depleted cells; lower right triangle: cells overexpressing *ctrA(D51E)*$\Delta\mathcal{3}\Omega$. Both datasets are taken from [15]. **C** Hi-C score versus inter-arm genomic distance for DnaA depleted cells (light blue dots). Dark blue dots: averages per inter-arm genomic distance bin. Orange dots: averages per inter-arm genomic distance bin for cells overexpressing *ctrA(D51E)*$\Delta\mathcal{3}\Omega$. Three distinct scaling regimes are identified, indicated by regions I-III. **D** Variance of Hi-C scores within an inter-arm genomic distance bin versus the average Hi-C score of this genomic distance bin for DnaA-depleted cells (blue line) and the cells overexpressing *ctrA(D51E)*$\Delta\mathcal{3}\Omega$ (orange line). **E** Probabilities of Hi-C score occurances in the noise floor regime, taken over pairs of genomic regions with an inter-arm genomic distance of at least 1.78 Mb. Blue line: DnaA-depleted cells; orange line: cells overexpressing *ctrA(D51E)*$\Delta\mathcal{3}\Omega$; black dots: averages per Hi-C score bin; dashed line: distribution for a poissonian process with a mean equal to the average of the two data sets and $\lambda = 4$; dash-dotted line: the same for $\lambda = 5$. **F** Hi-C score versus inter-arm genomic distance for wild-type cells (grey dots). Black dots: average per inter-arm genomic distance bin. Dash-dotted line: horizintally aligned with the minimum point of the average Hi-C scores. Red dashed line: $f_{\mathrm{av}}^{\mathrm{single}}(d)$ (from point $a$ to $c$) and the noise floor beyond point $c$ (see Supplementary text for more details). **G** Wild-type Hi-C scores after the filtering procedure is applied (grey dots) together with the averages per inter-arm genomic distance bin (black dots). **H** Wild-type Hi-C score map after the filtering procedure has been applied. Hi-C scores have been rescaled in the regime between the black and the grey lines. The noise floor region is enclosed withn the grey lines, where Hi-C scores have been randomly drawn from the distribution in **E**. **I** Correlations of contacts within an inter-arm genomic distance bin, between (1) the filtered wild-type data set, (2) the DnaA-depleted cells and (3) *ctrA(D51E)*$\Delta\mathcal{3}\Omega$ overexpressing cells.

# 6 Comparison of filter procedure for wild-type replicates

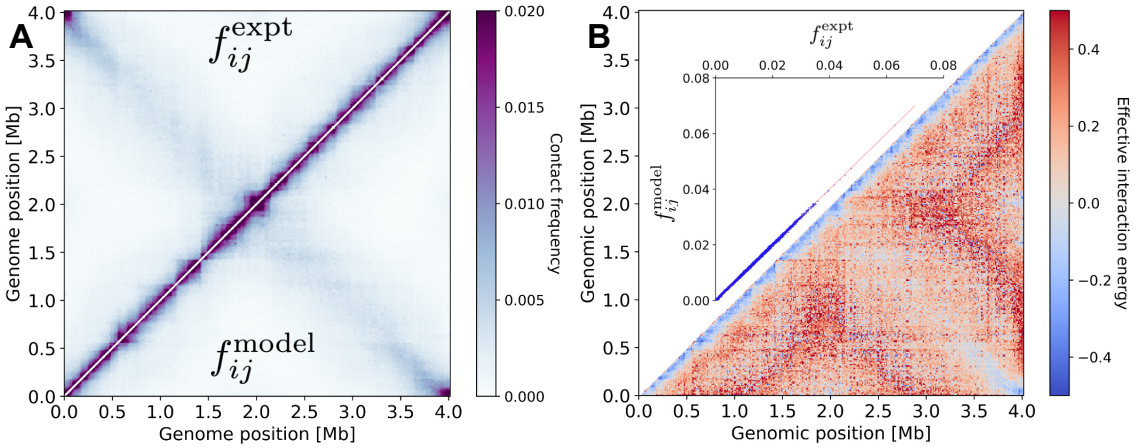



**Supplementary Figure 7: Comparison of data filtering procedure for wild-type replicates A** Unfiltered Hi-C scores as a function of inter-arm genomic distance for the three wild-type replicates published in [12]. The onsets of scaling regimes II and III as introduced in Supplementary Notes 5 are indicated by black vertical lines. **B** Hi-C scores versus inter-arm genomic distance for the three replicates after the filter procedure has been applied. **C** Upper left: Hi-C scores of the NcoI dataset before the filter procedure is applied. Lower right: Hi-C scores of the same data set after the filter procedure has been applied. **D** Upper left: Hi-C scores of the BglII replicate 2 dataset before the filter procedure is applied. Lower right: Hi-C scores of the same data set after the filter procedure has been applied.

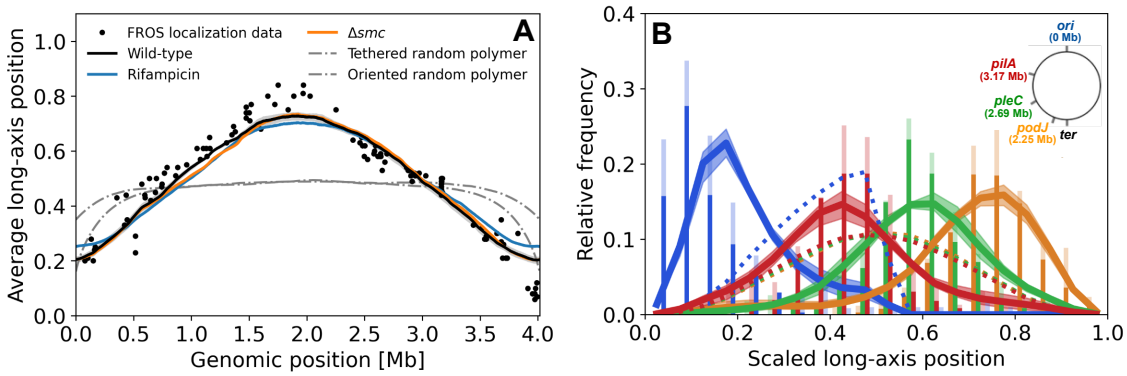
# 7 Results for MaxEnt model trained on unfiltered Hi-C data

To further investigate the effect of this data processing on the model results, we reran our analysis directly on the unprocessed Hi-C data. We find that the localizations of genomic regions (Supplementary Fig. 9), the orientational and radial correlations in positions of regions (Supplementary Fig. 10) and the local structure (Supplementary Fig. 11) are largely unaffected. The most significant difference is found in the localization of the *ter* region, which is now found to move throughout the

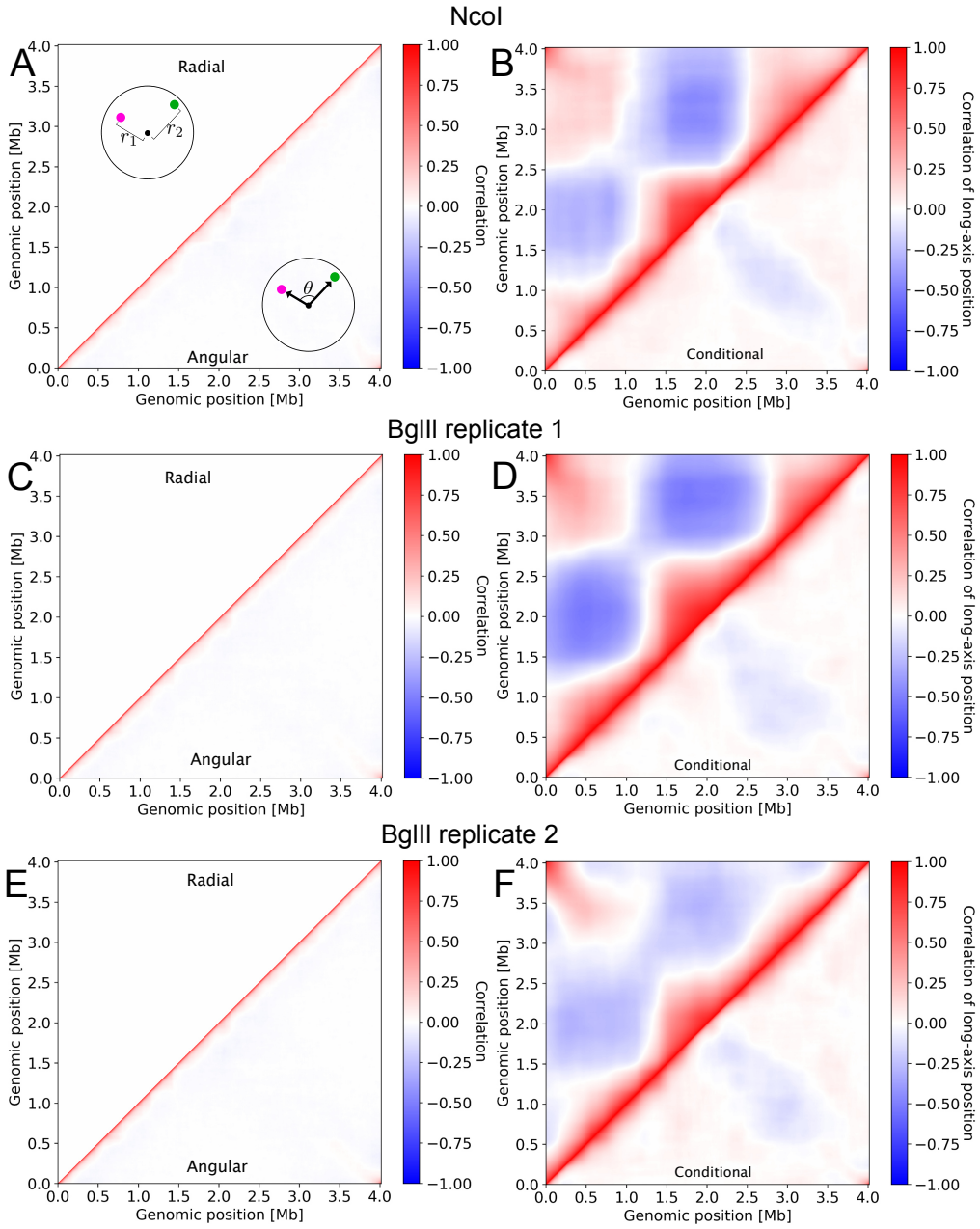
*ori* half of the cell in a minority of states (Supplementary Fig. 9). This movement of the terminus has an effect on the long-axis anti-correlations between the *ori* and *ter* regions (Supplementary Fig. 10), resulting in a modified long-axis correlation pattern compared to the filtered Hi-C data (Main Text Fig. 3B). However, if conditional long-axis correlations are computed, conditioned on the *ori* region (here defined as 3.75 Mb - 0.25 Mb) being in one half of the cell, and the *ter* region (here defined as 1.75 Mb - 2.25 Mb) being in the other half, the pattern of anti-correlations between the two juxtaposed chromosomal arms is restored (Supplementary Fig. 10).



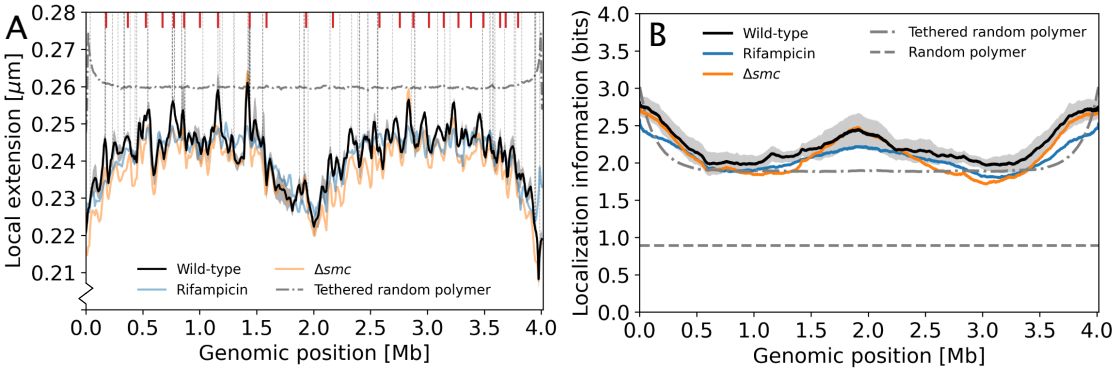

**Supplementary Figure 8: Results for Main Text Fig. 1, re-analyzed for the unfiltered Hi-C data of replicate 1 from [12].** **A** Comparison between experimental contact frequencies $f_{ij}^{\mathrm{expt}}$ (upper left corner, adapted from Ref. [12]) and contact frequencies obtained from our inferred MaxEnt model $f_{ij}^{\mathrm{model}}$ (lower right corner). **B** Associated inferred effective interaction energies $\epsilon_{ij}$ (lower right corner, white regions indicate $\epsilon_{ij} \to \infty$) together with a scatter plot of $f_{ij}^{\mathrm{expt}}$ vs. $f_{ij}^{\mathrm{model}}$ (inset).




**Supplementary Figure 9: Results for Main Text Fig. 2, re-analyzed for the three replicates from [12].** **A** Black solid line: average of the three data sets. Grey area: standard deviation of the three replicates, centered at the average. **B** Solid lines: averages of the three replicates. Shaded areas: standard deviations of the three replicates, centered at the average. Bars: experimental data from microscopy experiments (adapted from [3]). To indicate experimental variability, the solid/transparent bars indicate the minimum/maximum measured by two different methods: FROS or FISH.



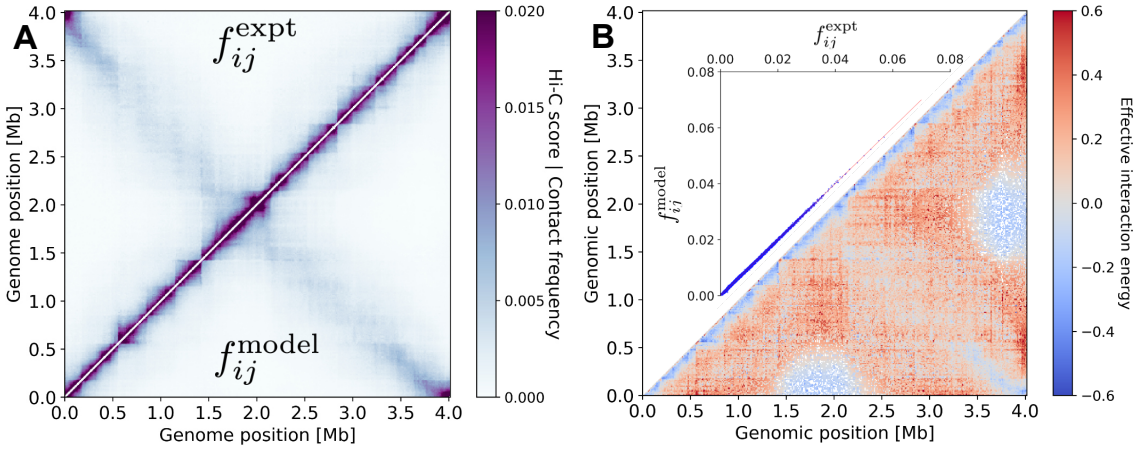

**Supplementary Figure 10: Two-point correlations for the three replicates from [12].** Plots **A** and **B** are for the NcoI replicate. **A** Upper left corner: two-point correlations in the radial positions between genomic regions. Lower right corner: two-point correlations in angular orientations around the long axis. **B** Upper left corner: two-point correlations between long-axis positions of genomic regions. Lower right corner: conditional long-axis correlations, conditioned on the *ori* region (here defined as 3.75 Mb - 0.25 Mb) being in one half of the cell, and the *ter* region (here defined as 1.75 Mb - 2.25 Mb) being in the other half. **C** and **D**: same as **A** and **B**, for BglII replicate 1. **E** and **F**: same as **A** and **B**, for BglII replicate 2.



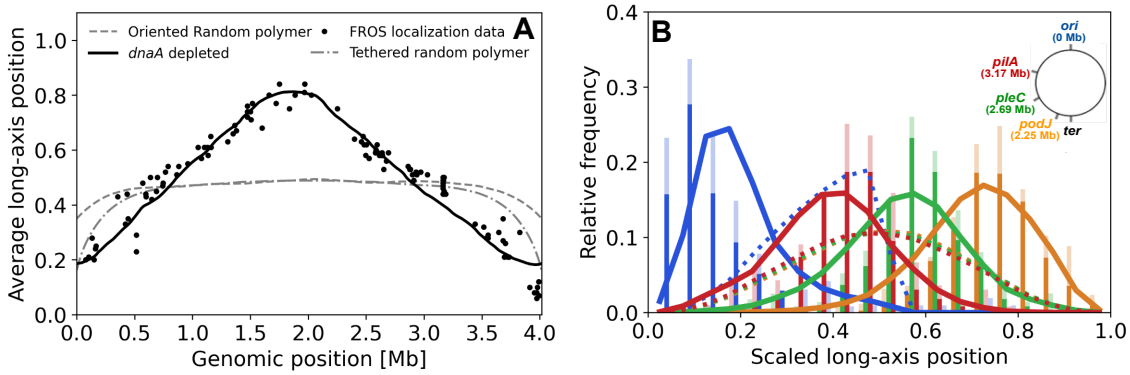

**Supplementary Figure 11: Results for Main Text Fig. 5, re-analyzed for the three replicates from [12].** **A** The local chromosome extension $\delta_i$ as a function of genomic position. Black solid line: average of the three replicates. Grey areas: standard deviation of the three replicates, centred at the average. **B** Localization information per genomic region in bits. Black solid line: average of the three replicates. Grey areas: standard deviation of the three replicates, centred at the average.

# 8 Results for MaxEnt model trained on replication-inhibited cells

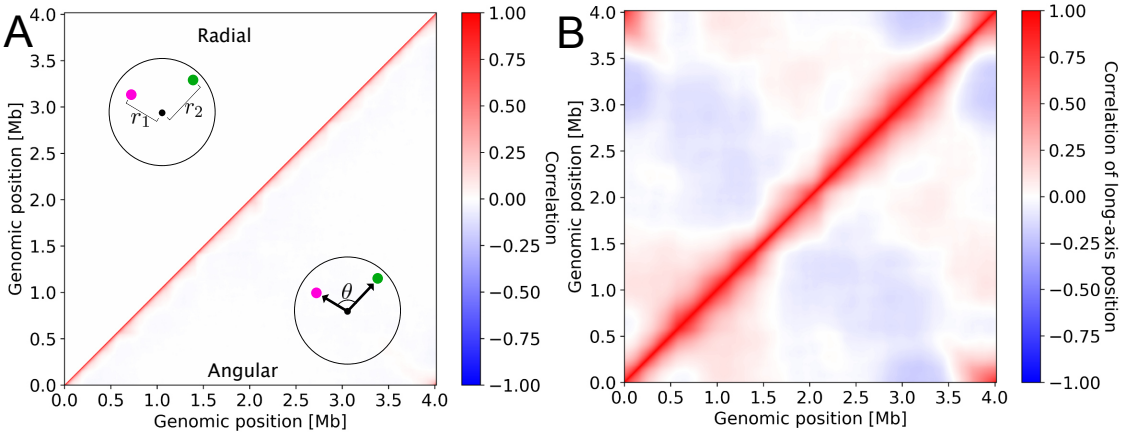
## 8.1 DnaA-depleted cells




**Supplementary Figure 12: Results for Main Text Fig. 1, re-analyzed for the DnaA-depleted cell data set from [15].** **A** Comparison between experimental contact frequencies $f_{ij}^{\mathrm{expt}}$ (upper left corner, adapted from Ref. [15] and contact frequencies obtained from our inferred MaxEnt model $f_{ij}^{\mathrm{model}}$ (lower right corner). **B** Associated inferred effective interaction energies $\epsilon_{ij}$ (lower right corner, white regions indicate $\epsilon_{ij} \to \infty$) together with a scatter plot of $f_{ij}^{\mathrm{expt}}$ vs. $f_{ij}^{\mathrm{model}}$ (inset).

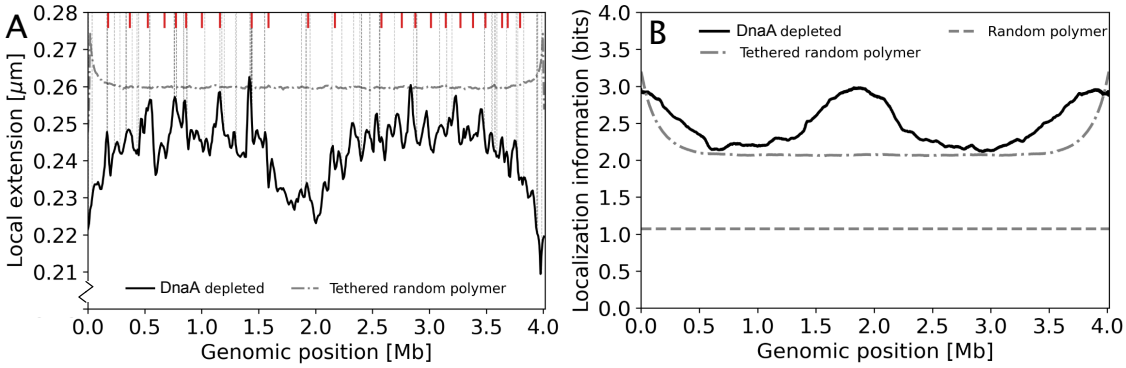



**Supplementary Figure 13: Results for Main Text Fig. 2, re-analyzed for the DnaA-depleted cell data set from [15].** **A** Black solid line: average long-axis positions of genomic regions for DnaA-depleted cells predicted by the MaxEnt model. **B** Solid lines: distribution of long-axis positions of chromosomal loci (blue: *ori*, red: *pilA*, green: *pleC*, orange: *podJ*) for DnaA-depleted cells predicted by the MaxEnt model, , together with previous experimental data from microscopy experiments (bars, adapted from [3]). To indicate experimental variability, the solid/transparent bars indicate the minimum/maximum measured by two different methods: FROS or FISH.

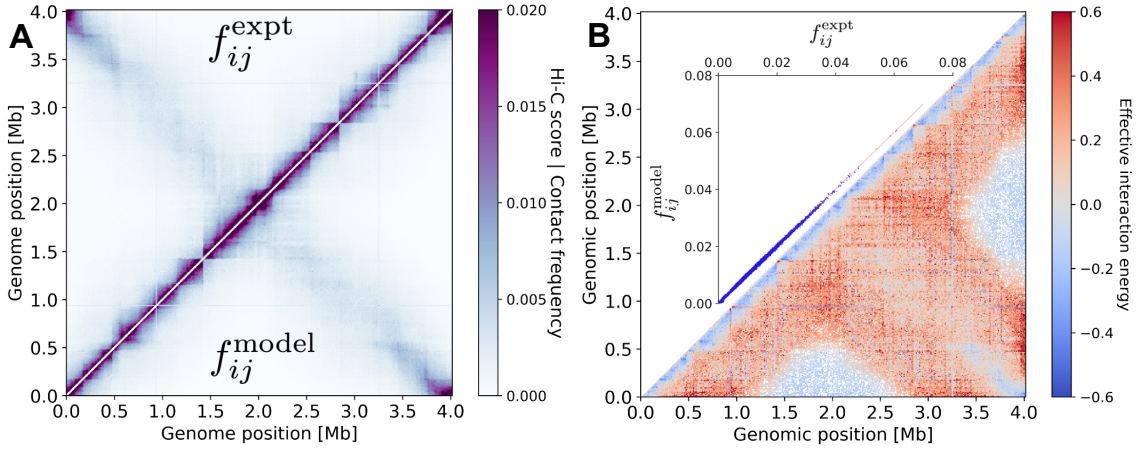



**Supplementary Figure 14: Results for Main Text Fig. 3, re-analyzed for the DnaA-depleted cell data set from [15].** **A** Upper left corner: two-point correlations in the radial positions between genomic regions. Lower right corner: two-point correlations in angular orientations around the long axis. **B** Two-point correlations between long-axis positions of genomic regions.

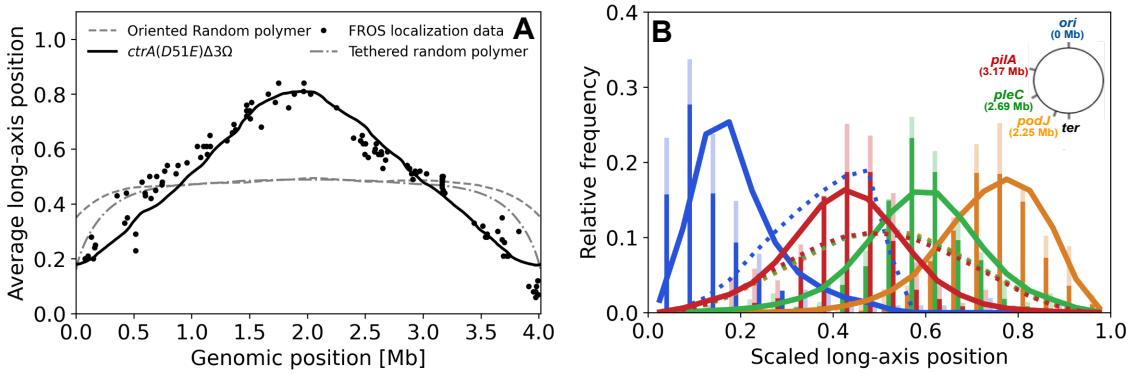
## 8.2 Cells overexpressing *ctrA(D51E)Δ3Ω*



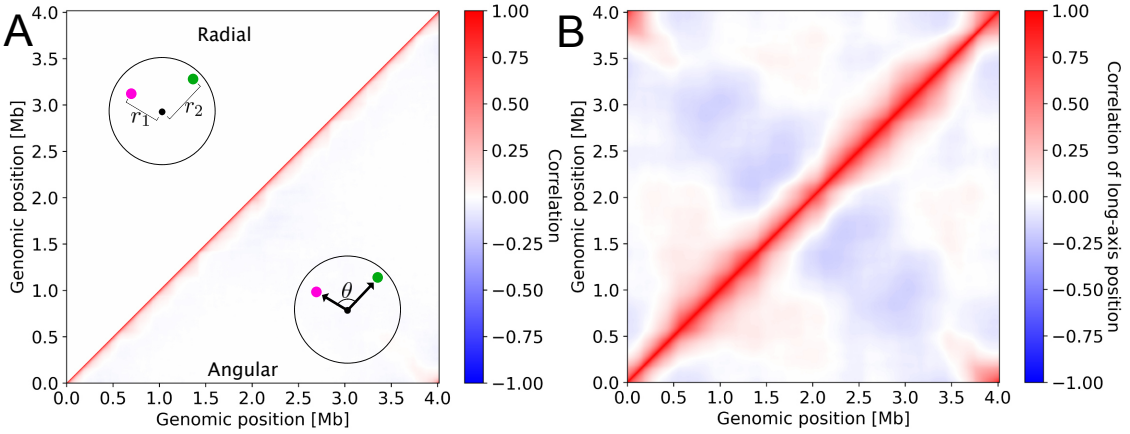

**Supplementary Figure 15: Results for Main Text Fig. 5, re-analyzed for the DnaA-depleted cell data set from [15].** **A** Black solid line: the local chromosome extension $\delta_i$ as a function of genomic position for DnaA-depleted cells as predicted by the MaxEnt model. **B** Black solid line: localization information per genomic region for DnaA-depleted cells as predicted by the MaxEnt model.



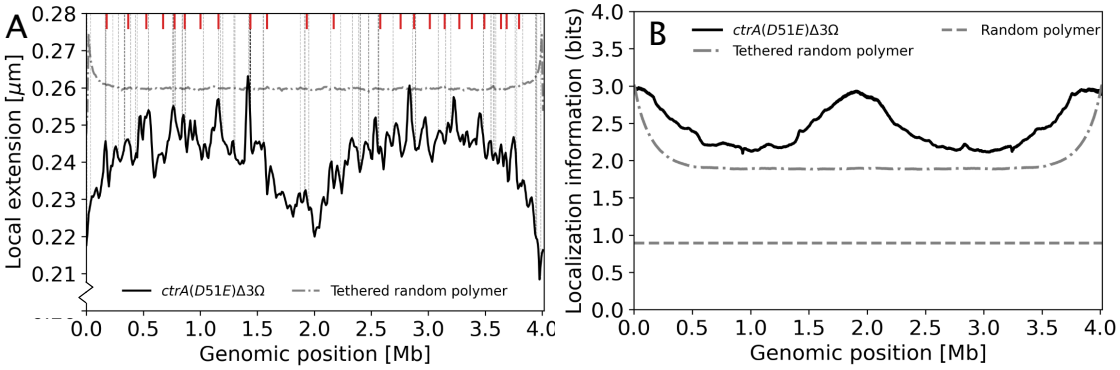

**Supplementary Figure 16: Results for Main Text Fig. 1, re-analyzed for the *ctrA(D51E)Δ3Ω* overexpressing cell data set from [15].** **A** Comparison between experimental contact frequencies $f_{ij}^{\text{expt}}$ (upper left corner, adapted from Ref. [15] and contact frequencies obtained from our inferred MaxEnt model $f_{ij}^{\text{model}}$ (lower right corner). **B** Associated inferred effective interaction energies $\epsilon_{ij}$ (lower right corner, white regions indicate $\epsilon_{ij} \to \infty$) together with a scatter plot of $f_{ij}^{\text{expt}}$ vs. $f_{ij}^{\text{model}}$ (inset).




**Supplementary Figure 17: Results for Main Text Fig. 2, re-analyzed for the *ctrA(D51E)*Δ3Ω overexpressing cell data set from [15].** **A** Black solid line: average long-axis positions of genomic regions for *ctrA(D51E)*Δ3Ω overexpressing cells predicted by the MaxEnt model. **B** Solid lines: distribution of long-axis positions of chromosomal loci (blue: *ori*, red: *pilA*, green: *pleC*, orange: *podJ*) for *ctrA(D51E)*Δ3Ω overexpressing cells predicted by the MaxEnt model, together with previous experimental data from microscopy experiments (bars, adapted from [3]). To indicate experimental variability, the solid/transparent bars indicate the minimum/maximum measured by two different methods: FROS or FISH.




**Supplementary Figure 18: Results for Main Text Fig. 3, re-analyzed for the *ctrA(D51E)*Δ3Ω overexpressing cell data set from [15].** **A** Upper left corner: two-point correlations in the radial positions between genomic regions. Lower right corner: two-point correlations in angular orientations around the long axis. **B** Two-point correlations between long-axis positions of genomic regions.




**Supplementary Figure 19: Results for Main Text Fig. 5, re-analyzed for the *ctrA(D51E)*$\Delta 3\Omega$ overexpressing cell data set from [15].** **A** Black solid line: the local chromosome extension $\delta_i$ as a function of genomic position for DnaA-depleted cells as predicted by the MaxEnt model. **B** Black solid line: localization information per genomic region for DnaA-depleted cells as predicted by the MaxEnt model.

# 9 Analysis of genomic Super Domains

## 9.1 Super Domain definition and long-axis exclusion analysis

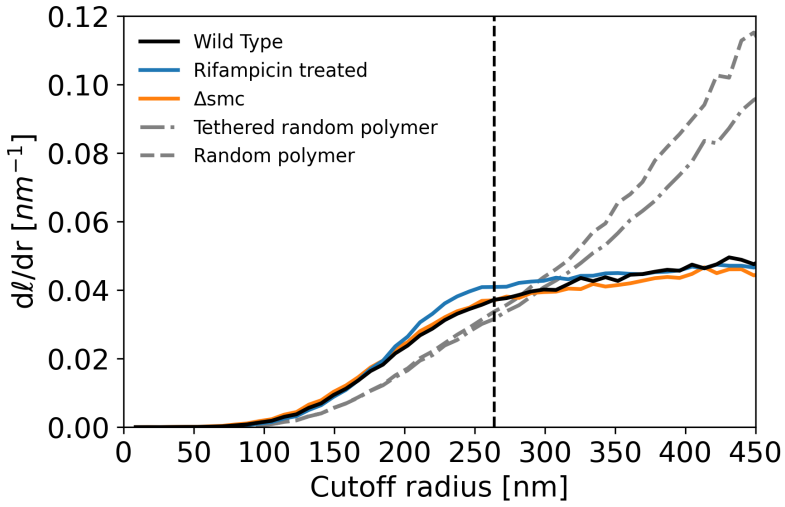
To define genomic Super Domains (SuDs), we first choose a cluster radius $r$. For each genomic region $i$, we consider a specific configuration of the chromosome and then calculate the length $\ell$ of the set of subsequent genomic regions (in both directions along the chromosome) that lie within the radius $r$ from the position of genomic region $i$ (illustrated by the black line in Main Text Fig. 4A). We observe that for each configuration of the chromosome, the genomic regions separate into a small number of domains, indicated by the blue and red areas in Main Text Fig. 4A. We identify a domain with each local maximum in $\ell$ (indicated by $L1 - L3$ and $R1 - R3$ in Main Text Fig. 4B); the peak location represents the genomic region at the center of a SuD and the peak value indicates the number of genomic regions within the domain.

To determine a natural choice for $r$, we perform a parameter sweep over $r$ and consider the change in the average value of $\ell$ with $r$: $\mathrm{d}\bar{\ell}/\mathrm{d}r$. We find that for the MaxEnt models on wild-type, rifampicin-treated and $\Delta smc$ cells, $\mathrm{d}\bar{\ell}/\mathrm{d}r$ initially increases with $r$, and then becomes approximately constant (Supplementary Fig. 20). For models unconstrained by Hi-C data (the 'random polymer', and the 'tethered random polymer'), such a transition to a plateau regime is not present. We interpret the transition to this plateau regime in the MaxEnt models as the genomic length scale at which the linear organization of the chromosome along the cell length starts dominating local fluctuations of loci(Main Text Fig. 2A&B). We take the crossover point between these two regimes to be $r = 264$ nm, indicated by the grey dashed line in Supplementary Fig. 20.

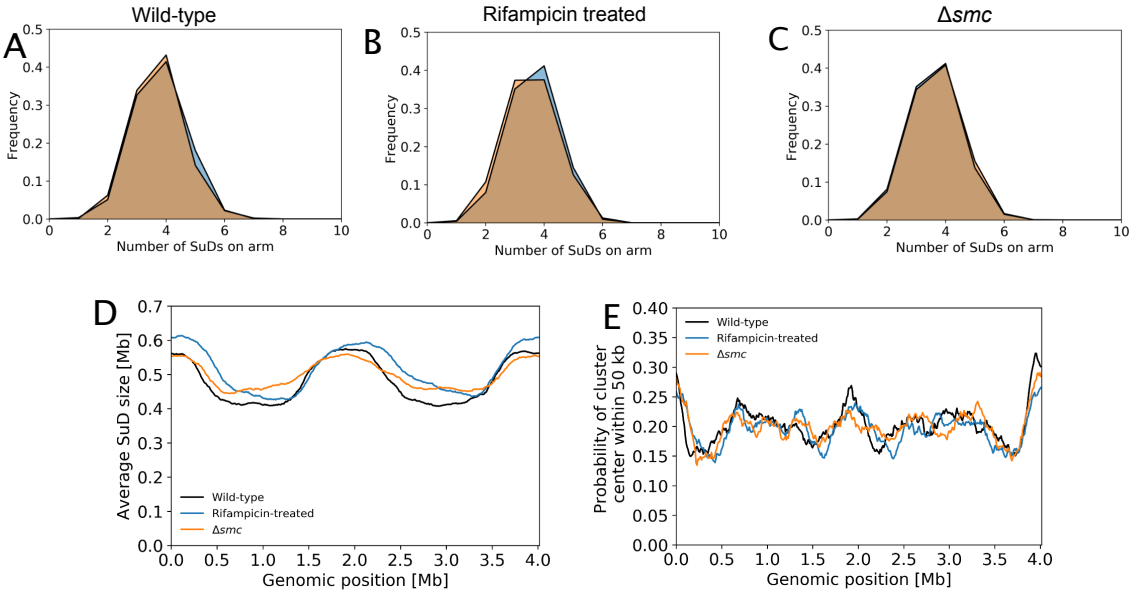
To quantify the degree of long-axis exclusion between SuDs, the distribution of long-axis positions of the genomic regions contained in each domain is computed (Main Text Fig. 4B). A long-axis position is assigned to a Super Domain based on the highest-occupied long axis coordinate of this cluster. The degree of overlap of long-axis positions is then computed for randomly paired left and right arm configurations and for correctly matched pairs.

## 9.2 Super Domain properties

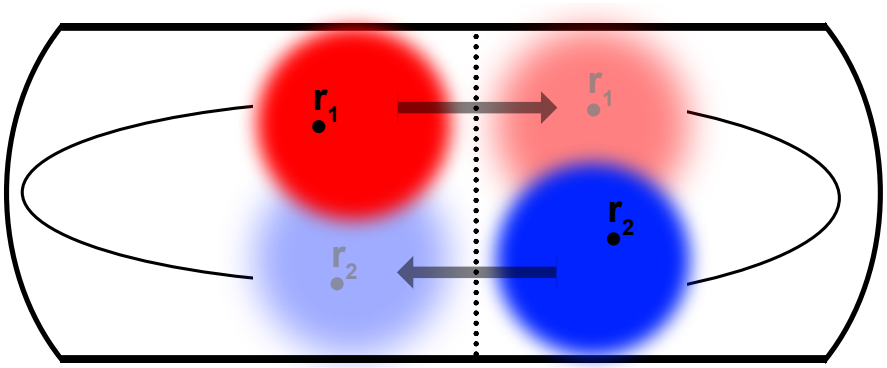
To quantify the distribution of SuD sizes and locations, we determined the average number of SuDs on each chromosomal arm, the average SuD size across genomic regions and the distribution of SuD center locations across the genome. The results are shown in Supplementary Figure 21. An illustration of the expected link between SuDs and the inferred anticorrelations between chromosomal arms is shown in Supplementary Figure 22




**Supplementary Figure 20: Super Domain cluster analysis.** Derivative of the average cluster size as a function of the cutoff radius $r$, for wild-type cells (black), rifampicin-treated cells (blue), a Δ*smc* mutant (orange), a tethered random polymer (dash-dotted line) and a random polymer (dashed line). The vertical dashed line indicates the chosen cutoff value.




**Supplementary Figure 21: Super Domain properties.** Distribution of the number of Super Domains across configurations for the left arm (blue) and the right arm (orange) for wild-type cells (**A**), rifampicin-treated cells (**B**) and a Δ*smc* mutant (**C**). **D** Average size of the SuD a genomic region is part of, given that it is part of a SuD, as a function of genomic position. **E** Probability of a cluster center being within 50 kb of a genomic region, as a function of genomic position.

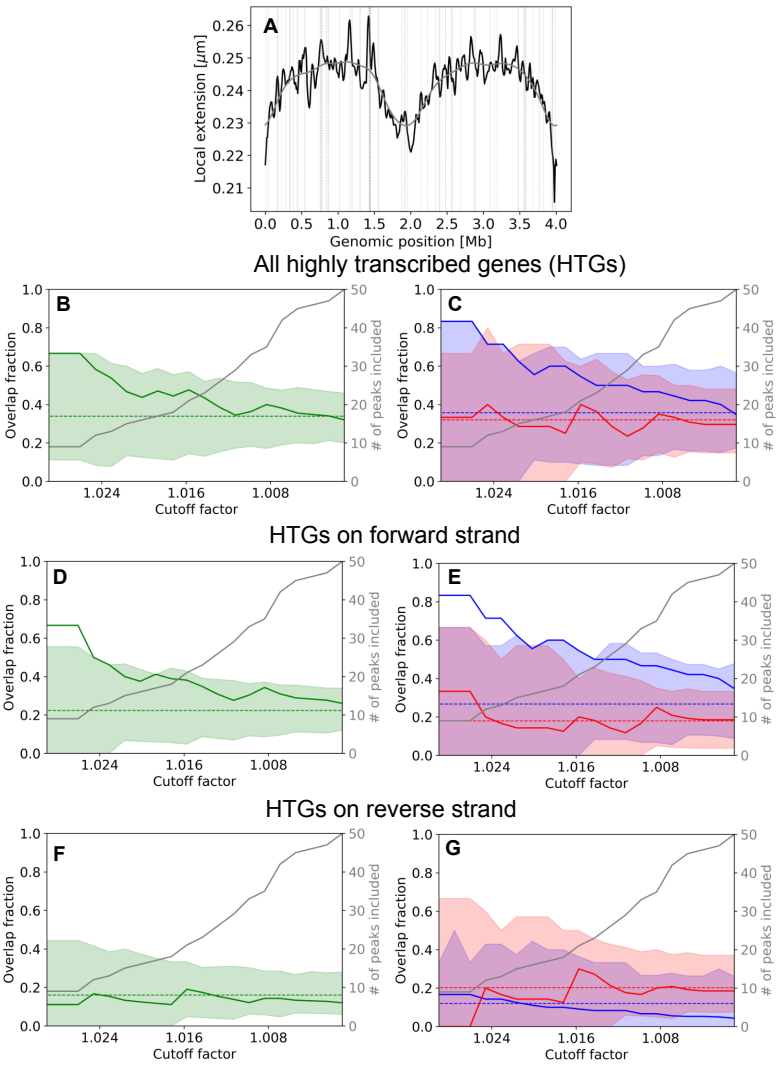



**Supplementary Figure 22: Illustration of SuDs inducing anticorrelations between chromosomal arms.** In this illustration, genomic regions $r_1$ and $r_2$ lie on different chromosomal arms but have the same average long-axis position (dashed line). The SuD that region $r_1$ is part of, has a tendency to avoid the SuD that region $r_2$ is part of (given that both regions are part of a SuD). This is expected to induce anticorrelations in the long-axis positions of regions $r_1$ and $r_2$.

## 10 Overlap analysis between local chromosome extension peaks and highly transcribed genes

To investigate the connection between peaks in the local extension profile and the locations of highly transcribed genes, we first construct a (nonlinear) trend line through the chromosome extension profile. This line is constructed by repeatedly applying a Gaussian smoothing filter over the data, incorporating periodic boundary conditions. The Gaussian smoothing is implemented by repeatedly applying a moving average over groups of 3 subsequent genomic regions. We find that 250 repeats to result in a satisfactory balance between smoothing out local peaks and keeping the larger-scale trend (grey line in Supplementary Figure 23A). Next, we select the subset of local extension peaks that lie a factor $\alpha$ above the trend line. We perform a sweep over $\alpha$ and calculate for each choice of $\alpha$ the fraction of incorporated peaks that coincide with the locations of highly transcribed genes. Additionally, for each $\alpha$ we simulate a number of randomly positioned peaks equal to the number of incorporated peaks. From this simulation, we calculate the expected fraction of overlap and the 95% confidence intervals.

We find that the fraction of overlap is significantly higher than expected for randomly positioned local extention peaks, if up to the 9 highest peaks are considered (Supplementary Figure 23B). If more peaks are incorporated, the fraction of overlap gradually decays to the level expected for random positions. Repeating this analysis for the right (0-2 Mb) and left (2-4 Mb) chromosomal arms seperately, we find that the fraction of overlap is only significantly higher than a random guess for the highest peaks of the right arm (Supplementary Figure 23C). For the left arm, by contrast, the fraction of overlap is close to the value expected by random guess for all values of $\alpha$.

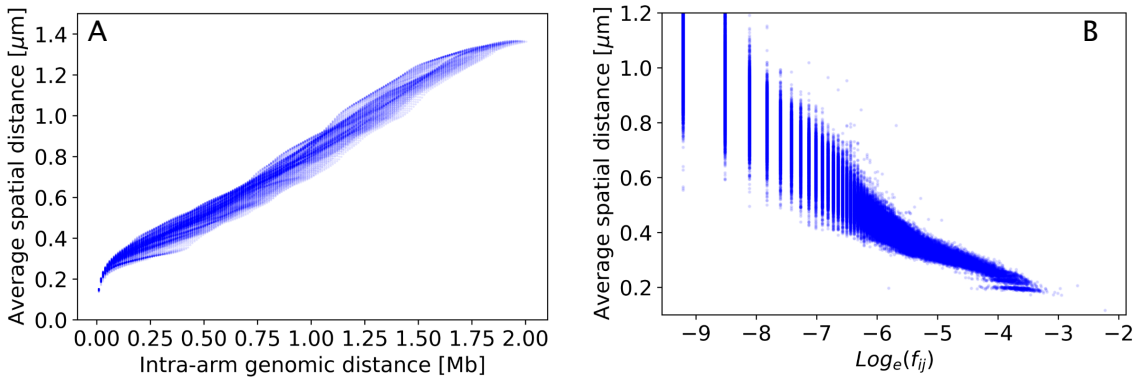



**Supplementary Figure 23: Analysis of the degree of overlap between peaks in local chromosome extension and the locations of highly transcribed genes. A** Wild-type local chromosome extension profile (black line), together with a trend line obtained from Gaussian smoothing (grey line) and the locations of highly transcribed genes (HTGs) (vertical dashed lines). **B** Green solid line: fraction of local extension peaks that coincide with the location of a highly transcribed gene, as a function of the cutoff factor $\alpha$. The dashed line indicates the expected fraction of overlap for randomly chosen locations of peaks, the light green area indicates the 95% confidence interval around this expected fraction. The grey line indicates the number of peaks included for a given cutoff factor (indicated on the right axis). **C** The same analysis as in B, performed separately for the right (0-2 Mb, blue) and left (2-4Mb, red) chromosomal arms. **D,E** The same analyses as in B and C, using only the positions of HTGs located on the forward strand of the chromosome. **F,G** The same analyses as in B and C, using only the positions of HTGs located on the reverse strand of the chromosome.

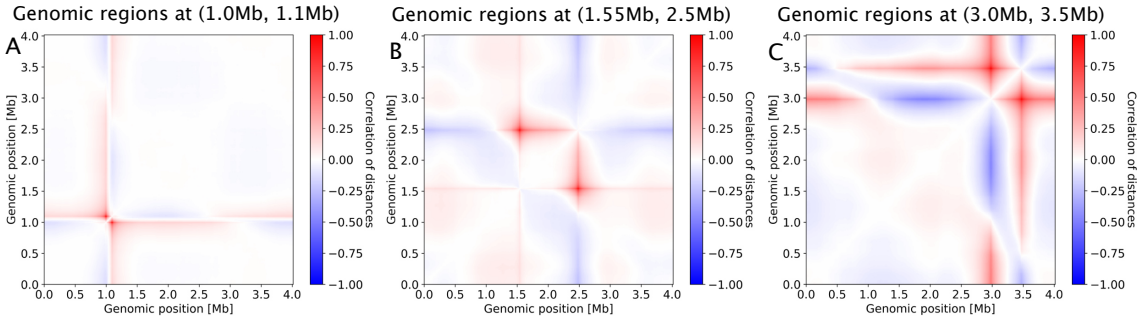
# 11 Relation between Hi-C scores and average distance and distance correlations

Previous modelling approaches for the *C. crescentus* chromosome used average distance based models to find typical chromosome configurations [13, 14]. In these approaches, an experimentally determined average linear relation between intra-arm genomic distances and average spatial distances was used to derive a functional relation between Hi-C contact scores and average spatial distances. Our MaxEnt model does not require this assumption, instead we can use the model to predict the relation between Hi-C scores and average distances. Interestingly, our MaxEnt model predicts an approximately linear relation between Hi-C scores and average distances, but with significant deviations from this average trend for individual pairs of genomic regions (Supplementary Figure 25A). Moreover, there are substantial deviations from a linear trend for small and large genomic distances. Finally, we also observe significant variations around an average trend for Hi-C scores versus spatial distances (Supplementary Figure 25B).

