Alma Mater Studiorum - Università di Bologna

DOTTORATO DI RICERCA IN

FISICA

Ciclo 33

Settore Concorsuale: 02/D1 - FISICA APPLICATA, DIDATTICA E STORIA DELLA FISICA

Settore Scientifico Disciplinare: FIS/07 - FISICA APPLICATA A BENI CULTURALI, AMBIENTALI, BIOLOGIA E MEDICINA

STOCHASTIC MODELING AND CORRELATION ANALYSIS OF OMICS DATA

Presentata da: Iva Budimir

Coordinatore Dottorato

Michele Cicoli

Supervisore

Gastone Castellani

Esame finale anno 2021

Acknowledgements

I am incredibly grateful for the opportunity to be one of the Marie Skłodowska-Curie Early Stage Researchers in the IMforFUTURE[1] project. It was an extraordinary experience which set me on the path of my research and gave me opportunity to meet fantastic people and see fantastic places.

I wish to express my deepest gratitude to my supervisor, Prof. Gastone Castellani, for his kind support and invaluable guidance and to Dr. Claudia Sala for her continuous help, patience and kindness.

I won't forget the insightful suggestions of my collegues, Prof. Daniel Remondini, Dr. Enrico Giampieri and Dr. Silvia Vitale, which helped in my research and here I wish to thank them.

The support of my family and friends was important to maintain my motivation even in hard times. I especially thank to my parents who were always there for me.

[1]The research leading to the results presented in this thesis was funded by the European Union's Horizon 2020 research and innovation programme IMforFUTURE (Innovative Training in Methods for Future Data) under H2020-MSCA-ITN grant agreement number 721815.

Abstract

We studied the properties of three different types of omics data: protein domains in bacteria, gene length in metazoan genomes and methylation in humans. Gene elongation and protein domain diversification are some of the most important mechanisms in the evolution of functional complexity. For this reason, the investigation of the dynamic processes that led to their current configuration can highlight the important aspects of genome and proteome evolution and consequently of the evolution of living organisms. The potential of methylation to regulate the expression of genes is usually attributed to the groups of close CpG sites. We performed the correlation analysis to investigate the collaborative structure of all CpGs on chromosome 21.

The long-tailed distributions of gene length and protein domain occurrences were successfully described by the stochastic evolutionary model and fitted with the Poisson Log-Normal distribution. This approach included both demographic and environmental stochasticity and the Gompertzian density regulation. The parameters of the fitted distributions were compared at the evolutionary scale. This allowed us to define a novel protein-domain-based phylogenetic method for bacteria which performed well at the intraspecies level. In the context of gene length distribution, we derived a new generalized population dynamics model for diverse subcommunities which allowed us to jointly model both coding and non-coding genomic sequences. A possible application of this approach is a method for differentiation between protein-coding genes and pseudogenes based on their length.

General properties of the methylation correlation structure were firstly analyzed for the large data set of healthy controls and later compared to the Down syndrome (DS) data set. The CpGs demonstrated strong group behaviour even across the large genomic distances. Detected differences in DS were surprisingly small, possibly caused by the small sample size of DS which reduced the power of statistical analysis.

Contents

Introduction

Biological entities from different omics disciplines can be described as dynamics systems. The omics measurements thus merely record the state of the changing system at the specific time point. Depending on the system and the feature which is being studied, the timescale of change can vary from seconds to thousands of years. As an example, the genome of an organism is stable during its lifetime but on a larger scale, it results from a myriad of changes which accumulated during the course of evolution. On the other hand, transcriptome or proteome content of the cell is constantly changing and consequently, the results of both transcriptomics and proteomics experiments are time-sensitive. To better understand both the current state of a dynamic system and dynamic processes which led to it, we develop the mathematical model of the system.

The mathematical model aims to describe the elements of a system, their states and their interactions. The model should be sufficiently precise as the representation of the system but at the same time simple enough to be solvable. Thus to successfully model complex systems one has to leave out many details. Viewed at this higher level, even the systems which are intrinsically deterministic have stochastic behaviour where stochasticity serves as a "container" for all unaccounted components of the system. Due to their complexity, the dynamics of biological systems thus have to be stochastically modeled [1].

We focus on three different biological systems with the origin in genomics, proteomics and methylomics. In Part I, we study the system of protein domains in bacteria where we observe the distribution of protein domain abundances. Similarly, in Part II, we observe the distribution of gene length in different metazoan species. These distributions are important as a measure of the diversity of the system recorded at a certain time point of the evolution. Following a different approach in Part III, instead of the dynamical properties, we model the interactions between the elements of the methylation system.

As we will observe in Part I and Part II, the distribution of protein domain abundances, as well as the gene length distribution, is long-tailed, with a large number of occurrences far from the centre of the distribution. It is known that long-tailed distribution is ubiquitous throughout genomic biology [2]. The occurrence of protein families and the number of transcripts per protein family are some examples of long-tailed distributions which are successfully fitted with the power-law function

[2]. The same type of distribution emerges as the relative species abundance (RSA) distribution in the context of ecological theories [3]. The RSA distribution has been extensively studied in the last 80 years resulting in great theoretical developments [3]. Thus, instead of developing a new framework to explain the mechanisms which led to the protein domain and the gene length distribution, we interpret them in the context of population dynamics. Specifically, we employ the model introduced by Engen and Lande in 1996 [4, 5]. They developed a stochastic model which includes both additive (demographic) and multiplicative (environmental) stochasticity and where, depending on the assumptions, the RSA distribution is Poisson Log-Normal or Negative Binomial. With respect to their stochastic model, the process of gene elongation, as well as the emergence of protein domains, are explained by the stochastic growth equation.

In the process of methylation the methyl group is covalently added to the cytosine (C) nucleotide and this epigenetic modification is well-studied for its potential to regulate gene expression. The cytosine methylation predominantly occurs at the CG dinucleotide sequences, and thus the target of methylomics studies are these so-called CpG sites. One of the available tools, Infinium 450K assay [6], measures the methylation level of more than 480 000 CpG sites spread across the genome. Groups of CpG sites are usually analyzed together [7, 8] as it is believed that the regulatory potential results from their group behaviour. Thus instead of a $\sim$480 000 independent entities, in the Part III, we observe the system of CpG sites and we aim to describe the interactions among them.

Stochastic processes

The materials in this first introductory chapter, unless stated otherwise, are taken from two textbooks about stochastic processes written by Karlin and Taylor [9, 10].

0.1 Elements of stochastic processes

A stochastic (or random) process is a family of random variables $\{X(t); t \in T\}$ indexed by a parameter t. To properly define a stochastic process, we need to specify the *state space* S, the *index parameter* T and the dependence relations among the random variables $X(t)$.

The state space S is the space in which all possible values of random variables $X(t)$ lie. If $S = \mathbb{N}_0 = \{0, 1, 2, \ldots\}$, then we call $X(t)$ a discrete state process. If $S = \mathbb{R}$, then $X(t)$ is a real-valued stochastic process. As with the codomain of any function, S is not uniquely defined but usually the appropriate choice for S is obvious.

The index $t, t \in T$ is usually interpreted as time parameter of the process and T is the set of all times in which we observe the stochastic process. If $T = \{0, 1, 2, \ldots\}$, then $X(t)$ is a discrete time stochastic process. In this case, we often write X_n instead of $X(t)$. If $T = [0, \infty\rangle$, then $X(t)$ is called a continuous time process.

To completely define a stochastic process, we need to specify all finite-dimensional distributions, that is the joint distributions of the variables $(X(t_1), \ldots, X(t_n))$, for all $t_1, \ldots, t_n \in T$ and for all $n \in \mathbb{N}$. However, if certain special relations among the random variable $X(t)$ hold, then a definition of a stochastic process becomes much more concise. These special relations among the variables define different classes of stochastic processes. Some of the most important classes are *stationary processes* and *Markov processes*, which we will discuss in more details.

Definition (Stationary process). A stochastic process $\{X(t); t \in \mathbb{R}\}$ is a stationary process if

$$P(X(t_1 + h) \leq x_1, \ldots, X(t_n + h) \leq x_n) = P(X(t_1) \leq x_1, \ldots, X(t_n) \leq x_n),$$
$$\text{for all } t_1, \ldots, t_n, h \in \mathbb{R} \text{ and for all } n \in \mathbb{N}. \quad (0.1)$$

In particular, a stationary process has stationary increments. In other words, for any $t, h \in \mathbb{R}$ the distribution of $X(t+h) - X(t)$ depends only on the length of h.

0.2 Markov process

A stochastic process with Markov property "suffers" from memorylessness. Roughly speaking, the future values of the process $X(s), s > t$ conditioned on the present $X(t)$ are not altered by additional knowledge of the past $X(u), u < t$. A precise definition of Markov process is given below.

Definition (Markov process). Markov process is a stochastic process $\{X(t); t \geq 0\}$ with the following property

$$\begin{aligned} &P(a < X(t) \leq b \mid X(t_1) = x_1, X(t_2) = x_2, \ldots, X(t_n) = x_n) = \\ &P(a < X(t) \leq b \mid X(t_n) = x_n), \quad \text{whenever } t_1 < t_2 < \cdots < t_n < t. \end{aligned} \tag{0.2}$$

Markov chain A Markov process with finite or denumerable state space is called a *Markov chain*. If we assume that the Markov chain is stationary, then we can define transition probabilities

$$P_{ij}(h) = P(X(t+h) = j \mid X(t) = i) \tag{0.3}$$

The Markov property asserts that $P_{ij}(h)$ satisfies:

1. $P_{ij} \geq 0$,
2. $\sum_j P_{ij}(h) = 1$,
3. $P_{ik}(h) = \sum_j P_{ij}(h) P_{jk}(h)$ (Chapman-Kolmogorov relation).

If additionally, $P_{ij}(h) = P_{ij}$ doesn't depend on the time difference h, then we have a homogeneous Markov chain.

0.2.1 Poisson process

Definition (Homogenouos Poisson process). Homogenous Poisson process with rate λ is a stochastic process $\{X(t); t \geq 0\}$ on the nonnegative integers which has the following properties:

1. $X(0) = 0$.
2. $\{X(t)\}$ has independent increments, that is, for every pair of disjoint time intervals $[t_1, t_2]$, $[t_1, t_2]$, with $t_1 \leq t_2 < t_3 \leq t_4$, the increments $X(t_4) - X(t_3)$ and $X(t_2) - X(t_1)$ are independent random variables.

3. $P(X(t+h) - X(t) = 0) = 1 - \lambda h + o(h)$ as $h \downarrow 0 \quad (x = 0, 1, 2, \ldots)$.

4. $P(X(t+h) - X(t) = 1) = \lambda h + o(h)$ as $h \downarrow 0$.

5. $P(X(t+h) - X(t) \geq 2) = o(h)$ as $h \downarrow 0$.

Remark. We say that an arbitrary function $f \in o(h)$ if $\lim_{h \downarrow 0} \frac{f(h)}{h} \to 0$.

Markov property The Markov property is a direct consequence of independent increments. For $t_1 < t_2 < \cdots < t_n < t$, we have

$$
\begin{aligned}
&P(X(t) \leq x \mid X(t_1) = x_1, X(t_2) = x_2, \ldots, X(t_n) = x_n) = \\
&P(X(t) - X(t_n) \leq x - x_{t_n} \mid X(t_2) - X(t_1) = x_2 - x_1, \ldots, X(t_n) = x_n) = \\
&P(X(t) - X(t_n) \leq x - x_{t_n} \mid X(t_n) = x_n) = \\
&P(X(t) \leq x \mid X(t_n) = x_n).
\end{aligned}
$$

Inhomogeneous Poisson process If instead of a constant, $\lambda(t) : [0, \infty\rangle \to [0, \infty\rangle$ is an integrable function, then we get an inhomogeneous (or nonhomogeneous) Poisson process. All of the properties (a)-(e) still hold replacing λ with $\lambda(t)$ as does the following relation

$$
X(t+h) - X(t) \sim Poisson\left(\int_t^{t+h} \lambda(\alpha) d\alpha\right). \tag{0.4}
$$

0.2.2 Brownian motion

Definition (Brownian motion). Brownian motion is a stochastic process $\{X(t); t \geq 0\}$ with the following properties:

1. Every increment $X(t+h) - X(t)$ is normally distributed with mean 0 and variance $\sigma^2 h$ with σ^2 being a constant.

2. $\{X(t)\}$ has independent increments, that is, for every pair of disjoint time intervals $[t_1, t_2]$, $[t_1, t_2]$, with $t_1 \leq t_2 < t_3 \leq t_4$, the increments $X(t_4) - X(t_3)$ and $X(t_2) - X(t_1)$ are independent random variables.

3. $X(0) = 0$ and $X(t)$ is almost surely continuous functions of t.

Remark. It can be shown that the paths $X(t)$ of the Brownian motion are nowhere differentiable.

Remark. If $\sigma^2 = 1$, then we are talking about *standard Brownian motion.* An arbitrary Brownian motion $X(t)$ with variance parameter σ^2 can easily be transformed to standard Brownian motion with $\tilde{X}(t) = X(t)/\sigma$. In the following discussions, we will use notation $B(t)$ for the standard Brownian motion.

Markov property Similar as in the case of the Poisson process, the Markov property is a direct consequence of independent increments.

0.3 Diffusion process and stochastic differential equation

0.3.1 Diffusion process

Diffusion processes constitute an important class of stochastic processes and are widely used in modeling of many physical, biological, economic or social phenomena. They can be defined in different ways, but here we use the following definition.

Definition (Diffusion process). Diffusion process is a Markov process $\{X(t); t \geq 0\}$ whose state space is an interval I with endpoints $-\infty \leq l < r \leq \infty$ with the following properties:

1. $\lim\limits_{h \downarrow 0} \frac{1}{h} P\left(|X(t+h) - x| > \epsilon \mid X(t) = x\right) = 0$ for every $\epsilon > 0$ and for all $x \in I$.
2. $\lim\limits_{h \downarrow 0} \frac{1}{h} E\left[X(t+h) - X(t) \mid X(t) = x\right] = \mu(x,t)$, where $\mu(x,t)$ is a continuous function of x and t.
3. $\lim\limits_{h \downarrow 0} \frac{1}{h} E\left[\left(X(t+h) - X(t)\right)^2 \mid X(t) = x\right] = \sigma^2(x,t)$, where $\sigma^2(x,t) \geq 0$ is a continuous function of x and t.

The function $\mu(x,t)$ is called the *infinitesimal mean* or *drift parameter*, and the function $\sigma^2(x,t)$ is called *infinitesimal variance* or *diffusion parameter*. In the following discussion, we will concentrate mostly on time homogeneous cases where $\mu(x,t) = \mu(x)$ and $\sigma^2(x,t) = \sigma^2(x)$ are independent of t.

In the study of diffusion processes, hitting times play an important role. Informally, the hitting time T_z of the state z is a first time the process reaches a state z. Similarly we can define the hitting time of any subset of a state space S.

Definition (Hitting time). Let $\{X(t); 0 \leq t < \zeta\}$ be a stochastic process. The hitting time T_z of z for the process $\{X(t)\}$ is defined as

$$T_z = \begin{cases} \infty, & \text{if } X(t) \neq z \text{ for } 0 \leq t < \zeta \\ \inf\left\{t \geq 0 \mid X(t) = z\right\}, & \text{otherwise.} \end{cases} \tag{0.5}$$

Regular process A diffusion process is *regular* if starting from any point in the interior of I any other point in the interior of I may be reached with positive probability. A diffusion process $\{X(t); t \geq 0\}$ whose state space is an interval I

with endpoints $-\infty \leq l < r \leq \infty$ is regular if $P\left(T_z < \infty \mid X(0) = x\right) > 0$ for every $l < x, z < r$. In the following discussion, without further mention, we should discuss only regular diffusion processes.

Brownian motion Brownian motion is a regular diffusion process on the interval $\langle -\infty, \infty \rangle$ with $\mu(x) = 0$ and $\sigma^2(x) = \sigma^2$, where σ^2 is a constant.

Here we state, without proof, two theorems which will be used in the next chapter. For the proof, see [10].

Theorem 1 (Transformation formula for a diffusion process). Let $\{X(t); t \geq 0\}$ be a regular diffusion process whose state space is an interval I having endpoints l and r, and suppose $\{X(t)\}$ has infinitesimal parameters $\mu(x)$ and $\sigma^2(x)$. Let g be a strictly monotone function on I with continuous second derivative $g''(x)$ for $l < x < r$. Then $Y(t) = g(X(t))$ defines a regular diffusion process on the interval with endpoints $g(l)$ and $g(r)$, and $\{Y(t)\}$ has infinitesimal parameters

$$\mu_Y(y) = \frac{1}{2}\sigma^2(x)g''(x) + \mu(x)g'(x) \tag{0.6}$$

$$\sigma_Y^2(y) = \sigma^2(x)\left[g'(x)\right]^2 \tag{0.7}$$

where $y = g(x)$.

Theorem 2. Let $\{X(t); t \geq 0\}$ be a time homogeneous regular diffusion process with infinitesimal mean $\mu(x)$ and variance $\sigma^2(x)$ whose state space is an interval I having endpoints $l < r$. Let a and b, $l < a < b < r$, be fixed states with the corresponding hitting times T_a and T_b. We define $T^* = min(T_a, T_b)$ as the first time the process reaches either a or b. Then a solution to the problem

$$w(x) = E\left[\int_0^{T^*} g(X(s))ds \mid X(0) = x\right], \quad a < x < b, \tag{0.8}$$

where g is a bounded and continuous function, can be written in the form

$$w(x) = \int_a^b G(x, \psi)g(\psi)d\psi, \tag{0.9}$$

where $G(x, \psi)$ is Green function of the process on the interval $[a, b]$ and is given with

$$G(x,\psi) = \begin{cases} 2\dfrac{[S(x) - S(a)][S(b) - S(\psi)]}{S(b) - S(a)}\dfrac{1}{\sigma^2(\psi)s(\psi)}, & a \leq x \leq \psi \leq b \\ 2\dfrac{[S(b) - S(x)][S(\psi) - S(a)]}{S(b) - S(a)}\dfrac{1}{\sigma^2(\psi)s(\psi)}, & a \leq \psi \leq x \leq b, \end{cases} \tag{0.10}$$

where

$$s(x) = \exp\left\{-\int^x \left[2\mu(\psi)/\sigma^2(\psi)\right] d\psi\right\}, \quad \text{for } l < x < r, \tag{0.11}$$

$$S(x) = \int^x s(\eta)d\eta, \quad \text{for } l < x < r. \tag{0.12}$$

Problem. If we want to determine the mean time prior to T^* that the process spends in the interval $[\psi, \psi+\Delta\rangle$, we are solving the problem from Theorem 2 for the function

$$g(x) = \begin{cases} 1, & \psi \leq x < \psi + \Delta, \\ 0, & \text{otherwise.} \end{cases} \tag{0.13}$$

Then the solution can be written in the form

$$w(x) = \int_{\psi}^{\psi+\Delta} G(x, \eta) d\eta. \tag{0.14}$$

Even thought the function g defined in (0.13) is not continuous as required by Theorem 2, (0.14) can still be obtained by introducing a suitable continuous approximation function. When $\Delta \to d\psi$, we get from (0.14) that $G(x, \psi)d\psi$ measures the expected time prior to T^* that the process spends in the infinitesimal interval $[\psi, \psi + d\psi\rangle$ given $X(0) = x$.

0.3.2 Stochastic differential equation (SDE)

Let $\{B(t); t \geq 0\}$ be a standard Brownian motion. The "process" $dB(t)/dt = W(t)$ is called a Gaussian white noise "process". We should recall that the Brownian motion paths $B(t)$ are nowhere differentiable and thus the white noise is not a stochastic process in the usual mathematical sense but an abstraction, the "generalized stochastic process". Roughly speaking, $dB(t)/dt$ can be described as a one-parameter collection of independent Gaussian random variables with mean zero and infinite variance [11].

The heuristical derivation of a stochastic differential equation (SDE) usually involves the inclusion of the stochastic white noise in a deterministic differential equation [11]. Although the white noise process is a mathematical abstraction, it proved to be a good estimation for a variety of noises or other random processes that occur naturally in physical and biological contexts. Furthermore, the solution of SDE, as discuss later, is a diffusion process which makes it amenable to mathematical analysis.

A classic stochastic differential equation has the following form

$$dX(t) = f(X(t), t)dt + g(X(t), t)dB(t). \tag{0.15}$$

The SDE (0.15) is actually a shorthand for the equation

$$X(t) - X(t_0) = \int_{t_0}^{t} f(X(s), s)ds + \int_{t_0}^{t} g(X(t), t)dB(t). \tag{0.16}$$

The mathematical ambiguity arises from the random integral

$$\int_{t_0}^{t} g(X(t), t)dB(t) \tag{0.17}$$

which doesn't exist in the usual Riemann-Stieltjes sense [11]. Consequently, a vast number of definitions for (0.17) exist each of which results in a different solution to the SDE [11]. The two prominent definitions are those associated to *Itô* and *Stratonovich*. We will present only the Itô's solution to the SDE (0.15). If the Itô's interpretation of the random integral (0.17) is used then the solution $X(t)$ to the SDE (0.15) is a diffusion process with infinitesimal mean

$$\mu(x,t) = f(x,t) \tag{0.18}$$

and infinitesimal variance

$$\sigma^2(x,t) = g^2(x,t). \tag{0.19}$$

Population dynamics

0.4 Relative species abundance (RSA)

In an ecological study of a community, one usually observes different species, each represented by a certain number of individuals. Some of the observed species are rare with only a few (or only one) members while others are more common and represented by many organisms. The number of observed individuals belonging to a species is called its abundance. The relative abundance is obtained if we divide it with the total number of individuals found in the community. To study the properties of a community, we divide the species into abundance classes and study the distribution of relative species abundances (RSA). The RSA distribution is an important measure of biodiversity which helps us to understand and classify the communities.

When we plot histogram of a species abundance distribution with abundances on the x-axis and number of collected species on y-axis, we observe a "hollow curve". More precisely, the distribution is long-tailed and approximately follows a power law $f(n) \approx Cn^{-m}$, where n is an abundance class and C and m are constants. The "hollow" RSA distribution with many rare species and just a few common species seems to be universal with no recorded case of a community which deviates from it [3]. This universal law of ecology inspired a lot of research and RSA has been a central topic in ecology in the last century [3].

In the beginnings of the RSA theories, Fisher *et al.* [12] introduced the Log-Series distribution as a fit for a long-tailed RSA distribution. Soon after, Preston [13] started plotting the histogram of RSA distributions on logarithmic scale (historically $\log_2$ scale). This allowed for a better examination of rare species and introduced a diversity since the Preston plots of RSA distributions visually differ among different communities. In the last 70 years, many new theoretical RSA models were developed which resulted in new RSA distributions with some of the most prominent ones being Log-Series, Negative Binomial and Poisson Log-Normal distribution. In their review, McGill *et al.* [3] counted more than 40 models which can be divided in five categories: *purely statistical models*, *branching processes*, *population dynamics models*, *niche partitioning models* and *spatial distribution models*. Interestingly, many of these models overlap and the same distribution can result from different models with

non-aligned assumption.

0.5 Population dynamics model for relative species abundance distribution

Using a stochastic model of population dynamics, Engen and Lande developed a general class of abundance models where species abundances are generated by an inhomogeneous Poisson process with rate $\lambda(x)$. In the first paper [4] they introduces a general model and showed that under certain conditions, the rate $\lambda(x)$ follows a Log-Normal distribution. In the second paper [5], under slightly different assumptions, they proved that the rate $\lambda(x)$ follows a Gamma distribution. As we will discuss in more details in the next sections, this means that under the same framework varying the assumptions of the model we may arrive to both the Poisson Log-Normal and the Negative Binomial RSA distribution. Moreover, since Log-Series distribution can be viewed as a special case of Negative Binomial distribution when dispersion parameter converges to zero, the three important RSA distributions can be derived from the Engen and Lande's abundance models.

0.5.1 Engen and Lande's general abundance model

If we assume that species enter the community at the times generated by inhomogeneous Poisson process with rate $\omega(t)$ and evolve independently of others, then their abundances are generated by an inhomogeneous Poisson process with rate

$$\lambda(x) = \int_0^\infty \omega(-t)p(t)f(x;t)dt, \tag{0.20}$$

where $p(t)$ is a probability that a species hasn't gone extinct and $f(x;t)$ is a distribution of its density.

Now we let the process for each species be a diffusion process which is a solution of a stochastic growth equation

$$\frac{dx}{dt} = rx - xg(x) + x\sigma_r(x)\frac{dB(t)}{dt}, \tag{0.21}$$

where $\sigma_r^2(x) = \sigma_e^2 + \sigma_d^2/x$ has two components: environmental and demographic stochasticity. Environmental stochasticity is due to environmental changes which act simultaneously on all individuals in population and demographic stochasticity reflects the differences among the individuals inside the population. Then the resulting diffusion process has the infinitesimal mean $m(x)$ and infinitesimal variance

$v(x)$ given by

$$m(x) = \left[r + \frac{1}{2}\frac{\sigma_d^2}{x} + \frac{1}{2}\sigma_e^2 \right] x - xg(x) \tag{0.22}$$

$$v(x) = \sigma_d^2 x + \sigma_e^2 x^2. \tag{0.23}$$

If we additionally assume that the speciation rate $\omega(t)$ is a constant, then the rate $\lambda(x)$ is given by

$$\lambda(x) = 2\omega \frac{1}{v(x)} \exp\left[\int_1^x \frac{2m(u)}{v(u)} du \right]. \tag{0.24}$$

We will prove all of the statements written above in three steps. The following derivations can be found in Engen and Lande's paper[4].

Proof (Species abundances are generated by inhomogenous Poisson process with rate $\lambda(x)$) We assume that species enter the community at the times generated by an inhomogeneous Poisson process with rate $\omega(t)$ and once they enter the community they evolve independently of each other. A species which entered the community at time t is still present in the community at time $t+s, s>0$ with probability $p(s)$ and, in case it hasn't gone extinct, its abundance comes from a distribution with density $f(x;s)$.

Let's observe two disjoint intervals Ω_1 and Ω_2 on the positive real axis. We now fix the time point t_0 and define random variables $Y_1(t)$ and $Y_2(t)$ as a number of species with abundances in intervals Ω_1 and Ω_2, respectively, at time t_0 which entered the community in the time interval $\langle t, t+\delta t\rangle, t < t_0$. The probability that no species enter the community in this interval is $1-\omega(t)\delta t + o(\delta t)$ and the probability that exactly one species enters the interval is $\omega(t)\delta t + o(\delta t)$. This leads to three different possibilities for the distribution of $(Y_1(t), Y_2(t))$

$$P\left(Y_1(t)=1, Y_2(t)=0\right) = \omega(t)p(t_0-t)\delta t \int_{\Omega_1} f(x; t_0-t)dx + o(\delta t), \tag{0.25}$$

$$P\left(Y_1(t)=0, Y_2(t)=1\right) = \omega(t)p(t_0-t)\delta t \int_{\Omega_2} f(x; t_0-t)dx + o(\delta t), \tag{0.26}$$

$$P\left(Y_1(t)=0, Y_2(t)=0\right) = 1-\omega(t)p(t_0-t)\delta t \int_{\Omega_1\cup\Omega_2} f(x; t_0-t)dx + o(\delta t). \tag{0.27}$$

The equation (0.25) comes from the fact that the $Y_1(t)=1$ and $Y_2(t)=0$ if exactly one species enters the community in the interval $\langle t, t+\delta t\rangle$, doesn't go extinct by time t_0 and its abundance falls in the interval Ω_1 in time t_0. All other possibilities of $(Y_1(t), Y_2(t))$ happen with probability $o(\delta t)$ and are thus neglected.

We introduce notation $I_j(t)$ for the integrals $\int_{\Omega_j} f(x; t_0 - t)dx$, $j = 1, 2$, and notice that $I_1 + I_2 = \int_{\Omega_1 \cup \Omega_2} f(x; t_0 - t)dx$. Now we calculate the joint moment generating function for $Y_1(t)$ and $Y_2(t)$

$$E\left(e^{uY_1(t)+vY_2(t)}\right) = 1+(e^u-1)\omega(t)p(t_0-t)I_1(t)\delta t+(e^v-1)\omega(t)p(t_0-t)I_2(t)\delta t+o(\delta t). \tag{0.28}$$

Taking the logarithm and using Taylor series approximation $\ln(1 + x) \approx x$, we get the corresponding cumulant generating function

$$K_t(u, v) = \left[(e^u - 1)I_1(t) + (e^v - 1)I_2(t)\right] p(t_0 - t)\omega(t)\delta t + o(\delta t). \tag{0.29}$$

To find the total cumulant generating function for the number of species with abundances in Ω_1 and Ω_2, we split the interval $\langle-\infty, t_0\rangle$ into intervals of length δt. Because of the species independance, the partial cumulant generating functions of type (0.29) will be independent and summing them we get

$$K(u, v) = (e^u - 1)\sum I_1(t)\omega(t)p(t_0 - t)\delta t + (e^v - 1)\sum I_2(t)\omega(t)p(t_0 - t)\delta t + o(\delta t), \tag{0.30}$$

where sum is taken over the partition of $\langle-\infty, t_0\rangle$ into intervals with length δt. If we now take the limit when $\delta t \to 0$, we obtain the integral

$$K(u, v) = (e^u - 1)\varphi_1 + (e^v - 1)\varphi_2, \tag{0.31}$$

where

$$\varphi_j = \int_{-\infty}^{t_0} I_j(t)\omega(t)p(t_0 - t)dt, \quad j = 1, 2, \tag{0.32}$$

From the form of the cumulant generating function (0.31), it follows that the number of species with abundances in Ω_1 and Ω_2 are independent Poisson random variables with parameters φ_1 and φ_2, respectively.

If, in particular, we choose $\Omega_1 = [x_0, x]$, we get

$$\varphi_1(x) = \int_{-\infty}^{t_0}\int_{x_0}^{x} f(y; t_0 - t)\omega(t)p(t_0 - t)dydt. \tag{0.33}$$

Since integrand function is non-negative, by Tonelli's theorem we are allowed to change the order of integration. The only requirement is that the integrand product is measurable with respect to Lebesgue measure, however this is much less strict assumption than continuity since practically all real functions which can be described are measurable. Upon change of the order of integration, taking the derivative with respect to x gives

$$\lambda(x) = \frac{d\varphi_1(x)}{dx} = \int_{-\infty}^{t_0} f(x; t_0 - t)\omega(t)p(t_0 - t)dt, \tag{0.34}$$

which after simple substitution can be written in the form

$$\lambda(x) = \int_0^\infty \omega(t_0 - t)p(t)f(x;t)dt. \tag{0.35}$$

Now it follows that the abundances are generated by an inhomogeneous Poisson process with rate $\lambda(x)$ given by (0.35). If we observe abundances in present taking $t_0 = 0$, we get exactly the equation (0.20).

□

Proof (Solution of the SDE)) To solve the SDE (0.21), we first write it in a the following concise form

$$\frac{dx}{dt} = r'x - xg(x), \tag{0.36}$$

where $r' = r + \sigma_r dB(t)/dt$ and $\sigma_r^2 = \sigma_e^2 + \sigma_d^2/x$. The function $g(x)$ is a density regulation term and will be discussed later.

Now we introduce the substitution $y = \ln x$. As a result of the substitution, we have $dy = dx/x$ and $x = e^y$ which together with (0.36) gives a SDE for y

$$\frac{dy}{dt} = [r - g(e^y)] + \sigma_r(e^y)\frac{dB(t)}{dt}. \tag{0.37}$$

Applying Itô approach, the process y is a diffusion process with infinitesimal mean $r - g(e^y)$ and infinitesimal variance $\sigma_r^2(e^y)$. Using the transformation formula for diffusion processes given by Theorem 1 for $x = e^y$, we get that x is a diffusion process with infinitesimal mean and variance given by (0.22) and (0.23).

□

Proof (The rate $\lambda(x)$ of Poisson process is given by equation (0.24)) Under the assumption that the speciation rate $\omega(t) = \omega_0$ is a constant the equation (0.20) becomes

$$\lambda(x) = \omega_0 \int_0^\infty p(t)f(x;t)dt. \tag{0.38}$$

Let's now assume that the process for each species is a diffusion process found as a solution to the SDE (0.21) which has the infinitesimal mean $m(x)$ and variance $v(x)$ given by (0.22) and (0.23), respectively. Then the integral in the equation (0.38) multiplied by dx represents the expected time the process is in $[x, x + dx\rangle$ and is called the Green function for the process. If we assume the new species enter the community at the abundance x_0 the Green function is $G(x_0, x) = \int_0^\infty p(t)f(x;t)$ and we have

$$\lambda(x) = \omega_0 G(x_0, x). \tag{0.39}$$

Let a and b, $a < b$ be the absorbing barriers for the process. Then the Theorem 2 gives the Green function

$$G(x_0, x) = 2\frac{[S(x_0) - S(a)][S(b) - S(x)]}{S(b) - S(a)}\frac{1}{v(x)s(x)}, \tag{0.40}$$

where

$$s(x) = \exp\left[-\int_a^x \frac{2m(u)}{v(u)}du\right] \tag{0.41}$$

and

$$S(x) = \int_a^x s(u)du. \tag{0.42}$$

For models with density regulation, that is, models where the increasing function $g(x)$ is included in the SDE (0.21), $m(x)$ becomes negative for large values of x, as can easily be concluded from the (0.22). Consequently, $lim_{x\to\infty}S(x) = \infty$. Since there is no upper theoretical limit for the abundance of a species, the upper barrier $b = \infty$. Inserting $b = \infty$ in the expression for $G(x_0, x)$, we get

$$G(x_0, x) = 2\frac{S(x_0) - S(a)}{v(x)s(x)}. \tag{0.43}$$

Let us define extinction to occur at $x = 1$, that is, we choose the lower boundary $a = 1$. Then we have $G(1, x) = 0$, meaning a species which enters the community at abundance 1 immediately goes extinct with probability 1. For this reason, we assume that species enter the community at the abundances $x_0 = 1 + \delta x$. Since $S(1 + \delta x) - S(1) = \int_1^{1+\delta x} s(u)du = s(1)\delta x + o(\delta x)$ as $\delta x \to 0$, we have

$$G(1 + \delta x, x) = 2\frac{s(1)\delta x}{v(x)s(x)} + o(\delta x), \tag{0.44}$$

as $\delta x \to 0$.

