A simple stochastic model to describe the evolution over time of core genome SNP GC content in prokaryotes

Jon Bohlin[1,2,3,*], Brittany Rose[1,4], Ola Brynildsrud[1,3] & Birgitte Freiesleben De Blasio[1,4]

August 27, 2019

[1]Division of Infection Control and Environmental Health, Norwegian Institute of Public Health, Oslo, Norway.

[2]Centre for Fertility and Health, Norwegian Institute of Public Health, Oslo, Norway.

[3]Department of Production Animals, Faculty of Veterinary Medicine, Norwegian University of Life Science, Oslo, Norway.

[4]Department of Biostatistics, Oslo Centre for Biostatistics and Epidemiology, Institute of Basic Medical Sciences, University of Oslo, Oslo, Norway.

[*]Corresponding author

Abstract

Genomes in living organisms consist of the nucleotides adenine (A), guanine (G), cytosine (C) and thymine (T). All prokaryotes have genomes consisting of double-stranded DNA, where the A's and G's (purines) of one strand bind respectively to the T's and C's (pyrimidines) of the other. As such, the number of A's on one strand nearly equals the number of T's on the other, and the same is true of one strand's G's and the other's C's. Globally, this relationship is formalized as Chargaff's first parity rule; its strandwise equivalent is Chargaff's second parity rule. Therefore, the GC content of any double-stranded DNA genome can be expressed as $\%GC = 100\% - \%AT$.

Variation in prokaryotic GC content can be substantial between taxa but is generally small within microbial genomes. This variation has been found to correlate with both phylogeny and environmental factors. Since novel single-nucleotide polymorphisms (SNPs) within genomes are at least partially linked to the environment, SNP GC content can be considered a compound measure of an organism's environmental influences, lifestyle and phylogeny.

We present a mathematical model that describes how SNP GC content in microbial genomes evolves over time as a function of the AT→GC and GC→AT mutation rates with Gaussian white noise disturbances. The model suggests that, in non-recombining bacteria, mutations can first accumulate unnoticeably and then abruptly fluctuate out of control. Thus, minuscule variations in mutation rates can suddenly become unsustainable, ultimately driving a species to extinction if not counteracted early enough. This model, which is suited specifically to symbiotic prokaryotes, conforms to scenarios predicted by Muller's ratchet and may suggest that this is not always a gradual, degrading process. It is also in agreement with some of the empirical evidence that motivated the formulation of the Red Queen hypothesis. We apply our model to different lineages of *Renibacterium salmoninarum* and find a substantial increase in SNP GC content within the most disseminated lineage, 1a. That increase could be due to a dramatic change in environment for this lineage.

1 Introduction

GC content varies considerably between prokaryotic species but is remarkably stable genome-wide, despite the fact that bacterial genomes are predominantly functional and expressed in some sense [1]. Bacteria can have an average genomic GC content of as low as 13.5% (*Candidatus* Zinderia insecticola) or of as high as 75% (*Anaeromyxobacter dehalogenans*) [2]. While both large and small bacteria can be either GC-rich or AT-rich, there seems to be a tendency—at least in some phylogentic groups—for symbionts with smaller genomes to be more AT-rich, while soil-dwelling bacteria with large genomes tend to be more GC-rich [4, 5].

The mechanisms responsible for GC richness in bacteria with large genomes are poorly understood; far more can be deduced from AT-rich bacteria with small genomes (see [5] for a general review of GC content in prokaryotes). For instance, it was conjectured [6] (before being later demonstrated [7]) that mutations are generally AT-biased due to frequent methylation of cytosine that can subsequently change to thymine. Bacteria in a symbiotic relationship with their host (often an insect) undergo reductive evolution through the loss of genes rendered unnecessary by the within-host environment. There is a clear evolutionary drive towards economizing energy expenditure [8, 21]. When host organisms have low effective population size (N_e) or density, genetic drift also influences the size and base composition of symbiont genomes [9, 10]. The outside environment can also affect genomic base composition in bacteria [12].

Phylogenetic relatedness, on the other hand, exerts strong pressure against changes in GC content. This is due in large part to the significant role that protein coding genes play in bacteria and to the fact that mutations in the first two positions of a codon can change the amino acid defined by that codon [13]. Phylogenetic influence on base composition in prokaryotes seems to be most prominent at the genus level and below [18].

There are several indicators that genome size reduction occurs before genomic GC content drops [9]. Loss of DNA mismatch repair (MMR) genes and proofreading enzymes can nevertheless lead to a relatively quick decrease in genomic GC content [14]. An increase in genomic GC content, on the other hand, can result in increased fitness [15], and this is associated with stronger selection on base composition [17, 18, 19]. Abundance of nitrogen, as in soil, has been identified as a driver for increased genomic GC content [16].

A recent study [2] found that single-nucleotide polymorphisms (SNPs) in microbial core genomes from different taxa were surprisingly GC-rich, except in cases where the genomes themselves were already among the most GC-rich. The study presented a mathematical model describing SNP GC content as a function of core genome GC content. The model indicated that GC→AT mutations occurred at roughly double the rate of AT→GC mutations, which suggests that most GC→AT mutations are lost prior to fixation [2].

In another recent study [3], it was shown that while GC→AT mutation rates are remarkably consistent across bacterial taxa, AT→GC mutation rates vary considerably. Since the environment exerts selective pressure on bacterial base composition [12, 13], it should, at least partly, be reflected in core genome SNPs, together with evolutionary history, lifestyle and taxon.

Stochastic events strongly impact the influence of the environment on genomic base composition in bacteria. Inspired by Motoo Kimura's seminal paper [11], we modify a previously described model [2] to investigate SNP GC content evolution with respect to time. Furthermore, we extend the model with the assumption of Gaussian white noise perturbations in the mutation rates. We assume that SNP GC content is subject to Chargaff's parity rules. In practice, this means that core genome SNP GC content depends on the bases that are selected (including through hitchhiking [37]) and not on random mutations that are purged before fixation. We employ Itô calculus to solve the stochastic differential equation (SDE) that accounts for the random perturbations in the AT→GC and GC→AT mutation rate parameters. We then discuss implications of the model and present outcomes that show striking concordance with Muller's ratchet [22] and with evolutionary mechanisms described by the Red Queen hypothesis [23, 24]. Finally, we apply the model to a genomic data set consisting of SNP GC content differences in lineages of the fish pathogen *Renibacterium salmoninarum* (taken from [20]).