In addition to these variations in average spatial distances, we also find significant correlations in deviations from these averages for individual configurations throughout the entire chromosome (Supplementary Figure 25). In previously used approaches [13, 14] such correlations could not be taken into account, which could explain the difference in predictions from our MaxEnt model.




**Supplementary Figure 24: Variations of average distance statistics between individual pairs of genomic regions. A** Average spatial distance versus genomic distance predicted by the MaxEnt model. **B** Average spatial distance versus the logarithm of the Hi-C score predicted by the MaxEnt model.

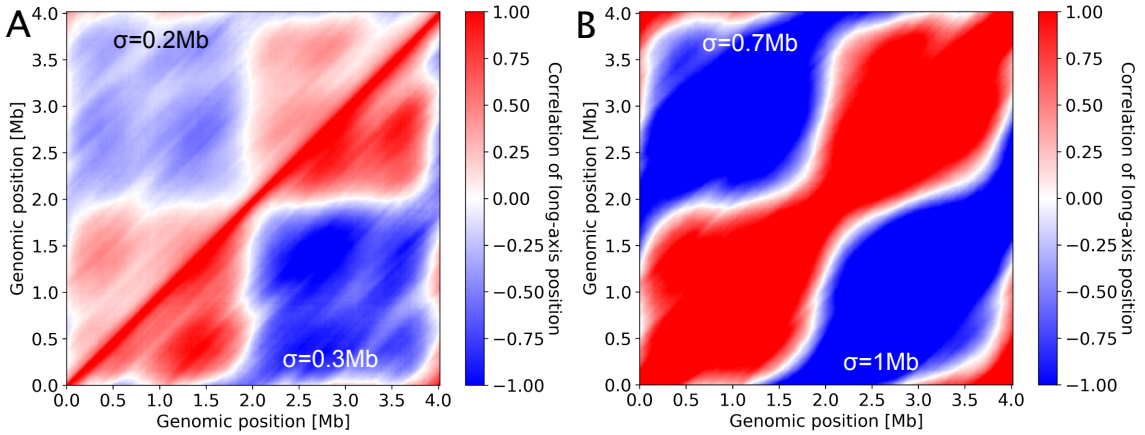



**Supplementary Figure 25: Correlations between distances of all pairs of genomic regions, and the distance between three sample pairs.** The chosen sample pairs are: **A** genomic regions at (1.0Mb, 1.1Mb), **B** genomic regions at (1.55Mb, 2.5Mb), **C** genomic regions at (3.0Mb, 3.5Mb).

# 12 A global rotation does not produce the observed long-axis correlation pattern

To illustrate the features of a long-axis correlation map that would be induced by a global rotation, we simulated the effects of such rotational fluctuations. Specifically, we took a set of configurations from our model, and generated an ensemble of new configurations by performing a rotational fluctuation with a random magnitude of all genomic regions along the polymers axial coordinate within each configuration. The magnitude of this rotation was drawn from a zero-average normal distribution, with the standard deviation $\sigma$ treated as a free parameter. For this new ensemble of configurations, including global rotation fluctuations, the long-axis correlations were calculated between all genomic regions. The resulting long-axis correlation maps for this rotational model for four choices of the standard deviation are shown in Supplementary Figure 26.

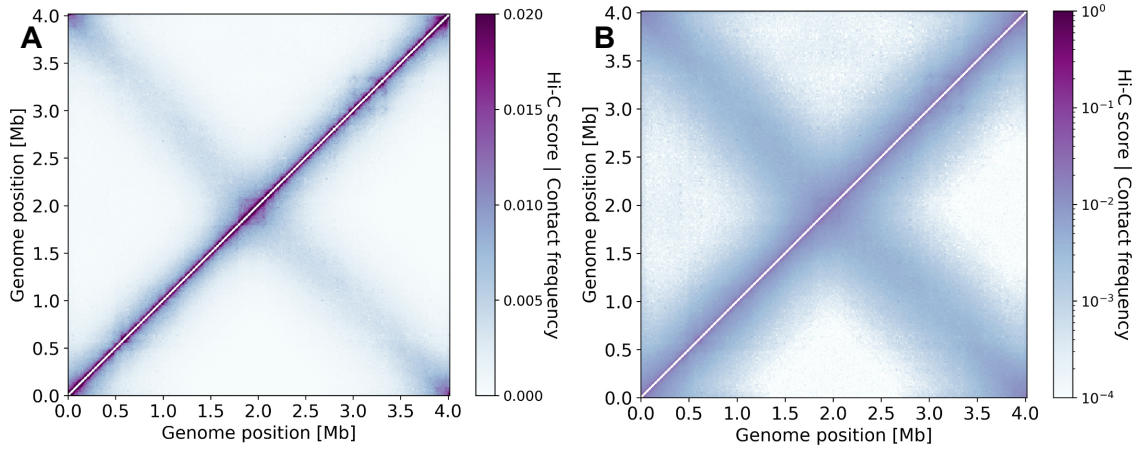
We see that for $\sigma = 0.2Mb$, the magnitude of correlations in the rotation model (Supplementary Figure 26A, upper left) is comparable to those observed in the original MaxEnt model (Main Text Fig. 3B, upper left). Importantly however, the anticorrelations in the rotation model are present between all genomic regions on opposite stretches of the chromosome. Thus, in this case, we see anti-correlation both between opposing genomic regions on the left and right chromosome arm and between opposing genomic regions near *ori* and *ter*. This is in contrast to the pattern observed in the original MaxEnt model, where the anticorrelations are only present between juxtaposed genomic regions lying on opposite sides of the left and right chromosome arms and opposing genomic regions near *ori* and *ter* exhibit positive correlations (Main Text Fig. 3B, upper left). For larger values of $\sigma$, the anticorrelation pattern in the rotation model initially remains qualitatively the same as for low $\sigma$, but the magnitude of correlations increases (Supplementary Figure 26A, lower right). For even larger values of $\sigma$, the long-axis correlation pattern starts to qualitatively change: the region of anticorrelation between *ori* and *ter* becomes larger (Supplementary Figure 26B). Furthermore, the magnitude of anticorrelations is much higher for these values of $\sigma$ than observed in the original MaxEnt model.



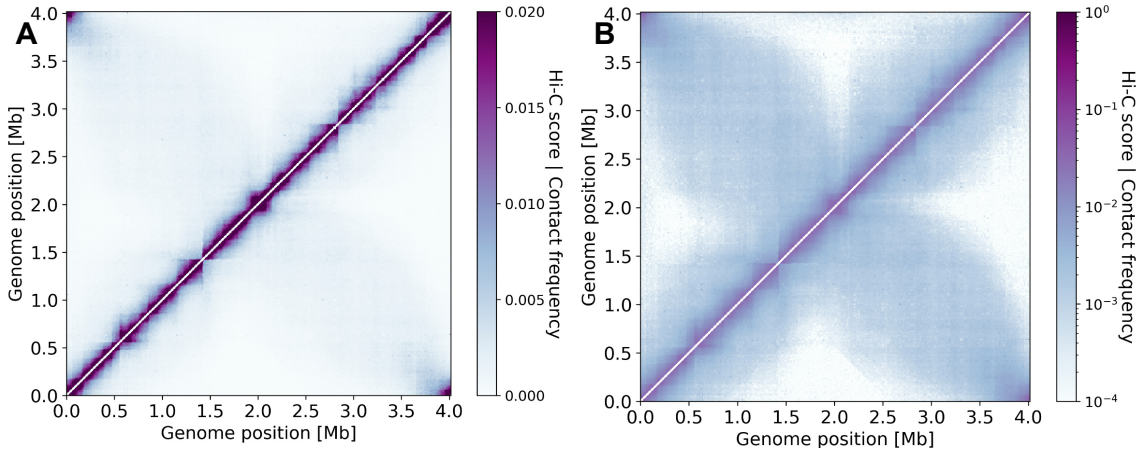

**Supplementary Figure 26: Long-axis correlations for chromosome configurations with global rotational fluctuations.** A) Upper left: long-axis correlations for model configurations with global rotational fluctuations along the polymer axis, drawn from a normal distribution with $\sigma$=0.2Mb. Lower right: the same for $\sigma$=0.3Mb. B) Upper left: same for $\sigma$=0.7Mb, lower right: same for $\sigma$=1Mb.

## 13 MaxEnt models for $\Delta smc$ cells and rifampicin-treated cells

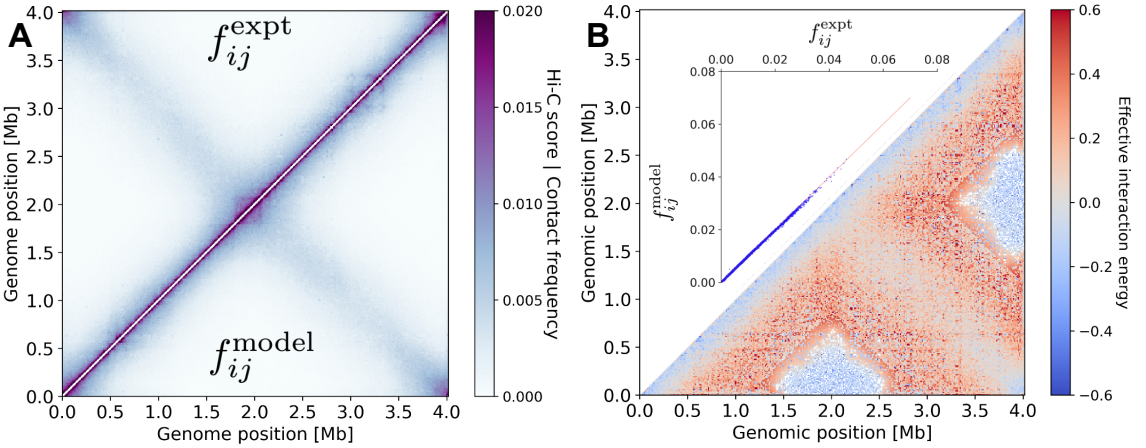
We apply the same approach to perform a Hi-C data analysis and MaxEnt model inference for rifampicin-treated cells and $\Delta smc$ cells. The prepossessing of Hi-C data is shown in Figs. 27 and 28, and the corresponding MaxEnt models are shown in Figs. 29 and 30. We show the results for the long-axis localization in Supplementary Figure 31 together with previously published experimental data, and various correlation functions are depicted in Supplementary Figure 32.



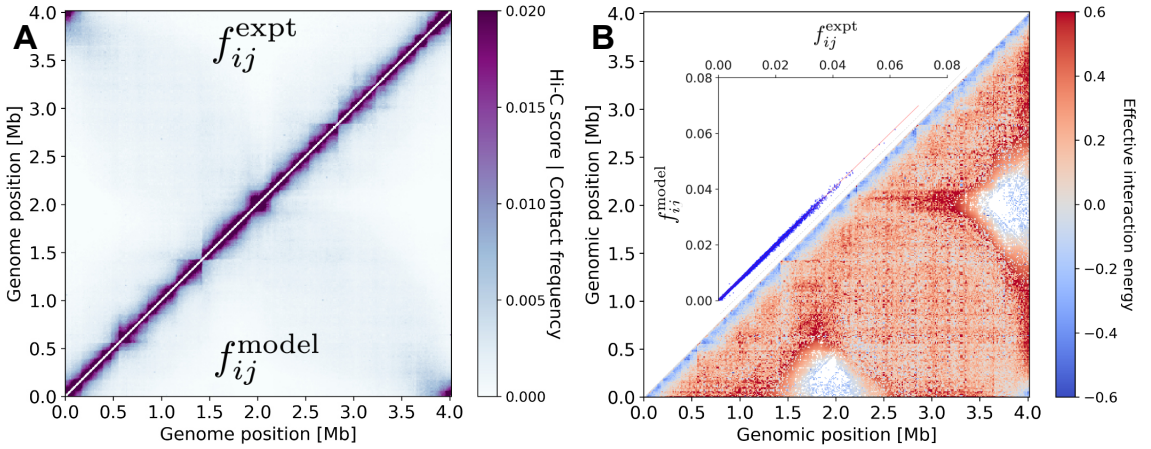

**Supplementary Figure 27: Hi-C scores for rifampicin-treated cells before and after correction.** Hi-C scores of rifampicin-treated cells before correction (upper left triangle), and after correction (lower right triangle) on a linear scale (A) and a logarithmic scale (B).



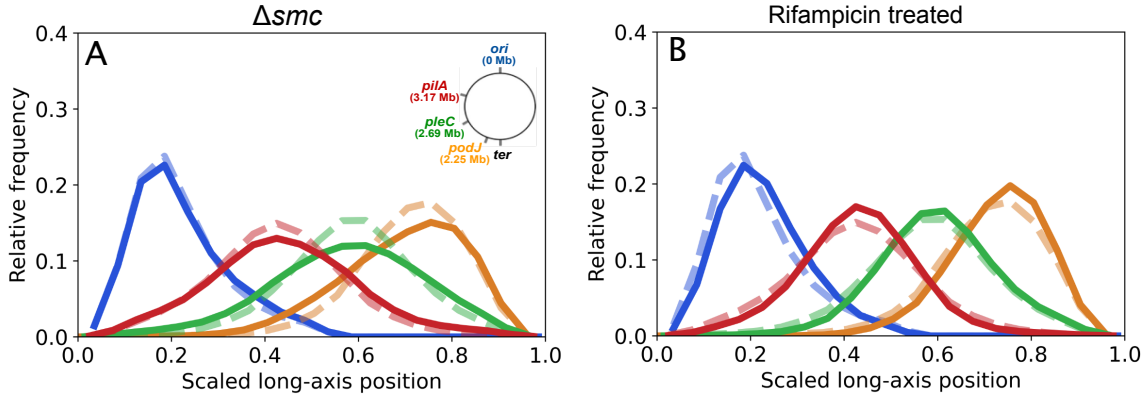

**Supplementary Figure 28: Hi-C scores for $\Delta smc$ cells before and after correction.** Hi-C scores of $\Delta smc$ cells before correction (upper left triangle), and after correction (lower right triangle) on a linear scale (A) and a logarithmic scale (B).



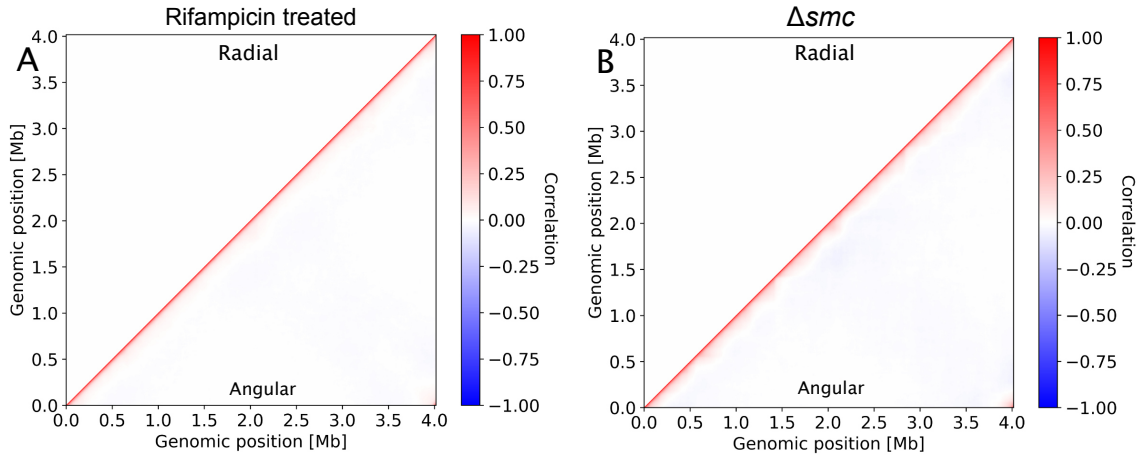

**Supplementary Figure 29: Maximum entropy model inferred for rifampicin-treated cells. A** Comparison between experimental contact frequencies $f_{ij}^{\text{expt}}$ (upper left corner, adapted from Ref. [12]) and contact frequencies obtained from our inferred MaxEnt model $f_{ij}^{\text{model}}$ (lower right corner). **B** Inferred effective interaction energies $\epsilon_{ij}$ (lower right corner) together with scatterplot of $f_{ij}^{\text{expt}}$ vs. $f_{ij}^{\text{model}}$ (inset).




**Supplementary Figure 30: Maximum entropy model inferred for Δ*smc* cells. A** comparison between experimental contact frequencies $f_{ij}^{\text{expt}}$ (upper left corner, adapted from Ref. [12]) and contact frequencies obtained from our inferred MaxEnt model $f_{ij}^{\text{model}}$ (lower right corner). **B** Inferred effective interaction energies $\epsilon_{ij}$ (lower right corner) together with scatterplot of $f_{ij}^{\text{expt}}$ vs. $f_{ij}^{\text{model}}$ (inset).




**Supplementary Figure 31: Distribution of long-axis positions for Δ*smc* and rifampicin-treated cells.** Comparison between inferred long-axes localization distributions for wild-type cells (dashed lines) and Δ*smc* mutants (**A**, solid lines) and rifampicin-treated cells (**B**, solid lines).




**Supplementary Figure 32: Radial and angular correlations for Δ*smc* and rifampicin-treated cells.** Correlations in the radial positions (upper left corner) and orientations around the long axis (lower right corner) between all pairs of genomic regions, for rifampicin-treated cells (**A**) and Δ*smc* mutants (**B**).

## 14 Estimates of localization information

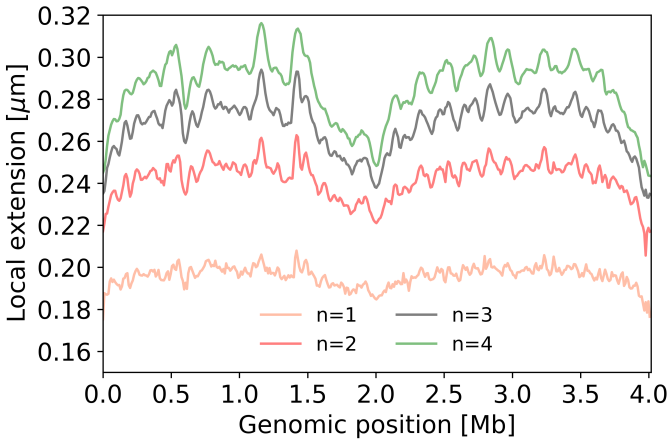
To compute the localization information for a genomic region, we first calculate the average occupation $P^{s,i}$ of each unit cell $s$ for each genomic region $i$ during a forward simulation. The localization entropy $S^i_{\text{loc}}$ in bits of site $i$ is then calculated by [16]

$$S^i_{\text{loc}} = -\sum_s P^{s,i} \log_2 P^{s,i}. \tag{S8}$$

The positional information is calculated by subtracting $S^i_{\text{loc}}$ from the localization entropy of a flat distribution.

A possible issue with calculating positional information within a coarse-grained model, is that the obtained value is an underestimate. This is the case if the localization is confined to a region approximately the size of a unit cell. Since we find the localizations of genomic regions to be significantly larger than this (Main Text Fig. 2B), we do not expect our estimate to be sensitive to the course graining scale.

## 15 Local extension interval and origin of *ori* and *ter* extensions




**Supplementary Figure 33: Change of local extension with genomic distance** Local extensions, defined as the average distance between the $n^{\text{th}}$ nearest neighbours of a genomic region, shown for $n = 1$ up to $n = 4$. The value of $n = 2$ is shown in the main text as its features are more prominent than those for $n = 1$, but less smoothened out than for higher values of $n$. The locations of the peaks are largely identical between these different choices for $n$.

A possible explanation for the low local extension of the *ori* and *ter* regions, would be the turning around of the average long-axis positions at these regions. As the local extension of a region is calculated as the average geometric distance between its $n^{th}$ neighbours, such an effect could cause the observed low local extension. To test if this is the case, we make use of the presence

of variations in the positions of the *ori* and *ter*; for a subset of states, these will not be the furthest regions along the long axis. If the inferred low local extension is indeed due to a 'turning around' of the chromosome at the *ori* and *ter*, the local extension would be expected to be higher for this subset of states.

Taking a conditional average of the local extension of the *ori* over states where the previous 5 or subsequent 5 genomic regions all have an equal or lower long-axis position than the *ori* region, we find an increase of only 2% compared to an average over all states. For the *ter* region, we find the same statistics (2% increase if either set of 5 neighboring regions has a higher or equal long-axis position than the *ter*). Thus, the inferred local density of the *ori* and *ter* regions reflect the intrinsic extensions of these regions, rather than artefacts due to a turning around of the average long axis positions at these sites.

# 16 Linear spatial organization of a polymer with juxtaposed chromosomal arms

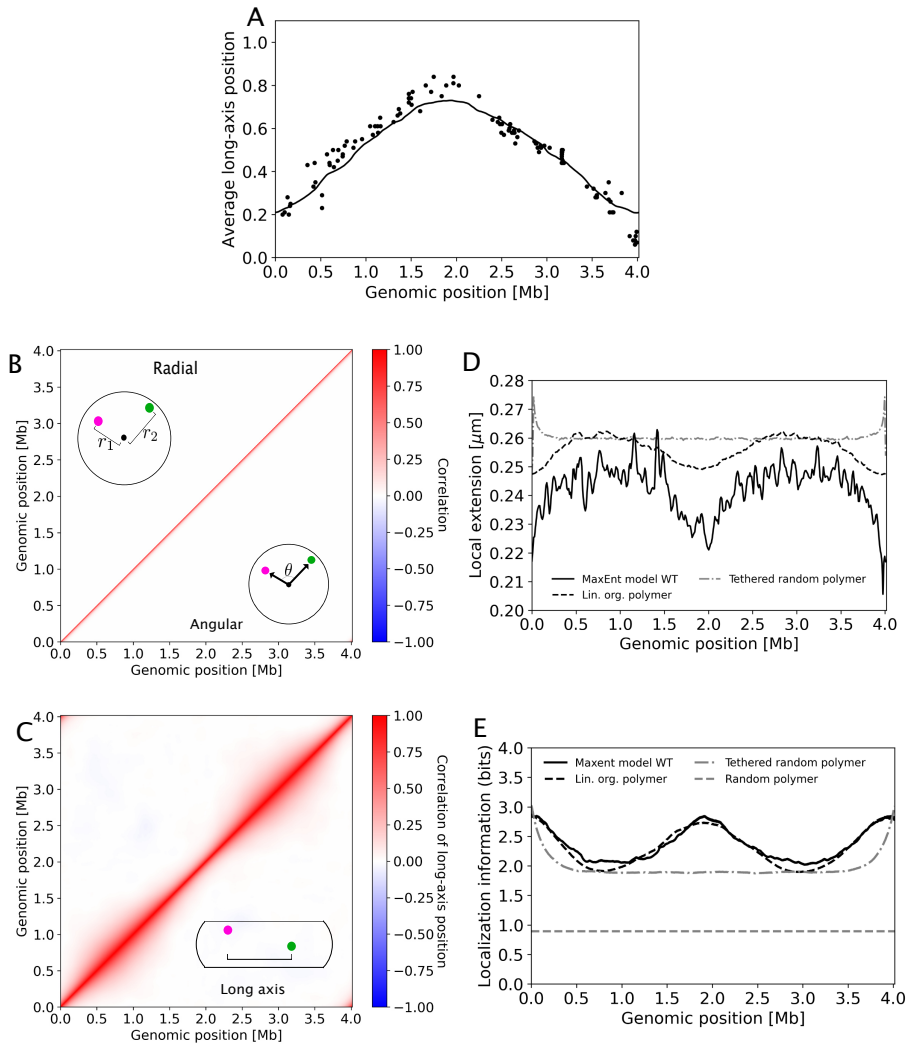
To investigate organizational features of a polymer with juxtaposed arms, but no additional structure, we derive a MaxEnt model taking average long-axis positions as the only constraints. This model we term the linearly organized polymer model. To enable a direct comparison to MaxEnt models learned from Hi-C data, we take the average long-axis positions predicted from these models and use these as constraints for the linearly organized polymer. This allows us to investigate to what extent features of the MaxEnt model based on Hi-C data are due to its linear organization throughout the cell. For the linearly organized polymer model, the entropy functional takes the following form:

$$\tilde{S} = -\sum_{\{\mathbf{r}\}} P(\{\mathbf{r}\}) \ln P(\{\mathbf{r}\}) - \sum_i \lambda_i \left( \sum_{\{\mathbf{r}\}} P(\{\mathbf{r}\}) z_i(\mathbf{r}) \quad - \langle z_i \rangle^{\text{cons}} \right) - \lambda_0 \left( \sum_{\{\mathbf{r}\}} P(\{\mathbf{r}\}) - 1 \right). \quad \text{(S9)}$$

Here, $z_i(\mathbf{r})$ denotes the long-axis position of region $i$ in configuration $\mathbf{r}$, and $\langle z_i \rangle^{\text{cons}}$ denotes the imposed average long-axis position of region $i$. Extremizing this entropy functional and solving for $P(\{\mathbf{r}\})$ yields

$$P(\{\mathbf{r}\}) = \frac{1}{Z} \exp\left[ -\sum_i \lambda_i z_i(\mathbf{r}) \right], \quad \text{(S10)}$$

with $Z = \exp[1 + \lambda_0]$ as in the main text. The solutions for $\lambda_i$ were found with an iterative Monte Carlo algorithm similar to the one presented in 3, where the update of $\lambda_i$ at each iteration of the inverse algorithm is now proportional to $\langle z_i \rangle^{\text{cons}} - \langle z_i \rangle^{\text{model}}$. The resulting organizational properties of the linearly organized polymer model are presented in Supplementary Figure 34.

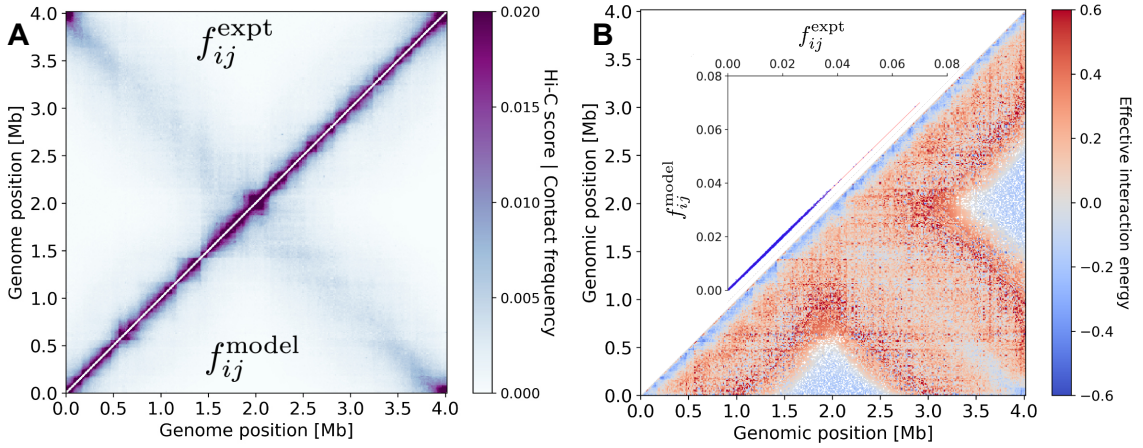



**Supplementary Figure 34: Model results for the linearly organized polymer model. A**) Average localization profile used as a constraint of the linearly organized polymer (line) together with the experimental FISH [3] data shown in Main Text Fig. 2A. (dots). **B**)) Radial correlations (upper left triangle) and angular correlations (lower right triangle) for the linearly organized polymer. **C**)) Long-axis correlations for the linearly organized polymer. **D**)) Results for the local extension as in Main Text Fig. 5A, together with those for the linearly organized polymer. **E**)) Results for the localization information as in Main Text Fig. 5B, together with those for the linearly organized polymer.

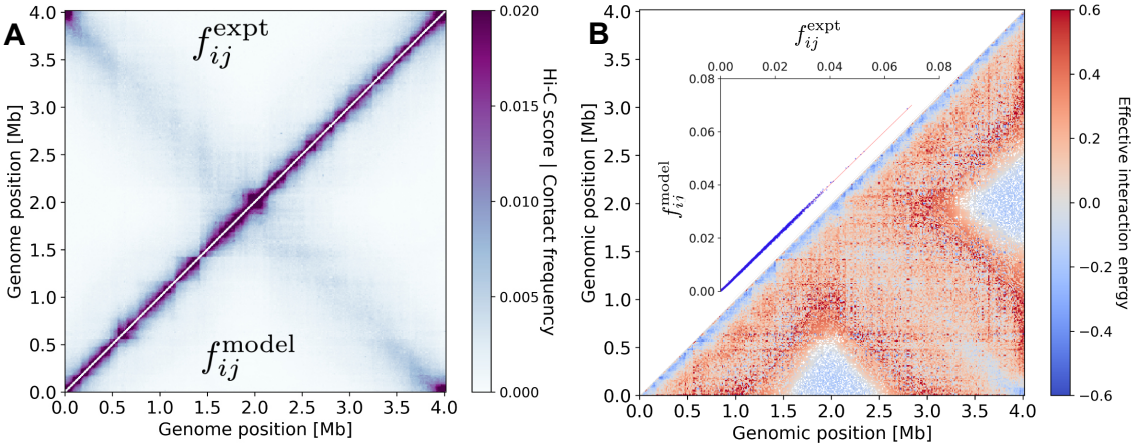
## 17 Independence of results for modified MaxEnt models

To test if our results are robust under minor model modifications, we inferred two alternative MaxEnt models: one with a slightly curved confinement, and one with a tethered *ori*. The former incorporates the typically observed *C. crescentus* cell shape, the latter enforces the experimentally measured long-axis distribution of the position of the *ori* locus. The inferred models are shown in Figs. 36 and 37.

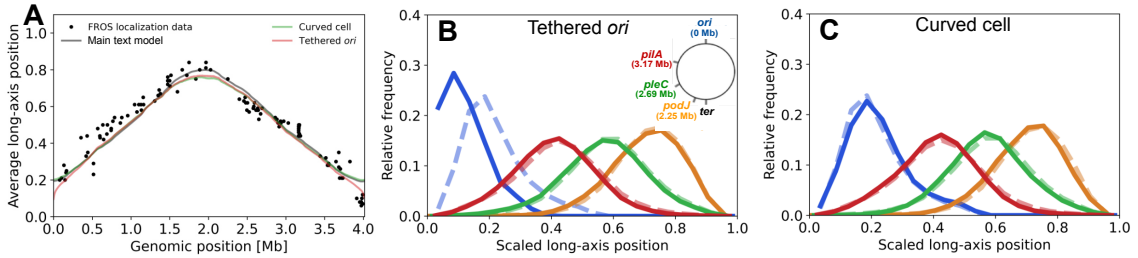

**Supplementary Figure 35: Top view of the curved cell shape used for analyses presented in this Note.** A lattice spacing corresponds to 88nm, as in the Main Text model.



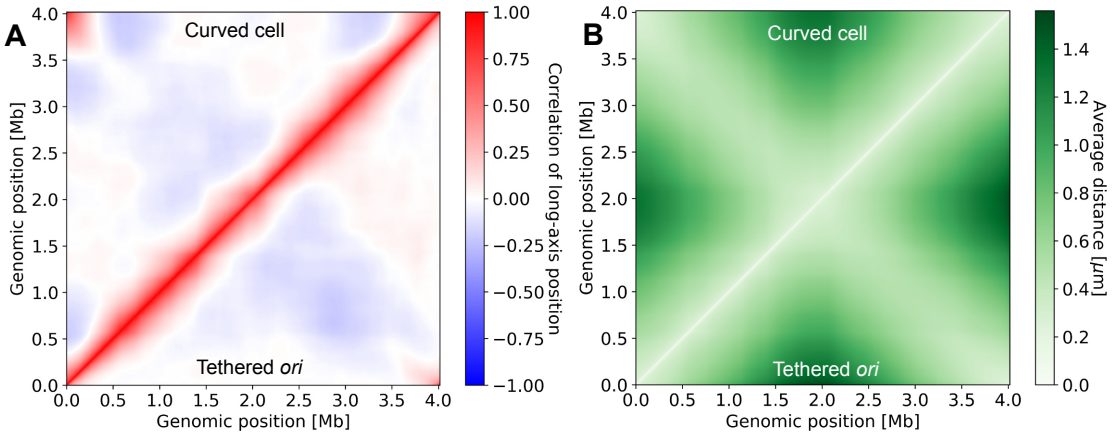

**Supplementary Figure 36: Results for Main Text Fig. 1, re-analyzed for a model with tethered *ori*.** **A** Comparison between experimental contact frequencies $f_{ij}^{\text{expt}}$ (upper left corner, adapted from Ref. [12] and contact frequencies obtained from our inferred MaxEnt model $f_{ij}^{\text{model}}$ (lower right corner). **B** Associated inferred effective interaction energies $\epsilon_{ij}$ (lower right corner, white regions indicate $\epsilon_{ij} \to \infty$) together with a scatter plot of $f_{ij}^{\text{expt}}$ vs. $f_{ij}^{\text{model}}$ (inset).



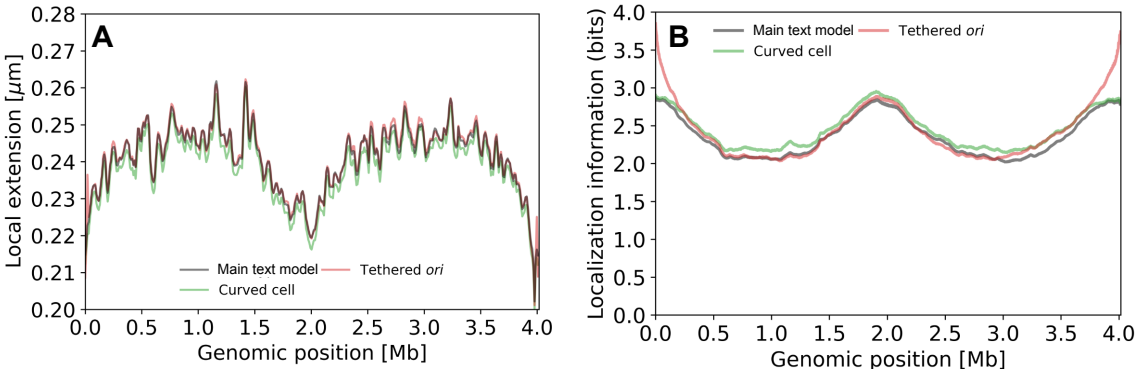

**Supplementary Figure 37: Results for Main Text Fig.1, re-analyzed for a model with a curved cell A** Comparison between experimental contact frequencies $f_{ij}^{\text{expt}}$ (upper left corner, adapted from Ref. [12] and contact frequencies obtained from our inferred MaxEnt model $f_{ij}^{\text{model}}$ (lower right corner). **B** Associated inferred effective interaction energies $\epsilon_{ij}$ (lower right corner, white regions indicate $\epsilon_{ij} \to \infty$) together with a scatter plot of $f_{ij}^{\text{expt}}$ vs. $f_{ij}^{\text{model}}$ (inset).




**Supplementary Figure 38: Results for Main Text Fig. 2, re-analyzed for a model with a tethered *ori* and a curved cell. A** Average scaled long-axis position predicted from MaxEnt models (solid lines) inferred for various MaxEnt models, including the model described in the main text (black), a model for a curved cell (green), and a model with a tethered *ori* (red), together with results from microscopy experiments (adapted from [3]). **B** Solid lines: localizations for a MaxEnt model with a tethered *ori*. Dashed lines: Maxent model results as presented in Main Text Fig. 2. **C** Solid lines: localizations for a MaxEnt model with a curved cell. Dashed lines: Maxent model results as presented in Main Text Fig. 2.




**Supplementary Figure 39: Long-axis correlations and average distances for Maxent models with a curved cell and a tethered *ori*.** **A** Long-axis correlations for a Maxent model with a curved cell (top left) and a tethered *ori* (bottom right). **B** Average distances for a Maxent model with a curved cell (top left) and a tethered *ori* (bottom right).




**Supplementary Figure 40: Results for Main Text Fig. 5, re-analyzed for a model with a tethered *ori* and a curved cell.** **A** The local chromosome extension $\delta_i$ as a function of genomic position. Model prediction are shown for the model described in the main text (black), a model for a curved cell (green), and a model with a tethered *ori* (red). **B** Localization information per genomic region in bits for the model described in the main text (black), a model for a curved cell (green), and a model with a tethered *ori* (red).

## References


[1] Lau, I. F. *et al.* Spatial and temporal organization of replicating *Escherichia coli* chromosomes. *Mol. Microbiol.* **49**, 731–743 (2004).

[2] Ely, B. Genetics of *Caulobacter crescentus*. *Methods in Enzymology* **204**, 372–384 (1991).

[3] Viollier, P. H. *et al.* Rapid and sequential movement of individual chromosomal loci to specific subcellular locations during bacterial DNA replication. *Proc. Natl. Acad. Sci. USA* **101**, 9257–9262 (2004).

[4] Tsai, J. W. & Alley, M. R. Proteolysis of the *Caulobacter* McpA chemoreceptor is cell cycle regulated by a ClpX-dependent pathway. *J. Bacteriol.* **183**, 5001–5007 (2001).

[5] University of Sussex. Image J Analysis: Single-molecule plugins (17.01.2020).

[6] Rueden, C. T. *et al.* ImageJ2: ImageJ for the next generation of scientific image data. *BMC Bioinformatics* **18**, 529 (2017). arXiv:1701.05940.

[7] Evinger, M. & Agabian, N. Envelope associated nucleoid from *Caulobacter crescentus* stalked and swarmer cells. *J. Bacteriol.* **132**, 294–301 (1977).

[8] Thanbichler, M., Iniesta, A. A. & Shapiro, L. A comprehensive set of plasmids for vanillate and xylose-inducible gene expression in *Caulobacter crescentus*. *Nucleic Acids Res.* **35**, e137–e137 (2007).

[9] De Martino, D., MC Andersson, A., Bergmiller, T., Guet, C. C. & Tkačik, G. Statistical mechanics for metabolic networks during steady state growth. *Nature Communications* **9**, 2988 (2018).

[10] Hartmann, R., Van Teeseling, M. C., Thanbichler, M. & Drescher, K. Bacstalk: a comprehensive and interactive image analysis software tool for bacterial cell biology. *Molecular Microbiology* (2020).

[11] Gan, L., Chen, S. & Jensen, G. J. Molecular organization of Gram-negative peptidoglycan. *Proc. Natl. Acad. Sci. USA* **105**, 18953–18957 (2008).

[12] Le, T. B. K., Imakaev, M. V., Mirny, L. A. & Laub, M. T. High-resolution mapping of the spatial organization of a bacterial chromosome. *Science* **342**, 731–734 (2013).

[13] Umbarger, M. A. *et al.* The three-dimensional architecture of a bacterial genome and its alteration by genetic perturbation. *Mol. Cell* **44**, 252–264 (2011).

[14] Yildirim, A. & Feig, M. High-resolution 3D models of *Caulobacter crescentus* chromosome reveal genome structural variability and organization. *Nucleic Acids Res.* **46**, 3937–3952 (2018).

[15] Le, T. B. & Laub, M. T. Transcription rate and transcript length drive formation of chromosomal interaction domain boundaries. *EMBO J.* **35**, 1582–1595 (2016).

[16] Dubuis, J. O., Tkacik, G., Wieschaus, E. F., Gregor, T. & Bialek, W. Positional information, in bits. *Proc. Natl. Acad. Sci. USA* **110**, 16301–16308 (2013).

## 2.2 Epilogue: Analytical contact frequencies for a lattice polymer subject to pairwise interaction energies

During the development of the Monte Carlo algorithm used in this chapter, we also derived analytical expressions for contact frequencies of a lattice polymer subject to pairwise interactions. These analytical expressions proved useful in validating the Monte Carlo algorithm by checking if the forward algorithm converged to the correct results for a given set of input energies. We present the developed analytics in this chapter epilogue.

### Contact frequencies for a free circular lattice polymer in 2D

We consider a circular polymer in 2 dimensions on a square lattice. We can construct such a polymer by starting on a lattice site which we call the origin. From there, a series of $N$ consecutive moves are made, where each move can be in four directions: up (u), down (d), left (l) and right (r). To ensure that the polymer forms a loop (i.e. that the last line returns at the origin again), the number of up moves must equal the number of down moves, and the number of right moves must equal the amount of left moves. Furthermore, it must hold that $N_u + N_d + N_r + N_l = N$, with $N_u$, $N_d$, $N_r$, $N_l$ representing the total number of up, down, right and left moves respectively.

We now derive an expression for the number of possible polymers we can construct in this way. To do this, we first realize that for a polymer with a number of up moves given by $N_u$ has

- $N_u$ down moves
- $\frac{N}{2} - N_u$ left moves
- $\frac{N}{2} - N_u$ right moves.

When the number of up moves $N_u$ is specified, the number of down, left and right moves are thus automatically determined as well. The question is then in how many ways these moves can be ordered.

Firstly, the number of ways we can distribute $N_u$ up moves over $N$ total moves, is given by $\binom{N}{N_u}$. The number of down moves (also given by $N_u$) we can then distribute over the remaining $N - N_u$ moves in $\binom{N-N_u}{N_u}$ ways. Subsequently, the $\frac{N}{2} - N_u$ left moves we can distribute in $\binom{N-2N_u}{\frac{N}{2}-N_u}$ ways. The places of the right-moves are uniquely determined after this. Putting this together, the total number of configurations (or multiplicity) $M_{tot}(N)$ of our polymer of length $N$ is given by

$$M_{tot}(N) = \sum_{N_u=0}^{N/2} \binom{N}{N_u} \times \binom{N-N_u}{N_u} \times \binom{N-2N_u}{\frac{N}{2}-N_u}. \tag{2.1}$$

Evaluating this expression, we obtain

$$M_{tot}(N) = \sum_{N_u=0}^{N/2} \frac{N!}{N_u!\,(N-N_u)!} \times \frac{(N-N_u)!}{N_u!\,(N-2N_u)!} \times \frac{(N-2N_u)!}{(\frac{N}{2}-N_u)!\,(\frac{N}{2}-N_u)!} \tag{2.2}$$

$$= \sum_{N_u=0}^{N/2} \frac{N!}{(N_u!\,)^2((\frac{N}{2}-N_u)!\,)^2}. \tag{2.3}$$

Evaluating the summation using Mathematica, the following expression is obtained:

$$M_{tot}(N) = \frac{4^N((\frac{1+N}{2}-1)!\,)^2}{\pi((\frac{N}{2})!\,)^2}. \tag{2.4}$$

We now consider the subset of states of the system where there is an overlap between two sites that are a distance of $d$ sites apart, as measured along the filament. This subset we can consider as comprising two circular polymers; one of length $d$ and one of length $N-d$. The total number of states $M^o_{tot}$ of this system is equal to the product of the number of states of each of these to smaller loops. It is thus given by

$$M^o_{tot}(N,d) = \frac{4^{(N-d)}((\frac{1+N-d}{2}-1)!\,)^2}{\pi((\frac{N-d}{2})!\,)^2} \times \frac{4^d((\frac{1+d}{2}-1)!\,)^2}{\pi((\frac{d}{2})!\,)^2}. \tag{2.5}$$

If we now divide equation (2.5) by equation (2.4), we obtain the fraction of time two sites at a distance of $d$ overlap, i.e. their contact frequency. This contact frequency $f_c(N,d)$ is given by

$$f_c(N,d) = \frac{((\frac{N-d}{2}-\frac{1}{2})!\,)^2 \times ((\frac{d}{2}-\frac{1}{2})!\,)^2 \times ((\frac{N}{2})!\,)^2}{\pi((\frac{N-d}{2})!\,)^2 \times ((\frac{d}{2})!\,)^2 \times ((\frac{N}{2}-\frac{1}{2})!\,)^2}. \tag{2.6}$$

We can approximate this expression using the fact that $\frac{(x-\frac{1}{2})!}{x!}$ can be expanded as $\sqrt{\frac{1}{x}} - \frac{(\frac{1}{x})^{3/2}}{8} + ...$ at $x \to \infty$. The expression for the contact frequency then becomes

$$f_c(N,d) \approx \frac{1}{\pi}\frac{2N}{d(N-d)}. \tag{2.7}$$

### Contact frequencies for a circular polymer with one pair of sites $(i, j)$ with an interaction energy

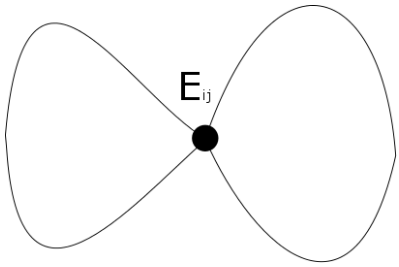



Figure 2.1: A polymer with one pair of sites $i, j$ with an interaction energy $E_{ij}$.

Consider a circular polymer on a lattice with one pair of sites, $i$ and $j$, that upon overlapping result in an energy gain of $E_{ij}$. To derive an expression for the contact frequencies of this polymer, we first write down the partition function. Each of the configurations of the polymer with an overlap between sites i and j has a weight of $e^{E_{ij}}$, whereas all other conformations have a weight of 1 (corresponding to a state with zero energy). We can thus write for the partition function

$$Z = f_c(N, d_{ij})e^{E_{ij}} + (1 - f_c(N, d_{ij})). \tag{2.8}$$

Here, $d_{ij}$ is the shortest distance along the polymer between sites $i$ and $j$. Note that we use the fraction of states with an overlap between $i$ and $j$, and not the total number of states, as a prefactor for the weight $e^{E_{ij}}$. We use this rather than the total number of states to facilitate computation, which leaves expectation values unchanged as long as this replacement is consistently applied.
We now compute the contact frequencies for a few categories of monomer pairs.

#### Contact frequencies between sites $i$ and $j$

To get the contact frequencies between sites $i$ and $j$, we want to single out exactly those states with a weight of $e^{E_{ij}}$. This frequency $P(C_{ij})$ is thus given by

$$P(C_{ij}) = \frac{1}{Z} f_c(N, d_{ij})e^{E_{ij}}. \tag{2.9}$$

#### Contact frequencies between two sites that are on the same side of $i$ and $j$

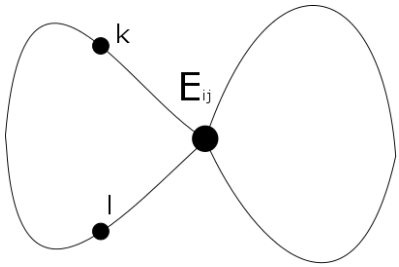



Figure 2.2: A polymer with one pair of sites $i, j$ with an interaction energy $E_{ij}$ where we consider the contact frequency between two sites $k, l$ that are on the same side of $i$ and $j$.

