UNIVERSITY OF PADUA
DEPARTMENT OF MATHEMATICS "TULLIO LEVI-CIVITA"
MASTER DEGREE IN MATHEMATICS

MASTER THESIS IN MATHEMATICS

STATISTICS OF UNSEEN VARIANTS IN GENOMICS: MATHEMATICAL MODELING AND DATA ANALYSIS

Supervisor:
Dr. Marco Formentin

Co-Supervisors:
Dr. Stefano Lise
Dr. Anna Tovo

Candidate:
Anna Fochesato
Student Number 1179361

ACADEMIC YEAR 2018-2019

Abstract

In the last years quantitative approaches have gained increasing importance in genomics research due to their ability of interpreting and characterizing the vast amount of genetic data that the new sequencing technologies have made available. Statistical tools in particular seem to play a central role in gaining information on the human DNA mutation variability. Such mutations may lead to tumor insurgence and progression, thus there is the exigency to develop analytical methods that could quantify the genetic heterogeneity of a tumor, whose knowledge may be crucial to design the best therapeutic setting. Within this framework, the present thesis aims to inference the statistical description of a genetic region taking as input few samples only. To this end, we present an ecological-inspired method to predict the number of mutations in a DNA sequence or in a whole tumor (global scale) from presence/absence information collected in a portion of the region (local scale). For our model, we have assumed to work under the neutral hypothesis of mutation demographic equivalence and within the parametric framework of a global RSA, i.e. frequency of mutations at given occurrence abundance, distributed according to a Negative Binomial. This latter choice has been justified by both the derivation of the Negative Binomial as steady state solution of a biological birth and death process and by the functional versatility Negative Binomial has in well accommodating different empirical RSA shapes (power law, Log-Series, unimodal). Under the hypothesis of demographic equivalence of mutations, it can be proved that the Negative Binomial is form invariant, i.e. a random subsample can still be described via a Negative Binomial distribution. In other words, the local scale RSA is a Negative Binomial if the global scale RSA is a Negative Binomial. It has followed a computable formula bridging the parameters of the RSA at the local scale to those at the global scale, which we have exploited to end up with an unbiased and consistent estimator of the number of global mutations in the genetic region of interest. Simulations on both DNA single-nucleotide polymorphism and synthetic spatial tumor growth datasets have been performed at last to test our framework. The promising results they have given back would confirm the stability and the reliability of the proposed method in genetic field.

Sommario

Negli ultimi anni l'utilizzo di approcci quantitativi in genomica è stato cruciale per interpretare e caratterizzare la vasta gamma di dati genetici resi disponibili dalle nuove tecniche di sequenziamento del DNA. In particolare, alcuni strumenti di indagine statistica si sono dimostrati indispensabili per ottenere un *insight* sulla variabilità genetica del DNA umano. Le mutazioni genetiche sono implicate nell'insorgenza e nella progressione dei tumori, da qui l'esigenza di sviluppare metodi analitici per quantificare l'eterogeneità genetica di un tumore, la cui conoscenza potrebbe avere valenza informativa nel design di terapie più efficaci. All'interno di questo *framework*, il presente lavoro di tesi si propone di inferire una completa descrizione statistica di una regione genetica, basandosi su pochi campionamenti della stessa. A tale scopo presentiamo un metodo di ispirazione ecologica per predire il numero di mutazioni presenti in un'intera sequenza di DNA o in un tumore (scala globale) partendo esclusivamente da informazioni di presenza/assenza sulle mutazioni in una porzione della regione scelta (scala locale). Per il nostro modello abbiamo assunto un *framework* parametrico in cui la RSA a scala globale, ovvero la frequenza delle mutazioni aventi una data occorrenza quantitativa, è distribuita secondo una Binomiale Negativa e in cui l'ipotesi neutrale di equivalenza demografica tra le mutazioni è soddisfatte. La scelta della Binomiale Negativa come famiglia di distribuzioni per il nostro metodo è giustificata dalla derivazione di tale distributione come soluzione all'equilibrio di un processo biologico stocastico di nascita e morte e dalla sua versalità nell'accomodare RSAs empiriche aventi forme diverse (power law, Log-Series, unimodali). Le ipotesi di lavoro sulla neutralità hanno permesso di provare l'invarianza in forma della Binomiale Negativa, ossia la proprietà per cui, data una Binomiale Negativa a scala globale, un suo campionamento è ancora descritto mediante una tale distribuzione. Da questa proprietà è seguita una formula computabile, collegante i parametri della RSA locale a quelli della RSA globale, che abbiamo utilizzato per derivare analiticamente l'espressione di uno stimatore corretto e consistente per il numero di mutazioni presenti a scala globale. Testato su datasets riguardanti polimorfismi del DNA e simulazioni spaziali della crescita tumorale, il metodo proposto ha restituito risultati promettenti che ne hanno certificano la bontà e la stabilità.

Contents

List of Figures

List of Tables

Introduction

During the last couple of decades, the development of new DNA sequencing technologies has provided researchers with a huge amount of genetic data and patterns. Thus, an increasing importance has been given in genomics field to computational and mathematical models, capable of interpreting such data and revealing unknown and counterintuitive biological principles overlooked by classical qualitative approaches (Byrne, 2010). Within this framework, we present a statistical method to infer the number of unseen variants in genomics region from local quantitative information. That is, assuming to know mutation occurrences in a portion of a genetic region - DNA sequence, tumor - representing the local scale, we aim at extrapolating statistical information on the number of mutations and their distribution in the whole region sample, i.e. at global scale.

Such a statistical tool to predict the number of unseen variants and their abundances is particularly informative in cancer research where the complete description of genetic profile of a tumor could lead to relevant clinical implications. Indeed, most effective therapies are those targeting at the most common genome alterations shared by cancer cells. Thus, the information on the statistics of tumor mutations from single-patient biopsies may be crucial to design therapies or to evaluate the clinical course following surgery. Nowadays, for ethical and economical reasons, diagnoses and treatment decisions are mainly based on results from few biopsy samples only and these are often unlikely to accurately capture the complete and precise mutational profile of the tumor. Therefore, we hope that being able to infer heterogeneity properties from such small tumor fraction, as our method does, could represent an useful tool in cancer research.

In order to tackle this challenge, in the present thesis we have adapted to genomics, and more specifically to oncology, a statistical method previously developed for biodiversity estimation (Tovo et al., 2017). Indeed, the ecology of cancer has recently emerged as a promising approach in capturing genetic statistical features. In particular, researchers have started to look at cancer from an ecological perspective as an evolutionary process in which new species adopting different surviving strategies try to invade new habitat (tissue) (Kareva, 2011). Following such an ecological new trend, in our approach we have supposed the problem of inferring the full extent of genetic mutations in a DNA sequence or in a tumor to be closely related to the 'unseen species problem', having its roots in ecology. Such a problem was proposed in the early 1940s by the British naturalist Corbet who listed species and number of individuals per species for all butterflies he had trapped during a journey in Malaysia and wondered how many new butterflies he would have seen if he had come back there. First tackled by Fisher, such a problem has arisen the interest of many other scientists throughout the years, leading to a myriad of estimators with applications in different fields, from bioscience to linguistics and in

the present thesis to genomics. As a consequence of its ecological background, many tools and patterns coming from ecology field - e.g. *species-abundance distribution* (SAD), telling how commonness and rarity are distributed among the species of an ecosystem, *species-area curve* (SAC), looking at how biodiversity changes with the sampled area, *relative-species abundance* (RSA), measuring the species frequency at given abundance - have been extensively used in our framework.

Due to the interdisciplinary nature of this thesis, Chapter 1 provides some basic biological knowledge that the reader will need in the following. In the first part an overview of DNA structure and cancer hallmarks is taken, whereas in the second some mathematical models and techniques looking at the tumor dynamics from an ecological perspective are presented. Within this last section, the attention will be paid in particular to the neutral theory. Fuelled by the empirical observation of similar macro-patterns arising from different ecological systems, the neutral theory has been developed to derive such universal traits at macro level from simple key rules of the microscopic dynamics. Its main finding is the explanation of the emergence of regular patterns by mean of simple models driven by stochasticities, disregarding species own identities and detailed biological features of the dynamics.

Chapter 2 is the mathematical core of the thesis, presenting the ecological-inspired framework we have adapted to upscale mutation richness in a genetic region. Within our framework, we have assumed the RSA at the global scale, i.e. the probability for a mutation to be observed n times, to be a Negative Binomial (NB) of parameters r and $0 < \xi < 1$ with normalization to (at least) one:

$$P(n) = \begin{cases} \frac{1}{1-(1-\xi)^r}\binom{n+r-1}{n}\xi^n(1-\xi)^r & \text{if } r > 0,\ n \geq 1 \\ \frac{1}{1-(1-\xi)^r \cdot \sum_{j=0}^{m-1}\binom{j+r-1}{j}\xi^j}\binom{n+r-1}{n}\xi^n(1-\xi)^r & \text{if } r \in (-m, -m+1),\ m \in \mathbb{N}^+,\ n \geq m \end{cases} \tag{1}$$

Observe that the value of r affects the support of the distribution in terms of the well-posedness of the binomial. Both stochastic and statistical reasons have driven us to select Negative Binomial as family of statistical distributions for our method:

- Negative Binomial arises naturally as steady-state distribution of a system modeled through a birth-death dynamics, with birth rate accounting weakly for intraspecific interactions;
- Negative Binomial is a versatile distribution, capable of well accommodating different functional shapes that may arise empirically: Log-Series for $r \to 0$, unimodal mode for $r > 1$, power law for $r < 0$ and bimodal mode with convex combinations;
- Negative Binomial is proved to satisfy the form invariance property under the neutral hypothesis on mutation equivalence. Such a property means that a sample of any size of a Negative Binomial with parameters (r, ξ) can still be described by a Negative Binomial having computable, new parameters.
  In particular, whereas ξ parameter scales according to the computable formula

$$\xi = \frac{\hat{\xi}_p}{p + (1-p)\hat{\xi}_p} = f\left(\hat{\xi}_p, p\right), \tag{2}$$

  r keeps invariant as scale varies, i.e. $r = \hat{r}_p$.