2 Mathematical model

2.1 Motivation

The mathematical model presented here is an extension of the model presented in [2]. The original model, which describes the change in core genome

SNP GC content with respect to core genome GC content, is

$$\frac{dF_{GC}(x)}{dx} = \alpha F_{GC}(x) + \beta(1 - F_{GC}(x)). \tag{1}$$

x represents core genome GC content, while $F_{GC}(x)$ represents SNP GC content. These terms are subject to the constraints $0 < x < 1$ and $0 < F_{GC}(x) < 1$. In [2], the parameters α and β were estimated by fitting the model to empirical data using either non-linear least square regression [2] or Bayesian inference [3].

In the present study, we are concerned with the change in SNP GC content with respect to time in a stochastic setting. That is, we are now interested in the relation

$$F_{t+\Delta t}(\omega) = F_t(\omega) + \alpha F_t(\omega)\Delta t + \beta(1 - F_t(\omega))\Delta t, \tag{2}$$

where $F_t(\omega)$ represents SNP GC content at time t. The change in $F_t(\omega)$ with respect to trajectory ω during time Δt is a parameter α times $F_t(\omega)$ times Δt plus a parameter β times $1 - F_t(\omega)$ (SNP AT content at time t) times Δt. In other words, the difference in SNP GC content with respect to time is assumed to be equal to parameter multiples of SNP GC content and SNP AT content. In classical calculus notation, we write

$$\frac{dF_t(\omega)}{dt} = \alpha F_t(\omega) + \beta(1 - F_t(\omega)), \tag{3}$$

Here, $F_t(\omega)$ is a stochastic process, and we let $\alpha = a + W_t(\omega)$ and $\beta = b + W_t(\omega)$, where $a, b \in \mathbb{R}$ and $W_t(\omega)$ is a Gaussian white noise process. Equation (3) is subject to the probability space $(\Omega, \mathcal{F}_t, P)$ as well as the measure space $(\mathbb{R}^+, \mathcal{G}, dt)$. Ω is the space of all trajectories ω, $\mathcal{F}_t$ is its filtration with respect to each time $t \in \mathbb{R}^+$ (*i.e.* $[0, \infty)$ of which $\mathcal{G}$ is the corresponding Borel algebra and dt Lebesgue measure), and P is a probability measure on Ω. We now have:

$$\begin{aligned} \frac{dF_t(\omega)}{dt} \quad &= (a + W_t(\omega))F_t(\omega) + (b + W_t(\omega))(1 - F_t(\omega)) \\ &= aF_t(\omega) + F_t(\omega)W_t(\omega) + \\ &+ b(1 - F_t(\omega)) + W_t(\omega)(1 - F_t(\omega)) \\ &= aF_t(\omega) + b(1 - F_t(\omega)) + W_t(\omega). \end{aligned}$$

Hence,

$$\frac{dF_t(\omega)}{dt} = aF_t(\omega) + b(1 - F_t(\omega)) + W_t(\omega). \tag{4}$$

It is important to note that, in the present form, this derivative does not exist in the classical sense or in the Radon–Nikodym sense for $F_t(\omega)$. However, if we assume that $F_t(\omega)$ is a continuous semimartingale (allowing for countable and bounded jumps), the Doob–Meyer decomposition theorem (pp. 129–133 of [25]) guarantees that $F_t(\omega) = F_0 + A(t) + X_t(\omega)$, where $A(t)$ is a function of bounded variation and $X_t(\omega)$ is a local martingale. Moreover, this decomposition is unique, and both $A(t)$ and $X_t(\omega)$ are adapted to $\mathcal{F}_t$. If we assume that $X_t(\omega)$ is a Brownian motion, then by chapter 3 of [26], (4) can be written as

$$dF_t(\omega) = (aF_t(\omega) + b(1 - F_t(\omega)))dt + dB_t(\omega). \tag{5}$$

Though the term $(aF_t(\omega)+b(1-F_t(\omega)))dt$ resembles (1), we must handle the Brownian motion term $dB_t(\omega)$ in a non-classical way. We allow for scaled volatility c, as it is not unreasonable to expect variance differences across organisms and/or environments in addition to time t. It can be shown that a scaled Brownian motion is also a Brownian motion: Let U_t be a Brownian motion (see, for instance, ch. 2 of [26]). Then,

$$\mathbb{E}(U_t) = \frac{1}{\sqrt{2\pi t}} \int_{\mathbb{R}} ue^{-\frac{u^2}{2t}} du.$$

Letting $u = cz$ and $\frac{du}{dz} = c$, it follows that

$$\frac{1}{\sqrt{2\pi t}} \int_{\mathbb{R}} ue^{-\frac{u^2}{2t}} du = \frac{1}{\sqrt{2\pi \frac{t}{c^2}}} \int_{\mathbb{R}} ze^{-\frac{c^2z^2}{2t}} cdz$$

$$= \frac{1}{\sqrt{2\pi \frac{t}{c^2}}} \int_{\mathbb{R}} cze^{-\frac{c^2z^2}{2t}} dz = \mathbb{E}(cZ_{\frac{t}{c^2}}).$$

We do not presume that $F_t(\omega)$ can see into the future. Thus, the martingale condition $\mathbb{E}(F_s(\omega)|\mathcal{F}_t) = F_t(\omega)$ with $s > t$ holds when $\mathbb{E}$ is the expectation operator with respect to probability measure $P(\omega)$, *i.e.* $\mathbb{E}(X) = \int_\Omega XdP$. As such, we assume that $F_t(\omega)$ is adapted to the filtration $\mathcal{F}_t$ for each t, which motivates the use of the Itô integral instead of the Fisk–Stratonovich integral [25]. It is therefore enough to assume that $F_t(\omega)$ is a càdlàg process, *i.e.* $\lim_{s\to t^+} F_s(\omega) = F_t(\omega)$ (left-continuous with right limits; see ch. 2 of [25]), implying that $F_t(\omega)$ has a countable number of bounded jumps. We can then use the Itô formula (see ch. 4 of [26]) to solve (5). Furthermore, since we assume that $0 < F_t(\omega) < 1$ and that a, b are finite constants, it is guaranteed that (5) has a strong and unique solution (see ch. 5 of [26]).