For two sites $k$ and $l$ that are on the same side of sites $i$ and $j$ (i.e. it is possible to walk from site $k$ to $l$ along the polymer without coming across site $i$ or $j$) we can distinguish the configurations in two classes: 1) those with a contact between $i$ and $j$ and 2) those without this contact. Weighing each of these cases by the prefactor obtained in (2.9) we obtain

$$P(C_{kl}) = P(C_{ij}) \times P(C_{kl}|C_{ij}) + (1 - P(C_{ij})) \times P(C_{kl}|NC_{ij}). \tag{2.10}$$

Here, $P(C_{kl}|C_{ij})$ represents the contact frequency between sites $k$ and $l$, given that sites $i$ and $j$ are in contact. It can be calculated using the expression for $f_c(N, d_{ij})$ given in equation (2.6) - with the adjustment that for $N$ we now take the length of the loop that includes sites $k$ and $l$ and is enclosed by sites $i$ and $j$. For $d_{ij}$ we then take the distance between $i$ and $j$ within this loop.
Similarly, $P(C_{kl}|NC_{ij})$ represents the contact frequency between sites $k$ and $l$ given that sites $i$ and $j$ are not in contact. This quantity can be calculated as

$$P(C_{kl}|NC_{ij}) = f_c(N, d_{kl}) \frac{1 - f_c(N', d'_{ij}|C_{kl})}{1 - f_c(N, d_{ij})} \tag{2.11}$$

Here, $f_c(N, d_{kl})$ represents the contact frequency between sites $k$ and $l$ for a free circular polymer (without interaction energies). This quantity has a correction factor to account for the fact that we are only considering conformations in which there is no contact between sites $i$ and $j$. Within this correction factor, the term $f_c(N', d'_{ij}|C_{kl})$ represents the contact frequency between sites $i$ and $j$ for a free circular polymer given that sites $k$ and $l$ are in contact. $N'$ then represents the length of the loop that includes sites $k$ and $l$ and is enclosed by sites $i$ and $j$. The term $d'_{ij}$ then represents the distance between $i$ and $j$ within this loop.

This correction factor can be derived as follows. The quantity $P(C_{kl}|NC_{ij})$ can be calculated as

$$P(C_{kl}|NC_{ij}) = \frac{M(C_{kl}) - M(C_{kl}, C_{ij})}{M_{tot} - M(C_{ij})} \tag{2.12}$$

Here, $M(C_{kl})$ is the multiplicity of all states that include a contact between sites $k$ and $l$, $M(C_{kl}, C_{ij})$ is the multiplicity of all states that include a contact between sites $k$ and $l$ and between sites $i$ and $j$. The logic here is that we first calculate the total number of states, and then subtract the number of states in which there is a contact between $i$ and $j$. This expression can be rewritten to

$$P(C_{kl}|NC_{ij}) = \frac{M(C_{kl})}{M_{tot}} \times \frac{1 - \frac{M(C_{kl}, C_{ij})}{M(C_{kl})}}{1 - \frac{M(C_{ij})}{M_{tot}}} \tag{2.13}$$

This expression is the same as equation (2.11), if the contact frequency notation is used. Note that for sufficiently large distances between $i$ and $j$ along the polymer, we can safely approximate $P(C_{kl}|NC_{ij}) \approx f_c(N, d_{kl})$.

#### Contact frequencies between two sites that are on opposite sides of $i$ and $j$

Calculating the contact frequencies for two sites that are on opposite sides of $i$ and $j$ is analytically highly nontrivial. For two sites $k$ and $l$ that are both close to sites $i$ and $j$ we can however make an approximation. As long as for both $k$ and $l$ it holds that the distance to the point of overlap of $i$ and $j$ is much smaller than the size of the loop enclosed by $i$ and $j$ that

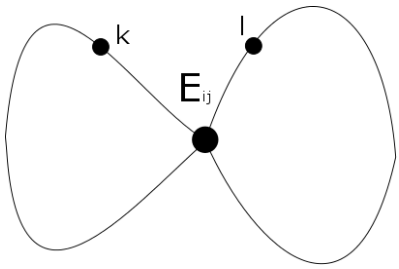



Figure 2.3: A polymer with one pair of sites $i, j$ with an interaction energy $E_{ij}$ where we consider the contact frequency between two sites $k, l$ that are on opposite sides of $i$ and $j$.

it is part of, we can treat the contact frequency between $k$ and $l$ as that of two free random walks on the lattice. The contact frequencies for these we can obtain as follows.

Consider two free polymers on the lattice, of lengths $L_1$ and $L_2$. The number of states in which the endpoints meet is given by $C_{tot}(L_1 + L_2)$. The total number of states these two polymers can be in is given by $4^{(L_1+L_2)}$. The contact frequency for these two polymers is given by the fraction of these two, i.e. by

$$P(C_{kl}|C_{ij}) = \frac{((\frac{1+L_1+L_2}{2} - 1)!)^2}{\pi((\frac{L_1+L_2}{2})!)^2}. \tag{2.14}$$

This can be approximated as

$$P(C_{kl}|C_{ij}) = \frac{2}{\pi}\frac{1}{L_1 + L_2}, \tag{2.15}$$

which we would also obtain if we make the approximation $N >> d$ in equation (2.7).
The full equation for the contact probability for the two sites on opposite sides of $i$ and $j$ is again

$$P(C_{kl}) = P(C_{ij}) \times P(C_{kl}|C_{ij}) + (1 - P(C_{ij})) \times P(C_{kl}|NC_{ij}). \tag{2.16}$$

The correction factor included in the expression for $P(C_{kl}|NC_{ij})$ is however slightly different now: it is given by

$$P(C_{kl}|NC_{ij}) = f_c(N, d_{kl})\frac{1 - P(C_{ij}|C_{kl})}{1 - f_c(N, d_{ij})}. \tag{2.17}$$

If for the quantity $P(C_{ij}|C_{kl})$ we again use the approximation used to obtain equation (2.14), we can set

$$P(C_{ij}|C_{kl}) = P(C_{kl}|C_{ij}) \tag{2.18}$$

And thus plug equation (2.14) into equation (2.17).

## Adding several sites with interaction energies

### The most general case: arbitrary interaction energies between all sites

Consider the most general case, where we have an interaction energy $E_{ij}$ between all pairs of sites $i, j$. The partition function for this system is as follows:

$$Z = M_N + \frac{1}{2}\sum_{i,j} M_{ij} e^{E_{ij}} + \frac{1}{8}\sum_{i,j,k,l} M_{ij,kl} e^{(E_{ij}+E_{kl})} + \cdots \tag{2.19}$$

The partition function is written out successively in states with zero contacts, with one contact, two contacts, etc. This allows us to group all microstates according to the energy of the microstate. The prefactors $C$ are equal to the number of microstates corresponding to a certain contact. Thus, $M_N$ is the number of states of the polymer for which there are no contacts, $M_{ij}$ is the number of states of the polymer in which there is only a contact between sites $i$ and $j$, and $M_{ij,kl}$ is the number of states in which there is only a contact between $i$ and $j$ and between $k$ and $l$.

The problem with this general approach is finding these numbers $M$. Just considering the first term, finding the value of $M_N$ amounts to finding the number of possible configurations for a self-avoiding circular lattice polymer of length $N$. Even in $2D$ this remains an open problem [129]. Instead of trying to solve for the contact frequencies for a general set of interactions, we therefore consider a few special cases that are more easily solvable.

### Non-crossing interaction energies

One set of energies for which contact frequencies are more readily computed, is a set in which there are no crossing interaction energies; for each nonzero $E_{ij}$ (with $i < j$) in the set all $E_{kl}$ must be zero if $k < i, i < l < j$ or $i < k < j, j < k$. An illustration of such a set of non-crossing interaction energies is given in figure 2.4.

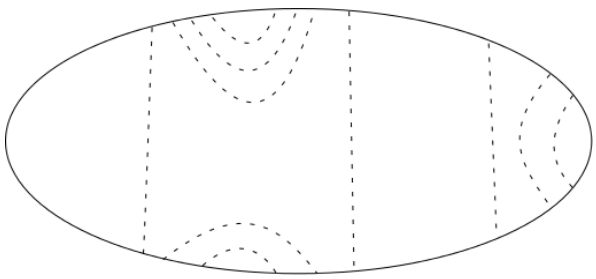


Figure 2.4: A circular polymer with only non-crossing interaction energies between sites. The pairs of sites with interaction energies between them are denoted by dashed lines.

What makes this set of interaction energies convenient, is that each possible combination of contacts between sites with interaction energies results in a configuration that is a series of closed loops. Within each of these loops, contact frequencies can be obtained exactly using equation (2.6). To calculate contact frequencies for the system as a whole, all that remains is to calculate the multiplicity of each possible combination of contacts between sites with interaction energies, and to calculate their statistical weight due to the energy gain from these contacts. We write down this calculation as follows.

Consider a set $E_{ij}, E_{kl}, ..., E_{yz}$ of sites with interaction energies between them. We can calculate the contact frequency between a pair of neutral sites $a$ and $b$ by considering all possible combinations of contacts between sites with interaction energies. For each of these combinations, we can calculate the contact probability between $a$ and $b$, given this specific combination of contacts. This conditional contact probability must then be weighted by the probability to be in a state with this specific combination of contacts between interaction sites. The expression for the contact probability between $a$ and $b$ is thus as follows:

$$\begin{aligned} P(C_{ab}|E_{ij}, E_{kl}, ..., E_{yz}) =& P(C_{ab}|C_{ij}, C_{kl}, ..., C_{yz}) \times P(C_{ij}, C_{kl}, ..., C_{yz}|E_{ij}, E_{kl}, ..., E_{yz}) \\ &+ P(C_{ab}|NC_{ij}, C_{kl}, ..., C_{yz}) \times P(NC_{ij}, C_{kl}, ..., C_{yz}|E_{ij}, E_{kl}, ..., E_{yz}) \\ &+ P(C_{ab}|C_{ij}, NC_{kl}, ..., C_{yz}) \times P(C_{ij}, NC_{kl}, ..., C_{yz}|E_{ij}, E_{kl}, ..., E_{yz}) \\ &+ \cdots \textit{(all possible combinations of contacts)} \\ &+ P(C_{ab}|NC_{ij}, NC_{kl}, ..., NC_{yz}) \times P(NC_{ij}, NC_{kl}, ..., NC_{yz}|E_{ij}, E_{kl}, ..., E_{yz}). \end{aligned} \tag{2.20}$$

As previously, $C_{ij}$ denotes a contact between sites $i$ and $j$ and $NC_{ij}$ denotes no contact between sites $i$ and $j$. We now proceed to calculating the terms on the right hand side of equation (2.20). We start with the probabilities of the form $P(C_{ij}, C_{kl}, ..., C_{yz}|E_{ij}, E_{kl}, ..., E_{yz})$, for which we first need an expression for the partition function.

### Partition function for the non-crossing system

We can construct the partition function for the non-crossing system by sequentially going over all possible combinations of contacts between sites with interaction energies. For each possible combination, we calculate its multiplicity and weigh it by the Bolzmann factor associated with this combination. We divide the entire partition function by the total multiplicity of the free polymer to reduce the size of all individual terms, which make its computation more efficient. This does not change the outcomes of any expectation values we calculate, as long as we employ this prefactor in the calculation of expectation values as well. The resulting expression for the partition function is:

$$\begin{aligned} Z = \frac{1}{M_{tot}} \Bigg[ & e^{E_{ij}} \cdot M(C_{ij}) \cdot \left(1 - \frac{M(C_{ij}, \cdots, C_{yz})}{M(C_{ij})}\right) \\ & + e^{E_{kl}} \cdot M(C_{kl}) \cdot \left(1 - \frac{M(C_{ij}, \cdots, C_{yz})}{M(C_{kl})}\right) \\ & + \cdots \\ & + e^{E_{ij}+E_{kl}} \cdot M(C_{ij}, C_{kl}) \cdot \left(1 - \frac{M(C_{ij}, \cdots, C_{yz})}{M(C_{ij}, C_{kl})}\right) \\ & + \cdots \textit{(all possible combinations of contacts)} \\ & + e^{E_{ij}+\cdots+E_{yz}} \cdot M(C_{ij}, \cdots, C_{yz}) \Bigg]. \end{aligned} \tag{2.21}$$

The factors $(1 - \frac{M(...)}{M(...)})$ in each term are derived in the same way as the correction factor discussed in section 2.2. They constitute correction factors to account for the fact that if a

specific set of contacts is considered, all the other sets of sites with interaction energies are explicitly taken not to be in contact. The numerator of $\frac{M(...)}{M(...)}$ is the multiplicity of states with all sites with interaction energies being in contact, the denominator is the multiplicity of states where the set of sites considered for that term are in contact, without imposing constraints on any other pairs of sites.

### Probabilities for sets of contacts between interaction energy sites

Using the partition function from Eq. 2.21, we can now write down expressions for the contact probabilities $P(C_{ij}, NC_{kl}, ..., C_{yz}|E_{ij}, E_{kl}, ..., E_{yz}), \cdots$. For example, the expression for $P(C_{ij}, NC_{kl}, ..., C_{yz}|E_{ij}, E_{kl}, ..., E_{yz})$ becomes

$$P(C_{ij}, NC_{kl}, ..., C_{yz}|E_{ij}, E_{kl}, ..., E_{yz}) = \frac{1}{Z} e^{E_{ij}+E_{mn}+...+E_{yz}} \cdot M(C_{ij}, C_{mn}, ..., C_{yz}) \times \\ \times \left(1 - \frac{M(C_{ij}, \cdots, C_{yz})}{M(C_{ij}, C_{mn}, ..., C_{yz})}\right). \tag{2.22}$$

To calculate these expressions, the multiplicities $M(...)$ and the conditional probabilities $P(C_{ab}|C_{ij}, C_{kl}, ..., C_{yz}), ...$ need to be determined. For a set of non-crossing interactions this can be done due to the way the polymer decomposes to a set of tethered rings that do not interact. In the following section this principle is illustrated using a few example configurations.

### Solving for contacts in a system of non-crossing interaction energies

Using the assumption of non-crossing interaction energies, we can obtain expressions for the different fractions of multiplicities included in equation (2.21). The procedure for obtaining these values is as follows.
Given a fraction $\frac{M(C_{ij},...,C_{vw})}{M(C_{kl},...,C_{rs})}$, we first let the indices of each $C$ be ordered with the smallest first and the largest second. Then, within each $M(...)$ we order the $C$'s by the first index in ascending order. After this, we compare the sequence of $C$'s in the numerator and the denominator, going from left to right. Each time we encounter a consecutive series of $C$'s that are contained in the numerator, but not in the denominator, this gives a contribution to the value of $\frac{M(C_{ij},...,C_{vw})}{M(C_{kl},...,C_{rs})}$. We illustrate the way this contribution is calculated in Figure 2.5, where a hypothetical term $\frac{M(C_1,C_2,C_3,C_4)}{M(C_1,C_3)}$ in $Z$ is depicted.

For example in Figure 2.5, we see that contacts $C_2$ and $C_4$ are not enforced in the denominator, resulting in two larger loops of lengths $L_2 + L_3$ and $L_4 + L_5$ being formed there. To calculate the fraction of multiplicities, we simply compute the combined multiplicity of the 5 loops included in the numerator, and the combined multiplicity of the 3 loops included in the denominator, and divide the two. Using equation (2.4) for each of the loops separately and

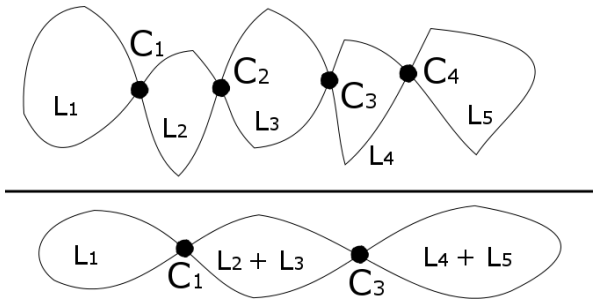



Figure 2.5: An illustration of a fraction of multiplicities that could appear in the expression for $Z$. In this case the fraction would be $\frac{M(C_1,C_2,C_3,C_4)}{M(C_1,C_3)}$

.

then multiplying the results, we obtain

$$\begin{aligned}\frac{M(C_1,C_2,C_3,C_4)}{M(C_1,C_3)} &= \frac{M(L_1)}{M(L_1)} \times \frac{M(L_2)M(L_3)}{M(L_2+L_3)} \times \frac{M(L_4)M(L_5)}{M(L_4+L_5)} \\ &= 1 \times \frac{((\frac{L_2-1}{2})!)^2((\frac{L_3-1}{2})!)^2((\frac{L_2+L_3}{2})!)^2}{\pi((\frac{L_2}{2})!)^2((\frac{L_3}{2})!)^2((\frac{L_2+L_3-1}{2})!)^2} \times \frac{((\frac{L_4-1}{2})!)^2((\frac{L_5-1}{2})!)^2((\frac{L_4+L_5}{2})!)^2}{\pi((\frac{L_4}{2})!)^2((\frac{L_5}{2})!)^2((\frac{L_4+L_5-1}{2})!)^2} \qquad (2.23) \\ &\approx \frac{1}{\pi^2}\frac{4(L_2+L_3)(L_4+L_5)}{L_2 \cdot L_3 \cdot L_4 \cdot L_5}. \qquad (2.24)\end{aligned}$$

If we now instead consider the case where contacts $C_2$ and $C_3$ are omitted from the denominator, we obtain configurations as illustrated in figure 2.6. The corresponding fraction of multiplicities are then given by

$$\begin{aligned}\frac{M(C_1,C_2,C_3,C_4)}{M(C_1,C_4)} &= \frac{M(L_1)}{M(L_1)} \times \frac{M(L_2)M(L_3)M(L_4)}{M(L_2+L_3+L_4)} \times \frac{M(L_5)}{M(L_5)} \\ &= 1 \times \frac{((\frac{L_2-1}{2})!)^2((\frac{L_3-1}{2})!)^2((\frac{L_4-1}{2})!)^2((\frac{L_2+L_3+L_4}{2})!)^2}{\pi^2((\frac{L_2}{2})!)^2((\frac{L_3}{2})!)^2((\frac{L_4}{2})!)^2((\frac{L_2+L_3+L_4-1}{2})!)^2} \times 1 \qquad (2.25) \\ &\approx \frac{1}{\pi^2}\frac{4(L_2+L_3+L_4)}{L_2 \cdot L_3 \cdot L_4}. \qquad (2.26)\end{aligned}$$

The general recipe for calculating these fractions of multiplicities follows the patterns of these two examples. Given an uninterrupted sequence $C_n, ..., C_{n'}$ in the numerator that does not appear in the denominator, the fraction of multiplicities gains approximately a factor of

$$\frac{1}{\pi^{(n'-n)}}\frac{2^{(n'-n)}(L_n + ... + L_{(n`+1)})}{L_n \times \cdots \times L_{(n`+1)}}. \qquad (2.27)$$

Using this procedure to calculate fractions of multiplicities, all terms in expression (2.21) for the partition function can be calculated. Now that we have a way to obtain $Z$, the probabilities in equation (2.20) can be calculated.

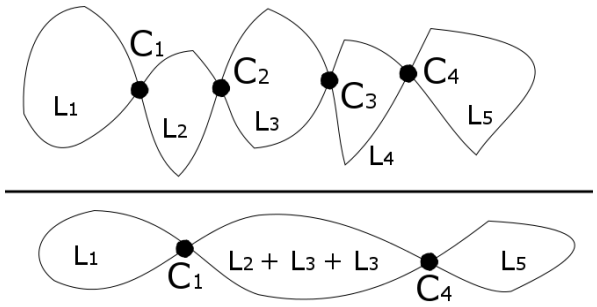



Figure 2.6: A second illustration of a fraction of multiplicities that could appear in the expression for $Z$. In this case the fraction would be $\frac{M(C_1,C_2,C_3,C_4)}{M(C_1,C_4)}$

.

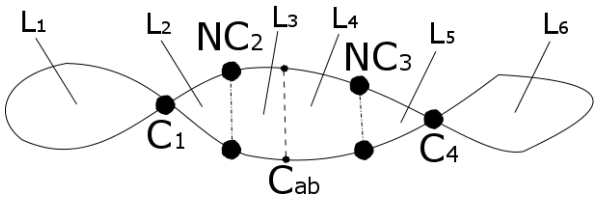



Figure 2.7: Illustration of the conditional contact probability $P(C_{ab}|C_1, NC_2, NC_3, C_4)$. Pairs of sites with dash-dotted lines between them represent sites with interaction energies that are explicitly taken to not be in contact, the two sites with dashed lines between them are the sites for which we would like to calculate the conditional contact probability. The quantities $L_1, ..., L_5$ are the lengths of the loops that would be formed if all sites with interaction energies would be in contact.

### Contact probabilities for the non-crossing system

To evaluate equation (2.20), we require expressions for all the conditional probabilities $P(C_{ab}|C_{ij}, NC_{kl}, ..., C_{yz}), \cdot$ and $P(C_{ij}, NC_{kl}, ..., C_{yz}|E_{ij}, E_{kl}, ..., E_{yz}), \cdots$. In figure 2.7 a schematic is drawn for the conditional contact probability $P(C_{ab}|C_1, NC_2, NC_3, C_4)$. To calculate this probability, we calculate the contact frequency between $a$ and $b$ given that they are part of a loop enclosed by $C_1$ and $C_4$. This is then multiplied by a correction factor to take into account that pair 2 and pair 3 are not in contact. Writing this out we obtain

$$
\begin{aligned}
&P(C_{ab}|C_1, NC_2, NC_3, C_4) = \\
&= \frac{M(C_{ab}) - M(C_{ab}, C_2, C_3) - M(C_{ab}, C_2, NC_3) - M(C_{ab}, NC_2, C_3)}{M_{tot} - M(C_2, C_3) - M(C_2, NC_3) - M(NC_2, C_3)} \\
&= \frac{M(C_{ab}) - M(C_{ab}, C_2, C_3) - M(C_{ab}, C_2) + M(C_{ab}, C_2, C_3) - M(C_{ab}, C_3) + M(C_{ab}, C_2, C_3)}{M_{tot} - M(C_2, C_3) - M(C_2) + M(C_2, C_3) - M(C_3) + M(C_2, C_3)} \\
&= \frac{M(C_{ab}) + M(C_{ab}, C_2, C_3) - M(C_{ab}) - M(C_{ab}, C_3)}{M_{tot} + M(C_2, C_3) - M(C_2) - M(C_3)} \\
&= \frac{M(C_{ab})}{M_{tot}} \times \left( \frac{1 - \frac{M(C_{ab},C_2)+M(C_{ab},C_3)-M(C_{ab},C_2,C_3)}{M(C_{ab})}}{1 - \frac{M(C_2)+M(C_3)-M(C_2,C_3)}{M_{tot}}} \right)
\end{aligned}
\tag{2.28}
$$

For each of the fractions of multiplicities included in equation (2.28) we can now obtain an expression in terms of the loop lengths, in the same way as has been done in section 2.2. This

way we obtain

$$P(C_{ab}|C_1, NC_2, NC_3, C_4) =$$
$$= \frac{2}{\pi}\frac{(L_1+L_2)}{L_1\cdot L_2} \times \left( \frac{1-\frac{2}{\pi}\left(\frac{L_1+L_2+L_3}{(L_1+L_2)\cdot L_3}+\frac{L_4+L_5+L_6}{(L_5+L_6)\cdot L_4}-\frac{2}{\pi}\frac{L_1+L_2+L_3}{(L_1+L_2)\cdot L_3}\frac{L_4+L_5+L_6}{(L_5+L_6)\cdot L_4}\right)}{1-\frac{2}{\pi}\left(\frac{L_{tot}}{(L_1+L_2)\cdot(L_3+L_4+L_5+L_6)}+\frac{L_{tot}}{(L_1+L_2+L_3+L_4)\cdot(L_5+L_6)}+\frac{2}{\pi}\frac{L_{tot}}{(L_1+L_2)\cdot(L_3+L_4)\cdot(L_5+L_6)}\right)} \right). \tag{2.29}$$

Here we have used the notation $L_{tot} = L_1 + L_2 + L_3 + L_4 + L_5 + L_6$. Analogous expressions for other combinations of contacts can be obtained using the same procedure.

# Chapter 3

# The spatial organization of a replicating bacterial chromosome, learned with a fully data-driven approach

*In collaboration with Chase P. Broedersz, Grzegorz Gradziuk, Janni Harju, Imesha R. Mudiyanselage, Muriel C.F. van Teeseling and Lucas Tröger*

## Summary

In the previous chapter, we saw how a Maximum Entropy chromosome model provides access to chromosome organization across genomic scales. This revealed a host of organizational features, from Super Domains to local extension patterns to the localization information contained by each genomic region. This model was constructed for a single, unreplicated chromosome, however in bacteria, replication is a ubiquitous chromosomal state. For a full understanding of spatial chromosome organization, a description of organization during replication is essential.

In this chapter, we expand our MaxEnt chromosome model to describe a replicating chromosome. We make use of Hi-C data sets at various times throughout the replication cycle, and modify the model phase space to match the replication progress at each time point. We find that only using Hi-C data is not sufficient to produce a model with predictive power; in this case we find a solution where the chromosomes do not segregate. In addition to the Hi-C constraints, we therefore impose the mean distance between replicated origins of replication (*ori*), as obtained via fluorescence microscopy. The choice of *ori* as the constrained region is biologically motivated; the newly replicated *ori* is known to be actively pulled across the cell, inducing chromosome segregation.

To validate the replicating MaxEnt model, we compare localizations of genomic regions across the chromosome and throughout the replication cycle between model and fluorescence microscopy. In making this comparison, we correct for a systematic bias in the experimental results due to the indistinguishability of fluorescent foci at short distance. We find that model and experiment closely match across comparisons, validating the predictive power of the replicating MaxEnt model.

Next, we investigate organizational features that are yet inaccessible to experiment. Our replicating MaxEnt model predicts a persistence of linear chromosomal organization throughout the replication cycle. A model containing only constraints on the *ori* positions, termed the *ori* pulling model, reveals that the linear organization of the replicating chromosomal segments

is largely explained by the pulling of replicated *ori*'s to opposite cell poles. The *ori* pulling model however fails to produce a linear organization within the unreplicated chromosome, which could be explained by the absence of SMC (Structural Maintenance of Chromosomes) proteins within this model, which induce a juxtaposition of chromosomal arms. Lastly, we discuss a few examples of more detailed organizational features that can be explored using the replicating MaxEnt model. The replicating MaxEnt model thus provides a principled approach to resolving spatial chromosome organization throughout the replication cycle, giving access to a wealth of organizational features and their change over time.

## 3.1 Introduction

Life requires the faithful replication of genetic material and the transfer of genomic copies to next generations. In bacteria, this genetic material is stored in a single circular chromosome, which is compacted by several orders of magnitude to fit inside its cellular confinement. During the replication process, the highly compacted chromosome must continue to facilitate transcription, replication, and segregation, and finally be faithfully passed on to daughter cells. The spatial organization of the chromosome during the replication process remains unclear; resolving this requires a characterization of the full distribution of single-cell chromosome configurations across the replication cycle, posing a major challenge for experiment and theory.

The phase of bacterial chromosome replication, termed the C period, is one of the most ubiquitous cell cycle phases under nutrient-rich conditions. In fact, for these conditions *Escherichia coli* and *Bacillus subtilis* are found to continuously replicate, with multifork replication allowing for a mass doubling time shorter than the chromosomal replication time [130]. In *Caulobacter crescentus*, multifork replication is not observed [131], however initiation of replication is observed shortly after synchronization of newborn swarmer cells [132]. Thus, for a full understanding of bacterial chromosome organization, a characterization of chromosome states during replication is essential.

Here, we study spatial chromosome organization throughout the replication cycle in *C. crescentus*. For this bacterium, newborn swarmer cells, which contain a single, unreplicated chromosome, initiate replication after the formation of a stalk at the cell pole [131]. At the onset of replication, two replication forks simultaneously move from the origin of replication (*ori*) to the terminus (*ter*), with one fork moving along each chromosomal arm [133]. Chromosomal segregation occurs in parallel with replication, where the newly replicated *ori* is actively transported via the ParB*S* system from the pole where the unreplicated *ori* is initially tethered to the opposite cell pole [51, 134, 135]. Division is regulated to be initiated after replication of the two chromosomes is completed [136].

In probing spatial chromosome organization in bacteria, a key experimental technique is Hi-C chromosome conformation capture [58, 59, 63–66], which yields average contact frequencies between pairs of genomic regions across the chromosome. This provides a wealth of information on organizational features, however, extracting the full distribution of three-dimensional chromosome configurations from this data is challenging. For non-replicating chromosomes, several approaches have been developed to perform such an inference, either by converting Hi-C scores to average distances [71–73], employing an equilibrium polymer model with pairwise interaction energies [74–76], or generating an ensemble of chromosome states that is consistent with experimental contact frequencies [77]. In [69] we developed a principled, fully data-driven Maximum Entropy approach to derive the full distribution of

chromosome configurations directly from Hi-C data, providing access to single-cell organizational features across genomic scales. This approach does not require any assumptions on a Hi-C score distance relation, is compatible with a chromosome out of thermal equilibrium, and is designed to find the least-structured ensemble that is consistent with experimental constraints.

The interpretation of Hi-C data for a replicating population of cells in terms of a replicating chromosome model is relatively unexplored terrain. In [137], consensus chromosome structures at different replication stages were constructed for *E. coli* using mixed-population Hi-C data containing cells at all replication stages. In this approach, each replication stage was learned separately on the same input data, and an assumed relation between Hi-C scores an average distances was employed. Furthermore, and assumption was made that Hi-C scores are dominated by interchromosomal contacts within each of the replicated segements. It is however unclear if Hi-C data from a mixed population is suitable as an input for a single replication state model. Furthermore, the assumed relation between Hi-C scores and average distances contains several weakening assumptions [69], and it is also unclear if interchromosomal contacts indeed dominate Hi-C scores. Therefore, a replicating chromosome model learned in a principled way, taking the full distribution of chromosome configurations into account is still lacking.

To elucidate spatial chromosomal organization during the replication process, we develop a fully data-driven Maximum Entropy model for a replicating *C. crescentus* chromosome. We build upon our approach developed for an unreplicated chromosome in the previous chapter, where we expand the model phase space to capture replication progression, and employ constraints from previously published Hi-C data [58] and fluorescence microscopy for a series of time points along the cell cycle. Our model constitutes a principled approach to derive the full distribution of chromosome configurations across the replication cycle, yielding insight into organizational features over space and time.

## 3.2 Results

### 3.2.1 Learning the Maximum Entropy model of a replicating chromosome

Here we develop a Maximum Entropy (MaxEnt) model for the spatial and temporal organization of a chromosome progressing through the replication cycle, constrained by chromosome capture experiments and live cell microscopy data. MaxEnt models have been used in a variety of biological contexts [74, 75, 81–85] to obtain the least-structured statistical model consistent with data. Recently [69], we developed a MaxEnt model for the spatial organization of a single, non-replicating bacterial chromosome. This provides a principled approach for inferring the statistics of chromosome structure in bacteria directly from experimental data. Here we expand this approach to describe a replicating chromosome in a growing cell. The model constraints consist of Hi-C data collected at various discrete time points during the cell cycle, as well as chromosomal localization data from live cell microscopy. The key idea of our generalized model is to apply these stroboscopic constraints to a MaxEnt chromosome model where the model phase space, determined by the cell size and progression of chromosomal replication, is constructed to match inferred cellular properties at each time point.

The derivation of the MaxEnt model of chromosome conformations corresponds to finding the probability distribution $P(\{\mathbf{r}\}, t)$ of chromosome states $\{\mathbf{r}\}$ at cell cycle progression time

$t$ that maximizes the Shannon entropy, given by

$$S(t) = -\sum_{\{\mathbf{r}\}} P(\{\mathbf{r}\}, t) \ln P(\{\mathbf{r}\}, t), \tag{3.1}$$

while satisfying time-dependent experimental constraints. This constrained maximization ensures that the least-structured model is found that is consistent with experimental data.

To construct a MaxEnt model for a replicating chromosome, we first define the set of allowed chromosome states $\{\mathbf{r}\}$. We employ the coarse-grained representation of the chromosome as a chain on a 3D cubic lattice within a cell-shaped confinement as described in [69], where the coarse-graining scale is set by the resolution of the Hi-C experiment. Within this description, a subset of $N$ monomers evenly spread along the chain represents the locations of the genomic regions corresponding to each Hi-C bin. Two genomic regions have a contact probability $\gamma$ if they occupy the same lattice site, in the sense that the proximity of the regions in space would lead to a count in a Hi-C experiment for this pair, and 0 otherwise. To capture the DNA replication process, we now generalize this description by including two replication forks, at positions $M(t)$ and $N - M(t)$. The replication forks are connected by three chromosomal segments: one segment of length $N - 2M(t)$ representing the unreplicated portion of the chromosome, and two segments of length $2M(t)$ representing the two identical copies of the replicated chromosome. Within this representation, a microstate $\sigma = \{\mathbf{r}_1, \mathbf{r}_2, ..., \mathbf{r}_M, \mathbf{r}_{M+1}, \mathbf{r'}_{M+1}, ..., \mathbf{r}_{N-M-1}, \mathbf{r'}_{N-M-1}, \mathbf{r}_{N-M}, ..., \mathbf{r}_N\} = \{\mathbf{r}\}$ is defined by the monomer positions $\mathbf{r}_i$ for the unreplicated portion of the chromosome, and the pairs of monomer positions $\mathbf{r}_j$,$\mathbf{r'}_j$ for the two copies of each genomic region on the replicated chromosome. To reflect cellular growth over the course of replication, the size of the cell-shaped confinement is made dependent on the cell cycle progression (SI 2.3).

For the MaxEnt model for an unreplicated chromosome at $t = 0$, Eq. 3.1 was maximized while enforcing the model contact frequencies to match experimental contact frequencies for each monomer pair. For a replicating chromosome however, the cell cycle dependent Hi-C data $f_{ij}^{\text{expt}}(t)$ does not distinguish between identical regions on each of the replicated chromosome copies. This forms one of the biggest challenges in interpreting Hi-C data on replicating chromosomes. To reflect this agnosticism intrinsic to the Hi-C data, we define the model contact frequencies $f_{ij}^{\text{model}}(t)$ as the combined frequency of the possible interchromosomal contacts $(\mathbf{r_i}, \mathbf{r_j})$, $(\mathbf{r'_i}, \mathbf{r'_j})$ and intrachromosomal contacts $(\mathbf{r'_i}, \mathbf{r_j})$, $(\mathbf{r_i}, \mathbf{r'_j})$. The first set of constraints we enforce on the replicating MaxEnt model is thus as follows:

$$\sum_{\{\mathbf{r}\}} P(\{\mathbf{r}\}, t)\gamma(\delta_{\mathbf{r}_i,\mathbf{r}_j} + \delta_{\mathbf{r'}_i,\mathbf{r}_j} + \delta_{\mathbf{r}_i,\mathbf{r'}_j} + \delta_{\mathbf{r'}_i,\mathbf{r'}_j}) \stackrel{!}{=} f_{ij}^{\text{expt}}(t), \tag{3.2}$$

for each $i, j$, where $\delta_{\mathbf{r}_i,\mathbf{r}_j}$ is the kronecker delta. If a genomic region $j$ is unreplicated, $\delta_{\mathbf{r}_i,\mathbf{r'}_j}$ and $\delta_{\mathbf{r'}_i,\mathbf{r'}_j}$ are equal to zero.

Imposing only Hi-C constraints turns out to be insufficient to obtain a MaxEnt model with predictive power: for this scenario, we find a solution where the replicated chromosome does not segregate (Appendix 3.4.4). This implies that the Hi-C data alone does not contain enough information to sufficiently constrain a replicating chromosome model, which is likely connected to the indistinguishability of inter- and intrachromosomal contacts within experiment. In addition to the Hi-C constraints, we therefore apply a positional constraint: the mean separation between replicated origin of replication (*ori*) regions should match experiment. The choice for *ori* as a constrained region is biologically motivated: in *C. crescentus*,

the *ori* is tethered to the cell pole [138, 139], and after replication the newly replicated ori is actively pulled to the opposite cell pole via the ParB$S$ system, driving chromosome segregation [51, 134, 135]. This constraint is therefore a promising candidate to induce segregation within the MaxEnt model. We thus impose:

$$\sum_{\{\mathbf{r}\}} P(\{\mathbf{r}\},t) d_{\mathrm{ori}}(\{\mathbf{r}\}) \stackrel{!}{=} \langle d_{\mathrm{ori}}^{\mathrm{expt}} \rangle(t), \tag{3.3}$$

where $d_{\mathrm{ori}}(\{\mathbf{r}\})$ is the distance between the two *ori* copies projected along the long cell axis, and $\langle d_{\mathrm{ori}}^{\mathrm{expt}} \rangle(t)$ is the experimentally measured mean long-axis *ori* distance.

To maximize Eq. 3.1 under constraints 3.3 and 3.2, we introduce the functional $\tilde{S}(t)$, with one Lagrange multiplier $\lambda_{ij}(t)$ for each Hi-C constraint, one Lagrange multiplier $\alpha(t)$ enforcing the separation between *ori*'s, and $\lambda_0(t)$ ensuring normalization:

$$\begin{aligned}\tilde{S}(t) = &- \sum_{\{\mathbf{r}\}} P(\{\mathbf{r}\},t) \ln P(\{\mathbf{r}\},t) - \sum_{ij} \lambda_{ij}(t) \Bigg( \sum_{\{\mathbf{r}\}} P(\{\mathbf{r}\},t) \gamma (\delta_{\mathbf{r}_i,\mathbf{r}_j} + \delta_{\mathbf{r'}_i,\mathbf{r}_j} + \delta_{\mathbf{r}_i,\mathbf{r'}_j} + \delta_{\mathbf{r'}_i,\mathbf{r'}_j}) - f_{ij}^{\mathrm{expt}}(t) \Bigg) \\ &- \alpha(t) \Bigg( \sum_{\{\mathbf{r}\}} P(\{\mathbf{r}\},t) d_{\mathrm{ori}}(\{\mathbf{r}\}) - \langle d_{\mathrm{ori}}^{\mathrm{expt}} \rangle(t) \Bigg) - \lambda_0(t) \Bigg( \sum_{\{\mathbf{r}\}} P(\{\mathbf{r}\},t) - 1 \Bigg).\end{aligned} \tag{3.4}$$

Setting $\frac{\partial \tilde{S}}{\partial P(\{\mathbf{r}\},t)} \stackrel{!}{=} 0$ we obtain:

$$P(\{\mathbf{r}\},t) = \frac{1}{Z} \exp \Big[ - \sum_{ij} \lambda_{ij}(t) \gamma (\delta_{\mathbf{r}_i,\mathbf{r}_j} + \delta_{\mathbf{r'}_i,\mathbf{r}_j} + \delta_{\mathbf{r}_i,\mathbf{r'}_j} + \delta_{\mathbf{r'}_i,\mathbf{r'}_j}) - \alpha(t) d_{\mathrm{ori}}(\{\mathbf{r}\}) \Big], \tag{3.5}$$

where $Z = \exp[1 + \lambda_0]$. This gives us the form of the probability distribution $P(\{\mathbf{r}\},t)$, where the values of the $\lambda_{ij}(t)$ and $\alpha(t)$ are determined by imposing constraints 3.3 and 3.2. Analogously to [69], a solution to this large set of highly nonlinear equations can be found by mapping Eq. 3.5 to an equilibrium polymer model. This equilibrium polymer model contains pairwise interaction energies $\epsilon_{ij}(t) = \gamma \lambda_{ij}(t)$, and a separation energy $\epsilon_{\mathrm{sep}}(t) = \alpha(t)$ that couples to the distance between replicated *ori*'s. This equilibrium polymer model is numerically solved using iterative Monte Carlo simulations for each time $t$ as in [69], with a modification to representatively sample replicating chromosome configurations (Appendix 3.4.3).

### 3.2.2 Applying the replicating MaxEnt model to Hi-C data on replicating *Caulobacter crescentus* cells

We apply our replicating MaxEnt model to experimental Hi-C data sets on replicating *Caulobacter crescentus* cells from [58]. Here, cell populations were synchronized to isolate swarmer cells, which initially contain a single unreplicated chromosome. Subsequently, at regular intervals after synchronization, Hi-C data was generated for this population of developing cells. Thus, each generated Hi-C data set contains cells at a similar cell cycle stage and replication fork progression.

Before our MaxEnt method can be applied to these data, known systematic biases in the raw data need to be corrected for, due to for instance the proportionality between the number of restriction sites in a genomic region and its Hi–C score [70]. For swarmer cells,

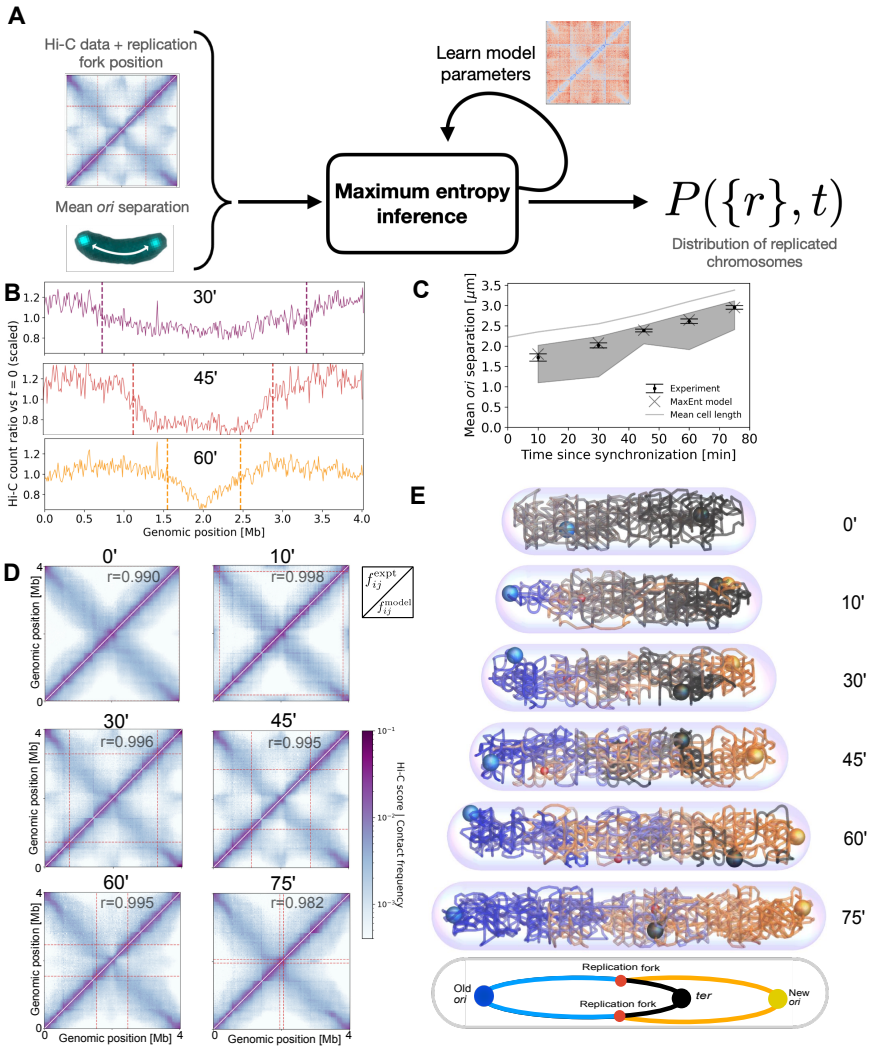



Figure 3.1: **Maximum entropy model inferred for a replicating chromosome in *C. crescentus*.** **A** Schematic of the MaxEnt inference procedure. **B** Ratios of total raw Hi-C counts per genomic region compared to $t = 0$ for three cell cycle times. The resulting profile is used to estimate the mean replication fork position (Appendix 3.4.2), indicated by the dashed vertical lines for each time point. **C** Black dots: measured mean distances between the two copies of *ori* obtained for each time since synchronization. Error bars indicate the error on the mean. Shaded areas: measured upper and lower standard deviations. Crosses: mean *ori* distances for the converged MaxEnt model. Grey line: mean cell length at each cell cycle time (Appendix 3.4.2). **D** Comparison between bias-corrected experimental contact frequencies $f_{ij}^{\text{expt}}$ (upper left corners) (Appendix 3.4.2), based on Hi-C data from [140]) and contact frequencies obtained from our inferred MaxEnt model $f_{ij}^{\text{model}}$ (lower right corners) for six replication cycle times. Inferred mean replication fork positions (Appendix 3.4.2) are indicated by the dashed lines. **E** Sample configurations of a replicating chromosome, shown for 6 times since synchronization. Blue sphere: old *ori*, yellow sphere: new *ori*, black sphere: *ter*, smaller red spheres: replication fork positions. The old and new chromosomes are determined as in Fig. 3.2.

such a correction is performed by normalization of the Hi-C map. This normalization assumes that the total contact frequency per genomic region is the same for all regions. This implies that

$$\sum_i f_{ij}^{\text{expt}}(0) = c \tag{3.6}$$

for each column $j$. For later cell cycle stages however, such normalization is not expected

to yield accurate contact frequencies, since the two copies of replicated regions together are expected to generate more contacts on average than single unreplicated regions.

Here we develop a bias correction procedure for later cell cycle stages, by using the Hi-C data for swarmer cells as a benchmark. Assuming that the systematic biases per genomic region are conserved throughout replication, we can rescale the total raw Hi-C count of each genomic region at a time $t$ by the raw count measured for swarmer cells. Specifically, we assume that the raw contact counts at time $t$ can be written as

$$f_{ij}^{\mathrm{raw}}(t) = b_i b_j f_{ij}^{\mathrm{expt}}(t) N_{\mathrm{cells}}(t), \tag{3.7}$$

with $b_i \in [0,1]$ the crosslinking bias of site $i$, and $N_{\mathrm{cells}}(t)$ the number of cells included in the measurement at time $t$. The sum over column $i$ in the raw data we can now approximate as

$$\sum_i f_{ij}^{\mathrm{raw}}(t) \approx b_j N_{\mathrm{cells}}(t) b_{\mathrm{av}} \sum_i f_{ij}^{\mathrm{expt}}(t), \tag{3.8}$$

where we assume that the crosslinking bias $b_i$ is uncorrelated with $f_{ij}^{\mathrm{expt}}(t)$, and $b_{\mathrm{av}}$ represents the average value of $b_i$ over all Hi-C bins. Using Eq.3.8 together with Eq.3.6, we can write the column sums of $f_{ij}^{\mathrm{expt}}(t)$ as

$$\sum_i f_{ij}^{\mathrm{expt}}(t) = c(t) \frac{\sum_i f_{ij}^{\mathrm{raw}}(t)}{\sum_i f_{ij}^{\mathrm{raw}}(0)}, \tag{3.9}$$

where we identify $c(t) = c\frac{N_{\mathrm{cells}}(0)}{N_{\mathrm{cells}}(t)}$. Thus, we obtain the column sums of the unbiased contact frequencies $f_{ij}^{\mathrm{expt}}(t)$ up to an overall scale factor $c(t)$ for each time point. From these column sums we construct input contact frequencies $\tilde{f}_{ij}^{\mathrm{expt}}(t)$ (Appendix 3.4.2), which are related to the $f_{ij}^{\mathrm{expt}}(t)$ via $\tilde{f}_{ij}^{\mathrm{expt}}(t) = \frac{1}{c(t)} f_{ij}^{\mathrm{expt}}(t)$.