A maximum likelihood method has been used to estimate the Negative Binomial parameters of the RSA at the local scale, whereas for the global parameter we have exploited Eq.(2). Finally, the formalism guaranteed by the assumptions on neutrality and on Negative Binomially RSAs has led us to the analytical derivation of our estimator for the number of mutations at global scale S. Such an estimator has been found to be

$$\hat{S} = \begin{cases} S_p \cdot \frac{1-(1-\xi)^r}{1-(1-\hat{\xi}_p)^r} & \text{when } r > 0 \\ S_p \cdot \frac{1-(1-\xi)^r \cdot \sum_{j=0}^{m-1} \binom{j+r-1}{j} \xi^j}{1-(1-\hat{\xi}_p)^r \cdot \sum_{j=0}^{m-1} \binom{j+r-1}{j} \hat{\xi}_p^j} & \text{when } r \in (-m, -m+1), m \in \mathbb{N} \setminus \{0\} \end{cases} \tag{3}$$

where S_p is the number of seen mutations at the local scale, $\hat{\xi}_p$ and r are the fitted local parameters and $\xi = f(\hat{\xi}_p, p)$ is a global parameter. Since the estimator we proposed has been proved to be unbiased and consistent, it could be profitably used for inferences.

Chapter 3 is focused on the application of our method to DNA single-nucleotide polymorphisms. In particular, it has been tested on three different datasets concerning observed variants in three independent regions of the X chromosome of 46 British males. We have considered a sample of $\bar{p} < 46$ people, counted the different mutations they carried and, on the basis of such local quantitative information, we have predicted the number of mutations in the whole ensemble of 46 individuals. The results of our estimates are here presented both numerically and graphically. They look promising, with an average error on prediction less than 10% and a predictable asymptotic behaviour.

In Chapter 4 our framework is applied to cancer heterogeneity. For ethical reasons, datasets on real tumors normally consist on presence/absence information for at most 10 human biopsies, which is a too small size to test our method with profit. Thus, our tests have been conducted on synthetic tumor growth datasets whose size we could control. The spatial stochastic model for tumor growth we have used to collect data is presented at first, then simulations on the model output data are performed. Again, a portion of the simulated tumor cells has been selected and a quantitative characterization of the mutations at that local scale has been done. Then, using our upscaling framework, estimates of the number of variants present in the whole tumor cells have been computed and corresponding results exposed. Even in this context, the results are good with relative error and estimate dispersion decreasing as the sampled tumor portion increases.

Chapter 1

Overview on cancer ecology and modeling

Highlighted the biological intents of the present thesis, we have decided to start our discussion with an overview on the main genetic and biological features that the reader will cross in the following. In particular, the first section of the chapter is a quick path across DNA strands, genes and mutations, whereas the second, dealing with the cancer hallmarks, proposes itself as a review of the some quantitative cancer models.

1.1 DNA structure and mutations

The cornerstone of genetics, the field of biology studying hereditary traits, is represented by DNA molecules which store all genetic information for a cell and an organism. Each human cell, with the exception of red blood cells, contains an entire copy of the genome, i.e. the genetic profile carried by DNA strands, which is passed to the daughter cells during the cell reproduction. These information are stored in the nucleus of the cells in the form of chromosomes, which are compact and highly spiral tangles of genetic material whose number and morphology depend on the species considered. DNA molecules display a double helix shape with two complementary twisted chains and are compound of different functional unities called genes. These latter are particular sections of DNA coding for the biochemical synthesis of specific aminoacid macromolecules - the proteins - whose tasks include cell signaling, immunological responses and cell cycle. More technically, a DNA chain is a linear sequence of four nucleotides - adenine (A), thymine (T), cytosine (C), guanine (G) - with

$$A \quad - - - \quad T \qquad \text{by mean of 2 hydrogen bonds,}$$

$$C \quad - - - \quad G \qquad \text{by mean of 3 hydrogen bonds.}$$

The A-T bond is weaker than the C - G one, thus it separates more easily (Durrett, 2010).

Some bases could be affected by alterations that could lead to a change in DNA basis sequence and basis pairing. These alterations affect the genotype of the individual, but may not produce discernible changes in the observable characteristics (phenotype). Moreover, their occurrences in somatic cells result in mutation transmission to daughter cells, while, more dramatically, occurrences in germinal cells, i.e. the gametes, are inherited by the individual's offspring.

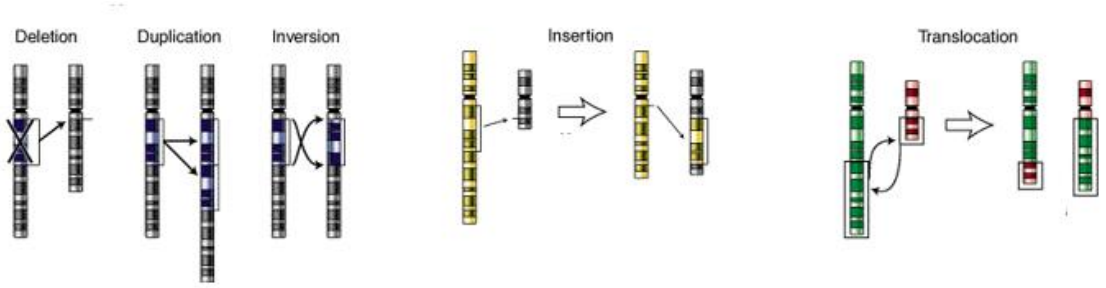


Figure 1.1: Different typologies of alterations in nucleic sequences

Different typologies of mutations are listed according to the nature of changes occurred in the nucleic sequence. They are:

- insertion mutations, which are due to the adding of one or more extra nucleotides into the DNA sequence (small-scale) or of chromosome sections into another chromosome (large-scale);

- deletion mutations, which consist in the removal of one or more nucleotides from the DNA sequence (small-scale) or in the drop of a portion of chromosome (large-scale);

- substitution mutations or traslocations, which are due to an exchange of a single nucleotide, like the transition $C \Leftrightarrow T, A \Leftrightarrow G$, (small-scale) or of different chromosome sections (large-scale).

All these mutation typologies can be driven by different factors, such as

- errors during the DNA replication with DNA polymerase, i.e. the enzyme synthesizing DNA molecules from deoxyribonucleotides, adding sometimes wrong nucleotide in the new DNA strand;

- errors during the DNA repairing with the introduction of alterations in the end rejoin process that follows the removal of wrong nucleotides;

- chemical endogenous agents and spontaneous events, like oxidative damages to the DNA strands or base alteration due to hydrogen shift;

- chemical and physical mutagens (X, Gamma and UV rays for example), that can cause base damages, insertions and deletions.

Our organism has a self-mechanism to repair mutations, that operates in two different manners: through the direct reversal of the chemical process generating the damage and through the replacement of damaged nucleotide bases. The majority of the alterations undergo such a mechanism, but others do not. Depending on where these latter occur and whether they alter the function of essential proteins, mutations could have various and serious effects on health, leading to genetic diseases in worst cases.

1.2 Genetic of cancer

One of the most common genetic disease is cancer, which is the consequence of random accumulation of mutations at molecular level (La Porta and Zapperi, 2017). Cancer develops from a single mutated cell and is fuelled by such a cell expansion process, during which daughter cells inherited the altered genetic heritage and may undergo other variations and selection events. However, not all the mutations in cell nuclei feed directly the development of cancer (Beerenwinkel et al., 2007). Some of them, the *passengers*, are neutral having no functional consequences, whereas others, the *drivers*, have a deleterious effect on growth advantages conferred to mutated clones and thus on these latter's spread. Cancer main features are the lack of growth control from the cell cycle regulatory mechanism due to gene alterations and the consequent abnormal cell expansion that may invade tissues, causing metastasis. Indeed, mutations fuelling tumor primarily alter (La Porta and Zapperi, 2017):

- *proto-oncogenes*, which shift from helping cell growth in normal conditions to promoting abnormal cell proliferation when they are over-present or hit by alterations;
- *oncosuppressors*, which stop to protect cell genome when mutated, by repressing cell cycle regulation and DNA repair system.

1.2.1 The hallmarks of cancer

To deep the insight, some sets of molecular, chemical and biological traits shared by the majority of cancers, called cancer hallmarks, have been identified. In Hanahan and Weinberg (2000) such essential features are listed to be:

- *Sustain of proliferative signaling:* the oncogenes simulate the effect of growth signals, which normally stimulate mitosis. This leads to an unregulated cancer cell proliferation due to self-sustained mechanisms (not to exogenous stimulation);
- *Insensibility to growth suppressor control:* cancer cells can deactivate growth-inhibitory signals, that normally block proliferation in order to guarantee a correct tissue homoeostasis;
- *Tissue invasion and metastasis:* angiogenic switch, whose task is regulating the process of blood vessels growth (*angiogenesis*, remains on, once activated, causing a continuous sprout of new vessels. Cancer cells then use these blood vessels (and the lymphatic ones) to flow towards new tissues, both adjacent and distant, causing tissue invasion and secondary neoplasms, which are responsible for 90% of cancer deaths;
- *Immortalization:* human cells have a self-regulated mechanism to limit the multiplication of cell clones, that induces the senescence, i.e. cell division ceasing. Cancer cells, not only disrupt cell-to cell proliferative signals, but also deactive the above surveillance mechanism, leading to an unlimited number of successive cell divisions;
- *Resistance to cell death:* cancer cells get more resistant to apoptosis, a self-programmed program which drives cells to death.