First, we must identify an integrating factor that removes $F_t(\omega)$ from the right-hand side. Let

$$\begin{aligned} dF_t(\omega) &= (aF_t(\omega) + b(1 - F_t(\omega)))dt + d\hat{B}_t(\omega) \\ &= (aF_t(\omega) - bF_t(\omega) + b)dt + d\hat{B}_t(\omega) \\ &= ((a-b)F_t(\omega) + b)dt + d\hat{B}_t(\omega), \end{aligned}$$

where $\hat{B}_t(\omega)$ is a c-scaled Brownian motion. Letting $g(t,x) = e^{(-(a-b)t)}x$, we get the integrating factor $g(t, F_t(\omega)) = Y_t(\omega) = e^{(-(a-b)t)}F_t(\omega)$. Applying Itô's formula (p. 44 of [26]), we see that

$$dY_t(\omega) = \frac{\partial g}{\partial t}(t, F_t(\omega))dt + \frac{\partial g}{\partial t}(t, F_t(\omega))dF_t(\omega) + \frac{1}{2}\frac{\partial^2 g}{\partial x^2}(t, F_t(\omega))(dF_t(\omega))^2. \tag{6}$$

Because $\frac{\partial^2 g}{\partial x^2}(t,x) = 0$, the last term of (6) is equal to zero. As a result,

$$\begin{aligned} dY_t(\omega) \quad &= \frac{\partial g}{\partial t}(t, F_t(\omega))dt + \frac{\partial g}{\partial x}(t, F_t(\omega))dF_t(\omega) \\ &= -(a-b)e^{(-(a-b)t)}F_t(\omega)dt + e^{(-(a-b)t)}dF_t(\omega) \\ &= -(a-b)e^{(-(a-b)t)}F_t(\omega)dt + \\ &\quad +e^{(-(a-b)t)}(((a-b)F_t(\omega) + b)dt + d\hat{B}_t) \\ &= be^{(-(a-b)t)}dt + e^{(-(a-b)t)}d\hat{B}_t. \end{aligned}$$

Thus, we have the differential

$$dY_t(\omega) = be^{(-(a-b)t)}dt + e^{(-(a-b)t)}d\hat{B}_t, \tag{7}$$

and so

$$dY_t(\omega) = d(e^{(-(a-b)t)}F_t(\omega)) = be^{(-(a-b)t)}dt + e^{(-(a-b)t)}d\hat{B}_t.$$

We can then find the formula for $F_t(\omega)$, by setting $s \in [0,t]$, and letting

$$d(e^{(-(a-b)t)}F_t(\omega)) = be^{(-(a-b)t)}dt + e^{(-(a-b)t)}d\hat{B}_t$$

which gives

$$e^{(-(a-b)t)}F_t(\omega) - F_0(\omega) = \int_0^t be^{(-(a-b)s)}ds + \int_0^t e^{(-(a-b)s)}d\hat{B}_s,$$

and

$$F_t(\omega) = F_0(\omega)e^{(a-b)t} + \int_0^t be^{(a-b)(t-s)}ds + \int_0^t e^{(a-b)(t-s)}d\hat{B}_s. \tag{8}$$

Assuming that $F_t(\omega)$ is a semimartingale, we have from the Doob–Meyer decomposition that $\int_0^t be^{(a-b)(t-s)}ds$ is of bounded variation and that

$$\int_0^t e^{(a-b)(t-s)}d\hat{B}_s$$

is a local martingale, which is in fact a martingale (see pp. 129–133 of [25]). While the latter martingale term must be solved numerically, the antiderivative of the bounded variation term can be solved using the chain rule:

$$\int_0^t be^{(a-b)(t-s)}ds = c_0 + \frac{b}{(a-b)}(e^{(a-b)t} - 1).$$

We thus obtain the explicit equation for $F_t(\omega)$:

$$F_t(\omega) = F_0(\omega)e^{(a-b)t} + \frac{b}{(a-b)}(e^{(a-b)t} - 1) + \int_0^t e^{(a-b)(t-s)}d\hat{B}_s$$

that can be written as:

$$F_t(\omega) = -\frac{b}{(a-b)} + (F_0(\omega) + \frac{b}{(a-b)})e^{(a-b)t} + \int_0^t e^{(a-b)(t-s)}d\hat{B}_s \qquad (9)$$

which is subject to the constraints $t \in [0, \infty)$ and $0 < F_t(\omega) < 1$. The integration constant c_0 is just assumed included in F_0. It should be noted that for $F_0 = 0$,

$$\mathbb{E}(F_t(\omega)) = \frac{b}{(a-b)}(e^{(a-b)t} - 1). \qquad (10)$$

Since the martingale term vanishes (see p. 30 of [26]), we get the solution to (1) when $t = x$ [2]. Furthermore, we do not need to bother with the martingale term when estimating parameters a and b. The variance is given by $\mathrm{Var}(F_t(\omega)) = \mathbb{E}((F_t(\omega) - \mathbb{E}(F_t(\omega))^2)$, which we can solve by setting

$$A := F_0(\omega)e^{(a-b)t} + \frac{b}{(a-b)}(e^{(a-b)t} - 1)$$

and

$$B := \int_0^t e^{(a-b)(t-s)}d\hat{B}_s.$$

This gives:

$$\begin{aligned}\mathrm{Var}(F_t(\omega)) \quad &= \mathbb{E}((F_t(\omega) - \mathbb{E}(F_t(\omega))^2 \\ &= \mathbb{E}((A+B)^2 - 2(A+B)A + A^2) \\ &= \mathbb{E}(A^2 + 2AB + B^2 - 2A^2 - 2AB + A^2) \\ &= \mathbb{E}(B^2) = \mathbb{E}((\int_0^t e^{(a-b)(t-s)}d\hat{B}_s)^2).\end{aligned}$$

The Itô isometry (see p. 26 of [26]) gives:

$$\mathbb{E}((\int_0^t e^{(a-b)(t-s)}d\hat{B}_s)^2) = \mathbb{E}(\int_0^t (e^{(a-b)(t-s)})^2 ds)$$

$$= \int_0^t e^{2(a-b)(t-s)}ds.$$

We can solve $\int_0^t e^{2(a-b)(t-s)}ds$ explicitly by calculating its antiderivative,

$$\int_0^t e^{2(a-b)(t-s)}ds = d_0 + \frac{1}{2(a-b)}(e^{2(a-b)t} - 1).$$

Hence, we recover the expectation for $F_t(\omega)$,

$$\mathbb{E}(F_t(\omega)) = F_0(\omega)e^{(a-b)t} + \frac{b}{(a-b)}(e^{(a-b)t} - 1), \tag{11}$$

and the corresponding variance (integration constant d_0 set to zero),

$$\mathrm{Var}(F_t(\omega)) = \frac{1}{2(a-b)}(e^{2(a-b)t} - 1). \tag{12}$$

2.2 The parameters a and b

We note that

$$0 < \mathbb{E}(F_t(\omega)) = F_0(\omega)e^{(a-b)t} + \frac{b}{(a-b)}(e^{(a-b)t} - 1) < 1. \tag{13}$$