To deal with the unknown scale factor $c(t)$, we employ the approach from [69] and treat $c(t)$ as an unknown parameter in the MaxEnt model. Absorbing $\gamma$ into $c(t)$, and setting $\tilde{c}(t) = \frac{c(t)}{\gamma}$, we extremize Eq. 3.4 with respect to $\tilde{c}(t)$, to obtain

$$\sum_{ij} \epsilon_{ij}(t) \tilde{f}_{ij}^{\mathrm{expt}}(t) = 0, \tag{3.10}$$

which is analogous to the condition for an unreplicated chromosome.

To construct the model phase space, for each input Hi-C map $\tilde{f}_{ij}^{\mathrm{expt}}(t)$ the mean replication fork position $M$ needs to be determined. We do this directly from the Hi-C data, by using variations in the column sums $\sum_i \tilde{f}_{ij}^{\mathrm{expt}}(t)$ of the bias-corrected data. For times around the middle of the cell cycle, we observe these column sums to form a plateau around *ori*, with a transition to a plateau at a lower value around *ter* (Fig. 3.1B). The inflection point of this transition is taken as the mean replication fork position (Appendix 3.4.2).

The constraints on the mean *ori* distance $\langle d_{\mathrm{ori}}^{\mathrm{expt}} \rangle$ for each replication stage we obtain via fluorescence microscopy. Following the procedure described in [58], we synchronized *C. crescentus* swarmer cells, which initially contain a single, unreplicated chromosome. Subsequently, we measured the mean distance between two fluorescently labelled *ori* loci at each cell cycle time for which a Hi-C data set was generated in [58] (Fig.3.1C, (Appendix 3.4.1). The obtained mean distances for each time point are used as model constraints.

With this model construction, we learn the replicating MaxEnt model for the six available Hi-C data sets from [58], which are evenly spread along the replication cycle. In all cases, the MaxEnt model converges to the input contact frequencies with high accuracy, reaching a Pearson's correlation coefficient of at least 0.98 in all cases (Fig. 3.1D). Through this model, we have access to the full distribution of single-cell chromosome configurations at each measurement time $t$, a sample of which is shown in Fig. 3.1E.

### 3.2.3 The replicating MaxEnt model closely matches chromosomal localization measurements

To validate our replicating MaxEnt model, we employ fluorescent microscopy to measure the time-dependent cellular localizations of eight genomic loci positioned roughly evenly along the chromosome. For each Hi-C time point, we thus obtain an ensemble of locus positions projected along the long cell axis. Before comparing localization statistics between model and experiment, a few analysis steps are performed: Firstly, to quantify mean long-axis localizations, an orientation of the cells needs to be chosen; the two cell poles are identical within the model, and are often indistinguishable in experiment. This orientation we do as follows. If a cell contains a single locus, the mean distance of this locus to the nearest cell pole is calculated. If the cell contains two loci, a distinction is made between the near and far chromosomal locus. The near locus sits closest to a cell pole, and the corresponding pole is termed the near pole. For both the near and the far locus, the mean distance to the near pole is calculated. An illustration of this orientation procedure is shown in Fig. 3.2A.

This orientation procedure thus yields separate statistics for cells containing either one or two copies of a chromosomal locus. The MaxEnt model contains only one of these categories at each time point, however in experiment a mix of focus counts is often observed (Appendix 3.4.5). To compare experimental localizations to the MaxEnt model despite these observed mixed populations, we compute conditional averages. For loci where the calculated mean replication fork position has not passed yet, an average is computed over all measured cells with one fluorescent focus. Conversely, if the mean replication fork position is calculated to have passed a tagged locus, an average is computed over all measured cells containing two foci.

A direct comparison between experimental and model locus counts is complicated by several factors: 1) imperfect synchronization of cell cycles, 2) imperfect labelling of tagged loci, and 3) indistinguishability of two loci if they are too close together on the experimental image (Appendix 3.4.5). The indistinguishability of fluorescent foci at short distance introduces a systematic bias in experimentally determined localizations, as short distances are removed from the statistics of cells containing two foci. To directly compare MaxEnt localizations to experiment despite this bias, we compute a bias-corrected MaxEnt prediction that emulates this indistinguishability at short distance. The cutoff distance for the onset of indistinguishability we obtain directly from experimental distance statistics between foci (Appendix 3.4.5).

With these procedures for cellular orientation, conditional averaging on the number of foci, and bias correction in the MaxEnt localization prediction, we compare localization profiles between model and experiment (Fig. 3.2). We find a close match for mean positions as well as standard deviations across genomic positions and across the replication cycle. The largest deviation between model and experiment is seen for the tagged locus at 134°after replication (top right Fig. 3.2). Surprisingly, the measured distance between these loci deviates from the general trend of decreasing average distance with decreased distance from *ter*. This deviation

from the overall trend could indicate that the localization of this site is affected by interactions with membrane-bound proteins; such direct interactions of chromosomal loci with membrane-bound proteins have been shown in *Escherichia coli* in [141]. Taken together, these results show that independent microscopy experiments provide strong validation for the predictive power of the replicating MaxEnt model.

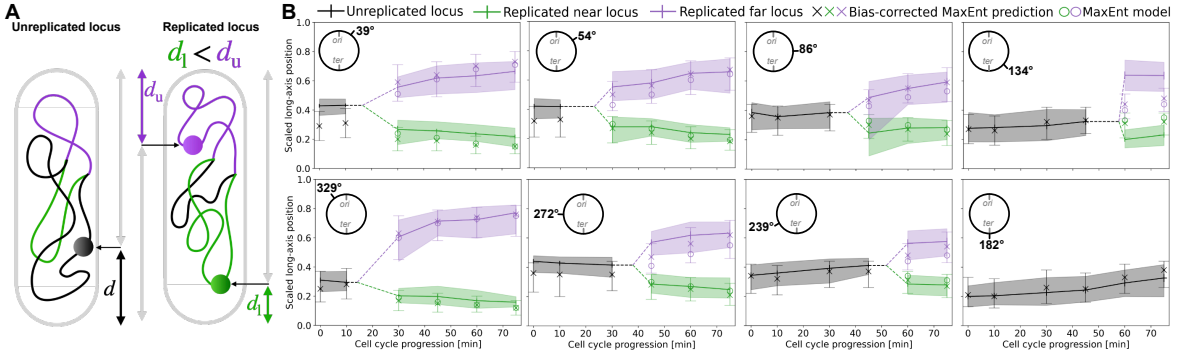



Figure 3.2: **Validation of MaxEnt model based on genomic localizations from microscopy data. A** Illustration of the orientation procedure used to compare localizations of genomic regions between measurement and MaxEnt model. Left: If the mean replication fork position has not passed the genomic region yet, the mean distance of the unreplicated region (black) to the nearest pole is calculated. In this case, only measured cells containing one focus are included. Right: If the mean replication fork position has passed the genomic region, the copy closest to a cell pole is identified as the 'near locus' (green), and the pole it is closest to is termed the 'near pole'. The other region is then termed the 'far locus' (purple). The distance of both loci to the near pole is calculated. In this case, only measured cells containing two foci are included. **B** Comparison of localizations between measurement (solid lines), bias-corrected MaxEnt prediction (crosses) and MaxEnt model (circles) for 7 chromosomal regions, with the chromosomal location indicated in the top left corner. For the bias-corrected MaxEnt prediction, a correction is made to account for the indistinguishability of fluorescent foci at short distance in experiment (Appendix 3.4.5). Vertical lines: measurement error on the mean. Shaded ares: measured standard deviations. Black bars: standard deviations from bias-corrected MaxEnt model. The meeting point of the dashed lines corresponds to the mean time at which a region is replicated (Appendix 3.4.2).

### 3.2.4 The replicating MaxEnt model yields full chromosomal localization profiles across replication stages

With our replicating MaxEnt model, we now have access to the full distribution of chromosome configurations across the replication cycle. To gain insight into its predictions on organization, we first consider the localization along the long cell axis of the entire chromosome. In contrast to the single-locus localizations of the previous section, this gives us insight into the simultaneous dynamic organization of the chromosome as a whole.

To characterize spatial chromsome organization of the replicating chromosome, we first note that the two copies of replicated chromosomal segments are treated identically within the MaxEnt model. Despite this equivalence, we can introduce a distinction between replicated segments by categorizing them based on the locations of genomic regions on these segments. We perform this categorization based on known organizational features of the old and new chromosomal regions. Fluorescent microscopy experiments have shown that the newly replicated *ori* migrates from the pole at which it is initially tethered to the opposite cell pole, where the terminus initially resides [142]. Upon completion of replication, the old chromosome typically extends over a longer distance than the new chromosome, in preparation for asymmetric division where the new chromosome is passed on to the smaller cell [132]. To distinguish between replicated chromosome copies in the MaxEnt model in a biologically

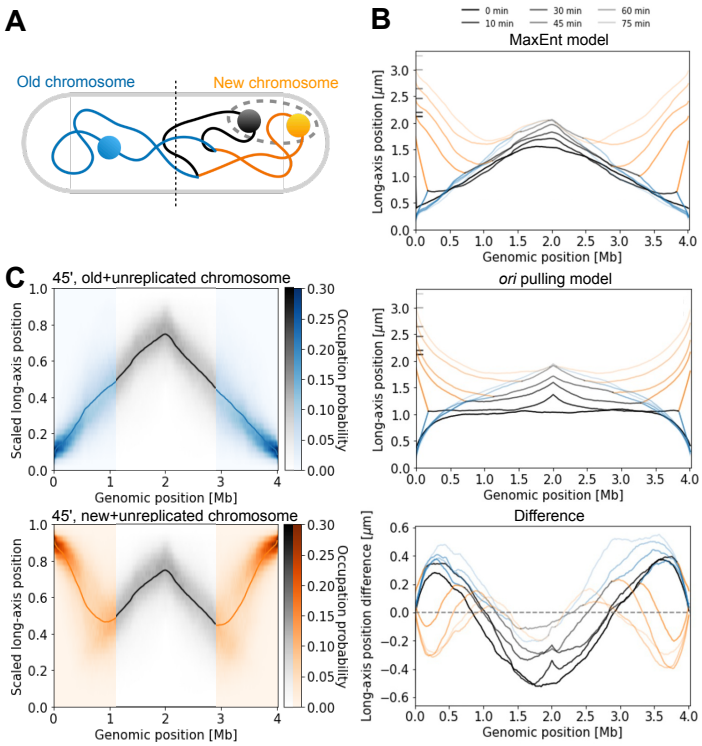



Figure 3.3: **Long-axis localization predicted by the replicating MaxEnt model A** The two replicated chromosome segments are divided into the new and old chromosome. The *ori* on the new chromosome shares its cell half with *ter*, whereas the *ori* on the old chromosome does not. In case both *ori*'s are in the same cell half, the orientation of Fig. 3.2A is used, where the near *ori* is identified als the old *ori*, and the far *ori* as the new *ori*. **B** Top: mean long-axis positions of the old (blue), new(orange) and unreplicated (black) chromosomes for each time since synchronization, indicated by the line transparencies. The grey lines at the top left indicate the mean cell length at each time point, with the transparancy again indicating the time since synchronization. Middle: same as top panel, but for a model where only the mean distance between replicated *ori*'s is used as a constraint. Bottom: the difference in mean long-axis positions between the MaxEnt model and the *ori* pulling model. **C** The full distribution of long-axis positions for the MaxEnt model at 45 minutes after synchronization. The occupation probability is indicated by the color intensity. The solid lines indicate the mean long-axis positions.

meaningful way, we categorize them based on similar features. We define the old *ori* as the *ori* that lies in the opposite cell half to *ter*, and define the new *ori* as the *ori* that lies in the same cell half as *ter*. In the rare case where both *ori*'s lie in the same cell half, the orientation of Fig. 3.2 is used, where we identify the old *ori* with the near locus, and the new *ori* with the far locus. Subsequently, we term the replicated chromosome segment that is connected to the old *ori* the old chromosome, and the chromosome segment connected to the new *ori* the new chromosome. An illustration of this orienting method is shown in Fig. 3.3A.

Computing chromosomal localizations using the old/new categorization scheme, we find several striking organizational features. The combined old and unreplicated chromosome maintains a linear organization throughout the entire cell cycle, with the mean distance between *ori* and *ter* increasing slightly as the cell grows. By contrast, the new and unreplicated chromosome combined exhibit a reversed linear organization at the replication fork position, with a colocalization of a linearly organized unreplicated region and a linearly organized new chromosome. Furthermore, for earlier time points the new chromosome extends over a longer stretch of cell length than the old chromosome. The MaxEnt model also gives us access to the full distribution of long-axis positions for each genomic locus, which we find to be tightly

localized around the mean values (Fig. 3.3C).

These findings are consistent with and expand upon aspects of replicating chromosome organization reported previously for *C. crescentus*. In [142], fluorescence microscopy revealed that the two copies of ori-proximal regions each maintain a linear organization at their respective cell poles shortly after being replicated. Our results suggest that this linear organization is maintained over the entire cell cycle throughout the replicated chromosomal segments, and is also preserved for the unreplicated segment of the chromosome. In [133], the two replication forks were found to closely colocalize, and gradually progresses towards midcell during replication, which is consistent with our localization results (Fig. 3.3B). Based on these findings, a model was proposed where the newly replicated DNA clusters at its cell pole and excludes the unreplicated chromosome, moving the terminus closer to midcell. Our results however suggest that the new chromosome is evenly spread along the long cell axis region it occupies, and partially shares cell space with the unreplicated chromosome. We do find that the terminus moves closer to midcell during replication, but predict this due to the combined old and unreplicated chromosome approximately maintaining its dimensions while the cell grows.

To understand to what extent our observed localization profiles are a physical consequence of the segregation force that pulls the two *ori*'s apart, we learned a model where only the mean separation between replicated *ori*'s is enforced, termed the *ori* pulling model. For the *ori* pulling model, we employ the same coarse-graining of the chromosome as for the replicating MaxEnt model, resulting in a model that is consistent with measured chromosome compaction at the Hi-C bin length scale [69], but is not subject to any constraints at larger length scales. We find that for this model, the linear organization of the replicated regions is maintained to some extent, although the mean long-axis positions decay to midcell faster (Fig.3.3B). By contrast, the unreplicated region shows a strong deviation between the two models, with a linear organization absent, especially for earlier cell cycle times. This suggests that the pulling of *ori* explains the localization profile of the replicated regions to a large extent, but is insufficient to explain the linear organization of the unreplicated chromosome segment. One explanation could be the absence of the loop extrusion motors SMC (Structural Maintenance of Chromosomes) [36, 39, 41, 143] in the *ori* pulling model, which induce a juxtaposed arrangement of chromosomal arms and may play a role in amplifying linear organization.

## 3.3 Outlook

With the replicating MaxEnt model, we can start exploring a wealth of organizational features throughout the replication cycle that are inaccessible to experiment. One example of such a feature is the pattern of inter- and intrachromsomal interactions between each of the replicated chromosome segments. Whereas in experimental Hi-C maps no distinction can be made between copies of replicated regions, within the MaxEnt model we can analyse each chromosomal segment individually. Furthermore, with experimental Hi-C data we cannot compare interactions over time, since each Hi-C map is subject to an unknown overall scale factor $c(t)$ (Eq. 3.9) With the replicating MaxEnt model however, we can directly compare the magnitude of interactions, since for each time point we obtain interactions up to a geometrical factor $\gamma$, which is identical for all time points.

As an illustration of such an analysis, we compute intrachromosomal contact maps for both the old and the new chromosome, as well as the intrachromosomal contact map between

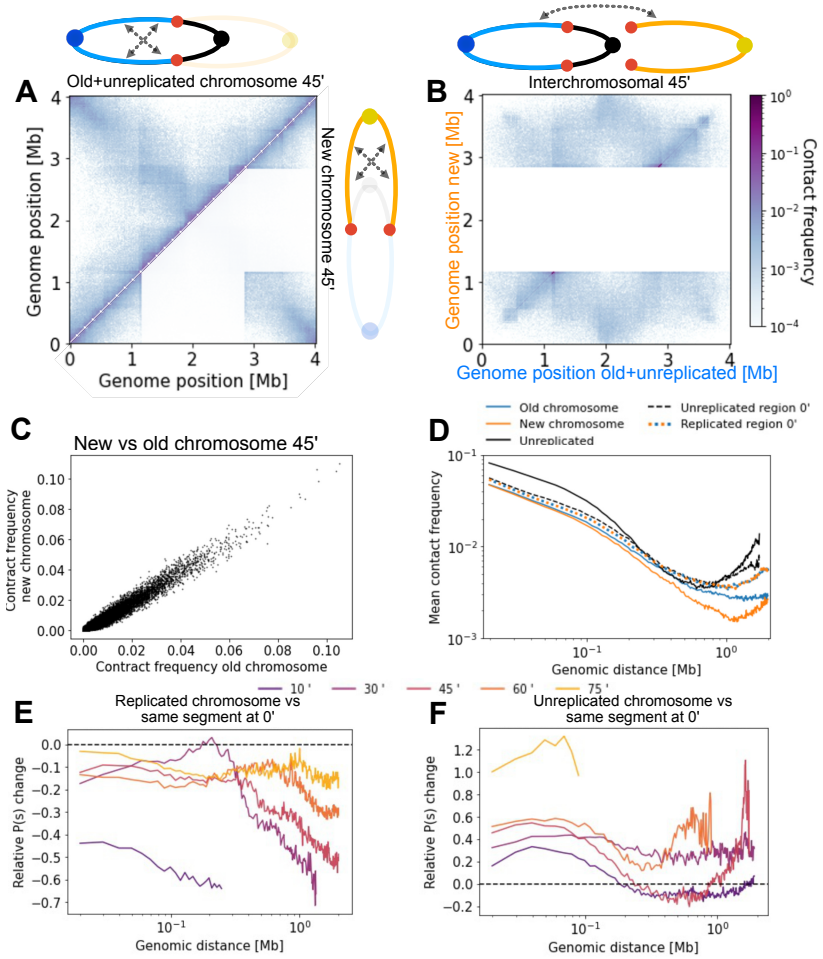



Figure 3.4: **Probing chromosomal interaction patterns with the MaxEnt model.** Results on panels A-D are shown for the MaxEnt model 45 minutes after replication. **A** Top left: model contact frequencies within the combined old and unreplicated chromosome. Bottom right: model contact frequencies within the new chromosome. **B** Model contact frequencies between (1) the combined old and replicated chromosome (horizontal axis) and (2) the new chromosome (vertical axis). **C** Direct comparison between contact frequencies on the old chromosome and the new chromosome. **D** Mean contact frequency $P(s)$ as a function of genomic distance $s$, shown for the old chromosome (blue), new chromosome (orange), unreplicated chromosome (black), as well as the mean contact frequency of the same regions on the unreplicated chromosome (blue&orange dashes for the replicated region, black dashes for the unreplicated region). **E** Relative change of P(s) on the replicated chromosome compared to the same segment at 0 minutes. **F** Relative change of P(s) on the unreplicated chromosome compared to the same segment at 0 minutes.

them (Fig. 3.4A). We find the interchromosomal contact maps to be highly correlated between replicated segments (Fig. 3.4C): we obtain a Person's r of at least 0.9 for each time point. For each of these contact maps, we also obtain the average contact frequency as a function of genomic distance, known as the $P(s)$ curve (Fig. 3.4D). Comparing P(s) curves between the replicated chromosomal segments and the same segments for $t = 0$, we find that the interactions within the replicated chromosome are systematically decreased (Fig. 3.4E). This effect is strongest directly after replication, and slowly moves back to the unreplicated chromosome values for later times. Comparing contact frequencies of the unreplicated chromosomal segment over time, we find that they systematically increase, with the strongest increase seen for the latest time point (Fig. 3.4F). This could suggest a compactification of the unreplicated chromosome as the replication fork progresses, and an opening up of the

chromosome after replication.

These are just a few features we can explore with the replicating MaxEnt model, but there are many more. Our model provides access to patterns of local extension and compaction of the chromosome, changes in chromosomal clustering and the Super Domain properties, and allows the study of chromosomal correlations and higher-order structure over time. Thus, the replicating MaxEnt model offers a window into the intricacies of chromosome organization over the replication cycle, with many further properties still waiting to be explored.

**Acknowledgements:** We thank the labs of Lucy Shapiro and Patrick Viollier for generous sharing of *C. crescentus* strains.

## 3.4 Appendix

### 3.4.1 Experimental procedures on *C. crescentus* cells

#### Bacterial strains and growth conditions

The *C. crescentus* strains used in this study (Table 3.1) were derived from the synchronizable wild-type CB15N (NA1000). Cells were grown in peptone-yeast extract (PYE) medium (Pointdexter, 1964) at 28°C under aerobic conditions (shaking at 190 rpm).

#### Synchronization of *C. crescentus* cultures

In order to analyze *C. crescentus* cells in a specific phase of their cell cycle, corresponding to a specific stage in their replication and segregation process, we synchronized the cells according to the protocol established in [144]. In brief, *C. crescentus* cells were grown to early exponential phase (OD ∼0.1) in PYE and induced for 2h with 2 µM xylose in order to express YFP and CFP that bind their respective arrays at specific chromosomal loci (see Table 3.1). Afterwards, cells were pelleted, resuspended in M2 salts buffer [145] and mixed 1:1 with Percoll. In a density centrifugation step, the newborn swarmer cells were separated from the stalked cells. The swarmer cells were collected and washed once in M2 salts, before being released into PYE (including 2 µM xylose) and allowed to grow at 28°C until they were analyzed by microscopy.

#### Experimental determination of cell sizes and intracellular locations of chromosomal loci throughout the cell cycle

To determine the dimensions of *C. crescentus* cells, as well as the copy number and intracellular location of specific fluorescently-labeled chromosomal loci at specific time points in their cell cycle, we subjected cells at specific time points after synchronization (see above) to fluorescence microscopy. To this end, cells were immobilized on pads made of 1% agarose in water and observed with a Nikon Ti2 Eclipse microscope. The microscope was equipped with an alpha Plan Apo $\lambda$ 100x/1.45 Oil ($\infty$)/0.17 WD 0.13 Ph3 objective (Nikon, Japan), a Spectra X Light Engine (Lumencor, USA) light source and a CFP-2432C and a YFP-2427B filter (Semrock, USA). Images were collected with an Orca-flash4.0LT Plus C11440-42U30 camera (Hamamatsu, Japan) and recorded with NIS Elements 5.30.02 (Nikon, Japan).

In order to extract the cellular dimensions and intracellular positions of the fluorescently-labeled loci, cells were segmented based on the phase contrast channel using MicrobeJ [146]. To be able to monitor cell cycle progression in each subset, both cell lengths and the percentage of cells showing constrictions (as detected via MicrobeJ's feature constriction option) were followed for each timepoint after synchronization for each strain. Fluorescent maxima were detected using the maxima detection, which determines the localizations of the maxima relative to the poles of the cell at sub-pixel resolution using a Gaussian fit.

Table 3.1: *C. crescentus* strains used in this study

| Strain | Genotype/description | Reference |
|---|---|---|
| CB15N | Synchronizable wild-type strain | Evinger & Agabian (1977) [147] |
| MvT171 | CB15N $P_{xyl}$::$P_{xyl}$-*lacI-cfp-tetR-yfp*10x tetO and 10x lacO spaced 10.0 kb apart at 108° | Messelink et al (2021) [69] |
| Tn3 | CB15N $P_{xyl}$::$P_{xyl}$-*lacI-cfp-tetR-yfp* $(lacO)_n$ integrated at the *ori* and $(tetO)_n$ integrated at bp 957206 (86°) | Viollier et al. (2004) [142] |
| Tn4 | CB15N $P_{xyl}$::$P_{xyl}$-*lacI-cfp-tetR-yfp* $(lacO)_n$ integrated at the *ori* and $(tetO)_n$ integrated at bp 2026048 (182°) | Viollier et al. (2004) [142] |
| Tn8 | CB15N $P_{xyl}$::$P_{xyl}$-*lacI-cfp-tetR-yfp* $(lacO)_n$ integrated at the *ori* and $(tetO)_n$ integrated at bp 1498826 (134°) | Viollier et al. (2004) [142] |
| Tn11 | CB15N $P_{xyl}$::$P_{xyl}$-*lacI-cfp-tetR-yfp* $(lacO)_n$ integrated at the *ori* and $(tetO)_n$ integrated at bp 3029646 (272°) | Viollier et al. (2004) [142] |
| Tn49 | CB15N $P_{xyl}$::$P_{xyl}$-*lacI-cfp-tetR-yfp* $(lacO)_n$ integrated at the *ori* and $(tetO)_n$ integrated at bp 596575 (54°) | Viollier et al. (2004) [142] |
| Tn72 | CB15N $P_{xyl}$::$P_{xyl}$-*lacI-cfp-tetR-yfp* $(lacO)_n$ integrated at the *ori* and $(tetO)_n$ integrated at bp 2673431 (239°) | Viollier et al. (2004) [142] |
| Tn85 | CB15N $P_{xyl}$::$P_{xyl}$-*lacI-cfp-tetR-yfp* $(lacO)_n$ integrated at the *ori* and $(tetO)_n$ integrated at bp 433392 (39°) | Viollier et al. (2004) [142] |
| Tn102 | CB15N $P_{xyl}$::$P_{xyl}$-*lacI-cfp-tetR-yfp* $(lacO)_n$ integrated at the *ori* and $(tetO)_n$ integrated at bp 3645906 (329°) | Viollier et al. (2004) [142] |

### 3.4.2 Model construction

As detailed in the Main Text, to construct the replicating MaxEnt model, the following input data are needed for each cell cycle progression time:

1. bias-corrected Hi-C data,
2. the mean replication fork position,
3. the mean cell length,
4. the mean separation between replicated *ori*'s.

Input 4 is obtained as described in Sec. 3.4.1. In this section, we describe how inputs 1-3 are obtained.

#### Input Hi-C data

For the input Hi-C data, the raw Hi-C counts cannot be used directly: these may contain biases due to varying crosslinker affinities between genomic regions, and to a lesser extent differences in GC content and the mappability of individual reads [58]. To remove such biases for the data set at $t = 0$ minutes after synchronization, a normalization procedure rescaling the Hi-C map such that all rows and columns add to 1 was applied in [58]. For Hi-C maps for later times after synchronization, we make use of the column sums in the raw data set at $t = 0$. The rescaling factor $f_i$ applied to each column $i$ in the data set at $t = 0$ encodes the read count bias of site $i$. Thus, for each data set at time $t$, dividing the sums of each column $i$ by the sums of column $i$ in the data set at $t = 0$, we obtain the bias-corrected relative column sums $\sigma_i^{\mathrm{corr}}$ of each column $i$ (Fig. 3.5). This procedure assumes that biases in crosslinker affinities and mappability of individual reads remain constant throughout the cell cycle.

To construct an input contact frequency data set $f_{ij}^{\mathrm{expt}}(t)$ that matches these bias-corrected column sums $\sigma_i^{\mathrm{corr}}$, we make use of an iterative procedure [58]. First, for a given cell cycle time $t$ we calculate the sum of the entire Hi-C matrix $\Sigma(t) = \sum_{i,j} m_{ij}(t)$, and the sum of each column $\sigma_i(t) = \sum_j m_{ij}(t)$. Each Hi-C matrix entry $m_{ij}(t)$ we then rescale according to $m_{ij}(t) \to m_{ij}(t)\frac{f_i f_j}{\sigma_i(t)\sigma_j(t)}\Sigma(t)$. This procedure is repeated until the target column sums are matched within an average relative deviation of 1 in 100000, which typically takes 15 iteration steps. Lastly, the entire Hi-C matrix is rescaled such that the sum of all entries equals the number of columns. The resulting input Hi-C maps $f_{ij}^{\mathrm{expt}}(t)$ for each cell cycle time $t$ are shown in Fig. 3.6.

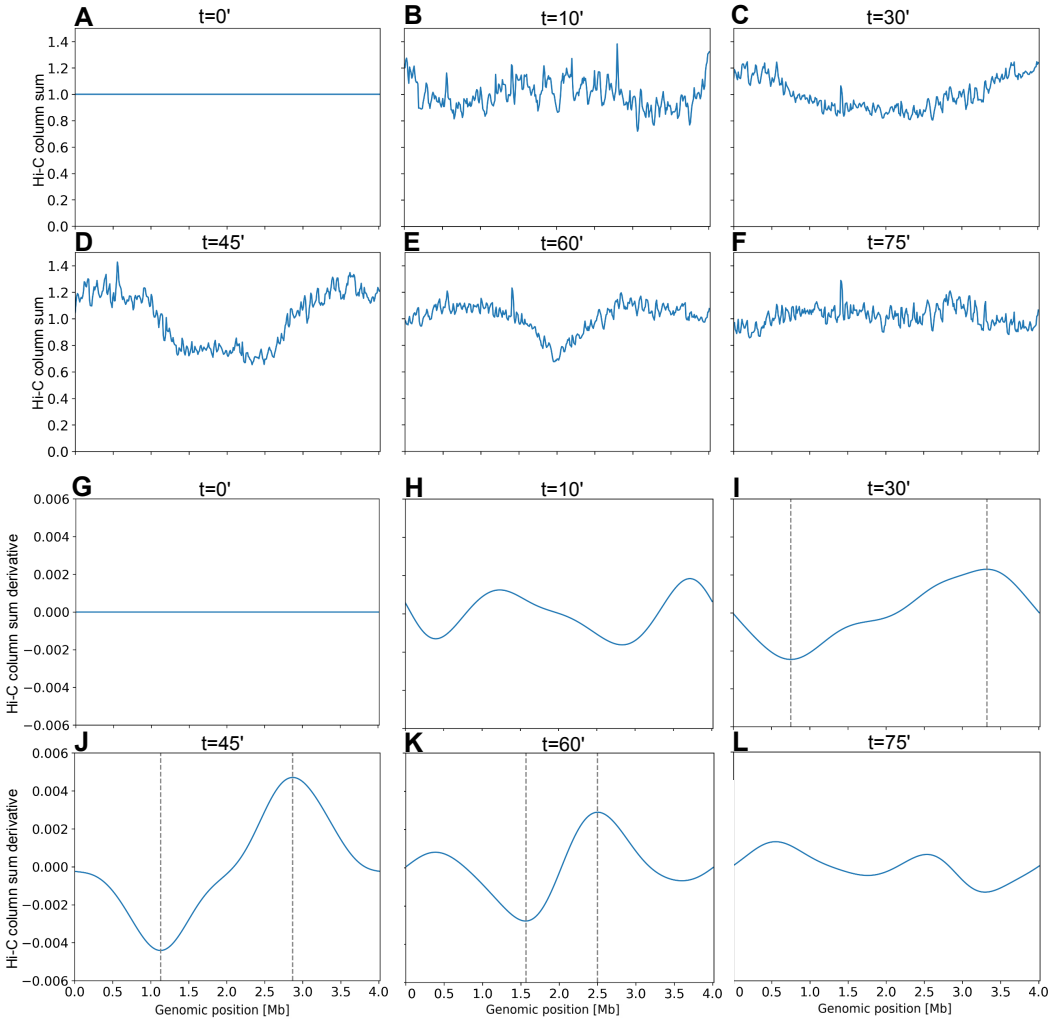



Figure 3.5: **Column sums for the input contact frequencies $f_{ij}^{\mathbf{expt}}(t)$, together with smoothed derivatives.** **A-F**: column sums $\sum_j f_{ij}^{\mathrm{expt}}(t)$ for each genomic position $i$ and each cell cycle progression time $t$. **G-L**: derivatives of the curves shown in (A-F), after applying a Gaussian smoothing with a sigma of 350kb. Dashed lines indicate the estimated replication fork positions for each chromosomal arm.

## Replication fork position

To estimate the mean replication fork position for each Hi-C data set, we use the bias-corrected column sums $\sigma_i^{\mathrm{corr}}$ calculated in Sec. 3.4.2. Due to a higher copy number of genomic regions per cell after replication, we expect the Hi-C score to be higher for regions where the replication fork has passed. Considering the column sums for the bias-corrected Hi-C data sets, we indeed find a clear transition between two Hi-C score regimes for the data sets at 30, 45 and 60 minutes after synchronization (Fig. 3.5).

Smoothing the column sums and taking a numerical derivative, we find an inflection point, which we interpret as the mean replication fork position (Fig. 3.5). Combining the estimated mean replication fork positions for the Hi-C data sets at 30,45, and 60 minutes after synchronization, we find that the inferred locations lie closely along a linear fit for both chromosomal arms 3.7. Importantly, the replication progress for both chromosomal

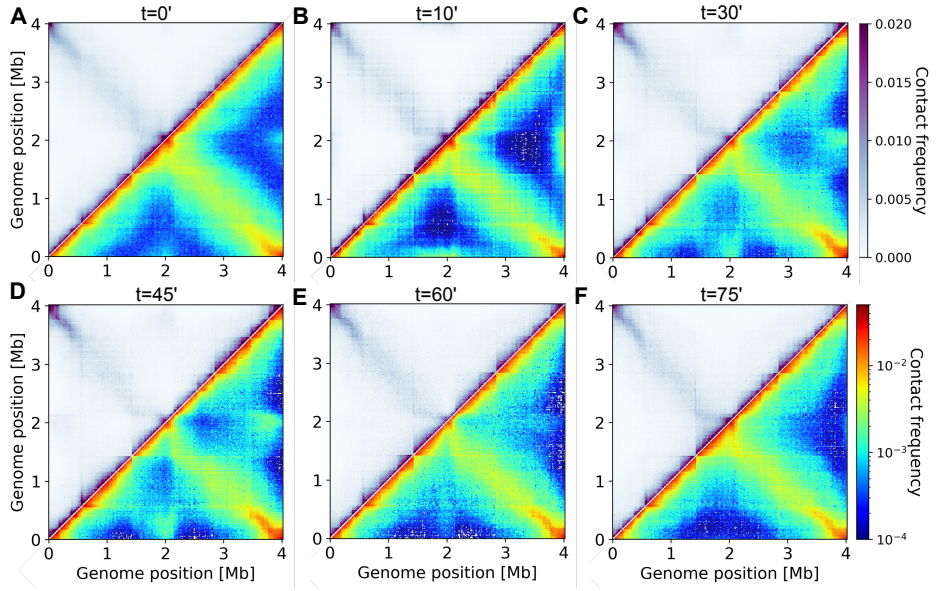



Figure 3.6: **Input Hi-C maps. A-F**: Input Hi-C maps $f_{ij}^{\text{expt}}(t)$ for *C. crescentus* cells for each cell cycle progression time, obtained via the bias correction procedure described in Sec. 3.4.2, applied to data from [58]. Upper left triangles: linear scale. Lower right triangles: logarithmic scale.

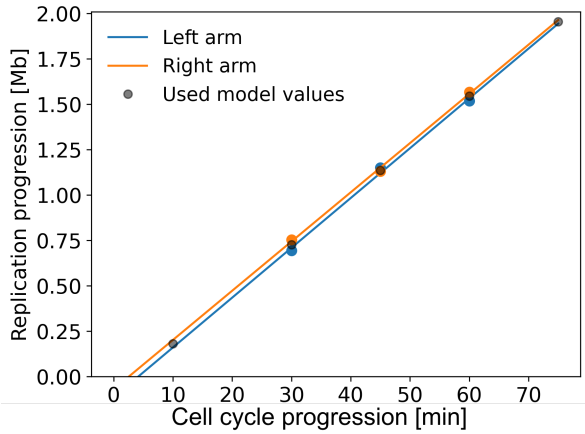



Figure 3.7: **Inferred progress of the mean replication fork position** Orange/blue dots: mean replication fork positions determined for the right/left chromosomal arms, as shown in Fig 3.6. Orange/blue lines: linear fits through the obtained mean fork positions for the right/left arms. Black dots: average of the two linear fits, shown for each time since synchronization for which a Hi-C data set was obtained by [58]. These averages are the positions used in the MaxEnt model construction.

arms is determined independently. Thus the close agreement between the two arms provides additional support for the accuracy of the inferred replication progress. The mean of the linear fits for both arms at each time since synchronization is used for the MaxEnt model construction.

### Cell size

To determine the mean cell size associated with each replication fork position, we make use of the fluorescent microscopy images described in Section 3.4.1. For each time since

synchronization, the mean cell length is determined (blue dots in Fig. 3.8). From this, the estimated cell envelope width [69] is subtracted, yielding the confinement lengths used as model inputs (black dots in Fig. 3.8). For the confinement width, a cylinder with rounded caps as used in [69] is applied, with the confinement width of 0.63 $\mu$m assumed to be constant throughout the cell cycle.

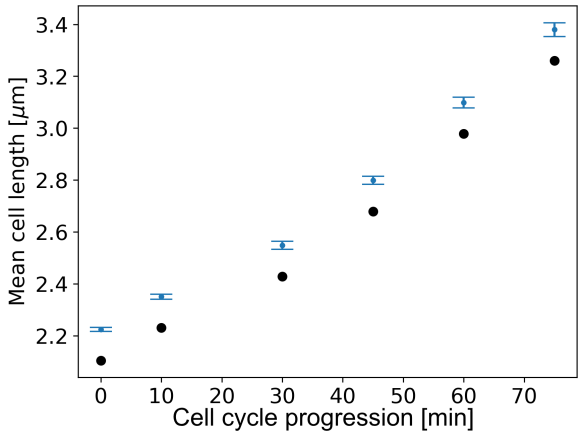



Figure 3.8: **Mean cell lengths at each time since synchronization** Blue dots: mean cell lengths determined for each time since synchronization. The error bars indicate two times the error on the mean. Black dots: cell lengths used as MaxEnt model inputs, with the estimated cell envelope width subtracted from the mean cell lengths.

### 3.4.3 Monte Carlo simulation

The Monte Carlo simulation of the lattice polymer is performed as in [69], with one extra move included to change the replication fork position. The added 'fork move', illustrated in Fig. 3.9, together with the loop move, kink move, and crankshaft move, forms the set of moves randomly chosen from at each step of the Monte Carlo simulation.

The addition of the fork moves preserves ergodicity, which we can see as follows. Considering the unreplicated portion of the chromosome and one of the replicated segments, which for convenience we name replicated segment 1, these together have the topology of a single unreplicated chromosome. This subset of monomers, which we term subset $S$, can be modified in the same way as the unreplicated chromosome in [69], which was shown to be ergodic under the loop, kink, and crankshaft moves, with the exception of the replication fork site $M$ which is only modified by the fork move. From the perspective of subset $S$ however, the fork move simply behaves as either a kink move or a loop move, with the constraint that the first monomer of replicated segment 2 is in the right position to allow these moves.

This constraint is satisfied as long as it is possible to put the first monomer of replicated segment 2 at each of the six possible sites around the replication fork site by a sequence of polymer moves. Here, replicated segment 2 can be considered as a linear polymer with fixed endpoints, where each of its monomers is subject to the kink, loop and crankshaft moves. Thus, segment 2 is subject to ergodic sampling. The only case where it is not possible to put the first monomer of replicated segment 2 at any position around site $M$, is therefore if the replicated segments are completely stretched, i.e. the L1 norm of the distance vector between the replication forks is equal to $2M$. As this is longer than our confinement length for all cell cycle times considered here, this exception is not relevant for our model. Thus, for our subset

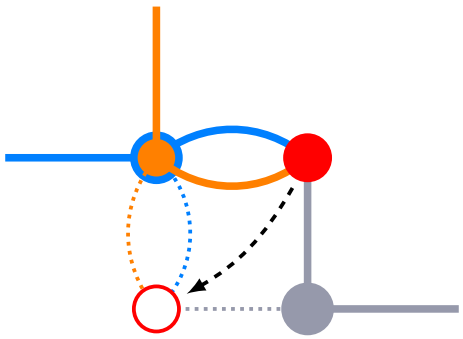


Figure 3.9: **Illustration of the fork move used in the Monte Carlo algorithm.** The fork move is performed on the junction site $M$ (red) if two of the three connected monomers overlap (orange and blue), and the third monomer (grey) is at a 90°angle to the other two.

$S$ we can consider the replication fork site $M$ to effectively move according to the kink and loop moves.

As the crankshaft move can be constructed as a combination of loop and kink moves, the replication fork site $M$ is effectively subject to the same moves as the other monomers in subset $S$, and thus the monomers in subset $S$ are subject to an ergodic sampling of states. As segment 1 and segment 2 are identical, both possible subsets that can be constructed are subject to ergodic sampling, thus this also holds for the chromosome as a whole.

### 3.4.4 MaxEnt model trained only on Hi-C data

In [69], a MaxEnt model for unreplicating *C.crescentus* cells was developed with Hi-C data as the only imposed constraint. Imposing only Hi-C constraints on our replicating chromosome model, we find that the model solution does not exhibit chromosomal segregation (Fig. 3.10). Thus, an additional constraint on the mean separation between replicating *ori*'s is imposed, as detailed in the Main Text.

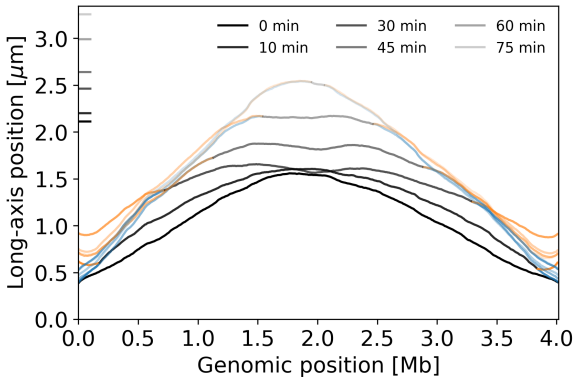



Figure 3.10: **Average long-axis localizations for a MaxEnt model trained only on Hi-C data.** Average long-axis positions shown for the old chromosome (blue), new chromosome (orange) and unreplicated chromosome (black). The transparency of the lines indicates the cell cycle progression time. The cellular orientation and the assignment of old and new chromosomes are done as in Main Text Fig. 3.

### 3.4.5 Correcting for indistinguishability of fluorescent foci at short distance

As described in the Main Text, we perform a conditional averaging on experimentally measured localizations to directly compare experimental and model predictions. This conditional averaging is required as the MaxEnt model contains either exactly one or exactly two copies of a chromosomal region at a given time point, whereas intermediate average values are typically observed in experiment. This discrepancy is attributed to a combination of imperfect synchronization of cells and imperfect label detection. One contribution to the imperfect label detection is the indistinguishability of fluorescent foci at short distance, which results in a systematic bias in inferred mean label positions. In this section we estimate the magnitude of this bias, and describe how we compute a modified MaxEnt localization prediction that incorporates the experimental bias.

#### Estimating the distance at which fluorescent foci become indistinguishable

To estimate the distance at which fluorescent foci become indistinguishable, we construct a histogram of observed pairwise focus distances 3.11. In this histogram, we find a sudden cutoff to zero counts for distances below 0.32 $\mu$m. This cutoff is consistent with our assumption of indistinguishability of fluorescent foci at short distance. The observed minimum distance of 0.32 $\mu$m is used as a cutoff value in the calculation of bias-corrected MaxEnt localizations (SI 3.4.5).

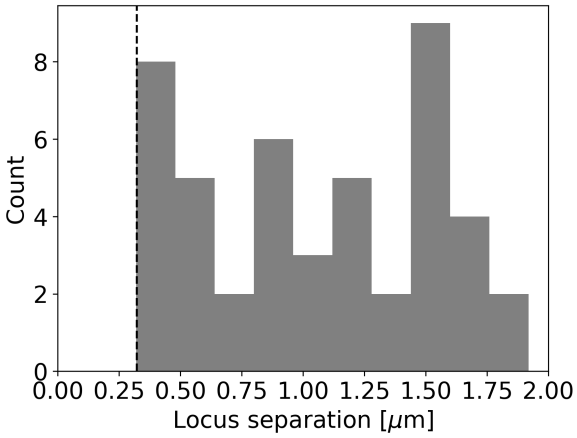



Figure 3.11: **Histogram of measured focus distances at 86°.** Data are taken from all time points from the Tn3 data set. The dashed vertical line indicates the smallest observed distance, equal to 0.32 $\mu$m, which is used as input for the bias-corrected MaxEnt localization prediction (SI 3.4.5).

#### Calculating corrected MaxEnt localization profiles

To calculate the bias-corrected MaxEnt localization profiles shown in Main Text Fig. 2, we generate an ensemble of MaxEnt chromosome configurations for each time point. For each configuration, we compute 2D distances between the experimentally tagged regions, projected along the cell length and cell width coordinates. Only loci pairs who's projected 2D distance is above the threshold determined in Sec. 3.4.5, are included to obtain the bias-corrected localization profiles shown in Main Text Fig. 2.

### Fraction of cells with two labels over time

In Fig. 3.12, the measured fractions of cells with two foci are shown for all measurement conditions, together with the MaxEnt prediction taking the indistinguishability of foci at short distance into account. The results suggest that this indistinguishability strongly contributes to the observed fractions. Furthermore, we find that the onset of replication for tagged regions, taken as the onset of non-zero fractions, matches well between model and experiment.

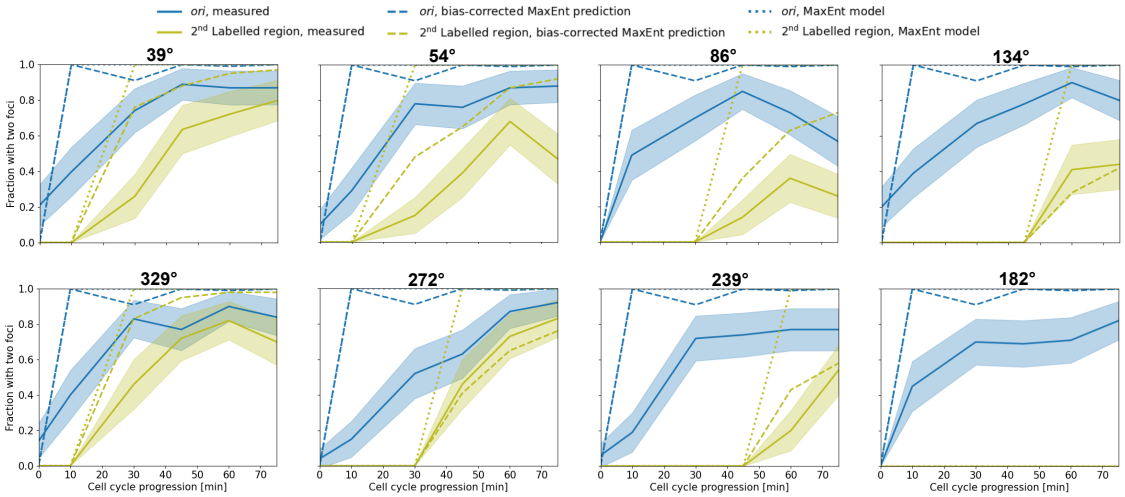



Figure 3.12: **Fraction of cells with two foci for each measured locus, together with MaxEnt prediction.** Blue solid lines: measured fraction of cells with two *ori*'s. Blue dashed lines: predicted measured fraction by the MaxEnt model, given the indistinguishability of two foci at short distance. Blue dotted lines: locus counts in the MaxEnt model. Yellow solid lines: measured fraction of cells with two copies of the locus indicated at the top left. Yellow dashed lines: predicted fraction for the same region by the MaxEnt model. Yellow dotted lines: locus counts of the same region in the MaxEnt model. The shaded areas indicate the error margins (2*SEM) on the experimental data, calculated by assuming binomial sampling.