Recent studies (Hanahan and Weinberg, 2011) have improved the above list with the adding of two emerging hallmarks:

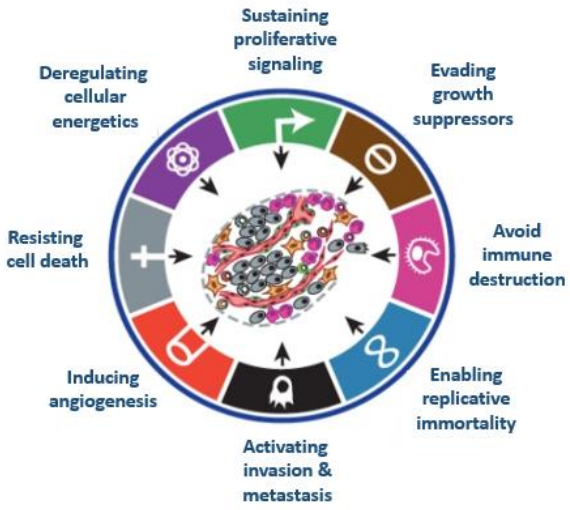


Figure 1.2: List of cancer hallmarks and characteristics.

- *Obstacle to immune destruction:* cancer cell succeed in limiting the detective action of immune system, which works to recognize and delete tumor formation;
- *Deregulation of cellular energy:* cancer cells can alter their energy acquisition in order to better fuel and feed tumor progression, by limiting the glucose metabolism to glycolysis.

and two enabling characteristics, which are essential conditions for theabove mentioned hallmarks:

- *Genome instability and mutation:* DNA stability, normally ensured by an efficient genome maintenance system, is altered by a succession of random genetic mutations having selective advantages in growth process;
- *Promotion of inflammation:* inflammatory response associated to cancer can fuel tumor evolution by increasing hallmarks capabilities through the supply of growth/survival/pro-angiogenic factors.

1.3 Ecological perspective in cancer dynamics

In the last few years, the ecology of cancer has emerged as a promising approach in cancer modeling (Altrock et al., 2015). In particular, it has been so far recognized the utility of regarding cancer as an evolutionary and ecological process that may be profitable described by tools coming from mathematical ecology (Durrett and Levin, 1994). This latter field, which borrows most of its approach from mechanical statistics (Azaele et al., 2016), represents a natural framework when one attempts to model the dynamics of large families of interacting units. Tumorgenensis, indeed, may be regarded as the result of a Darwinian struggle between interacting species (healthy and mutated cells) of an ecosystem, involved in exploitative interactions. Thus, the coexistence of several cell types, whose dynamics is driven by mutualistic ecological interactions and stochasticities, suggests that approaches from ecology, population dynamics and evolutionary game theory should be accounted for a reliable and highly detailed

cancer modeling.
However, even dropping such a high grade of magnitude from the dynamics description, interesting conclusions can be drawn by mean of these ecological approaches. Indeed, empirical observations and census data have revealed the outstanding emergence of regular and universal statistical features in ecosystems, despite of the contrasting biological conditions affecting them. By suitably simplifying above models disregarding biological details and stochasticities, a theoretical explanation for the arising of such common, global traits from few key features follows. Scientists refer to this theoretical idea as neutral theory. We will see that it will represent a fundamental hypothesis for the derivation of our statistical framework.

In what follows, we will present in details same quantitative models of ecological inspiration used in cancer dynamics at first, then our focus will be shifted to the neutral theory and its claiming.

1.3.1 Simple population models for cancer dynamics

Tumors are not homogeneous, but are composed by many genetically different cells whose own level of variability and fitness affect the evolution of the system and lead to cell competitions for space and resources. From an ecological point of view, such an evolutionary process can be look as an attempt of new species (cancer cells), having specific metabolic and proliferative strategies and rates, to overcome the resident species (somatic cells) and invade new habitat (tissue) (Kareva, 2011). Thus, competition approaches, mutuated from population dynamics and dealing with the architecture and the magnitude of the interaction networks, seem to mathematically well-suit such a dynamics.

Historically, the first models to treat species interaction in terms of interspecific competition are those of Lotka and Volterra. They are entirely based on Gause law, stating that two species, asking for the same limiting resources in a close environment not affected by external perturbations, can not coexist for a long time. If we named $N_1(t)$ the population size of specie 1 (which can be healty cells or cancer cells) at time t and $N_2(t)$ the one of species 2, then the simplest Lotka-Volterra dynamics is described by the equations

$$\begin{cases} \dot{N}_1(t) = s_1 N_1 \left(1 - \frac{N_1}{K_1} - \frac{b_{1,2}}{K_1} N_2\right) \\ \dot{N}_2(t) = s_2 N_2 \left(1 - \frac{N_2}{K_2} - \frac{b_{2,1}}{K_2} N_1\right) \end{cases} \tag{1.1}$$

For $i, j \in \{1, 2\}$, $s_i N_i$ describes the intrinsic population size growth with s_i meaning the available resources in the area inhabited by the species i, $\frac{N_i^2}{K_i}$, with K_i carrying capacity, accounts for the self-competition between the individuals of the same species and consequently avoids the population size to diverge, $-s_i \frac{b_{i,j}}{K_i} N_i N_j$ models how species i interacts with species j according to the sign and the magnitude of the interaction effect constant $b_{i,j}$. Depending on the sign of this latter three scenarios are plausible: cancer cells overcome the normal ones giving rise to a tumor mass, cancer cells are stopped in their progression (tumor mass is still present but it does not expand), cancer cells are reduced and driven to death. In the last two cases, we guess that interaction terms should account for drug benefits, which lead to the modification of the system interaction constants.
Mathematically speaking, by performing the following change of coordinates

$$x = \frac{N_1}{K_1}, \quad y = \frac{N_2}{K_2}, \quad \tau = s_1 t, \quad \rho = \frac{s_2}{s_1}, \quad \alpha_{1,2} = b_{1,2} \frac{K_2}{K_1}, \quad \alpha_{2,1} = b_{2,1} \frac{K_1}{K_2} \tag{1.2}$$

we find that the biological meaningful steady states of the system are:

- $A = (1, 0)$, which is stable and attractive for $\alpha_{2,1} > 1$ and occurs when only species 1 survives,
- $B = (0, 1)$, which is stable and attractive for $\alpha_{1,2} > 1$ and occurs when only species 2 survives,
- $C = \left(\frac{1-\alpha_{1,2}}{1-\alpha_{1,2}\alpha_{2,1}}, \frac{1-\alpha_{2,1}}{1-\alpha_{2,1}\alpha_{1,2}}\right)$, which is stable for $(\alpha_{1,2} < 1, \alpha_{1,2} < 1)$ and arises when the two species coexist.

A more current approach to the modeling of the struggle among different strategies is provided by the evolutionary game theory. Such a theory deals with the understanding of the long-term proliferation of well-mixed and infinite cells in the context of a game. In cancer biology, such a game can be formulated as a table that ascribes fitness values (pay-offs) to every pairwise interaction between cell phenotypes (strategies) (Altrock et al., 2015): cells with best strategies will spread in the population. Indeed, under normal conditions, cells in healthy tissue move toward an evolutionarily stable strategy (ESS). However, if one of the present cell types escapes its biological constraints and rates, the balance may be destroyed and one species may overcome the others. For a mathematical translation of the above evolutionary process, let us consider two population of cells with initial conditions on the abundance. Let A be the population of healthy cells, having an initial population equal to $N_A(0)$ and α as growth rate, and B be the population of tumor cells, with an initial cardinality of $N_B(0)$ and a growth rate equal to β. Under the assumption that the resource availability needs a constant population size, the dynamics of the two populations are respectively

$$\dot{N_A}(t) = N_A(t)(\alpha - \omega), \tag{1.3}$$

and

$$\dot{N_B}(t) = N_B(t)(\beta - \omega), \tag{1.4}$$

where ω is the average growth rate. Observe that if α or β is greater than the average rate, then the corresponding population will overgrow the other one. The constant-size for the whole population ($N_A(t) + N_B(t) = N_T = constant\ \forall t > 0$) leads to

$$N_T\omega = N_A(t)\alpha + N_B(t)\beta, \tag{1.5}$$

which enables us to rewrite the system of Eq.(1.3) and Eq.(1.4) into an unique dynamical equation:

$$\dot{N_A}(t) = N_A(t)(N_T - N_A(t))(\alpha - \beta). \tag{1.6}$$

If population size is $\gg 1$, we can convert the latter equation into another one concerning frequency x; that is the replicator equation

$$\dot{x}(t) = x(t)(1 - x(t))(\alpha - \beta). \tag{1.7}$$

Observe that, now, the growth rates α and β can be thought as a measure for fitness. Within this framework, an improvement of the description can be performed by accounting for the dependence of fitness from reproduction rates and population frequencies. Then, making the

substitutions $\alpha \to \phi_A(x)$ and $\beta \to \phi_B(x)$ to explicate the fitness-frequency dependence, we obtain the following mean-field inspired equation

$$\dot{x}(t) = x(t)(1 - x(t))(\phi_A(x) - \phi_B(x)). \tag{1.8}$$

After the computation of the pay-off matrix for $N_A \leftrightarrow N_B$ game, $\begin{bmatrix} p_{AA} & p_{AB} \\ p_{BA} & p_{BB} \end{bmatrix}$, where $p_{i,j}$, $i, j \in \{A, B\}$ is the output for a i cell type individual after its interaction with one of population j, we end up with the average fitness

$$\phi_A(x) = xp_{AA} + (1 - x)p_{AB}, \tag{1.9}$$

$$\phi_A(x) = xp_{BA} + (1 - x)p_{BB}, \tag{1.10}$$

that we can substitute in Eq.(1.8) to solve it.