For $t = 0$ we see from condition (13) that $0 < F_0(\omega) < 1$. For $(a-b) > 0$ $e^{(a-b)t}$ approaches inifinity so this condition is not reasonable. We are therefore left with the condition $(a-b) \leq 0$. Since $0 < F_0 < 1$ we get

$$0 < F_0(\omega)e^{(a-b)t} + \frac{b}{(a-b)}(e^{(a-b)t} - 1) < 1$$

Letting $t \to \infty$ we see that

$$0 < \frac{b}{b-a} < 1$$

which implies that $b > 0$ and that $a < 0$. For $a = b$ the bounded variation term $A(t)$ in eq.(9) collapses into a linear equation:

$$\begin{aligned} A(t) &= \frac{b}{a-b}(e^{(a-b)t} - 1) \\ &= \frac{b}{a-b}(1 + (a-b)t + \frac{(a-b)^2t^2}{2!} + \cdots + \frac{(a-b)^nt^n}{n!} + \cdots - 1) \\ &= b(\frac{1}{a-b} + t + \frac{(a-b)^1t^2}{2!} + \cdots + \frac{(a-b)^{n-1}t^n}{n!} + \cdots - \frac{1}{a-b}) \\ &= b(t + \frac{(a-b)^1t^2}{2!} + \cdots + \frac{(a-b)^{n-1}t^n}{n!} + \cdots) \\ &= bt \end{aligned} \tag{14}$$

We will henceforth assume that $F_0 > 0$ and $(a-b) < 0$.

2.3 The martingale term

We use Gaussian white noise to model perturbations in the AT→GC (a) and GC→AT (b) mutation rates. We also allow for scaling of $c > 0$, as mentioned above. The scale can be determined by factors such as species/strain, environment, host and presence of MMR genes. The martingale term,

$$\int_0^t e^{(a-b)(t-s)} d\hat{B}_s, \tag{15}$$

depends on the parameters a and b as well as on the duration of the time period. Since we assume that $(a-b) < 0$, the martingale term approaches 0 as $t \to \infty$ and Brownian motion $\hat{B}_t(\omega)$ for $a = b$. For $(a-b) < 0$ it can be seen that (15) increases as $s \to t$.

We can reach the same conclusion by examining the variance of $F_t(\omega)$ (described in (12) above). The Brownian motion is assumed to have mean $\mu = 0$ and variance $\mathbb{E}(\hat{B}_t^2(\omega)) = t$. Thus, the variance of Browninan motion is in general expected to increase with time t. Since there is no simple way to calculate the integral in (15) analytically, we do so numerically:

$$\int_0^t e^{(a-b)(t-s)} d\hat{B}_s = \sum_{s_0}^{s_N} e^{(a-b)(t-s_i)} (\hat{W}_{s_{i+1}}(\omega) - \hat{W}_{s_i}(\omega)) \Delta s_i, \tag{16}$$

where $\hat{W}_s(\omega)$ is c-scaled white noise, $\Delta s_i = s_{i+1} - s_i$, and $s_0 = 0, \ldots, s_i = t_i, \ldots, s_N = t$.

2.4 The Girsanov transform

Equation (7) can be written as

$$dF_t(\omega) = ((a-b)F_t(\omega) + b)dt + d\hat{B}_t(\omega).$$

Since we know from (9) that

$$F_t(\omega) = -\frac{b}{(a-b)} + (F_0(\omega) + \frac{b}{(a-b)})e^{(a-b)t} + \int_0^t e^{(a-b)(t-s)} d\hat{B}_s.$$

If we let

$$Z_t(\omega) = \int_0^t e^{(a-b)(t-s)} d\hat{B}_s,$$

then $Z_t(\omega)$ is an Itô process (see [26]). After some rearrangements we can set

$$K_t(\omega) = (a-b)((F_0(\omega) + \frac{b}{a-b})e^{(a-b)t} + Z_t(\omega)),$$

and since $Z_t(\omega)$ is a martingale, we know from the Doob–Meyer decomposition that $K_t(\omega)$ is also a martingale. We can thus write

$$dF_t(\omega) = K_t(\omega)dt + d\hat{B}_t(\omega).$$

The Girsanov theorem allows us to compute the Radon–Nikodym derivative (see ch. 3, p. 146 of [25]) of a measure Q with respect to the probability measure P as follows:

$$\frac{dQ}{dP} = exp(-\int_0^t K_s(\omega)d\hat{B}_s - \frac{1}{2}\int_0^t K_s^2(\omega)ds).$$

This means that $F_t(\omega)$ is a Brownian motion under the measure Q, since we assume that $(a-b) < 0$ which implies that Kazamaki's (and hence Novikov's condition) apply $\forall t$ (see chs. 4 and 8 of [26]).

2.5 Further generalizations

The model describing SNP GC content can be made more general if we assume that the parameters a and b are functions. It is important to note that if a and b are functions with respect to time, obtaining an analytical solution may be impossible. While up to this point we have assumed that variation in the model is described by a white noise process, a more complicated noise term X_t could also be used. For instance, if we let

$$\frac{dF_t(\omega)}{dt} = (a + X_t(\omega))F_t(\omega) + (b + X_t(\omega))(1 - F_t(\omega)),$$

we have

$$\frac{dF_t(\omega)}{dt} = aF_t(\omega) + X_t(\omega)F_t(\omega) + b + X_t(\omega) - (b + X_t(\omega))F_t(\omega).$$

This reduces to

$$\frac{dF_t(\omega)}{dt} = (a - b)F_t(\omega) + b + X_t(\omega),$$

where

$$X_t(\omega) = \theta(t,\omega) + \kappa(t,\omega)\hat{W}_t(\omega).$$

Thus,

$$\frac{dF_t(\omega)}{dt} = aF_t(\omega) + b(1 - F_t(\omega)) + (\theta(t,\omega) + \kappa(t,\omega)\hat{W}_t(\omega)),$$

and after rearranging:

$$dF_t(\omega) = ((a - b)F_t(\omega) + \theta(t,\omega) + b)dt + \kappa(t,\omega)dB_t(\omega). \tag{17}$$