Combining (0.44) with (0.38), we get

$$\lambda(x) = 2(\omega_0\delta x)\frac{s(1)}{v(x)s(x)} + o(\delta x). \tag{0.45}$$

If we now let $\omega_0 \to \infty$ and $\delta x \to 0$ so that $\omega_0\delta x \to \omega > 0$ and using $s(1)/s(x) = \exp\{\int_1^x 2m(u)/v(u)du\}$, the abundance model is given exactly by equation (0.24).

□

0.5.2 Log-Normal species abundance distribution

To obtain the Poisson process rate $\lambda(x)$, we need to solve the integral in (0.24). If we introduce a new variable $\epsilon = \sigma_d^2/\sigma_e^2$ as a ratio between the demographic and environmental variance, we get

$$\begin{aligned} 2\int \frac{m(u)}{v(u)}du &= 2\int\left[\frac{r + \frac{1}{2}\frac{\sigma_d^2+\sigma_e^2 u}{u} - g(u)}{\sigma_d^2 + \sigma_e^2 u}du\right] \\ &= 2\frac{r}{\sigma_e^2}\int\frac{du}{u+\epsilon} + \int\frac{du}{u} - 2\frac{1}{\sigma_e^2}\int\frac{g(u)}{u+\epsilon}du \\ &= 2\frac{r}{\sigma_e^2}\ln(u+\epsilon) + \ln u - 2\frac{1}{\sigma_e^2}\int\frac{g(u)}{u+\epsilon}du \end{aligned} \tag{0.46}$$

Let us now assume that the density regulation is given by Gompertz curve. This means that the function $g(x)$ is a logarithmic type of function. Keeping in mind form of the integral (0.46), we choose $g(x) = \gamma \ln(x+\epsilon)$, with γ being a constant. Now we have $\int g(u)du/(u+\epsilon) = (\gamma/2)\ln^2(u+\epsilon) + const.$ Together with (0.46), we get

$$2\int_1^x \frac{m(u)}{v(u)}du = \ln(x) + \frac{2r}{\sigma_e^2}\left[\ln(x+\epsilon) - \ln(1+\epsilon)\right] - \frac{\gamma}{\sigma_e^2}\left[\ln^2(x+\epsilon) - \ln^2(1+\epsilon)\right]. \tag{0.47}$$

Inserting (0.47) in (0.24), we get

$$\lambda(x) = \frac{\alpha\omega}{x+\epsilon}\exp\left[-\frac{1}{2}\frac{[\ln(x+\epsilon) - r/\gamma]^2}{\sigma_e^2/2\gamma}\right], \tag{0.48}$$

where

$$\alpha = \frac{2}{\sigma_e^2}\exp\left\{\frac{\gamma}{\sigma_e^2}\left[\ln(1+\epsilon) - \frac{r}{\gamma}\right]^2\right\}. \tag{0.49}$$

From the form of (0.48) we recognize that this is a Log-Normal abundance model with a translation $-\epsilon$, mean r/γ and variance $\sigma_e^2/2\gamma$. Unless a demographic variance σ_d^2 is significantly larger that the environmental variance σ_e^2, $x = \epsilon$ represents a very small abundance and for $x \gg \epsilon$ the translation may be ignored in which case we have the standard Log-Normal abundance model with

$$\lambda(x) \sim LogNormal\left(\mu = \frac{r}{\gamma}, \sigma^2 = \frac{\sigma_e^2}{2\gamma}\right). \tag{0.50}$$

Remark. Endgen and Lande [4] showed that certain generalizations of the model, such as introducing the heterogeneity of the species or normally distributed growth rate r, still generate a Log-Normal species abundance model.

0.5.3 Gamma type species abundance distribution

Here, as proposed in [5], we derive the resulting Poisson process rate $\lambda(x)$ for a different density regulation function $g(x) = \eta x$, where η is a constant. Now the integral in (0.46) equals

$$\begin{aligned}\int \frac{g(u)}{u+\epsilon}du &= \eta\int \frac{u}{u+\epsilon}du \\ &= \eta\left[u - \epsilon\ln(u+\epsilon)\right] + const.\end{aligned} \tag{0.51}$$

Together with (0.46), we have

$$\begin{aligned}2\int_1^x \frac{m(u)}{v(u)}du &= ln(x) + \frac{2r}{\sigma_e^2}\left[\ln(x+\epsilon) - \ln(1+\epsilon)\right] - \frac{2\eta}{\sigma_e^2}\left[(x - \epsilon\ln(x+\epsilon)) - (1 - \epsilon\ln(1+\epsilon))\right] \\ &= \ln(x) - \frac{2\eta}{\sigma_e^2}(x-1) + 2\frac{r+\eta\epsilon}{\sigma_e^2}\ln\left(\frac{x+\epsilon}{1+\epsilon}\right)\end{aligned} \tag{0.52}$$

which gives

$$\lambda(x) = a\omega(x+\epsilon)^{\frac{2(r+\eta\epsilon)}{\sigma_e^2}-1} e^{-\frac{2\eta}{\sigma_e^2}(x+\epsilon)}, \tag{0.53}$$

where

$$a = \frac{2e^{2(1+\epsilon)\eta/\sigma_e^2}}{\sigma_e^2(1+\epsilon)^{2(r+\eta\epsilon)/\sigma_e^2}}. \tag{0.54}$$

This distribution is a Gamma distribution with a translation $-\epsilon$, shape $2(r+\eta\epsilon)/\sigma_e^2$ and rate $2\eta/\sigma_e^2$. Similarly as in the Log-Normal model, unless demographic variance is much larger than the environmental variance, $x = \epsilon$ represents very small abundances so we can assume that $x \gg \epsilon$. In this case, we arrive at the familiar Gamma abundance model

$$\lambda(x) \sim Gamma\left(\alpha = \frac{2(r+\eta\epsilon)}{\sigma_e^2}, \beta = \frac{2\eta}{\sigma_e^2}\right). \tag{0.55}$$

0.5.4 Adding variability of Poisson sampling

Under Engen and Lande's Poisson abundance model, the relative species abundance is distributed as $\lambda(x)$, where $\lambda(x)$ can be a Log-Normal or Gamma distribution, depending on a chosen density regulation function. To be more precise, the abundances of species are generated by an inhomoheneous Poisson process with rate $\lambda(x)$. Let call this process $\{N(x); x \geq 0\}$. Then probability that a new abundance entering the "community" of abundances is in the interval $[x, x+dx\rangle$ is

$$P(N(x+dx) - N(x) = 1) = \lambda(x)dx + o(dx), \tag{0.56}$$

which means that the RSA distribution has density $\lambda(x)$.

However, since the RSA distribution is in fact a discrete distribution we consider $\lambda(x)$ to be an idealized mean RSA distribution. The empirical RSA distribution is constructed from a random sample from the population where probability that a random species will be represented with r individuals is a compound Poisson distribution with mean $\lambda(x)$

$$P_r = \int_0^\infty \frac{x^r e^{-x}}{r!}\lambda(x)dx. \tag{0.57}$$

Speaking roughly, we are summing over the theoretical RSA distribution the probabilities that the species with theoretical RSA $\lambda(x)$ is represented by r individuals. The procedure of Poisson sampling was introduced in the beginnings of the RSA theories [12]. Poisson sampling can be thought of as a process of discretization, however it has additional advantage of adding the variability of random sampling into the model [14].

When we add the variability of Poisson sampling to Engen and Lande's model, then the Log-Normal abundance model results in Poisson Log-Normal RSA distribution with scale μ and location σ^2

$$RSA \sim PoiLN\left(\mu = \frac{r}{\gamma}, \sigma^2 = \frac{\sigma_e^2}{2\gamma}\right). \tag{0.58}$$

On the other hand, the Gamma abundance model gives a compound Poisson Gamma RSA distribution, better known as Negative Binomial distribution with dispersion $\alpha > 0$ and success rate $p \in [0, 1]$

$$RSA \sim NB\left(\alpha = \frac{2(r + \eta\epsilon)}{\sigma_e^2}, p = \frac{\sigma_e^2}{2\eta + \sigma_e^2}\right). \tag{0.59}$$

0.5.5 Log-Series as a limit case of Negative Binomial

To fit the observed long-tailed RSA distributions, in 1943 Fisher [12] introduced the Log-Series (also known as Logarithmic) distribution. In his derivation, Fisher used the mixture of Poisson and Gamma distributions and obtained a Negative Binomial distribution

$$f(k; \alpha, p) = \frac{(k + \alpha - 1)!}{k!(\alpha - 1)!}(1 - p)^\alpha p^k \tag{0.60}$$

which he then observed in the limit when dispersion parameter $\alpha \to 0$ and obtained

$$f(k; \alpha \to 0, p) = \frac{p^k}{k}. \tag{0.61}$$

Adding a normalization constant the new Log-Series distribution has a density

$$f(k; p) = \frac{-1}{\ln(1 - p)}\frac{p^k}{k}. \tag{0.62}$$

In his purely statistical derivations, Fisher didn't use any theoretical model to justify the use of Gamma distribution. Nonetheless, the one-parameter Log-Series distribution fitted data well and even though it is usually outperformed by the newer RSA models [3], it is still a valuable example of a RSA distribution.

Part I

Evolutionary model of protein domains: towards a better resolution in bacterial taxonomy

Chapter 1

Introduction

1.1 Protein domains

Proteins are biological macromolecules essential to life. They are the cell's building blocks and executors of the majority of the cell's functions. From a structural point of view, a protein molecule is made from a long chain of amino acids. There are 20 different amino acids which are coded for directly in an organism's DNA molecule. The gene gives instructions on how to create a protein through the processes of transcription and translation. During the transcription the DNA molecule is copied into RNA which then in the process of translation serves as a template for the synthesis of the protein. With many details being left out, these are the fundamentals of the central dogma of molecular biology [15].

As a result of various chemical interactions, the synthesised amino acid sequence folds into the final three-dimensional conformation of the protein. Interestingly, experiments indicate that the amino acid sequence contains all of the information needed for specifying the three-dimensional shape of a protein. The resulting three-dimensional structure of the protein, its conformation, is crucial for the protein's chemistry. Namely, to perform its tasks a protein has to bind to other molecules and the binding sites of the protein are affected by its conformation [15].

Most proteins are between 50 and 2 000 amino acids long. These large molecules are often composed of smaller modular units called protein domains which are between 40 and 350 amino acids long. By definition, protein domains are components of protein that can fold independently into a compact and stable structure. The different domains of a protein are often associated with different functions thus making protein domains the functional units of the protein [15, 16]. It is estimated that two-thirds of proteins consist of two or more domains in prokaryotes and an even larger fraction in eukaryotes [16]. On the other end, the same domain can be a part of multiple proteins. A subset of domains, called protein modules, are especially dispersed and are found in many different proteins. The omnipresence of protein modules is a consequence of their special three-dimensional structure which

facilitated their mobility during the evolution and spread them across the genome [15].

During the course of evolution, the genes and genomes have evolved to create the great diversity of life forms on our planet. Inspecting the modifications of the genes and consequently the encoded proteins, it has been shown that protein three-dimensional conformation is much more conserved than its sequence of amino acids [17]. Thus, instead of genes, we look at protein domains as the tools of evolution which helped to create diverse assembly of proteins from likely an initially relatively limited set of domains [17]. Domain shuffling/recombination, gene sequence duplication and divergence are the main mechanisms that allow the outbreak of proteins with new functionalities, thereby contributing to the emergence of complexity [18]. In the process of evolution, different domains have been duplicated to different extent. Specifically, the abundances of protein domains families follow the power low with a few highly abundant domains and many low abundant domains [16]. This inspires us to investigate the distribution of protein domain abundance within the population dynamics paradigm. The investigation of the dynamic processes that led to the current configuration of protein domains can highlight the important aspects of the proteome evolution and consequently of the evolution of living organisms.

1.2 Evolutionary model of protein domains

In this work, we investigate the evolution of protein domains using the approach of ecological theory. To correlate the world of protein domains with the usual subjects of ecological studies, we redefine the meaning of *species* and *community*. The communities which we are going to study are proteomes inside which we observe the protein domains. Thus protein domains have the meaning of species and their frequencies are so-called species abundances. Using this approach, we obtain the relative species abundance (RSA) distribution of protein domains.

To model the processes leading to RSA distribution of protein domains, we use Engen and Lande's population dynamics model [4, 5] which is described in an introductory Section 0.5. It is assumed that a process for every species is a diffusion process which solves the stochastic growth equation

$$\frac{dx}{dt} = rx - xg(x) + x\sigma_r(x)\frac{dB(t)}{dt}, \tag{1.1}$$

where r is a constant growth rate, $g(x)$ is a density regulation function and $\sigma_r^2(x) = \sigma_e^2 + \sigma_d^2/x$ is a stochastic variance consisting of environmental and demographic stochasticity. In the model of protein domains, the growth rate can be interpreted as rate by which a protein domain species gets a new copy inside the genome. Demographic variance captures small differences between different individual domains which belong to the same protein domain species, while environmental variance presents larger scale disruptions of the proteome which act on all species of pro-

tein domains simultaneously. Events such as horizontal gene transfer are an example of environmental stochasticity. The role of density regulation $g(x)$ is to keep species abundance from "exploding". This is usually a consequence of limited resources in the community, but when talking about proteome it can be interpreted as the limit in total genome, and consequently proteome, size [19]. We will consider two different density regulation functions. The first one is a Gompertzian function $g(x) = \gamma \ln(x + \epsilon)$, where $\epsilon = \sigma_d^2/\sigma_e^2$, which leads to the Poisson Log-Normal RSA distribution

$$RSA \sim PoiLN\left(\mu = \frac{r}{\gamma}, \sigma^2 = \frac{\sigma_e^2}{2\gamma}\right). \tag{1.2}$$

The other type of density regulation with the function $g(x) = \eta x$ results in Negative Binomial RSA distribution

$$RSA \sim NB\left(\alpha = \frac{2(r + \eta\epsilon)}{\sigma_e^2}, p = \frac{\sigma_e^2}{2\eta + \sigma_e^2}\right). \tag{1.3}$$

We will also consider the Log-Series RSA distribution as a special case of Negative Binomial distribution when the dispersion parameter $\alpha \to 0$.

In our data set, we have protein domain counts for different bacteria. Separately for every bacterium, we will fit RSA of protein domains with different distributions, specifically with Log-Series, Negative Binomial and Poisson Log-Normal distribution. Choosing the model which outperforms the others in most of bacteria, we infer the theoretical evolutionary model which best describes our data. Moreover, fitting the RSA protein domain distribution, we will get estimates of distribution parameters for all bacteria. We will then use these parameters to classify bacteria and to introduce our RSA phylogenetic method.

1.3 Bacterial phylogeny

To evaluate our approach of bacterial phylogeny, we will compare the results with bacterial taxonomy and 16S rRNA gene-based bacterial phylogeny. In the taxonomic *tree of life*, Bacteria is one of the three domains together with Archaea and Eukarya. Some of the lower taxonomic ranks are phylum, class, order, family, genus and species. While taxonomic classification of bacteria has well-organised hierarchical structure, it lacks the details of phylogenetic approach. This is especially important on the extremely diverse intraspecies level of bacteria. Twenty years ago Lan and Reeves [20] wrote about extensive intraspecies variation in bacteria which is manifested both in genomic and phenotypic differences. As more bacteria have been sequenced, this became even more apparent. An important aspect of this diversity are different pathogenic properties of strains belonging to the same bacterial species. Namely, different strains can infect different hosts [21–23], or even more extremely, some bacterial species can have pathogenic as well as non-pathogenic isolates [20,

24]. In the following paragraph, we briefly summarize other methods commonly used for the phylogenetic reconstruction at the subspecies level.

16S rRNA gene sequence is widely used in bacterial classification since mutations of hypervariable region of the 16S rRNA gene serve as a good estimate of evolutionary time. However, bacterial classification based on 16S rRNA gene has low phylogenetic power at the subspecies level where functional diversification of strains is faster than random mutations of 16S rRNA gene [25]. While there are many other methods for inferring bacterial phylogenies, some of which are specifically developed for lower taxonomic levels, there is not one method which is considered to be the gold standard. In his textbook about phylogenetic inference [26], published in 2004, Felsenstein estimated there are about 3 000 papers on methods for inferring phylogenies. The most widely used methods can be classified in the following manner. One class of methods is based on distance matrix. In these type of methods, the estimates of distances among bacterial molecular sequences are used for phylogenetic reconstruction. Other prominent methods use maximum parsimony, maximum likelihood or Bayesian inference in phylogenetic tree reconstruction. In the background of these methods is the substituton models which describes the evolution of molecular sequences through the probabilistic modeling of random mutations. But beside the phylogenetic metodology, the appropriate choice of the data to which it is applied is of great relevance. Different molecular sequences can be used for the inference. However, the chosen sequences have to satisfy two conditions for the succesful phylogenetic reconstruction: orthologs of sequences have to be shared by tested bacteria and have to contain the vertical phylogenetic signal. Approaches of using a sequence of one gene, such as 16S rRNA, may work at higher taxonomic levels, but are usually too generic to differentiate species or strains. The popular method which overcomes this problem while still controlling the phylogenic noise is multilocus sequence analysis (MLSA) [27, 28], the method which simultaneously uses sequences of seven housekeeping genes. However, the choice of genes used by this or similar methods is of great importance since phylogenetic signal is not equally distributed across the genome and some classes of genes perform better at phylogenic inference [29]. From this short summary of methods, it is clear that phylogenic inference in bacteria is a challenging problem, especially at the lower taxonomic levels where experts should be included at various steps of analysis.

Chapter 2

Materials and methods

2.1 Data retrieval

This part of work was performed by our collaborators, Dr. Edoardo Saccenti and Dr. Maria Suarez Diez from the Laboratory of Systems and Synthetic Biology, Wageningen University Research (the Netherlands). The genome sequences of 3 370 bacteria were downloaded from the NCBI database [30]. Draft genome sequences were discarded and only the higher quality fully circular genome sequences were retained. GeneBank files containing genome sequences and existing annotations were retrieved from the NCBI database and imported into the Semantic Annotation Platform for Prokaryotes [31] using the EMBL/GBK to RDF SAPP module. *De novo* identification of genetic elements (gene calling) was performed using Prodigal (2.6) [32] with codon table 11. Dedicated SPARQL queries were built to extract proteins and their sequences from the RDF triplestore used by SAPP to store the intermediate results. InterProScan [33] was used to identify protein domains in the corresponding sequences. Due to the high number of distinct protein sequences to be analyzed, the SURFsara GRID was used (Grid reference) to concurrently analyze the sequences. Dedicated SPARQL queries were used to retrieve the identified domains and assign them to the originating protein and bacterial genome. Finally, the matrix generating module from SAPP was used to generate a matrix containing the number of instances of the detected domain (domain abundance) for each of the studied genomes and for each of the identified protein domains. Overall 3 370 bacterial genomes were analyzed and 13 934 distinct domains were identified.

2.2 Model fitting and selection

Protein domains RSAs were fitted with the Maximum Likelihood Estimation method implemented in R "sads" package v.0.4.2 [34]. Data were modeled with a truncated Poisson Log-Normal, a truncated Log-Series and a truncated Negative Binomial

distribution, so that to test and compare different ecological hypothesis. In all cases, truncation was performed to exclude the zero abundance class, that is not observable in empirical data. Two bacterial genomes which couldn't be fitted due to extreme RSA distributions are GCA_000200735 and GCA_000831405. After removal of these bacteria, we proceeded with the analysis of the remaining 3 368 bacterial genomes. Akaike Information Criterion (AIC) [35] and R-squared were computed to assess the models performances and for model selection.

2.3 Calculation of RSA model-based distance matrix

Fitting truncated Poisson Log-Normal distribution to protein domain RSA, we obtained estimates μ and σ for 3 368 bacteria. In addition to Poisson Log-Normal parameters, we calculated protein domain density for every bacteria as ratio between total number of protein domains present in the genome and the genome length. RSA distance between each pair of bacteria was calculated as 3D euclidean distance in the scaled space of μ, σ and protein domain density. Scaling was performed independently at each dimension subtracting the mean and dividing by standard deviation.

2.4 Calculation of 16S rRNA gene-based distance matrix

In order to calculate phylogenetic distanced based on 16S rRNA gene, we used the silva database [36] to retrieve the 16S rRNA reference sequences of the bacterial species for which protein domain data were available. For 48 bacteria, the 16S rRNA sequence was not present, so we considered only the remaining 3 320 bacteria for the following analysis. Since the same bacterial genome can have multiple different copies of 16S rRNA gene, as 16S rRNA distance between a pair of bacteria we used the mean pairwise distance between all pairs of 16S rRNA sequences within two genomes [37]. The alignment of 16S rRNA sequences and calculation of distances was performed with mothur [38] "pairwise.seqs" function using default options.

2.5 Comparison of different clustering solutions

For each taxonomic level, we considered only bacteria for which the classification was known and which belonged to a taxonomic group with at least 10 members. As a result, we analysed 3 270, 3 148, 3 032, 2 714, 2 139 and 161 bacteria at phylum, class, order, family, genus and species level, respectively. They belonged to 14, 20,

48, 63, 54 and 48 different taxonomic groups. Hierarchical clustering was performed on both RSA and 16S rRNA distance matrices. For RSA clustering we used Ward's minimum variance method while for 16S rRNA we used average linkage method. The Ward's method was used to minimize the total in-cluster variance [39]. Since this method is based on Euclidean distance, it couldn't be applied to 16S rRNA distance matrix. To get clustering solutions for RSA and 16S rRNA, we cut the hierarchical tree fixing the number of clusters to the number of taxa at the selected taxonomic level. Finally, we used the NMI score as a measure of clustering agreement between the three approaches: RSA, 16S and predefined taxonomy. To calculate the baseline for the NMI score, we compared the taxonomy with clustering solutions based on simulations. For the fixed number of clusters and data points, each simulation assigned random cluster to each data point. Purity of a RSA cluster with respect to taxonomy is calculated as the ratio between the size of the cluster's most abundant taxonomic group and the cluster size. The total purity of the RSA clustering solution is a weighted average of its clusters' purities where weight is a cluster size.

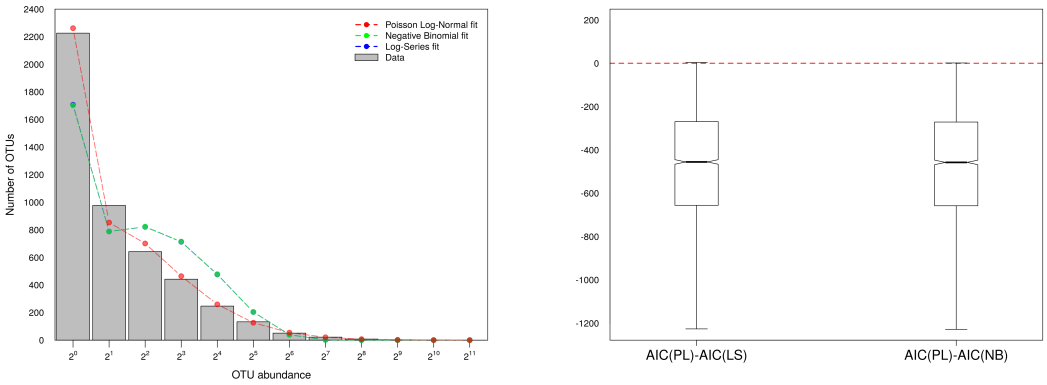
Chapter 3

Results and discussion

3.1 Protein domains RSA follows a Poisson Log-Normal distribution

The protein domains RSA distributions of 3 368 bacterial genomes were obtained as detailed in the Materials and methods section. Three evolutionary hypotheses were tested by fitting the RSA distributions with the Log-Series, the Negative Binomial and the Poisson Log-Normal distribution (Figure 3.1-a). According to the Akaike Information Criterion (AIC) [35], in 99.97% of bacteria the selected model was the Poisson Log-Normal (Figure 3.1-b). This model performed better than both the Log-Series and the Negative Binomial and described the data well, with an average R^2 of 0.97 and a minimum R^2 of 0.86. The results imply that the Gompertzian density regulation $g(x) = \gamma \ln(x + \epsilon)$ is better choice for protein domains RSA. The Gompertzian function is a weaker density regulator than the alternative $g(x) = \eta x$ which leads to the Negative Binomial distribution. This suggests that the over-abundant protein domains are not strictly regulated as they would be in case of $g(x) = \eta x$. As a result, we expect to see a heavier long tail of the RSA distribution. If we calculate the mean abundance of protein domains we notice that the majority of protein domains have small abundances. Namely, the maximum mean abundance taken across all bacteria is 10.34. On the other side, when we calculate the maximum abundance of protein domains we observe some extremely abundant protein domains which results in heavily long-tailed RSA. Precisely, the mean of maximum abundance across all bacteria is 536.6.

From the Figure 3.1-a, we may notice that Negative Binomial fit and Log-Series fit overlap. Since Log-Series distribution is a limiting case of Negative Binomial when the dispersion parameter tends to zero, this observation implies that the dispersion parameter of the Negative Binomial distribution is close to zero. The estimated dispersion parameter has mean 2.67×10^{-4} and median 2.62×10^{-7} which is in agreement with the observed overlap.



(a) Preston plot. **(b)** AIC differences.

Figure 3.1: **(a)** Example of protein domains Preston plot fitted with three different distributions: the Poisson Log-Normal, the Negative Binomial and the Log-Series. Results refer to the bacterial genome GCA_000717515. **(b)** Distribution of the difference between the AIC obtained with the Poisson Log-Normal model (PL) and the Log-Series (LS) or the Negative Binomial (NB) model, considering all the 3 368 bacterial genomes.

3.2 Comparing RSA and taxonomy

The Poisson Log-Normal distribution is characterized by two parameters: a scale parameter μ and a location parameter σ^2. By fitting Poisson Log-Normal distribution to protein domain relative species abundances, we obtain estimates of μ and σ^2 for each bacterium in our data set. The scatter plot of these parameters (Figure 3.2) shows two properties of our estimates, the negative value of parameter μ and the negative relationship between σ^2 and μ. From the population dynamics model which led to Poisson Log-Normal RSA distribution, we recall the form of parameters $\mu = r/\gamma$ and $\sigma^2 = \sigma_e^2/2\gamma$, where r is a growth rate, σ_e^2 is environmental noise and γ is a positive multiplicative constant from the Gompertzian function.

Negative value of μ implies negative value of the growth rate r . For the fitted Negative Binomial distribution the dispersion parameter converged to zero. The dispersion parameter $\alpha = 2(r + \eta\epsilon)/\sigma_e^2$, where $\eta\epsilon > 0$, converges to zero only in case of very small or negative growth rate r confirming our observation about negative r. The growth rate r can be expressed as difference between a birth rate and a death rate, $r = b - d$. This means that the death rate is greater than the birth rate. In the evolutionary model of protein domains, birth rate b has the meaning of duplication rate of a certain protein domain species while death rate d is a rate at which these protein domains die. Since death, or rather deactivation, of protein domain can be a result of many different events which disrupt the coding sequence of protein domain, the explanation of the negative r may be in fact that the protein domain death

happens at faster rate than the duplication of the whole protein domain sequence. After simple algebraic manipulation of μ and σ^2 equations, we obtain the following relation $\mu = 2r\sigma^2/\sigma_e^2$ which explains the negative linear relationship between these two parameters.

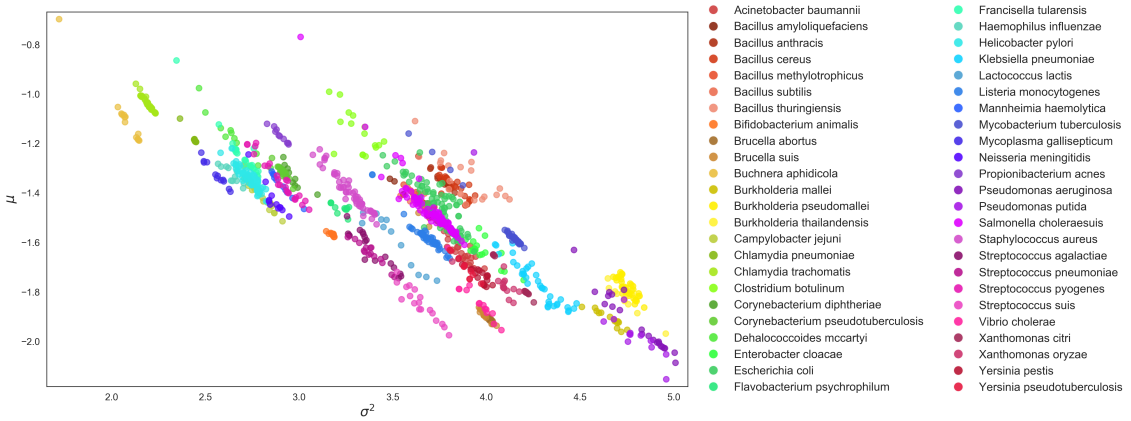


Figure 3.2: Scatter plot of Poisson Log-Normal parameters μ versus σ^2 obtained fitting the protein domains RSAs. Figure shows only those species which are represented with at least 10 different strains in our data set. There are in total 1 173 bacteria which belong to 48 different species. Different colors represent different species, as indicated in the legend.

Furthermore, the μ versus σ^2 plot shows the presence of roughly parallel stripes, which suggests a cluster structure of the data (Figure 3.2). When we depict strains belonging to the same species using the same color, it emerges that the stripes are related to the bacterial taxonomy. From the RSA model point of view, this indicates preservation of protein domain dynamics at the species level. Additionally, the cluster structure of data motivates us to introduce the new approach to bacterial phylogeny using the estimated parameters μ and σ^2.

3.3 Protein domain RSA and evolutionary distance

To assess the ability of protein domain RSA model in estimating the evolutionary distance, we compared our results with bacterial taxonomy and 16S rRNA gene-based phylogeny. Including both taxonomy and phylogeny in comparison may seem redundant because of their intrinsic connection. Not only is the modern microbial taxonomy mostly based on 16S rRNA gene [25], but the cutoffs used in 16S rRNA phylogeny originated from the phenotype-based taxonomy [40]. However, we use complementary information of both approaches. Taxonomy provides predefined levels of classification together with their phenotypic properties. On the other hand,

both parameters of RSA model and 16S rRNA gene alignments produce pairwise distance for each pair of bacteria which allows for comparison between the methods at arbitrary levels. Moreover, we are especially interested in intraspecies level of bacteria where phylogeny is necessary for comparison.

Poisson Log-Normal parameters, μ and σ, and the densities of protein domains are used for calculation of pairwise distances between bacteria. The density serves as additional estimate of protein domain dynamics describing to which extends is the whole bacterial genome populated with protein domains. We refer to these distances as RSA distances. 16S rRNA distances are calculated for aligned sequences of 16S rRNA genes following the standard procedure [38]. Taxonomic levels included in comparison are: phylum, class, order, family, genus and species.

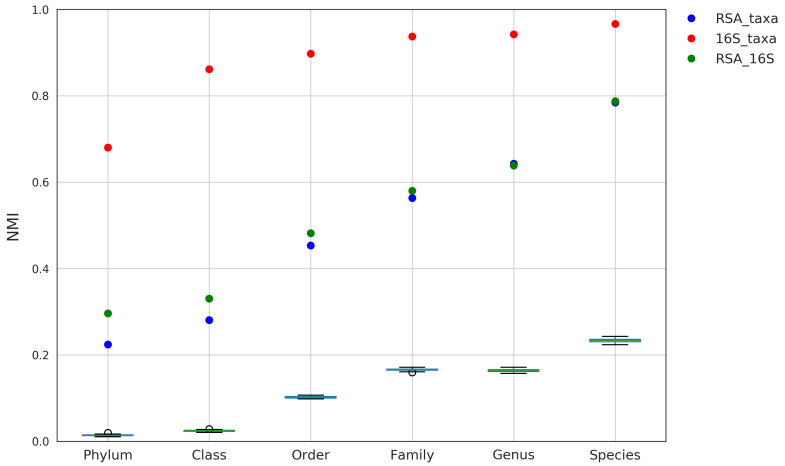


Figure 3.3: Comparison between three clustering solutions on different taxonomic levels: phylum, class, order, family, genus and species (*x-axis*). NMI scores (*y-axis*) are calculated as a measurement of agreement between clusters based on: RSA method and taxonomy (*blue*), 16S rRNA gene and taxonomy (*red*), RSA method and 16S rRNA gene (*green*). The boxplots represent the baselines of NMI score and are based on simulations.