Thus, if tumor cells adopt the best strategy in terms of resources exploitation and proliferation, then they will win the competition, leading to a successful invasion of tissue and to the formation of a primary tumor mass.
Once a tumor insurgence occurs other quantitative models help us to analyze and monitor the probabilistic growth of its size. In this context, the models providing the more realistic evolution are those capable of accounting for both demographic (population size) and environmental (drug effects) stochasticities that cancers undergo. Indeed, whereas the effect of demographic stochasticity decreases with population size, the effect of environmental factors does not, thus they should be included in the dynamical description (Haccou et al., 2005). Within this stochastic framework, the branching processes are among the most spread probabilistic tools to track the tumor growth. A branching process is a Markov process (see Chapter 2, Subsection 2.2.2 for details) characterized by the assumption that each cell events occur at given rates, independently of the population size or composition, or of the time. After a time equal to Δt, each cell is supposed to have generated a random number of offspring and, as mutations accumulate in tumor cells, each new cell typology has acquired new event rates (Altrock et al., 2015). Named n_1 the number of cells harbouring one mutation and n_2 the number of those harbouring two mutations, such properties can be translated into the following transition probabilities:

$$\begin{cases} P\left(n_1(t+\Delta t) = a+1, n_2(t+\Delta t) = b | n_1(t) = a, n_2(t) = b\right) \approx \lambda_1(1-u)a\Delta t \\ P\left(n_1(t+\Delta t) = a-1, n_2(t+\Delta t) = b | n_1(t) = a, n_2(t) = b\right) \approx \mu_1 a\Delta t \\ P\left(n_1(t+\Delta t) = a, n_2(t+\Delta t) = b+1 | n_1(t) = a, n_2(t) = b\right) \approx \left(\lambda_2 + \lambda_1 u\right) b\Delta t \\ P\left(n_1(t+\Delta t) = a, n_2(t+\Delta t) = b-1 | n_1(t) = a, n_2(t) = b\right) \approx \mu_2 b\Delta t \end{cases} \tag{1.11}$$

where λ_i and μ_i, $i \in \{1, 2\}$ are the birth and death rates for the two cell types respectively, while u is the mutation rate from the first cell type to the second.
Among branching processes, a particular interest has been paid to Moran process, which allows to incorporate a mutational fitness distribution in a constant-size population, instead of a fixed fitness value . Let us suppose that there are n species, $i \in \{1, 2, ..., n\}$, having abundance $N_1, N_2, ..., N_n$, with $N_1 + N_2 + N_n = N$ constant and be $f_1, f_2, ..., f_n$ their fitness values, respectively. Then, at each time step, a Moran process works by choosing a cell of type i for a single reproduction with a probability proportional to its fitness f_i and by removing randomly an individual-type, say j, from the population in order to keep the population size

constant. Thus, the transition probability for the individual-type i to increase and for the individual-type j to decrease is

$$P\left(N_i = N_i + 1, N_j = N_j - 1\right) = \frac{N_i f_i}{N_i f_i + N_j f_j} \frac{N_j}{N}. \tag{1.12}$$

1.3.2 Macroscopic and universal traits in ecological dynamics: the neutral theory

All the above models provide us with a reliable and detailed description of the evolution of stochastic populations affected by species interactions. Both evolutionary game theory and competitor approaches in cancer modeling strongly account for the architecture of cells interaction network. This latter naturally insert into the system description a sort of identification for the mutations: each of them need to be somehow labelled in order to assign the corresponding interaction rates and to model the interplay dynamics by mean of the effect that the presence of mutation i has on the behaviour of mutation j, for instance. Such a high grade of details is useful to track the evolutionary trajectory of the system, but is not treatable enough to explain the emergence of common traits that has been empirically observed. (Azaele et al., 2016). Indeed, from data collected in various ecosystems, the observation of similar macropatterns despite of the particularities of the systems has occurred. Thus, there has been the exigency for many researchers to understand this counterintuitive tendency. They have found that if one looks at the system dynamics backward, dropping both the stochastical differences occurring among mutations and the mutation indexing in favour of a demographic equivalence for all the mutations, then a justification of such a universal, statistical behaviour can be obtained. This is the framework of the neutral theory, an individual-based stochastic theory claiming at describing the population dynamics as based on random drift, with individuals in the community carrying the same per capita probabilities of giving birth, dying and speciating and having the derivation of several macro-ecological patterns from few fundamental mechanisms as major strength. (Azaele et al., 2016).

A pioneering attempt to explain the emergence of universal ecological traits was made by MacArthur and Wilson in (MacArthur and Wilson, 1967). They have observed the population dynamics at two different scales:

- local scale $\rightarrow$ *specie community*: a group of similar species competing for the same resources in a local area,
- global scale $\rightarrow$ *metacommunity*: a group of similar species living in a set of local communities,

and proposed a model in which species richness at local community scale is the result of both immigration of new species from the metacommunity and the extinction phenomena that the species may undergone. Moreover, they have described a dynamics not noticeably affected by competition: species with survival advantages are still present but are determined by random dispersal and stochastic local extinction. The role of the stochasticity in the dynamics, the balance between immigration and extinction in determining species abundance and the demographic equivalence of the species have represented a crucial departure from the previous competitive theories (Azaele et al., 2016).

The neutral theory has found its maximum formalization in *The Unified Neutral Theory of*

Biodiversity and Biogeography of Stephen Hubbell (Hubbell, 2001). Here, under the hypothesis of:

- *stochasticity:* species dynamics is driven by random events that lead to a birth and death process;
- *independence:* no interspecific interactions affect species relations;
- *neutrality:* no demographic differences occur among species individuals,

Hubbell has derived various macroscopic patterns, like the functional shape of the empirical global RSA (a static measure to describe the heterogeneity and richness of an ecosystem, i.e. the frequency of species at given abundance).
For instance, adding the assumptions of species independence and random spatial distribution for species to those of neutrality and setting the density of individuals in a ecological region equal to ρ, we have that the probability for a species to have abundance n in an area of size a is $\sim$ Poisson(ρa). To model metacommunity dynamics and derive a reliable RSA distribution, Hubbell has exploited Markovian modeling approaches, which generally enable to both insert and control the stochasticities and their effects.

Definition 1.1. A stochastic process $\{X(t)\}_{t\in\mathbb{R}^+}$ having support in ϵ is a continuous-time Markov chain if $\forall\ x_0, x_1, ..., x_n, x_{n+1} \in \epsilon$, $\forall\ 0 < t_1 < t_2 < ... < t_n < t_{n+1} \in \mathbb{R}^+$ the following identity (sometimes called "memorylessness") holds

$$P(X(t_{n+1}) = x_{n+1}|X(t_n) = x_n, ..., X(t_0) = x_0) = P(X(t_{n+1}) = x_{n+1}|X(t_n) = x_n) \quad (1.13)$$

Thus, let $P_{n,s}(t)$ be the probability for the species labeled with s, $s \in \{1, ..., S\}$, to have exactly n individuals at time t in the metacommunity. Then, such a state probability follows a birth and death dynamics governed by the following Kolmogorov forward equation

$$\begin{cases} \frac{\partial}{\partial t}P_{n,s}(t) = P_{n-1,s}(t)\, b_{n-1,s} + P_{n+1,s}(t)\, d_{n+1,s} - P_{n,s}(t)\, b_{n,s} - P_{n,s}\,(t)d_{n,s} & n \geq 1 \\ \frac{\partial}{\partial t}P_{0,s}(t) = -P_{0,s}\, b_{0,s} + P_{1,s}\, d_{1,s} & n = 0 \end{cases} \quad (1.14)$$

where $b_{n,s}$ and $d_{n,s}$ are birth and death rates for species s occurred n times, respectively. Observe that n=0 case is equivalent to set $b_{-1,s} = d_{0,s} = 0$, meaning that no births could occur if initial population is equal to -1 and no deaths are possible with null population.
Since the RSA is a static tool, we need to study the above system at the equilibrium. It becomes

$$\begin{cases} P_{n,s}\ (b_{n,s} + d_{n,s}) = P_{n-1,s}\, b_{n-1,s} + P_{n+1,s}\, d_{n+1,s} & n \geq 1 \\ P_{0,s}\, b_{0,s} = P_{1,s}\, d_{1,s} & n = 0 \end{cases} \quad (1.15)$$

Then, the steady-state probability is given by

$$P_{n,s} = P_{0,s} \prod_{i=0}^{n-1} \frac{b_{i,s}}{d_{i+1,s}}, \quad (1.16)$$

where $P_{0,s}$ must be selected equal to $\left(\sum_{n=1}^{\infty} \prod_{i=0}^{n-1} \frac{b_{i,s}}{d_{i+1,s}}\right)^{-1}$ to normalize.
Indeed, from the initial condition of the Kolmogorov system follows that

- $P_{1,s} = \frac{b_{0,s}}{d_{1,s}} P_{0,s}$.

Substituting this expression into the differential equation for $n = 1$, we end up with

- $P_{2,s} = \frac{P_{1,s}(b_{1,s}+d_{1,s}) - P_{0,s}\, b_{0,s}}{d_{2,s}} = \frac{P_{1,s}(b_{1,s}+d_{1,s}) - P_{1,s}\, d_{1,s}}{d_{2,s}} = \frac{b_{1,s}}{d_{2,s}} P_{1,s} = \frac{b_{0,s}\, b_{1,s}}{d_{1,s}\, d_{2,s}} P_{0,s}$.