We could, for instance, let $X_t(\omega)$ be a mean-reverting Ornstein–Uhlenbeck process, *i.e.*

$$\frac{dX_t(\omega)}{dt} = GC_0 - F_t(\omega) + \hat{W}_t(\omega).$$

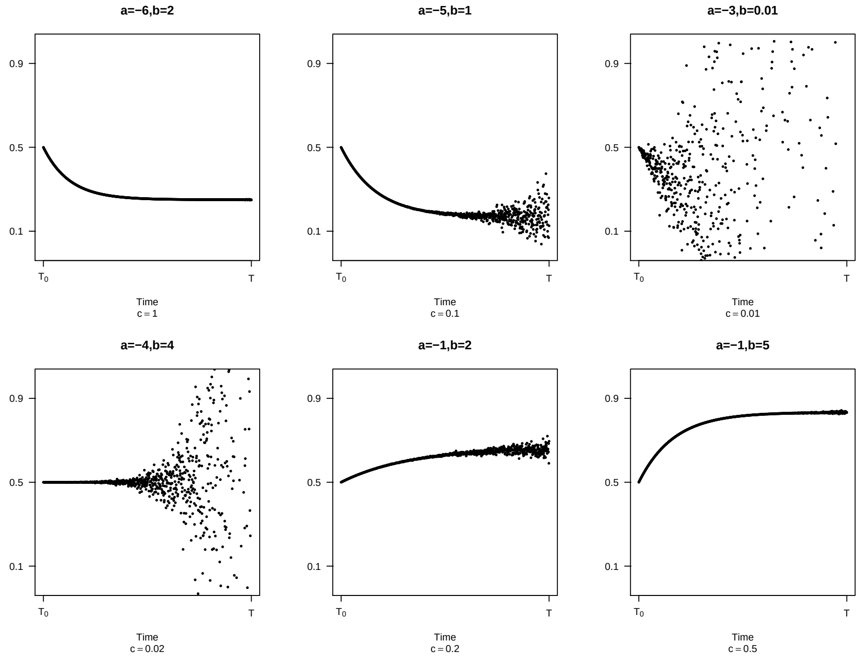

Figure 1: The model (9) with different combinations of parameters a, b and Brownian scaling coefficient c, all starting at $F_0 = 0.5$. The vertical axis describes SNP GC content, while the horizontal axis describes time t from $T_0 = 0$ to $T = 1$.

Hence, we let $\theta(t, \omega) = GC_0 - F_t(\omega)$ and $\kappa(t, \omega) = 1$. Plugging these into (17), we see

$$dF_t(\omega) = ((a - b)F_t(\omega) + GC_0 + b)dt + d\hat{B}_t(\omega). \tag{18}$$

We can now use the integrating factor $g(t, F_t(\omega)) = Y_t(\omega) = e^{(-(a-b)t)}F_t(\omega)$ to solve (18) in a similar fashion to (8).

3 Results and Discussion

Equation (9) describes a model for core genome SNP GC content in prokaryotes. It obeys Chargaff's parity laws [27]. As discussed in section 1, SNPs are subject to natural selection, which is in turn mediated by the environment of the organism(s) at hand. While a drop in SNP GC content could indicate relaxed selective pressures with ensuing mutations from genetic drift and AT mutational bias [7], increased negative or purifying selection may favor GC-biased SNPs [1, 17, 18, 19]. Selective pressure for improved fitness could also lead to increased GC content [15]. Carbon starvation [41] and/or nitrogen abundance [16] have also been found to have an effect on genomic base composition.

All in all, microbial organisms in the same environments often acquire the same nucleotide biases if enough time is allowed to pass [12, 13]. Such environmental signatures become particularly evident in SNPs since, as discussed above and in section 1, these polymorphisms arise as a consequence of natural selection regulated by the environment. By fitting (9) to empirical data with either non-linear regression models or Bayesian inference, we can estimate the relative proportions of mutation from AT$\rightarrow$GC (a) and GC$\rightarrow$AT (b) over time. The model described by (1) estimates analogous parameters for AT$\rightarrow$GC (α) and GC$\rightarrow$AT (β), but with respect to core genome GC content rather than time.

3.1 The accumulating effects of stochastic processes

Figure 1 shows different paths of SNP GC content with respect to time; increase, decrease and stasis for various Brownian scaling coefficients c.

All stochastic fluctuations observed in the curves in figure 1 are a consequence of the Brownian motion term (15). These SNP GC content curves all become more unstable as time passes, to varying degrees depending on c. The mathematical mechanisms behind these stochastic fluctuations are outlined in section 2.3. Equation (15) indicates that both the mutation parameters a and b are responsible for how the stochastic fluctuations progress with respect to time (see also section 2.2).

Some of the paths in figure 1 initially exhibit barely visible stochastic fluctuations, but these grow in magnitude as the SNP GC content mutation rates start to vary out of control, especially for low values of c.

The progression of the stochastic fluctuations in the SNP GC content curves for low c is not at all expected *a priori*. Below, we demonstrate that the nature of the abruptly exploding mutation rates is supported both in theory [22, 23, 24] and in practice. In the following examples, we focus on mechanisms resulting in genome reduction [8, 9] and subsequent AT bias in the base composition of many symbionts. We end this section with a case study utilizing real genetic data on SNP GC content in different lineages of the fish pathogen *R. salmoninarum*.

3.2 Evolution of microbial obligate symbionts

Free-living bacteria that develop a sustained symbiotic relationship with a host [34] will often, over time, undergo genome reduction [8]. This process of genome reduction is preceded by a phase of pseudogenization, in which the symbiont's genome retains its usual size [29], but the genes are not under selective constraints imposed by the host and thus become abundant. Accumulated mutations eventually render many genes defective [9, 28]. The process of pseudogenization may drag on for a long time [30], but eventually non-expressed genetic regions will be excised due to energy economization [21], lack of recombination, and/or the absence of streamlining due to low population density and reduced selective pressure from the environment [10, 22]. The first genes lost are typically those least conserved within a species [35]. It is only after a continuous symbiotic relationship and the pseudogenization phase that a drop in genomic GC content seems to occur, most likely because of the loss of MMR genes that counter the AT mutational bias [8].

After the decrease in genomic GC content, there is usually no return to a free-living lifestyle for the bacterium [30]. Non-recombining symbionts ultimately disintegrate, as described by the concept of Muller's ratchet [22]. According to some recent findings, the host, which eventually becomes dependent on the symbiont, can establish similar relationships with other bacteria [28, 33, 30, 9, 34].