The RSA and 16S rRNA distance matrices were used to perform hierarchical clustering on bacteria. Discrete system of taxonomic classification implied flat cuts of hierarchical clusterings. The cuts were done for different taxonomic levels: phylum, class, order, family, genus and species, fixing the total number of clusters. While enabling the comparison (Figure 3.3), this is arguably not the best way to utilize the information obtained by hierarchical clustering. At each taxonomic levels the Normalized Mutual Information (NMI) is used as a measurement of agreement between different clustering solutions [41]. The theoretical range of the NMI score is interval $[0, 1]$, but NMI is biased towards clustering solutions with more clusters and fewer data points [42]. Consequently, the baseline of NMI score in practise is not zero and relatively high NMI scores can be an artifact caused by the low ratio

between number of bacteria and number of taxonomic groups. To make comparison fair, we used simulations to calculate the baseline of NMI, as shown by the boxplots in Figure 3.3. Looking at colored points in the same figure, we observe that taxonomy and phylogeny have expected high agreement. RSA model is not performing as good, but NMI is still evidently higher than the baseline signifying that the RSA model captures phylogenetic signal to a certain degree. Particularly, observing the differences between the obtained NMI and the baseline, we notice that the performance of RSA method increases on lower taxonomic levels reaching the maximum at species level. The total purity of RSA clustering solution at species level is 0.65 signifying that 65% of bacteria are correctly classified by RSA taking taxonomy as ground truth. However, looking at sizes of clusters and their content we notice that there are some clusters with only 1 bacteria and that some species are abundant in multiple clusters. This indicates that cutting hierarchical tree by fixing the total number of clusters results in information lost. Motivated by this observation and the fact that RSA method performs the best at the species level, we decided to look into the intraspecies level of bacteria.

3.4 Short evolutionary distance: identification of divergent strains

Here we present a few interesting results at the subspecies level identified by the RSA method. As discussed in the introduction, there are many different phylogenetic methods specially modeled for bacterial classification. Placing the RSA method inside the framework of phylogenic inference, it could be classified as distance matrix method, but keeping on mind that RSA-based distances greatly differ from the commonly used distances based on multiple sequence alignments. As the first step of RSA method, we estimate the proteome evolution of a bacterium with two numbers (μ, σ^2) and then using only the three numerical values: μ, σ^2 and the protein domain density, we perform the clustering. Therefore, RSA method is not designed for a specific taxonomic level or for a certain species and can thus be generally applied to all species present in our data set. For the following discussion, we considered only 48 species from our data set which had at least 10 different strains. Since RSA model is estimating the proteome evolution, we expect to see differences even between recently diverged strains. The problem we encountered is the validation of our results at the intraspecies level.

Since the performance of other methods on our data set is out of scope of this work, we focused on the published literature. But even if the true phylogenetic tree is known, and existent, we wouldn't expect for RSA method to perfectly reconstruct it. Instead, we are focused on finding interesting patterns at the intraspecies level such as clustering of subspecies or identification of divergent isolates. Below we discuss 6 bacterial species in more details. For the remaining 42 species we couldn't

find any literature explaining the found patterns.

3.4.1 Intraspecies clustering

Different strains of the same bacterial species can have a different host range. Here we present two examples in which RSA method is able to identify host-based subspecies separation.

Xanthomonas citri: two subspecies of the plant pathogen
Xanthomonas citri subsp. *citri* (XCC) is a causal agent of citrus canker type A, a bacterial disease affecting different plants from the genus *Citrus*. While citrus canker A infects most citrus species, two of its variants, A* and A^W, have a much more limited host range with XCC pathotype A^W infecting only Key lime (*C. aurantifolia*) and alemow (*C. macrophylla*) [21]. Even though three different pathotypes of XCC are currently known, in our data set we have only 17 strains of pathotype A and 5 strains of pathotype A^W [21]. RSA-based clustering of 22 XCC strains identifies two clearly separated clusters (Figure 3.4-a, left) which coincide with two XCC pathotypes. Concurrently, clustering based on 16S rRNA gene failed to identify two pathotypes of XCC (Figure 3.4-a, right). This suggests that even though pathotypes A and A^W have different hosts, their diversification is not distant enough to be reflected by variability of the 16S rRNA gene. On the other hand, protein domain dynamics of two pathotypes succesfully describe different functions of their proteomes.

Another important aspect of the citrus canker is geographical spread of the disease. 22 strains of XCC from our data set have diverse geographical origin. While all A^W strains were sampled from USA, strains of pathotype A originate from USA, Brazil and China. RSA clustering of 17 A-type strains colored by their sampling location shows interesting geographical pattern (Figure S1.5). The similar pattern is obtained by Patane et al. [21] using maximum likelihood tree based on 1785 concatenated unicopy genes. The only strain with different behaviour is jx-6 (GCA_001028285) coming from China. However, this can be explained by the fact that while protein domains of other 21 strains come from one circular chromosome and two plasmids, in strain jx-6 only protein domains from the chromosome were taken into account.

Chlamydia pneumoniae: a different host of the human pathogen
Another bacteria whose host range is identified by the RSA method is *Chlamydia pneumoniae* (*Cpn*). This obligate intercellular parasite is wide-spread in human population causing acute respiratory disease. Besides humans, different animal species can be infected with *Chlamydia pneumoniae*. In our data set, we have nine strains which infect humans (*Homo sapiens*) and one strain isolated from koala (*Phascolarctos cinereus*). Comparison of their 16S rRNA genes shows only a small

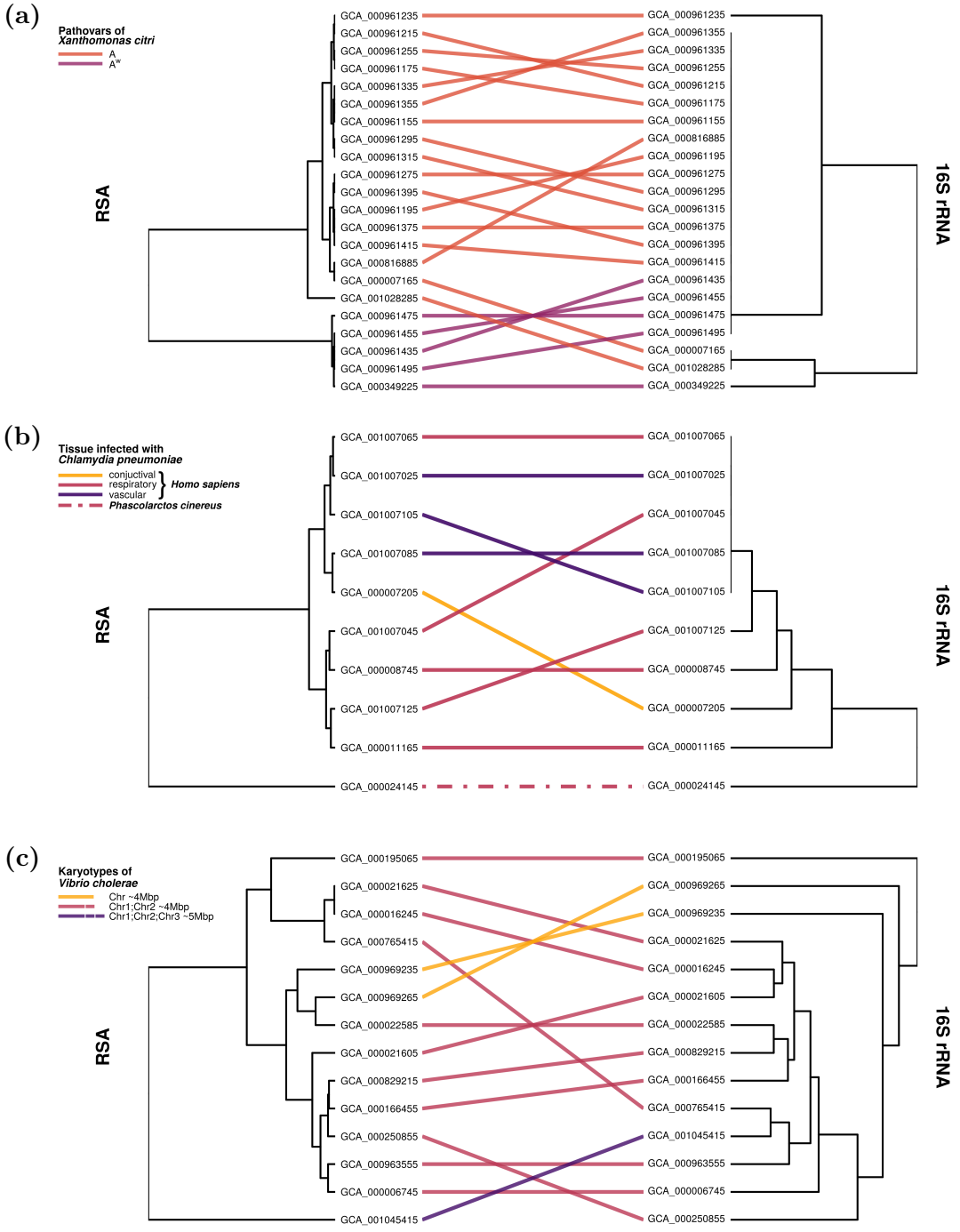

(a)
Pathovars of
Xanthomonas citri
A
A^w
RSA
16S rRNA
GCA_000961235
GCA_000961215
GCA_000961255
GCA_000961175
GCA_000961335
GCA_000961355
GCA_000961155
GCA_000961295
GCA_000961315
GCA_000961275
GCA_000961395
GCA_000961195
GCA_000961375
GCA_000961415
GCA_000816885
GCA_000007165
GCA_001028285
GCA_000961475
GCA_000961455
GCA_000961435
GCA_000961495
GCA_000349225
GCA_000961235
GCA_000961355
GCA_000961335
GCA_000961255
GCA_000961215
GCA_000961175
GCA_000961155
GCA_000816885
GCA_000961195
GCA_000961275
GCA_000961295
GCA_000961315
GCA_000961375
GCA_000961395
GCA_000961415
GCA_000961435
GCA_000961455
GCA_000961475
GCA_000961495
GCA_000007165
GCA_001028285
GCA_000349225
(b)
Tissue infected with
Chlamydia pneumoniae
conjuctival
respiratory
vascular
Homo sapiens
Phascolarctos cinereus
RSA
16S rRNA
GCA_001007065
GCA_001007025
GCA_001007105
GCA_001007085
GCA_000007205
GCA_001007045
GCA_000008745
GCA_001007125
GCA_000011165
GCA_000024145
GCA_001007065
GCA_001007025
GCA_001007045
GCA_001007085
GCA_001007105
GCA_001007125
GCA_000008745
GCA_000007205
GCA_000011165
GCA_000024145
(c)
Karyotypes of
Vibrio cholerae
Chr ~4Mbp
Chr1;Chr2 ~4Mbp
Chr1;Chr2;Chr3 ~5Mbp
RSA
16S rRNA
GCA_000195065
GCA_000021625
GCA_000016245
GCA_000765415
GCA_000969235
GCA_000969265
GCA_000022585
GCA_000021605
GCA_000829215
GCA_000166455
GCA_000250855
GCA_000963555
GCA_000006745
GCA_001045415
GCA_000195065
GCA_000969265
GCA_000969235
GCA_000021625
GCA_000016245
GCA_000021605
GCA_000022585
GCA_000829215
GCA_000166455
GCA_000765415
GCA_001045415
GCA_000963555
GCA_000006745
GCA_000250855

Figure 3.4: (Previous page.) Hierarchical clustering of bacteria at the intraspecies level, comparing solutions obtained by RSA and 16S rRNA method. Each subplot shows a tanglegram with RSA-based dendrogram on the left and 16S rRNA-based dendrogram on the right. Lines connect the same bacteria from two dendrograms. The color/type of the line represents the feature of the bacterium it connects. **(a)** 22 strains of *Xanthomonas citri* belong to two different pathovars: A (*orange*) and A^W (*purple*). **(b)** 10 strains of *Chlamydia pneumoniae* are isolated from different tissues: conjuctival (*yellow*), respiratory (*magenta*) and vascular (*violet*). 9 strains represented with solid line are human (*Homo sapiens*) pathogens while the one strain represented by dashed line is koala (*Phascolarctos cinereus*) pathogen. **(c)** 14 strains of *Vibrio cholerae* are colored based on their karyotype. 11 strains have two circular chromosomes Chr1 ($\sim$ 3 Mb) and Chr2 ($\sim$ 1 Mb) (*magenta*). 2 strains have one $\sim$ 4 Mb long circular chromosome (*yellow*). One strain has three chromosomes Chr1 ($\sim$ 3 Mb), Chr2 ($\sim$ 1 Mb) and Chr3 ($\sim$ 1 Mb) (*violet*).

distance between the koala strain LPCoLN (GCA_000024145) and human-derived strains (Figure 3.4-b, right). Incorporating more genomic information, study based on whole-genome sequencing of four human-derived isolates and strain LPCoLN [43] observed much higher variation between human and koala-derived strains than within the human-derived strains. Additionally, they presented strong evidence that the strain LPCoLN is basal to human isolates. Evolutionary separation of animal and human isolates is reflected in diverse protein domain evolution of their genomes and is consequently recognized by the RSA method. RSA-based clustering clearly separates one animal isolate from the group of highly similar human isolates (Figure 3.4-b, left).

Tissue tropism in *Chlamydia pneumoniae* was focus of the study conducted by Weinmaier et al. They compared whole-genome sequences of multiple *Cpn* strains isolated from different human anatomical sites using animal isolates as an outgroup [22]. The human-derived strains can be divided into conjuctival, raspiratory and vascular based on their tissue of origin. While Weinmaier et al. found a very good agreement between the anatomical origin of strains and the maximum likelihood phylogenetic tree based on all SNPs, they didn't manage to achieve a clear separation between anatomical subgroups of *Cpn*. Morever, small variation in the phlylogenetic tree reconstruction method led to somewhat different phylogenetic tree. This demonstrates how delicate is the process of phylogenetic reconstruction on the low phylogenetic levels where the differences between bacterial genomes are very subtle. RSA method (Figure 3.4-b, left) shows certain correlation with bacterial tissue of origin, but it is also not able to identify a clear pattern. However, since we used only nine human-derived strains it is hard to draw a meaningful conclusion about correlation between RSA-based phylogenetic tree and tissue tropism in *Cpn*.

3.4.2 Different genome composition

A great difference between two bacterial genomes is undoubtedly linked to different dynamics of their protein domains. In fact, we observe strong correlation between parameter σ^2 of protein domain RSA distribution and the genome length of bacteria (Figure S1.1). One explanation for this could be in the definition of the parameter σ^2 which is equal to $\sigma^2 = \sigma_e^2/2\gamma$. Since smaller genome represents a scarcer environment for the populations of protein domains, we expect to see stronger density regulation. The multiplicative constant in the Gompertzian density regulation function γ would thus be inversely correlated to the genome length of bacteria. Under reasonable assumption that environmental noise of protein domain dynamics σ_e^2 is independent of the genome length, we obtain that parameter σ^2 is as well inversely correlated to γ and as a consequence positively correlated to the genome length. However, this general trend is lost at the subspecies level where genome lengths are almost constant (Figure S1.1). Still, there are some bacterial species with great genomic diversity among strains and here we present two examples.

Vibrio cholerae: a strain with different karyotype composition
The causative agent of cholera disease, bacteria *Vibrio cholerae* has an interesting genomic composition. While majority of bacterial genomes consist of one circular chromosome, *Vibrio cholerae* is well studied for its bipartite genome. The usual genome of *Vibrio cholerae* consists of two chromosomes of unequal size, Chr1 with length of $\sim$ 3 Mb and significantly smaller Chr2 with length of $\sim$ 1 Mb. However, not all strains of *Vibrio cholerae* have the same karyotype. Two strains, 1154-74 (GCA_000969235) and 10432-62 (GCA_000969265), underwent the process of chromosomal fusion and have only one $\sim$ 4 Mb long circular chromosome. Whole genome comparison showed a high degree of synteny between two single-chromosome strains and standard two-chromosome strains [44]. On the other hand, strain TSY216 (GCA01045415) has $\sim$ 1 Mb long replicon in addition to Chr1 ($\sim$ 3 Mb) and Chr2 ($\sim$ 1 Mb). Whole genome comparison with the representative of two-chromosome *Vibrio cholerae* strain revealed that Chr1 and Chr2 share almost identical gene content, but the third replicon does not share conserved regions with Chr1 and Chr2 [45]. From the proteome point of view, we expect for single and two-chromosome strains to be highly similar while three-chromosome strain should have almost $\sim$ 25% more coding sequences. This large-scale diversity of *Vibrio cholerae* proteomes is recognized by RSA method which clearly identifies strain TSY216 as an outlier (Figure 3.4-c, left). On the contrary, mutations of 16S rRNA gene don't reflect the changes in karyotype if the 16S rRNA gene-carrying chromosome retains high similarity. *Vibrio cholerae* strains from our data set have from 7 to 10 copies of 16S rRNA gene and all of them are located on $\sim$ 3 Mb long chromosome. Since this chromosome shows high synteny across all strains, 16S rRNA gene-based clustering doesn't identify any significant difference between strains (Figure 3.4-c, right).

Buchnera aphidicola: a strain with abundance of pseudogenes
We should keep on mind that the RSA method is estimating evolution of proteome through the dynamics of its protein domain content. While the abundance of protein domains in bacteria is generally well approximated with their genome length, in reality bacterial proteome and genome represent two different points of view. Interesting examples of this incogruity are bacterial genomes which have unusually large proportion of pseudogenes which significantly reduces their proteomes. Here we present results about *Buchnera aphidicola*, bacterial species which is in mutualistic endosymbiotic relationship with different aphids (members of superfamily *Aphidoidea*). As many endosymbionts, it underwent the process of genome reduction as an adaptation to the host-associated lifestyle and has a genome with length < 1 Mb. One of the main processes which contributed to genome reduction is gene inactivation followed by progressive gene disintegration [46]. Pseudogenization is thus an intermediate step between an ancestral complete genome and a modern reduced genome. Among 13 strains of *Buchnera aphidicola* in our data set, for 12 of them number of pseudogenes ranges from 7 to 63 with the thirteenth strain JF98 (GCA_000183305) having remarkable 176 pseudogenes. Since all strains have a similar total number of genes (protein coding and pseudogenes), proteome of the strain JF98 is significantly smaller. This is detected by the RSA method which identifies strain JF98 as an obvious outlier (Figure S1.2-a). While RSA-based phylogeny does good job in detecting a strain with significantly smaller preoteome, it doesn't perform very well in host reconstruction. Specifically, RSA-based phylogeny of *Buchnera aphidicola* is not congruent with the phylogeny of its hosts, even though it does show certain correlation. In this regard, 16S rRNA-based clustering performs much better (Figure S1.2-b). However, this is not surprising since 16S rRNA-based phylogenies of endosymbionts are in general congruent with the phylogenies of their hosts. The reason behind the success of 16S rRNA method is the fact that the endosymbiont and its host co-evolve together from the moment their endosymbiotic relationship starts [23].

3.4.3 Distinct bacterial isolates

For several species in our data set, RSA-based phylogenic tree indicates outliers. While we believe that there is a biological reason for these strains to diverge, reading the literature we were able to explain the clustering pattern for only two such species. Other possibility for existance of an outlier would be a notable mistake in whole-genome sequencing. Identification of such a mistake would also be a valuable result, albeit hard to verify.

Listeria monocytogenes: two clonal strains
Listeria monocytogenes is a food-borne pathogenic bacterium which causes listeriosis in humans. From the 48 *Listeria monocytogenes* strains present in our data set,

RSA method identifies a subgroup of two strains (Figure S1.3). These two strains, La111 (GCA_000382925) and N53-1 (GCA_000382945), seem to have very similar proteome composition which differs from proteome compositions of other strains. Holch et al. investigated strains La111 and N53-1 in their study of bacterial persistence in *Listeria monocytogenes* [47]. They found that these two strains which were isolated 6 years apart from different Danish fish processors, are extremely similar and collectively different from other analyzed strains based on whole-genome analysis. Moreover, they found that they differ only in 2 proteins which explains our results.

Francisella tularensis: an attenuated strain of the pathogenic bacteria
Francisella tularensis is intracellular bacterium which is a causal agent of tularemia. In our data set we have 25 *Francisella tularensis* strains from different subspecies: *tularensis*, *holarctica*, *mediasiatica* and *novicida*. While neither RSA nor 16S rRNA perform well in separating subspecies of *Francisella tularensis*, RSA method identifies one outlier, strain TIGB03 (GCA_000248415) (Figure S1.4). Unlike other 24 virulent strains, strain TIGB03 is an attenuated *tularensis* strain. This strain was described by Modise et al. as an attenuated O-antigen mutant of the virulent strain TI0902 (GCA_000248435) [24]. Indeed, we notice that 16S rRNA genes of these two strains are identical (Figure S1.4) and thus 16S rRNA-based clustering is unable to detect a mutant strain. Comparing whole-genome sequences of strains TIGB03 and TI0902, Modise et al. found 31 nonsynonymous point mutations and 75.9 kb long duplicated region in the mutant strain TIGB03. It may seem that this difference in genome length is the main cause why our method manages to distinguish mutant strain from the others. However, genome lengths of 25 strains we included in our study range from 1.86 to 2.05 Mb and the genome of strain TIGB03 is far from extreme with the length of 1.97 Mp. Thus, we believe that in fact nonsynomymous mutations led to different proteome composition which is recognized by our method.

Chapter 4

Conclusion

Engen and Lande's population dynamics model proved to work well for the protein domain abundances in bacteria. Poisson Log-Normal fit of the protein domain RSA distribution outperformed the alternative Negative Binomial and Log-Series fits. This suggests that the Gompertzian density regulation is a better assumption for protein domains than the alternative linear regulation function. Since Gompertzian function has a weaker effect on abundant protein domains, we expect to observe some very abundant protein domains. This is in fact true, since on average a bacteria from our data set has at least one protein domain with more than 500 copies in the bacterial genome. This is in concordance with our prior knowledge about especially abundant protein domains called protein modules. Another interesting observation is the overlap of Negative Binomial and Log-Series fitting distribution which suggests that the dispersion parameter of the Negative Binomial distribution converges to zero. Interpreting the dispersion parameter in the context of Engen and Lande's model, this means that the growth rate r of protein domains is very close to zero or even negative. The negative growth rate r is as well inferred from the estimated parameters of the Poisson Log-Normal model. Since the growth rate r is a difference between the birth rate and the death rate, it seems that protein domain deactivation is happening at the faster rate than the duplication.

Moreover, estimated parameters μ and σ^2 of the Poisson Log-Normal RSA fit reflect the evolutionary distance of bacteria. This is visible from the cluster structure of bacteria plotted in the (μ, σ^2)-space. Adding the additional parameter of protein domain density, we defined the RSA phylogenetic distance as Euclidean distance in the normalized three-dimensional space. The use of Euclidean distance has benefits in the subsequent analysis because it allows for the use of methods such as Ward's minimal variance method which is especially modeled for the Euclidean distance. RSA phylogeny proves to be in good agreement with both taxonomy and 16S rRNA-based phylogeny at different taxonomic levels with the best agreement at the species level.

In this work, we are specially interested in the intraspecies level of bacteria. The

bacteria shows great diversity at the intraspecies level where certain strains of the same species can have different phenotypic properties. As an important example, in some species only some strain are pathogenic. However, this diversity is not classified by the classical taxonomy. For this reason, at the intraspecies level we rely on different phylogenetic methods to quantify the differences among bacteria. Unfortunately, the choice of an appropriate phylogenetic method is not an easy task. The generic methods, such as 16S rRNA method, which work well for higher taxonomic ranks usually have low phylogenetic power at the intraspecies level. On the other hand, phylogenetic methods designed for a special bacterial species may work well on the species they are designed for, but generally can't be extended to other bacterial species. RSA phylogenetic distance is a generic method which can be applied on any bacterium with sequenced genome. Despite the generality of the method, it managed to detect some interesting patterns at the intraspecies level which are in comparison missed by other generic 16S rRNA method.

In certain special cases of intraspecies diversity, the RSA method performs well. As example, in the case of clear separation of bacterial subspecies, as with *Xanthomonas citri*, the RSA method manages to correctly classify bacteria in the subgroup. Additionally, great differences in the genome, such as different karyotype composition or excessive pseudogenization, are reflected in RSA phylogenetic distance. Finally, the RSA phylogeny works best in separating outlier strains. We believe that the strength of RSA phylogenetic method is exactly at the intraspecies level in detecting the outlier strains. In case that the clear phenotypic explanation for the outlier is missing, it may potentially suggest the mistake in genome sequencing.

Finally, our results suggest that studying the biodiversity of protein domains in bacteria through the population dynamics RSA model gives interesting insights into the evolution of bacterial proteome and consequently the evolution of bacteria. The most interesting result is the new approach to phylogenetic distance which performs well at the intraspecies level of bacteria.

Part II

Evolutionary model of gene length: generalization of Engen and Lande's model in presence of diverse subcommunities

Chapter 5

Introduction

5.1 The evolution of genome

All living organisms which inhabit our planet are specified by their genomes. A genome contains the biological information needed to build the organism and maintain its biological functions. This information is stored in every cell of the living organism and is divided into multiple molecules called chromosomes. From the structural point of view, genome is made of DNA. A molecule of DNA is a linear polymer built from four nucleotides: adenine (A), cytosine (C), guanine (G) and thymine (T). In living cells, two linear DNA molecules bond forming a double helix. Chemical bond between two single stranded DNA molecules happens following a predefined bonding pattern where adenine forms a hydrogen bond with thymine and cytosine forms a hydrogen bond with guanine. The deterministic bonding pattern between nucleotides creates A-T and C-G base pairs (bp). The biological information stored in a genome sequence is written with these four letters. For it to become useful for the cell, the written genetic code has to be expressed. The complex process of genome expression starts with the transcription. Small pieces of the genome, called genes, are individually copied into RNA molecules, one stranded polymers with similar structure as DNA. In case a transcribed gene is coding for a protein, in the following process of translation a protein is synthesized from the given RNA template. In more complex organisms, the transcribed RNA usually undergoes a process of alternative splicing which enables for the same gene to produce multiple different proteins called splice isoforms [48].

First cellular life forms appeared on our planet billions of years ago (∼3.5 bya). During the course of evolution, an incredible diversity of life evolved from the simple polynucleotides. Diversification of increasingly complex genomes is the basis of the ongoing process of evolution. Genome changes result from the accumulation of small-scale alterations of the genome sequence. These alterations of short nucleotide sequences in the genome are called mutations. Point mutation occurs when a single nucleotide is replaced by another nucleotide. The other types of mutations

are caused by insertion or deletion of one or a few nucleotides. Mutations generally arise from errors in DNA replication but can also be caused by different chemical or physical mutagens. Many mutations are corrected by the cell's repair mechanisms. If the remaining mutation proves to be lethal for the cell, it is wiped out from the genome together with the cell in which it occurred. In other cases, the effect of persisting mutation varies in intensity and it can even be neutral. Other larger in scale alterations of the genome sequence can happen, such as recombination and transposition. These two processes result in a rearrangement of the DNA segments within the genome. For any genome alteration to be inherited, in the case of multicellular organism, it has to happen in a germ cell. Through the course of evolution, some of the genome alterations become fixed in the population while others are deleted [48].

From the vast genome sequence, which can reach billions of base pair in length, we are mostly interested in the small transcribable sections called genes. While an existing gene gains new functions as a result of gene sequence alterations cause by mutations, a new gene enters the genome in two different ways. Firstly, it can be a product of gene duplication. After the duplication, the new gene is "free" from the selective pressure of evolution and can acquire a new function. Genes can also be acquired from the other species in the process called lateral gene transfer. In bacteria and archaea new genes are frequently gained this way. While much less frequent, lateral gene transfer still happens in eukaryotic species [48].

With the aim to study the evolution of living organisms and to understand the existence of life, scientist sequenced genomes of multiple species. To make sense of it, they developed the tools to locate the genes inside the genome and to assign the function to found genes. To deepen the understanding, many different aspect of genes are studied such as gene expression levels or gene alternative splicing such as many complementary information like genome methylation [48]. Among all the layers of information, one simple gene attribute seems to be a good estimator of much more complex genetic properties. Grishkevic and Yanai [49] showed that the gene length together with expression level influences the genetic novelties in human and mouse. Two mechanisms of genetic novelties, gene duplication and alternative splicing, are negatively correlated meaning that genes with large genetic families have small number of splice isoforms. However, if we account for the length of the gene and its expression level, then the negative correlation is lost. On its own, gene length is positively correlated with alternative splicing and negatively correlated with gene family size. Other study showed the negative relationship between the gene length and genome-wide expression levels [50]. During the evolution the genes increased in length, due in part to the insertion of transposable elements (TEs) [49]. The gene elongation is considered to be one of the most important mechanisms in the evolution of functional complexity [51]. It is well known that the mean gene length is much larger in eukaryots than in prokaryotes. The orthologous genes found both in prokaryotes and eukaryotes are on average longer in eukaryotes, suggesting that the genes underwent the process of elongation in their evolutionary path. However, the average gene length in different eukaryotic species seems to be highly preserved,

despite the great variability in the lengths of different genes [51]. The distribution of gene length is long-tailed with many shorter genes and a few especially long genes. In this work, we study the dynamic processes which led to the observed gene length distribution in metazoan genomes.

5.2 Evolutionary model of gene length

In all metazoan genomes available at Ensembl [52], we observed the long-tailed gene length distribution, as will be discussed in the Results. To better understand the processes which led to it, we employ the population dynamics model for relative species abundance (RSA) distribution. In this approach, we redefine the meaning of *species* and *community*. The communities which we are going to study are different metazoan genomes. More precisely, we observe the genome only as a collection of coding sequences, neglecting the existence of the rest of the genome. This approach doesn't violate the assumptions of population dynamics since in a regular population dynamics study one generally observes only a certain subcommunity, such as animals or plants. Inside the genome, the nucleotides are interpreted as individuals and genes as species. All nucleotides in the gene sequence belong to the single gene species and their number is the abundance of the gene. In other words, the length of a gene is the abundance of its nucleotides. Using this approach, the gene length distribution is modeled as the relative species abundance (RSA) distribution.

We use Engen and Lande's population dynamics model for RSA distribution [4, 5] to model the dynamic processes leading to the gene length distribution. This model is described in an introductory Section 0.5 and here we discuss only the interpretation for the gene length dynamics. We assume that new species enter the community at some rate. If we neglect the effects of the lateral gene transfer, then the rate by which new genes enter the genome is the rate of gene duplication. A gene is acquiring mutations from its entry to the genome until its deactivation. The process of gene elongation is assumed to be a diffusion process which solves the stochastic growth equation

$$\frac{dx}{dt} = rx - xg(x) + x\sigma_r(x)\frac{dB(t)}{dt}, \tag{5.1}$$

where r is a constant growth rate, $g(x)$ is a density regulation function and $\sigma_r^2(x) = \sigma_e^2 + \sigma_d^2/x$ is a stochastic variance consisting of environmental and demographic stochasticity. The growth rate can be written as a difference between birth and date rates $r = b - d$. Birth rate of a nucleotide base is in fact an insertion rate of a single nucleotide. Similarly death rate is a deletion rate. In this model, all other types of mutations are approximated by a single nucleotide insertion and deletion. Insertions (or deletions) of larger sequences can be described as a sum of multiple single nucleotide mutation. Demographic stochasticity is due to possible differences in insertion/deletion rates for four nucleotide bases, while environmental stochas-

ticity can be caused by recombinational events such as exon shuffling. Namely, in the case of random recombination the larger genes have greater probability to be affected.

The role of density regulation $g(x)$ is to keep species abundances from "exploding". This is usually a consequence of limited resources in the community, but when talking about genome it can be interpreted as the limit in total genome size [19]. Since the average gene length is highly conserved in different eukaryotic species despite their different evolutionary history [51], it seems that barrier for the length of a gene exists even though the mechanisms and reasons of its existence are not clear. We will consider two different density regulation functions. The first one is a Gompertzian function $g(x) = \gamma \ln(x + \epsilon)$, where $\epsilon = \sigma_d^2/\sigma_e^2$, which leads to the Poisson Log-Normal RSA distribution

$$RSA \sim PoiLN\left(\mu = \frac{r}{\gamma}, \sigma^2 = \frac{\sigma_e^2}{2\gamma}\right). \tag{5.2}$$

The other type of density regulation with the function $g(x) = \eta x$ results in Negative Binomial RSA distribution

$$RSA \sim NB\left(\alpha = \frac{2(r + \eta\epsilon)}{\sigma_e^2}, p = \frac{\sigma_e^2}{2\eta + \sigma_e^2}\right). \tag{5.3}$$

We will also consider the Log-Series RSA distribution as a special case of Negative Binomial distribution when the dispersion parameter $\alpha \to 0$.

In this work, we analyze ∼300 metazoan gene length distributions. If the population dynamics approach is an appropriate choice for the gene length dynamics, then we expect to get a good fit with Log-Series, Negative Binomial or Poisson Log-Normal distribution.