This suggests that we might have $P_n = \frac{b_{n-1,s}}{d_{n,s}} P_{n-1}$ for $n \geq 1$. Assuming this, we have then

$$P_{n+1} = \frac{(b_{n,s} + d_{n,s})P_{n,s} - b_{n-1,s}P_{n-1,s}}{d_{n+1,s}} = \frac{(b_{n,s} + d_{n,s})P_{n,s} - d_{n,s}P_{n,s}}{d_{n+1,s}} = \frac{b_{n,s}}{d_{n+1,s}} P_{n,s}. \quad (1.17)$$

Therefore, substituting recursively Eq.(1.16) arises. as wanted.
At this point, observe that the demographic equivalence for species, driven by the assumption of the neutral theory, allows us to drop the s index in Eq.(1.16). In other word, we can focus on the probability that a given species - do not care which, they are supposed to be identical - has abundance n. Now, depending on the functional form of $b_{n,s}$ and $d_{n,s}$, various distribution for $P_{n,s}(t)$ at metacommunity scale can be worked out. In his work, Hubbell has made the ecologically meaningful selection of:

$$b^M_{n,s} = b_s\, n + \delta_{n,0}\nu \quad \text{as the birth rate,} \quad (1.18)$$

$$d^M_{n,s} = d_s\, n \quad \text{as the death rate} \quad (1.19)$$

with the additional boundary condition term $\delta_{0,s} = \nu$, which ensures to work with a community of size at least one. Such a choice, for which the system is governed by ecological drift and random speciation, leads to an empirically reliable Log-Fisher distribution for RSA.
Indeed, by naming $x_s = \frac{b_s}{d_s} \in (0, 1)$ and substituting the rates into Eq.(1.16), the steady-state solution is

$$P_{n,s} = P_{0,s} \frac{\nu}{b_s} \frac{x_s^n}{n}. \quad (1.20)$$

The $P_{0,s}$ constant can be actually be determined by imposing

$$1 = \sum_{n=1}^{\infty} P_{n,s} = P_{0,s} \frac{\nu}{b_s} \sum_{n_1}^{\infty} \frac{x_s^n}{n} = P_{0,s} \frac{\nu}{b_s} [-\log(1 - x_s)], \quad (1.21)$$

where special series $\sum_{i=1}^{\infty} \frac{x^i}{i} = \frac{1}{\log(1-x)}$ with $|x| \geq 1, x \neq 1$ has been used in the last equality. Thus, it results that $P_n = P(n) \sim 1-$normalized Log-Series distribution with probability mass function given by

$$P(n) = -\frac{1}{\log(1 - x)} \frac{x^n}{n}. \quad (1.22)$$

In the next chapter, which concerns the mathematical core of the thesis, we will derive the statistical distribution chosen for our method in the same manner of above using suitable and ecological-driven birth and death rates.

To complete the description of the macro-patterns resulting from the unified neutral theory, Hubbell has treated the ecological dynamics at the local scale, i.e. local communities, as well. Under the following assumptions:

- the timescale is faster at the local scale than at the global one;

- migration and random drift only govern the dynamics;

he has described the dynamics as based on the following rules:

- **Initialization.** Have a local species community dipped in a metacommunity representing an infinite pool of species.
- Two different pathways are now possible.
  RULE 1
  - **Step 1** → *Birth and death events:* with a probability of $1 - m$, select randomly two individuals from local community: if they belong to the same species go back to initialization, otherwise go to step 2.
  - **Step 2** → with a probability of $\frac{1}{2}$, remove one individual between the two selected and replace it with an immediately-mature individual of the other species.

  RULE 2
  - **Step 1** → *Death and migration events:* with a probability of m, pick randomly an individual and go to step 2.
  - **Step 2** → with a probability of $\frac{1}{2}$, remove the individual and substitute it with one belonging to a species present in the metacommunity randomly chosen according to a probability proportional to its abundance at global scale.

Volvok has been managed to translate such rules into the following analytical birth and death rates for the k species (Volkov et al., 2013):

$$b^L_{n,k} = (1-m)\frac{n}{S^L}\frac{S^L-n}{S^L-1} + m\frac{n^M_k}{S^M}\left(1-\frac{n}{S^L}\right), \tag{1.23}$$

$$d^L_{n,k} = (1-m)\frac{n}{S^L}\frac{S^L-n}{S^L-1} + M\left(1-\frac{n^M_k}{S^M}\right)\frac{n}{S^L}, \tag{1.24}$$

where n^M_k is the abundance of k species at metacommunity scale, S^L is the number of species in the local community, whereas S^M is the one in the metacommunity. Observe that the first terms of both Eq.(1.23) and Eq.(1.24) account for the first rule with a k-species birth related to a different-species death, while the second ones regard second rules with an increasing of k-species due to migration events from metacommunity. Substituting Eq.(1.23) and Eq.(1.24) into Eq.(1.16) we end up with the expression for RSA at local community scale:

$$P_{n,k} = \frac{S^L!}{n!(J-n)!}\cdot\frac{\Gamma(n+\lambda_k)}{\Gamma(_k)}\cdot\frac{\Gamma(\theta_k-n)}{\Gamma(\theta_k-S^L)}\cdot\frac{\Gamma(\lambda_k+\theta_k-S^L)}{\Gamma(\lambda_k+\theta_k)} \tag{1.25}$$

with

$$\lambda_k = \frac{m}{1-m}(S^L-1)\frac{n^M_k}{S^M}, \tag{1.26}$$

$$\theta_k = S^L + \frac{m}{1-m}(S^L-1)\left(1-\frac{n^M_k}{S^M}\right). \tag{1.27}$$

Neutral theory in genomics

For the genetic purposes of the present thesis we need to check if neutral theory holds in genomics field as well. Indeed, the ecological-inspired statistical framework we aim at adapting is heavily based on neutrality. Thus, to guarantee the same mathematical formalism and exploit a similar ensemble approach, neutral assumptions of demographic equivalence and stochastic effects must be satisfied. From actual observations and experiments, the biologist M. Kimura (Kimura, 1986) has confirmed these hypothesis hold even at DNA level, stating that

- almost all mutations occurring at DNA molecular level depend on random fixation and neutral mutant selectivity and not on Darwinian species selection acting on best-competitor-mutants;
- natural selection still affects genetic evolution, but only a small percentage of DNA changes is adaptive in nature;
- most of intraspecific variability present at molecular level is neutral. Thus, DNA polymorphic alleles are kept in the species by some form of balancing selection involving mutational input and random extinction.

Thus, the neutral theory properties in genomics are empirically satisfied and consequently we can be legitimated to assume them as basic hypothesis for our framework.

Chapter 2

Statistical inference of unseen variants

In this chapter we present in details the upscaling method we will develop on, providing information on the statistical distribution used as well. We have adapted a statistical framework firstly implemented for the estimation of the species richness in the Amazonian forests (Tovo et al., 2017). Tested on both simulated and real forests, such a method gave back the best results among those given by various other ecological estimators, thus we have selected it for our piece of work.

To describe and visualize the heterogeneity and the spread of mutations over desired unities (tumor cells, tumor biopsies or individuals' sequenced sections of DNA), *Species-Abundance Distribution* (SAD) pattern may be profitably used in genomics. Such a curve provides quantitative information on the number of mutations occurring in exactly n unities, with $n \in \{1, 2, 3, ...\}$, i.e. it lists the observable mutations within a tumour region along with the number of occurrences for mutations.
Unfortunately, in clinical applications the medical ethics and the high costs of sequencing impose for this curve to be measured locally only (for example taking into account biopsies that cover just a fraction of the whole tumor or sequencing only few people). Therefore, one only has information on mutation occurrences in a small region. An upscaling method consists on inferring the SAC and estimating the total number of occurred mutations, S, at the global scale. An object of interest for our statistical approach, strongly connected to the SAD, is the ecologically-borrowed *Relative Species Abundance* distribution (RSA), which is the probability for a mutation to occur exactly n times.

Hereafter the RSA at global scale is postulated to be distributed according to

- a Negative Binomial (NB) of parameters $r > 0$ and $0 < \xi < 1$, whose probability mass function is

$$P(n|1) = \underbrace{\frac{1}{1-(1-\xi)^r}}_{:=c(r,\xi)} \underbrace{\binom{n+r-1}{n} \xi^n (1-\xi)^r}_{:=P(n|r,\xi)} \qquad n \geq 1$$

  or

- an extended Negative Binomial of parameters $r \in (-m, -m+1)$ with $m \in \mathbb{N} \setminus \{0\}$ and $0 < \xi < 1$, whose probability mass function is

$$P(n|1) = \underbrace{\frac{1}{1-(1-\xi)^r \cdot \sum_{j=0}^{m-1} \binom{j+r-1}{j} \xi^j}}_{:=\tilde{c}(r,\xi)} \binom{n+r-1}{n} \xi^n (1-\xi)^r \qquad n \geq m$$

depending on the r parameter value.
Notice that in the first case the support is set to start from one because only mutations with abundance at least one can be seen and counted, whereas in the second case the support needs to start from m to have the well-posedness of both binomial coefficients.
Moreover, given that binomials can be written in terms of Gamma functions, as

$$\binom{n+r-1}{n} = \frac{\Gamma(n+r)}{\Gamma(r)\Gamma(n+1)}, \tag{2.1}$$

it follows that we need to split the negative domain for r into $(-m, -m+1)$ open intervals because the Gamma function, having singularities in 0 and in $\mathbb{Z}^-$, is not defined in the whole negative real line. For the extended Negative Binomial, we have developed the general theory for r parameter lying in any negative intervals $(-m, -m+1)$, even if in the datasets analyzed in Chapter 3 and Chapter 4 only the $m = 1$ formula has emerged. It can be proved that such a formula coincides with the analytical expression of the 1-normalized Negative Binomial with $r > 0$.

Remark 1. To streamline the notations, in the whole dissertation we have preferred to write the probability mass functions using binomials instead of Gamma functions. Moreover, for the same purpose, we will refer to the extended Negative Binomial of parameters $r \in (-m, -m+1)$ and ξ as Negative Binomial with $r \in (-m, -m+1)$ or as m-extended Negative Binomial.

The choice of the statistical framework $\{(r, \xi, S), r \in \mathbb{R} \setminus \{0\}, 0 < \xi < 1, S \in \mathbb{N}\}$ for our model has been driven by both biological and mathematical reasons that are explained in the next section, which closely follows Tovo et al. (2017) in the model derivation and justifications.

2.1 Negative Binomial as steady-state distribution of a birth and death dynamics

To start off, it can be proved that the Negative Binomial arises naturally as the steady-state distribution of system undergoing a birth and death dynamics having specific event rates that weakly account for interspecific interactions. As first approximation, cancer cells evolution - as any population growth - can be profitably modeled by such a dynamics due to its proliferative mechanism.

Definition 2.1. Let be $X(t)$ a birth and death process, i.e. the variable describing the temporal evolution of the process, having birth and death rates equal to λ_n and μ_n, respectively. Then, $X(t)$ is a continuous-time Markov chain with support in $\mathbb{N}$ if it satisfies the following properties:

- $P(X(t+h) - X(t) = 1 | X(t) = n) = \lambda_n h + o(h)$
- $P(X(t+h) - X(t) = -1 | X(t) = n) = \mu_n h + o(h)$
- $P(|X(t+h) - X(t)| \geq 2 | X(t) = n) = o(h)$

In other words, a birth and death process is a particular Markov process in which jumps between neighboring states only are allowed. For our genetic purposes, it means that the population size can only increase by one or decrease by one at each step. The probability of transition from state n to state $n+1$ is given by $\lambda_n \Delta t + o(\Delta t)$, while the probability of moving from state n to state $n-1$ is given by $\lambda_n \Delta t + o(\Delta t)$ in an infinitesimal time.