# Chapter 4

# The unusual single-cell growth of *Corynebacterium glutamicum*

## Chapter summary

In this chapter, we shift perspective from the organization of a replicating chromosome throughout the cell cycle, to the growth of its enclosing cell. Replication must be tightly coordinated with cellular growth, to ensure a conserved chromosome density across generations and a faithful transfer of genetic material to each daughter cell. So far, single bacterial cells have been found to grow predominantly exponentially, which implies the need for tight growth regulation mechanisms to ensure cell size homeostasis. In this chapter, we investigate the single-cell growth behaviour of *Corynebacterium glutamicum*, which has several highly atypical growth characteristics. The cell wall forms a thick meshwork around the cell, and cell wall growth occurs exclusively at the cell poles. Furthermore, *C. glutamicum* lacks many common growth-regulatory mechanisms. Therefore, it is a promising candidate to search for novel single-cell growth modes.

From detailed single-cell microscopy experiments, we obtain growth statistics over time and across generations. Extracting single-cell growth behaviour from this is challenging however, due to noise and intrinsic variability from cell to cell. We develop an inference procedure to extract average single-cell elongation profiles, using the noise-reducing properties of multi-cell averaging, while carefully avoiding inspection bias effects. Our method is validated with simulated cells following various single-cell growth modes in the presence of noise.

From this inference procedure, we learn that single *C. glutamicum* cells do not follow the generally observed exponential single-cell growth. Rather, cells initially increase their elongation rate, but transition to a regime of approximately linear growth for later growth times. To understand this growth behaviour, we model single-cell growth as being rate-limited by the apical growth mechanism, termed the Rate-Limiting Apical Growth (RAG) model. Within this model, the initial acceleration of growth is due to the maturation of the new cell pole, while a linear growth regime is reached once the new pole has matured. We find this model be consistent with the observed elongation rate curves. Elongation measurements on a ΔrodA mutant, where part of the apical cell wall insertion mechanism is inhibited, reveal an overall downward shift of elongation rates, as would be expected based on the RAG model.

To investigate the implications of asymptotically linear growth for cell size regulation, we simulate a population of growing cells performing either exponential or linear growth, with all growth parameters taken directly from our data set. We find that while asymptotically

linear growth results in a relatively narrow distribution of cell sizes, whereas exponential growth with the same parameters yields a long-tailed distribution of cell lengths. This offers an evolutionary explanation for *C. glutamicum*'s lack of of many common size-regulation mechanisms.

## 4.1 Publication

# Single-cell growth inference of Corynebacterium glutamicum reveals asymptotically linear growth

by

**Joris J.B. Messelink[*,1], Fabian Meyer[*,2,3], Marc Bramkamp[2,3], Chase P. Broedersz[1,4]**

[*]Equal contribution.
[1]Arnold Sommerfeld Center for Theoretical Physics and Center for NanoScience, Department of Physics, Ludwig Maximilian University Munich, Munich, Germany.
[2]Ludwig-Maximilians-Universität München, Fakultät Biologie, Planegg-Martinsried, Germany
[3]Christian-Albrechts-Universität zu Kiel, Institut für allgemeine Mikrobiologie, Kiel, Germany
[4]Department of Physics and Astronomy, Vrije Universiteit Amsterdam, Amsterdam, Netherlands

**Reprinted on pages 110 - 148 from**

**eLife 10:e70106 (2021),**
**doi:10.7554/eLife.70106**

RESEARCH ARTICLE

# Single-cell growth inference of *Corynebacterium glutamicum* reveals asymptotically linear growth

**Joris JB Messelink[1†], Fabian Meyer[2,3†], Marc Bramkamp[2,3*], Chase P Broedersz[1,4*]**

[1]Arnold-Sommerfeld-Center for Theoretical Physics, Ludwig-Maximilians-Universität München, Munich, Germany; [2]Ludwig-Maximilians-Universität München, Fakultät Biologie, Planegg-Martinsried, Germany; [3]Christian-Albrechts-Universität zu Kiel, Institut für allgemeine Mikrobiologie, Kiel, Germany; [4]Department of Physics and Astronomy, Vrije Universiteit Amsterdam, Amsterdam, Netherlands

**Abstract** Regulation of growth and cell size is crucial for the optimization of bacterial cellular function. So far, single bacterial cells have been found to grow predominantly exponentially, which implies the need for tight regulation to maintain cell size homeostasis. Here, we characterize the growth behavior of the apically growing bacterium *Corynebacterium glutamicum* using a novel broadly applicable inference method for single-cell growth dynamics. Using this approach, we find that *C. glutamicum* exhibits asymptotically linear single-cell growth. To explain this growth mode, we model elongation as being rate-limited by the apical growth mechanism. Our model accurately reproduces the inferred cell growth dynamics and is validated with elongation measurements on a transglycosylase deficient Δ*rodA* mutant. Finally, with simulations we show that the distribution of cell lengths is narrower for linear than exponential growth, suggesting that this asymptotically linear growth mode can act as a substitute for tight division length and division symmetry regulation.

***For correspondence:**
bramkamp@ifam.uni-kiel.de (MB);
c.p.broedersz@vu.nl (CPB)

[†]These authors contributed equally to this work

**Competing interests:** The authors declare that no competing interests exist.

**Funding:** See page 15

**Preprinted:** 26 May 2020
**Received:** 06 May 2021
**Accepted:** 01 October 2021
**Published:** 04 October 2021

**Reviewing editor:** Aleksandra M Walczak, École Normale Supérieure, France



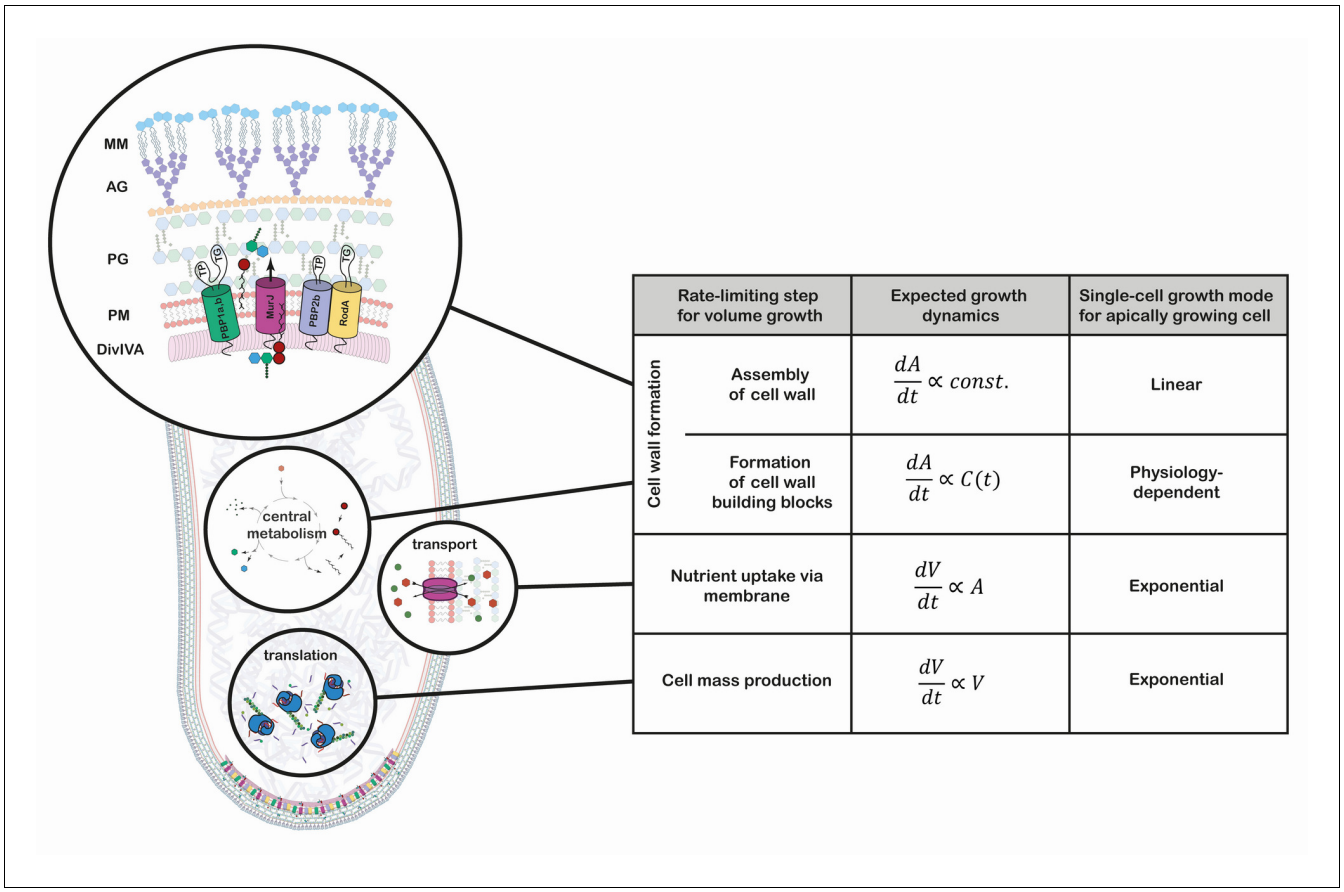
## Introduction

Regulated single-cell growth is crucial for the survival of a bacterial population. At the population level, fundamental laws of growth were discussed as early as the beginning of the 20th century, and distinct population growth phases were identified and attributed to bacterial growth (***Lane-Claypon, 1909***; ***Buchanan, 1918***; ***Monod, 1949***). At the time, however, growth behavior at the single-cell level remained elusive. This changed only over the last decade, as evolving technologies enabled detailed measurements of single-cell growth dynamics. Extensive work was done on common model organisms, including *Escherichia coli*, *Bacillus subtilis*, and *Caulobacter crescentus*, revealing that averaged over the cell cycle, single cells grow exponentially for these species (***Taheri-Araghi et al., 2015***; ***Mir et al., 2011***; ***Iyer-Biswas et al., 2014***; ***Yu et al., 2017***; ***Godin et al., 2010***).

Single-cell exponential growth is expected if cellular volume production is proportional to the protein content (***Amir, 2014***), as shown to be the case for *E. coli* (***Belliveau et al., 2020***). Importantly, however, such a proportionality will only be present if cellular volume production is rate-limiting for growth. Cells with different rate-limiting steps could display distinct growth behavior. Recently, detailed analysis of the mean growth rate throughout the cell cycle revealed deviations from pure exponential growth. For *B. subtilis* (***Nordholt et al., 2020***), a biphasic growth mode was observed, where a phase of approximately constant elongation rate is followed by a phase of increasing elongation rate. For *E. coli*, a new method provides evidence for super-exponential in the later stages of the cell cycle (***Kar et al., 2021***).

A promising candidate for uncovering strong deviations from exponential growth is the Gram-positive *Corynebacterium glutamicum*. This rod-shaped bacterium grows its cell wall exclusively at the cell poles, allowing, in principle, for deviations from exponential single-cell growth (***Figure 1***). The dominant growth mode depends on the rate-limiting step for growth, which is presently unknown for this bacterium. Non-exponential growth modes may have important implications for growth regulatory mechanisms: while exponential growth requires checkpoints and regulatory systems to maintain a constant size distribution (***Mir et al., 2011***), such tight regulation might not be needed for other growth modes.

*Corynebacterium glutamicum* is broadly used as a production-organism for amino-acids and vitamins and also serves as model organism for the taxonomically related human pathogens *Corynebacterium diphteriae* and *Mycobacterium tuberculosis* (***Hermann, 2003***; ***Antoine et al., 1988***; ***Schubert et al., 2017***). A common feature of Corynebacteria and Mycobacteria is the existence of a complex cell envelope. The cell wall of these bacteria is a polymer assembly composed of a classical bacterial peptidoglycan (PG) sacculus that is covalently bound to an arabinogalactan (AG) layer (***Alderwick et al., 2015***). Mycolic acids are fused to the arabinose and form an outer membrane like




| | Rate-limiting step for volume growth | Expected growth dynamics | Single-cell growth mode for apically growing cell |
|---|---|---|---|
| Cell wall formation | Assembly of cell wall | $\frac{dA}{dt} \propto const.$ | Linear |
| Cell wall formation | Formation of cell wall building blocks | $\frac{dA}{dt} \propto C(t)$ | Physiology-dependent |
| | Nutrient uptake via membrane | $\frac{dV}{dt} \propto A$ | Exponential |
| | Cell mass production | $\frac{dV}{dt} \propto V$ | Exponential |

**Figure 1.** Growth mode analysis for four possible rate-limiting steps for cellular volumegrowth in the apically growing *C. glutamicum*. Here, $V$ is the cellular volume, $A$ is the cell wall area, and $C(t)$ is the concentration of membrane building blocks in the cytoplasm. A constant cell width is assumed throughout, implying $\frac{dA}{dt} \propto \frac{dV}{dt}$. A fixed production capacity per unit volume is assumed for the rate-limiting steps 'cell mass production' and 'formation of cell wall building blocks'. For the rate-limiting step 'assembly of cell wall', a constant insertion area at the cell poles is assumed. For an analysis of the single-cell growth mode if cell wall building block formation is the rate-limiting step for growth, see Appendix 1. Cell mass production, specifically ribosome synthesis, has previously been indicated as the rate-limiting step for growth in *E. coli* (***Belliveau et al., 2020***; ***Scott et al., 2010***; ***Amir, 2014***). Linear growth is observed if the rate-limiting step for volume growth is the cell wall assembly (shown here in a simplified representation). The protein DivIVA serves as a scaffold at the curved membrane of the cell pole for the recruitment of the Lipid-II flippase MurJ and several mono- and bi-functional trans-peptidases (TP) and -gylcosylases (TG). In the process of elongation, peptidoglycan (PG) precursors are integrated into the existing PG sacculus, which serves as a scaffold of the synthesis of the arabinogalactan-layer (AG) and the mycolic-acid bilayer (MM).

bilayer, rendering the cell surface highly hydrophobic (**Puech et al., 2001**). The mycolic acid membrane (MM) is an efficient barrier that protects the cells from many conventional antibiotics.

*C. glutamicum*'s growth and division behavior is vastly different to that of classical model species. In contrast to rod-shaped firmicutes and γ-proteobacteria, where cell-wall synthesis is dependent on the laterally acting MreB, members of the *Corynebacterianeae* lack a *mreB* homologue and elongate apically. This apical elongation is mediated by the protein DivIVA, which accumulates at the cell poles and serves as a scaffold for the organization of the elongasome complex (**Letek et al., 2008**; **Hett and Rubin, 2008**; **Sieger et al., 2013**; **Figures 1** and **2A,B**). Furthermore, a tightly regulated division-site selection mechanism is absent in this species. Without harboring any known functional homologues of the Min- and nucleoid occlusion (Noc) system, division typically results in unequally sized daughter cells (**Donovan et al., 2013**; **Donovan and Bramkamp, 2014**). Lastly, the spread in growth times between birth and division is much wider than in other model organisms, suggesting a weaker regulation of this growth feature (**Donovan et al., 2013**). These atypical growth properties suggest that this bacterium is an interesting candidate to search for novel growth modes. To reveal the underlying growth regulation mechanisms, it is necessary to study the elongation dynamics of *C. glutamicum*.

Here, we measure the single-cell elongations within a proliferating population of *C. glutamicum* cells, and develop an analysis procedure to infer their growth behavior. We find that *C. glutamicum* deviates from the generally assumed single-cell exponential growth law. Instead, these *Corynebacteria* exhibit asymptotically linear growth. We develop a mechanistic model, termed the rate-limiting apical growth (RAG) model, showing that these anomalous elongation dynamics are consistent with the polar cell wall synthesis being the rate-limiting step for growth. Finally, we demonstrate a connection between mode of growth and the impact of single-cell variability on the cell size distribution of the population. For an asymptotically linear grower, these variations have a much smaller impact on this distribution than they would for an exponential grower, which may suggest an evolutionary explanation for the lack of tight regulation of single-cell growth in *C. glutamicum.*

# Results

## Measuring elongation trajectories using microfluidic experiments

To measure the development of single *C. glutamicum* cells over time, we established a workflow combining single-cell epifluorescence microscopy with semi-automatic image processing. Cells were grown in a microfluidic device. We used wild type cells and cells expressing the scaffold protein DivIVA as a translational fusion to mCherry. DivIVA is used as a marker for cell cycle progression, since it localizes to the cell poles and to the newly formed division septum in *C. glutamicum* (**Letek et al., 2008**; **Donovan et al., 2013**).

For the choice of microfluidic device, we deviate from the commonly used Mother Machine (**Long et al., 2013**), which grows bacteria in thin channels roughly equaling the cell width. The Mother Machine is not ideally suited for *C. glutamicum* growth, as the characteristic V-snapping at division could lead to shear forces and stress during cell separation, affecting growth (**Zhou et al., 2019**). Indeed, in some cases, the mother machine has been shown to affect growth properties even in cells not exhibiting V-snapping at division, due to mechanical stresses inducing cell deformation (**Yang et al., 2018**). Therefore, we instead used microfluidic chambers that allow the growing colony to expand without spatial limitations into two dimensions for several generations (**Figure 2C,D**, Materials and methods). Within the highly controlled environment of the microfluidic device, a steady medium feed and a constant temperature of 30˚C was maintained. We extracted bright-field- and fluorescent-images over 3-min intervals, which were subsequently processed semi-automatically with a workflow developed in FIJI and R (**Schindelin et al., 2012**; **R Development Core Team, 2003**). For each individual cell per time-frame, the data set contains the cell's length, area and estimated volume, the DivIVA-mCherry intensity profile, and information about generational lineage (**Figure 2E–G**). We used these data sets to further investigate the growth behavior of our bacterium. Thus, using this procedure, we obtained data sets containing detailed statistics on single-cell growth of *C. glutamicum*.

For subsequent analysis, the measured cell lengths were used, because of their low noise levels as compared to other measures (**Appendix 2—figure 1B**). Importantly, the increases in cell length

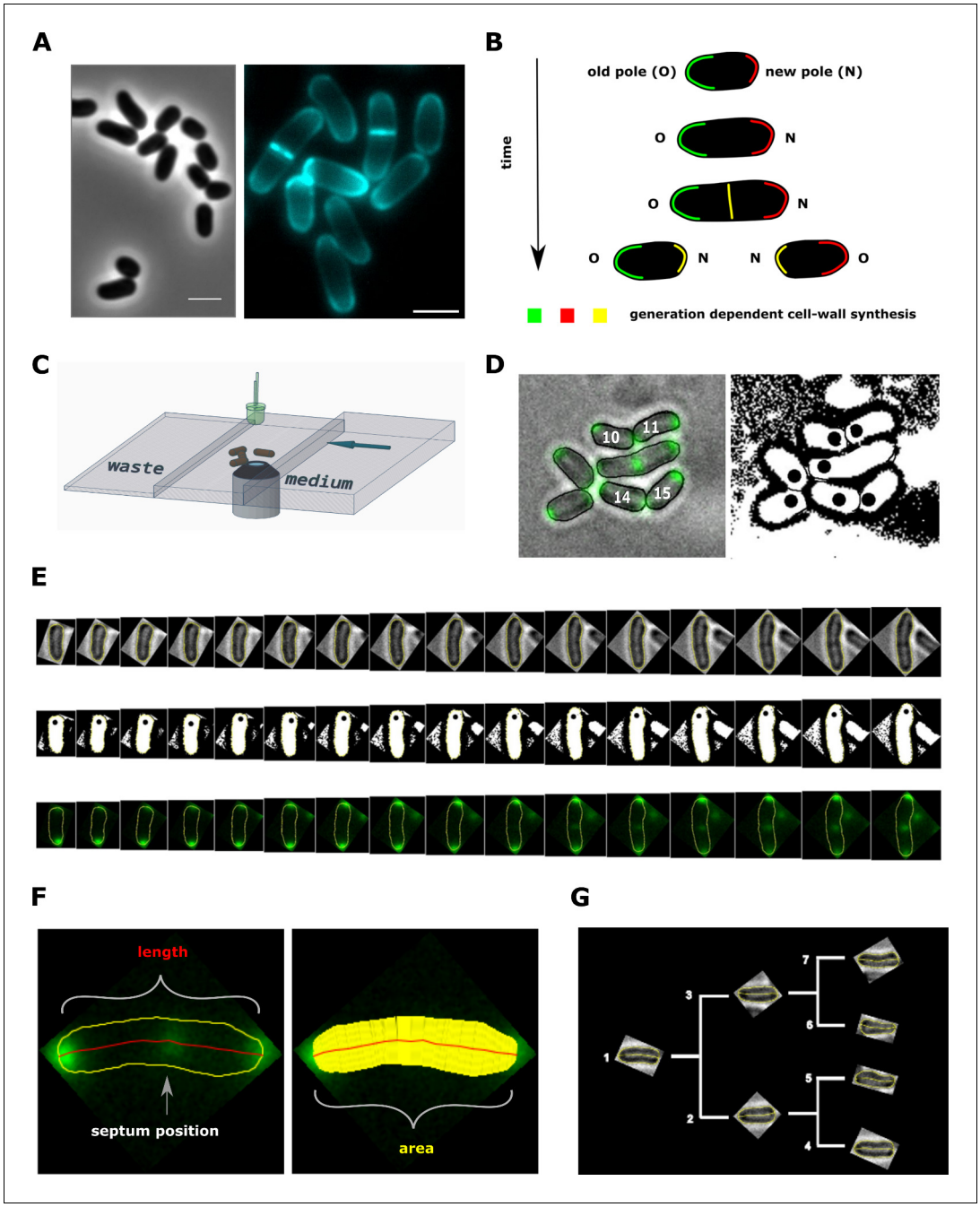



**Figure 2.** Experimental procedure and image analysis. (**A**, left) Phase contrast image of *C. glutamicum* in logarithmic growth phase, indicating the variable size of daughter cells. (**A**, right) HADA labeling of nascent peptidoglycan (PG), indicating the asymmetric apical growth where the old cell-pole always shows a larger area covered compared to the new pole. The labeling also reveals the variable septum positioning; Scale bar: 2 µm (**B**) Schematic showing the generation-dependent sites of PG synthesis in *C. glutamicum*, including the maturation of a new to an old cell-pole. (**C**) Illustration of the microfluidic device for microscopic monitoring of a growing colony. (**D**) Example screen-shot of the developed method to extract individual cell cycles from a multi-channel time-lapse micrograph. The left panel shows a merging of the bright-field channel and the mCherry-tagged DivIVA together with an individual ID# that is assigned to cells right after division. The black dots in the right panel indicate the new cell pole. (**E**) Example of an extracted individual cell cycle from birth (left) until prior to division (right), showing the bright-field (top), the orientation (middle) and the localization of mCherry-tagged DivIVA (bottom). (**F**) Example of the developed single cell analysis algorithm, measuring the length according to the cell's geometry, as well as the cell's area and the septum position relative to the new pole. (**G**) Dendrogram providing the rationale for identification of single cells in a growing colony.

are proportional to the increases in cell area (***Appendix 2—figure 1A***), suggesting that cellular length increase is also proportional to the volume increase. This proportionality is expected since the rod-shaped *C. glutamicum* cells insert new cell wall material exclusively at the poles, while maintaining a roughly constant cell width over the cell cycle (***Schubert et al., 2017***; ***Daniel and Errington, 2003***).

## Population-average test suggests non-exponential growth for *C. glutamicum*

A standard way of characterizing single-cell bacterial growth, is to determine the average relation between birth length $l_b$ and division length $l_d$ (***Amir, 2014***). For *C. glutamicum*, we find an approximately linear relationship between these birth and division lengths, with a slope of 0.91±0.16 (2xSEM, ***Figure 3A***). This indicates that on a population level, *C. glutamicum* behaves close to the *adder* model, in which cells on average grow by adding a fixed length before dividing (***Jun and Taheri-Araghi, 2015***; ***Amir, 2014***).

To investigate the growth dynamics from birth to division, we first tested if our cells conform to the generally observed exponential mode of single-cell growth. To this end, we applied a previously developed analysis on bacterial elongation data (***Logsdon et al., 2017***), by plotting $\ln\left(\frac{l_d}{l_b}\right)$ versus the growth time (***Figure 3B***). For an exponential grower, with the same exponential growth rate $\alpha$ for all cells, the averages of $\ln\left(\frac{l_d}{l_b}\right)$ per growth time bin are expected to lie along a straight line with slope $\alpha$ intersecting the origin. By contrast, there appears to be a systematic deviation from this trend, with cells with shorter growth times lying above this line and cells with longer growth times lying below it, suggesting non-exponential elongation behavior. However, the significance and implications of these deviations for single-cell growth behavior are not clear from this analysis. There are several quantities that could be highly variable between cells that are averaged out in this representation, such as possible variations in exponential growth rate as a function of birth length, or variations in growth mode over time. Furthermore, it was recently shown that exponentially growing cells can appear non-exponential with this test in the presence of noise in the exponential growth rate (***Kar et al., 2021***). Thus, a more detailed analysis of the growth trajectories is needed to rule out exponential growth, and to quantitatively characterize the growth dynamics.

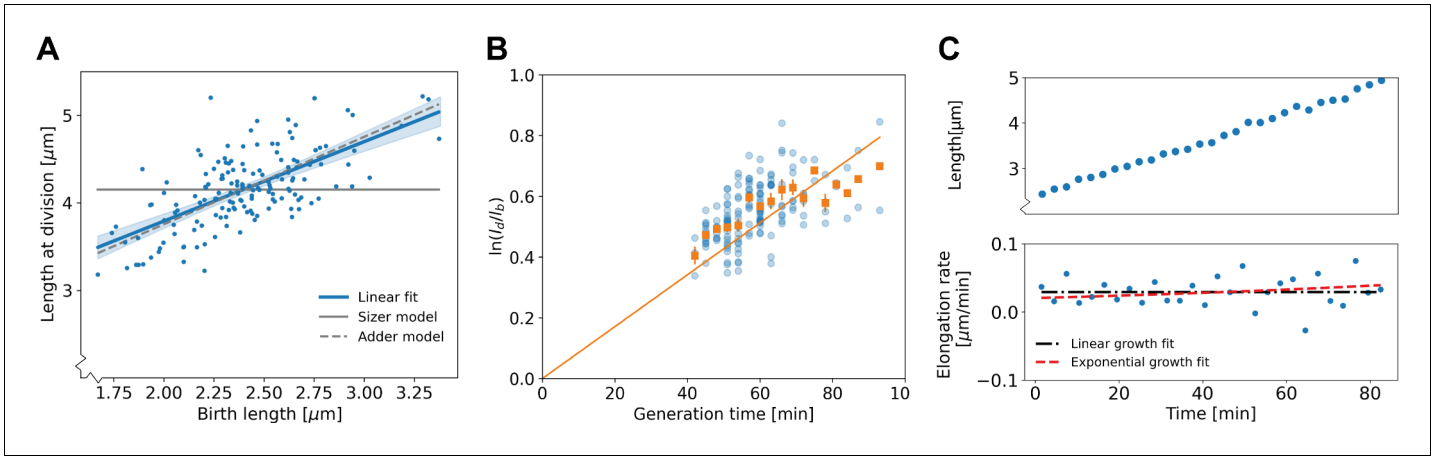



**Figure 3.** Population-level and single-cell level growth analysis. (**A**) Birth length $l_b$ plotted against division length $l_d$ for all measured cells, together with a linear fit (blue line), which has a slope of 0.91±0.16. Gray solid line: best fit assuming a pure sizer (slope 0). Gray dashed line: best fit assuming a pure adder (slope 1). The 95% confidence intervals of the linear fit, obtained via bootstrapping, are indicated by the blue shaded region. (**B**) Generation time versus $\ln\left(\frac{l_d}{l_b}\right)$ for all cells (blue dots) and the average per generation time (orange squares), with the standard error of the mean shown for all generation times for which at least three data points are available. The orange line represents a linear fit through the generation time averages that passes through the origin. For exponential growth, the averages would lie along this line, and the slope would be equal to the exponential growth rate. (**C**) Growth trajectory for a single cell (upper panel), together with its derivative for each measurement interval (lower panel). Fits to the derivative are shown for linear growth (black dash-dotted line) and exponential growth (red dashed line).

The variability of key growth parameters is not easily extracted from individual growth trajectories due to the inherent stochasticity of the elongation dynamics and measurement noise (***Figure 3C***). In fact, it has been estimated that to distinguish between exponential and linear growth for an individual trajectory, the trajectory needs to be determined with an error of ~6% (***Cooper, 1998***). Distinguishing subtler growth features may require an even higher degree of accuracy, which is presently experimentally unavailable (Appendix 3). Therefore, an analysis method is needed that is less noise-sensitive than an inspection of the single-cell trajectories, but simultaneously does not average out potentially relevant growth features such as time-dependence and birth length variability.

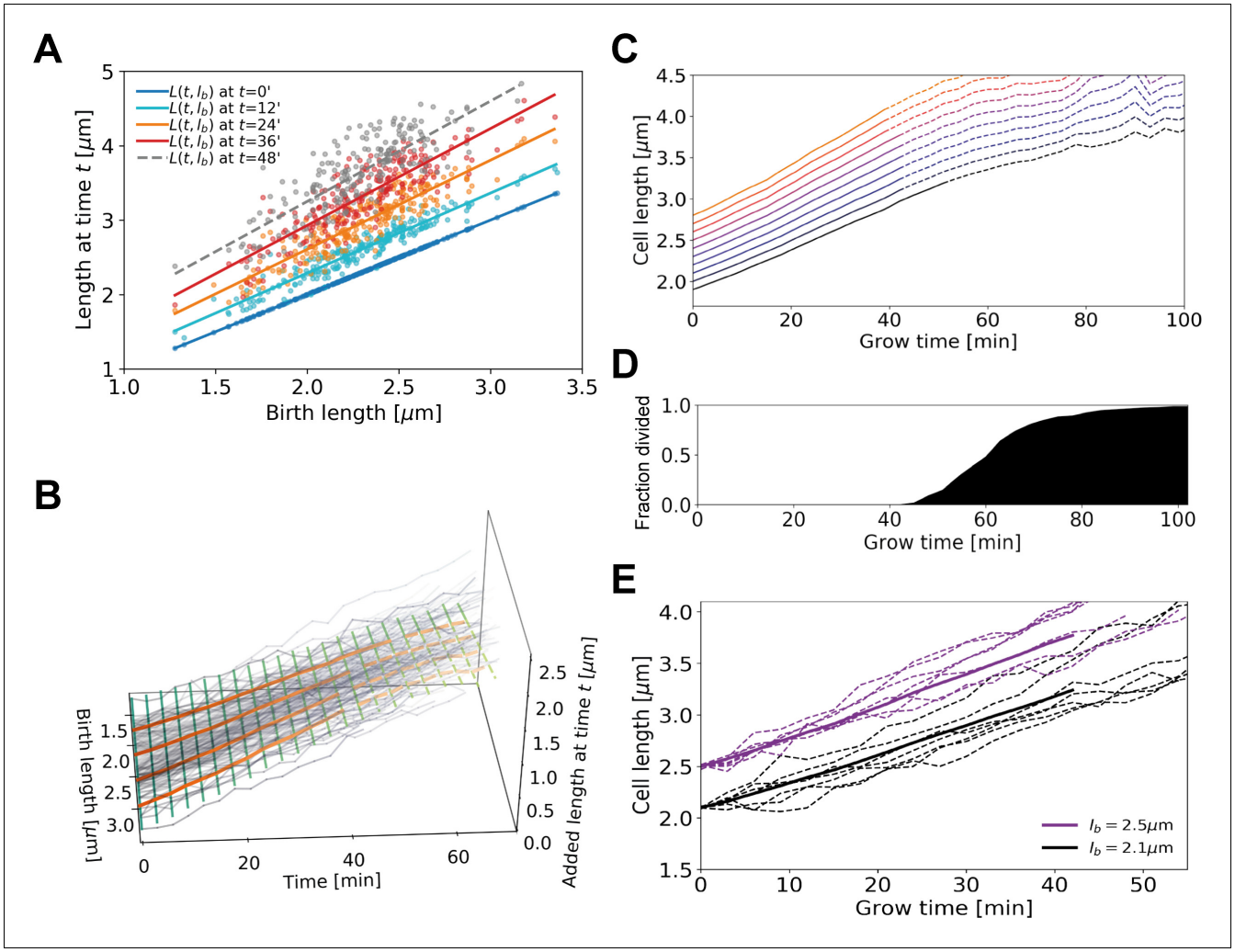



**Figure 4.** Average elongation curve inference procedure. (**A**) For each cell, the length $L(t)$ at different times $t$ since birth is plotted as a function of birth length $l_b$. A linear fit of the resulting 'wave front' is performed for each time $t$. This allows us to determine average cell length $L(t,\ l_b)$ at time $t$ as a function of birth length $l_b$. (**B**) 3D representation of the inference method of average length trajectories, with the added length $L(t,\ l_b) - l_b$ on the z-axis. Elongation trajectories for individual cells are indicated in gray, linear fits through all cell lengths at each timestamp are indicated by green lines. The orange lines represent four sample average length trajectories, obtained by connecting all values of the green lines associated with one birth length. Dotted lines represent regimes where averages are biased due to dividing cells. (**C**) Average elongation trajectories obtained from the fits shown in (A) for a range of birth lengths, starting at 1.9 µm with steps of 0.1 µm (solid lines). The dashed lines represent regions where the inferred elongation curves are biased due to dividing cells, and are excluded from subsequent analysis. (**D**) Cumulative fraction of cells divided as a function of grow time. (**E**) Elongation trajectories for cells with birth lengths close to 2.5 µm (purple dashed lines) and birth lengths close to 2.1 µm (black dashed lines) together with their respective inferred average trajectories (purple solid line and black solid line).

## Growth-inference method yields average elongation rate curves

To obtain quantitative elongation rate curves as a function of time and birth length, despite the high degree of individual variation, we developed a data analysis procedure that exploits the noise-reducing properties of multiple-cell conditional averaging. The key idea is to obtain an average dependence of the cellular length $L(t, l_b)$ on the time $t$ since birth and birth length $l_b$, by first obtaining the average dependence of $L(t, l_b)$ on $l_b$ for each discrete value of $t$ individually. This yields an average elongation curve for each birth length $l_b$, without the need to perform inference on noisy $L(t)$ single-cell curves.

The analysis procedure is as follows. First, for all cells in our data set, we determine the time since birth $t$, the cellular length $L$ at time $t$, and the birth length $l_b$. Subsequently, we relate the length at time $t$ to the birth length, yielding a series of scatter plots for each measurement time (***Figure 4A***). Importantly, these scatterplots suggest a simple apparently linear relationship between $L$ and $l_b$. For each such plot, we thus make a linear fit through the data, yielding a family of curves for each time since birth $t$ (***Figure 4B***). Higher-order fitting functions result in a negligible improvement of the goodness-of-fit, while increasing the mean error on inferred elongation rates (***Appendix 2—figure 3***). Note that for both purely linear and purely exponential growth, would depend linearly on: for linear growth $L(t,\ l_b) = \alpha t + l_b$, whereas for exponential growth $L(t,\ l_b) = l_b \exp(\alpha t)$ (***Appendix 2—figure 3***). From the family of relations, we compute a series of points $\{L(t_0,\ l_b),\ L(t_1,\ l_b),\ L(t_2,\ l_b)\}$ yielding the average growth trajectory of a cell starting out at length $l_b$ (***Figure 4C***). Note, we must remove a bias in the $l_b$ associated with each average trajectory, arising from measurement noise in the cell lengths at birth (Appendix 4). In summary, this procedure allows us to obtain an unbiased interference of the average elongation trajectories as a function of the cell's birth length.

## Elongation rate inference reveals asymptotically linear growth mode

Our inference approach yields the functional dependence of the average added length on growth time and birth length. We find that the average length steadily increases initially, but levels off and shows pronounced fluctuations for larger growth times (***Figure 4C***). This late-time behavior (dashed lines in ***Figure 4C***) is caused by decreasing cell numbers due to division events (***Figure 4D***), which also introduces a bias in the averaging procedure. After the first division event, the average inferred growth would be conditioned on the cells that have not divided yet. For a given birth length, faster-

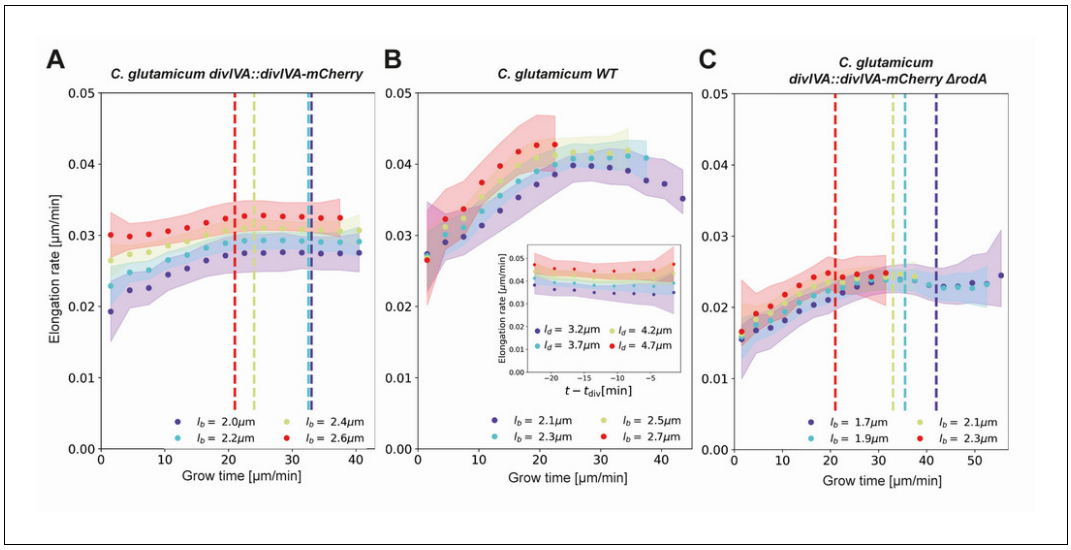



**Figure 5.** Inferred average elongation rates. (**A**) Average elongation rates for four birth lengths (dots), for the DivIVA-labeled cells. The 2σ confidence intervals obtained by bootstrapping are indicated by the shaded areas. Vertical dashed lines: average onset of septum formation per birth length. (**B**) Average elongation rate trajectories for the wild-type cells, confidence intervals shown as in (**A**). Inset: average elongation trajectories as a function of the time until division. (**C**) Average elongation rate trajectories for the Δ*rodA* mutant, confidence intervals shown as in (**A**).

growing cells divide earlier than slower-growing cells (***Appendix 2—figure 2***) causing this conditional average to underestimate cellular elongation rates for the whole population after the first division. Because our aim is to infer elongation curves that characterize the whole population, ranging from slow to fast growers, for further analysis only the part of each trajectory before the first division event is used (***Figure 4D***). Sub-population elongation curves can also be obtained that extend past the first division event, but only if the entire analysis for these curves is performed only on these slower-dividing cells (***Appendix 2—figure 4***).

We obtain elongation rate curves by taking a numerical derivative of smoothed growth trajectories (Appendix 5). To determine the associated error margins of the elongation rates, we use a custom bootstrapping algorithm (***Efron, 1979***). The resulting 2σ bounds are shown as semitransparent bands. Despite the high noise level of individual elongation trajectories, the inferred average elongation rates have an error margin of around 8%. Thus, our approach robustly infers average elongation trajectories from single-cell growth data. Elongation rates of cells with larger birth length are consistently higher than the elongation rates of cells with smaller birth length. Strikingly, the elongation rate curves initially increase, but then gradually level off toward a linear growth mode (***Figure 5***). We note a slight difference in the cell elongation rates between the strain expressing DivIVA-mCherry (***Figure 5A***) and wild type cells (***Figure 5B***). Importantly, this difference does not qualitatively change the mode of growth, but does show that a translational fusion to DivIVA tends to lower elongation rates. This likely reflects a disturbance in the interaction between RodA or bifunctional PBPs and the DivIVA-mCherry fusion protein, indicating that the DivIVA-mCherry fusion is not fully functional. This is consistent with findings we reported earlier (***Donovan et al., 2013***).

To further test if the linear growth mode persists until division, we adapt our inference procedure to obtain average elongation curves $L(t-t_d,\ l_d)$ as a function of the time until division $t-t_d$ and division length $l_d$. The construction is analogous to that of $L(t,\ l_\mathrm{b})$ (Appendix 6). Calculating the corresponding elongation rate curves, we find that that linear growth indeed extends until the division time across division lengths (inset ***Figure 5B,SI***, ***Appendix 6—figure 1***). Note that with this construction, elongation rates become biased once $|t-t_d|$ exceeds the shortest single-cell total growth time. Hence, for our analysis we only consider elongation rate curves until this point.

To test the performance of our proposed inference method, we simulated a population of growing cells with a presumed growth mode from which we sample cells lengths as in our experiments, including measurement noise (Appendix 3). We ran simulations for cells performing linear growth, exponential growth, and the growth mode inferred here for DivIVA-labeled cells (***Figure 5A***). We find that our inference method is able to recover the input growth mode with high precision in all cases (Appendix 4, Appendix 7), demonstrating the accuracy and internal consistency of our inference method.

## Onset of the linear growth regime does not consistently coincide with septum formation

A central feature of the obtained elongation rate curves is a transition from an accelerating to a linear growth mode after approximately 20–25 min (***Figure 5***). One possibility is that this levelling off is connected with the onset of division septum formation. Given that the FtsZ-dependent divisome propagates the invagination of the septum under the consumption of cell wall precursors (e.g. Lipid-II), we hypothesized that the appearance of the additional sink for cell-wall building blocks could lead to coincidental leveling-off of the elongation rates (***Scheffers and Tol, 2015***). To test this hypothesis, we used the moment of a sharp increase in the average DivIVA-mCherry signal at the cell center as a proxy for the moment of onset of septum formation (***Appendix 2—figure 7***): the inward growing septum introduces a negative curvature of the plasma membrane, leading to the accumulation of DivIVA (***Lenarcic et al., 2009***; ***Strahl and Hamoen, 2012***). We observe that the onset of septum formation does not consistently coincide with the moment at which the elongation rate levels off (***Figure 5A***): for smaller cells, the onset of septum formation occurs much later. Therefore, it seems implausible that the observed linear growth regime is due the septum acting as a sink for cell-wall building blocks.

## Polar cell wall formation is the rate-limiting step for growth, leading to a linear growth regime

To provide insight into the anomalous single-cell growth behavior, we model single-cell elongation as being rate-limited by the apical cell wall formation mechanism. To formulate this rate-limiting apical growth (RAG) model, we first consider the biochemical pathway that leads to cell wall formation in *C. glutamicum*, as illustrated in ***Figure 1***. The key process for cell wall formation in *C. glutamicum* is polar peptidoglycan (PG) synthesis. PG intermediates are provided by the substrate Lipid-II, and the integration of new material into the PG-mesh is mediated by transglycosylases (TGs) located at the cell pole. At the TG sites, Lipid-II is translocated across the plasma membrane by the Lipid-II flippase MurJ (***Sham et al., 2014***; ***Kuk et al., 2017***; ***Butler et al., 2013***). After PG building blocks provided by Lipid-II are incorporated into the existing cell wall by transglycolylation, transpeptidases (TP) conduct the crosslinking of peptide subunits, which contributes to the rigidity of the cell wall (***Scheffers and Pinho, 2005***; ***Valbuena et al., 2007***; ***Schleifer and Kandler, 1972***). During growth, the area of the PG sacculus, and thus the number of TG sites, is extended by RodA and bifunctional penicillin binding proteins (PBPs), recruited by DivIVA (***Letek et al., 2008***; ***Sieger et al., 2013***).

To model this growth mechanism, we assume that the rate of new cell wall formation is proportional to the number of TG sites. We describe the interaction between Lipid-II and TG sites by Michaelis-Menten kinetics (***Figure 6A***). Specifically, if the cell length added per unit time is proportional to the cell wall area added per unit time, we find

$$\frac{dL(t)}{dt} = \alpha \frac{C(t)N(t)}{K_m + C(t)} \tag{1}$$

with $L(t)$ the cell length at time $t$, $C(t)$ the concentration of Lipid-II, $K_m$ the Michaelis constant for this reaction, and $\alpha$ is a proportionality constant.

To gain insight into the cell-cycle-dependence of $N(t)$ and $C(t)$, we made use of the cyan fluorescent D-alanine analogue HADA (see Materials and methods) to stain newly inserted peptidoglycan. Exponentially growing *C. glutamicum* cells were labeled with HADA for 5 min before imaging. The HADA stain will mainly appear at sites of nascent PG synthesis. As expected, HADA staining resulted in a bright cyan fluorescent signal at the cell poles and at the site of septation. Still images were obtained with fluorescence microscopy and subjected to image analysis (***Figures 2A*** and ***6B***, Materials and methods).

We first verify that the HADA intensity profile at the cell poles can be used as a measure for the peptidoglycan insertion rate. To do this, we assume that the HADA intensity profile has two relevant contributions: fluorescent probe present in the cell plasma, and fluorescent probe attached to newly inserted peptidoglycan. We use the minimum of the HADA intensity profile, consistently located around mid-cell in non-dividing cells, as an estimate of the contribution from the cell plasma in each cell, and subtract this from the entire cellular profile to obtain the corrected HADA profile (***Appendix 2—figure 8***). We then define the polar regions where we use the corrected HADA intensity to measure newly inserted peptidoglycan as the portions of the cell within 0.78 μm of the cell tips. Our results are, however, not strongly dependent on this polar region definition (***Appendix 2—figure 10***). Subsequently, we compute a moving average of the corrected polar HADA intensity as a function of cell length (***Figure 6C***). These polar HADA intensities are approximately proportional to the inferred average single-cell elongation rates (Appendix 8), as shown in the inset of ***Figure 6C***. Thus, the polar HADA intensities can be used as a measure for the cellular elongation rate. Assuming a proportional relationship between elongation rate and peptidoglycan insertion rate, this implies the polar HADA intensities are also approximately proportional to the peptidoglycan insertion rate. Deviations of ~10% from proportionality within the error margins observed over the range of tip intensities do not affect subsequent conclusions from the HADA intensity data.