So far, we considered for all genes in a genome to form a single community. However, there are different groups of genes and for a proper model it is necessary to better understand their biological functions and their mutual connection.

5.3 Genome as a mixture of different gene types

A typical genome of the multicellular organism consists of two parts: the nuclear genome and the mitochondrial genome. The mitochondrial genome is a small circular DNA molecule which has multiple copies and is found inside the mitochondria, the cell organelle. Nuclear genome is a large DNA molecule divided into multiple chromosomes [48]. When we refer to the genome in this work, we refer only to the nuclear genome.

Protein-coding genes are small sections of the genome which contain the instructions for the protein synthesis. Because of their protein-coding ability, they are the most studied and best annotated of all gene types. However, there are many other

important gene types scattered across the genome. The genes which undergo the process of transcription, but are not translated into proteins are called noncoding RNA genes. Noncoding RNAs have many important functions in the cell some of which have been recently discovered and many which still remain to be found. Noncoding RNAs shorter than 200 nucleotides are called *short noncoding RNAs* (sncRNAs) and the longer ones are called *long noncoding RNAs* (lncRNAs). Two most important groups of sncRNAs are ribosomal RNAs (rRNAs) and transfer RNAs (tRNAs). The rRNAs are components of ribosomes, the structures within which the protein synthesis takes place. The tRNAs are also involved in the process of protein synthesis, where they carry the amino acids to the ribosomes and ensure that the protein is correctly synthesized. Some of the other sncRNAs are microRNAs (miRNAs), short interfering RNAs (siRNAS), small nuclear RNAs (snRNAs), small nucleolar RNAs (snoRNAs), piwi-interacting RNAs (piRNAs) and vault RNAs. They have diverse functions inside the cell with some of them controlling the expression of their target genes. The role of many sncRNAs is still not known and the new discoveries are constantly being made. The lncRNAs are even more enigmatic and despite the evidence that certain diseases are associated with the transcription of lncRNAs, there are still doubts in the scientific communities about their usefulness [48]. We may note that the lncRNA serves as an umbrella term for many different categories of nonoding RNA which are longer than 200 nucleotides and which may have very different function [53].

Another important group to consider are the *pseudogenes*. Pseudogene is a small section of the genome whose sequence resembles protein-coding gene, but which doesn't code for a functional RNA or protein. They are derived from the genes which lost their protein-coding ability during the evolution. By their origin, they are divided into nonprocessed and processed pseudogenes. Nonprocessed pseudogenes arise from the mutation which disrupted the coding sequence of the gene. If the mutation happened to the gene which is a part of multigene family, then it is further call duplicated pseudogenes. In this case the other genes from the gene family are still active and the event is usually not deleterious to the organism and a pseudogene is retained. If the mutation happens to the gene which is not part of the multigene family then the resulting pseudogene is called unitary pseudogene. Since in the process of pseudogenization the function of the gene is lost the effects on the organisms can be substantial which makes unitary pseudogenes rare. Processed pseudogenes are products of the reverse transcription. They are derived when the mature mRNA copy of the gene is reinserted into the genome. Consequently, the processed pseudogenes lack the introns. The pseudogenes are often called the evolutionary relics and they can be valuable when studying the evolutionary history of the genome. In the recent years, the growing number of pseudogenes are found to have important biological role and their significance is being reevaluated by the scientific community [48, 53].

In an evolutionary study of the genome, it clearly makes sense to consider not only protein-coding genes but also the short and long noncoding RNAs as well as

pseudogenes. These groups of genes which inhabit the same genome have intertwined evolutionary paths. The pseudogenization process converts once protein-coding genes into pseudogenes. However, it is not exclusively one-way process since there are evidences of "resurrected" pseudogenes which regained the protein-coding ability [54]. On the other hand, a long noncoding RNA can also be a product of "resurrected" pseudogene and some of these lncRNAs are reported to be associated with human diseases [55]. Moreover, several lncRNAs in the human genome originated from ancestral protein-coding genes [56]. In the future, the definitions of some of these gene types may change with new discoveries of their functional potential and, in fact, the scientist are already questioning some of the currently used terminology [53].

In the literature, the term "gene" is often exclusively used for protein-coding genes. However, for simplicity, in this work we extend the definition of gene to all sections of genome which have protein-coding, regulatory or evolutionary relevance. Thus in the following chapters, we talk about four different gene types/biotypes: coding genes, short noncoding genes, long noncoding genes and pseudogenes.

Chapter 6

Materials and methods

6.1 Data retrieval

Gene length data for multiple metazoan species was downloaded from Ensembl, a genome annotation and dissemination platform. Specifically, data for 170 species was downloaded from Ensembl release 101 [52] and data for 106 species was downloaded from Ensembl Metazoa release 48 [57]. Ensembl Metazoa is part of Ensembl Genomes which offers data for metazoan non-vertrebate species. For the downloading process, we used R "biomaRt" package v.2.44.4 [58] which offers programmatic access to the Ensembl BioMart data management system [59]. Data for certain species were not downloaded. Assemblies for which the genome build method was described as "projection build" were not downloaded. Additionally, some species had multiple available assemblies of which we downloaded only one. The list of all such species and the corresponding list of chosen assemblies is given as supplementary table S2.1. Finally, the data for *Saccharomyces cerevisiae* which is available at Ensembl is not downloaded since we were interested only in metazoan genomes. For each of 276 species, we obtained a table with the list of genes described by the attributes ("ensembl_gene_id", "description", "chromosome_name", "start_position", "end_position", "gene_biotype").

Ensembl gene biotypes are divided into multiple groups (coding, pseudogene, short noncoding RNA, long noncodingRNA). To obtain the definition of different Ensembl gene biotype groups, we used the Ensembl REST API [60]. The full list of all gene biotypes found in the data and the groups to which they belong, can be found in the supplementary table S2.2. As a reference for the chromosome name, we downloaded the list of toplevel sequences for every species using Ensembl REST API [60].

For every species, we obtained the taxonomic classification at multiple taxonomic levels (Phylum, Class, Order, Family, Genus). This information is obtained with the help of the function *taxonomy* from the R package "taxize" v.0.9.99 [61] which is based on the powerful R package "myTAI" v.0.9.2 [62]. The "db" parameter of

the "taxonomy" function was set to the NCBI database. A few missing taxonomic classifications were filled based on The Catalogu of Life [63].

6.2 Data preparation

For every species separately, we cleaned the data in the following steps. Since Ensembl list of genes may contain multiple transcripts per gene, we firstly kept only the genes with the unique Ensembl gene ID. Concerning the gene biotypes, we removed the transposable elements and "TEC" (to be experimentally confirmed) genes. Based on the list of dowloaded toplevel sequences, we kept only those genes which are located on the toplevel sequences. For the genome assembly, the toplevel sequences consist of chromosomes and any unlocalised or unplaces scaffolds which together define a primary assembly. In this preparation step, 6402 genes were removed for *Homo sapiens*, 779 for *Mus musculus* and 4721 for *Danio rerio*. Additionally, we removed all mitochondrial genes.

In this work, the *gene length* is defined as the number of nucleotide bases (base pairs) in the gene. It is calculated as the difference between the end and start position of the gene plus 1.

For the decisions of biological relevance in the steps of data retrieval and data preparation, we consulted our collaborator Dr. Maria Giulia Bacalini from the IRCCS Istituto delle Scienze Neurologiche di Bologna (Italy).

6.3 Fitting the gene length distribution for a single gene biotype

For every species, genes were divided in four groups based on their biotype: coding, pseudogene, short noncoding and long noncoding. Each of the biotype groups was separately analyzed. In case at least 100 genes of a certain biotype were found, the gene length distribution was fitted with the Maximum Likelihood Estimation method implemented in R "sads" package v.0.4.2 [34]. Data were modeled with a truncated Poisson Log-Normal, a truncated Log-Series and a truncated Negative Binomial distribution, so that to test and compare different ecological hypothesis. Two species, *Mus caroli* and *Mus pahari*, reported error in fitting of protein-coding genes with Negative Binomial distribution. In total, gene length distribution of coding, pseudogene, short noncoding and long noncoding genes was successfully fitted for 274, 153, 267 and 106 species, respectively.

Akaike Information Criterion (AIC) [35] and Kolmogorov-Smirnov (KS) statistic were computed to assess the models performances and for model selection. The KS statistic for two empirical cumulative distribution function is given by

$$KS(F_{1,n}, F_{2,m}) = \sup_x |F_{1,n}(x) - F_{2,m}(x)|, \tag{6.1}$$

where n and m are the numbers of observed samples. It is obvious that even for samples coming from the same distribution, the KS statistic depends on the number of observed samples. Thus, when calculating the KS statistic, we randomly sampled 100 gene lengths. Then we calculated the KS statistic between this sample and a sample from the fitted distribution of the same size. Repeating the process 10 times, we obtained the mean KS statistics (Figure 7.1-b). Using a fixed sample size of 100 was important so that we can compare the values of KS statistic for different species and different biotypes since some gene length distributions had more than 20000 genes while others had nearly the minimum 100. Finally, we used simulations to determine the minimum empirically possible value of KS statistic with the number of observations equal to 100. For 1000 times we drew two samples from the same Poisson Log-Normal distribution and calculated the KS statistics. The μ and σ^2 parameters were based on mean parameter values estimated for different biotypes. Combining all KS statistics, we got the mean and standard deviations which are used in the Figure 7.1-b.

6.4 Modeling the RSA distribution in presence of diverse subcommunities: generalization of Engen and Lande's model

Here we discuss the generalization of Engen and Lande's model [4, 5] when the community is divided into diverse mutually disjoint subcommunities. Let us assume that the species are divided into K groups where each group corresponds to a certain subcommunity. For a random observed species, let p_i be the probability that the species belongs to the ith group. Then the membership to subcommunity is a categorical random variable with vector of probabilities $(p_1, \ldots, p_K), \sum_{i=1}^{K} p_i = 1$.

Suppose that every subcommunity grows according to the Engen and Lande's model. Then the abundance model for the ith community $\lambda_i(x)$, with parameters of $\lambda_i(x)$ depending on the populaton dynamics inside the subcommunity, will be of Log-Normal or Gamma form depending on the density regulation function.

If we now again observe the total community as the whole, then the abundance model of the community is a mixture

$$\lambda(x) = \sum_{i=1}^{K} p_i \lambda_i(x). \tag{6.2}$$

Adding Poisson sampling variability, as explained in an introductory chapter, the RSA distribution of the community is the Poisson distribution with the parameter given with (6.2). However, this is the same as the mixture of $Poisson(\lambda_i)$ distribu-

tions, as is clear from the derivation

$$P_r = \int \frac{x^r e^{-x}}{r!} \left(\sum_{i=1}^{K} p_i \lambda_i(x) \right) dx$$
$$= \sum_{i=1}^{K} p_i \int \frac{x^r e^{-x}}{r!} \lambda_i(x) dx, \tag{6.3}$$

where P_r is a probability that the RSA distribution takes value r. If we assume that the density regulation function for the total community is of Gompertzian type then

$$RSA \sim \sum_{i=1}^{K} p_i \, PoiLN(\mu_i, \sigma_i). \tag{6.4}$$

Similarly, for the linear density regulation function, the RSA is a mixture of Negative Binomial random variables. The mixture of both Poisson Log-Normal and Negative Binomial distributions is also possible, in case of different density regulations within the subcommunities.

Remark. Note that the crucial assumption of this generalization is that the species are partitioned into mutually disjoint groups. Otherwise, if we were to allow the possibility of a species which belongs to more subcommunities, the equation (6.2) wouldn't hold. As an example, if the same species has the abundance λ_1 in one subcommunity and abundance λ_2 in another subcommunity then when observing total community, this species would have abundance of $\lambda_1 + \lambda_2$. Thus this model is not applicable when joining observations from multiple communities with shared species. However, it is applicable to the gene length RSA model because, by definition of the species, the nucleotides of one gene can't belong to different genomic subcommunities since we observe a gene as unity.

6.5 Bayesian modeling of the gene length distribution

We used Python "pymc3" package v.3.9.3 [64] for the Bayesian data analysis. Since the Poisson Log-Normal distribution is not implemented in the package, we extended the package adding our implementation of the Poisson Log-Normal distribution. The new class *PoissonLognormal* was defined as a child class of the "pymc3" class *Discrete*. For a log-likelihood method of the class, we took the logarithm of the Bulmer's approximation of the Poisson Log-Normal probability function [14]

$$PoiLN(r; \mu, \sigma^2) \approx \frac{(2\pi\sigma^2)^{-1/2}}{r} e^{-\frac{(\ln r - \mu)^2}{2\sigma^2}} \left[1 + \frac{1}{2r\sigma^2} \left\{ \frac{(\ln r - \mu)^2}{\sigma^2} + \ln r - \mu - 1 \right\} \right]. \tag{6.5}$$

The approximation (6.5) is highly accurate for large values of r with the relative error less than 10^{-3} when $r \geq 10$ [14], while for smaller values of r numerical calculations should be used instead. However, since the genes are generally longer than 10 nucleotides, we use the approximation formula (6.5) for all values of r.

Suppose the observed community is a collection of K diverse communities. Then depending on a chosen population dynamics model, the RSA distribution for the total community is a mixture of K Poisson Log-Normal or K Negative Binomial variables. We fit the K-Poisson Log-Normal mixture using the following Bayesian model [65]

$$\begin{aligned} \mu_i &\sim Normal(mu = 0, sd = 10), \quad i = 1, \ldots, K \\ \sigma_i &\sim Gamma(mu = 1, sd = 10), \quad i = 1, \ldots, K \\ (p_1, \ldots, p_K) &\sim Dirichlet(a = (\alpha_1, \ldots, \alpha_K)) \\ data &\sim \sum_{i=1}^{K} p_i \, PoiLN(\mu_i, \sigma_i^2), \end{aligned} \tag{6.6}$$

where the hyperparameter a specifies our prior belief about relative sizes of the subcommunities.
Similarly, we define K-Negative Binomial mixture with

$$\begin{aligned} \mu_i &\sim Gamma(mu = 10^4, sd = 10^3), \quad i = 1, \ldots, K \\ \alpha_i &\sim Gamma(mu = 1, sd = 10), \quad i = 1, \ldots, K \\ (p_1, \ldots, p_K) &\sim Dirichlet(a = (\alpha_1, \ldots, \alpha_K)) \\ data &\sim \sum_{i=1}^{K} p_i \, NB(\mu = \mu_i, \alpha = \alpha_i). \end{aligned} \tag{6.7}$$

The "pymc3" implementation of the Negative Binomial variable is parameterized by the mean $\mu = \frac{p\alpha}{1-p}$ and dispersion α, where p is a success rate.

The same Bayesian models can be used in case of a homogeneous community with $K = 1$. In this case the Dirichlet variable is not needed.

We use a traceplot to monitor the convergence of Monte Carlo Markov Chains (MCMC) [64, 65]. In every performed sampling in this work, we drew 2 000 samples with the default burin period of additional $000 samples. Initialization method for the NUTS sampler w$
$adapt\_diag$". $To assess the model accuracy, we per formed the posterior predictive checks \, citekruschke 2$

6.5.1 Human, mouse and zebrafish

To fit human and mouse gene length distribution we used Poisson Log-Normal and Negative Binomial mixture model with $K = 3$. In both models, we specified $a = (10, 40, 50)$ as a hyperparameter of Dirichlet distribution. For zebrafish, we used the same mixtures with $K = 2$ and Dirichlet hyperparameter $a = (10, 90)$.

For every distribution fitting, we drew a random subsample from the gene length distribution of size 3 000 and sampled the trace using the subsample as observed data.

6.5.2 Mammals: pseudogenes and protein-coding genes

For every mammalian species, we observed two different distributions: gene length distribution for protein-coding genes (coding distribution) and gene length distribution for both pseudogenes and protein-coding genes (pseudo-coding distribution). Instead of the whole distributions, we used a random subsample of size 3000 for MCMC sampling. For coding distribution we used a single Poisson Log-Normal distribution ($K = 1$) and for pseudo-coding distribution we used a mixture of Poisson Log-Normal distributions with $K = 2$ with Dirichlet hyperparameter $a = (40, 60)$.

To assess the goodness of fit, we calculated Kolmogorov-Smirnov (KS) statistic and Wasserstein distance between random subsamples of gene length distribution ($n = 3\,000$) and subsamples from posterior predictive distribution ($n = 3\,000$) . Gene length subsamples were taken from the genes which weren't use in model fitting. The final KS statistic and Wasserstein distance are mean values from 10 repetitions.

To obtain a point estimate of a parameter from Bayesian model, we took the mean value of its trace.

6.6 Differentiation between pseudogenes and protein-coding genes

We propose a model for distinguishing between pseudogenes and protein-coding genes based on their length. From the mixture of pseudogenes and protein-coding genes, we take a subsample of size n and fit it with Poisson Log-Normal mixture for $K = 2$. In case of Bayesian fitting, we then take mean values from posterior distribution to obtain point estimates $\mu = (\mu_1, \mu_2)$, $\sigma^2 = (\sigma_1^2, \sigma_2^2)$ and $p = (p_1, p_2)$, with $p_1 + p_2 = 1$. The fitted pseudo-coding distribution is

$$P(r; \mu, \sigma^2, p) = p_1\, PoiLN(r; \mu_1, \sigma_1^2) + p_2\, PoiLN(r; \mu_2, \sigma_2^2), \tag{6.8}$$

where $PoiLN(r; \mu_1, \sigma_1^2)$ is a Poisson Log-Normal probability function for parameters μ_1 and σ_1^2.

For a new gene coming from pseudo-coding mixture with length r_0, the probability of belonging to the first distribution $PoiLN(r; \mu_1, \sigma_1)$ follows from Bayes rule

$$P(1; r_0, \mu, \sigma^2) = \frac{p_1\, PoiLN(r_0; \mu_1, \sigma_1^2)}{p_1\, PoiLN(r_0; \mu_1, \sigma_1^2) + p_2\, PoiLN(r_0; \mu_2, \sigma_2^2)} \tag{6.9}$$

with the similar equation being valid for $P(2; r_0, \mu, \sigma^2)$. Then the gene is classified as pseudogene or coding gene based on the following expression

$$\begin{aligned} \text{biotype} &= \begin{cases} \text{pseudogene}, & P(1; r_0, \mu, \sigma^2) > P(2; r_0, \mu, \sigma^2) \\ \text{coding}, & \text{otherwise} \end{cases} \\ &= \begin{cases} \text{pseudogene}, & p_1 \, PoiLN(r_0; \mu_1, \sigma_1^2) > p_2 \, PoiLN(r_0; \mu_2, \sigma_2^2) \\ \text{coding}, & \text{otherwise.} \end{cases} \end{aligned} \tag{6.10}$$

The model is tested for human and mouse genome. We used a subsample ($n = 3\,000$) from the mixed distribution of pseudocoding and protein-coding gene lengths and fitted it with Poisson Log-Normal mixture with $K = 2$. The remaining genes which weren't used for fitting are used to assess performance of the model. The same division into training and test set is used to assess the performance of the logarithmic regression in predicting the gene biotype based on its length.

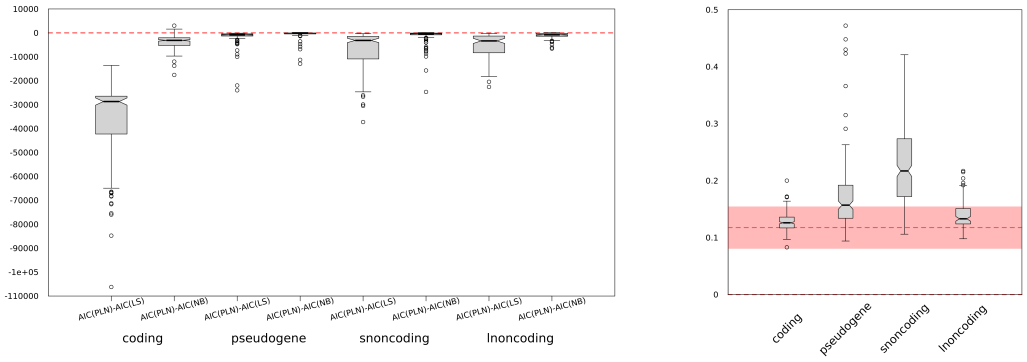
Chapter 7

Results and discussion

7.1 Gene length distribution is Poisson Log-Normal across all biotypes

The gene length distributions for 276 metazoan species were obtained as detailed in Materials and methods section. The genes were divided in four biotype groups: coding, pseudogenes, short noncoding and long noncoding. For every species, we separatelly fitted the gene length distribution for different biotypes. This way we observed the gene biotypes as different subcommunities of the genome which underwent separate evolution. For every biotype, we tested three evolutionary hypotheses fitting the gene length distribution with the Log-Series, the Negative Binomial and the Poisson Log-Normal distribution. According to the Akaike Information Criterion (AIC) [35], the Poisson Log-Normal model outperformed both the Negative Binomial and Log-Series for all biotypes (Figure 7.1-a). Specifically, for protein-coding genes the Poisson Log-Normal was selected over Negative Binomial in 97.45% species, and in 89.54%, 96.23% and 98.13% of species for pseudogenes, long noncoding and short non coding genes, respectively. Moreover, it was selected over Log-Series in all species for every biotype. The results imply that the Gompertzian density regulation $g(x) = \gamma \ln(x + \epsilon)$ is a better choice than the alternative $g(x) = \eta x$. Because of the extreme lengths of certain genes, these results are expected since logarithmic density regulation has lesser effect on the restrictive evolution of the long genes than the alternative linear function.

Kolmogorov-Smirnov (KS) statistic was calculated to assess the goodness of fit of the Poisson Log-Normal model. The KS statistic between two empirical distributions depends on the number of observations and in case of small number of observations its empirical minimum is significantly larger than the theoretical zero. Thus we fixed the number of observations when calculating KS statistics for different biotypes and we estimates the empirical minimum for the fixed number of observations. Obtained mean KS statistics for every species are represented by boxplots



(a) AIC differences.

(b) KS statistic.

Figure 7.1: **(a)** Distribution of the difference between the AIC obtained with the Poisson Log-Normal model (PLN) and the Log-Series (LS) or the Negative Binomial (NB) model, separately calculated for protein-coding genes, pseudogenes, short noncoding genes and long noncoding genes. **(b)** Distribution of the mean KS statistic between gene length distribution and the Poisson Log-Normal fit, separately calculated for different gene biotypes. The dashed red line together with its standard deviation represents the value of KS statistic between two random samples coming from the same distribution and can be thought of as the empirical minimum of the KS values represented in the boxplot.

(Figure 7.1-b), separately for every biotype. We observe that Poisson Log-Normal fits both protein-coding genes and long noncoding genes well. The fit of pseudogenes is not as good, probably due to small number of annotated pseudogenes in several species. Namely, only 787.4 pseudogenes are annotated on average, compared with 20026.9 protein-coding, 2272.3 short noncoding and 2878.8 long noncoding genes (not taking into account species with less than 100 genes which were discarded before fitting). Finally, short noncoding genes are not very well fitted by the Poisson Log-Normal model. Since by definition short non coding genes have length smaller than 200 nucleotides, their gene length distribution is right-truncated. Since the 200 nucletides is human-made threshold between short and long non-coding genes, from the population dynamics approach it may make more sense to group all noncoding genes together.

7.2 Protein-coding gene length distribution shows evolutionary trend in metazoan species

The length of a gene generally increases in evolutionary time [49]. We estimate the evolutionary dynamics of gene length with two parameters of Poisson Log-Normal

distribution, μ and σ. Engen and Lande's model [4] specifies that $\mu = r/\gamma$ and $\sigma^2 = \sigma_e^2/2\gamma$, where r is a growth rate, γ is a multiplicative constant from Gompertzian function and σ_e^2 is an environmental variance. To assess our results from the evolutionary perspective, we plot all species as points in (μ, σ)-space, where μ and σ^2 are parameters obtained fitting the protein-coding genes with Poisson Log-Normal distribution (Figure 7.2). Protein-coding genes were chosen over other biotypes because of their coding ability. Natural selection makes sequences of protein-coding genes more preserved in related organisms than the other sequences in the genome [48]. Moreover, protein-coding genes show the best agreement with Poisson Log-Normal model across all biotypes (Figure 7.1-b).

Different taxonomic classes shown in Figure 7.2 are not clearly separated by the μ and σ, but nevertheless the evolutionary trend is evident. More complex species such as mammals (*Mammalia*), birds (*Aves*) and reptiles (*Reptilia*) have the largest μ and σ values and roundworms (*Chromadorea*) have the smallest μ and σ values. Invertebrates (*Arachnida*, *Insecta*, *Branchiopoda*, *Gastropoda*) show high variation in parameters, but with μ smaller than in more complex species. Lastly, fish (*Actinopteri*) have μ and σ values between those obtained for invertebrates and those obtained for mammals, birds and reptiles.

The μ parameter has values in range $\langle 6.5, 10.5 \rangle$ which signifies high ratio between growth rate r and density regulation constant γ. This is expected since gene species have high abundance of nucleotides. Namely, taking into account the assumption that a gene enters the community at the abundance 1, high growth rate is required for the gene to reach the length of tens of thousands of nucleotides. Manipulating expressions for μ and σ, we get $\mu = 2r\sigma^2/\sigma_e^2$ which explains observed positive correlation between the μ and σ (Figure 7.2).

7.3 Multimodal gene length distribution corresponds to different biotypes

Because of their intrinsic connection, we now observe subcommunities of different gene biotypes together as one community. Namely, as discussed in the introduction, evolutionary paths of different gene biotypes are intertwined. Moreover, short and long noncoding genes are arbitrary separated with the threshold of 200 nucleotides and it makes sense to observe them together. To model the joint gene length distribution for all biotypes, we use the generalization of Engen and Lande's model in which different biotypes correspond to diverse subcommunities. The resulting gene length distribution is thus a mixture of Poisson Log-Normal variables or Negative Binomial variables, depending on the chosen density regulation function.

In the process of genome annotation, not all gene biotypes are equally covered. When an assembly of a genome is published, it is just a first version of the assembly which requires a lot of additional work and multiple revisions. In fact, it can be

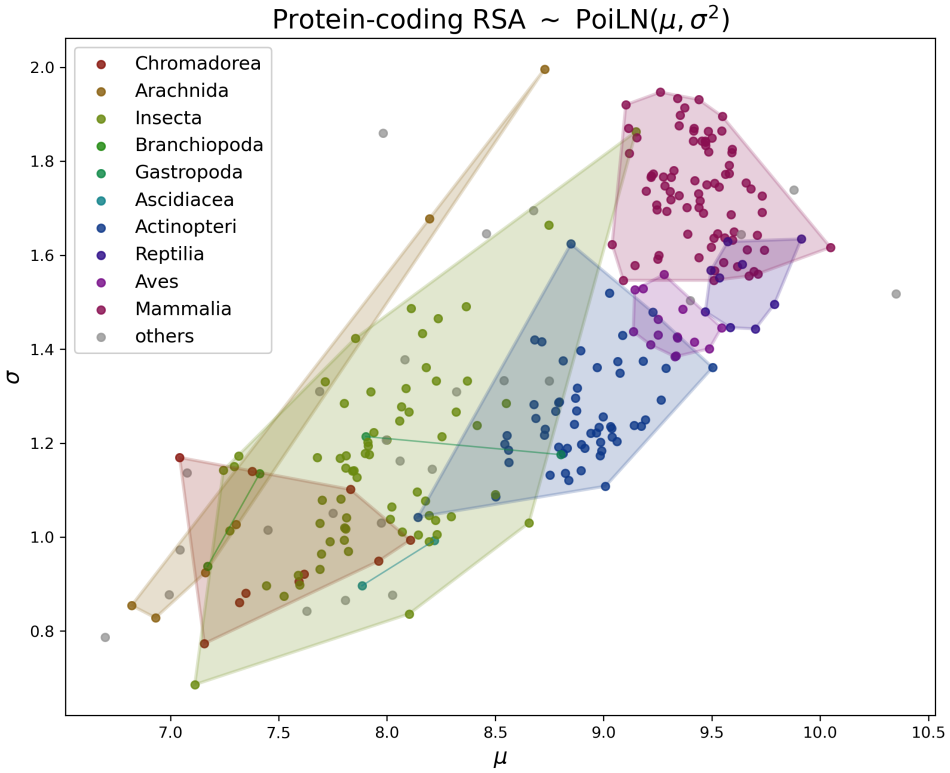


Figure 7.2: Scatter plot of Poisson Log-Normal parameters μ versus σ obtained fitting the protein-coding gene length distribution. Different colors represent different taxonomic classes, as indicated in the legend. Taxonomic classes to which only one species belongs are grouped together in the class *others*. In total, 274 species are represented.

said that the genome is never fully annotated since it is impossible to capture all the information written in its sequence [48]. As expected, protein-coding genes are usually of the biggest interest and thus best annotated. Since in this section we wish to model gene length dynamics for all biotypes, we have to observe a genome which has good coverage across all biotypes. Human genome as well as the genomes of model organisms are constantly being updated and are thus well annotated in comparison to others. Here we observe the genomes of human (*Homo sapiens*), mouse (*Mus musculus*) and zebrafish (*Danio rerio*).

Human gene length distribution is trimodal (Figure 7.3-a) with the modes corresponding to small noncoding genes, pseudogenes and protein-coding genes. The long noncoding genes lie in-between the pseudo and coding genes, but in the join mixture they don't contribute to the separate mode. Mouse gene lengths exhibit similar pattern as in human (Figure 7.3-c), but the long noncoding genes almost completely overlap with protein-coding genes. The distribution for zebrafish is bimodal (Figure 7.3-e), due to almost nonexistent pseudogenes. Long noncoding genes

again overlap in length with protein-coding genes.

It is known that the number of pseudogenes is uneven in different metazoan genomes and that mammals have high number of pseudogenes [54]. The observed bimodal/trimodal distributions are consequence of these differences in the genome compositions. While it is obvious why short noncoding genes correspond to the left mode, it may be less apparent why are pseudogenes shorter than protein-coding genes. By their origin, pseudogenes can be processed or unprocessed. Processed pseudogene is a product of the reverse transcripton of the mRNA and is thus lacking introns which explains its short length. Unprocessed pseudogene arises when the mutation disrupts the coding ability of a duplicated gene. The reason behind the short length of an unprocessed gene lies in the correlation between duplication patterns and gene length. Duplications proved to be increasingly incomplete for longer genes and thus recently duplicated genes are shorter than the average gene [49]. Consequently, the unprocessed pseudogenes will be shorter in length.

We fitted human and mouse gene length distribution with a mixture of three Poisson Log-Normal (3-PoiLN) variables and with a mixture of three Negative Binomial (3-NB) variables. The zebrafish gene length was fitted with the same type of mixtures of two variables (2-PoiLN and 2-NB). The distribution fitting was performed using Bayesian inference. Both mixtures converged (PoiLN traceplots shown in Figure S2.1-b,d,f), but the Poisson Log-Normal mixture outperformed the Negative Binomial. From the comparison of cumulative distribution functions (Figure 7.3-b,d,f) we notice that Poisson Log-Normal fits the data very well, while Negative Binomial shows divergence in tails. For human and mouse, Negative Binomial misses the first mode which corresponds to short noncoding genes (S2.1-b,d,f). In all three cases, the right tail of Negative Binomial is too short and doesn't fit the long gene lengths well (Figure S2.1-a,c,e). This is a consequence of the stronger density regulation in Negative Binomial model which overregulates the extremely long genes.