Assumed that our mutational spectrum is composed by S independent mutations, let us name $P_{n,s}(t)$ the probability for the mutation s, $s \in \{1, 2, ..., S\}$, to have abundance n, i.e. it occurs in n samples, at time t. Then, within an evolutionary ecological perspective one may assume that the probability function follows the same dynamics of Chapter 1, Subsection 1.3.2.
Thus, the master equation for $P_{n,s}$ is

$$\frac{\partial}{\partial t} P_{n,s}(t) = P_{n-1,s}(t)\, b_{n-1,s} + P_{n+1,s}(t)\, d_{n+1,s} - P_{n,s}(t)\, b_{n,s} - P_{n,s}\,(t) d_{n,s} \tag{2.2}$$

where $b_{n,s}$ and $d_{n,s}$ are the birth and death rate for mutation s occurred n times, respectively. The formula at the equilibrium is

$$P_{n,s} = P_{0,s} \prod_{i=0}^{n-1} \frac{b_{i,s}}{d_{i+1,s}} \tag{2.3}$$

again, with $P_{0,s}$ normalization constant. In order to deal with a Negative Binomial distribution we need to set the following biological inspired birth and death rates:

- $b_{n,s} = b_s(n + r_s)$,
- $d_{n,s} = d_s\, n$,

where b_s and d_s are density independent per-capita birth and death rates, while r_s is a clustering parameter that accounts for intraspecfic interactions or migrations. By substituting relations above into Eq.(2.3) and setting $\xi_s = \frac{b_s}{d_s}$ we end up with

$$\begin{aligned} P_{n,s} = P_{0,s} \prod_{i=0}^{n-1} \frac{b_{n,s}}{d_{n,s}} &= \xi_s^n \left[r_s \cdot \frac{1 + r_s}{2} \cdot \frac{2 + r_s}{3} \cdot ... \cdot \frac{n - 1 + r_s}{n} \right] \\ &= \frac{(r_s + n - 1)!}{(r_s - 1)! n!}\, \xi_s^n = \binom{n + r_s - 1}{n} \xi_s^n, \end{aligned} \tag{2.4}$$

where the definition of rising factorial is used to prove the third equality.
The constant $P_{0,s}$ can be determined through the following non-zero abundance normalization

$$1 = \sum_{n=1}^{\infty} P_{n,s} = P_{0,s} \sum_{n=1}^{\infty} \binom{n + r_s - 1}{n} \xi_s^n = P_{0,s} \frac{1 - (1 - \xi_s)^{r_s}}{(1 - \xi_s)^{r_s}}. \tag{2.5}$$

Therefore, at the equilibrium, the probability for the th mutation to have n occurrences, i.e. the RSA, is given by a Negative Binomial of parameters $(r_s > 0, 0 < \xi_s < 1)$ and normalization to one:

$$P_{n,s} = \underbrace{\frac{1}{1-(1-\xi_s)^{r_s}}}_{:=c(r_s,\xi_s)} \underbrace{\binom{n+r_s-1}{n} \xi_s^n \, (1-\xi_s)^{r_s}}_{:=P(n|r_s,\xi_s)} . \tag{2.6}$$

Within the neutral hypothesis, all mutations have the same probability of proliferating, dying and speciating, so that s index can be neglected from the Eq.(2.6).

In particular, observe that the birth rate giving a Negative Binomial distribution and the one giving a Log-Fisher (see Eq.(1.18) and Eq.(1.19)) differ only for the presence of the cluster parameter r_s. Thus, as we will analytically prove, Log-Fisher can be thought as a special case of a Negative Binomial distribution having parameter r going to 0.

Finally, corresponding SAD can be found through

$$SAD(n) = \mathbb{E}[S_1(n)] = \sum_{n=1}^{\infty} P_{n,s} = S \cdot P_n, \tag{2.7}$$

where $S_1(n)$ is the number of mutations occurring n times at the global scale $p = 1$ and where s label has been removed. One then ends with the functional shapes for the SAD given by:

$$SAD(n) = S \cdot c(r,\xi) \binom{n+r-1}{n} \xi^n (1-\xi)^r \quad \text{with } n \geq 1$$

if the RSA $\sim$ 1-normalized Negative Binomial of parameters $r > 0$ and ξ,

$$SAD(n) = S \cdot \alpha(x) \frac{x^n}{n} \quad \text{with } n \geq m$$

if the RSA $\sim$ 1-normalized Log Series of parameter α.

2.2 Form flexibility of Negative Binomial

It can be showed that the Negative Binomial, both with $r > 0$ and $r \in (-m, -m+1)$, displays different functional shapes according to the values assumed by the parameters. Such a flexibility is resulted to well accommodate our genetic-inspired RSAs and SADs. Indeed, we guess these latter may have bumps if a set of mutations is common at the considered scale or may have an initial boost followed by a noticeable decreasing if hyper-rare mutations are dominant. Then, we need a distribution versatile enough to cover all these possibilities, i.e. showing at least modal and power law behaviours that would well capture the described frameworks.

2.2.1 Functional shapes accommodated by a single Negative Binomial

Proposition 2.2.1. *The tail of a Negative Binomial of parameter r and ξ goes like a power law x^δ of exponent $\delta = -1 + r$ with an exponential cut-off weighted by ξ.*

Proof. From the probability mass function of the Negative Binomial, we deduce that its behavior depends on the binomial coefficient, which can be written in term of the Gamma function (see Eq.(2.1)). An asymptotic approximation (Flajolet and Robert Sedgewick, 2009) for the Gamma function is

$$\lim_{n\to\infty} \frac{\Gamma(n+c)}{\Gamma(n)n^c} = 1 \quad \forall c \in \mathbb{C}. \tag{2.8}$$

Then, when $n \gg 1$ binomial coefficient becomes

$$\binom{n+r-1}{n} = \frac{\Gamma(n+r)}{\Gamma(n+1)\Gamma(r)} = \frac{\Gamma(n+r)}{n\Gamma(n)\Gamma(r)} \sim \frac{1}{\Gamma(r)} n^{-1+r}. \tag{2.9}$$

Therefore, asymptotically, a Negative Binomial behaves like

$$k \cdot n^{-1+r} e^{n \ln(\xi)}, \tag{2.10}$$

where k is a constant depending on r and ξ parameters. □

The proposition can be analytically proved for $n \gg 1$ only. By graphically observing that a Negative Binomial of parameters (r, ξ) and a power law of exponent $-1 + r$ match from small n, one concludes that the relationship holds for all $n > 0$.

It follows that different shapes for Negative Binomial SADs are displayed according to the r domain.

- Case: $r > 1$. When r parameter runs in $(1, \infty)$, the 1-normalized Negative Binomial distribution we would deal with shows an unimodal mode. Its peak acquires the characteristic shape and gets more noticeable as $r \gg 1$.

- Case: $r < 0$. When $r < 0$, the m-extended Negative Binomial we would have display a power law behaviour, with lower exponent as r gets more negative.

- Case: $r \to 0$. When $r \to 0$, a special case of 1-normalized Negative Binomial, that is the 1-normalized Log-Series, arises;

  To justify this latter affirmation we have the following:

  Proposition 2.2.2. *In the limit $r \to 0$, a 1-normalized Negative Binomial of parameters $r > 0$ and ξ becomes a 1-normalized Log-Series distribution of parameter α with $\alpha = \frac{1}{\log(1-\xi)}$, $\xi = \frac{b(n+r)}{dn}$.*

  Proof. When $r \approx 0$ we have

$$\begin{aligned} \binom{n+r-1}{r} &= \frac{\Gamma(n+r)}{\Gamma(n+1)\Gamma(r)} \\ &= \frac{(n+r-1)(n+r-2)\cdots r}{n!} \\ &= \frac{r}{n}\frac{n+r-1}{n-1}\frac{n+r-2}{n-2}\cdots\frac{r+1}{1}. \end{aligned} \tag{2.11}$$

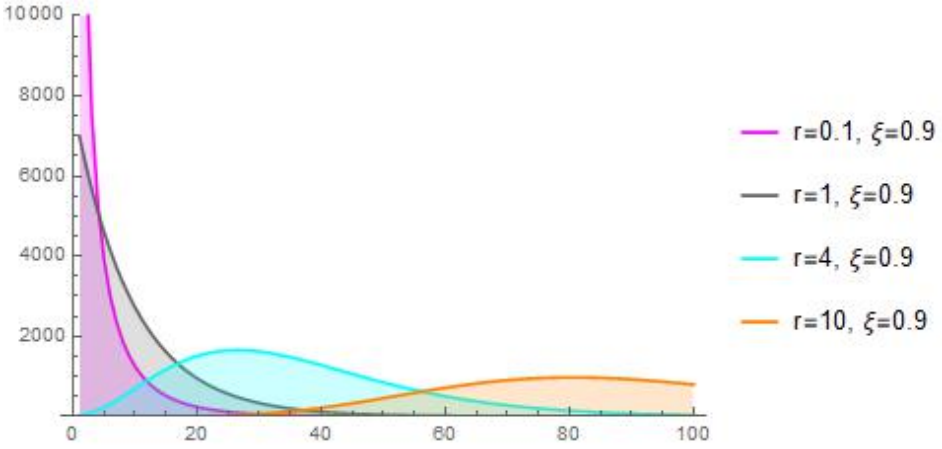


Figure 2.1: Single Negative Binomial SADs with different $r > 0$ and fixed $\xi = 0.9$.