Analyzing the HADA intensity profile for smaller segments within the polar region, we find that the increase in intensity is unevenly distributed (***Figure 6D***). Close to the cell tip, the HADA intensity remains approximately constant across cell lengths, whereas a linear increase over cell lengths is seen further from the tip. Considering the implications of these measured intensities for $C(t)$ and $N(t)$ within our model in ***Equation (1)***, we argue for a scenario where either $C(t)$ is constant or $C(t) \gg K_m$. Our reasoning is as follows. From ***Equation (1)***, we see that the approximately constant intensity at the cell tip can be produced in two ways: (1) $C(t) \gg K_m$ or $C(t)$ is constant across cell

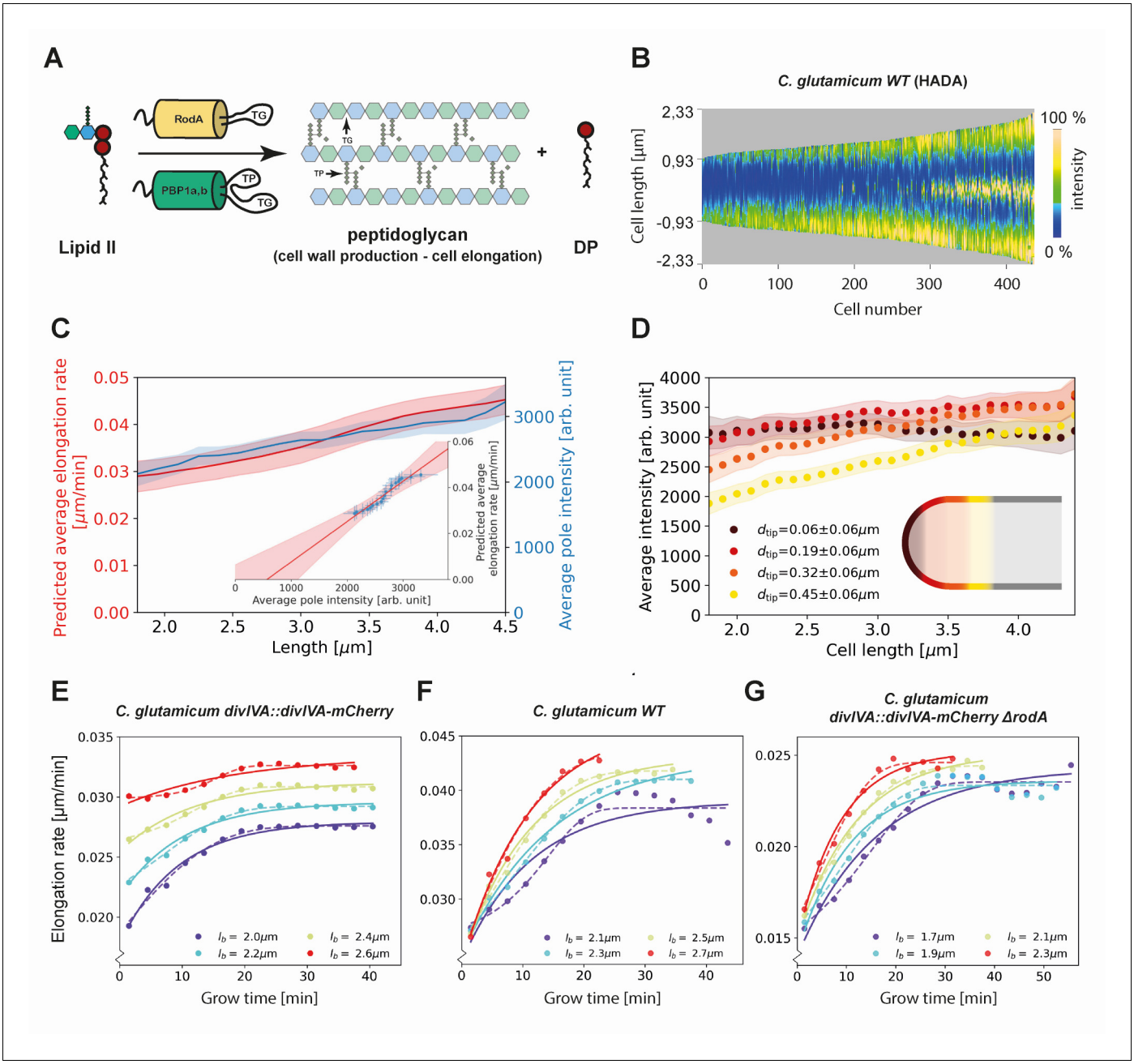



**Figure 6.** Modeling of average elongation rates using HADA staining results. (**A**) Schematic depicting cell wall formation via Lipid-II and transglycosylases (TG's). The corresponding Michaelis-Menten equation describes the change of length over time as function of the Lipid-II concentration $C(t)$ and the number of the TG sites $N(t)$. (**B**) Demograph of *C. glutamicum* cells stained with HADA. Cell are ordered by length, with the stronger signal oriented downwards. (**C**) Average elongation rate as a function of cell length (red), predicted from obtained average elongation rate curves (Appendix 8), together with the average HADA staining intensity at the cell pole after background correction (blue). The cell pole is defined here as the region within 0.77 µm (60 pixels) of the cell tip. The shaded regions indicate the 2XSEM bounds. For both curves, a moving average over cells within 0.7 µm of each x-coordinate is applied over the underlying data. Inset: predicted average elongation rate versus average HADA staining intensity (blue dots). A linear fit through the result (red line) is consistent with a proportional relationship. (**D**) Average HADA intensity as a function of cell length, shown for four regions close to the cell tip. A moving average over cells within 0.7 µm of each x-coordinate is applied over the underlying data. (**E-G**) Dots: average elongation rate curves as shown in ***Figure 5A***. Solid lines: best fit of elongation model from ***Equation (2)***, which assumes constant transglycosylase recruitment. Dashed lines: best fit of elongation model from ***Equation (3)***, which assumes an exponential increase of transglycosylase recruitment.

lengths, and the number of transglycosylases at the tip $N_{\mathrm{tip}}(t)$ is constant, or (2) $N_{\mathrm{tip}}(t)$ and $C(t)$ anti-correlate in such a way to produce constant insertion.

However, we consider constant $N_{\mathrm{tip}}(t)$ as biologically the most plausible scenario. This is supported by noting that the concentration of Lipid-II is the same directly before and after division, such that $C(t)$, and by implication $N_{\mathrm{tip}}(t)$, is similar for the shortest and the longest cell lengths (***Appendix 2—figure 9***). In our subsequent analysis, we will therefore assume that either $C(t)$ is constant, or $C(t) \gg K_m$. This implies that $\frac{dL(t)}{dt}$ in ***Equation (1)*** is directly proportional to $N(t)$.

To derive an expression for $N(t)$, we first note that the old and new cell pole in the cell need to be treated differently. We assume the number of polar TG-sites to saturate within one cellular life-cycle, such that the new pole initiates with $N(t)$ below saturation, while the old pole – inherited from the mother cell – is saturated. Letting the number of TG sites increase proportional to the number of available sites, we arrive at the following kinetic description for $N(t)$

$$\frac{dN(t)}{dt} = \beta(N^{\max} - N(t)) \tag{2}$$

Here, $N^{max}$ is the maximum number of sites at the cell poles, and $\beta$ is a rate constant. This result, together with ***Equation (1)***, defines our RAG model. The predicted elongation rates provide a good fit to the experiment for all studied genotypes (***Figure 6E–G***), although the data appear to exhibit a stronger inflection.

Instead of assuming a constant recruitment of TG enzymes, we can construct a more refined model that takes TG recruitment dynamics into account. There is evidence that transglycosylase RodA and PBPs are recruited to the cell pole via the curvature-sensing protein DivIVA (***Letek et al., 2008***; ***Sieger et al., 2013***). As shown in ***Lenarcic et al., 2009***, DivIVA also recruits itself, leading to the exponential growth of a nucleating DivIVA cluster. Therefore, we let the recruitment rate of TG enzymes be proportional to the number of DivIVA proteins $N_{\mathrm{D}}(t) = N_{\mathrm{D}}(0)e^{\gamma t}$. This results in a modified kinetic description for $N(t)$ (***Equation (2)***):

$$\frac{dN(t)}{dt} = \beta e^{\gamma t}(N^{\max} - N(t)) \tag{3}$$

This refined model can capture more detailed features of the measured elongation rate curves (***Figure 6E–G***), including the stronger inflection, with an additional free parameter, $\gamma$, encoding the self-recruitment rate of DivIVA.

The central assumption of our RAG model is that the growth of the cell poles, mediated via accumulation of TG enzymes, is the rate-limiting step for cellular growth. To test this assumption, we repeated our experiment with a *rodA* knockout (***Sieger et al., 2013***). The SEDS-protein RodA is a mono-functional TG (***Meeske et al., 2016***; ***Emami et al., 2017***; ***Sjodt et al., 2018***), whose deletion results in a phenotype with a decreased population growth rate in the shaking-flask (***Sieger et al., 2013***). The cells' viability is nonetheless backed up by the presence of bifunctional class A PBPs capable of catalyzing transglycoslyation and transpeptidation reactions. We expect this knockout to lower the efficiency of polar cell wall formation, thus slowing down the rate-limiting step of growth. Specifically, we expect the knockout of *rodA* to mainly affect the efficiency of Lipid-II integration into the murein sacculus. Within our RAG model, this translates to a lowering of the cell wall production per transglycosylase site $\alpha$. This would imply elongation rate curves of similar shape for the Δ*rodA* mutant, only scaled down by a factor $\alpha^{WT}/\alpha^{\Delta rodA}$. Indeed, we observe such a scaling down of the elongation rate curves (***Figure 5C***), lending further credence to our model for *C. glutamicum* growth.

A striking feature observed across growth conditions and birth lengths, is the onset of a linear growth regime after approximately 20 min (***Figure 5A–C***). The robustness of this timing can be understood from the RAG model: the regime of linear growth is reached via an exponential decay of the number of available TG sites until saturation is reached. This exponential decay makes the moment of onset of the linear growth regime relatively insensitive to variations in $N(0)$ and $N^{\max}$. Specifically, from ***Equation (2)***, it can be shown that the difference between $N(t)$ and $N^{\max}$ is halved every $\ln(2)\beta$ minutes, which amounts to ~8 min given fitted value of $\beta$ (***Appendix 9—table 1***).

Finally, our RAG model makes a prediction for the degree of transglycosylase saturation of the cell poles at birth, relative to the saturation in the linear growth regime. We find that this saturation

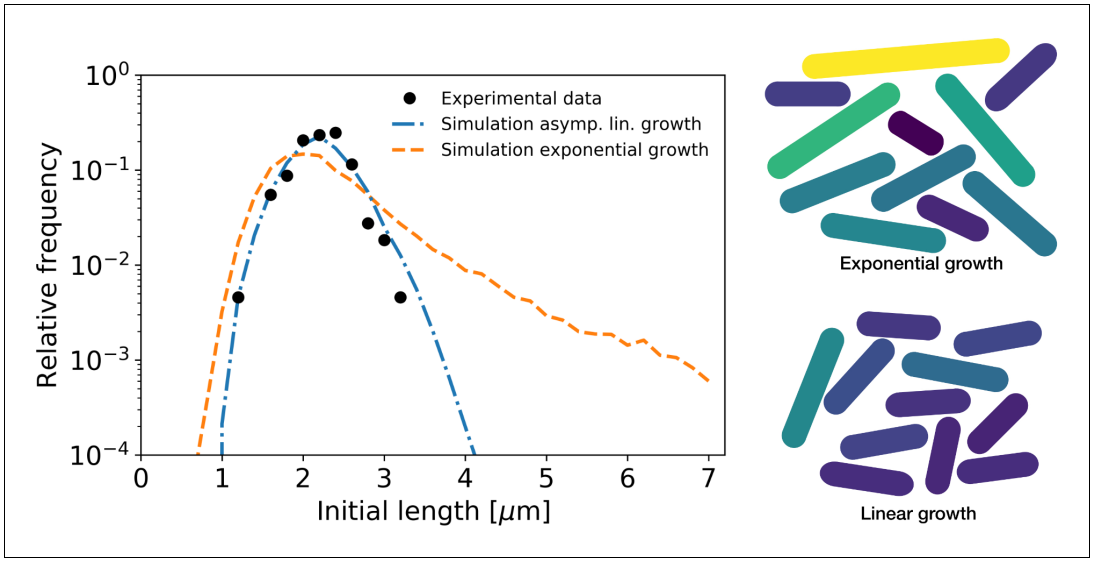



**Figure 7.** Simulation of population growth for asymptotically linear and exponential growth. Left: birth length distribution for simulated asymptotically linear growth (blue dash-dotted line), and for simulated exponential growth (orange dashed line). For both simulations, all relevant growth parameters and distributions are obtained directly from the experimental data. Black dots: experimental birth length distribution. Right: sample of 11 cells from the exponential and asymptotically linear growth simulations, color coded according to length.

is comparable between wild-type and the Δ*rodA* mutant (~65% on average), but significantly higher for DivIVA-labeled cells (~80% on average) (***Appendix 9—tables 1*** and ***2***). Note that the percentage of the saturation levels are relative values and do not suggest that in the DivIVA-mCherry fusion more transglycosylase sites are present in absolute numbers.

## Birth length distribution of linear growers is more robust to single-cell growth variability

After obtaining average single-cell growth trajectories, we next asked how this growth behavior at the single cell level affects the growth of the colony. It was shown that asymmetric division and noise in individual growth times results in a dramatic widening of the cell-size distribution for a purely exponential grower (***Marantan and Amir, 2016***). For an asymptotically linear grower, however, we would expect single-cell variations to have a much weaker impact.

To quantify the difference between asymptotically linear growth and hypothetical exponential growth for *C. glutamicum*, we performed population growth simulations for both cases. For the asymptotically linear growth, we assumed the elongation rate curves obtained from our model. For exponential growth, we assumed the final cell size to be given by $l_d = l_b \exp(\alpha(t_t + \Delta t)) + \Delta l$, with $\alpha$ the exponential elongation rate, $t_t$ the target growth time, $\Delta t$ a time-additive noise term and $\Delta l$ a size-additive noise term. All growth parameters necessary for the simulation were obtained directly from the experimental data (Appendix 10). From this simulation, the distribution of initial cell lengths was determined for each scenario.

The resulting distribution of birth lengths for the asymptotically linear growth case closely matches the experimentally determined distribution (***Figure 7***). By contrast, the distribution for exponential growth is much wider, and exhibits a broad tail for longer cell lengths. This suggests a strong connection between growth mode and the effect of individual growth variations on population statistics. *C. glutamicum* has a high degree of variation of division symmetry (***Appendix 10—figure 1C***) and single-cell growth times, but due to the asymptotically linear growth mode, the population-level variations in cell size are still relatively small. This indicates that linear growth can act as a regulator for cell size.

## Discussion

By developing a novel growth trajectory inference and analysis method, we showed that *C. glutamicum* exhibits asymptotically linear growth, rather than the exponential growth predominantly found in bacteria. The obtained elongation rate curves are shown to be consistent with a model of apical cell wall formation being the rate-limiting step for growth. The RAG model is further validated by experiments with a Δ*rodA* mutant, in which the elongation rate curves look functionally similar, but with a downward shift compared to wild type (***Figure 5B,C***), as expected based on our model. For *C. glutamicum*, apical cell wall formation is a plausible candidate for the rate-limiting step of growth, because synthesis of the highly complex cell wall and lipids for the mycolic acid membrane is cost intensive and a major sink for energy and carbon in *Corynebacteria* and *Mycobacteria* (***Brennan, 2003***).

An analysis of elongation rates as a function of time and birth length has previously been done in *B. subtilis* by binning cells based on birth length (***Nordholt et al., 2020***). Applying this method to our data set yields elongation rates averaged over cells within a binning interval (***Appendix 2—figure 5***). Averaging our inferred elongation rates over the same bins, we find the two methods to yield consistent results. The binning method, however, involves a tradeoff: a smaller bin width results in a larger error on the inferred elongation rates, whereas a larger bin width averages out all variation within a larger birth length interval. Our method does not suffer from this binning-related tradeoff, and it provides detailed elongation rate curves at any given birth length. In other recent work (***Kar et al., 2021***), average growth rate curves were calculated as a function of cell phase. Our method provides additional detail by extracting the dependence of elongation rate on birth length as well as time since birth.

Our proposed growth model shares some similar features to recent experimental observations on polar growth in Mycobacteria (***Hannebelle et al., 2020***). Polar growth was shown to follow 'new end take off' (NETO) dynamics (***Hannebelle et al., 2020***), in which the new cell pole makes a sudden transition from slow to fast growth, leading to a bilinear polar growth mode. In our proposed growth model for *C. glutamicum* however, the new pole gradually increases its average elongation rate before saturating to a constant maximum. The deviation of *C. glutamicum* from NETO dynamics can also be seen by comparing each of the pole intensities in the HADA staining experiment, which does not show any signatures of NETO-like growth (***Appendix 2—figure 11***). It remains unclear which molecular mechanisms produce the differences in growth between such closely related species. However, the mode of growth described here for *C. glutamicum* might well be an adaption to enable higher growth rates.

To investigate the implications of our inferred single-cell growth mode for cell-size homeostasis throughout a population of cells, we performed simulations of cellular growth and division over many generations. We found that our asymptotically linear growth model accurately reproduces the experimental distribution of cell birth lengths. By contrast, a model of exponential growth predicts a much broader distribution with a long tail for larger birth lengths. This indicates a possible connection between mode of growth and permissible growth-related noise levels for the cell. Indeed, if single-cell growth variability is reduced by a factor 3, the distributions corresponding to both growth modes show a similarly narrow width (***Appendix 10—figure 2***). However, an asymptotically linear grower is able to maintain a narrow distribution of cell sizes even for higher noise levels, whereas for an exponential grower this distribution widens dramatically (***Figure 7***).

The enhanced robustness of the length distribution of linear growers is interesting from an evolutionary point of view. Most rod-shaped bacteria use sophisticated systems, such as the Min system, to ensure cytokinesis precisely at midcell (***Bramkamp et al., 2009***; ***Lutkenhaus, 2007***). Bacteria encoding a Min system grow by lateral cell wall insertion. In contrast, rod-shaped bacteria in the *Actinobacteria* phylum such as *Mycobacterium* or *Corynebacterium* species, grow apically and do not contain a Min system, nor any other known division site selection system (***Donovan and Bramkamp, 2014***). *C. glutamicum* rather couples division site selection to nucleoid positioning after chromosome segregation via the ParAB partitioning system (***Donovan et al., 2013***), and has a broader distribution of division symmetries. We speculate that due to *C. glutamicum's* distinct growth mechanism, a more precise division site selection mechanism is not necessary to maintain a narrow cell size distribution.

The elongation rates reported in this work reflect the increase in cellular volume over time. However, the increase in cell *mass* is not necessarily proportional to cellular volume. In exponentially growing *E. coli*, the cellular density was recently reported to systematically vary during the cell cycle, while the surface-to-mass ratio was reported to remain constant (**Oldewurtel et al., 2019**). It is unknown how single-cell mass increases in *C. glutamicum*, but it would follow exponential growth if mass production is proportional to protein content. This raises the question how linear volume growth and exponential mass growth are coordinated. The presence of a regulatory mechanism for cell mass production that couples to cell volume is implied by the elongation rate curves obtained for the Δ*rodA* mutant. As the elongation rate is lower in this mutant, average mass production needs to be lowered compared to the WT in order to prevent the cellular density from increasing indefinitely.

Our growth trajectory inference method is not cell-type specific, and can be used to obtain detailed growth dynamics in a wide range of organisms. The inferred asymptotically linear growth of *C. glutamicum* deviates from the predominantly found exponential single-cell bacterial growth, and suggests the presence of novel growth regulatory mechanisms.

# Materials and methods

### Key resources table

| Reagent type (species) or resource | Designation | Source or reference | Identifiers | Additional information |
|---|---|---|---|---|
| Gene (include species here) | 'divIVA'; 'rodA' | KEGG | 'cg2361'; 'cg0061' | |
| Strain, strain background (*Corynebacterium glutamicum*) | 'ATCC 13032'; 'RES 167' | 'ATCC'; '**Tauch et al., 2002**' | '13032';"RES 167' | |
| Genetic reagent (*Corynebacterium glutamicum*) | 'RES 167 divIVA::divIVA-mCherry';"RES 167 Δ rodA, divIVA::divIVA-mCherry' | '**Donovan et al., 2012**'; '**Sieger et al., 2013**' | 'CDC010'; 'BSC002' | |
| Chemical compound, drug | HADA stain | Tocris Bioscience | 6647/5 | |
| Software, algorithm | MorpholyzerGT | This paper | | see Materials and methods |
| Other | CellASIC microfluidic System | Millipore | B04A | |

## Culture and live-cell time-lapse imaging

Exponentially growing cells of *C. glutamicum WT*, *C. glutamicum divIVA::divIVA-mCherry* and *C. glutamicum divIVA::divIVA-mCherry* Δ*rodA* respectively, grown in BHI–medium (Oxoid) at 30˚C and 200 rpm shaking, were diluted to an $OD_{600}$ of 0.01. According to the manufacturer's manual cells were loaded into a CellASIC- microfluidic plate type B04A (Merck Milipore) and mounted on a Delta Vision Elite microscope (GE Healthcare, Applied Precision) with a standard four-color InSightSSI module and an environmental chamber heated to 30˚C. Images were taken in a three-minute interval for 10 hr with a 100×/1.4 oil PSF U-Plan S-Apo objective and a DS-red-specific filter set (32% transmission, 0.025 s exposure).

## Staining of newly inserted peptidoglycan and visualization in demographs

For the staining of nascent PG, 1 ml of exponentially growing *C. glutamicum ATCC 13032* cells, cultivated in BHI–medium (Oxoid) at 30˚C and 200 rpm, were harvested, washed with PBS and resuspended in 25 µl PBS, together with 0.25 µl of 5 mM HADA dissolved in DMSO. The cells were incubated at 30˚C in the dark for 5 min, followed by a two-time washing step with 1 ml PBS and finally resuspended in 100 µl PBS. To obtain still- phase-contrast and fluorescent micrographs, 2 µl of the cell suspension were immobilized on an agarose pad. For microscopy, an Axio Imager (Zeiss) equipped with EC Plan-Neofluar 100x/1.3 Oil Ph3 objective and a Axiocam camera (Zeiss) was used together with the appropriate filter sets (ex: 405 nm; em: 450 nm). For single-cell analysis and the

visualization in demographs, custom algorithms, developed in FIJI and R (***Schindelin et al., 2012***; ***R Development Core Team, 2003***), were used. The code is available upon request.

### Image analysis

For image analysis, a custom-made algorithm was developed using the open-source programs FIJI and R (***Schindelin et al., 2012***; ***R Development Core Team, 2003***). During the workflow unique identifiers to single-cell cycles are assigned. The cell outlines are determined manually. Individual cells per timeframe are extracted then from the raw image and further processed automatically. The parameters length, area and relative septum position are extracted and stored together with the genealogic information and the timepoint within the respective cell cycle. The combination of image analysis and cell cycle dependent data structuring yields a list that serves as a base for further analysis. The documented code is available at: https://github.com/Morpholyzer/MorpholyzerGenerationTracker (copy archived at swh:1:rev:d01d362ea53b9be6027f29fb85668a0ed418398a, ***Morpholyzer, 2021***).

## Acknowledgements

This work was further funded by grants from the Deutsche Forschungsgemeinschaft (project P05in TRR174, granted to MB and project P06 in TRR174, granted to CB). JM is supported by a DFG fellowship within the Graduate School of Quantitative Biosciences Munich (QBM). We thank our colleagues from CB and MB groups for discussions, feedback and comments on the manuscripts.

## Additional information

### Funding


| Funder | Grant reference number | Author |
|---|---|---|
| Ludwig-Maximilians-Universität München | Graduate Student Stipend | Joris JB Messelink |
| Deutsche Forschungsgemeinschaft | TRR 174 project P06 | Joris JB Messelink<br>Chase P Broedersz |
| Deutsche Forschungsgemeinschaft | TRR 174 project P05 | Fabian Meyer<br>Marc Bramkamp |

The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.


### Author contributions

Joris JB Messelink, Software, Formal analysis, Investigation, Visualization, Writing - original draft, Writing - review and editing; Fabian Meyer, Data curation, Investigation, Visualization, Methodology, Writing - original draft, Writing - review and editing; Marc Bramkamp, Chase P Broedersz, Conceptualization, Supervision, Writing - original draft, Writing - review and editing

### Author ORCIDs

Joris JB Messelink https://orcid.org/0000-0002-7986-4527
Fabian Meyer https://orcid.org/0000-0002-8305-0390
Marc Bramkamp https://orcid.org/0000-0002-7704-3266
Chase P Broedersz https://orcid.org/0000-0001-7283-3704

### Decision letter and Author response

Decision letter https://doi.org/10.7554/eLife.70106.sa1
Author response https://doi.org/10.7554/eLife.70106.sa2

## Additional files

### Supplementary files

- Source data 1. HADA staining data.
- Source data 2. Elongation measurement data.
- Transparent reporting form

### Data availability

All data generated during this study are included in the manuscript and supporting files.

## References


**Alderwick LJ**, Harrison J, Lloyd GS, Birch HL. 2015. The Mycobacterial Cell Wall–Peptidoglycan and Arabinogalactan. *Cold Spring Harbor Perspectives in Medicine* **5**:a021113. DOI: https://doi.org/10.1101/cshperspect.a021113, PMID: 25818664

**Amir A**. 2014. Cell size regulation in bacteria. *Physical Review Letters* **112**:208102. DOI: https://doi.org/10.1103/PhysRevLett.112.208102

**Antoine I**, Coene M, Cocito C. 1988. Size and homology of the genomes of leprosy-derived corynebacteria, *Mycobacterium leprae*, and other corynebacteria and mycobacteria. *Journal of Medical Microbiology* **27**:45–50. DOI: https://doi.org/10.1099/00222615-27-1-45, PMID: 3050108

**Belliveau NM**, Chure G, Hueschen CL, Garcia HG, Kondev J, Fisher DS, Theriot JA, Phillips R. 2020. Fundamental limits on the rate of bacterial growth. *bioRxiv*. DOI: https://doi.org/10.1101/2020.10.18.344382

**Bramkamp M**, van Baarle S, Baarle Svan. 2009. Division site selection in rod-shaped bacteria. *Current Opinion in Microbiology* **12**:683–688. DOI: https://doi.org/10.1016/j.mib.2009.10.002, PMID: 19884039

**Brennan PJ**. 2003. Structure, function, and biogenesis of the cell wall of *Mycobacterium tuberculosis*. *Tuberculosis* **83**:91–97. DOI: https://doi.org/10.1016/s1472-9792(02)00089-6, PMID: 12758196

**Buchanan RE**. 1918. Life phases in a bacterial culture. *Journal of Infectious Diseases* **23**:109–125. DOI: https://doi.org/10.1086/infdis/23.2.109

**Butler EK**, Davis RM, Bari V, Nicholson PA, Ruiz N. 2013. Structure-function analysis of MurJ reveals a solvent-exposed cavity containing residues essential for peptidoglycan biogenesis in *Escherichia coli*. *Journal of Bacteriology* **195**:4639–4649. DOI: https://doi.org/10.1128/JB.00731-13, PMID: 23935042

**Cooper S**. 1998. Length extension in growing yeast: is growth exponential?–yes. *Microbiology* **144 (Pt 2)**:263–265. DOI: https://doi.org/10.1099/00221287-144-2-263, PMID: 9493363

**Daniel RA**, Errington J. 2003. Control of cell morphogenesis in bacteria: two distinct ways to make a rod-shaped cell. *Cell* **113**:767–776. DOI: https://doi.org/10.1016/s0092-8674(03)00421-5, PMID: 12809607

**Donovan C**, Sieger B, Krämer R, Bramkamp M. 2012. A synthetic Escherichia coli system identifies a conserved origin tethering factor in actinobacteria. *Molecular Microbiology* **84**:105–116. DOI: https://doi.org/10.1111/j.1365-2958.2012.08011.x, PMID: 22340668

**Donovan C**, Schauss A, Krämer R, Bramkamp M. 2013. Chromosome segregation impacts on cell growth and division site selection in *Corynebacterium glutamicum*. *PLOS ONE* **8**:e55078. DOI: https://doi.org/10.1371/journal.pone.0055078, PMID: 23405112

**Donovan C**, Bramkamp M. 2014. Cell division in Corynebacterineae. *Frontiers in Microbiology* **5**:132. DOI: https://doi.org/10.3389/fmicb.2014.00132, PMID: 24782835

**Efron B**. 1979. Bootstrap methods: another look at the jackknife. *The Annals of Statistics* **7**:1–26. DOI: https://doi.org/10.1214/aos/1176344552

**Emami K**, Guyet A, Kawai Y, Devi J, Wu LJ, Allenby N, Daniel RA, Errington J. 2017. RodA as the missing glycosyltransferase in *Bacillus subtilis* and antibiotic discovery for the peptidoglycan polymerase pathway. *Nature Microbiology* **2**:16253. DOI: https://doi.org/10.1038/nmicrobiol.2016.253, PMID: 28085152

**Godin M**, Delgado FF, Son S, Grover WH, Bryan AK, Tzur A, Jorgensen P, Payer K, Grossman AD, Kirschner MW, Manalis SR. 2010. Using buoyant mass to measure the growth of single cells. *Nature Methods* **7**:387–390. DOI: https://doi.org/10.1038/nmeth.1452, PMID: 20383132

**Hannebelle MTM**, Ven JXY, Toniolo C, Eskandarian HA, Vuaridel-Thurre G, McKinney JD, Fantner GE. 2020. A biphasic growth model for cell pole elongation in mycobacteria. *Nature Communications* **11**:452. DOI: https://doi.org/10.1038/s41467-019-14088-z, PMID: 31974342

**Hermann T**. 2003. Industrial production of amino acids by coryneform bacteria. *Journal of Biotechnology* **104**:155–172. DOI: https://doi.org/10.1016/s0168-1656(03)00149-4, PMID: 12948636

**Hett EC**, Rubin EJ. 2008. Bacterial growth and cell division: a mycobacterial perspective. *Microbiology and Molecular Biology Reviews* **72**:126–156. DOI: https://doi.org/10.1128/MMBR.00028-07, PMID: 18322037

**Iyer-Biswas S**, Wright CS, Henry JT, Lo K, Burov S, Lin Y, Crooks GE, Crosson S, Dinner AR, Scherer NF. 2014. Scaling laws governing stochastic growth and division of single bacterial cells. *PNAS* **111**:15912–15917. DOI: https://doi.org/10.1073/pnas.1403232111, PMID: 25349411

**Jun S**, Taheri-Araghi S. 2015. Cell-size maintenance: universal strategy revealed. *Trends in Microbiology* **23**:4–6. DOI: https://doi.org/10.1016/j.tim.2014.12.001, PMID: 25497321

**Kar P**, Tiruvadi-Krishnan S, Männik J, Männik J, Amir A. 2021. To bin or not to bin: analyzing Single-Cell growth data. *bioRxiv*. DOI: https://doi.org/10.1101/2021.07.27.453901

**Kuk AC**, Mashalidis EH, Lee SY. 2017. Crystal structure of the MOP flippase MurJ in an inward-facing conformation. *Nature Structural and Molecular Biology* **24**:171–176. DOI: https://doi.org/10.1038/nsmb.3346, PMID: 28024149

**Lane-Claypon JE**. 1909. Multiplication of Bacteria and the Influence of Temperature and some other Conditions thereon. *The Journal of Hygiene* **9**:239–248. DOI: https://doi.org/10.1017/s0022172400016260, PMID: 20474395

**Lenarcic R**, Halbedel S, Visser L, Shaw M, Wu LJ, Errington J, Marenduzzo D, Hamoen LW. 2009. Localisation of DivIVA by targeting to negatively curved membranes. *The EMBO Journal* **28**:2272–2282. DOI: https://doi.org/10.1038/emboj.2009.129, PMID: 19478798

**Letek M**, Ordóñez E, Vaquera J, Margolin W, Flärdh K, Mateos LM, Gil JA. 2008. DivIVA is required for polar growth in the MreB-lacking rod-shaped actinomycete *Corynebacterium glutamicum*. *Journal of Bacteriology* **190**:3283–3292. DOI: https://doi.org/10.1128/JB.01934-07, PMID: 18296522

**Logsdon MM**, Ho PY, Papavinasasundaram K, Richardson K, Cokol M, Sassetti CM, Amir A, Aldridge BB. 2017. A Parallel Adder Coordinates Mycobacterial Cell-Cycle Progression and Cell-Size Homeostasis in the Context of Asymmetric Growth and Organization. *Current Biology* **27**:3367–3374. DOI: https://doi.org/10.1016/j.cub.2017.09.046, PMID: 29107550

**Long Z**, Nugent E, Javer A, Cicuta P, Sclavi B, Cosentino Lagomarsino M, Dorfman KD. 2013. Microfluidic chemostat for measuring single cell dynamics in bacteria. *Lab on a Chip* **13**:947–954. DOI: https://doi.org/10.1039/c2lc41196b, PMID: 23334753

**Lutkenhaus J**. 2007. Assembly dynamics of the bacterial MinCDE system and spatial regulation of the Z ring. *Annual Review of Biochemistry* **76**:539–562. DOI: https://doi.org/10.1146/annurev.biochem.75.103004.142652, PMID: 17328675

**Marantan A**, Amir A. 2016. Stochastic modeling of cell growth with symmetric or asymmetric division. *Physical Review E* **94**:1–18. DOI: https://doi.org/10.1103/PhysRevE.94.012405

**Meeske AJ**, Riley EP, Robins WP, Uehara T, Mekalanos JJ, Kahne D, Walker S, Kruse AC, Bernhardt TG, Rudner DZ. 2016. SEDS proteins are a widespread family of bacterial cell wall polymerases. *Nature* **537**:634–638. DOI: https://doi.org/10.1038/nature19331, PMID: 27525505

**Mir M**, Wang Z, Shen Z, Bednarz M, Bashir R, Golding I, Prasanth SG, Popescu G. 2011. Optical measurement of cycle-dependent cell growth. *PNAS* **108**:13124–13129. DOI: https://doi.org/10.1073/pnas.1100506108, PMID: 21788503

**Monod J**. 1949. The growth of bacterial cultures. *Annual Review of Microbiology* **3**:371–394. DOI: https://doi.org/10.1146/annurev.mi.03.100149.002103

**Morpholyzer**. 2021. GenerationTracker. *Software Heritage*. swh:1:rev:d01d362ea53b9be6027f29fb85668a0ed418398a. https://archive.softwareheritage.org/swh:1:dir:ad7989d68e72f9d96279597f5f0b5e174b61c86e;origin=https://github.com/Morpholyzer/MorpholyzerGenerationTracker;visit=swh:1:snp:e702272c2f8ff27d5065a1ebd9022d6e0017d79f;anchor=swh:1:rev:d01d362ea53b9be6027f29fb85668a0ed418398a

**Nordholt N**, van Heerden JH, Bruggeman FJ. 2020. Biphasic Cell-Size and Growth-Rate Homeostasis by Single *Bacillus subtilis* Cells. *Current Biology* **30**:2238–2247. DOI: https://doi.org/10.1016/j.cub.2020.04.030, PMID: 32413303

**Oldewurtel ER**, Kitahara Y, Cordier B, Özbaykal G, Teeffelen S. 2019. Bacteria control cell volume by coupling Cell-Surface expansion to Dry-Mass growth. *bioRxiv*. DOI: https://doi.org/10.1101/769786

**Puech V**, Chami M, Lemassu A, Lanéelle MA, Schiffler B, Gounon P, Bayan N, Benz R, Daffé M. 2001. Structure of the cell envelope of corynebacteria: importance of the non-covalently bound lipids in the formation of the cell wall permeability barrier and fracture plane. *Microbiology* **147**:1365–1382. DOI: https://doi.org/10.1099/00221287-147-5-1365, PMID: 11320139

**R Development Core Team**. 2003. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/

**Scheffers DJ**, Pinho MG. 2005. Bacterial cell wall synthesis: new insights from localization studies. *Microbiology and Molecular Biology Reviews* **69**:585–607. DOI: https://doi.org/10.1128/MMBR.69.4.585-607.2005, PMID: 16339737

**Scheffers DJ**, Tol MB. 2015. LipidII: just another brick in the wall? *PLOS Pathogens* **11**:e1005213. DOI: https://doi.org/10.1371/journal.ppat.1005213, PMID: 26679002

**Schindelin J**, Arganda-Carreras I, Frise E, Kaynig V, Longair M, Pietzsch T, Preibisch S, Rueden C, Saalfeld S, Schmid B, Tinevez JY, White DJ, Hartenstein V, Eliceiri K, Tomancak P, Cardona A. 2012. Fiji: an open-source platform for biological-image analysis. *Nature Methods* **9**:676–682. DOI: https://doi.org/10.1038/nmeth.2019, PMID: 22743772

**Schleifer KH**, Kandler O. 1972. Peptidoglycan types of bacterial cell walls and their taxonomic implications. *Bacteriological Reviews* **36**:407–477. DOI: https://doi.org/10.1128/br.36.4.407-477.1972, PMID: 4568761

**Schubert K**, Sieger B, Meyer F, Giacomelli G, Böhm K, Rieblinger A, Lindenthal L, Sachs N, Wanner G, Bramkamp M. 2017. The Antituberculosis Drug Ethambutol Selectively Blocks Apical Growth in CMN Group Bacteria. *mBio* **8**:e02213-16. DOI: https://doi.org/10.1128/mBio.02213-16, PMID: 28174310

**Scott M**, Gunderson CW, Mateescu EM, Zhang Z, Hwa T. 2010. Interdependence of cell growth and gene expression: origins and consequences. *Science* **330**:1099–1102. DOI: https://doi.org/10.1126/science.1192588, PMID: 21097934

**Sham LT**, Butler EK, Lebar MD, Kahne D, Bernhardt TG, Ruiz N. 2014. Bacterial cell wall. MurJ is the flippase of lipid-linked precursors for peptidoglycan biogenesis. *Science* **345**:220–222. DOI: https://doi.org/10.1126/science.1254522, PMID: 25013077

**Sieger B**, Schubert K, Donovan C, Bramkamp M. 2013. The lipid II flippase RodA determines morphology and growth in *Corynebacterium glutamicum*. *Molecular Microbiology* **90**:966–982. DOI: https://doi.org/10.1111/mmi.12411, PMID: 24118443

**Sjodt M**, Brock K, Dobihal G, Rohs PDA, Green AG, Hopf TA, Meeske AJ, Srisuknimit V, Kahne D, Walker S, Marks DS, Bernhardt TG, Rudner DZ, Kruse AC. 2018. Structure of the peptidoglycan polymerase RodA resolved by evolutionary coupling analysis. *Nature* **556**:118–121. DOI: https://doi.org/10.1038/nature25985, PMID: 29590088

**Strahl H**, Hamoen LW. 2012. Finding the corners in a cell. *Current Opinion in Microbiology* **15**:731–736. DOI: https://doi.org/10.1016/j.mib.2012.10.006, PMID: 23182676

**Taheri-Araghi S**, Bradde S, Sauls JT, Hill NS, Levin PA, Paulsson J, Vergassola M, Jun S. 2015. Cell-size control and homeostasis in bacteria. *Current Biology* **25**:385–391. DOI: https://doi.org/10.1016/j.cub.2014.12.009, PMID: 25544609

**Tauch A**, Kirchner O, Löffler B, Götker S, Pühler A, Kalinowski J. 2002. Efficient electrotransformation of corynebacterium diphtheriae with a mini-replicon derived from the corynebacterium glutamicum plasmid pGA1. *Current Microbiology* **45**:362–367. DOI: https://doi.org/10.1007/s00284-002-3728-3, PMID: 12232668

**Valbuena N**, Letek M, Ordóñez E, Ayala J, Daniel RA, Gil JA, Mateos LM. 2007. Characterization of HMW-PBPs from the rod-shaped actinomycete *Corynebacterium glutamicum*: peptidoglycan synthesis in cells lacking actin-like cytoskeletal structures. *Molecular Microbiology* **66**:643–657. DOI: https://doi.org/10.1111/j.1365-2958.2007.05943.x, PMID: 17877698

**Yang D**, Jennings AD, Borrego E, Retterer ST, Männik J. 2018. Analysis of Factors Limiting Bacterial Growth in PDMS Mother Machine Devices. *Frontiers in Microbiology* **9**:1–12. DOI: https://doi.org/10.3389/fmicb.2018.00871, PMID: 29765371

**Yu FB**, Willis L, Chau RM, Zambon A, Horowitz M, Bhaya D, Huang KC, Quake SR. 2017. Long-term microfluidic tracking of coccoid cyanobacterial cells reveals robust control of division timing. *BMC Biology* **15**:1–14. DOI: https://doi.org/10.1186/s12915-016-0344-4, PMID: 28196492

**Zhou X**, Rodriguez-Rivera FP, Lim HC, Bell JC, Bernhardt TG, Bertozzi CR, Theriot JA. 2019. Sequential assembly of the septal cell envelope prior to V snapping in *Corynebacterium glutamicum*. *Nature Chemical Biology* **15**:221–231. DOI: https://doi.org/10.1038/s41589-018-0206-1, PMID: 30664686

# Appendix 1

## Single-cell growth mode for apical cell wall formation as a rate-limiting step for growth

To study growth limited by polar cell wall formation, we start by considering the Michaelis-Menten equation describing this formation process (Main Text ***Equation (1)***):

$$\frac{dL(t)}{dt} = \alpha \frac{C(t)N(t)}{K_m + C(t)}, \tag{A1}$$

with $L(t)$ the cell length at time $t$, $C(t)$ the concentration of cell wall building blocks in the cytosol, $N(t)$ the number of transglycosylases at the cell pole, $K_m$ the Michaelis constant for this reaction, and $\alpha$ a proportionality constant.

In Main Text ***Figure 1***, we consider two scenarios. (1) Abundant availability of cell wall building blocks, that is $C(t) \gg K_m$, and (2) scarcity of cell wall building blocks, that is, $C(t) < K_m$.

### A1.1 Building block insertion as a rate-limiting step for growth

In scenario (1), ***Equation (A1)*** reduces to $\frac{dL(t)}{dt} = \alpha N(t)$. In the regime of a constant number of transglycosylases at the pole, this implies that $\frac{dL(t)}{dt}$ is constant, resulting in linear growth.

### A1.2 Building block availability as a rate-limiting step for growth

In scenario (2), the dynamics of building block creation, usage, and dilution need to be considered to determine the cellular elongation rate behavior. For the number of building blocks in the cytosol as a function of time $n(t)$, we can write the following differential equation:

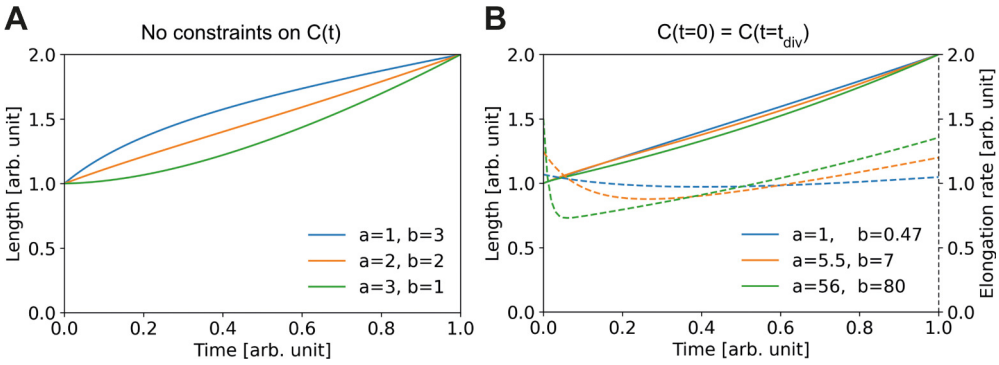
$$\frac{dn(t)}{dt} = aV(t) - b\frac{dV(t)}{dt}. \tag{A2}$$

Here, $a$ encodes building block production rate per unit volume, and $b$ encodes building block usage by the cell wall formation mechanism, making use of $\frac{dA(t)}{dt} \propto \frac{dV(t)}{dt}$. To connect ***Equation (A2)*** to ***Equation (A1)***, we note that $C(t) = \frac{n(t)}{V(t)}$. Restricting ourselves to the regime $C(t) \ll K_m$, we can rewrite ***Equation (A1)*** to

$$\frac{dV(t)}{dt} = c\frac{n(t)}{V(t)}, \tag{A3}$$

where we made use of $\frac{dL(t)}{dt} \propto \frac{dV(t)}{dt}$. Here, $c$ encodes the proportionality between volume increase and the concentration of building blocks.

Combining ***Equation (A2)*** with ***Equation (A3)***, we obtain a set of coupled nonlinear differential equations governing the time-evolution of $V(t)$. These equations have no simple analytic solution; however, we can numerically explore the dependence of $V(t)$ on the differential equation parameters. To do this, we first absorb $c$ into $n(t)$, leaving us with two free parameters and two boundary conditions. The boundary conditions we set by imposing $V(0) = 1$ and $V(1) = 2$. In ***Appendix 1—figure 1A***, we see that depending on the choice for $a$ and $b$ we can have either sublinear, approximately linear, or superlinear growth. This demonstrates that the single-cell growth mode is dependent on the physiology of building block creation and depletion in the cell.

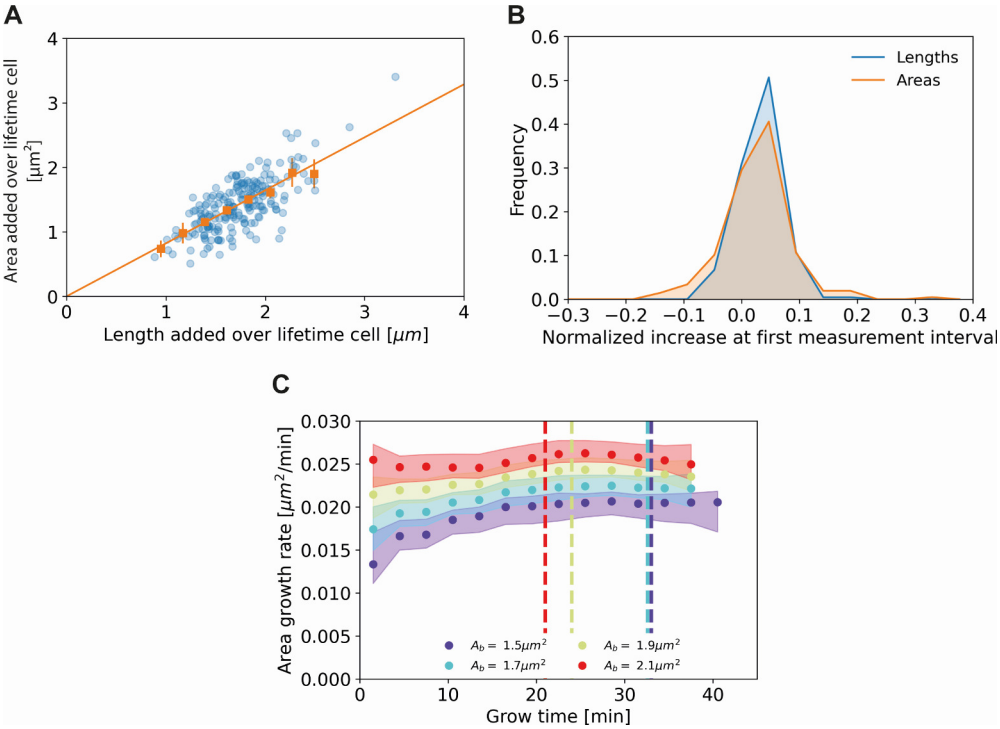



**Appendix 1—figure 1.** Elongation curves assuming building block availability is the limiting step for growth. (**A**) Numerically obtained solutions for $V(t)$, from the set of coupled differential equations ***Equation (A2)*** and ***Equation (A3)***. For all solutions, $V(0) = 1$ and $V(1) = 2$ are imposed. (**B**) Solutions as in (**A**), but with the additional constraint that the concentrations before and after division are the same, i.e. $C(t = 0) = C(t = t_{\mathrm{div}})$. Solid lines: solutions for $V(t)$. Dashed lines: corresponding $\frac{dV(t)}{dt}$, which are proportional to the concentration $C(t)$ per ***Equation (A3)***.