7.4 Differentiation between pseudogenes and protein-coding genes based on the length

The similarity between pseudogenes and their functional paralogs often results in misannotation of pseudogenes as protein-coding genes by computational genome annotation systems [66]. Wrongly annotated pseudogenes are problematic for any subsequent analysis which relies on the correct biotype annotation such as the design of locus-specific microarrays [66]. Moreover, with new discoveries, the importance of pseudogenes is being recognized [55] and they aren't anymore simply discarded as "junk" DNA. Thus the differentiation between pseudogenes and protein-coding genes is an important task. Consequently, new computational methods are being developed for the task of pseudo-coding gene differentiation since the alternative

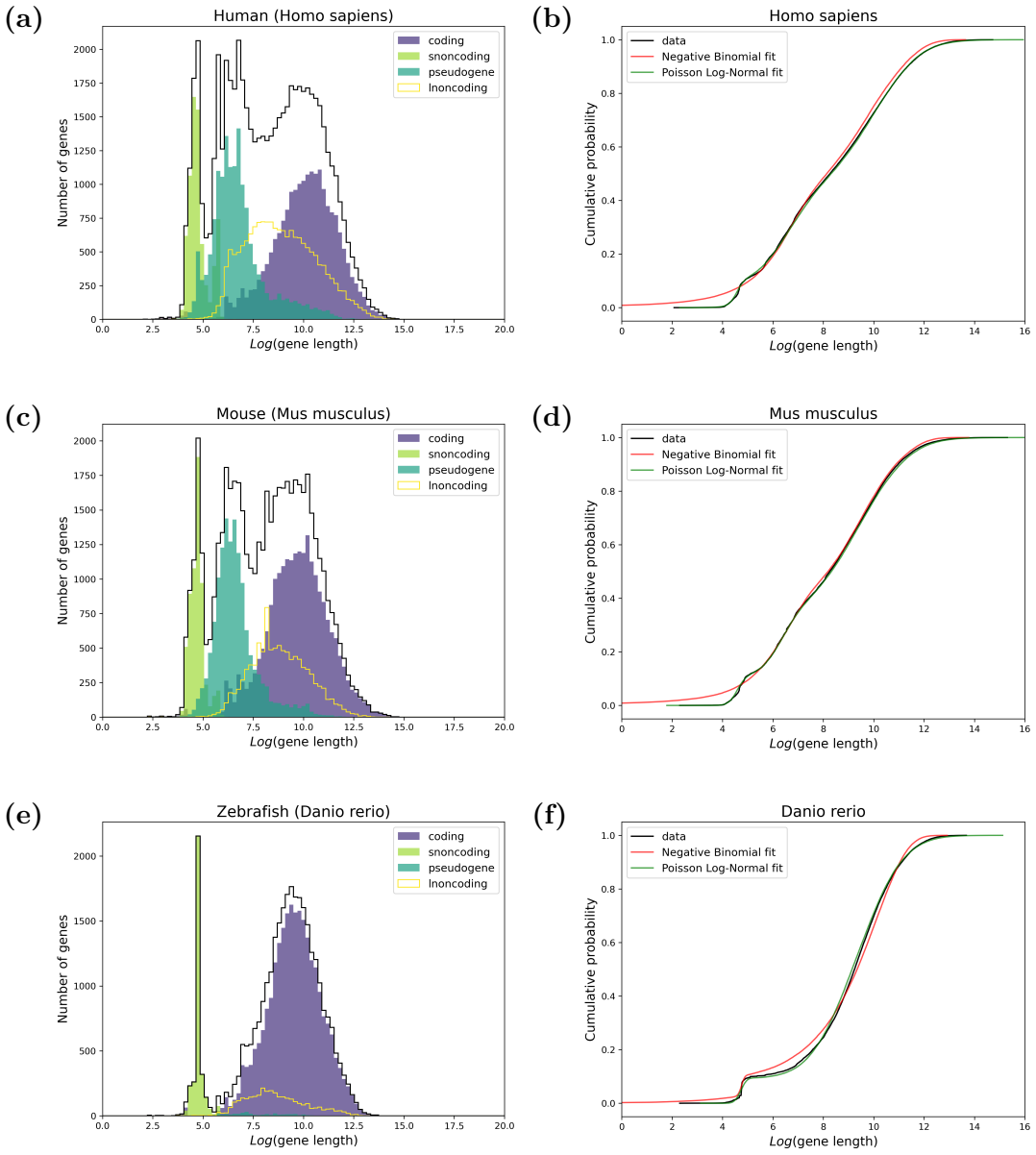


Figure 7.3: **(a)** Gene length distribution of different gene biotypes in *Homo sapiens*. **(b)** Comparison of the Negative Binomial and Poisson Log-Normal fit for the gene length distribution in *Homo sapiens* using empirical cumulative distribution functions (CDFs). A fitting distribution is calculated as a sample from posterior predictive distribution. **(c-d)** Distribution of gene length for different biotypes and CDF fit comparison for *Mus musculus*. **(e-f)** Distribution of gene length for different biotypes and CDF fit comparison for *Danio rerio*.

manual curation is a time-consuming process. These methods are generally based on the alignments of the sequence of a putative pseudogene. As an example, Yao *et al.* developed the method which uses transcript-based and protein-based sequence alignments to infer different types of pseudogenes. The *PPFINDER* method is another alignment-based method which identifies the processed pseudogenes searching for their parental gene and comparing it to the putative pseudogene [67]. Another method for the identification of processed pseudogenes *PseudoDomain* based on protein domain classification was developed as an alternative to other alignment-based methods which are problematic in case of a poorly annotated genome [68]. Here we propose a new pseudo-coding differentiation method based solely on the length of a gene. While this method by itself is not accurate enough to decide about the gene's coding ability, it may be incorporated as the first step of another more complicated method for pseudogene identification.

For species with high number of pseudogenes, such as human and mouse, the gene length distribution is trimodal (Figure 7.3-a,c). Classifying genes into different biotype groups, we notice that the three modes of distribution correspond to small noncoding genes, pseudogenes and protein-coding genes, from left to right. If instead, we observe only pseudogenes and protein-coding genes we get the bimodal distribution of gene length with the pseudogenes corresponding to the left mode.

A random subsample from the pseudo-coding gene length distribution may be used to estimate the parameters of the Poisson Log-Normal mixture of two variables. For a gene coming from the pseudo-coding mixture, we can now easily calculate the probability of it belonging to the first or to the second variable from the mixture. If the former probability is greater then the latter, the gene is classified as pseudogene. Otherwise, it is classified as protein-coding gene. Neglecting the non coding RNA genes, the pseudo-coding community has two distinctive subcommunities and the estimated mixture parameters represent differences in the gene length dynamics between the two subcommunities.

To test our pseudo-coding prediction model, we used human and mouse genome and compared our PoiLN model with the logistic regression. The 3000 genes were used as the training set. With the remaining genes, we tested the performance of both PoiLN method and logistic regression. Since both human and mouse are well annotated organisms and in lack of a better approach, the Ensembl biotype classification of a gene was considered to be the true biotype.

The results for human can be found in Table 7.1. Compared to logistic regression, the PoiLN method shows better balance between the precision and recall statistics in both the protein-coding and pseudogene prediction and has higher accuracy (0.882 compared to 0.82). Similar results can be found for mouse, with even higher difference in accuracy (0.888 compared to 0.767, Table S2.3). The precision of PoiLN method is 0.894 for protein-coding prediction and 0.872 for pseudogene prediction (taken as average between human and mouse). Overall, PoiLN method outperforms the logistic regression for both human and mouse.

Important difference between the two methods is the fact that logistic regression is a supervised method which depends on the accurate annotation of genes in the training set while the PoiLN method is unsupervised. This makes the PoiLN method applicable for genomes with poorly characterized biotypes. However, the genome still needs to have high enough number of located pseudo-coding genes for the accurate Poisson Log-Normal mixture fitting. If we use the PoiLN method just as the first step of a more advanced pseudo-coding classification method, then instead of classification of genes as pseudogenes or as coding genes, one can report the probability of each biotype which can be used in the future steps of the method.

Table 7.1: Contingency table for coding-pseudogene prediction methods for *Homo sapiens*. Two methods are compared, PoiLN method and logistic regression. **(i)** Prediction of protein-coding genes. PoiLN method has precision of 0.891 with recall of 0.906. Logistic regression has precision of 0.926 with recall of 0.745. **(ii)** Prediction of pseudogenes. PoiLN method has precision of 0.87 with recall of 0.85. Logistic regression has precision of 0.728 with recall of 0.92. **(iii)** Accuracy of PoiLN method is 0.882 compared with logistic regression accuracy of 0.82.

	PoiLN method		logistic regression		
true biotype	coding	pseudo	coding	pseudo	total
coding	16957	1768	13941	4784	18725
pseudo	2085	11810	1116	12779	13895

7.5 Gene length distribution for protein-coding genes and pseudogenes in mammals

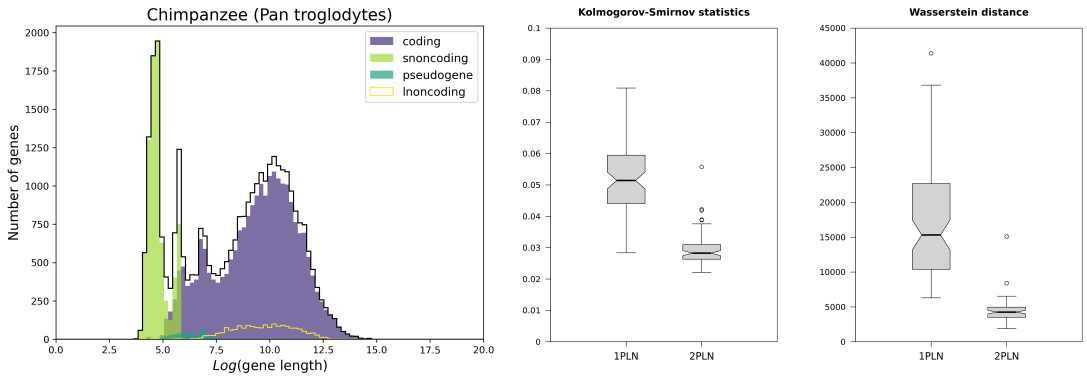
Well annotated mammalian genomes have similar number of pseudogenes and protein-coding genes [53]. As example, human genome has 20410 protein-coding genes and 15210 pseudogenes. On the other hand, some mammalian genomes have very small number of annotated pseudogenes. The genome of chimpanzee (*Pan troglodytes*) has 23521 protein-coding genes and only 485 pseudogenes. The chimpanzee and human diverged from the common ancestry only 6 million years ago and consequently their genomes show high similarity with nucleotide sequence identity within the coding DNA greater than 98.5% [48]. Thus the extremely small number of chimpanzee pseudogenes is likely caused by incomplete annotation and possibly also by wrong annotations of pseudogenes as coding genes. To strengthen the argument, mouse has 22506 coding genes and 13656 pseudogenes which is much more similar in number to human despite their earlier evolutionary divergence.

Comparing the gene length distribution of chimpanzee (Figure 7.4-a) to the

one for human (Figure 7.3-a), we notice that the distribution in chimpanzee is also roughly trimodal, but with the middle mode corresponding to protein-coding genes. Thus we hypothesize that at least some of the pseudogenes in chimpanzee are wrongly annotated as protein-coding.

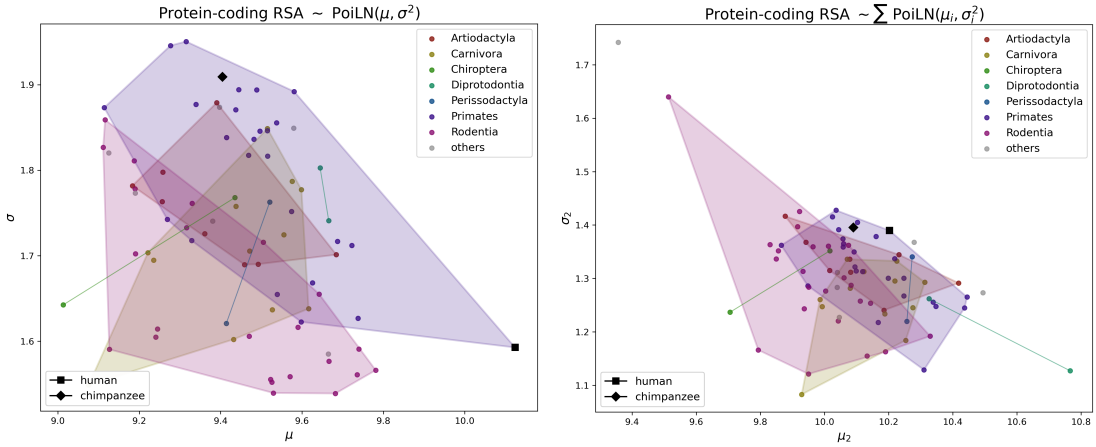
We observed that the Poisson Log-Normal parameters obtained for protein-coding genes follow the evolutionary trend in metazoan genomes (Figure 7.2). For mammals, being abundant in pseudogenes, we wish to test whether inclusion of both protein-coding genes and pseudogenes will result in even more evident evolutionary patterns. For every mammalian species, we fitted protein-coding genes with Poisson Log-Normal and the pseudo-coding mixture with the mixture of two Poisson Log-Normal variables. If all mammals were to have similar gene length distribution as human and mouse, the parameters μ_2 and σ_2 for the second variable in the Poisson Log-Normal mixture ($\mu_2 > \mu_1$) would correspond to protein coding genes and would be similar to the parameters μ and σ of the single Poisson Log-Normal protein-coding distribution. On the other hand, for gene length distribution similar to chimpanzee, μ_2 and σ_2 would in fact be better representatives of true protein-coding genes.

To assess the goodness of fit for both the Poisson Log-Normal (1PLN) fit of coding genes and Poisson Log-Normal 2-variable mixture (2PLN) for pseudo-coding genes, we calculated Kolmogorov-Smirnov (KS) statistic and Wasserstein distance between the data and the fit (Figure 7.4-b,c). The Poisson Log-Normal mixture fits the pseudo-coding distribution better than the single Poisson Log-Normal fits only the coding genes. However, this could have been expected simply because of the greater number of parameters. Nevertheless, small values of both KS statistic



(a) Chimpanzee gene length. **(b)** KS statistic. **(c)** Wasserstein dist.

Figure 7.4: **(a)** Gene length distribution of different gene biotypes in *Pan troglodytes*. **(b)** Distribution of the mean KS statistic for different mammalian species. Two models are compared, fitting protein-coding genes with Poisson Log-Normal (1PLN) and fitting pseudogenes and protein-coding genes with mixture of 2 Poisson Log-Normal distributions (2PLN). **(c)** Distribution of the mean Wasserstein distance for different mammalian species with the same models as for KS statistic.



(a) (μ, σ)-plot for PoiLN model.

(b) (μ_2, σ_2)-plot for 2-PoiLN model.

Figure 7.5: Scatter plots of Poisson Log-Normal parameters for 81 mammalian species. Different colors represent different taxonomic orders, as indicated in the legend. Taxonomic orders to which only one species belongs are grouped together in the order *others*. Human (*Homo sapiens*) and chimpanzee (*Pan troglodytes*) are additionally marked with different symbols. **(a)** μ and σ are obtained fitting gene length distribution of protein-coding genes with Poisson Log-Normal. **(b)** μ_2 and σ_2 are obtained fitting gene length distribution of protein-coding genes and pseudogenes with Poisson Log-Normal 2-mixture and they represent the parameters of the second part of the mixture ($\mu_2 > \mu_1$).

and Wasserstein distance suggest that Poisson Log-Normal mixture fits the pseudo-coding lengths very well.

To compare our results with taxonomic classification of mammalian species, we plotted them in both (μ, σ)-space based on Poisson Log-Normal fit of protein-coding genes (Figure 7.5-a) and in (μ_2, σ_2)-space based on Poisson Log-Normal mixture fit of pseudo-coding genes (Figure 7.5-b). The distance between human and chimpanzee indeed decreased in (μ_2, σ_2)-space as expected if our hypothesis of wrongly annotated genes is correct. However, from the visual inspection of plots, the separation of taxonomic orders is not improved for (μ_2, σ_2) compared to (μ, σ).

The ratios between intracluster and intercluster distances for different parameter spaces and for both taxonomic order and taxonomic family are shown in Table 7.2. For random clustering, the expected ratio between intracluster and intercluster distances is equal to 1 so in all cases species show clustering patterns in concordance with taxonomy, albeit not very strong. At the order level the best clustering is achieved in (μ_2, σ_2)-space, while at family level the best clustering is achieved for (μ, σ). For well annotated species, that is those species which have total number of pseudogenes and coding genes greater than 20000, the clustering is in better agreement with taxonomy. This was expected since for the poorly annotated species the gene length distribution is not the true representation of the RSA distribution

based on Engen and Lande's model. In case of well annotated species, the best ratio is obtained for taxonomic order classification in the $(\mu_1, \sigma_1, \mu_2, \sigma_2)$-space. The corresponding dendrogram shows that some taxonomic orders, such as Primates, are better clustered than the others (Figure S2.2).

The gene length distribution for multiple mammalian species seems to be biased because of the poor genome annotation and thus the parameters of the Poisson Log-Normal distribution are just rough approximations of the equations $\mu = r/\gamma$ and $\sigma^2 = \sigma^2 e/2\gamma$, where r, μ and σ_e^2 are parameters from the population dynamics model for gene length. While this approximation is close enough to identify evolutionary trend at higher taxonomic ranks (Figure 7.2), at lower taxonomic levels evolutionary trend is hardly visible. The separation of species at different taxonomic ranks is still better than in the completely random model (as shown by the ratio of intracluster and intercluster distance), but not good enough for accurate taxonomy-based clustering of mammals.

Table 7.2: Ratio between mean intracluster distances and mean intercluster distances for different clustering solutions. The left table is calculated for all 81 mammalian species. The right table is calculated for 61 mammalian species which had at least 20000 pseudo-coding genes. The distances are calculated in three different Euclidean spaces: 1PLN=(μ, σ), 2PLN$_2$ =(μ_2, σ_2) and 2PLN=$(\mu_1, \sigma_1, \mu_2, \sigma_2)$. The clusters are defined as taxonomic orders or as taxonomic families.

	mammals			mammals ($> 2 \times 10^4$ genes)		
	1PLN	2PLN$_2$	2PLN	1PLN	2PLN$_2$	2PLN
Order	0.90	0.87	0.89	0.84	0.74	0.71
Family	0.79	0.9	0.81	0.74	0.89	0.8

7.6 The extremely abundant multigene family of olfactory receptors violates the assumption of the Poisson Log-Normal model

Olfactory receptor (OR) genes are responsible for detection of odors in the environment. The olfaction is essential for survival of many mammals and OR genes consitute the largest multigene family in mammals characterized by frequent gene gains and losses [69]. Due to the difference in lifestyle, the OR gene number varies greatly among species with an extreme example of African elephant which has ∼2000 coding OR genes and > 2200 OR pseudogenes. [69]

We found several mammalian species whose gene length distribution had the

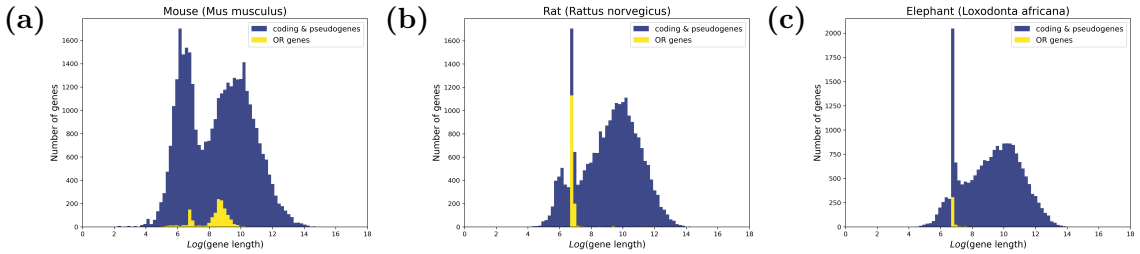


Figure 7.6: Gene length distribution for protein-coding genes and pseudogenes, with olfactory receptor (OR) genes in yellow, for three species: **(a)** mouse (*Mus musculus*), **(b)** rat (*Rattus norvegicus*), **(c)** elephant (*Loxodonta africana).*

extreme peak at the gene length of $\sim e^7$. This peak proved to correspond to the OR genes (both protein-coding genes and pseudogenes). For rat (*Rattus norvegicus*), based on the Ensembl gene description, we found 1 333 coding OR genes and 48 OR pseudogenes. The distribution of these genes perfectly matched with the described peak (Figure 7.6-b). On the other hand, 1 198 coding OR genes and 226 OR pseudogenes were found in mouse, but their lengths were much more diverse (Figure 7.6-a). The OR gene numbers for mouse are in agreement with findings from Niimuri *et al.* [69], since they reported 1 130 coding and 236 pseudo OR genes. For the rat, they found 1 207 coding, and 508 pseudo OR genes (with additional 52 unclassified pseudo/coding OR genes). The discrepancy in numbers may be due to the wrongly annotated pseudogenes in the Ensembl database and some yet unrecognized OR genes. The difference between the "blue" and "yellow" peak in Figure 7.6-b may be partially due to these missing OR genes and partially due to the other genes of the same length.

African elephant (*Loxodonta africana*) has extreme peak at the similar position as the rat (Figure 7.6-c), yet Ensembl describes only 351 coding genes as OR genes and none of the pseudogenes. This is in contradiction with the results from Namurii et al. [69] where they identified extremely expanded OR gene family in the African elephant. Thus we believe that the extreme "blue" peak in the figure 7.6-c belongs to the OR genes, despite the missing Ensembl description.

We are not sure if the extreme difference between gene length distribution of the OR genes in mouse and rat can be caused by sequencing or annotation errors. However, the OR peaks such as those found in rat and elephant cannot be fitted with the Poisson Log-Normal distribution. In the Poisson Log-Normal, as well as Negative Binomial, model we assumed that new genes enter the community in times specified by Poisson process. The extremely abundant gene families, such as OR genes in rat and elephant, are violating this assumption since too many genes enter the genome community in too short time. Thus for the successful Poisson Log-Normal fitting of such species we would have to remove the OR genes. Since many OR genes have missing description, we can't systematically remove them. However, we keep in mind that the μ and σ parameters for these species are not well-estimated.

Chapter 8

Conclusion

The population dynamics model developed by Engen and Lande has proved to be a good choice as the model for gene elongation. For all biotypes, the Poisson Log-Normal fit outperformed the alternative Negative Binomial and Log-Series fits in the majority of metazoan genomes. This suggests that the Gompertzian density regulation function is a better choice than the alternative linear regulation function. The results are in agreement with observed genes of extreme lengths since the Gompertizian function has weaker regulatory effect which allows for their emergence. While Poisson Log-Normal fit prevailed as the best of three models for all biotypes, for short non coding RNAs it showed certain divergence from the true gene length distribution. This is probably due to the definition of short noncoding RNAs which limits their length to 200 nucleotides causing the gene length distribution to be right-truncated. For protein-coding genes and long noncoding RNAs the fit was very good, while some metazoan genomes which had small number of annotated pseudogenes showed divergence also for this biotype group.

From the evolutionary perspective, separation of genes into different biotypes may not be the best approach since their evolutionary paths are intertwined. We derived the generalization of Engen and Lande's model for the community with diverse subcommunities. The assumption is that the subcommunities may have different dynamics and when observed as a single community the resulting RSA distribution is a mixture of Poisson Log-Normal or Negative Binomial distributions. The partition of the community into subcommunities doesn't include any additional assumption of independence since in the original Endgen and Lande's model species are assumed to be independent of each other. McGill *et al.* wrote about the multimodal RSA distribution which was reported in a few studies but without any theoretical derivation for the observed distribution [3]. The multiple modes of the RSA distribution are observable only on the log scale so the RSA distribution is still long-tailed. The subcommunities model can be applied for the distribution of gene length when we observe all biotypes together defining different biotypes as genomic subcommunities. The mixture of three Poisson Log-Normal variables was a good

fit for human and mouse gene length distribution. The three modes corresponded to short noncoding RNAs, pseudogenes and protein-coding genes. For the zebrafish genome, which has a small number of pseudogenes, the mixture of two Poisson Log-Normal variables was a good fit. Two modes corresponded to small noncoding RNAs and protein coding genes. This suggests that the generalization of Engen and Lande's model with the Gompertzian density regulation is good model for gene length distribution with subcommunities corresponding to different biotypes. The long noncoding RNAs didn't contribute to the separate mode. Because of their overlap with the protein-coding genes is it infeasible to infer their contribution to the mixture model. In fact, it is possible that there are even more subcommunities in the genome, such as different types of pseudogenes, which are grouped together in the gene length distribution.

We used the bimodality of gene length distribution for the pseudo-coding mixture to define the classification method. The method differentiates between the pseudogenes and protein coding genes based on their length and has accuracy of 0.882 in human and 0.888 in mouse. For this method to work it is necessary that enough pseudogenes and protein-coding genes are located on the genome so that their gene length distribution is bimodal. The advantage of the method is that it works in unsupervised manner and is thus not affected by wrongly annotated genes. Instead of the classification of the gene as coding, the method can also assign the coding probability to each gene so it can be used in pipeline with other more refined methods.

Estimated Poisson Log-Normal parameters μ and σ for protein-coding genes showed evolutionary trend for metazoan genomes. Different taxonomic classes were clustered together in (μ, σ)-space, with some overlap. The trend was lost at the lower taxonomic levels for mammalian genomes, thought it outperformed the random clustering. The parameters should reflect the evolutionary dynamics of the protein-coding genes, however some mammalian species have unusual gene length distributions probably caused by poor annotation. As an example, the gene length distribution of chimpanzee greatly differs from the one for human, despite their close evolutionary relations. Additionally, the extremely expanded multigene families, such as olfactory receptor genes in rat and elephant, violate assumptions of Engen and Lande's model and thus can't be fitted with the Poisson Log-Normal, nor the Negative Binomial, distribution.

The number of the Poisson Log-Normal distributions in the mixture was determined by visual inspections of the gene length distribution. A possible extension of our work is to implement the Dirichlet Process Mixture of Poisson Log-Normal variables. In case of the successful fitting, this model may even detect the hidden subcommunities in the mixture distribution which are not visible from the histogram of gene lengths.

Part III

Characterization of DNA methylation correlation structure in Down syndrome: study of chromosome 21

Chapter 9

Introduction

The genome of an organism contains coding sequences for all RNA and protein molecules produced in its cells. While the same DNA information is part of every cell, we witness great diversity among different cell types of the same organism. A cell generally expresses only a fraction of its genes, and observed variety of cell types is caused by different expression patterns. The gene expression can also be altered as response to the external signals. In theory the expression of a gene can be controlled at many steps from the transcription to the regulation of protein activity, however for most genes the initiation of RNA transcription is the most important point of regulation. The gene transcription can be regulated by specialized proteins or by the covalent modifications of the DNA sequence [15].

The heritable covalent modifications of the genome are part of epigenome inheritance. Epigenetics studies heritable alterations in a cell or organism's phenotype that do not involve changes in nucleotide sequence of DNA such as histone modifications. DNA methylation is another epigenetic mechanism which has potential to regulate gene expression. In the process of methylation a methyl groups is added to the cytosine (C) nucleotide which results in covalently altered 5-methylcytosine (5-mC). This alteration predominantly occurs on cytosines which are part of the CG dinucleotide. The CG nucleotide is often called CpG site where "p" indicates a phosphate linkage to distinguish it from a CG base pair. DNA methylation underlies the phenomenon of genomic imprinting which is observed in mammals, in which a gene is expressed depending on whether it was inherited from the mother or the father [15].

As a result of enzyme activities during the course of evolution, many CG dinucleotides disappeared from the genome and the remaining CG dinucleotides are unevenly distributed in the genome. The selected regions with high density of CG dinucleotides are called CpG islands (CPI). The human genome contains more than 20 000 CpG islands with the average length of 1 000 nucleotides. CpG islands often contain gene promoters and their methylation status has a potential to alter the expression of a gene. Thus the methylation of CpG islands is the main target of many methylation studies [15].

Different technologies are available for the interrogation of the methylation status of CpG sites. Their chemistry is based on methylation-specific enzyme digestion, affinity enrichment, chemical treatment with bisulphite (BS) or the combination of the mentioned techologies [8]. A popular cost-efficient techology for DNA methylation profiling is Illumina Infinium HumanMethylation450 BeadChip [6] which genotypes BS-treated DNA to reveal the ratio between methylated and unmethylated alleles [8]. The Illumina 450K microarray measures methylation levels of more than 480 000 CpG sites across the whole genome [6]. The measured Illumina 450K CpG sites were carefully selected and they cover 99% of RefSeq genes and 96% of CpG islands. In total 485 577 sites are interrogated, of which 482 421 CpG sites, 3 091 non-CpG sites ("ch" sites) and 65 random SNPs (rs probes). Additional 850 control probes are included in the microarray design. The non-CpG sites correspond to cytosines which may be methylated, but which are not part of the CG dinucleotide. The beta-values are calculated as the ratio between the intensity of the methylated allele and the total intensity (specifically $\beta = M/(M + U + 100)$, where M is intensity of the methylated allele and U is intensity of the unmethylated allele) [6]. The methylation studies, excluding the single-cell methylation studies, measure the methylation of population of cells. Thus the beta-value of 0.5 can be obtained in various scenarios, with the extremes when all alleles in all cells are 50% methylated or when 50% of cells are completely methylated while other half is completely unmethylated [8].

Methylation studies aim to infer the difference in methylation among different groups of samples. It is possible to inspect individual differentially methylated CpG sites in case of single-base resolution assays such as Infinium 450K using some form of the statistical test. However, differentially methylated regions (DMRs) are usually of greater interest because of the better prediction power. The methods for DMR detection can be designed for predefined CpG regions or can define their own CpG regions. CpG islands are a popular choice for methods based on the analysis of predefined regions. The methods which define novel regions usually find clusters of CpGs based on their chromosomal proximity testing for different criteria [8]. Regardless of the chosen method for differential methylation analysis, the great attention should be payed to the preprocessing steps [70].

Studies of methylation patterns reported the bimodal behaviour where different clusters of CpG sites tend to be either hypermethylated or hypomethylated, rarely existing in intermediate states [7]. The observed group behaviour suggests that CpG sites are not independent, but rather that the methylation profile is guided by the complicated network structure of CpG sites. In this work, we study the correlation structure of CpG sites. We observe the general trend of methylation correlation structure in the control data set with large sample size. Subsequently we compare it to the methylation correlation structure found for 29 Down syndrome patients, their unaffected siblings and their mothers. For computational reasons, we restricted our analysis on the chromosome 21 (HSA21)

Chapter 10

Materials and methods

10.1 Data retrival and description of data sets

Two DNA methylation data sets were downloaded from the Gene Expression Omnibus (GEO) data repository [71]: *control data set* (accession number GSE87571) and *Down syndrome data set* (accession number GSE52588). Both data sets contain measurements from Illumina Infinium HumanMethylation450 BeadChip [6] for the whole blood samples which were converted into beta-values.

The control data set has beta values for 728 healthy samples with age ranging from 14 to 94 years. The big sample size and the wide age span of the control data set were originally used to study the effects of aging on the human DNA methylome [72]. The female to male ratio of the data set is 388:340.

The Down syndrome (DS) data set has a family-based structure with measurements from a Down syndrome patient, one unaffected sibling and the mother for each of 29 families included in the study [73]. We observe these measurements as three separate data sets of 29 Down syndrome patients (DSP), 29 unaffected siblings (DSS) and 29 mothers (DSM). The age span of DSP, DSS and DSM is 10-43, 9-52 and 42-83 with the female to male ratio of 11:8, 22:7 and 29:0, respectively.

The description of CpG sites measured by Infinium 450K methylation array was found in Illumina "HumanMethylation450_15017482_v1-2.csv" manifest file (https://support.illumina.com/array/array_kits/infinium_humanmethylation450_beadchip_kit/downloads.html). We additionally downloaded the list of all UCSC RefSeq genes using UCSC Table Browser [74] and the list of all UCSC CpG islands using R package "AnnotationHub" v.2.20.2 [75]. R package "BSgenome.Hsapiens.UCSC.hg19" v.1.4.3. [76] was used to search the DNA sequence for the CG dinucleotides and "N" nucleotides which represent part of genome which is unsequenced. The Infinium 450K methylation array is based on *Homo sapiens* (human) genome assembly GRCh37 (hg19) [6] and hence the same assembly was used for all retrieved data.

10.2 Data preparation and preprocessing

Control and Down syndrome (DS) data sets with beta values were separately preprocessed following the same steps. We firstly checked for any CpG site which is not of "cg" or "ch" type and we excluded them together with the CpGs with missing values thus performing complete-case analysis. We then used the *meffil.estimate.cell.counts.from.betas* (with reference data set "blood gse35069 complete") function from R package "meffil" v.1.1.1 [77] to estimate the white blood cell counts for every sample. In the next step we removed the problematic CpGs. 60 466 CpG sites (file "hm450.hg19.manifest.tsv" file downloaded from `http://zwdzwd.github.io/InfiniumAnnotation#current`) were identified as problematic by Zhou *et al.* [78] and were recommended to be masked. Additionally, the multimodality of beta-values can indicate that the CpG site is in close proximity to SNP which is affecting the accuracy of methylation measurement [72]. We used the implementation of the DBSCAN (density-based spatial clustering of applications with noise) algorithm [79] from the R package "dbscan" v.1.1-5 [80] to check for bi-, tri- and multimodality of gender-specific beta-value distribution for every CpG in control data set. The same multimodal CpGs which were detected and removed from control data set were later removed from the DS data set. The M-values were calculated from beta values using the transformation formula $M = \log_2 [\beta/(1-\beta)]$. M-values were reported to have better performance for a statistical analysis of CpG methylation values [81]. The beta-values which were equal to zero, if present, were replaced with the smallest positive beta-value found in the data set prior to M-value conversion. There weren't any beta values equal to one, but they could have been similarly replaced with the maximum beta-value. Finally, we used the "regRCPqn" algorithm developed by Sala *et al.* [70] for the renormalization of already preprocessed beta-values. The algorithm aims to mitigate the data set-dependent effects of preprocessing and thus enables the joint analysis of different data sets.

The controls data set originally contained beta-values for 483 586 CpG sites. 59 598 CpGs (from 60 466 problematic CpGs) were masked and additional 91 multimodal CpGs were removed. During the "regRCPqn" renormalization, all sites from X and Y chromosome were removed and 414 046 CpGs remained. For DS data set 65 SNP probes (rs probes) and 23 437 CpGs with missing values were removed from the original 485 577 CpG sites. Additional 57 528+91 problematic sites were filtered. Finally the removal CpGs on X and Y chromosome left 394 792 CpGs.