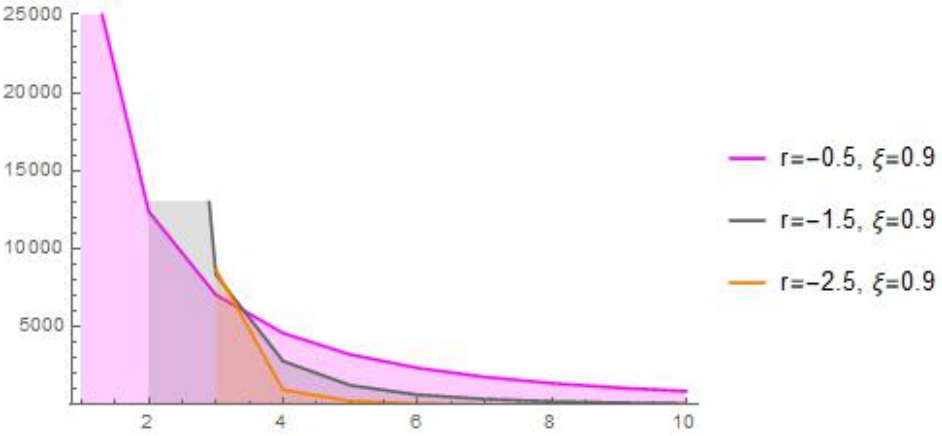


Figure 2.2: 1-, 2-, 3-extended Negative Binomial SADs with different $r < 0$ and fixed $\xi = 0.9$.

By substituting into the limit we end up with

$$\begin{aligned}\lim_{r\to 0} c(r,\xi)\binom{n+r-1}{r}\xi^n(1-\xi)^r &= \lim_{r\to 0}\frac{(1-\xi)^r}{1-(1-\xi)^n}\binom{n+r-1}{n}\xi^n \\ &= \lim_{r\to 0} -\frac{\frac{1-\xi)^r}{r}}{\frac{(1-\xi)^r-1}{r}}\frac{r}{n}\frac{n+r-1}{n-1}\frac{n+r-2}{n-2}\cdots\frac{r+1}{1} \\ &= -\underbrace{\frac{1}{\log(1-\xi)}}_{:=\alpha}\frac{\xi^n}{n}.\end{aligned} \tag{2.12}$$

which is the form of a 1-normalized Log-Series distribution of parameter α. □

2.3 Definition of form invariance property

In what follows, we give the definition of form invariance and we prove that Negative Binomial distribution arising both when $r > 0$ and $r \in (-m, -m+1)$ satisfies it. The turning idea of our method consists on exploiting such a property of the Negative Binomial distribution to link local information to global one and deduce the unknown parameters from such a relationship.

For us, form invariance means that if we binomially subsampling a Negative Binomial, new data are still described via a Negative Binomial. That is, the RSA at any local scale is a Negative Binomial if the RSA at global scale is a Negative Binomial.
More technically, if we have a binomially distributed conditional probability $P(k|n,p)$ for a mutation to occur k times in a sample of size p, given that it occurs n times at global scale, then the formalism to derive the emergence of a Negative Binomial distribution even at local scale is guaranteed. To achieve such a binomially conditional probability we need to work under the *mean-field hypothesis*. This latter deals with the absence of spatial correlation due to both interspecific and intraspecific interactions in the sampling and with the demographic equivalence of species. Observe that from the lack of anisotropies and inhomogeneties in the region of interest follows that the probability for a mutation to occur in a region sampling of area $a = pA$, where A is the total area, is exactly p, i.e. the probability is proportional to the sampling size. Moreover, the demographic equivalence of mutations, resulting from the setting of similar event rates, allows to have the above probability equal for each mutations. To visualize more clearly the *mean field hypothesis* suppose to partition the whole genetic region of interest into several units having the same area and to throw the similar supposed mutations in the region according to a Poisson process, in the same way we would throw balanced marbles. Then, each "balanced" mutation has the same chance to lie in a specific unit which is proportional to the size of the unit itself. Under these assumptions, the sampling is binomial, i.e. the conditional probability $P(k|n,p)$ is binomially distributed with parameters (n,p)

$$P_{binom}(k|n,p) = \begin{cases} \binom{n}{k} p^k (1-p)^{n-k} & n \geq k \\ 0 & n < k \end{cases} \tag{2.13}$$

and the form invariance property for Negative Binomial can be derived. In our genetic framework, the drop of mutation identities and stochastic differences follows from the neutral theory which we have already proved to be reliable in genomics. To achieve the hypothesis on spatial uncorrelation as well, we need to check the level of homogeneity characterizing the genetic region analyzed. We guess that in our datasets such a level guarantees a sufficient uncorrelation magnitude to satisfy the assumption. Observe that in case of inhomogeneties, a larger number of located samples is needed to cover all possible thickenings and to work with a statistical reliable sampling.

Before providing an analytical proof of the form invariance property of the Negative Binomial resulting from binomial sampling of data, we make an important observation.
Remark 2. In statistics, a widely mentioned property for distributions is the scale invariance property. Reader should pay attention to not confuse the two properties, the form invariance one and the scale invariance one, since they have different statistical meanings.
Indeed, a distribution f is said to be scale invariant if it satisfies

$$f(\lambda x) = g(\lambda) f(x). \tag{2.14}$$

In this respect, it holds:

Proposition 2.3.1. *A distribution is scale invariant if and only if it is a power law.*

Proof. ($\Leftarrow$) Be $f(x) = C \cdot x^{-\alpha}$ a power law. Then, it follows

$$f(\lambda x) = C(\lambda x)^{-\alpha} = \lambda^{-\alpha} \cdot C \cdot x^{-\alpha} = \lambda^{-\alpha} f(x). \tag{2.15}$$

($\Rightarrow$) Be $f(x)$ such that $f(\lambda x) = g(\lambda) f(x)$. Computing the previous formula in $x = 1$ gives $g(\lambda) = \frac{f(\lambda)}{f(1)}$. By substituting the multiplicative constant into main expression we deal with

$$f(\lambda x) = \frac{f(\lambda)}{f(1)} f(x), \tag{2.16}$$

whose derivative with respect to λ is

$$x f^{'}(\lambda x) = \frac{f^{'}(\lambda)}{f(1)} f(x) \tag{2.17}$$

If we compute this latter in $\lambda = 1$, we end with the differential equation

$$f(x) = f(1) x^{\frac{f(1)}{f^{'}(1)}} \tag{2.18}$$

whose only solution is a power-law distribution. □

2.3.1 Proof of form invariance property for 1-normalized Negative Binomial with $r > 0$

Lemma 2.3.2. *Named $P(n|1)$ the probability for a mutation to occur $n \geq 1$ times at the global scale, i.e. the global RSA, and $P(k|n,p)$, with $p \in (0,1)$, the conditional probability that a variant appears in k samples at the local scale p, given that it occurs in n at the global scale, let us assume that*

- $P(n|1) \sim$ *1-normalized Negative Binomial* $(r > 0, \xi \in (0,1))$,
- $P(k|n,p) \sim$ *Binomial* $(n, p \in (0,1))$.

Then, the RSA at the local scale p, $P_{sub}(k|p)$, is a 1-normalized Negative Binomial for $k \geq 1$ as well, with parameters $(\hat{r}_p, \hat{\xi}_p)$ given by

$$\begin{cases} \hat{\xi}_p = \frac{p\xi}{1-\xi(1-p)} \\ \hat{r}_p = r \end{cases} \tag{2.19}$$

Proof. First, let $k \geq 1$.

$$
\begin{aligned}
P_{sub}^{NB}(k|p) &= \sum_{n\geq k} P(k|n,p)\cdot P(n|1) \\
&= \sum_{n\geq k} \binom{n}{k} p^k (1-p)^{(n-k)} \cdot \underbrace{\frac{1}{1-(1-\xi)^r}}_{c(r,\xi)} \binom{n+r-1}{n} \xi^n (1-\xi)^r \\
&= (\xi p)^k (1-\xi)^r c(r,\xi) \sum_{n\geq k} \binom{n}{k} \binom{n+r-1}{n} (\xi(1-p))^{n-k} \\
&= (\xi p)^k (1-\xi)^r c(r,\xi) \sum_{n\geq k} \underbrace{\frac{r\cdots(r+k-1)}{k!}}_{\binom{k+r-1}{k}} \cdot \frac{(r+k)\cdots(r+n-1)}{(n-k)!} \cdot (\xi(1-p))^{n-k} \\
&= (\xi p)^k (1-\xi)^r c(r,\xi) \binom{k+r-1}{k} \sum_{l\geq 0} \frac{(\xi(1-p))^l}{l!} (r+k)\cdots(r+l+k-1) \\
&= (\xi p)^k (1-\xi)^r c(r,\xi) \binom{k+r-1}{k} \sum_{l\geq 0} (\xi(1-p))^l \binom{r+l+k-1}{l} \\
&= c(r,\xi) \binom{k+r-1}{k} \underbrace{\left(\frac{p\xi}{1-\xi(1-p)}\right)^k}_{:=(\hat{\xi}_p)^k} \underbrace{\left(\frac{1-\xi}{1-\xi(1-p)}\right)^r}_{:=(1-\hat{\xi}_p)^r}
\end{aligned}
\tag{2.20}
$$

where the special series $\sum_{i=0}^{\infty} \binom{i+m}{i} x^i = \frac{1}{(1-x)^{m+1}}$ has been used in the last equality.
Now let $k = 0$. By taking the complementary, we have

$$
\begin{aligned}
P_{sub}^{NB}(0|p) &= 1 - \sum_{k\geq 1} P_{sub}^{NB}(k|p) \\
&= 1 - c(r,\xi) \cdot \sum_{k=1}^{\infty} \underbrace{\binom{k+r-1}{k} \hat{\xi}_p^k (1-\hat{\xi}_p)^r}_{P^{NB}(k|r,\hat{\xi}_p)} \\
&= 1 - \frac{c(r,\xi)}{c(r,\hat{\xi}_p)}
\end{aligned}
\tag{2.21}
$$

where the normalization $\sum_{k=1}^{\infty} c(r,\hat{\xi}_p)\cdot P(k|\hat{r}_p = r, \hat{\xi}_p) = 1$ has been exploited.
The local RSA is then

$$
\begin{aligned}
P(k|p) &= \frac{P_{sub}^{NB}(k|p)}{\sum_{k'=1}^{\infty} P_{sub}^{NB}(k'|p)} \\
&= \frac{c(r,\xi)\cdot P^{NB}(k|r,\hat{\xi}_p)}{\sum_{k'=1}^{\infty} c(r,\xi)\cdot P^{NB}(k'|r,\hat{\xi}_p)} \\
&= c(r,\hat{\xi}_p)\cdot P^{NB}(k|r,\hat{\xi}_p)
\end{aligned}
\tag{2.22}
$$

□

To stretch out the result: starting with a global RSA distributed according to a Negative Binomial distribution, a binomially sampling of any size can still be statistically described via a Negative Binomial having coefficients depending on the global scale parameters and on the sampling area through an explicit formula.