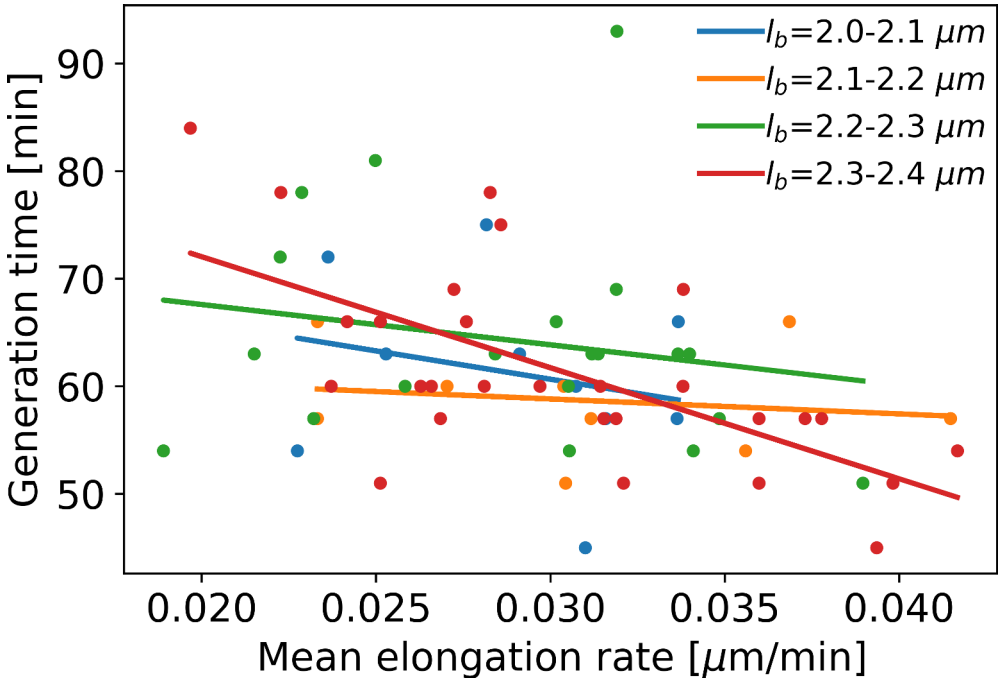
We can further constrain the solution space by demanding that the concentration of building blocks $C(t) = \frac{n(t)}{V(t)}$ is the same at birth and division. In this scenario, the observed variation in elongation curves is smaller (solid lines ***Appendix 1—figure 1B***), however the corresponding elongation rates (dashed lines ***Appendix 1—figure 1B***) still show marked qualitative differences between parameter choices.

# Appendix 2

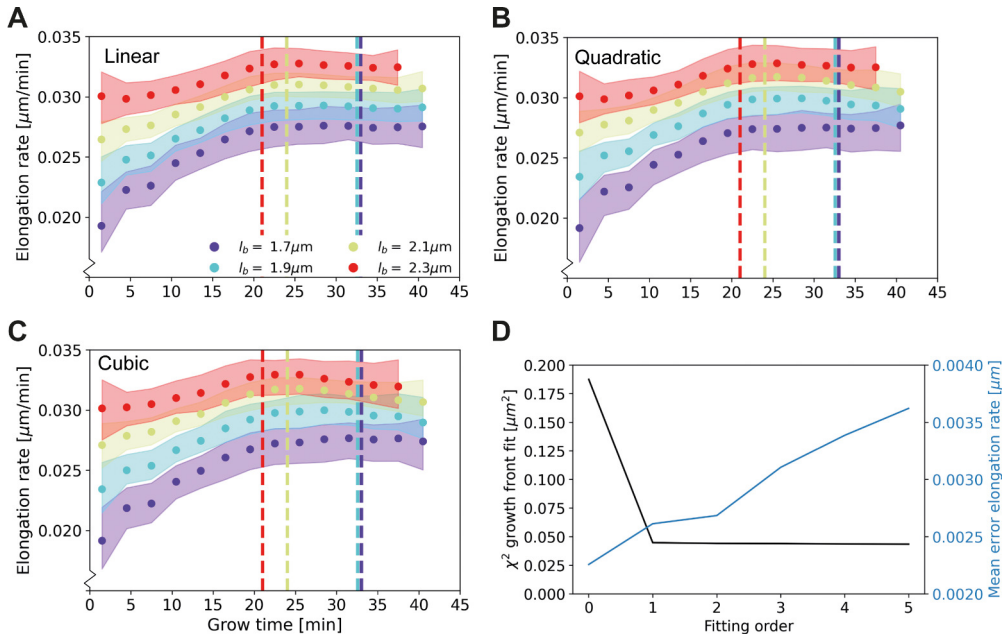
## Supplementary figures




**Appendix 2—figure 1.** Comparing cell length and cell area measurements. (**A**) Length added versus area added over the cell lifetime for all cells included in our analysis (blue dots), together with averaged values at 0.2 μm intervals (orange squares) and 95% confidence intervals (orange vertical lines). The results are consistent with a proportional relationship (orange line). (**B**) Histogram of the normalized increase at first measurement interval using cell lengths (blue) and areas (orange). For the cellular lengths, this quantity is defined as $\frac{L(t=3min)-L(t=0)}{\langle l_b \rangle}$, whereas for the areas it is defined as $\frac{A(t=3min)-A(t=0)}{\langle A_b \rangle}$, with $A(t)$ the area at time $t$ and $A_b$ the birth area. The wider distribution for the areas suggests a higher measurement noise for this quantity. (**C**) Area growth rate for DivIVA-labeled cells using estimated cell areas. The trajectories are consistent with those obtained from cell lengths (Main Text ***Figure 5A***).




**Appendix 2—figure 2.** Mean elongation rate versus generation time for cells in four different birth size bins. Linear fits are indicated by solid lines. As generation times within a birth size bin tend to be shorter for faster-growing cells, the elongation rate curves obtained with our method become biased after the first division event. This justifies only using the part of the elongation rate curves until the first division event for further analysis.

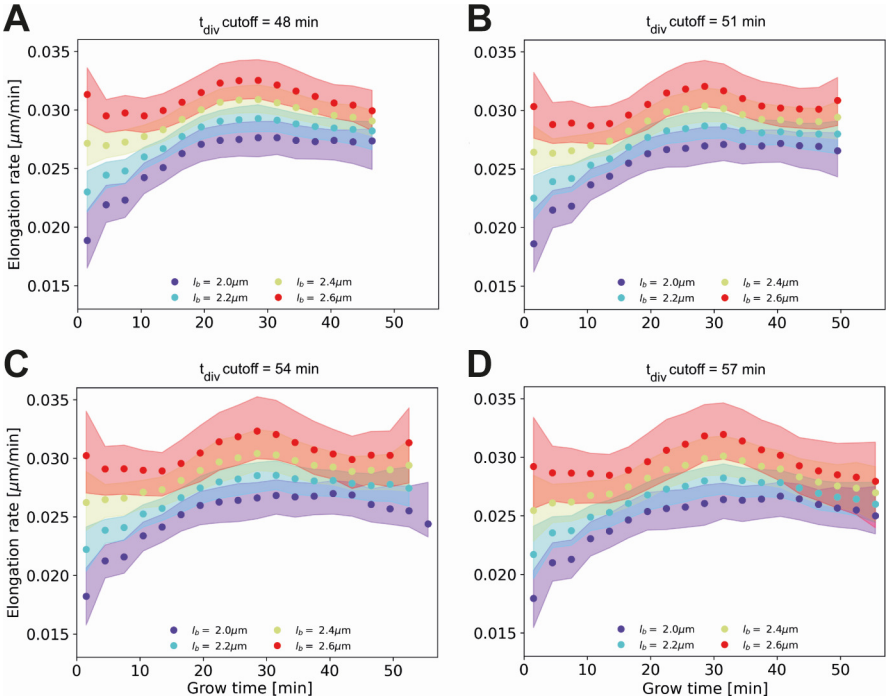



**Appendix 2—figure 3.** Elongation rate curves for different orders of the wave front fit of Main Text ***Figure 3A***: Linear (**A**), quadratic (**B**), and cubic (**C**). (**D**) $\chi^2$ of the fit of the wave front of Main Text ***Figure 3A*** for different fitting orders, together with the mean error on the elongation rate curves.

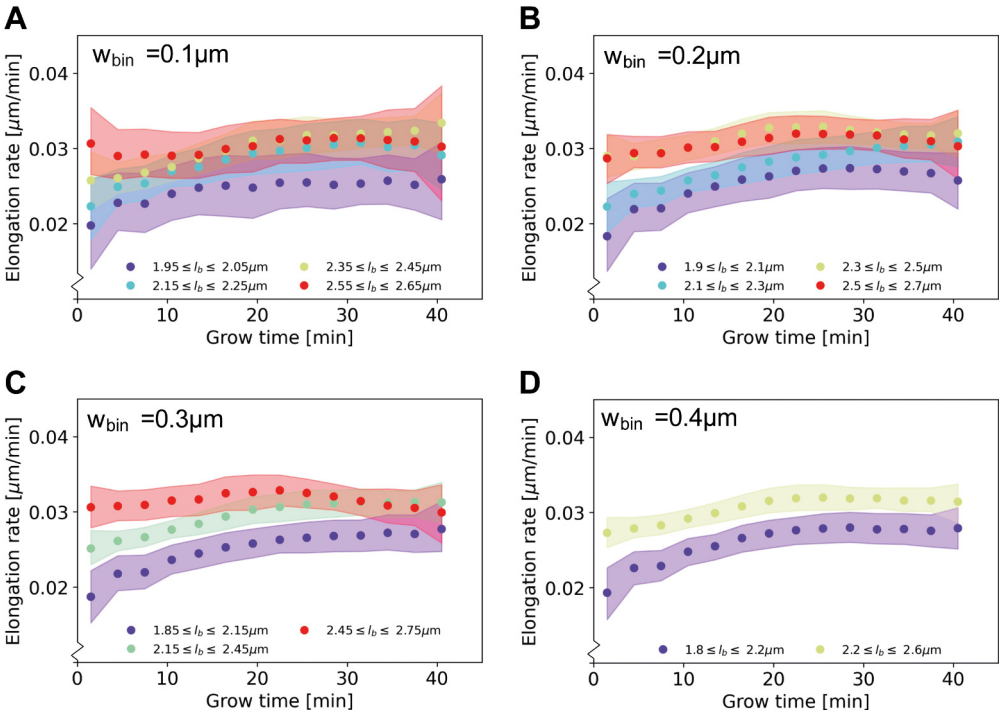
*Appendix 2—figure 3 continued on next page*

*Appendix 2—figure 3 continued*

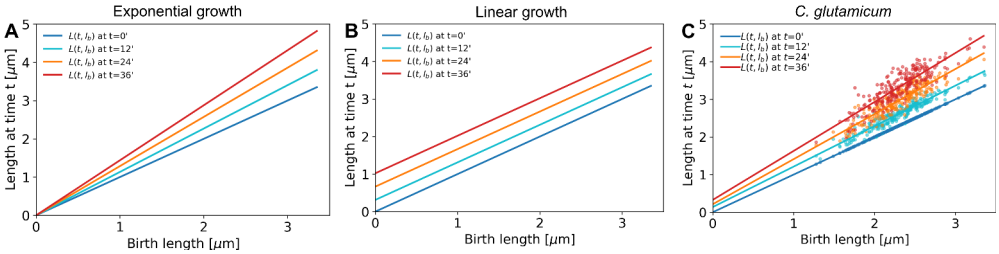
The negligible improvement of the goodness-of-fit after the first order justifies the use of a linear fit for further analysis.



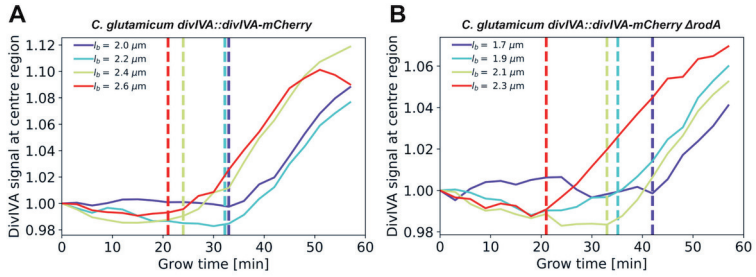

**Appendix 2—figure 4.** Conditional elongation rate curves, conditioned on DivVIA-labeled cells that have a generation time larger than a set cutoff value: 48 min (**A**), 51 min (**B**), 54 min (**C**) and 57 min (**D**). The inferred elongation rate curves still display similar growth behavior to the unconditioned population (Main Text ***Figure 5A***), but exhibit an overall downwards shift with increasing cutoff times. For larger cutoff times, the number of cells included decreases, resulting in larger errors on the inferred elongation rates. The linear growth phase observed until the cutoff time for the unconditioned population is seen to persist for longer grow times.



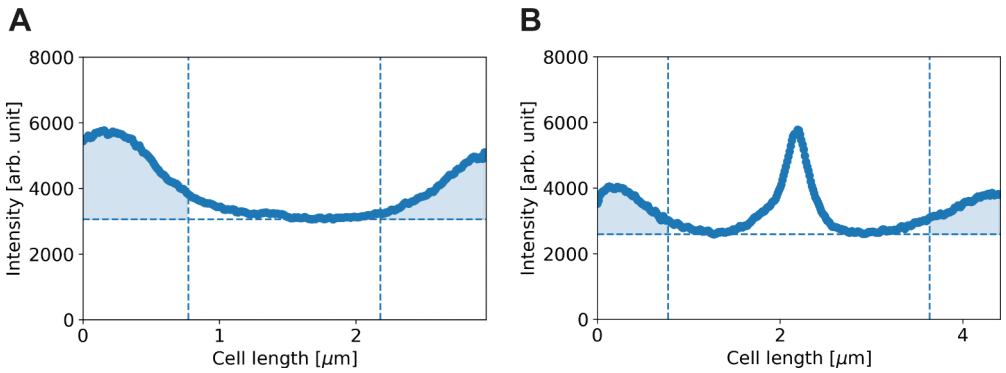

**Appendix 2—figure 5.** Elongation rate curves obtained through a binning procedure. Cells are divided into birth length bins, and for each bin the average length as a function of grow time is calculated. The resulting elongation curves are smoothened according to the same procedure as the elongation curves presented in the main text (see Appendix 5). From the smoothened elongation curves, elongation rates are calculated as a function of grow time. Results are shown for a bin width of 0.1 µm (**A**), 0.2 µm (**B**), 0.3 µm (**C**), 0.4 µm (**D**), where each $l_b$ indicates the center of the birth length bin.



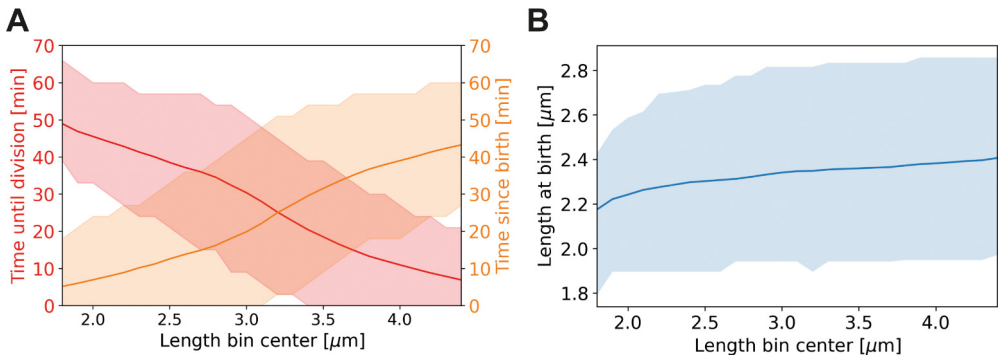

**Appendix 2—figure 6.** A linear fit through the cell lengths at each time step would be enough to describe exponential growth (**A,** offset is zero for all time stamps) as well as linear growth (**B,** slope is equal to 1 for all time stamps). *C. glutamicum* (**C**) matches neither of these growth modes.




**Appendix 2—figure 7.** The average DivIVA-mCherry signal from the cell center over time is shown for DivIVA-labeled cells (**A**) and Δ*rodA* DivIVA-labeled cells (**B**). The cell center is here defined as the region between 20% and 80% of the total cell length. The onsets of septum formation, derived from the DivIVA signal-mCherry signal, are indicated by the dashed lines; these do not consistently coincide with the levelling off of elongation rates (Main Text ***Figure 5A***). This is inconsistent with the leveling off being due to a competition between polar growth and septum formation.




**Appendix 2—figure 8.** Calculation of corrected polar HADA intensity, illustrated for two HADA profiles. Solid line: HADA intensity profile. Dashed horizontal line: minimum of HADA profile. Dashed vertical lines: boundary of polar region. Shaded area: calculated total polar intensity. Results shown for a cell with a length of 2.3 μm (**A**) and 4.4 μm (**B**).

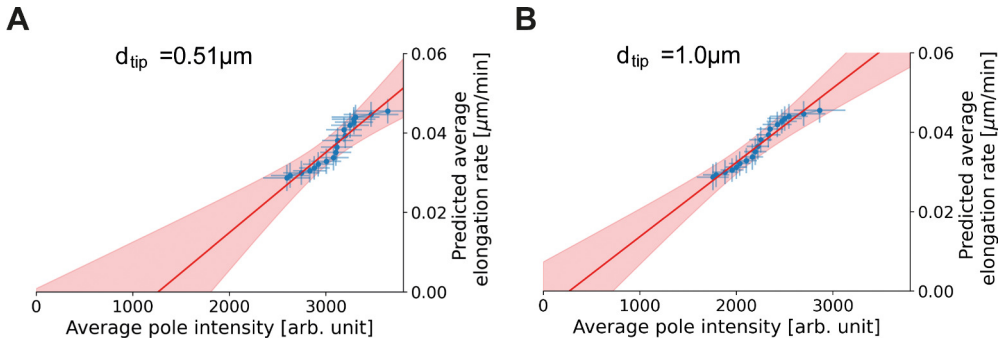



**Appendix 2—figure 9.** Average properties of wild-type cells as a function of length. Values are shown over the range of observed lengths in the HADA staining experiment, using a moving average with the same width (±0.7 μm) as in Main Text ***Figure 6C***. (**A**) Red line: average time until division, together with the two standard deviation bounds (red shaded area). Orange line: average

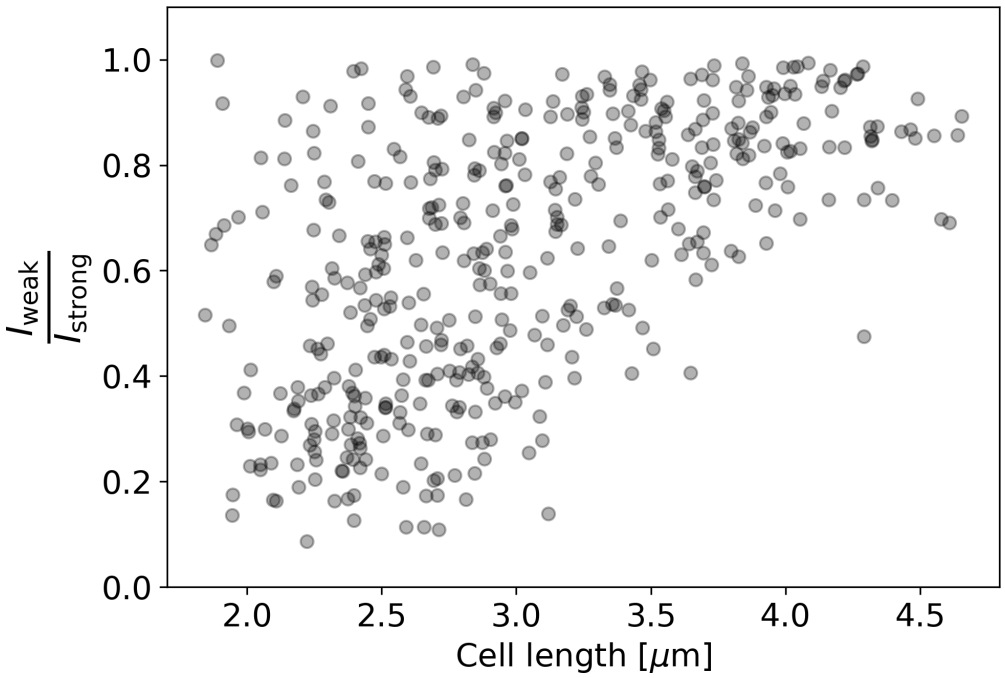
*Appendix 2—figure 9 continued on next page*

*Appendix 2—figure 9 continued*

time since birth, together with two standard deviation bounds (orange shaded area). (**B**) Blue line: average birth length for each birth length bin (blue line), together with the two standard deviation bounds (blue shaded area).




**Appendix 2—figure 10.** Proportionality between average pole intensity and predicted average elongation rate for different polar region definitions. Average elongation rate as a function of cell length (red), predicted from obtained average elongation rate curves, together with the average HADA staining intensity at the cell pole after background correction (blue). Results are shown for a polar region defined to be within 0.51 µm (**A**) and 1.0 µm (**B**) of the cell tip.




**Appendix 2—figure 11.** Ratio of intensities between the weaker and the stronger pole of each cell in the HADA staining experiment. Polar intensities are calculated as described in ***Appendix 2—figure 8***. Here, $I_{\mathrm{weak}}$ denotes the intensity of the cell pole with the weaker HADA intensity signal, and $I_{\mathrm{strong}}$ denotes the intensity of the pole with the stronger signal. For NETO-like growth (***Hannebelle et al., 2020***), a clustering of values around 0 (before new end take off) and 1 (after new end take off) would be expected, which is not observed here.

# Appendix 3

## Measurement noise estimate

To obtain an estimate for the measurement noise from our time-series growth data, we make use of length measurements at subsequent time intervals. For short enough time intervals, the variance of the length differences between intervals can be used as a measure of the measurement noise. However, since we expect cellular growth to also significantly contribute to this variance within the 3 min measurement interval, we have to separate out the two contributions.

To separate out the two contributions to the variance in subsequent length measurements, we write this variance as

$$\mathrm{Var}(l_m(t+\Delta t)-l_m(t)) \;=\; \mathrm{Var}(l(t+\Delta t)-l(t))+2\sigma_{\mathrm{n}}^2 \tag{A7}$$

with $l_m(t)$ the measured length at time $t$, $l(t)$ the actual length at time $t$, and $\sigma_{\mathrm{n}}$ the standard deviation of the measurement noise. This expression can be derived by noting that for a single elongation trajectory, we have

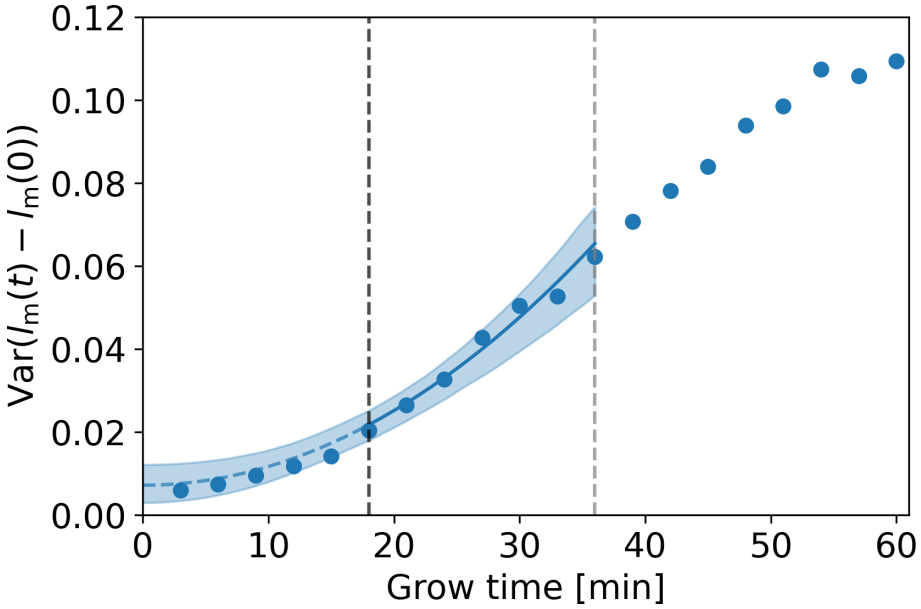
$$l_m(t+\Delta t)-l_m(t) \;=\; l(t+\Delta t)+\xi-(l(t)+\xi)=l(t+\Delta t)-l(t)+\sqrt{2}\xi, \tag{A8}$$

with $\xi$ the measurement noise. A solution for $\sigma_{\mathrm{n}}$ can be found if the functional form of $\mathrm{Var}(l(t),l(t+\Delta t))$ is known, by obtaining values for multiple $\Delta t$ and treating $\sigma_{\mathrm{n}}$ as a fitting parameter. To obtain this functional form, we make use of the observed linear growth regime after ~20 min (Main Text ***Figure 5***). We observe that the elongation rate is approximately constant in this regime for cells of all birth lengths, and now assume that this is also true for cells individually within this regime. The contrary would imply that non-constant single-cell elongation rates precisely cancel out across time and birth lengths to produce linear growth, which seems biologically implausible.

For linearly growing single cells, the standard deviation of $l(t+\Delta t)-l(t)$ is proportional to $\Delta t$, implying that the term $\mathrm{Var}(l(t),l(t+\Delta t))$ is of the form

$$\mathrm{Var}(l(t+\Delta t)-l(t))=\; c\Delta t^2, \tag{A9}$$

with $c$ an unknown parameter. To simultaneously obtain $c$ and $\sigma_{\mathrm{n}}$, we fit ***Equation (A7)*** under substitution of ***Equation (A9)*** to the DivIVA-labeled cell data over the regime between the onset of linear growth (18 min, black dashed line ***Appendix 3—figure 1***) and the first division event (36 min, gray dashed line ***Appendix 3—figure 1***). From this fit, we obtain the estimates $\sigma_n =$ 0.060 $\pm$ 0.018 $\mu$m and $c=4.5x10^{-5}\pm 0.47\mathrm{xm}^2\ \mathrm{min}^{-2}$, where the error margins are determined via bootstrapping. This value of $\sigma_n$ is used in the correction procedure for assigned birth lengths described in Appendix 4.




**Appendix 3—figure 1.** Estimation of measurement noise procedure. Blue dots: variance of $l_m(t) - l_m(0)$ as a function of grow time for the DivIVA-labeled cells, with $l_m(t)$ the measured cellular length at grow time $t$. A fit of ***Equation (A7)*** under substitution of ***Equation (A9)*** (blue line) is made to the points between the onset of linear growth (black dashed line) and the moment of first division (gray dashed line). The value of the extrapolated fit (blue dashed line) at t=0 is equal to $2\sigma_n^2$, with $\sigma_n$ the standard deviation of the measurement noise. The 95% confidence intervals of the model fit (blue shaded area) are obtained via bootstrapping.

# Appendix 4

## Bias correction procedure for assigned birth lengths

Before calculating average elongation rate curves, a statistical bias arising in the assignment of birth lengths to each curve needs to be corrected for. This bias is not specific to the inference method introduced in this paper, but arises in any procedure involving the assignment of lengths to a cells within a population, if there is noise in the measurement of individual cell lengths.

Due to measurement noise, cells will be assigned to birth lengths that systematically differ from their actual birth lengths. Specifically, given that the birth lengths in the population follow a symmetric, unimodal distribution, cells with a measured birth length larger than the population mean will on average be assigned a birth length that is larger than their actual length. Conversely, cells with a birth length smaller than the population mean will on average be assigned a birth length that is smaller.

The magnitude of the systematic deviation in the assignment of birth lengths is calculated as follows. Given that the cellular birth lengths follow a Gaussian distribution $P_l(l_b)$ with mean $\mu_l$ and standard deviation $\sigma_l$, and the measurement noise follows a Gaussian distribution $P_n(\Delta l)$ with mean 0 and standard deviation $\sigma_n$, the distribution of measured lengths will again be a Gaussian, with mean $\mu_m = \mu_l$ and standard deviation $\sigma_m = \sqrt{\sigma_l^2 + \sigma_n^2}$.

For a given measured birth length $l_m$, we now consider the probability distribution of corresponding actual birth lengths $P_l(l_b|l_m)$. This distribution is given by

$$P_l(l_b|l_m) = P_l(l_b)P_n(l_m - l_b). \tag{A4}$$

The product of two Gaussian distributions is again Gaussian, with a mean equal to

$$\langle l_b|l_m\rangle = \frac{\sigma_n^2\mu_l + \sigma_l^2\int l_b P_n(l_m - l_b)\mathrm{d}l_b}{\sigma_n^2 + \sigma_l^2} = \frac{\sigma_n^2\mu_l + \sigma_l^2 l_m}{\sigma_n^2 + \sigma_l^2}. \tag{A5}$$

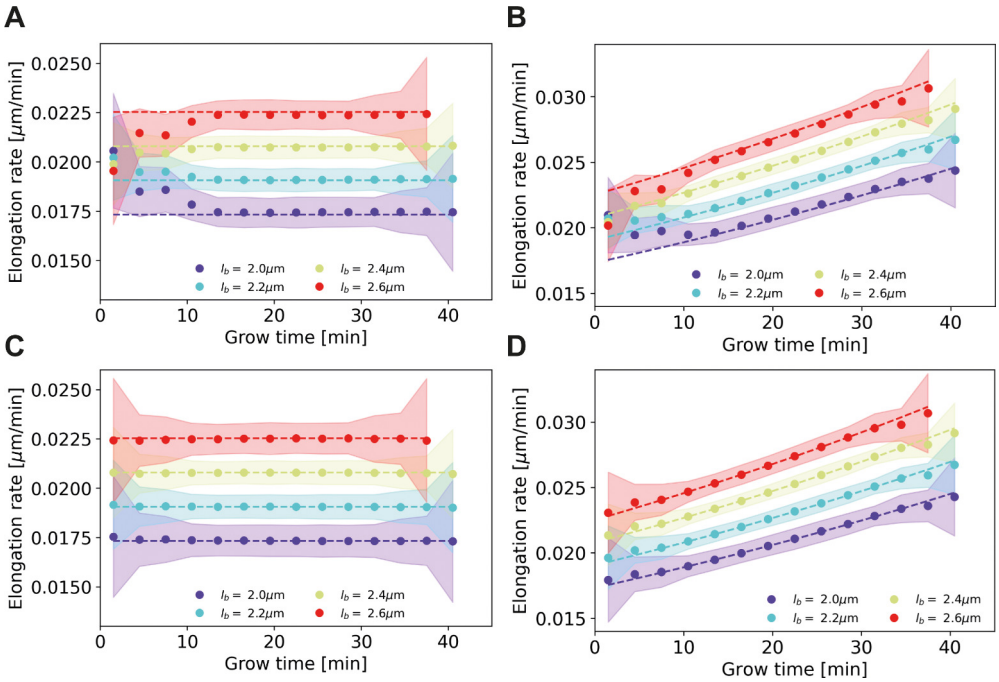
***Equation (A5)*** thus provides the transformation needed to remove the systematic bias in the assignment of birth lengths, and to determine the most likely birth length $l_b$ to a cell with a measured birth length $l_m$. For an estimation of the experimental measurement noise, see Appendix 3.

For the length increase since birth, there is no systematic bias once the bias in birth length has been removed. We can see this as follows. For each single-cell elongation trajectory, the measured length $l_m(t)$ at time $t$ is given by

$$l_m(t) = l_b + \Delta l_t + \xi, \tag{A6}$$

with $\xi$ the measurement noise and $\Delta l_t$ the length increase since birth at time $t$. As the measurement noise $\xi$ has a zero mean, there is no systematic bias in length increases after birth, provided that we have an unbiased estimate for the birth length $l_b$.

To test the derived correction procedure for assigned birth lengths, we performed a simulation of a population of growing cells, with the length measurement subject to noise. The measurement noise was sampled from a Gaussian, with the same standard deviation as estimated for experiment (Appendix 3). The single-cell growth mode was chosen as an input parameter. We analyzed two choices for input growth mode: linear (***Appendix 4—figure 1A,C***, dashed lines) and exponential (***Appendix 4—figure 1B,D***, dashed lines), with elongation rates comparable in magnitude to measured elongation rates.




**Appendix 4—figure 1.** Elongation rate inference on simulated data sets, with and without bias correction procedure for assigned birth lengths. For all panels: dashed lines: input elongation rates. Dots: mean inferred average elongation rates, obtained by applying our inference procedure to 1000 simulated data sets. Shaded areas: 2σ bounds on the inferred elongation rates. For all simulated data sets, the measurement noise is drawn from a Gaussian distribution with a standard deviation of 0.075 μm, matching the estimated experimental noise (Appendix 4). The population size and birth length distribution are chosen to match those observed for the DivIVA-labeled cells. Simulation conditions: (**A**) Linear input elongation rates constructed by setting $l(t) = l_b + 0.26 l_b t$. No bias correction procedure for assigned birth lengths is applied. (**B**) Exponential input elongation rates constructed by setting $l(t) = l_b e^{0.26t}$. No bias correction procedure for assigned birth lengths is applied. (**C**) Input elongation rates as in (**A**). The bias correction procedure for assigned birth lengths is applied. (**D**) Input elongation rates as in (**B**). The bias correction procedure for assigned birth lengths is applied.

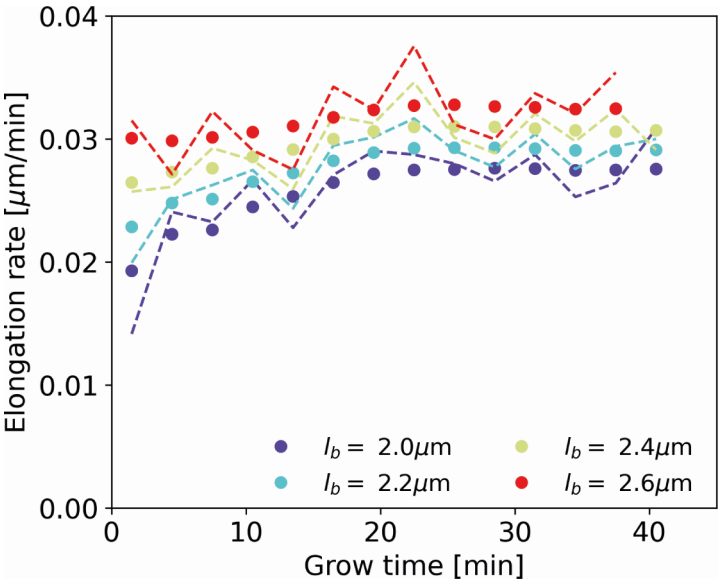
For each single-cell growth mode, we applied our elongation rate inference procedure to simulated cell lengths subject to measurement noise. Without correcting for a bias in assigned birth lengths, we find a systematic deviation between inferred elongation rates and input elongation rates in both cases (***Appendix 4—figure 1A,B***). With the implementation of the correction for assigned birth lengths, the input elongation rates are, however, accurately recovered (***Appendix 4—figure 1C,D***).

Minor deviations from the input elongation rates can still be seen for exponentially growing cells (***Appendix 4—figure 1D***), arising from applying a Gaussian smoothing to elongation curves that are locally nonlinear due to limited time resolution. However, this effect is small compared to the uncertainty on the inferred elongation rates.

# Appendix 5

## Smoothing of elongation curves

We obtain elongation rate curves (Main Text ***Figure 5*** and ***Figure 6C***) by taking a numerical derivative of smoothed growth trajectories. For the smoothing, a Gaussian smoothing procedure was used. In this procedure, a moving average is applied twice over groups of three subsequent time stamps of average elongation curves. As a check of the validity of the smoothing procedure, we also compare elongation rates before and after smoothing (***Appendix 5—figure 1***).

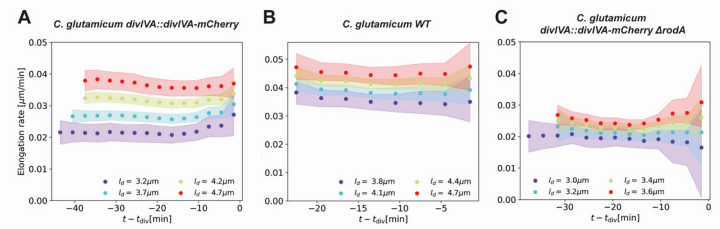



**Appendix 5—figure 1.** Average elongation rate curves obtained after Gaussian smoothing of the inferred average elongation curves (dots), together with average elongation rate curves obtained from unsmoothed average elongation curves (dashed lines).

# Appendix 6

## Calculating mean elongation curves as a function of time until division

The construction of the average elongation curves $L(t-t_d, l_{\mathrm{d}})$ as a function of the time until division $t-t_d$ and division length $l_d$ is as follows. We relate the length at time $t-t_d$ to the division length $l_{\mathrm{d}}$ for all cells, and use linear fits to obtain a family of curves $L_{t-t_d}(l_{\mathrm{d}})$ for each $t-t_d$. From this family of relations $L_{t-t_d}(l_{\mathrm{d}})$, we can subsequently compute $L(t-t_d, l_{\mathrm{d}})$ for any choice of $l_{\mathrm{d}}$. The resulting mean elongation rate curves are shown in ***Appendix 6—figure 1***.

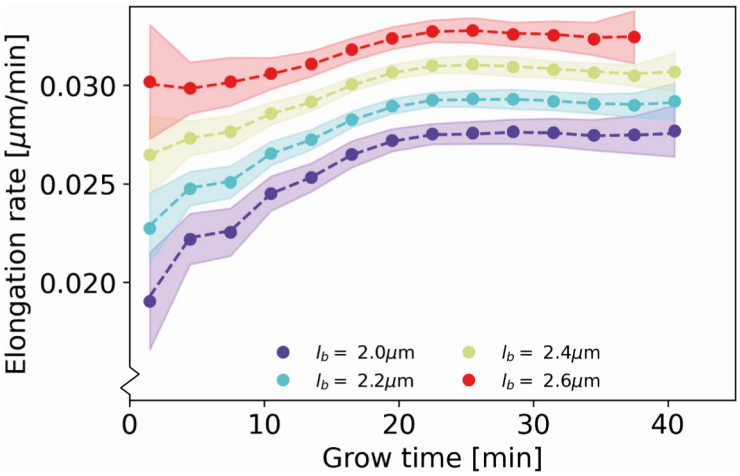



**Appendix 6—figure 1.** Inferred elongation rates as a function of the time until division, shown for DivIVA-labeled cells (**A**), wild-type cells (**B**) and the Δ*rodA* mutant (**C**). To obtain these curves, the elongation rate inference procedure described in the Main Text was applied, with the modification that $L(t-t_{\mathrm{div}}, l_{\mathrm{d}})$ was calculated, rather than $L(t, l_{\mathrm{b}})$. This yields average elongation rate curves as a function of division length, which are unbiased until the growth time of the shortest-lived cell (left endpoints of the elongation rate curves). The inferred linear growth regime for later grow times persists until division.

# Appendix 7

## Testing the elongation rate inference procedure

To test our elongation rate inference procedure, we generated a simulated data set with elongation rates as inferred by our inference procedure for DivIVA-labeled cells (Main Text ***Figure 5***). The distribution of birth lengths and division lengths of the simulated cells are taken to match the experimentally observed distributions. On each simulated data point, a measurement noise as determined in Appendix 3 is applied. On the simulated data set subject to noise, we apply the assigned birth length correction procedure as described in Appendix 4, and subsequently apply our elongation rate inference procedure. We find that the input elongation rates are accurately recovered (***Appendix 7—figure 1***), demonstrating the internal consistency of our inference approach.




**Appendix 7—figure 1.** Recovery of inferred elongation rates from simulated growth Dashed lines: input elongation rates, as inferred for DivIVA-labeled cells (Main Text ***Figure 5A***). Dots: average of elongation rates inferred from simulated growth experiment. Shaded areas: 95% confidence intervals inferred from simulated growth experiment, obtained via bootstrapping.

# Appendix 8

## Prediction of average elongation rate as a function of cell length

To calculate the predicted average elongation rates shown in Main Text ***Figure 6C***, we make use of our time-series data for wild-type cells, and the inferred mean elongation rates shown in Main Text ***Figure 5B***.

We start by calculating the time-averaged elongation rate $\bar{l}'_i$ for each cell $i$ in the wild-type data set, where the prime denotes a time derivative, by dividing the length added between birth and division by the total growth time. We then assume that the elongation rate for a cell at a time $t$ is approximately given by a rescaling of the population-averaged elongation rates $L'(t,l_b)$ by the time-averaged elongation rate of the cell $\bar{l}'_i$. Specifically, we calculate the estimated elongation rate at time $t$ by

$$l'_i(t) = L'(t,l_b)\,\frac{\bar{l}'_i n_{\mathrm{i}}}{\sum_{t=0}^{t^i_{\mathrm{div}}} L'(t,l_b)}, \tag{A10}$$

with $n_{\mathrm{i}}$ the number of time intervals in the growth trajectory of cell $i$, and $t^i_{\mathrm{div}}$ its division time. For times $t$ later than the first population division event $T_{\mathrm{div}}$, we obtain a value for $L'(t,l_b)$ by extrapolating the linear growth regime, setting $L'(t,l_b) = < L'(t,l_b) >_{20\,\mathrm{min}<t<T_{\mathrm{div}}}$.

From the ensemble $\{l'_i(t)\}$ of estimated elongation rates of all cells at each time since birth, we calculate the average elongation rate as a function of cell length by taking a moving average over the corresponding measured $\{l_i(t)\}$. The standard error on the mean is calculated from the standard deviation and the number of cells of each moving average bin.

# Appendix 9

## RAG model fitting procedure

The model fits shown in Main Text ***Figure 6E–G*** are obtained via the ParametricNDSolve function in Mathematica. The obtained parameter values are shown in ***Appendix 9—tables 1*** and ***2***.

**Appendix 9—table 1.** Parameter values obtained by fitting Main Text ***Equation (2)*** to inferred elongation rate curves.
The values shown in column 4 and 6 are an average over the four birth lengths of each condition.

| Genotype | $l_b$ [$\mu$m] | $\beta$ [$t^{-1}$] | $\langle\beta\rangle$ [$t^{-1}$] | $\frac{N(t=0)}{N^{max}}$ | $\left\langle\frac{N(t=0)}{N^{max}}\right\rangle$ |
|---|---|---|---|---|---|
| *wild-type* | 2.1 | 0.088 | 0.085 | 0.67 | 0.62 |
| | 2.3 | 0.068 | | 0.62 | |
| | 2.5 | 0.093 | | 0.62 | |
| | 2.7 | 0.089 | | 0.58 | |
| *divIVA::divIVA-mCherry* | 2.0 | 0.109 | 0.088 | 0.69 | 0.80 |
| | 2.2 | 0.100 | | 0.77 | |
| | 2.4 | 0.087 | | 0.84 | |
| | 2.6 | 0.054 | | 0.88 | |
| *divIVA::divIVA-mCherry ΔrodA* | 1.7 | 0.063 | 0.087 | 0.61 | 0.64 |
| | 1.9 | 0.094 | | 0.65 | |
| | 2.1 | 0.084 | | 0.64 | |
| | 2.3 | 0.11 | | 0.65 | |

**Appendix 9—table 2.** Parameter values obtained by fitting Main Text ***Equation (3)*** to inferred elongation rate curves.
The values shown in columns 5 and 7 are an average over the four birth lengths of each condition.

| Genotype | $l_b$ [$\mu$m] | $\beta$ [$t^{-1}$] | $\gamma$ [$t^{-1}$] | $\langle\beta e^{\gamma t}\rangle_{t<20min}$ [$t^{-1}$] | $\frac{N(t=0)}{N^{max}}$ | $\left\langle\frac{N(t=0)}{N^{max}}\right\rangle$ |
|---|---|---|---|---|---|---|
| *wild-type* | 2.1 | 0.016 | 0.162 | 0.13 | 0.72 | 0.67 |
| | 2.3 | 0.039 | 0.086 | | 0.67 | |
| | 2.5 | 0.058 | 0.080 | | 0.65 | |
| | 2.7 | 0.082 | 0.050 | | 0.62 | |
| *divIVA::divIVA-mCherry* | 2.0 | 0.072 | 0.06 | 0.14 | 0.71 | 0.82 |
| | 2.2 | 0.050 | 0.09 | | 0.79 | |
| | 2.4 | 0.025 | 0.14 | | 0.86 | |
| | 2.6 | 0.005 | 0.25 | | 0.92 | |
| *divIVA::divIVA-mCherry ΔrodA* | 1.7 | 0.023 | 0.094 | 0.12 | 0.67 | 0.68 |
| | 1.9 | 0.039 | 0.092 | | 0.69 | |
| | 2.1 | 0.064 | 0.050 | | 0.67 | |
| | 2.3 | 0.064 | 0.100 | | 0.68 | |

# Appendix 10

## Population simulation method

The goal of the population growth simulations is to obtain the distribution of cellular birth lengths assuming two different growth modes: asymptotically linear and exponential elongation. Both simulations extract all necessary growth parameters and distributions from the experimental data. For the asymptotically linear growth mode, the simulation serves as a check whether the assumed growth mode indeed recovers the correct cellular length distribution. For the exponential growth scenario, the simulation reveals the cellular length distribution an exponential grower would have if it had inherent noise levels similar to *C. glutamicum* allowing for a fair comparison. Both simulations start with a single cell and continue for 20 generations, after which the birth lengths of the last generation are binned and plotted. Repeated simulations with different lengths of the starting cell do not show discernable differences.

## Exponential growers

For the exponential growers, cells are assumed to elongate according to

$$l(t) = l_b \exp(\alpha t) + \zeta(t) \tag{A11}$$

The exponential growth rate $\alpha$ is chosen as the slope of the linear fit of $\ln\left(\frac{l_t}{l_b}\right)$ versus $t_d$ that intersects the origin, as shown in Main Text ***Figure 3B***. A size-additive noise term is indicated by $\zeta(t)$, which will be specified below at the time of division. For a cell with a given birth length $l_b$, the target final length $l_t$ is determined via a linear fit of $l_b$ versus $l_d$, as shown in Main Text ***Figure 3A***. The target growth time $t_t$ is then given by $t_t = \frac{1}{\alpha}\ln\left(\frac{l_t}{l_b}\right)$. A time additive noise term $\Delta t$ is added to $t_t$ according to experimentally observed growth time variations (***Appendix 10—figure 1D***). Additionally, a size-additive noise term $\Delta l$ encodes the division length variation due to $\zeta(t)$, which is also directly obtained from experiment (***Appendix 10—figure 1E***).

The full expression for the division length $l_d$ is then given by

$$l_d = l_b \exp(\alpha(t_t + \Delta t)) + \Delta l \tag{A12}$$

At division, the characteristic V-snap of *C. glutamicum* occurs, separating the two daughter cells. During this V-snap, the length of the daughter cells rapidly increases: the average measured birth length is 0.57 times the average measured division length (2.3 µm and 4.0 µm respectively), instead of the expected ratio of 0.5. To account for this V-snap effect, we calculate the distribution of added lengths during the V-snap. We find that the average added length depends on the division length: longer cells add less length during the V-snap than shorter cells (***Appendix 10—figure 1B***). To take this length dependence into account, we subdivide the data set into three division length bins, and obtain a distribution of added lengths during the V-snap for each bin. When a simulated cell divides, an added length during V-snap is randomly drawn from the distribution corresponding to its division length.

After division, the length asymmetry of the two daughter cells is chosen by drawing a random value from the experimentally observed division asymmetry distribution (***Appendix 10—figure 1C***) corresponding to the obtained division length. This distribution is found to be narrower for the shortest birth lengths (***Appendix 10—figure 1C***), thus two distributions are used.