For the following analysis, only the CpG sites on the human chromosome 21 (HSA21) were kept, 3 691 for control data set and 3 540 for DS data set with 3 536 overlapping CpGs. For control data set, the normalized M-values were fitted to linear regression model with age, gender and cell counts as independent variables for every CpG separately. The residuals were calculated as difference between observed and predicted M-values. For the Down syndrome data set, the residuals were obtained as the difference between the normalized M-values and the predicted M-values based on the linear regression model for controls. The 3 691 control residuals and 3 536

DS residuals were used in the subsequent analyses. Only the overlapping 3 536 CpG sites were used in comparisons between data sets.

10.3 Characterization of the correlation structure of CpG sites on HSA21

The correlation between two CpG sites was calculated as a Pearson correlation between the corresponding residuals. The Pearson correlation was chosen because of the approximately normal distribution of residuals. Moreover, the Pearson correlation allows us to use the t-test for the significance of correlation.

10.3.1 General characteristics of correlation

The correlation between 3 691 CpG sites was calculated for controls. To assess the stability of the calculated correlations, we randomly sampled with replacement 728 control subject for 100 times. In each iterations, we calculated correlation matrix between all 3 691 CpGs. From 100 correlation values, we calculated the mean (*mean bootstrap correlation*) and standard deviation (*standard deviation of bootstrap correlation*).

To identify the group of highly correlated CpGs, we performed the hierarchical clustering with the distance measure defined by $\text{dist} = 1 - \max(0, \text{cor})$ and the "average" method for clustering. The dendrogram was cut at the 0.5 height.

10.3.2 Partial correlation approach

98 CpG sites were chosen from the resulting hierarchical clusters in the following manner: 33+15 CpGs were part of two biggest clusters and additional 50 CpGs were picked from the singletons, clusters with only one member. The function *cor2pcor* from the R package 'corpcor' v.1.6.9 [82] was used to calculate the partial correlation matrix from the Pearson correlation matrix for the 98 chosen CpG sites.

10.3.3 Testing for significance of correlation

For 3536 CpG sites which remained after the preprocessing steps in both control and DS data sets, we calculated all 6 249 880 pairwise correlations. P-values for the two-tailed t-test for correlation, as well as Benjamini-Hochberg and Hochberg adjusted p-values, were calculated using the function *corr.test* from the R package "psych" v.2.0.9 [83]. The significance level of 0.05 was used.

10.3.4 Estimating the variability of correlation with respect to the small sample size

To assess the effect of the small sample size on the estimation of correlation, we randomly sampled 29 from 728 control subject for 100 times. In each iterations, we calculated correlation matrix between all 3 536 CpGs. From 100 correlation values, we calculated the mean (*mean subsampling correlation*) and standard deviation (standard deviation of subsampling correlation). The sample size of 29 was chosen based on the sample size of DS data sets.

10.4 Comparison of correlation structure between DS data sets and controls

For two correlation matrices, M_1 and M_2, we calculate the *correlation matrix distance* [84] with formula

$$d(M_1, M_2) = 1 - \frac{\mathrm{tr}(M_1 \cdot M_2)}{\|M_1\|_2 \|M_2\|_2}, \tag{10.1}$$

where tr is a trace operator and $\|.\|_2$ is Frobenius norm.

10.4.1 Small-variance correlations

We refer to those correlations whose standard deviation of subsampling was smaller than 0.11 as *small-variance correlations*. The corresponding CpG sites are thus called *small-variance CpGs*. The minimal standard deviation of subsampling for weak correlations ($|r| < 0.2$) was 0.1145 and thus the threshold of 0.11 is chosen to exclude the weak correlations. For the community detection based on the correlation matrix for small-variance CpGs, we used three different approaches:

- *Small-variance unweighted* method defines an unweighted network with only small-variance correlations as edges.
- *Positive correlation weighted* method defines a weighted network where positive correlations correspond to weights and negative correlations are neglected.
- *Positive correlation unweighted* method defines an unweighted network with all positive correlations as edges.

For the community detection, we used the *walktrap.community* function from the R package "igraph" v.1.2.6 [85].

10.4.2 CpG islands

We visually inspected the correlation structure for all CpG islands on HSA21 which have at least 3 CpG sites measured both in controls and DS data sets. For this purpose, we plotted next to each other beta-values of the measured CpG sites and correlation matrices for all data sets. Additionally, we calculated the correlation matrix distances between control, DSP, DSS and DSM correlation matrices. The plotted beta-values were obtained from the normalized M-values applying the transformation $\beta = 2^M/(2^M + 1)$. On beta-value plots, we additionally plotted the positions of all CG dinucleotides and the positions of exons. Exon positions were retrieved from the RefSeq list of genes based on "exonStarts" and "exonEnds" columns. The exons were included because the role of methylation in alternative splicing was reported [86], however the exact regulatorx mechanisms are not clear.

Chapter 11

Results and discussion

11.1 General characteristics of DNA methylation correlation structure on HSA21

Chromosome 21 (HSA21) is the smallest human chromosome with the total number of 380 444 CpG sites of which 4 205 are measured by Illumina 450K microarray. The unsequences DNA of HSA21 spans over two wide regions of nucleotides, 1−9 411 193 and 11 188 130−14 338 129, with the remaining unsequenced regions having the average length of 30 804 bp. The HSA21 coordinate plots, which we use to show the chromosomal positions of CpG sites, are covering the region from the first known CpG site (9 411 552$^{\text{nd}}$ nucleotide) till the last known CpG site (48 119 686$^{\text{th}}$ nucleotide) instead of the whole chromosome which excludes the first unsequenced region (Figure 11.1-d,e). The coverage of the Illumina 450K microarray is good for all sequenced genomic regions except for the additional "gap" before the centar of the chromosome. Since the choise of Illumina CpG sites was based on several criteria [6], such as gene and CpG island coverage, this region is probably not populated with important genomic sequences.

The correlation structure of 3 691 HSA21 CpG sites based on methylation measurements for controls was examined. Generally, the correlation between a pair of CpG sites decreases with the distance (Figure 11.1-a), however several pairs of CpGs show high correlation despite the huge base pair distance between them. The CpG correlation doesn't seem strictly dependant on the distance with many CpG islands whose CpG sites are almost completely uncorrelated (an example shown in Figure 11.2-a) and groups of highly correlated CpGs which are spread across the chromosome (Figure 11.2-b). Clustering CpGs into groups of highly correlated CpGs with the average intracluster correlation greater than 0.5, we found 155 clusters of 528 CpGs. The remaining CpG sites remained as single-member clusters. Generally, the larger clusters contained the CpGs from across the chromosome while the smaller clusters contained only few close CpGs (Figure S3.1).

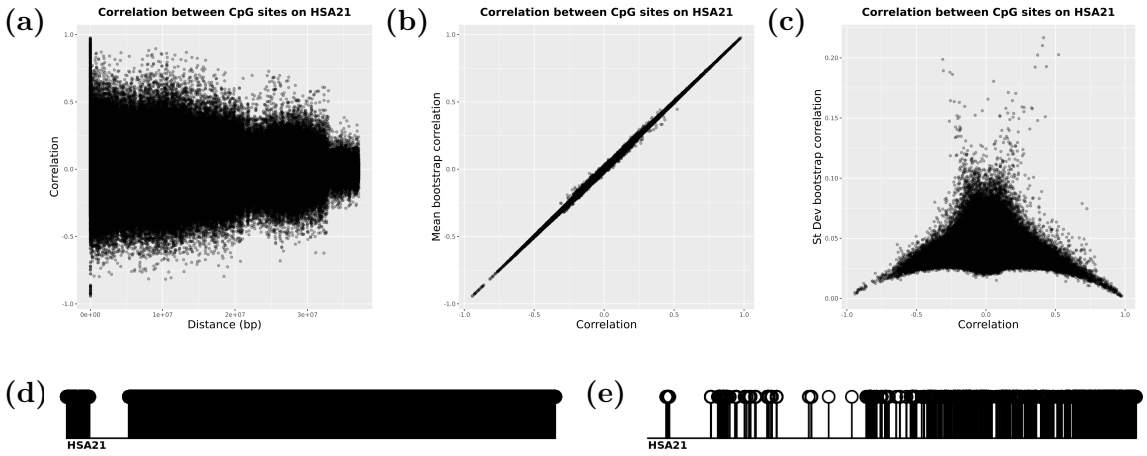
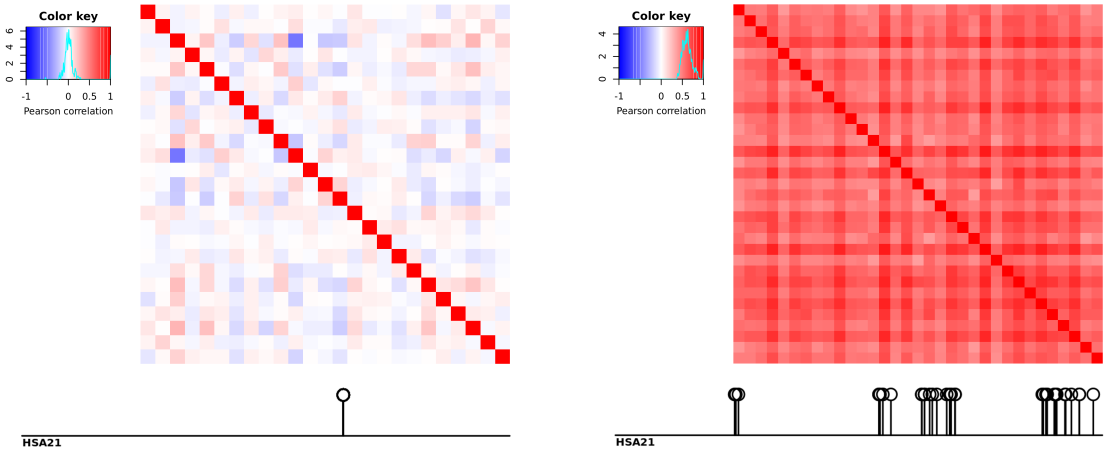


Figure 11.1: **(a)** Correlation versus distance for the pairs of CpG sites on HSA21. **(b)** Mean bootstrap correlation versus the correlation. The bootstrap mean is obtained from 100 repetitions of resampling with replacement. **(c)** Standard bootstrap deviation of correlation versus correlation. **(d)** HSA21 coordinates of all CpG sites, obtained from the sequence of HSA21. **(e)** HSA21 coordinates of all CpG sites measured by Illumina HumanMethylation450 BeadChip.

To assess the variability of CpG correlation, we performed random resampling with replacement of 728 control samples. The calculated mean bootstrap correlation value was highly similar to the correlation obtained for all 728 samples (Figure 11.1-b). The standard deviation of bootstrap correlation measurements was compared with the correlation obtained for all samples (Figure 11.1-c). The smaller values of correlation showed higher variability while the strong correlation coefficients were highly preserved.

11.2 Partial correlation ignores the pattern of highly correlated groups of CpG sites

Penalization methods are often employed for estimation of the inverse covariance matrix (precision matrix) in high-dimensional omics data sets. Two penalty functions are widely used, the ℓ_1 (lasso) penalty [87] which produces sparse precision matrices and the ℓ_2 (ridge) penalty [88] which shrinks the elements of precision matrix proportionally. With many available generalizations of these two methods, the choice of appropriate penalization method should be carefully considered. When adequate estimation of precision matrix is obtained, the partial correlation matrix is easily calculated from the estimate. The exact partial correlation can be obtained only if the cavoariance matrix is not singular which is not the case in high-dimensional setting and thus the penalization method is necessary.



(a) CpG island. **(b)** Group of highly correlated CpGs.

Figure 11.2: Correlation matrix and HSA21 coordinates for group of CpG sites. **(a)** 25 CpGs which belong to CpG island "chr21:34914303-34915906". Their HSA21 coordinates overlap in the figure. **(b)** Group of 33 highly correlated CpGs.

The partial correlation measures the correlation between two variables while removing the confounding effect of the remaining controlling variables. Prior to the choice of penalization method, we decided to assess whether the partial correlation is the right measurement of correlation between CpG sites. Control data set is high-dimensional ($3\,691 > 728$) and hence the exact partial correlation can't be calculated for all CpGs. However, for a smaller subset of CpGs the exact inverse covariance matrix and consequently partial correlation matrix can be calculated. For this purpose, 98 CpGs were chosen as detailed in the Materials and methods. Among the 98 CpGs, there were two groups of highly correlated CpGs while the remaining CpGs were uncorrelated to all other CpGs (Figure 11.3-a, left). The observed correlation structure of the 98 CpGs is chosen to represent different correlation patterns found on the whole chromosome.

Comparison between partial correlation and Pearson correlation for pairs of 98 CpGs (Figure 11.3-b) shows that partial correlation decreased both positive and negative correlations. The partial correlation of originally uncorrelated CpGs and of those which were strongly correlated is almost the same which makes them indistinguishable. In fact, the cluster structure of the chosen CpGs is completely lost when partial correlation is used instead of Pearson correlation (Figure 11.3-b, right). The observed over-shrinkage of partial correlation is caused by the presence of highly correlated groups of CpGs. These CpGs act as confounding variables when partial correlation is computed so the originally found correlation is removed together with the confounding effect. This makes the partial correlation a poor choice in presence of highly correlated clusters of variables. Since highly correlated clusters of CpGs are present on the HSA21, we decided not to use the partial correlation approach.

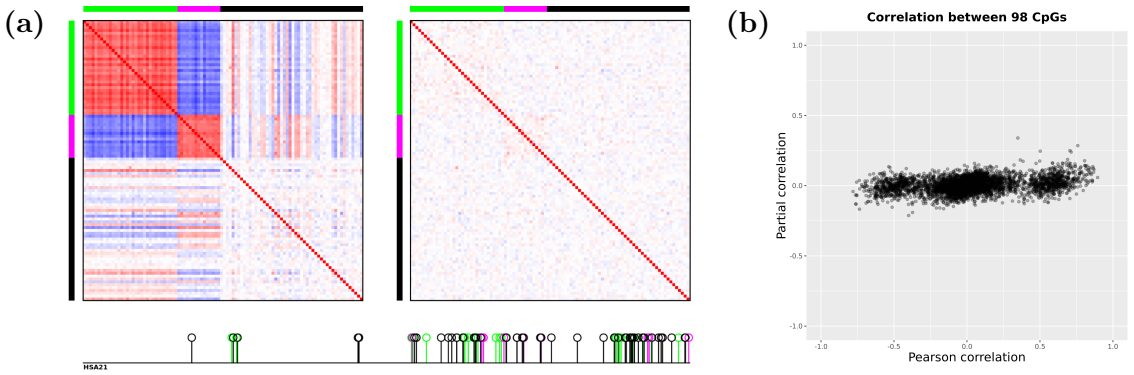


Figure 11.3: 98 CpG sites were used to compare partial correlation with Pearson correlation. **(a)** Pearson correlation matrix (*left*) and partial correlation matrix (*right*) for the same list of 98 CpGs. In respect to Pearson correlation, there were two clusters of highly correlated CpGs, of size 33 and 15, and 50 CpGs which weren't strongly correlated to other CpGs. The HSA21 coordinates of CpGs are shown belowe the heatmaps, colored by their cluster membership. **(b)** Scatter plot of partial correlation versus Pearson correlation for all pairs of the 98 CpGs.

11.3 Testing the significance of correlation

Chromosome 21 has 3 536 CpG sites which were measured in both control and Down syndrome data set and kept after the preprocessing steps. From the corresponding 6 249 880 pairwise correlations, we needed to identify those which are significant for the subsequent analysis. Statistically significant correlations may be inferred from the t-test which tests whether the correlation is significantly different from zero. The normality of residuals required by the test can be assumed, especially for the control data set which has large number of samples ($n = 728$). However, because high number of correlations were simultaneously tested the multiple comparison adjustment is necessary.

We used two different approaches to adjust for the high number of correlations in control data set. The Benjamini-Hochberg (BH) adjustment which controls the false discovery rate and the Hochberg adjustment which controls the family-wise error rate [89]. 1 320 802 correlations had BH-adjusted p-value less than 0.05 with the minimum absolute significant correlation being equal to 0.096. For the Hochberg adjustment 168 134 correlations were significantly different from zero with the significance level of 0.05 and the minimum absolute significant correlation of 0.212. Thus if we were to use BH adjustment all positive and negative correlations whose absolute value is greater than 0.096 would be significant with the similar conclusion about Hochberg adjustment and the 0.212 as the significant correlation threshold. The observed low significant values of correlation emerge because of the large sample size of control data set and thus we want to check if they are reproducible for the data set of smaller sample size such as one of the DS data sets ($n = 29$).

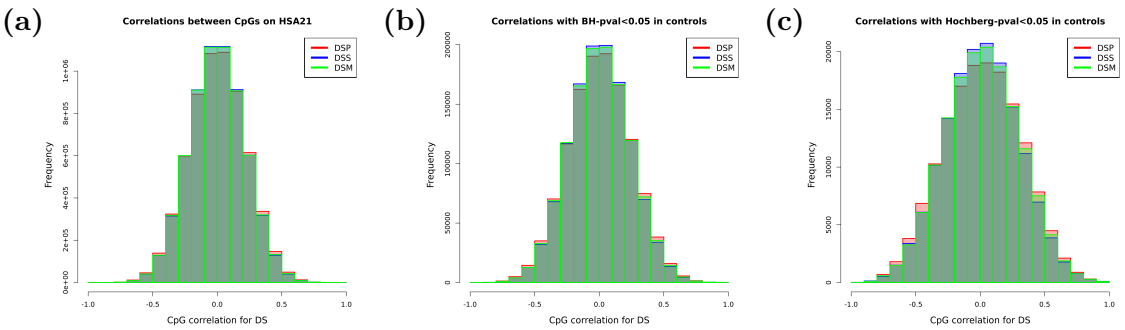


Figure 11.4: Distribution of correlation between CpG sites in DS data sets. **(a)** Correlations between all measured CpGs on HSA21. **(b)** The CpG correlations with Benjamini-Hochberg adjusted p-value smaller than 0.05. **(c)** The CpG correlations with Hochberg adjusted p-value smaller than 0.05

To compare the results with the DS data sets, we plot the distributions of CpG correlation in different DS data sets, DSP, DSS and DSM, under three condition. Firstly we compare the distributions for all correlations and then we compare them only for the correlations which are significant in respect to BH and Hochberg adjustment (Figure 11.4-a,b,c, respectively). We observe that Down syndrome distributions were similar in all three versions with the slight difference between DSP and the other two distributions. Namely, DSP distribution had lower number of low correlations ($|r| < 0.2$) and consequently higher number of strong correlations ($|r| > 0.5$) than DSS and DSM. The difference was more pronounced in the case of strong family-wise Hochberg adjustment. This suggests that the DSP correlations are in better agreement with control data set than correlations calculated for DSS or DSM. Namely, as the multiple comparison adjustment for controls becomes more restrictive, the corresponding DSP correlations increase in absolute value. However, the observed difference is too small to make a definitive conclusion.

11.4 Problem with the small sample size of the Down snydrome data set

The comparison between correlation values in control data set and in one of DS data sets shows great variability for all three DS data sets (Figure 11.5-a,b,c). In general, only correlations which are very high in controls ($|r| > 0.5$) are preserved in DS data sets. Comparing the correlations among the DS data sets we observe a similar pattern where only the high correlations seem to be constant, but the shape of correlation points in $[-1, 1]\times[-1, 1]$-space is different (Figure S3.2-a). Because of the larger sample size the correlation distribution in controls is narrower than in DS data sets so the observed difference in shapes is expected.

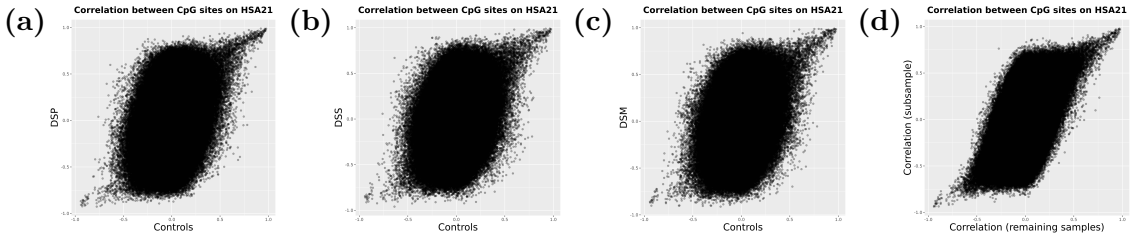


Figure 11.5: Comparison of correlation among different data sets for all measured CpG pairs on HSA21. **(a)** DSP vs. controls. **(b)** DSS vs. controls. **(c)** DSM vs. controls. **(d)** Random subsample of 29 controls versus the remaining 699 control samples.

Since the observed variability is present in all DS data sets, it can't be caused by Down syndrome condition. To test if the variability is caused by the small sample size of DS data sets or by the difference in the data collection and normalization which wasn't removed by preprocessing and renormalization, we randomly selected 29 control samples and compared them to the remaining 699 samples (Figure 11.5-d) and to the DS data sets (Figure S3.2-b). CpG correlation values for the random subsample again show great variability in respect to the correlations calculated for the larger (699) sample size, but the shape for all correlations is less dispersed than in case of DS data sets. This suggests that the major cause of the variability is in the small sample size, however some variability is still caused by the differences between control data set and Down syndrome data set. These differences which

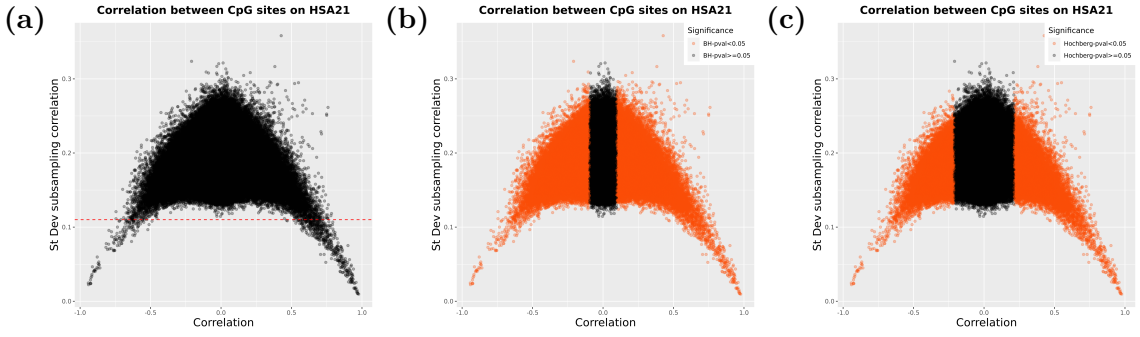


Figure 11.6: **(a)** Scatter plot of standard deviation of CpG correlation subsampling versus CpG correlation. The subsampling was repeated for 100 times, each time taking random subsample of 29 control samples from. The red horizontal line represent standard deviation of subsampling equal to 0.11. **(b)** The CpG correlations with Benjamini-Hochberg adjusted p-value smaller than 0.05 are plotted in *red*. **(c)** The CpG correlations with Hochberg adjusted p-value smaller than 0.05 are plotted in *red*.

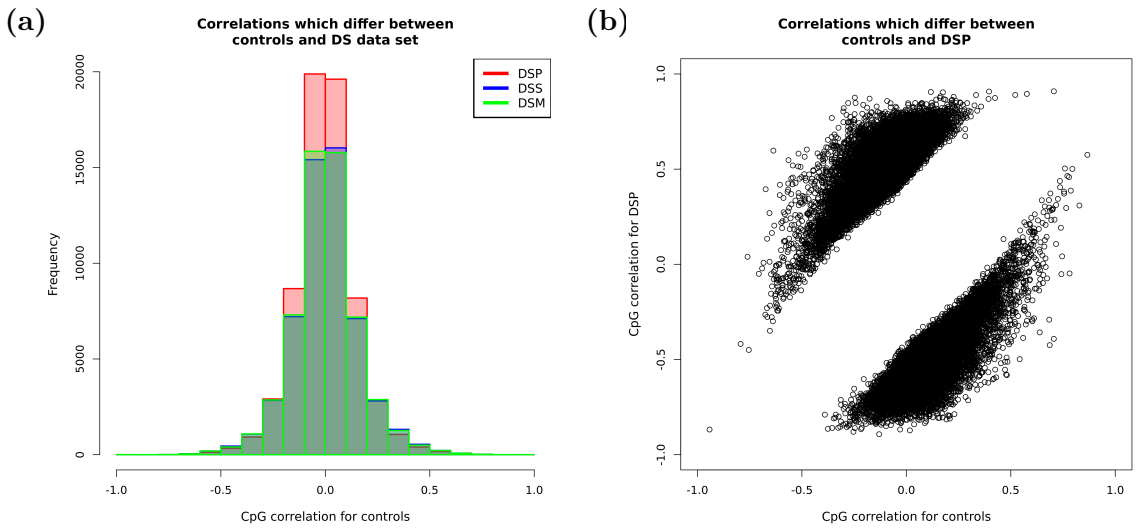


Figure 11.7: CpG correlations which differ from the corresponding correlations for controls by more than three bootstrap standard deviations. **(a)** Distribution of correlation in controls for the CpG correlations which differ between DS data sets and controls. **(b)** Correlation comparison between controls (*x-axis*) and DSP (*y-axis*) for the CpG correlations which differ between DSP and controls.

remained after preprocessing and renormalization may be caused by variability in data preparation or by unknown confounding variable that differs in the two studies.

With the aim to better understand the variability caused by small sample size ($n = 29$), we calculated the standard deviation of subsampling for every correlation between pairs of CpG sites (Figure 11.6-a). This confirmed our observation based on a single subsample that high correlations remain preserved even in small sample size. However, not all weak correlations has the same level of variability. It seems that sensitivity to the reduction in sample size is the property of a pair of CpG sites and is not simply dependant only on the correlation value. In fact, only highly correlated pairs of CpGs are consistently robust with respect to small sample size. Even the correlations which are significant after the BH or Hochberg adjustment show high variability in case of the small sample size (Figure 11.6-b,c). Thus, if we wish to compare DS data sets to the control data set we should use different approach which takes into account the variability caused by small sample size.

To identify only the meaningful differences in correlation between DS data sets and controls, we search for correlations which differ by at least three standard subsampling deviations. 65 235, 55 382 and 55 684 correlations which differ between controls and DSP, DSS and DSM, respectively, were detected. We were interested to know what is the value of these correlations (Figure 11.7-a). It seems they are roughly normally distributed around the zero which is in agreement with the observed pattern of variability (Figure 11.6-a). However, the DSP have higher number

of differentially correlated pairs of CpGs than DSS and DSM. Moreover, the DSP correlations which differ from controls are mostly those correlations which were uncorrelated or weakly correlated in controls. This suggests that for Down syndrome patients, new correlations are formed between CpG sites. This is again observed in Figure 11.7-b which is a subfigure of Figure 11.5-a. The difference in number of differentially correlated CpG pairs of ∼10 000 between DSP and DSS/DSM implies slightly different shape between correlation plots (Figure 11.5-a and Figure 11.5-b,c). However, this difference is hardly noticeable from the plots.

11.5 Correlation structure for the robust correlations

The correlation noise caused by the small sample size of DS data sets makes the comparison of correlation structure between DS and controls difficult. In fact, such comparison at the whole chromosome level could lead to unreliable conclusions. Thus we reduced our analysis only to those correlations which are robust with respect to small sample size. We defined robust correlations as those correlations whose standard deviation of subsampling was smaller than 0.11 (Figure 11.6-a, red line). The value of 0.11 was chosen because we didn't want to include robust correlations with values close to zero, that is robustly uncorrelated pairs of CpGs. 588 correlation, 457 positive and 131 negative, satisfied the criteria. In the following text, we refer to them as *small-variance correlations*. The corresponding 284 CpG

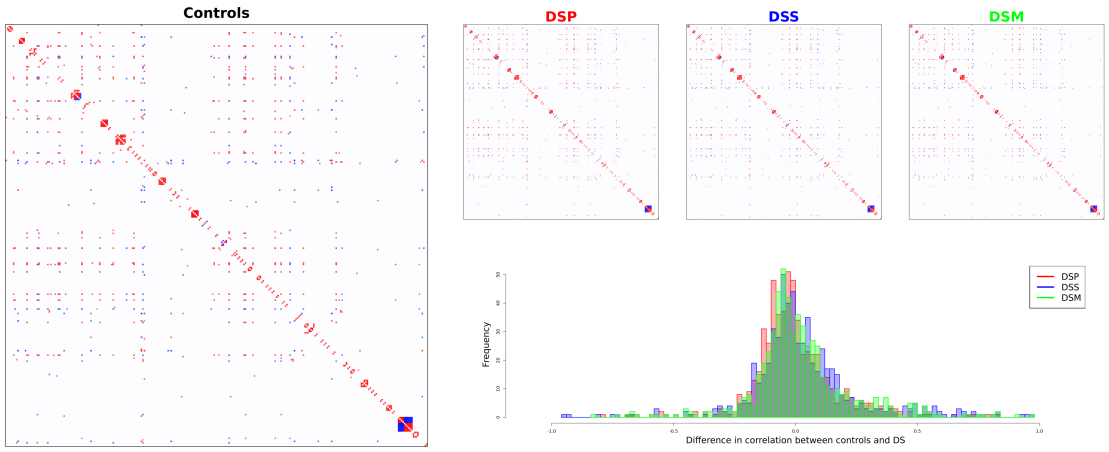


Figure 11.8: Correlation matrices of 284 small-variance CpG sites, calculated for controls, DSP, DSS and DSM. The CpG sites are ordered by their location on HSA21 and only small-variance correlations are shown with negative correlations in *blue* and positive correlations in *red*. Histogram shows distribution of correlation differences between controls and the DS data sets, for shown small-variance correlations.

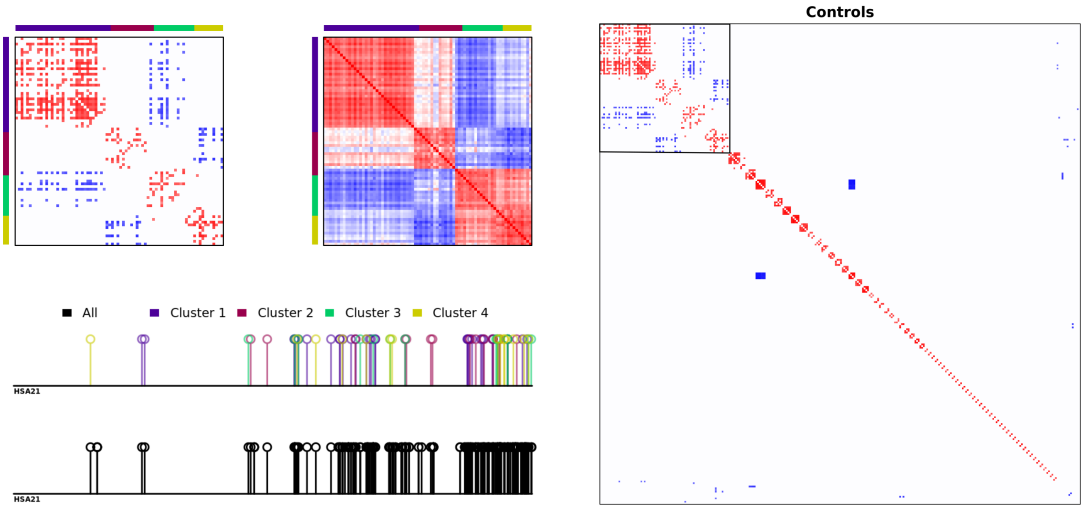


Figure 11.9: Correlation matrix of 284 small-variance CpG sites, calculated for controls, is shown on the right. The order of CpGs is based on cluster membership for *small-variance unweighted* clustering. 76 CpGs from the first four clusters are reshown in separate smaller correlation matrices. The left smaller matrix contains only small-variance correlations, while the right smaller matrix contains all correlations. Below the matrices, the HSA21 coordinates of all 284 CpGs and the coordinates of CpGs in clusters are plotted.

sites are called *small-variance CpG sites*.

The correlation structure for the small-variance correlations was investigated in DS data sets and compared to the one found in controls (Figure 11.8). We observe high similarity of correlation structure among all data sets. The differences in small-variance correlations between controls and DS data sets are similarly distributed for all DS data sets (Figure 11.8, histogram). However, small difference may be noticed for DSP which show the highest similarity to controls. In fact, if we calculate the sum of element-wise distances between the sparse matrices of small-variance correlations, we get the distances of 71.146, 89.503 and 82.979 for DSP, DSS and DSM, respectively.