Intracellular pathogens, on the other hand, do not appear to engage in symbiotic relationships with a host, most likely due to the increased constraints of a pathogen-host relationship [36]. As such, though these pathogens may undergo genome reduction, they do not seem to experience the same dramatic gene loss observed in some symbionts, which are reduced to mere organelles in their hosts [28]. It is not uncommon, however, for the genomic base composition of intracellular pathogens to be AT-biased [36].

There do appear to be some similarities between the evolutionary mechanisms of symbionts and those of free-living bacteria that undergo changes in environment even if not through attachment to a host. There are only a few documented examples of free-living bacteria that experience genome reduction with subsequent genomic AT bias after a change in environment/niche. One of these is the cyanobacterium *Prochlorococcus spp.* [31], whose high-light ecotypes living close to the water surface are more AT-rich and have smaller genomes than the low-light ecotypes living at greater depths [32]. Indeed, genomic GC content and genome size increase, respectively, from 30.8% and 1.66 megabase pairs (Mbp) in the high-light strains to 50.0% and 2.68 Mbp in the low-light strains [32].

3.3 Modeling AT bias in microbial genomes

As mentioned above, microbial genomes appear to become more AT-rich after the loss of MMR genes, regardless of niche and/or environment. This is most likely due to AT mutational bias [7], and it may be mediated by genetic drift in light of relaxed selective pressures [8].

The model in (15) was formulated in a recent study [2] that assumed that the change in SNP GC content with respect to core genome GC content was a constant multiple of SNP GC content and another constant multiplied by SNP AT content. In the present study, we investigated how SNP GC content evolves over time, allowing for stochastic fluctuations. We modeled these fluctuations using a Gaussian white noise process $\hat{W}_t(\omega)$, which is subject to a scaled $c > 0$, in the AT→GC and GC→AT mutation rates. We introduced the scaling to account for differences between species, environment/niche, and selective pressures or lack thereof.

From figure 1, it can be seen that the mutation rates remain fairly stable, at least in the increasing and decreasing curves, before abruptly fluctuating out of control. Once we decrease c the random fluctuations start sooner and escalate a bit more. Since the mutation rates fluctuate so drastically as time $t \to T$, it is natural to expect that the outcome predicted by Muller's ratchet will be achieved [22], *i.e.* that the bacterial population will go extinct. However, (9) suggests that although the random fluctuations start relatively late, the species' fate may be sealed far earlier, before any stochastic fluctuation can be observed.

A loss of MMR genes could imply that the scaling parameter c adds more weight to the martingale term (2.1), which triggers the amplification of the stochastic fluctuations. However, the similarity of the mutation rate parameters a and b can also influence the magnitude of the stochastic fluctuations. Indeed, from (15), it can be seen that low mutation rates magnify the effect of the martingale term as $a - b \to 0$, since $e^{(a-b)} \to 1$.

3.4 Connections with theories from evolutionary biology

Leigh Van Valen wanted a model to confirm that extinction rates correlate with age in the fossil record. However, after testing this hypothesis, he found no such correlation [23]. Thus, he formulated the Red Queen hypothesis, taking its name from Lewis Carroll's 1871 book *Through the Looking-Glass, and What Alice Found There.* In that book, the Red Queen utters to Alice about the nature of Looking-Glass Land, "Now, here, you see, it takes all the running you can do, to keep in the same place."

Later on, the Red Queen hypothesis was expanded to account for molecular data as well [24]. The model presented in (9) demonstrates related mechanisms for prokaryotes and sheds light on the case of microbial symbionts that have undergone genome reduction with a subsequent drop in GC content. If mutations are not kept in check, extinction will ensue. In other words, the martingale term (15) must be kept as low as possible in order to avoid the random fluctuations that lead to extinction. Since the choice of c is shaped by factors such as species, environment, host and mutation rates, extinction rates will differ between populations, as predicted by the Red Queen hypothesis (See also Figure 1). Furthermore, non-recombining clonal organisms will sooner or later accumulate deleterious mutations that decrease the organisms' fitness to the point of driving their species to extinction.

In the previous section, we discussed how microbial symbionts undergo genome reduction together with a drop in GC content, most likely as a consequence of lost MMR genes. The genomes of these symbionts eventually

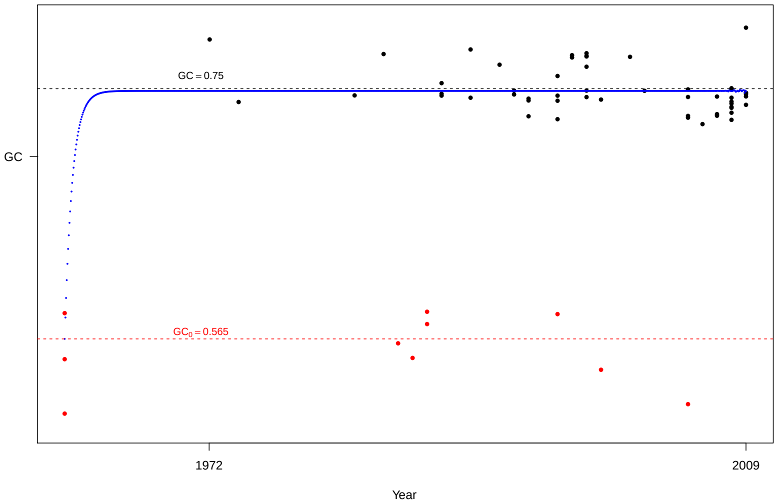

Figure 2: *R. salmoninarum* lineage 1a SNP GC content (vertical axis) plotted against year (horizontal axis). SNP GC content of lineages 1b and 2 is similar to *R. salmoninarum* genomic GC content (red line).

disintegrate due to accumulated hitchhiking effects [37] and genetic drift, as posited by Muller's ratchet [22]. There are experimental findings to support these hypotheses [44]. Our model in (9) provides insight to this by delineating the stochastic fluctuation in mutation rates that will ultimately spiral out of control, depending on the mutation parameters a and b and on the scaling parameter c.

3.5 Mutation rates in *R. salmoninarum*

The fish pathogen *R. salmoninarum* is the causative agent of bacterial kidney disease (BKD), which predominantly afflicts salmonoids. *R. salmoninarum* belongs to the GC-rich, gram-positive Actinobacteria family. It is an intracellular pathogen with a genome size of approximately 3.15 Mbp and a genomic GC content of 56.5%. Its genome is remarkably well conserved and thus appears not to recombine [20].