## Asymptotically linear growers

For asymptotically linear growers, cells are assumed to elongate according to

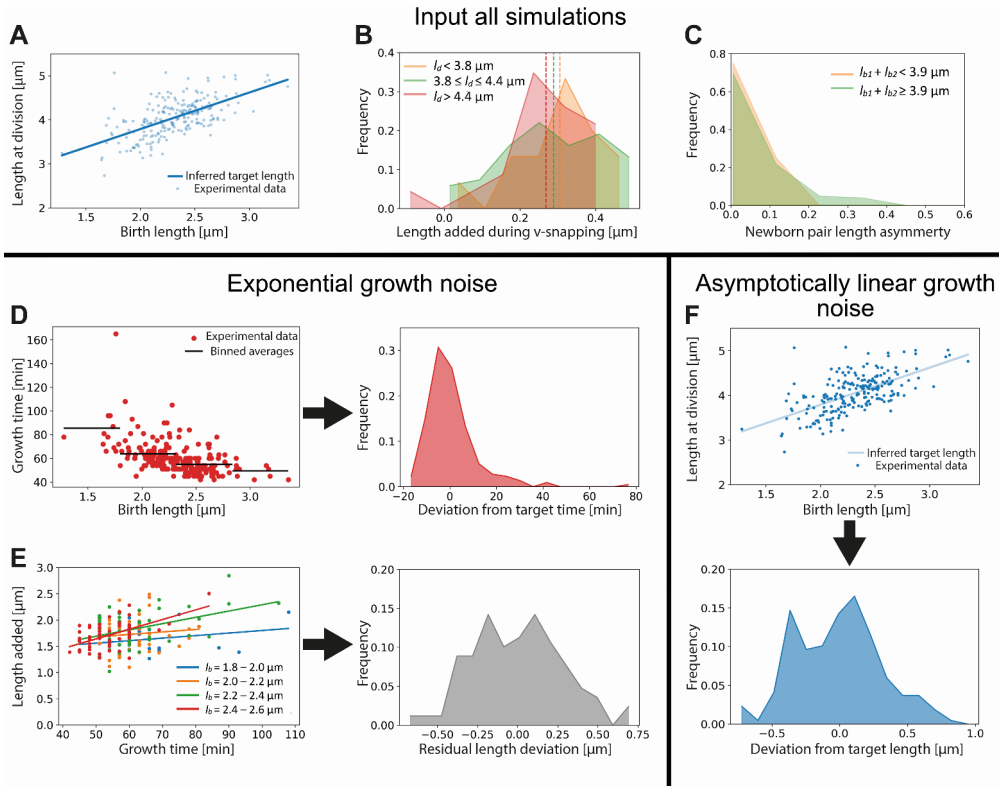
$$l(t) = l_b + \lambda t + \gamma(\exp(-\beta t) - 1) + \eta(t), \tag{A13}$$

which is obtained by inserting Main Text ***Equation 3*** into Main Text ***Equation 1***, integrating and grouping constant terms into $\lambda$ and $\gamma$. An additive noise term $\eta(t)$ is added to this to account for single-cell variability around the inferred average growth trajectory. We assume the cells to have the

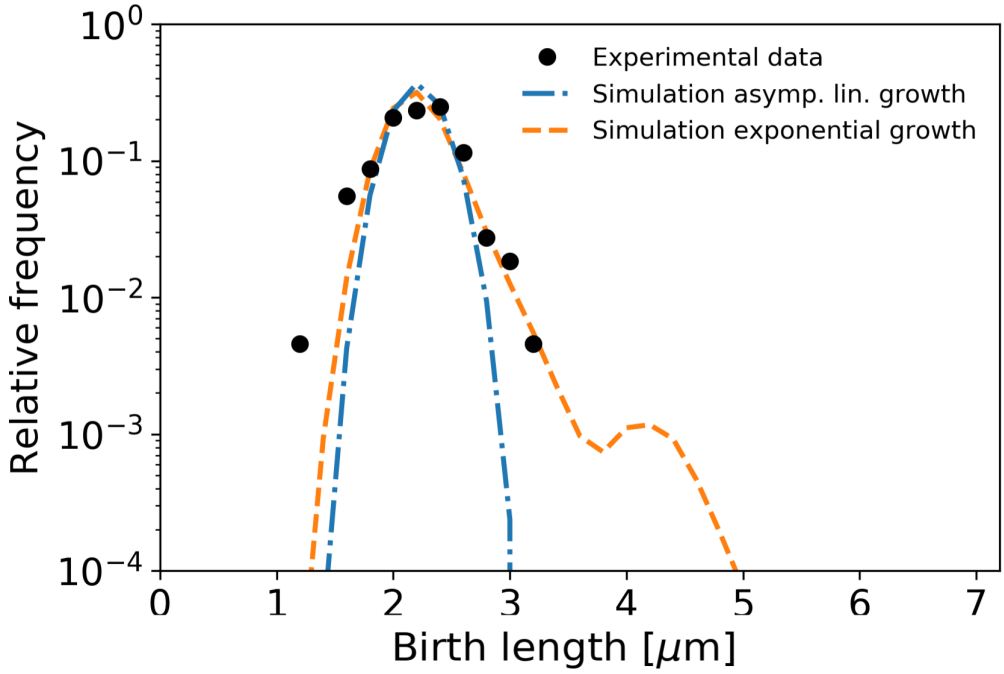
same target final length $l_t$ as in the exponentially growing scenario, determined via a linear fit of $l_b$ versus $l_d$. For cells close to observed division times, the term proportional to $\gamma$ can be approximated as being constant in time, simplifying the growth mode to linear growth (Main Text ***Figure 5A***). A time-additive noise term $\Delta t$ will then act as size-additive noise and can thus be absorbed into one additive noise term $\Delta l$, obtained from the experimental distribution of final sizes $l_d$ around the target final sizes (***Appendix 10—figure 1F***). The expression for the division length is thus given by

$$l_d = l_t + \Delta l. \tag{A14}$$

The division asymmetry and V-snap effect are incorporated in the same way as for the exponential grower simulation.




**Appendix 10—figure 1.** Input used for simulations of exponential and asymptotically linear growth. For both simulations, a linear fit of the division length versus birth length is used to define a target length (**A**). The length added during the V-snap at division is randomly drawn from the distribution corresponding to the division length of the simulated cell (**B**). The experimental data is divided into three subpopulations according to division length (red, green, and orange distributions), as the average length added during V-snap decreases with division length (dashed lines). The asymmetry of the daughter cells is randomly drawn from the distribution corresponding to the combined length of the simulated daughters (**C**). As the asymmetry is lower for the smallest daughter cells, the experimental data is divided into two subpopulations (red and green distributions). For the simulation of exponential growth, two noise sources are needed as input. The time-additive noise is randomly drawn from the distribution of deviations from target growth times (**D**). This distribution is obtained from the deviations of single-cell growth times from the average of their birth length bin. All growth variability not captured by growth time variations is calculated for four narrow birth length bins (blue, orange, green, and red points) (**E**). From the distribution of deviations of added lengths from a linear fit for each initial size bin, a size-additive noise term is randomly drawn. For the linear growth simulation, only a single additive noise term is required, which is randomly drawn from the distribution of deviations of cells lengths at division from the target division length (**F**).




**Appendix 10—figure 2.** Birth length distributions as in Main Text ***Figure 7***, but with single-cell variability in division symmetry, growth time, and (residual) length deviation reduced by a factor 3. The second peak in the length distribution of exponential growth is attributed to the large time deviation of one single cell seen in ***Appendix 10—figure 1D***.

# Conclusions & outlook

In this thesis we investigated spatial chromosome organization and cell size throughout the bacterial cell cycle. Using a Maximum Entropy inference procedure, we uncovered organizational features across genomic scales, firstly for an unreplicating chromosome, and then for several stages throughout the replication cycle. We then uncovered a novel bacterial single-cell growth mode, which was found to act as a regulator for cell size at the population level.

The Maximum Entropy chromosome model developed in chapter 2 constitutes a principled approach to infer the full distribution of chromosome configurations directly from experimental Hi-C data. We saw how the MaxEnt model for *Caulobacter crescentus* correctly predicts genomic localizations along the long cell axis, despite the model input only consisting of two-point contact frequencies. The MaxEnt model predicted a striking pattern of positional correlations along the long cell axis, which were explained by large genomic clusters termed Super Domains (SuDs), which tend to exclude each other if they lie on opposite chromosomal arms. On smaller genomic scales, we found a pattern of local extensions that correlates with the locations of highly-transcribed genes, but only for one chromosomal arm. Lastly, we quantified the localization information contained by each genomic region, which could be used by the cell to localize proteins and protein droplets.

In chapter 3, we expanded the approach from chapter 2 to describe a replicating chromosome. We made use of Hi-C data collected at various times throughout the cell cycle in *C. crescentus*, and adapted our model phase space to describe a replicating chromosome. A model trained only on Hi-C data turned out to be insufficiently constrained: this model yielded a replicating chromosome that does not segregate, which is likely due to the Hi-C data not distinguishing between inter- and intrachromosomal contacts. Thus, we added an additional constraint on the separation of replicated origins of replication (*ori*), motivated by the biologically observed active pulling of the newly replicated *ori*, which induces segregation. The resulting model was found to predict measured localizations of genomic regions with high accuracy across the chromosome. Furthermore, the replicating MaxEnt model provides insight into organizational features not yet accessible to experiment. We found a persistence of linear organization throughout the replicated chromosome, for all replication stages. A model containing only constraints on the positions of the replicated *ori*'s, termed the *ori* pulling model, showed that the linear organization of the replicated segment of the chromosome is largely explained by the pulling of *ori*. The linear organization of the unreplicated chromosome was not reproduced by this model however, which could be explained by the absence of loop extrusion motors in the *ori* pulling model. The replicating MaxEnt model provides access to many more organizational features yet to be explored, a few of which were discussed at the end of the chapter.

In chapter 4, we shifted perspective from the chromosome inside a growing cell to the dimensions of the cell itself. We studied single-cell elongation over time in the atypically growing *Corynebacterium glutamicum*, which forms a new cell wall exclusively at the cell poles, has a thick meshed cell wall structure, and lacks many common size regulation mechanisms. These

properties make this bacterium a promising candidate for uncovering novel single-cell growth modes that deviate from the commonly found exponential single-cell growth. From detailed single-cell measurements, we inferred average elongation rates despite noise and intrinsic variability in single-cell growth. Our inference procedure achieves this by using the noise-reducing properties of multi-cell averaging, while carefully avoiding inspection bias effects. We found that *C. glutamicum* deviates from the commonly observed exponential single-cell growth; mean elongation rates initially increase, but then level off to a linear growth regime. We found that this growth mode is consistent with the apical cell wall formation mechanism being the rate-limiting step for growth. Lastly, with population growth simulations we showed that asymptotically linear growth acts as a cell size regulation mechanism at the population level, offering an evolutionary explanation for the lack of many common growth regulation mechanisms in this bacterium.

This work leaves open several avenues for further exploration. In chapter 2, we saw that the MaxEnt model predicts the presence of large genomic clusters, termed Super Domains, while superresolution experiments confirmed the clustered nature of the chromosome. It is however still unclear if there is any substructure within the SuDs, although super-resolution experiments on *B. subtilis* [63] and *E. coli* [148] revealing similar structures would suggest this. In *B. subtilis*, high-density chromosomal regions (HDRs) were found to change their cellular positioning in the absence of ParB or SMC, and the number of HDRs appeared proportional to the estimated genomic content throughout the cell cycle and across growth conditions[63]. In *E. coli*, blob-like Mbp-size domains were observed in a chromosome within a broadened cell, which undergo major dynamic rearrangements at the minute timescale [148]. An interesting extension would be to search for similar features within the MaxEnt model for *C. crescentus*, as well as via direct experimental measurement. This would enhance our insight into large-scale organization in *C. crescentus*, but also shed light on the similarities between large-scale cluster organizations in different bacteria.

Although we learned a MaxEnt model for *C. crescentus*, our approach can readily be adapted to study other bacterial species and growth conditions. Single-chromosome Hi-C data sets have been published for *C. crescentus* cells under nutrient starvation [67], depletion of ParA and ParB [40], and replication-inhibited cells with increasing cell lengths [40]. For *Bacillus subtilis*, single-chromosome data has also been obtained for a mutant with a single ParS site [39]. With our MaxEnt approach, we can investigate chromosomal organization for these species and conditions in detail.

For the inference of chromosome structure from Hi-C data, a class of approaches has previously been developed that converts Hi-C scores to average distances, from which consensus structures are then calculated [72, 73, 149]. For *C. crescentus*, we found in chapter 2 that such distance-based models and the MaxEnt model do not agree on all organizational features; in [72, 149] a helical chromosomal structure was predicted, which is not observed within the MaxEnt model. This discrepancy may be explained by substantial region-to-region deviations from the mean relation between Hi-C scores and average distances, as well as significant correlations in the distances between genomic regions as predicted by the MaxEnt model. This raises the question however: under which conditions do distance-based models yield reliable predictions on consensus structures? A study employing various computationally generated chromosome ensembles and corresponding Hi-C maps could shed light on the necessary circumstances for the input model to be recovered, for distance-based methods as well as the MaxEnt model.

In our work on a replicating *C. crescentus* chromosome in chapter 3, we left several aspects

of organization unexplored. It would be interesting to study the properties of SuDs throughout the replication cycle, especially close to the replication fork position. Furthermore, with this model we can study patterns of local extension over time, which are likely affected by the pulling of the replicated *ori*. Lastly, quantifying localization information over time could provide insight into changes in the degree of order in the chromosome over time, as well as localization patterns that could be formed via the specific binding of proteins or protein droplets to genomic regions.

A promising expansion of the replicating chromosome model, is the adaption to Hi-C data sets for unsynchronized cells. For many bacteria, a synchronization as performed for *C. crescentus* is not possible, thus for these cases Hi-C data can only be obtained for a mixed population across cell cycle stages. Such mixed-population Hi-C data sets have been obtained for *Escherichia coli* [59], *Bacillus subtilis* [62, 63], the genome-reduced *Mycoplasma pneumoniae* [64], *Vibrio cholerae* [65], and *Corynebacterium glutamicum* [66]. An extension of the MaxEnt model for such mixed populations will require a modification of the model phase space to describe a distribution of replication stages and cell sizes. An important question for this extension, is whether Hi-C data contains enough information to sufficiently constrain a model of a mixed population. For our stroboscopic replicating chromosome model for *C. crescentus* in chapter 3, we already saw that a positional constraint on the replicated *ori* distances is required to yield a model with predictive power. It is therefore possible that such positional constraints are also required for a mixed MaxEnt model, and potentially additional types of constraints could also be needed. Overcoming these challenges will however provide insight into spatial chromosome organization across a wealth of bacterial species throughout the replication cycle.

At the start of chapter 4, we discussed how different rate-limiting steps for single-cell bacterial growth imply different growth modes. We then saw how the inferred asymptotically linear elongation rates for *C. glutamicum* are consistent with a model of the apical cell wall formation mechanism being the rate-limiting step for growth. This does not imply however that this mechanism is rate-limiting across growth conditions. In fact, under systematic lowering of the nutrient concentration a tipping point might be expected, where nutrient uptake becomes rate-limiting. This would imply a switch to exponential growth, which in turn could entail a sudden widening of of the cell size distribution. Detailed experiments over a range of nutrient conditions could reveal such a speculated nutrient-induced phase transition of the bacterial growth mode, and possibly uncover novel growth regulatory mechanisms.

While for rich nutrient conditions we found *C. glutamicum*'s cell length to grow asymptotically linearly, the increase in cell mass is not necessarily bound to this growth mode. In [150], dry-mass density was found to vary significantly throughout the cell cycle in *E. coli* and *C. crescentus*, as cells were shown to expand their surface, rather than volume, in proportion to biomass growth. Whether this proportionality also holds for the asymptotically linearly growing *C. glutamicum* is an open question; it would be broken however if mass production is proportional to protein content, as this implies an exponential mass increase. Thus, measuring the dry-mass increase over time in *C. glutamicum* could shed light on how linear volume growth and exponential mass growth are coordinated.

In a broader context, this work illustrates how principled inference methods can create deep understanding of biological mechanisms directly from experimental data. The proclamation 'mathematics is biology's next microscope' [151] rings true in this work, with chromosome organization and cellular growth further illuminated by the light of analytical inference.

# Bibliography


[1] Mavridou, D. A., Gonzalez, D., Kim, W., West, S. A. & Foster, K. R. Bacteria Use Collective Behavior to Generate Diverse Combat Strategies. *Current Biology* **28**, 345–355.e4 (2018).

[2] Wagner, A. *Arrival of the fittest: solving evolution's greatest puzzle* (Penguin, 2014).

[3] Barkai, N. & Leibler, S. Robustness in simple biochemical networks. *Nature* **387**, 913–917 (1997).

[4] Berg, H. & Purcell, E. Physics of chemoreception. *Biophysical Journal* **20**, 193–219 (1977).

[5] Bialek, W. & Setayeshgar, S. Physical limits to biochemical signaling. *Proceedings of the National Academy of Sciences* **102**, 10040–10045 (2005).

[6] Mehta, P. & Schwab, D. J. Energetic costs of cellular computation. *Proceedings of the National Academy of Sciences* **109**, 17978–17982 (2012).

[7] Halatek, J., Brauns, F. & Frey, E. Self-organization principles of intracellular pattern formation. *Philosophical Transactions of the Royal Society B: Biological Sciences* **373**, 20170107 (2018).

[8] Brauns, F. *et al.* Bulk-surface coupling identifies the mechanistic connection between Min-protein patterns in vivo and in vitro. *Nature Communications* **12**, 3312 (2021).

[9] Sun, M., Wartel, M., Cascales, E., Shaevitz, J. W. & Mignot, T. Motor-driven intracellular transport powers bacterial gliding motility. *Proceedings of the National Academy of Sciences* **108**, 7559–7564 (2011).

[10] Wadhwa, N. & Berg, H. C. Bacterial motility: machinery and mechanisms. *Nature Reviews Microbiology* (2021).

[11] Nirody, J. A., Sun, Y.-R. & Lo, C.-J. The biophysicist's guide to the bacterial flagellar motor. *Advances in Physics: X* **2**, 324–343 (2017).

[12] Celani, A. & Vergassola, M. Bacterial strategies for chemotaxis response. *Proceedings of the National Academy of Sciences* **107**, 1391–1396 (2010).

[13] Reichenbach, T., Mobilia, M. & Frey, E. Mobility promotes and jeopardizes biodiversity in rock–paper–scissors games. *Nature* **448**, 1046–1049 (2007).

[14] Bauer, M. & Frey, E. Multiple scales in metapopulations of public goods producers. *Physical Review E* **97**, 042307 (2018).

[15] Luria, S. E. & Delbrück, M. Mutations of Bacteria from Virus Sensitivity to Virus Resistance. *Genetics* **28**, 491–511 (1943).

[16] Lenski, R. E. & Travisano, M. Dynamics of adaptation and diversification: a 10,000-generation experiment with bacterial populations. *Proceedings of the National Academy of Sciences* **91**, 6808–6814 (1994).

[17] Wiser, M. J., Ribeck, N. & Lenski, R. E. Long-Term Dynamics of Adaptation in Asexual Populations. *Science* **342**, 1364–1367 (2013).

[18] Wang, X., Llopis, P. M. & Rudner, D. Z. Organization and segregation of bacterial chromosomes. *Nature Reviews Genetics* **14**, 191–203 (2013).

[19] Postow, L., Hardy, C. D., Arsuaga, J. & Cozzarelli, N. R. Topological domain structure of the Escherichia coli chromosome. *Genes & Development* **18**, 1766–1779 (2004).

[20] Nöllmann, M., Crisona, N. J. & Arimondo, P. B. Thirty years of Escherichia coli DNA gyrase: From in vivo function to single-molecule mechanism. *Biochimie* **89**, 490–499 (2007).

[21] Luttinger, A. The twisted 'life' of DNA in the cell: bacterial topoisomerases. *Molecular Microbiology* **15**, 601–606 (2006).

[22] Drlica, K. Control of bacterial DNA supercoiling. *Molecular Microbiology* **6**, 425–433 (1992).

[23] Worcel, A. & Burgi, E. On the structure of the folded chromosome of Escherichia coli. *Journal of Molecular Biology* **71**, 127–147 (1972).

[24] Higgins, N. P., Yang, X., Fu, Q. & Roth, J. R. Surveying a supercoil domain by using the gamma delta resolution system in Salmonella typhimurium. *Journal of Bacteriology* **178**, 2825–2835 (1996).

[25] Deng, S., Stein, R. A. & Higgins, N. P. Organization of supercoil domains and their reorganization by transcription. *Molecular Microbiology* **57**, 1511–1521 (2005).

[26] Kavenoff, R. & Ryder, O. A. Electron microscopy of membrane-associated folded chromosomes of Escherichia coli. *Chromosoma* **55**, 13–25 (1976).

[27] Arold, S. T., Leonard, P. G., Parkinson, G. N. & Ladbury, J. E. H-NS forms a superhelical protein scaffold for DNA condensation. *Proceedings of the National Academy of Sciences* **107**, 15728–15732 (2010).

[28] Dame, R. T. H-NS mediated compaction of DNA visualised by atomic force microscopy. *Nucleic Acids Research* **28**, 3504–3510 (2000).

[29] Dame, R. T., Noom, M. C. & Wuite, G. J. Bacterial chromatin organization by H-NS protein unravelled using dual DNA manipulation. *Nature* **444**, 387–390 (2006).

[30] Dame, R. T. The role of nucleoid-associated proteins in the organization and compaction of bacterial chromatin. *Molecular Microbiology* **56**, 858–870 (2005).

[31] Cosgriff, S. *et al.* Dimerization and DNA-dependent aggregation of the Escherichia coli nucleoid protein and chaperone CbpA. *Molecular Microbiology* **77**, 1289–1300 (2010).

[32] Schneider, R. An architectural role of the Escherichia coli chromatin protein FIS in organising DNA. *Nucleic Acids Research* **29**, 5107–5114 (2001).

[33] Dame, R. T., Rashid, F.-Z. M. & Grainger, D. C. Chromosome organization in bacteria: mechanistic insights into genome structure and function. *Nature Reviews Genetics* **21**, 227–242 (2020).

[34] Wang, X., Tang, O. W., Riley, E. P. & Rudner, D. Z. The SMC Condensin Complex Is Required for Origin Segregation in Bacillus subtilis. *Current Biology* **24**, 287–292 (2014).

[35] Jensen, R. B. & Shapiro, L. The Caulobacter crescentus smc gene is required for cell cycle progression and chromosome segregation. *Proceedings of the National Academy of Sciences* **96**, 10661–10666 (1999).

[36] Ganji, M. *et al.* Real-time imaging of DNA loop extrusion by condensin. *Science* **360**, 102–105 (2018).

[37] Kim, E., Kerssemakers, J., Shaltiel, I. A., Haering, C. H. & Dekker, C. DNA-loop extruding condensin complexes can traverse one another. *Nature* **579**, 438–442 (2020).

[38] Pradhan, B. *et al.* Smc complexes can traverse physical roadblocks bigger than their ring size. *BioRxiv* (2021).

[39] Wang, X., Brandão, H. B., Le, T. B. K., Laub, M. T. & Rudner, D. Z. *Bacillus subtilis* SMC complexes juxtapose chromosome arms as they travel from origin to terminus. *Science* **355**, 524–527 (2017).

[40] Tran, N. T., Laub, M. T. & Le, T. B. K. SMC progressively aligns chromosomal arms in *Caulobacter crescentus* but is antagonized by convergent transcription. *Cell* **20**, 2057–2071 (2017).

[41] Miermans, C. A. & Broedersz, C. P. Bacterial chromosome organization by collective dynamics of SMC condensins. *Journal of the Royal Society Interface* **15**, 20180495 (2018).

[42] Lin, D. C.-H. & Grossman, A. D. Identification and Characterization of a Bacterial Chromosome Partitioning Site. *Cell* **92**, 675–685 (1998).

[43] Murray, H., Ferreira, H. & Errington, J. The bacterial chromosome segregation protein Spo0J spreads along DNA from parS nucleation sites. *Molecular Microbiology* **61**, 1352–1361 (2006).

[44] Breier, A. M. & Grossman, A. D. Whole-genome analysis of the chromosome partitioning and sporulation protein Spo0J (ParB) reveals spreading and origin-distal sites on the Bacillus subtilis chromosome. *Molecular Microbiology* **64**, 703–718 (2007).

[45] Gerdes, K., Howard, M. & Szardenings, F. Pushing and Pulling in Prokaryotic DNA Segregation. *Cell* **141**, 927–942 (2010).

[46] Leonard, T. A., Butler, P. J. & Löwe, J. Bacterial chromosome segregation: structure and DNA binding of the Soj dimer ? a conserved biological switch. *The EMBO Journal* **24**, 270–282 (2005).

[47] Vecchiarelli, A. G. *et al.* ATP control of dynamic P1 ParA-DNA interactions: a key role for the nucleoid in plasmid partition. *Molecular Microbiology* no–no (2010).

[48] Vecchiarelli, A. G., Mizuuchi, K. & Funnell, B. E. Surfing biological surfaces: exploiting the nucleoid for partition and transport in bacteria. *Molecular Microbiology* **86**, 513–523 (2012).

[49] Broedersz, C. P. *et al.* Condensation and localization of the partitioning protein ParB on the bacterial chromosome. *Proc. Natl. Acad. Sci. USA* **111**, 8809–8814 (2014).

[50] Hanauer, C., Bergeler, S., Frey, E. & Broedersz, C. P. Theory of Active Intracellular Transport by DNA Relaying. *Physical Review Letters* **127**, 138101 (2021).

[51] Lim, H. C. *et al.* Evidence for a DNA-relay mechanism in ParABS-mediated chromosome segregation. *eLife* **3** (2014).

[52] Wiggins, P. A., Cheveralls, K. C., Martin, J. S., Lintner, R. & Kondev, J. Strong intranucleoid interactions organize the *Escherichia coli* chromosome into a nucleoid filament. *Proc. Natl. Acad. Sci. USA* **107**, 4991–4995 (2010).

[53] Surovtsev, I. V., Campos, M. & Jacobs-Wagner, C. DNA-relay mechanism is sufficient to explain ParA-dependent intracellular transport and patterning of single and multiple cargos. *Proceedings of the National Academy of Sciences* **113**, E7268–E7276 (2016).

[54] Shebelut, C. W., Guberman, J. M., van Teeffelen, S., Yakhnina, A. A. & Gitai, Z. Caulobacter chromosome segregation is an ordered multistep process. *Proceedings of the National Academy of Sciences* **107**, 14194–14198 (2010).

[55] Lee, P. S. & Grossman, A. D. The chromosome partitioning proteins Soj (ParA) and Spo0J (ParB) contribute to accurate chromosome partitioning, separation of replicated sister origins, and regulation of replication initiation in Bacillus subtilis. *Molecular Microbiology* **60**, 853–869 (2006).

[56] Birnie, A. & Dekker, C. Genome-in-a-Box: Building a Chromosome from the Bottom Up. *ACS Nano* **15**, 111–124 (2021).

[57] Lieberman-Aiden, E. *et al.* Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science* **326**, 289–293 (2009).

[58] Le, T. B., Imakaev, M. V., Mirny, L. A. & Laub, M. T. High-resolution mapping of the spatial organization of a bacterial chromosome. *Science* **342**, 731–734 (2013).

[59] Lioy, V. S. *et al.* Multiscale structuring of the *E. coli* chromosome by nucleoid-associated and condensin proteins. *Cell* **172**, 771–783.e18 (2018).

[60] Mercier, R. *et al.* The MatP/matS Site-Specific System Organizes the Terminus Region of the E. coli Chromosome into a Macrodomain. *Cell* **135**, 475–485 (2008).

[61] Espeli, O., Mercier, R. & Boccard, F. DNA dynamics vary according to macrodomain topography in the E. coli chromosome. *Molecular Microbiology* **68**, 1418–1427 (2008).

[62] Wang, X. *et al.* Condensin promotes the juxtaposition of DNA flanking its loading site in *Bacillus subtilis*. *Genes Dev.* **29**, 1661–1675 (2015).

[63] Marbouty, M. *et al.* Condensin- and replication-mediated bacterial chromosome folding and origin condensation revealed by Hi-C and super-resolution imaging. *Molecular Cell* **59**, 588–602 (2015).

[64] Trussart, M. *et al.* Defined chromosome structure in the genome-reduced bacterium Mycoplasma pneumoniae. *Nature Communications* **8**, 14665 (2017).

[65] Marbouty, M. *et al.* Metagenomic chromosome conformation capture (meta3C) unveils the diversity of chromosome organization in microorganisms. *eLife* **3** (2014).

[66] Böhm, K. *et al.* Chromosome organization by a conserved condensin-ParB system in the actinobacterium Corynebacterium glutamicum. *Nature Communications* **11**, 1485 (2020).

[67] Le, T. B. K. & Laub, M. T. Transcription rate and transcript length drive formation of chromosomal interaction domain boundaries. *EMBO* **35**, 1582–1595 (2016).

[68] Brandão, H. B., Ren, Z., Karaboja, X., Mirny, L. A. & Wang, X. DNA-loop-extruding SMC complexes can traverse one another in vivo. *Nature Structural & Molecular Biology* **28**, 642–651 (2021).

[69] Messelink, J. J., van Teeseling, M. C., Janssen, J., Thanbichler, M. & Broedersz, C. P. Learning the distribution of single-cell chromosome conformations in bacteria reveals emergent order across genomic scales. *Nature communications* **12**, 1–9 (2021).

[70] Pal, K., Forcato, M. & Ferrari, F. Hi-C analysis: from data generation to integration. *Biophysical Reviews* **11**, 67–78 (2019).

[71] Umbarger, M. A. *et al.* The three-dimensional architecture of a bacterial genome and its alteration by genetic perturbation. *Mol. Cell* **44**, 252–264 (2011).

[72] Yildirim, A. & Feig, M. High-resolution 3D models of *Caulobacter crescentus* chromosome reveal genome structural variability and organization. *Nucleic Acids Res.* **46**, 3937–3952 (2018).

[73] Oluwadare, O., Highsmith, M. & Cheng, J. An overview of methods for reconstructing 3-D chromosome and genome structures from Hi-C data. *Biological Procedures Online* **21**, 7 (2019).

[74] Zhang, B. & Wolynes, P. G. Topology, structures, and energy landscapes of human chromosomes. *Proc. Natl. Acad. Sci. USA* **112**, 6062–6067 (2015).

[75] Di Pierro, M., Zhang, B., Aiden, E. L., Wolynes, P. G. & Onuchic, J. N. Transferable model for chromosome architecture. *Proc. Natl. Acad. Sci. USA* **113**, 12168–12173 (2016).

[76] Abbas, A. *et al.* Integrating Hi-C and FISH data for modeling of the 3D organization of chromosomes. *Nature Communications* **10**, 2049 (2019).

[77] Tjong, H. *et al.* Population-based 3D genome structure analysis reveals driving forces in spatial genome organization. *Proc. Natl. Acad. Sci. USA* **113**, E1663–1667 (2016).

[78] Javer, A. *et al.* Persistent super-diffusive motion of Escherichia coli chromosomal loci. *Nature Communications* **5**, 3854 (2014).

[79] Weber, S. C., Spakowitz, A. J. & Theriot, J. A. Nonthermal ATP-dependent fluctuations contribute to the in vivo motion of chromosomal loci. *Proc Natl Acad Sci USA* **109**, 7338–7343 (2012).

[80] Smith, K., Griffin, B., Byrd, H., MacKintosh, F. C. & Kilfoil, M. L. Nonthermal fluctuations of the mitotic spindle. *Soft Matter* **11**, 4396–4401 (2015).

[81] Tkačik, G. *et al.* The simplest maximum entropy model for collective behavior in a neural network. *Journal of Statistical Mechanics: Theory and Experiment* **2013**, P03011 (2013).

[82] Bialek, W. *et al.* Statistical mechanics for natural flocks of birds. *Proc Natl Acad Sci USA* **109**, 4786–4791 (2012).

[83] Marks, D. S. *et al.* Protein 3D structure computed from evolutionary sequence variation. *PloS One* **6**, e28766 (2011).

[84] Mora, T., Walczak, A. M., Bialek, W. & Callan, C. G. Maximum entropy models for antibody diversity. *Proc Natl Acad Sci USA* **107**, 5405–5410 (2010).

[85] De Martino, D., MC Andersson, A., Bergmiller, T., Guet, C. C. & Tkačik, G. Statistical mechanics for metabolic networks during steady state growth. *Nature Communications* **9**, 2988 (2018).

[86] Shannon, C. E. A Mathematical Theory of Communication. *Bell System Technical Journal* **27**, 379–423 (1948).

[87] Jaynes, E. T. Information Theory and Statistical Mechanics. *Physical Review* **106**, 620–630 (1957).

[88] Carter, T. An introduction to information theory and entropy. *http://csustan.csustan.edu/ tom/Lecture-Notes/Information-Theory/info-lec.pdf* (2014).

[89] Jaynes, E. On the rationale of maximum-entropy methods. *Proceedings of the IEEE* **70**, 939–952 (1982).

[90] Monod, J. The growth of bacterial cultures. *Annual Review of Microbiology* **3**, 371–394 (1949).

[91] Buchanan, R. E. Life Phases in a Bacterial Culture. *Journal of Infectious Diseases* **23**, 109–125 (1918).

[92] Lane-Claypon, J. E. Multiplication of Bacteria and the Influence of Temperature and some other conditions thereon. *Journal of Hygiene* **9**, 239–248 (1909).

[93] Wang, P. *et al.* Robust Growth of Escherichia coli. *Current Biology* **20**, 1099–1103 (2010).

[94] Young, J. W. *et al.* Measuring single-cell gene expression dynamics in bacteria using fluorescence time-lapse microscopy. *Nature Protocols* **7**, 80–88 (2012).

[95] Taheri-Araghi, S., Brown, S. D., Sauls, J. T., McIntosh, D. B. & Jun, S. Single-Cell Physiology. *Annual Review of Biophysics* **44**, 123–142 (2015).

[96] Marantan, A. & Amir, A. Stochastic modeling of cell growth with symmetric or asymmetric division. *Physical Review E* **94**, 012405 (2016).

[97] Amir, A. Cell Size Regulation in Bacteria. *Physical Review Letters* **112**, 208102 (2014).

[98] Campos, M. *et al.* A Constant Size Extension Drives Bacterial Cell Size Homeostasis. *Cell* **159**, 1433–1446 (2014).

[99] Deforet, M., van Ditmarsch, D. & Xavier, J. B. Cell-Size Homeostasis and the Incremental Rule in a Bacterial Pathogen. *Biophysical Journal* **109**, 521–528 (2015).

[100] Taheri-Araghi, S. *et al.* Cell-Size Control and Homeostasis in Bacteria. *Current Biology* **25**, 385–391 (2015).

[101] Fievet, A. *et al.* Single-Cell Analysis of Growth and Cell Division of the Anaerobe Desulfovibrio vulgaris Hildenborough. *Frontiers in Microbiology* **6** (2015).

[102] Logsdon, M. M. *et al.* A Parallel Adder Coordinates Mycobacterial Cell-Cycle Progression and Cell-Size Homeostasis in the Context of Asymmetric Growth and Organization. *Current Biology* **27**, 3367–3374.e7 (2017).

[103] Oishi, M., Yoshikawa, H. & Sueoka, N. Synchronous and Dichotomous Replications of the Bacillus subtilis Chromosome During Spore Germination. *Nature* **204**, 1069–1073 (1964).

[104] Yoshikawa, H., O'Sullivan, A. & Sueoka, N. Sequential replication of the bacillus subtilis chromosome, iii. Regulation of initiation. *Proceedings of the National Academy of Sciences* **52**, 973–980 (1964).

[105] Schaechter, M., MaalOe, O. & Kjeldgaard, N. O. Dependency on Medium and Temperature of Cell Size and Chemical Composition during Balanced Growth of Salmonella typhimurium. *Journal of General Microbiology* **19**, 592–606 (1958).

[106] Donachie, W. D. Relationship between Cell Size and Time of Initiation of DNA Replication. *Nature* **219**, 1077–1079 (1968).

[107] Hill, N. S., Kadoya, R., Chattoraj, D. K. & Levin, P. A. Cell Size and the Initiation of DNA Replication in Bacteria. *PLoS Genetics* **8**, e1002549 (2012).

[108] Si, F. *et al.* Invariance of Initiation Mass and Predictability of Cell Size in Escherichia coli. *Current Biology* **27**, 1278–1287 (2017).

[109] Koppes, L. J., Meyer, M., Oonk, H. B., de Jong, M. A. & Nanninga, N. Correlation between size and age at different events in the cell division cycle of Escherichia coli. *Journal of Bacteriology* **143**, 1241–1252 (1980).

[110] Bipatnath, M., Dennis, P. P. & Bremer, H. Initiation and Velocity of Chromosome Replication in Escherichia coli B/r and K-12. *Journal of Bacteriology* **180**, 265–273 (1998).

[111] Sharpe, M. E., Hauser, P. M., Sharpe, R. G. & Errington, J. Bacillus subtilis Cell Cycle as Studied by Fluorescence Microscopy: Constancy of Cell Length at Initiation of DNA Replication and Evidence for Active Nucleoid Partitioning. *Journal of Bacteriology* **180**, 547–555 (1998).

[112] Wallden, M., Fange, D., Lundius, E. G., Baltekin, Ö. & Elf, J. The Synchronization of Replication and Division Cycles in Individual E. coli Cells. *Cell* **166**, 729–739 (2016).

[113] Boye, E., Stokke, T., Kleckner, N. & Skarstad, K. Coordinating DNA replication initiation with cell growth: differential roles for DnaA and SeqA proteins. *Proceedings of the National Academy of Sciences* **93**, 12206–12211 (1996).

[114] Zheng, H. *et al.* Interrogating the Escherichia coli cell cycle by cell dimension perturbations. *Proceedings of the National Academy of Sciences* **113**, 15000–15005 (2016).

[115] Helmstetter, C., Cooper, S., Pierucci, O. & Revelas, E. On the Bacterial Life Sequence. *Cold Spring Harbor Symposia on Quantitative Biology* **33**, 809–822 (1968).

[116] Sompayrac, L. & Maaløe, O. Autorepressor Model for Control of DNA Replication. *Nature New Biology* **241**, 133–135 (1973).

[117] Ho, P.-Y. & Amir, A. Simultaneous regulation of cell size and chromosome replication in bacteria. *Frontiers in Microbiology* **6** (2015).

[118] Taheri-Araghi, S. Self-Consistent Examination of Donachie's Constant Initiation Size at the Single-Cell Level. *Frontiers in Microbiology* **6** (2015).

[119] Willis, L. & Huang, K. C. Sizing up the bacterial cell cycle. *Nature Reviews Microbiology* **15**, 606–620 (2017).

[120] Menikpurage, I. P., Woo, K. & Mera, P. E. Transcriptional Activity of the Bacterial Replication Initiator DnaA. *Frontiers in Microbiology* **12** (2021).

[121] Skarstad, K. & Katayama, T. Regulating DNA Replication in Bacteria. *Cold Spring Harbor Perspectives in Biology* **5**, a012922–a012922 (2013).

[122] Amir, A. Is cell size a spandrel? *eLife* **6** (2017).

[123] Mir, M. *et al.* Optical measurement of cycle-dependent cell growth. *Proceedings of the National Academy of Sciences* **108**, 13124–13129 (2011).

[124] Iyer-Biswas, S. *et al.* Scaling laws governing stochastic growth and division of single bacterial cells. *Proceedings of the National Academy of Sciences* **111**, 15912–15917 (2014).

[125] Yu, F. B. *et al.* Long-term microfluidic tracking of coccoid cyanobacterial cells reveals robust control of division timing. *BMC Biology* **15**, 11 (2017).

[126] Godin, M. *et al.* Using buoyant mass to measure the growth of single cells. *Nature Methods* **7**, 387–390 (2010).

[127] Nordholt, N., van Heerden, J. H. & Bruggeman, F. J. Biphasic Cell-Size and Growth-Rate Homeostasis by Single Bacillus subtilis Cells. *Current Biology* **30**, 2238–2247.e5 (2020).

[128] Kar, P., Tiruvadi-Krishnan, S., Männik, J., Männik, J. & Amir, A. Distinguishing different modes of growth using single-cell data. *eLife* **10** (2021).

[129] Guttmann, A. J. Self-avoiding walks and polygons–an overview. *arXiv preprint arXiv:1212.3448* (2012).

[130] Wang, J. D. & Levin, P. A. Metabolism, cell growth and the bacterial cell cycle. *Nature Reviews Microbiology* **7**, 822–827 (2009).

[131] Skerker, J. M. & Laub, M. T. Cell-cycle progression and the generation of asymmetry in Caulobacter crescentus. *Nature Reviews Microbiology* **2**, 325–337 (2004).

[132] Jensen, R. B. Coordination between Chromosome Replication, Segregation, and Cell Division in Caulobacter crescentus. *Journal of Bacteriology* **188**, 2244–2253 (2006).

[133] Jensen, R. B. A moving DNA replication factory in Caulobacter crescentus. *The EMBO Journal* **20**, 4952–4963 (2001).

[134] Mohl, D. A. & Gober, J. W. Cell cycle-dependent polar localization of chromosome partitioning proteins in *Caulobacter crescentus*. *Cell* **88**, 675–84 (1997).

[135] Jalal, A. S. B. & Le, T. B. K. Bacterial chromosome segregation by the ParABS system. *Open Biology* **10**, 200097 (2020).

[136] Gogou, C., Japaridze, A. & Dekker, C. Mechanisms for Chromosome Segregation in Bacteria. *Frontiers in Microbiology* **12** (2021).

[137] Wasim, A., Gupta, A. & Mondal, J. A Hi–C data-integrated model elucidates E. coli chromosome's multiscale organization at various replication stages. *Nucleic Acids Research* **49**, 3077–3091 (2021).

[138] Ebersbach, G., Briegel, A., Jensen, G. J. & Jacobs-Wagner, C. A Self-Associating Protein Critical for Chromosome Attachment, Division, and Polar Organization in Caulobacter. *Cell* **134**, 956–968 (2008).

[139] Bowman, G. R. *et al.* A Polymeric Protein Anchors the Chromosomal Origin/ParB Complex at a Bacterial Cell Pole. *Cell* **134**, 945–955 (2008).

[140] Le, T. B. K., Imakaev, M. V., Mirny, L. A. & Laub, M. T. High-resolution mapping of the spatial organization of a bacterial chromosome. *Science* **342**, 731–734 (2013).

[141] Brameyer, S. *et al.* DNA-binding directs the localization of a membrane-integrated receptor of the ToxR family. *Communications Biology* **2**, 4 (2019).

[142] Viollier, P. H. *et al.* Rapid and sequential movement of individual chromosomal loci to specific subcellular locations during bacterial DNA replication. *Proc. Natl. Acad. Sci. USA* **101**, 9257–9262 (2004).

[143] Bürmann, F. & Gruber, S. SMC condensin: Promoting cohesion of replicon arms. *Nature Structural and Molecular Biology* **22**, 653–655 (2015).

[144] Tsai, J. W. & Alley, M. R. Proteolysis of the Caulobacter McpA chemoreceptor is cell cycle regulated by a ClpX-dependent pathway. *J. Bacteriol.* **183**, 5001–5007 (2001).

[145] Ely, B. Genetics of *Caulobacter crescentus*. *Methods Enzymol.* **204**, 372–384 (1991).

[146] Ducret, A., Quardokus, E. M. & Brun, Y. V. Microbej, a tool for high throughput bacterial cell detection and quantitative analysis. *Nature microbiology* **1**, 1–7 (2016).

[147] Evinger, M. & Agabian, N. Envelope associated nucleoid from *Caulobacter crescentus* stalked and swarmer cells. *J. Bacteriol.* **132**, 294–301 (1977).

[148] Wu, F. *et al.* Direct imaging of the circular chromosome in a live bacterium. *Nature Communications* **10**, 2194 (2019).

[149] Umbarger, M. A. *et al.* The Three-Dimensional Architecture of a Bacterial Genome and Its Alteration by Genetic Perturbation. *Molecular Cell* **44**, 252–264 (2011).

[150] Oldewurtel, E. R., Kitahara, Y. & van Teeffelen, S. Robust surface-to-mass coupling and turgor-dependent cell width determine bacterial dry-mass density. *Proceedings of the National Academy of Sciences* **118**, e2021416118 (2021).

[151] Cohen, J. E. Mathematics Is Biology's Next Microscope, Only Better; Biology Is Mathematics' Next Physics, Only Better. *PLoS Biology* **2**, e439 (2004).

# Acknowledgements

First of all, I want to thank my supervisor Chase Broedersz for your enthusiasm and support during this scientific journey. Under your guidance, I developed into a mature scientist. You pushed me to never stop asking: what is the big deal? Why are your results important? Why should we care? Maintaining this big-picture view has made my research more relevant, more interesting, and more fulfilling. You also taught me how to think rigorously and critically about science and my own research, how to write well to maintain clarity and interest, and developed my skills in giving captivating and memorable talks.

My graduate school QBM I want to thank for helping me come to Munich, and starting my transition from a pure theoretical physicist to an interdisciplinary scientist at the interface of physics and biology with helpful courses and lectures. What I cherish most of all though is the group of friends I've made through the QBM program. I want to thank David, David, Alex, Kimbu, Zhenya and Lina for forming such a close friend group. You made Munich feel like home. I relished all the trips and holidays together, the countless evenings hanging out and chatting at a restaurant, bar, or at a dinner party at home. I also want to thank David for all the small day-to-day interactions at university, all the adventures, stupid discussions, and close friendship.

I also want to thank all the flatmates I've lived with over the last years; thank you Kimbu, Alex, Vanessa, Olaf, Joeri, Max and Søren for giving me a cozy home base. It was great to always find someone at home up for a chat or some casual banter. During the covid-19 lockdown periods, the WG kept things fun & cozy, and in a way pulled all of us through it. Also during the intense final sprint of this thesis, it was very nice to experience so much support from you guys.

Marc Bramkamp and Fabian Meyer I want to thank for taking me into the world of microbiology, taking the time to explain countless biological details, and experiencing what it's like to really fuse physics approaches and biological knowledge. Our project was a clear case of the result being more than the sum of its parts.

Muriel and Grześ I want to thank for being part of the final project of my PhD. I really enjoyed the dynamic of working on this project all together, with everyone doing excellent work and solving problems at breakneck speed. Thanks also to Grześ for helping me make the cover illustration of this thesis.

A further thank you to all the master students I've worked with over the years: Jacqueline, Johannes, Lucas and Janni. It was a pleasure to work together on each of our projects, and I learned a lot from thinking about projects from a supervisor perspective. Thank you also for your patience while I was also still finding out how to do science myself.

I'd like to thank my fellow group members over the years with whom I've had many fun discussions, dinners, conference visits, or just a chat over a cup of coffee: Hugo, Mareike, George, Karsten, Felix, Timo, David M, Manon, Isabella, Fridtjof, Federico, Federica, Grześ,

Estelle, Moritz, Patrick, Phillipp, Emanuel, Laeschi, Silke, Jonas, Raphaela, Matthew, Tobias, Johannes and Tom. A special thanks to Karsten for countless interesting discussions about everything under the sun, creative painting sessions, and entertaining dinners.

Finaly, I want to thank my parents and sister for their unrelenting support, and always being there for me with a listening ear and advice.