The small-variance CpG sites in Figure 11.8 are ordered by their chromosomal position. Thus small clusters on the diagonal correspond to the CpG sites which are in proximity to each other on HSA21. The other correlations are "scattered" across the matrix. To gain a better insight into the observed correlation structure, we wish to cluster the CpG sites into groups of highly correlated CpGs. Using the *small-variance unweighted* method of clustering, we found 85 CpG communities (Figure 11.9). The clustering uncovered an interesting correlation structure of 284 small-variance CpGs. Majority of clusters are "isolated" from the other clusters and some are strongly negatively correlated to other clusters. Especially interesting is the group of four bigger clusters of CpGs. Among them, the first cluster is negatively

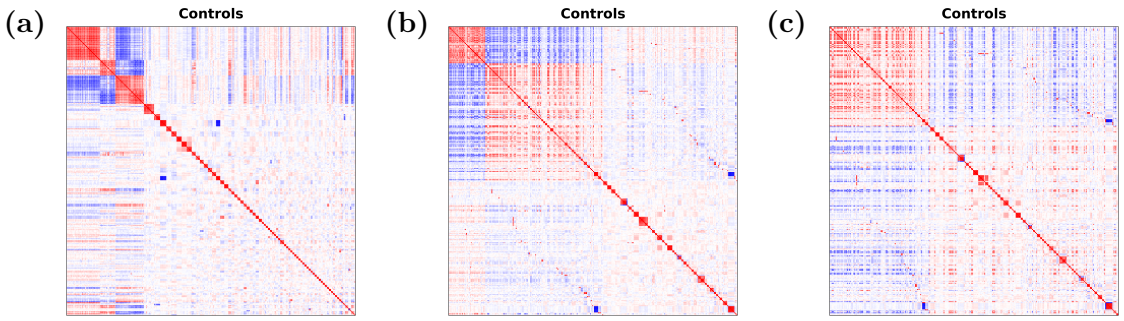


Figure 11.10: Correlation matrices of 284 small-variance CpG sites for controls with different orders of CpGs. The order of CpGs is based on cluster membership for **(a)** *small-variance unweighted* clustering, **(b)** *positive correlation weighted* clustering, **(c)** *positive correlation unweighted* clustering.

correlated to the third and the second is negatively correlated to the fourth (Figure 11.9, left zoomed matrix). However, if we show all correlations among the CpGs in these four groups, it seems that the third and fourth clusters form a single cluster (Figure 11.9, right zoomed matrix). The HSA21 coordinates for CpGs which belong to the four clusters are scattered across the chromosome and don't show any special pattern (Figure 11.9, left bottom).

The full correlation matrix of all 284 CpGs ordered with respect to membership to 85 found clusters shows interesting "stripe" pattern (Figure 11.10-a). This suggests that some of the smaller cluster may in fact be part of the first four clusters. Thus we try other clustering methods. The *positive correlation weighted* method finds three big clusters of CpGs (Figure 11.10-b). The first two clusters are groups of moderately to strongly correlated CpGs and the two groups are mutually negatively correlated. While the third group may be added to the first based on correlation pattern, correlation between its CpGs and the other CpGs is too weak. Finally, the *positive correlation unweighted* method finds 61 cluster of CpGs (Figure 11.10-c), but from the sorted correlation matrix we may see that these smaller clusters are in fact part of two big weakly correlated clusters.

If we visualize correlation matrices for DS data sets with the same order of CpG sites (Figure S3.3), we notice that only the small-variance correlations are consistently preserved in the DS data sets. To assess the difference between correlation structure in DS data sets and in controls, we calculate the distances between the correlation matrices for 284 small-variance CpGs. Since the distance between the correlation matrices is invariant to the order of variables, we may use any of the full correlation matrices. The distance between controls and DSP, DSS and DSM is 0.449, 0.504 and 0.502, respectively. To better understand the meaning of these values, we computed the distribution of correlation matrix distances for different data sets (Figure 11.11). The results suggest that for a random subset of CpG sites, the

distance between controls and DS data sets is approximately the same. However, for the small-variance CpGs the correlation structure of controls is surprisingely better preserved in DSP than in DSS and DSM.

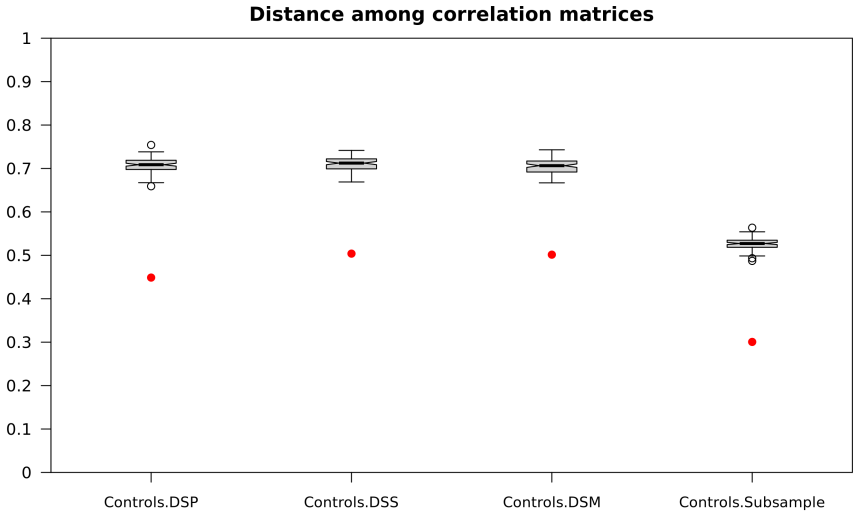


Figure 11.11: Distribution of distance between CpG-correlation matrices. Distances were calculated between controls and different data sets: DSP, DSS, DSM and controls subsample. The red points are distances obtained for 284 small-variance CpGs. The boxplots show distances calculated for 284 randomly chosen CpGs, with 100 repetitions. The controls subsample consists of 29 randomly chosen control samples. The subsampling was repeated 100 times, and the mean distances are reported in boxplots and as the red point.

11.6 Correlation structure of the CpG island

There are 365 CpG islands on the chromosome 21 of which 325 are covered by Illumina 450K microarray. We visually inspected all 234 CpG islands which had at least 3 measured CpGs inside their region. Majority of them didn't show any evident difference between data sets. However, we found an interesting example where a single CpG expressed negative correlation with the other CpGs of the same island only in DSP (Figure 11.12-a). In general, as was the case at the chromosome level, strong correlations remained preserved in all data sets (Figure 11.12-b,c). While for the majority of CpG islands, the CpG sites are not strongly correlated (Figure 11.2-a), the cases with the strong positive or negative correlations are the most interesting for any subsequent analysis. To better understand the displayed plots we would need to add another layer of information such as the genomic characterization of nearby regions.

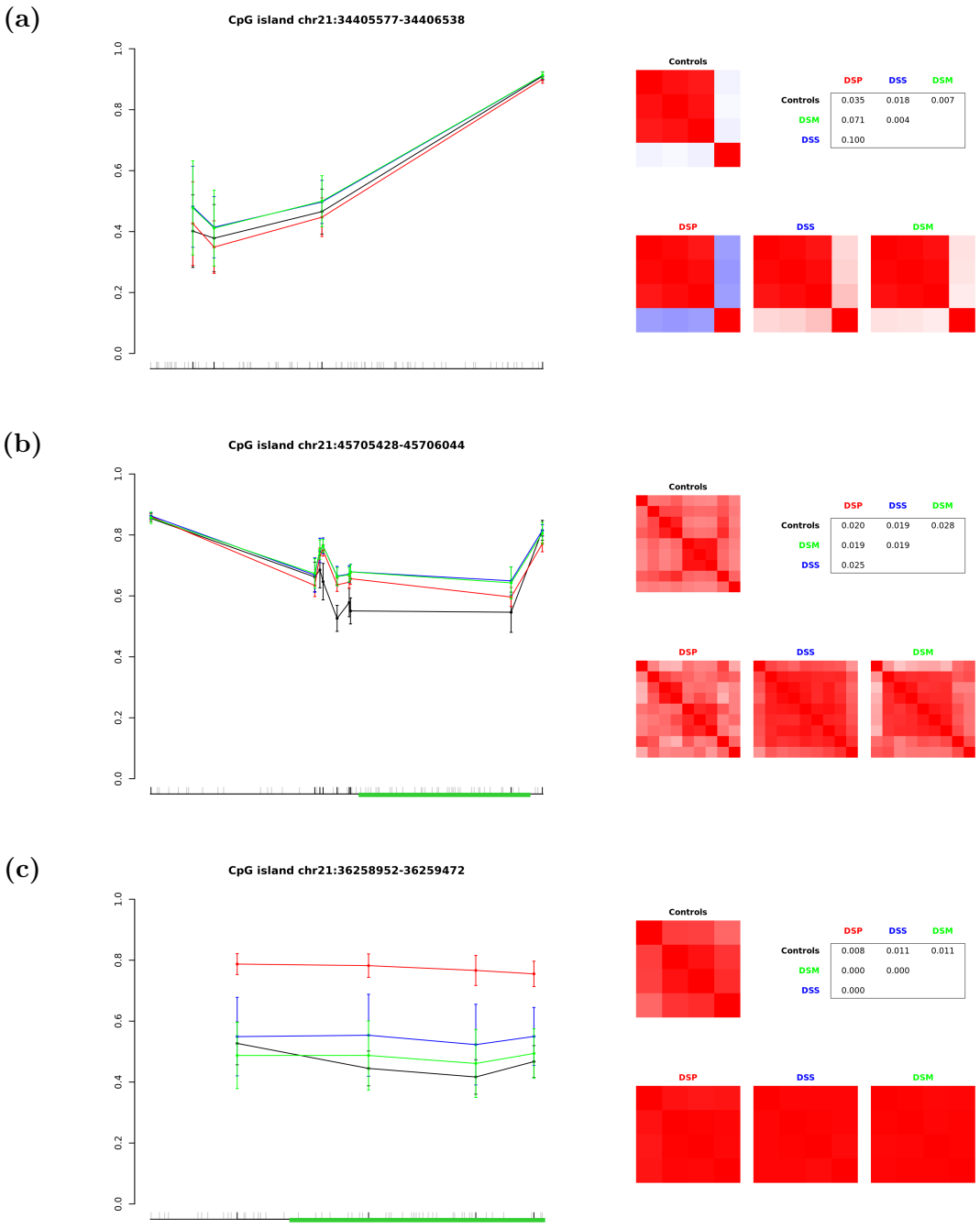


Figure 11.12: Left. beta values for different data sets: controls (*black*), DSP (red), DSS (*blue*) and DSM (*green*). Grey ticks on the *x-axis* are positions of all CG dinucleotides, while the green polygons represent positions of RefSeq-defined exons. **Right**. Correlation matrices for different data sets together with the correlation matrix distance between them. The plotted CpG island are **(a)** chr21:34405577-34406538, **(b)** chr21:45705425-45706044, **(c)** chr21:36258952-36259472.

Chapter 12

Conclusion

Comprehensive characterization of the methylation correlation structure proved to be a challenging problem. Nevertheless in this work we characterized the main aspects of correlation structure, explained some of the problems with the data and hopefully set the course for the future research. We focused on the chromosome 21 because of its connection to the Down syndrome condition and because of its small size which made computations faster. However, the conclusions which are based on the control data set can, with caution, be extended to other chromosomes.

For a pair of CpG sites, the relationship between their correlation and their chromosomal distance isn't straightforward. While for close CpGs, there is indeed a greater probability for strong correlation, this is not the universal rule. We observed many distant CpG pairs which were strongly correlated to each other as well as many CpG islands made of uncorrelated or weakly correlated CpGs.

We considered the partial correlation approach. Different penalization methods are frequently used in omics data sets to estimate the partial correlation matrix among the elements of the system. However, the methylomics greatly differs from the omics disciplines which are directly or indirectly based on the system of protein coding genes, such as genetics, proteomics or transcriptomics. At the first glance, the size of the system differs greatly with $\sim$20 000 protein-coding genes and more than 480 000 CpGs measured by Illumina 450K array. But while the problem of high dimension could be solved with more computational power, we believe that the bigger problem is the intrinsic difference between the systems. While the interactions between genes can be viewed as interactions between individuals, the observed group behaviour [7] of CpG sites suggests that they in fact function as groups and individually have very little biological power. Our findings are in agreement with this hypothesis. When we sorted the correlation matrix for small-variance CpG sites using different clustering algorithms, we noticed groups of CpGs which form on diagonal. Interestingly these groups behaved as single entities. Namely, when we observed correlations between the groups they had united pattern. Thus on the larger scale we in fact got the network of the groups of CpGs. This type of correlation

structure, based on our small-scale study, is not suited for the partial correlation approach because while accounting for all the"spurious" correlation, partial correlation shrinks all correlations to almost zero.

This work aimed to find the distinctive characteristics of the methylation correlations structure in Down syndrome. An interesting family-based design of the Down syndrome data set provided two controls data sets, DSS and DSM, for the DSP. However, the sample size of 29 subjects is not big enough to unravel the full complexity of the system of CpGs sites, even when restricted to chromosome 21. We verified this intuitive observation about the small sample size using the control data set. Randomly sampling from the control data set, we demonstrated the instability of correlations when sample size is equal to 29. In fact, only the strong correlations remained preserved.

Statistical test for the significance of correlation could potentially be used to find the meaningful patterns of correlation. However, when comparing two different data sets one of which has a small sample size, these test could lead to many false positive conclusions, even after the adjustment for multiple comparison. Thus we took different approach and analyzed only correlations which are robust with respect to the small sample size. However, because of their invariability, these correlations are highly conserved in different data sets and we may miss some important differences in correlation if we restrict our analysis only to them. The "right" approach is probably somewhere in the middle, as a trade-off between committing the type I error and the loss of power.

We did, however, manage to find some small differences between DSP and DSS/DSM. Our results suggest that in DSP, with respect to controls, new correlations are being created. On the other hand, for strong correlations, the correlation structure seems to be more similar between DSP and controls than between DSS/DSM and controls. This is a surprising discovery and we are not sure what could cause it, especially since the differences are small. It may in fact be the consequence of some latent behaviour of the system or merely a noise in the data.

For the future work it would be interesting to include other data sets, possibly of the larger size, and try to reproduce the observed behaviour of the system. Another possible extension would be to include a multi-omics data set which could offer a new layer of information.

Conclusion

In this thesis, we aimed to describe three different biological systems with the origin in genomics, proteomics and methylomics. The gene length distribution was successfully described as the relative species abundance (RSA) distribution of the population of genes where the dynamic process of gene elongation corresponds to the growth equation. Similarly, the distribution of protein domain abundances was described as the RSA distribution in the population of protein domains where new protein domains enter the community in time specified by Poisson process. For both systems, the population dynamics model developed by Engen and Lande proved to work well. This model includes both demographic and environmental noise as well as the density regulation function. Both distributions were best fitted with the Poisson Log-Normal distribution which emerges from the Gompertzian density regulation. Gompertzian function allows the occurrence of more abundant species in comparison with the alternative linear density regulation which yields the Negative Binomial model. Thus we conclude that the population of genes and the population of protein domains are best described when both demographic and environmental stochasticity are included in the model with the Gompertzian density regulation. Moreover, when besides the protein-coding genes, we include into the model non-coding RNAs and pseudogenes, we obtain the multimodal Poisson Log-Normal distribution. To describe this extended dynamic system, we generalized Engen and Lande's model for the case when diverse subcommunities are observed together.

RSA distribution is a measure of biodiversity of the community at the certain time point. Thus the gene length distribution and protein domain RSA distribution are measure of diversity at the certain point of evolution. In Part I, we defined the RSA phylogenetic method in bacteria which is based on the parameters of the protein domain RSA distribution. The method gave some interesting results at the intraspecies level of bacteria. In Part II, the description of metazoan species in terms of their gene length distribution followed an evolutionary trend. Moreover, the subcommunity parameter of the multimodal distribution was a good predictor for different gene biotypes.

In Part III, we focused on the methylation data and we aimed to characterize the methylation correlation structure in Down syndrome. For the system of CpG sites on the chromosome 21, we considered the pattern of their interactions. We observed that the CpG sites work in small highly correlated groups which can be

viewed as unique entities in the methylation network. Interestingly, the nodes of this network, that is, the groups of highly correlated CpGs were often scattered across the chromosome. Thus the approach of grouping together highly correlated CpG sites has different implications that the usual grouping of CpGs into CpG island.

From these three examples, we may conclude that an appropriate model for the biological system offers valuable insights into the behaviour of the system which may be missed if we merely observe elements of the system as separate independent biological entities. Moreover, if we wish to estimate dynamical processes of the complex system, including the stochasticity improves the resolution of the system as it accounts for all unobserved variables in the system.

Supplementary material

S1 Supplementary material of Part I

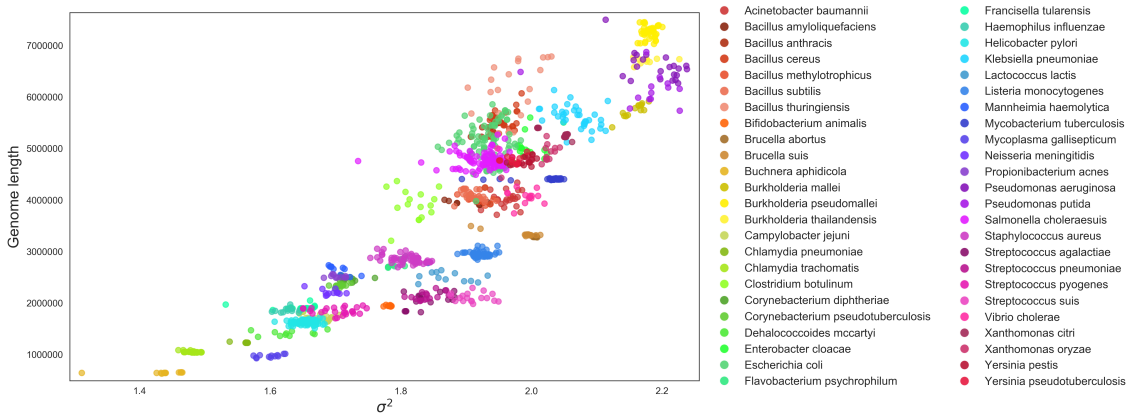


Figure S1.1: Scatter plot of bacterial genome length versus Poisson Log-Normal parameter σ^2 obtained fitting the protein domains RSAs. Figure shows only those species which are represented with at least 10 different strains in our data set. There are in total 1173 bacteria which belong to 48 different species. Different colors represent different species, as indicated in the legend.

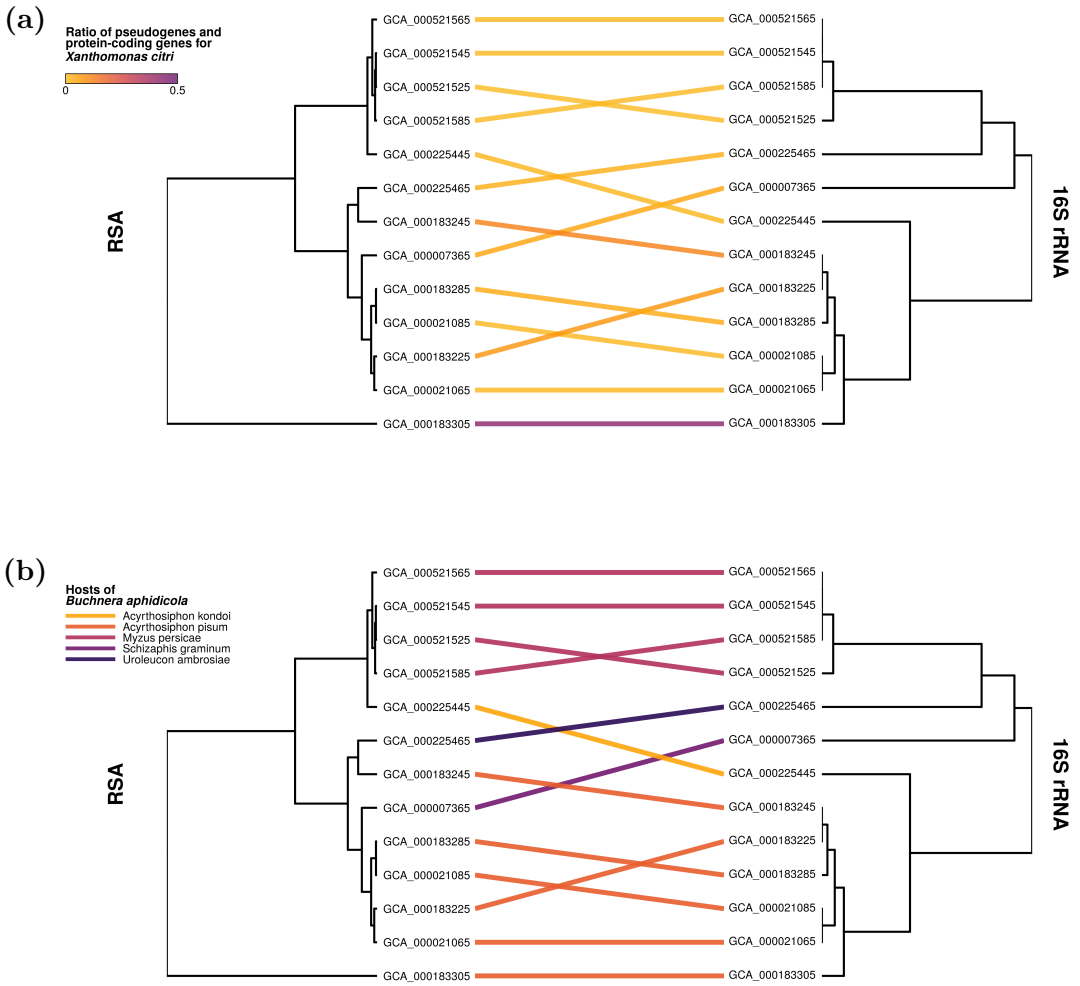


Figure S1.2: Hierarchical clustering of 13 strains of *Buchnera aphidicola*, by RSA method (*left*) and based on 16S rRNA gene (*right*). **(a)** Strains are colored by the ratio between number of pseudogenes and protein-coding genes. We notice that 12 stains have relatively low ratio while strain GCA_000183305 has ratio equal to 0.429. **(b)** 13 strains have different aphid hosts: *Acyrthosiphon kondoi* (*yellow*), *Acyrthosiphon pisum* (*orange*), *Myzus persicae* (*magenta*), *Schizaphis graminum* (*violet*) and *Uroleucon ambrosiae* (*navy*).

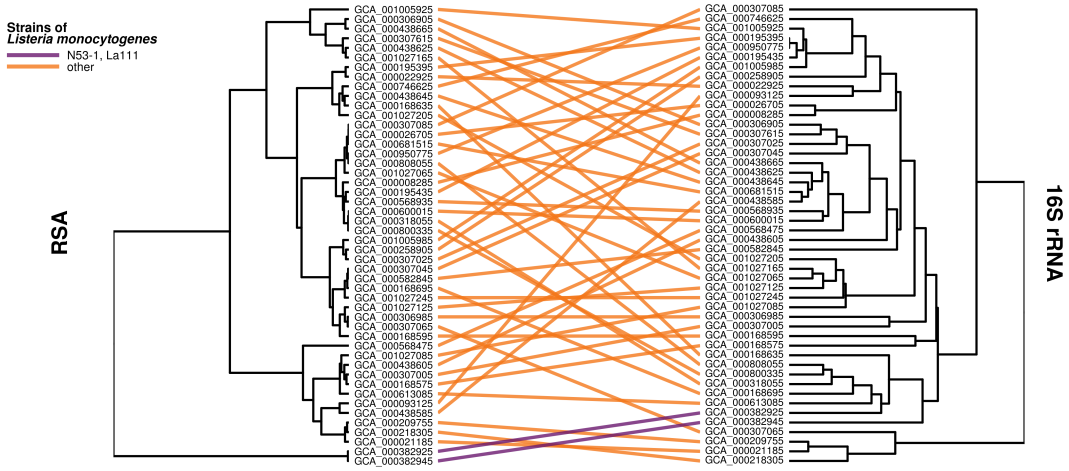


Figure S1.3: Hierarchical clustering of 48 strains of *Listeria monocytogenes*, by RSA method (*left*) and based on 16S rRNA gene (*right*). RSA method identifies a separate cluster of two strains: N53-1 (GCA_000382945) and La111 (GCA_000382925) (*violet*).

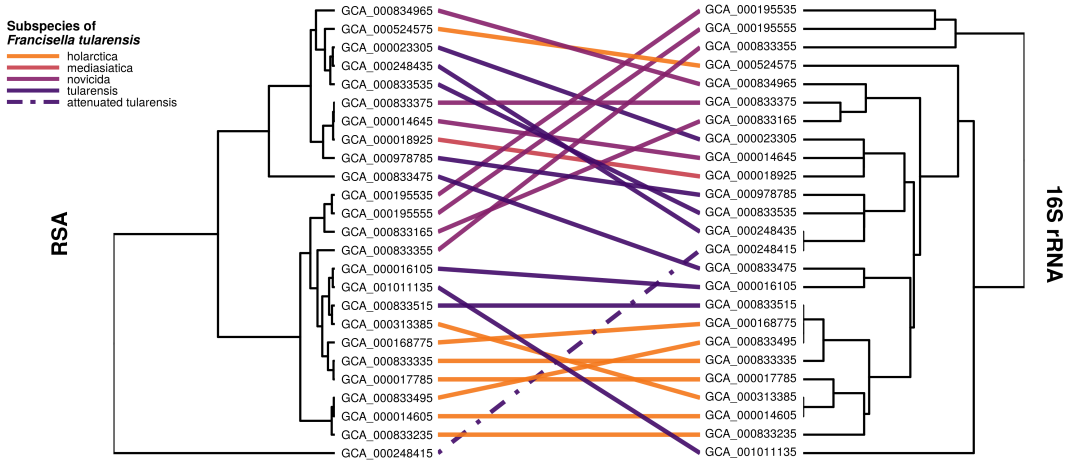


Figure S1.4: Hierarchical clustering of 25 strains of *Francisella tularensis*, by RSA method (*left*) and based on 16S rRNA gene (*right*). The strains belong to different subspecies: *holarctica* (*yellow*), *mediasiatica* (*orange*), *novicida* (*magenta*) and *tularensis* (*violet*). Strain TIGB03 (GCA_000248415) (*dashed violet*) is an attenuated *tularensis* strain. More precisely, it is an attenuated O-antigen mutant of virulent strain TI0902 (GCA_000248435).

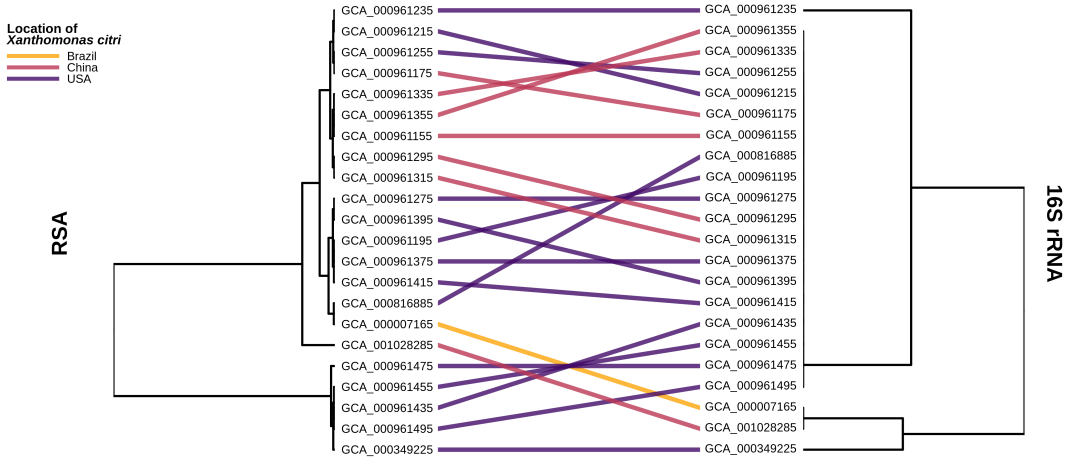


Figure S1.5: Hierarchical clustering of 22 strains of *Xanthomonas citri*, by RSA method (*left*) and based on 16S rRNA gene (*right*). The strains have different origin: Brazil (*yellow*), China (*magenta*) and USA (*violet*).

S2 Supplementary material of Part II

Table S2.1: Species with multiple assemblies available [52]. The column *Assembly* specifies the assembly used in this work. For species from the classes Actinopteri and Mammalia, the reference genome is chosen. The only exception is the sheep assembly *Oar_v3.1*, which is chosen over the reference one since the reference genome wasn't available for download. For species from the classes Arachnica and Insecta, we chose one of the available strains. We should note that the common name *mosquito* is used for all species belonging to the taxonomic family Culicidae, to which genus *Anopheles* belongs.

Species	**Species (common name)**	**Assembly**
Mus musculus	Mouse	GRCm38.p6
Canis lupus familiaris	Dog	CanFam3.1
Capra hircus	Goat	ARS1
Ovis aries	Sheep (texel)	Oar_v3.1
Cricetulus griseus	Chinese hamster	CHOK1GS_HDv1
Heterocephalus glaber	Naked mole-rat	HetGla_female_1.0
Oryzias latipes	Japanese medaka	ASM223467v1
Cyprinus carpio	Common carp	common_carp_genome
Astyanax mexicanus	Mexican tetra	Astyanax_mexicanus-2.0
Anopheles coluzzii	Mosquito	AcolM1
Anopheles sinensis	Mosquito	AsinS2
Anopheles stephensi	Mosquito	AsteS1
Ixodes scapularis	Deer tick	IscaW1

Table S2.2: Definition and classification of gene biotypes, as defined by Ensembl [52]. Only the biotypes found in the analyzed metazoan genomes are listed. Note that some biotypes have multiple spellings, probably caused by a spelling mistake in genome annotation.

Biotype group	**Ensembl biotype**
protein coding	protein_coding, IG_C_gene, IG_D_gene, IG_J_gene, IG_LV_gene, IG_V_gene, TR_C_gene, TR_D_gene, TR_J_gene, TR_V_gene, non-translating_CDS, polymorphic_pseudogene, unknown_likely_coding
pseudogene	pseudogene, processed_pseudogene, unprocesses_pseudogene, unitary_pseudogene, transcribed_processed_pseudogene, transcribed_unprocessed_pseudogene, transcribed_unitary_pseudogene, translated_processed_pseudogene, translated_unprocessed_pseudogene, rRNA_pseudogene, tRNA_pseudogene, IG_pseudogene, IG_C_pseudogene, IG_D_pseudogene, IG_J_pseudogene, IG_V_pseudogene, TR_J_pseudogene, TR_V_pseudogene
small non-coding RNA	miRNA, pre_miRNA, rRNA, tRNA, sRNA, snRNA, snoRNA, scRNA, scaRNA, SRP_RNA, Mt_rRNA, Mt_tRNA, ncRNA, piRNA, RNase_MRP_RNA, RNase_P_RNA, vault_RNA, vaultRNA, Y_RNA, misc_RNA
long non-coding RNA	lncRNA, lincRNA, macro_lncRNA, bidirectional_promoter_lncrna, bidirectional_promoter_lncRNA, 3prime_overlapping_ncRNA, antisense, antisense_RNA, sense_intronic, sense_overlapping, processed_transcript, ribozyme
unknown	TEC, transposable_element

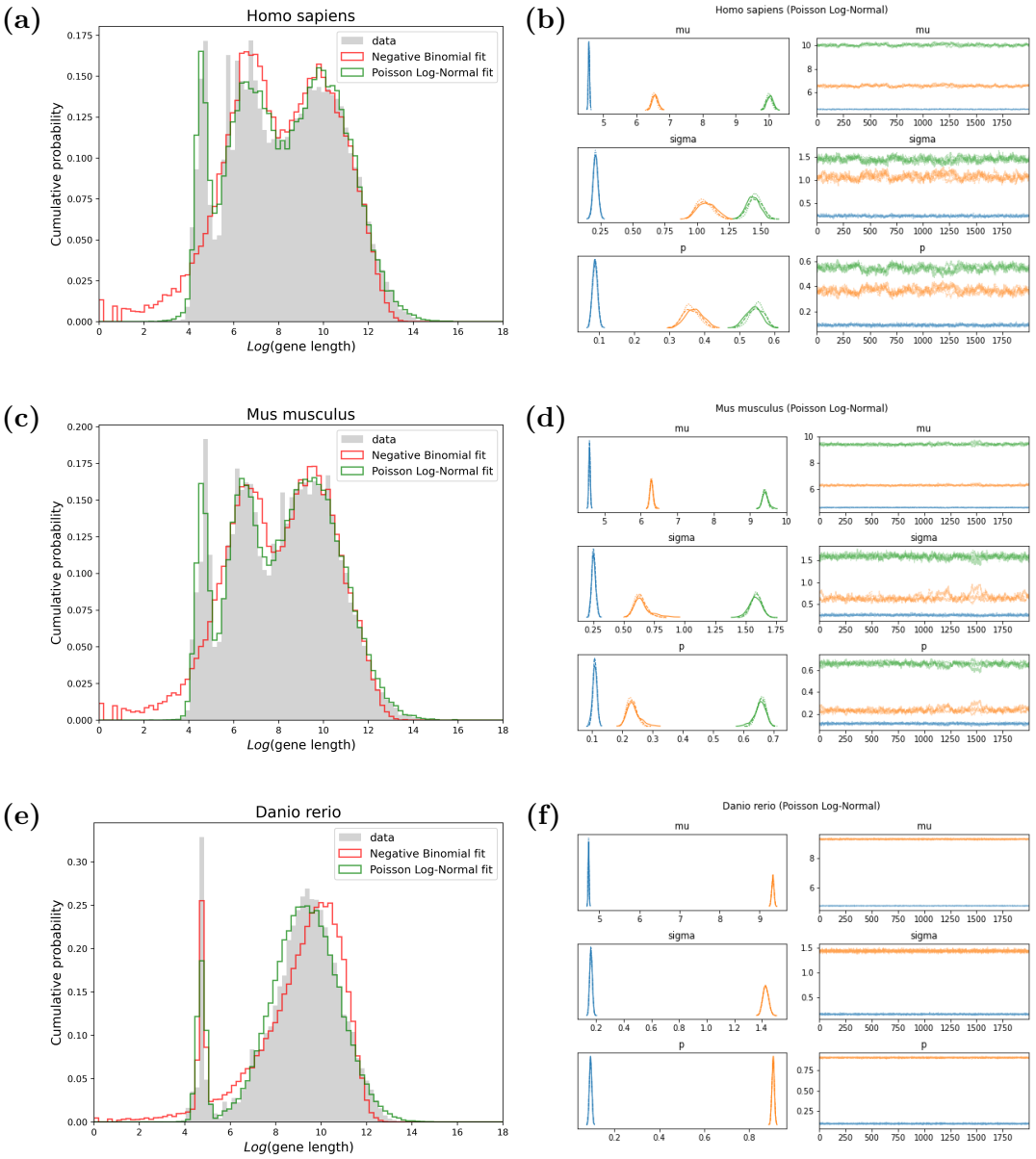


Figure S2.1: **(a)** Histogram comparison of the Negative Binomial and Poisson Log-Normal fit for the gene length distribution in *Homo sapiens*. A fitting distribution is calculated as a sample from posterior predictive distribution. **(a)** Traceplot of the Bayesian Poisson Log-Normal mixture model for *Homo sapiens* gene length distribution (without the burn-in period). **(c-d)** Histogram comparison and Poisson Log-Normal traceplot for *Mus musculus*. **(e-f)** Histogram comparison and Poisson Log-Normal traceplot for *Danio rerio*.