In a previous publication, we examined the SNP GC content over time of *R. salmoninarum* lineage 1, consisting of sublineages 1a (isolated from North America and Europe) and 1b (isolated from North America), and of lineage 2 (isolated from the UK and Norway) [20]. While sublineage 1b (3 isolates) and lineage 2 (7 isolates) are more endemic to particular environments, sublineage 1a (52 isolates) is widely dispersed across North America and Norway. We found that SNP GC content in sublineage 1b and in lineage 2 was equal to 56.5%, *i.e.* the genomic GC content. The SNP GC content of sublineage 1a, however, was approximately 75% (see figure 2).

Therefore, we here set $c = 1$ and $F_0 = 0.565$ to correspond to the genomic GC content. We estimate a and b using Bayesian inference, selecting non-informative uniform distributions as priors for both parameters. The median posterior estimates are $a = -22.668$ and $b = 67.421$, which suggests that the martingale term (15) holds minimal influence. Furthermore, the GC→AT mutation rate is substantially higher than the AT→GC mutation rate; there is a ratio of almost 3:1 between them ($67.421/22.668 = 2.98$), which is highly unusual [3].

Though the increased SNP GC content is puzzling, it may indicate that sublineage 1a is subject to stronger selective pressure than sublineage 1b and lineage 2 [17, 19], since recombination is not known to take place in *R. salmoninarum*. Recent publications argue that nitrogen abundance and carbon starving, which can occur at great ocean depths, may push for increased GC content [16, 41], but further research is needed before any conclusion can be drawn.

4 Conclusions

We have presented a mathematical model that describes change in SNP GC content over time as a function of mutation parameters a and b. The model contains a stochastic term that describes how minuscule, random changes in mutation rates early on can lead to abrupt, disastrous fluctuations later. We treated examples of this phenomenon in host-associated and symbiotic bacteria.

The model, with its incorporated stochastic term and corresponding scaling parameter c, shows remarkable congruence with at least some parts of the Red Queen hypothesis. In the model, c must not be too large to avoid genomic disintegration. Varying c among species implies that the lifespan of a species need not correlate with the time to its extinction. Furthermore, the model demonstrates how Muller's ratchet operates and suggests that extinction may occur rapidly, depending on c, as opposed to arising slowly.

When we applied the model to different lineages of *R. salmoninarum*, we found that one sublineage, 1a, exhibited a dramatic increase of approximately 20% in SNP GC content, while no differences were detected in sublineage 1b or lineage 2. Dramatic drops in genomic GC content have been documented in both host-associated and free-living bacteria; increases in genomic GC content are less common. The substantial rise in SNP GC content observed in sublineage 1a may thus be the start of a process leading to increased genomic GC content. Since recombination is absent, or at least very rare, in *R. salmoninarum*, the increase in SNP GC content may also indicate that the genome is subject to increased selective pressure, which drives the genomic GC content upwards. Alternatively, sublineage 1a could have moved to a different environmental niche than sublineage 1b and lineage 2, one in which carbon is scarce [41].

Our use of stochastic differential equations, which allow for deterministic modeling of random processes, revealed how bacterial mutation rates may be influenced by stochastic fluctuations. Although simple, the model described here provides novel insight into evolutionary processes with mathematical rigour.

5 Materials and Methods

The genomes utilized in our study were taken from a previous publication [20]. They are all available from the European Bioinformatics Institute (accession number: PRJEB4487). The genomic data files were assembled using MAQ 0.7.1 [38] against the reference *R. salmoninarum* ATCC33209 (NCBI accession number: NC 010168.1), as described in [20].

In the present study, 6 isolates were excluded due to missing date information or poor assembly quality. The removed isolates were Rs3, 5223, 684, MT3106, Cow-chs-94 and NCIMB 1111 (see [20] for details). SNPs were extracted using parSNP from HarvestTools [39], and Seaview [42] was used to examine the base composition of the SNPs to confirm that Chargaff's parity laws were followed, *i.e.* to verify that there were approximately similar numbers of A's and T's and of G's and C's. Both sublineage 1a (52 isolates, > 1400 SNPs) and sublineage 1b (3 isolates, > 400 SNPs) conformed to these rules within 2%, while a 5% deviation was found for lineage 2 (7 isolates, > 500 SNPs). This deviation could be due to recent mutations, natural selection and/or sequencing/assembly errors, as SNP base composition was similar to genomic GC content (*i.e.* 56.5%).

All figures were generated and statistical analyses performed in R [40]. Bayesian parameter estimates were obtained using JAGS [43]. Non-informative uniform priors from –100 to 100 were assumed for both a and b, and model precision was set to 1.0E–2. The Markov chain ran for 5,000,000 iterations with thinning set to 1,000. 12,500 iterations were saved. All chains converged.

References

[1] Rocha, E. P., & Feil, E. J. (2010). *Mutational patterns cannot explain genome composition: are there any neutral sites in the genomes of bacteria?* PLoS genetics, **6(9)** e1001104

[2] Bohlin, J., Eldholm, V., Brynildsrud, O., Petterson, J. H. O., & Alfsnes, K. *Modeling of the GC content of the substituted bases in bacterial core genomes.* BMC genomics, **19(1)** 589 (2018)

[3] Bohlin, J., Rose, B., & Petterson, J. H. O. *Estimation of AT and GC content distributions of nucleotide substitution rates in bacterial core genomes.* Big Data Analytics, – — (2019)

[4] Bohlin, J., Sekse, C., Skjerve, E., & Brynildsrud, O. *Positive correlations between genomic % AT and genome size within strains of bacterial species.* Environmental microbiology reports, **6(3)** 278-286 (2014)

[5] Agashe, D., & Shankar, N. *The evolution of bacterial DNA base composition.* Journal of Experimental Zoology Part B: Molecular and Developmental Evolution, **322(7)** 517-528 (2014)

[6] Bentley, S. D., & Parkhill, J. *Comparative genomic structure of prokaryotes.* Annu. Rev. Genet., **38** 771-791 (2004)

[7] Hershberg, R., & Petrov, D. A. *Evidence that mutation is universally biased towards AT in bacteria.* PLoS genetics, **6(9)** e1001115 (2010)

[8] McCutcheon, J. P., & Moran, N. A. *Extreme genome reduction in symbiotic bacteria.* Nature Reviews Microbiology, **10(1)** 13 (2012)

[9] Wernegreen, J. J. *In it for the long haul: Evolutionary consequences of persistent endosymbiosis.* Current opinion in genetics & development, **47** 83-90 (2017)