Table S2.3: Contingency table for coding-pseudogene prediction methods for *Mus musculus*. Two methods are compared, PoiLN method and logistic regression. **(i)** Prediction of protein-coding genes. PoiLN method has precision of 0.896 with recall of 0.928. Logistic regression has precision of 0.959 with recall of 0.654. **(ii)** Prediction of pseudogenes. PoiLN method has precision of 0.874 with recall of 0.822. Logistic regression has precision of 0.626 with recall of 0.954. **(iii)** Accuracy of PoiLN method is 0.888 compared with logistic regression accuracy of 0.767.

	PoiLN method		logistic regression		
true biotype	coding	pseudo	coding	pseudo	total
coding	19153	1482	13496	7139	20635
pseudo	2224	10303	576	11951	12527

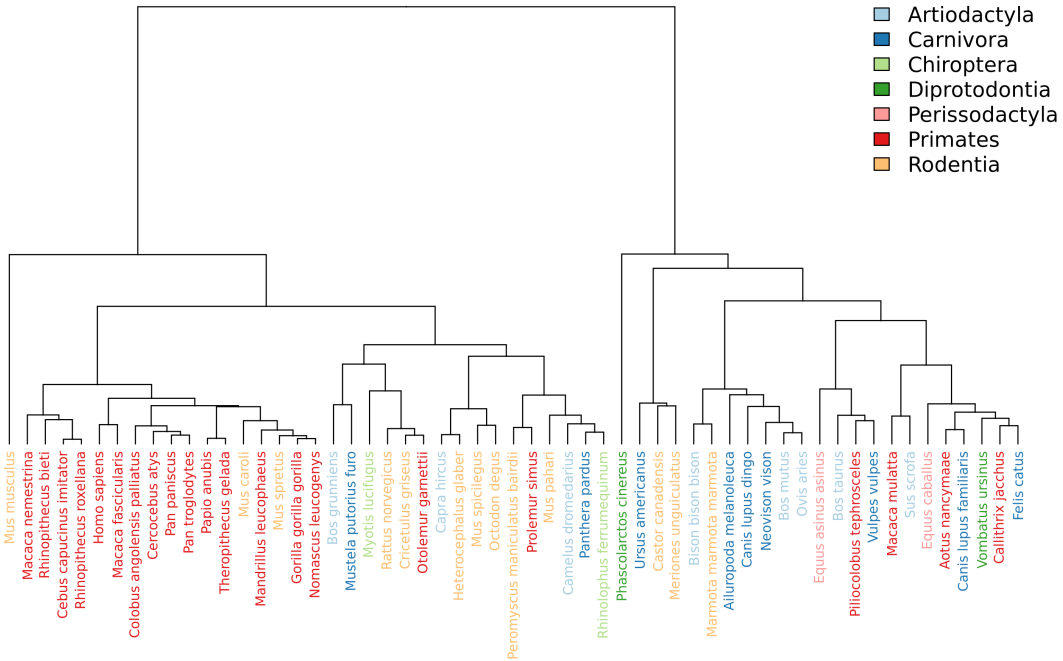


Figure S2.2: Dendrogram of 57 mammalian species. Distance is an Euclidean distance in standardized $(\mu_1, \sigma_1, \mu_2, \sigma_2)$-space, with $mean = 0, variance = 1$ standardization. Different colors represent different taxonomic orders, as indicated in the legend. Taxonomic orders to which only one species belonged are removed together with all species which had less than 20000 annotated pseudo-coding genes.

S3 Supplementary material of Part III

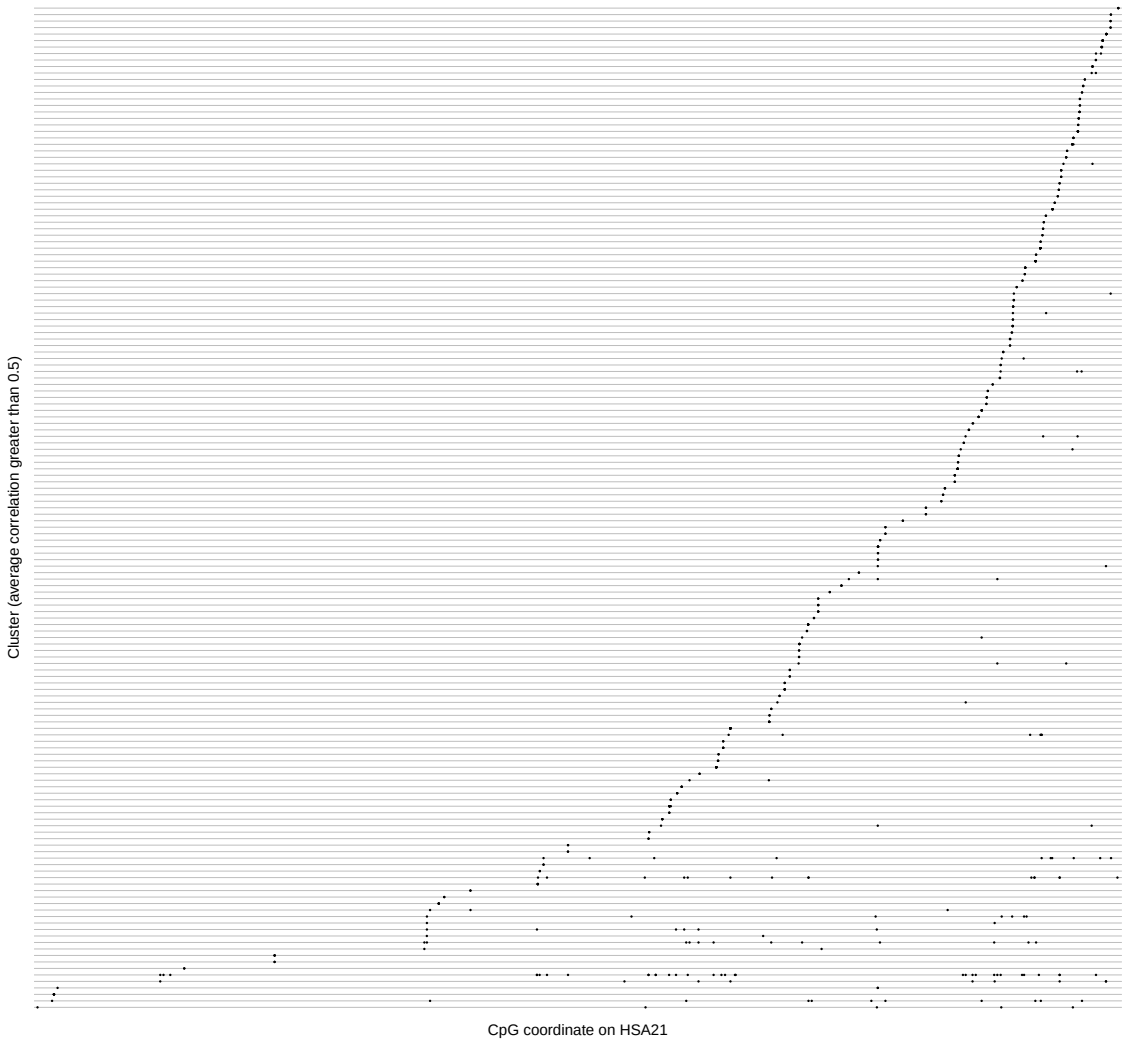


Figure S3.1: Clusters of CpG sites on HSA21. Every point represents a CpG site with HSA21 coordinate on *x-axis* and cluster number on *y-axis*. Hierarchical clustering with *average* linkage method was performed on the distance matrix among 3691 CpG sites, with dist $= 1 - \max(0, \text{cor})$. Dendrogram was cut at the height of 0.5, which resulted in 155 clusters with 528 CpGs in total and 3163 singletons. Singletons are the CpG sites which aren't part of any cluster and are not shown in the figure.

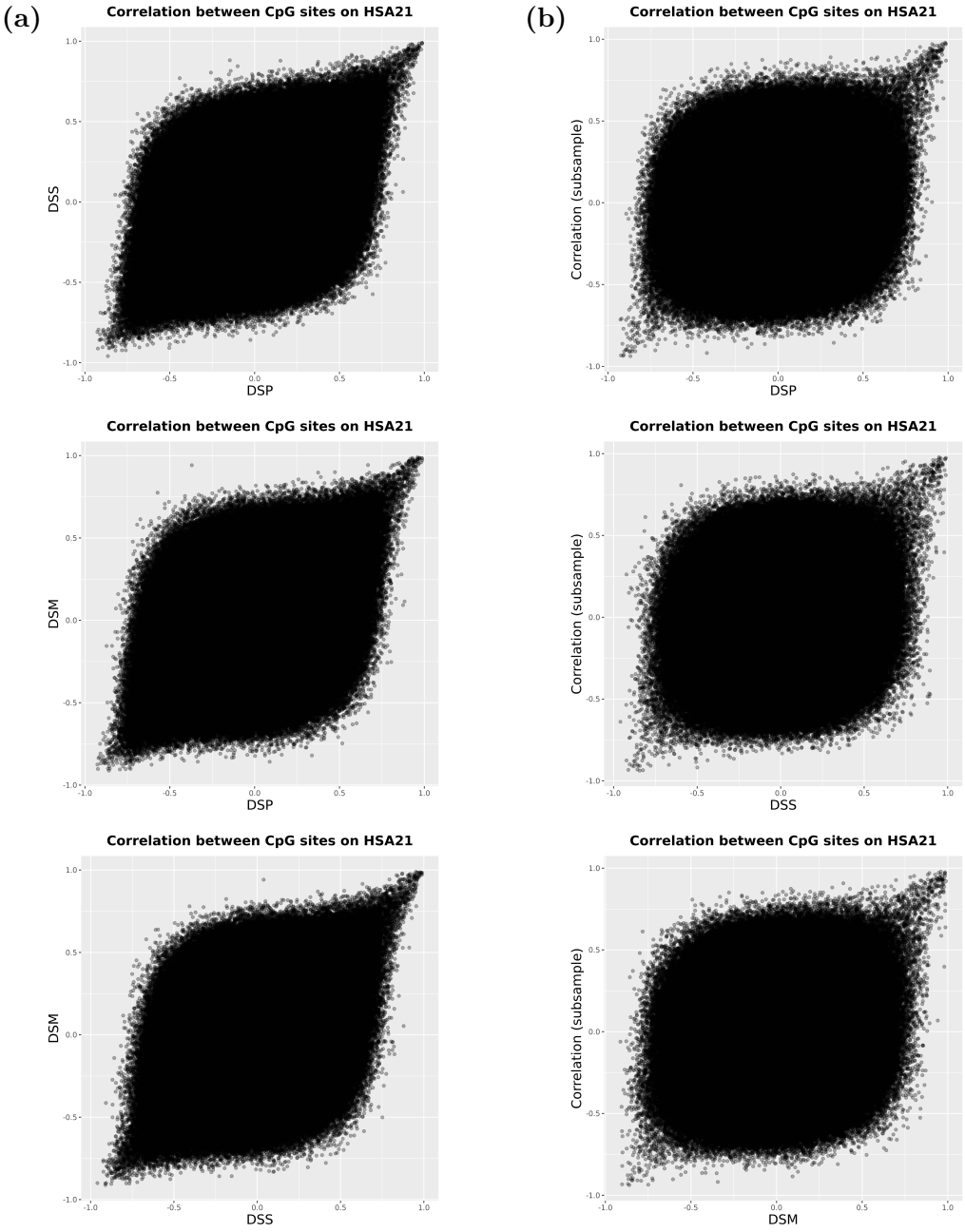


Figure S3.2: Correlations between CpG sites on HSA21 were calculated for different data sets. **(a)** Correlation comparison among Down syndrome data sets. **(b)** Correlation comparison between Down syndrome data sets and a random control subsample. 29 control samples were randomly chosen among 728 control samples and correlations among CpGs were calculated using only the chosen random subsample.

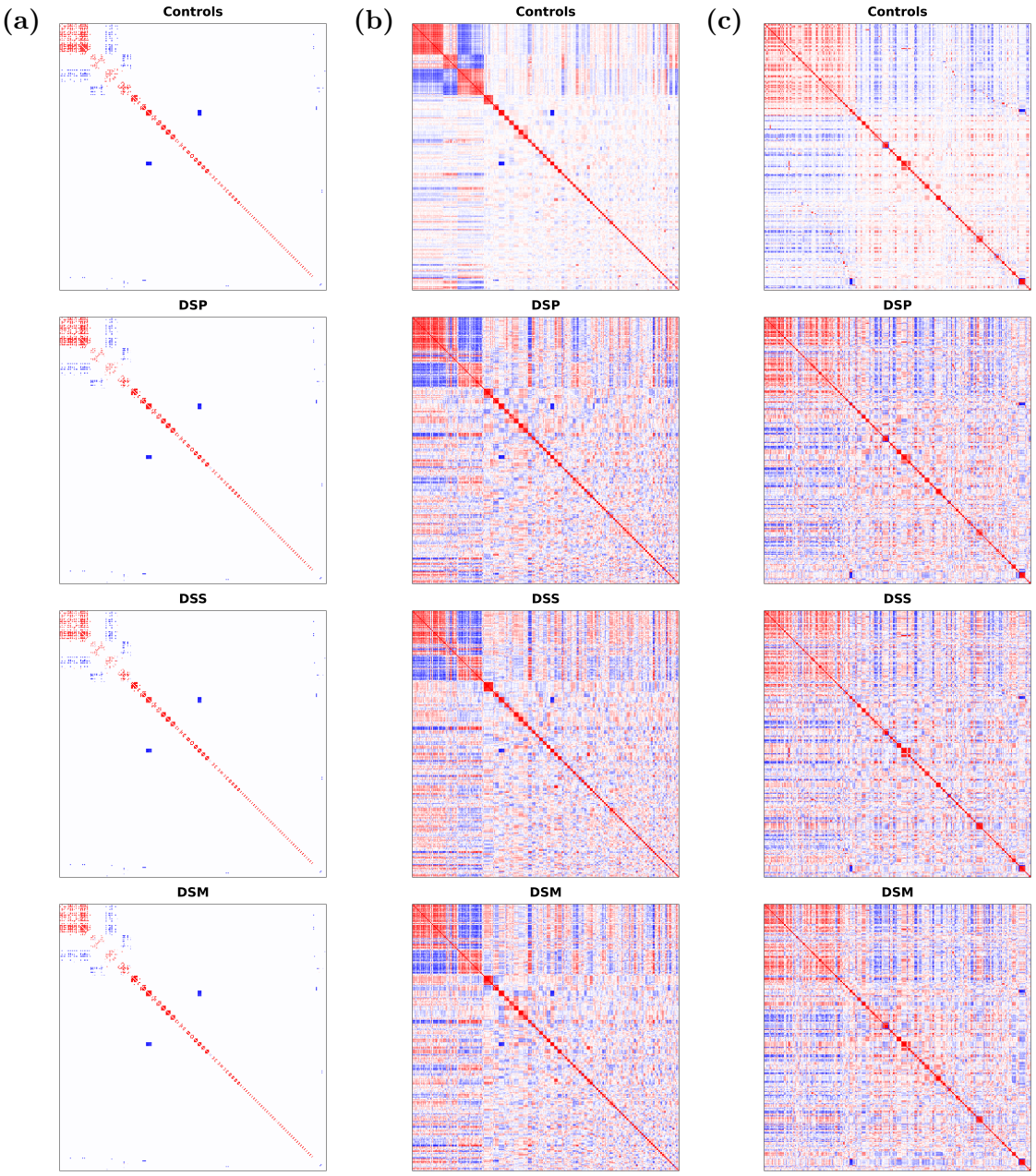


Figure S3.3: Correlation matrices for different data sets: controls, DSP, DSS and DSM. All matrices show the same 284 CpG sites, possibly in different order, with negative correlations in *blue* and positive correlations in *red*. **(a)** Only small-variance correlations are shown. The order of CpGs is based on cluster membership for *small-variance unweighted* clustering. **(b)** The same order of CpGs as in (a), with all correlations shown. **(c)** All correlations shown with the order of CpGs based on cluster membership for *positive correlation unweighted* clustering.

Bibliography

[1] Darren J Wilkinson. *Stochastic modelling for systems biology*. CRC press, 2018.

[2] Nicholas M Luscombe et al. "The dominance of the population by a selected few: power-law behaviour applies to a wide variety of genomic properties". In: *Genome biology* 3.8 (2002), pp. 1–7.

[3] Brian J McGill et al. "Species abundance distributions: moving beyond single prediction theories to integration within an ecological framework". In: *Ecology letters* 10.10 (2007), pp. 995–1015.

[4] Steinar Engen and Russell Lande. "Population dynamic models generating the lognormal species abundance distribution". In: *Mathematical biosciences* 132.2 (1996), pp. 169–183.

[5] Steinar Engen and Russell Lande. "Population dynamic models generating species abundance distributions of the gamma type". In: *Journal of Theoretical Biology* 178.3 (1996), pp. 325–331.

[6] Marina Bibikova et al. "High density DNA methylation array with single CpG site resolution". In: *Genomics* 98.4 (2011), pp. 288–295.

[7] Cecilia Lövkvist et al. "DNA methylation in human epigenomes depends on local topology of CpG sites". In: *Nucleic acids research* 44.11 (2016), pp. 5123–5132.

[8] Mark D Robinson et al. "Statistical methods for detecting differentially methylated loci and regions". In: *Frontiers in genetics* 5 (2014), p. 324.

[9] Samuel Karlin and Howard M Taylor. *A first course in stochastic processes*. 2nd ed. Academic Press, 1975.

[10] Samuel Karlin and Howard M Taylor. *A second course in stochastic processes*. 1st ed. Academic Press, 1981.

[11] Michael Turelli. "Random environments and stochastic calculus". In: *Theoretical population biology* 12.2 (1977), pp. 140–178.

[12] Ronald A Fisher, A Steven Corbet, and Carrington B Williams. "The relation between the number of species and the number of individuals in a random sample of an animal population". In: *The Journal of Animal Ecology* (1943), pp. 42–58.

[13] Frank W Preston. “The commonness, and rarity, of species”. In: *Ecology* 29.3 (1948), pp. 254–283.

[14] MG Bulmer. “On fitting the Poisson lognormal distribution to species-abundance data”. In: *Biometrics* (1974), pp. 101–110.

[15] Bruce Alberts et al. “Molecular biology of the cell”. In: (2014).

[16] Christine Vogel et al. “Structure, function and evolution of multidomain proteins”. In: *Current opinion in structural biology* 14.2 (2004), pp. 208–216.

[17] Christoph P Bagowski, Wouter Bruins, and Aartjan JW te Velthuis. “The nature of protein domain evolution: shaping the interaction network”. In: *Current genomics* 11.5 (2010), pp. 368–376.

[18] Xueying Xie, Jing Jin, and Yongyi Mao. “Evolutionary versatility of eukaryotic protein domains revealed by their bigram networks”. In: *BMC evolutionary biology* 11.1 (2011), p. 242.

[19] Oriane Hidalgo et al. “Is there an upper limit to genome size?” In: *Trends in Plant Science* 22.7 (2017), pp. 567–573.

[20] Ruiting Lan and Peter R Reeves. “Intraspecies variation in bacterial genomes: the need for a species genome concept”. In: *Trends in microbiology* 8.9 (2000), pp. 396–401.

[21] José SL Patané et al. “Origin and diversification of Xanthomonas citri subsp. citri pathotypes revealed by inclusive phylogenomic, dating, and biogeographic analyses”. In: *BMC genomics* 20.1 (2019), p. 700.

[22] Thomas Weinmaier et al. “Genomic factors related to tissue tropism in Chlamydia pneumoniae infection”. In: *BMC genomics* 16.1 (2015), p. 268.

[23] Paul Baumann. “Biology of bacteriocyte-associated endosymbionts of plant sap-sucking insects”. In: *Annu. Rev. Microbiol.* 59 (2005), pp. 155–189.

[24] Thero Modise et al. *Genomic comparison between a virulent type A1 strain of Francisella tularensis and its attenuated O-antigen mutant*. 2012.

[25] J Michael Janda and Sharon L Abbott. “16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls”. In: *Journal of clinical microbiology* 45.9 (2007), pp. 2761–2764.

[26] Joseph Felsenstein. *Inferring phylogenies*. 2nd ed. Sinauer Associates is an imprint of Oxford University Press, 2004.

[27] Dirk Gevers et al. “Re-evaluating prokaryotic species”. In: *Nature Reviews Microbiology* 3.9 (2005), p. 733.

[28] William P Hanage, Christophe Fraser, and Brian G Spratt. “Sequences, sequence clusters and bacterial species”. In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 361.1475 (2006), pp. 1917–1927.

[29] Iñaki Comas, Andrés Moya, and Fernando González-Candelas. "Phylogenetic signal and functional categories in Proteobacteria genomes". In: *BMC evolutionary biology* 7.S1 (2007), S7.

[30] David L Wheeler et al. "Database resources of the national center for biotechnology information". In: *Nucleic acids research* 28.1 (2000), pp. 10–14.

[31] Jasper J Koehorst et al. "SAPP: functional genome annotation and analysis through a semantic framework using FAIR principles". In: *Bioinformatics* 34.8 (2018), pp. 1401–1403.

[32] Doug Hyatt et al. "Prodigal: prokaryotic gene recognition and translation initiation site identification". In: *BMC bioinformatics* 11.1 (2010), p. 119.

[33] Philip Jones et al. "InterProScan 5: genome-scale protein function classification". In: *Bioinformatics* 30.9 (2014), pp. 1236–1240.

[34] Paulo I. Prado, Murilo Dantas Miranda, and Andre Chalom. *sads: Maximum Likelihood Models for Species Abundance Distributions*. R package version 0.4.2. 2018.

[35] Kenneth P Burnham, David R Anderson, and Kathryn P Huyvaert. "AIC model selection and multimodel inference in behavioral ecology: some background, observations, and comparisons". In: *Behavioral ecology and sociobiology* 65.1 (2011), pp. 23–35.

[36] Christian Quast et al. "The SILVA ribosomal RNA gene database project: improved data processing and web-based tools". In: *Nucleic acids research* 41.D1 (2012), pp. D590–D596.

[37] Tomáš Větrovsk and Petr Baldrian. "The variability of the 16S rRNA gene in bacterial genomes and its consequences for bacterial community analyses". In: *PloS one* 8.2 (2013), e57923.

[38] Patrick D Schloss et al. "Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities". In: *Applied and environmental microbiology* 75.23 (2009), pp. 7537–7541.

[39] Joe H Ward Jr. "Hierarchical grouping to optimize an objective function". In: *Journal of the American statistical association* 58.301 (1963), pp. 236–244.

[40] Frederick M Cohan. "What are bacterial species?" In: *Annual Reviews in Microbiology* 56.1 (2002), pp. 457–487.

[41] Pablo A Estévez et al. "Normalized mutual information feature selection". In: *IEEE Transactions on neural networks* 20.2 (2009), pp. 189–201.

[42] Simone Romano et al. "Standardized mutual information for clustering comparisons: one step further in adjustment for chance". In: *International Conference on Machine Learning*. 2014, pp. 1143–1151.

[43] GSA Myers et al. “Evidence that human Chlamydia pneumoniae was zoonotically acquired”. In: *Journal of bacteriology* 191.23 (2009), pp. 7225–7233.

[44] Gary Xie et al. “Exception to the rule: genomic characterization of naturally occurring unusual Vibrio cholerae strains with a single chromosome”. In: *International journal of genomics* 2017 (2017).

[45] Kazuhisa Okada et al. “Characterization of 3 megabase-sized circular replicons from Vibrio cholerae”. In: *Emerging Infectious Diseases* 21.7 (2015), p. 1262.

[46] Laura Gómez-Valero, Amparo Latorre, and Francisco J Silva. “The evolutionary fate of nonfunctional DNA in the bacterial endosymbiont Buchnera aphidicola”. In: *Molecular Biology and Evolution* 21.11 (2004), pp. 2172–2181.

[47] Anne Holch et al. “Genome sequencing identifies two nearly unchanged strains of persistent Listeria monocytogenes isolated at two different fish processing plants sampled 6 years apart”. In: *Applied and environmental microbiology* 79.9 (2013), pp. 2944–2951.

[48] Terence A. Brown. *Genomes 4*. 4th ed. Garland Science, 2017.

[49] Vladislav Grishkevich and Itai Yanai. “Gene length and expression level shape genomic novelties”. In: *Genome research* 24.9 (2014), pp. 1497–1503.

[50] Francesca Chiaromonte, Webb Miller, and Eric E Bouhassira. “Gene length and proximity to neighbors affect genome-wide expression levels”. In: *Genome research* 13.12 (2003), pp. 2602–2608.

[51] Lin Xu et al. “Average gene length is highly conserved in prokaryotes and eukaryotes and diverges only between the two kingdoms”. In: *Molecular biology and evolution* 23.6 (2006), pp. 1107–1108.

[52] Andrew D Yates et al. “Ensembl 2020”. In: *Nucleic acids research* 48.D1 (2020), pp. D682–D688.

[53] Seth W Cheetham, Geoffrey J Faulkner, and Marcel E Dinger. “Overcoming challenges and dogmas to understand the functions of pseudogenes”. In: *Nature Reviews Genetics* 21.3 (2020), pp. 191–201.

[54] Ilenia D’Errico, Gemma Gadaleta, and Cecilia Saccone. “Pseudogenes in metazoa: origin and features”. In: *Briefings in Functional Genomics* 3.2 (2004), pp. 157–167.

[55] Thomas C Roberts and Kevin V Morris. “Not so pseudo anymore: pseudogenes as therapeutic targets”. In: *Pharmacogenomics* 14.16 (2013), pp. 2023–2034.

[56] Hadas Hezroni et al. “A subset of conserved mammalian long non-coding RNAs are fossils of ancestral protein-coding genes”. In: *Genome biology* 18.1 (2017), pp. 1–15.

[57] Kevin L Howe et al. “Ensembl genomes 2020—enabling non-vertebrate genomic research”. In: *Nucleic acids research* 48.D1 (2020), pp. D689–D695.

[58] Steffen Durinck et al. “Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt”. In: *Nature protocols* 4.8 (2009), p. 1184.

[59] Rhoda J Kinsella et al. “Ensembl BioMarts: a hub for data retrieval across taxonomic space”. In: *Database* 2011 (2011).

[60] Andrew Yates et al. “The Ensembl REST API: Ensembl data for any language”. In: *Bioinformatics* 31.1 (2015), pp. 143–145.

[61] Scott A Chamberlain and Eduard Szöcs. “taxize: taxonomic search and retrieval in R”. In: *F1000Research* 2 (2013).

[62] Hajk-Georg Drost et al. “myTAI: evolutionary transcriptomics with R”. In: *Bioinformatics* 34.9 (2018), pp. 1589–1590.

[63] Y. Roskov et al. “Species 2000 ITIS Catalogue of Life, 2019 Annual Checklist”. Digital resource at www.catalogueoflife.org/annual-checklist/2019. In: *Species 2000: Naturalis, Leiden, the Netherlands. ISSN 2405-884X.* (2019).

[64] John Salvatier, Thomas V Wiecki, and Christopher Fonnesbeck. “Probabilistic programming in Python using PyMC3”. In: *PeerJ Computer Science* 2 (2016), e55.

[65] John Kruschke. *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. 2nd ed. Academic Press, 2014.

[66] Alison Yao, Rosane Charlab, and Peter Li. “Systematic identification of pseudogenes through whole genome expression evidence profiling”. In: *Nucleic acids research* 34.16 (2006), pp. 4477–4485.

[67] Marijke J van Baren and Michael R Brent. “Iterative gene prediction and pseudogene removal improves genome annotation”. In: *Genome research* 16.5 (2006), pp. 678–685.

[68] Yuan Zhang and Yanni Sun. “PseudoDomain: identification of processed pseudogenes based on protein domain classification”. In: *Proceedings of the ACM Conference on Bioinformatics, Computational Biology and Biomedicine*. 2012, pp. 178–185.

[69] Yoshihito Niimura, Atsushi Matsui, and Kazushige Touhara. “Extreme expansion of the olfactory receptor gene repertoire in African elephants and evolutionary dynamics of orthologous gene groups in 13 placental mammals”. In: *Genome research* 24.9 (2014), pp. 1485–1496.

[70] Claudia Sala et al. “Evaluation of pre-processing on the meta-analysis of DNA methylation data from the Illumina HumanMethylation450 BeadChip platform”. In: *Plos one* 15.3 (2020), e0229763.

[71] Tanya Barrett et al. “NCBI GEO: archive for functional genomics data sets—update”. In: *Nucleic acids research* 41.D1 (2012), pp. D991–D995.

[72] Åsa Johansson, Stefan Enroth, and Ulf Gyllensten. "Continuous aging of the human DNA methylome throughout the human lifespan". In: *PloS one* 8.6 (2013), e67378.

[73] Maria Giulia Bacalini et al. "Identification of a DNA methylation signature in blood cells from persons with Down Syndrome". In: *Aging (Albany NY)* 7.2 (2015), p. 82.

[74] Donna Karolchik et al. "The UCSC Table Browser data retrieval tool". In: *Nucleic acids research* 32.suppl_1 (2004), pp. D493–D496.

[75] Martin Morgan and Lori Shepherd. *AnnotationHub: Client to access AnnotationHub resources*. R package version 2.20.2. 2020.

[76] Team TBD. *BSgenome.Hsapiens.UCSC.hg19: Full genome sequences for Homo sapiens (UCSC version hg19, based on GRCh37.p13)*. R package version 1.4.3. 2020.

[77] Matthew Suderman, Gibran Hemani, and Josine Min. *meffil: Efficient algorithms for DNA methylation*. R package version 1.1.1. 2020.

[78] Wanding Zhou, Peter W Laird, and Hui Shen. "Comprehensive characterization, annotation and innovative use of Infinium DNA methylation BeadChip probes". In: *Nucleic acids research* 45.4 (2017), e22–e22.

[79] Martin Ester et al. "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise". In: *KDD'96*. Portland, Oregon: AAAI Press, 1996, pp. 226–231.

[80] Michael Hahsler, Matthew Piekenbrock, and Derek Doran. "dbscan: Fast Density-Based Clustering with R". In: *Journal of Statistical Software* 91.1 (2019), pp. 1–30.

[81] Pan Du et al. "Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis". In: *BMC bioinformatics* 11.1 (2010), p. 587.

[82] Juliane Schafer et al. *corpcor: Efficient Estimation of Covariance and (Partial) Correlation*. R package version 1.6.9. 2017.

[83] William Revelle. *psych: Procedures for Psychological, Psychometric, and Personality Research*. R package version 2.0.9. Northwestern University. Evanston, Illinois, 2020.

[84] Markus Herdin and Ernst Bonek. "A MIMO correlation matrix based metric for characterizing non-stationarity". In: *IST Mobile Wireless Communications Summit*. Lyon, France, 2004.

[85] Gabor Csardi and Tamas Nepusz. "The igraph software package for complex network research". In: *InterJournal* Complex Systems (2006), p. 1695.

[86] Galit Lev Maor, Ahuvi Yearim, and Gil Ast. “The alternative role of DNA methylation in splicing regulation”. In: *Trends in Genetics* 31.5 (2015), pp. 274–280.

[87] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. “Sparse inverse covariance estimation with the graphical lasso”. In: *Biostatistics* 9.3 (2008), pp. 432–441.

[88] Wessel N Van Wieringen and Carel FW Peeters. “Ridge estimation of inverse covariance matrices from high-dimensional data”. In: *Computational Statistics & Data Analysis* 103 (2016), pp. 284–303.

[89] Shi-Yi Chen, Zhe Feng, and Xiaolian Yi. “A general introduction to adjustment for multiple comparisons”. In: *Journal of thoracic disease* 9.6 (2017), p. 1725.