[10] Lynch, M., Ackerman, M. S., Gout, J. F., Long, H., Sung, W., Thomas, W. K., & Foster, P. L. *Genetic drift, selection and the evolution of the mutation rate.* Nature Reviews Genetics, **17(11)** 704 (2016)

[11] Kimura, M. *A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences.* Journal of Molecular Evolution, **16(2)** 111-120 (1980)

[12] Foerstner, K. U., Von Mering, C., Hooper, S. D., & Bork, P. *Environments shape the nucleotide composition of genomes.* EMBO reports, **6(12)** 1208-1213 (2005)

[13] Reichenberger, E. R., Rosen, G., Hershberg, U., & Hershberg, R. *Prokaryotic nucleotide composition is shaped by both phylogeny and the environment.* Genome biology and evolution, **7(5)** 1380-1389 (2015)

[14] Lind, P. A., & Andersson, D. I. *Whole-genome mutational biases in bacteria.* Proceedings of the National Academy of Sciences, **105(46)** 17878-17883 (2008)

[15] Raghavan, R., Kelkar, Y. D., & Ochman, H. *A selective force favoring increased G+C content in bacterial genes.* Proceedings of the National Academy of Sciences, **109(36)** 14504-14507 (2012)

[16] Seward, E. A., & Kelly, S. *Dietary nitrogen alters codon bias and genome composition in parasitic microorganisms.* Genome biology, **17(1)** 226

[17] Hildebrand, F., Meyer, A., & Eyre-Walker, A. (2010). *Evidence of selection upon genomic GC-content in bacteria.* PLoS genetics, **6(9)** e1001107 (2016)

[18] Bohlin, J., Eldholm, V., Pettersson, J. H., Brynildsrud, O., & Snipen, L. *The nucleotide composition of microbial genomes indicates differential patterns of selection on core and accessory genomes.* BMC genomics, **18(1)** 151 (2017)

[19] Bobay, L. M., & Ochman, H. *Impact of recombination on the base composition of bacteria and archaea.* Molecular biology and evolution, **34(10)** 2627-2636 (2017)

[20] Brynildsrud, O., Feil, E. J., Bohlin, J., Castillo-Ramirez, S., Colquhoun, D., McCarthy, U., ... & Verner-Jeffreys, D. W. *Microevolution of Renibacterium salmoninarum: evidence for intercontinental dissemination associated with fish movements.* The ISME journal, **8(4)** 746 (2014)

[21] Lane, N., & Martin, W. *The energetics of genome complexity.* Nature, **467(7318)** 929 (2010)

[22] Moran, N. A. *Accelerated evolution and Muller's rachet in endosymbiotic bacteria.* Proceedings of the National Academy of Sciences, **93(7)** 2873-2878 (1996)

[23] Van Valen, L. *A new evolutionary law.* Evol Theory, **1** 1-30 (1973)

[24] Van Valen, L. *Molecular evolution as predicted by natural selection.* Journal of molecular evolution, **3(2)** 89-101 (1974)

[25] Protter, P. E. *Stochastic differential equations* Springer, Berlin, Heidelberg

[26] Øksendal, B. (2003). *Stochastic differential equations.* Springer, Berlin, Heidelberg (2005)

[27] Elson, D., & Chargaff, E. *Regularities in the composition of pentose nucleic acids.* Nature, **173(4413)** 1037 (1954)

[28] Moran, N. A., & Bennett, G. M. *The tiniest tiny genomes.* Annual review of microbiology, **68** 195-215 (2014)

[29] Klasson, L. *The unpredictable road to reduction.* Nature ecology & evolution, **1(8)** 1062 (2017)

[30] Hosokawa, T., Ishii, Y., Nikoh, N., Fujie, M., Satoh, N., & Fukatsu, T. *Obligate bacterial mutualists evolving from environmental bacteria in natural insect populations.* Nature microbiology, **1(1)** 15011 (2016)

[31] Martínez-Cano, D. J., Reyes-Prieto, M., Martínez-Romero, E., Partida-Martínez, L. P., Latorre, A., Moya, A., & Delaye, L. *Evolution of small prokaryotic genomes.* Frontiers in microbiology, **5** 742 (2015)

[32] Batut, B., Knibbe, C., Marais, G., & Daubin, V. *Reductive genome evolution at both ends of the bacterial population size spectrum.* Nature Reviews Microbiology, **12(12)** 841 (2014)

[33] Wernegreen, J. J. *Endosymbiont evolution: predictions from theory and surprises from genomes.* Annals of the New York Academy of Sciences, **1360(1)** 16-35 (2015)

[34] Boscaro, V., Kolisko, M., Felletti, M., Vannini, C., Lynn, D. H., & Keeling, P. J. *Parallel genome reduction in symbionts descended from closely related free-living bacteria.* Nature ecology & evolution, **1(8)** 1160 (2017)

[35] Bolotin, E., & Hershberg, R. *Bacterial intra-species gene loss occurs in a largely clocklike manner mostly within a pool of less conserved and constrained genes.* Scientific reports, **6** 35168 (2016)

[36] Weinert, L. A., & Welch, J. J. *Why might bacterial pathogens have small genomes?* Trends in ecology & evolution, **32(12)** 936-947 (2017)

[37] Smith, J. M., & Haigh, J. *The hitch-hiking effect of a favourable gene.* Genetics Research, **23(1)** 23-35 (1974)

[38] Li, H., Ruan, J., & Durbin, R. *Mapping short DNA sequencing reads and calling variants using mapping quality scores.* Genome research, gr-078212 (2008)

[39] Treangen, T. J., Ondov, B. D., Koren, S., & Phillippy, A. M. *The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes.* Genome biology, **15(11)** 524 (2014)

[40] R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[41] Hellweger, F. L., Huang, Y., & Luo, H. *Carbon limitation drives GC content evolution of a marine bacterium in an individual-based genome-scale model.* The ISME journal, **12(5)** 1180 (2018)

[42] Gouy, M., Guindon, S., & Gascuel, O. *SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building.* Molecular biology and evolution, **27(2)** 221-224 (2009)

[43] Plummer, M. *JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling.* In Proceedings of the 3rd international workshop on distributed statistical computing, **Vol. 124** No. 125.10 (2003)

[44] Woods, R. J., Barrick, J. E., Cooper, T. F., Shrestha, U., Kauth, M. R., & Lenski, R. E. *Second-order selection for evolvability in a large Escherichia coli population.* Science, **331(6023)** 1433-1436 (2011)