2.3.2 Proof of form invariance property for m-extended Negative Binomial with $r \in (-m, -m+1)$

Lemma 2.3.3. *Let $P(n|1)$ be the probability for a mutation to occur $n \geq m$ times at the global scale, i.e. the global RSA, and $P(k|n,p)$, with $p \in (0,1)$, the conditional probability that a variant appears in $k \geq m$ samples at the local scale p, given that it occurs in n at the global scale. Let us now assume that*

- *$P(n|1) \sim$ m-normalized Negative Binomial $(-m < r < -m+1$ with $m \in \mathbb{N} \setminus 0, \xi \in (0,1))$,*
- *$P(k|n,p) \sim$ Binomial $(n, p \in (0,1))$.*

Then, the RSA at the local scale p, $P_{sub}(k|p)$, is a m-normalized Negative Binomial for $k \geq m$ as well, with parameters $(\hat{r}_p, \hat{\xi}_p)$ given by

$$\begin{cases} \hat{\xi}_p = \frac{p\xi}{1-\xi(1-p)} \\ \hat{r}_p = r. \end{cases} \tag{2.23}$$

Proof. The probability mass function of Negative Binomial when $r > 0$ and when $r \in (-m, -m+1)$ differs only for the normalization constant. For the m-extended Negative Binomial arising when $r \in (-m, -m+1)$ such a constant is

$$\tilde{c}(r,\xi) = \frac{1}{1-(1-\xi)^r \cdot \sum_{j=0}^{m-1} \binom{j+r-1}{j} p^j}. \tag{2.24}$$

When $k \geq m$ computations follow closely those of the proof of Lemma 2.3.2. We have:

$$\begin{aligned} P_{sub}^{m-NB}(k|p) &= \sum_{m \leq k \leq n} P(k|n,p) \cdot P(n|1) \\ &= \tilde{c}(r,\xi) \binom{k+r-1}{k} \underbrace{\left(\frac{p\xi}{1-\xi(1-p)}\right)^k}_{(\hat{\xi}_p)^k} \underbrace{\left(\frac{1-\xi}{1-\xi(1-p)}\right)^r}_{(1-\hat{\xi}_p)^r} \end{aligned} \tag{2.25}$$

and

$$\begin{aligned} P_{sub}^{m-NB}(0|p) + P_{sub}^{m-NB}(1|p) + ... + P_{sub}^{m-NB}(m-1|p) &= 1 - \sum_{k \geq m} P_{sub}^{m-NB}(k|p) \\ &= 1 - \sum_{k \geq m} \tilde{c}(r,\xi) \underbrace{\binom{k+r-1}{k} (\hat{\xi}_p)^k (1-\hat{\xi}_p)^r}_{P^{m-NB}(k|r,\hat{\xi}_p)} \\ &= 1 - \frac{\tilde{c}(r,\xi)}{\tilde{c}(r,\hat{\xi}_p)}. \end{aligned} \tag{2.26}$$

Finally, the local RSA at scale p is

$$\begin{aligned} k \geq m: \qquad P(k|p) &= \frac{P_{sub}^{m-NB}(k|p)}{\sum_{k'=m}^{\infty} P_{sub}^{m-NB}(k'|p)} \\ &= \frac{\tilde{c}(r,\xi) \cdot P^{m-NB}(k|r,\hat{\xi}_p)}{\sum_{k'=m}^{\infty} \tilde{c}(r,\xi) \cdot P^{m-NB}(k'|r,\hat{\xi}_p)} \\ &= \tilde{c}(r,\hat{\xi}_p) \cdot P^{m-NB}(k|r,\hat{\xi}_p). \end{aligned} \tag{2.27}$$

□

2.3.3 Proof of the form invariance property for 1-normalized Log-Series with parameter α

Lemma 2.3.4. *Let $P(n|1)$ be the probability for a mutation to occur $n \geq 1$ times at the global scale, i.e. the global RSA, and $P(k|n,p)$, with $p \in (0,1)$, the conditional probability that a variant appears in k samples at the local scale p, given that it occurs in n at the global scale, again. Suppose that*

- $P(n|1) \sim$ *1-normalized Log-Series* $(\alpha(x))$,
- $P(k|n,p) \sim$ *Binomial* $(n, p \in (0,1))$.

Then, the RSA at the local scale p, $P_{sub}(k|p)$, is a 1-normalized Log-Series for $k \geq 1$ as well, with parameter

$$\alpha(\hat{x}_p) = \frac{1}{\log(1-\hat{x}_p)} \quad \textit{where} \quad \hat{x}_p = \frac{px}{1-x(1-p)}. \tag{2.28}$$

Proof. First, be $k \geq 1$.

$$\begin{aligned} P_{sub}^{LS}(k|p) &= \sum_{n \geq k} P(k|n,p) \cdot P(n|1) \\ &= \sum_{n \geq k} \binom{n}{k} p^k (1-p)^{n-k} \cdot \alpha(x) \frac{x^n}{n} \\ &= \alpha(x)(px)^k \sum_{n \geq k} \binom{n}{k} \frac{1}{n} \cdot (x(1-p))^{n-k} \\ &= \alpha(x)(px)^k \sum_{n \geq k} \frac{(n-1)\cdots(n-k+1)}{k(k-1)\cdots 1} \cdot (x(1-p))^{n-k} \\ &= \alpha(x)(px)^k \sum_{l \geq 0} \frac{(l+k+1)\cdots(l+1)}{k \cdot (k-1)\cdots 1} (x(1-p))^l \\ &= \alpha(x)(px)^k \sum_{l \geq 0} \frac{1}{k} \binom{l+k+1}{l} (x(1-p))^l \\ &= \alpha(x) \underbrace{\left(\frac{px}{1-x(1-p)}\right)^k}_{:=(\hat{x}_p)^k} \frac{1}{k}, \end{aligned} \tag{2.29}$$

where the special series $\sum_{i=0}^{\infty}\binom{i+m}{i}x^i = \frac{1}{(1-x)^{m+1}}$ has been used in the last equality again.
Now be $k=0$. Passing to the complementary, we obtain

$$\begin{aligned} P_{sub}^{LS}(0|p) &= 1 - \sum_{k\geq 1} P_{sub}^{LS}(k|p) \\ &= 1 - \alpha(x)\cdot \sum_{k=1}^{\infty} \underbrace{\frac{\hat{x}_p}{k}}_{P^{LS}(k|\hat{x}_p)} \\ &= 1 - \frac{\alpha(x)}{\alpha(\hat{x}_p)}, \end{aligned} \tag{2.30}$$

where the normalization equality $\sum_{k=1}^{\infty}\alpha(\hat{x}_p)\cdot\frac{\hat{x}_p}{k} = 1$ has been used.
It follows that the RSA at the local scale p is

$$\begin{aligned} P(k|p) &= \frac{P_{sub}^{LS}(k|p)}{\sum_{k'=1}^{\infty} P_{sub}^{LS}(k'|p)} \\ &= \frac{\alpha(x)\cdot P^{LS}(k|\hat{x}_p)}{\sum_{k'=1}^{\infty}\alpha(x)\cdot P^{LS}(k'|\hat{x}_p)} \\ &= \alpha(\hat{x}_p)\cdot P^{LS}(k|\hat{x}_p). \end{aligned} \tag{2.31}$$

□

2.4 A statistical model for mutation inference

2.4.1 Statistical framework and working hypothesis

Up to now the technical tools needed to derive our upscaling method have been introduced. As already highlighted, the key result for us is the form invariance property Negative Binomial satisfies, which allows to derive an explicit formula for the global mutation number in the three cases in which the local RSA is distributed according to a 1-normalized Negative Binomial with $r>0$, a m-extended Negative Binomial with $r\in(-m,-m+1)$ and a Log-Series.
In the previous section we have seen that such a form invariance property is verified if a binomial sampling can be performed. Thus, we need to assume as working hypothesis for our method those conditions which enable us to have a binomial sampling. In Chapter 2, these latter are listed to be the lack of spatial correlations and the demographic equivalence of mutations, which we have already said to be plausible in genomics. Thus, Lemmas 2.3.2, 2.3.3, 2.3.4 and their claiming hold and can be used in our upscaling procedure.

2.4.2 Estimator of the number of global mutations for 1-normalized Negative Binomial method with $r>0$

We have showed that sampling a fraction p of data distributed according to a Negative Binomial of parameters $r>0$ and ξ results in another Negative Binomial having same cluster parameter r and a rescaled parameter $\hat{\xi}_p$ given by Eq.(2.19). By inverting this latter equation, we can write global parameter ξ as function of the local parameter $\hat{\xi}_p$ and the sampling fraction p

$$\xi = \frac{\hat{\xi}_p}{p+\hat{\xi}_p(1-p)}. \tag{2.32}$$

Observe that above formula can be generalized to a computable expression linking parameters at any two different scales p and p^*

$$\hat{\xi}_{p^*} = \frac{p^*\hat{\xi}_p}{p + \hat{\xi}_p(p^* - p)} = U(p^*|p, \hat{\xi}_p). \tag{2.33}$$

For $p^* = 1$ we obviously restore Eq.(2.32).

Recall that our scope is inferring the global number of mutations, S, from the available information at the local scale $p \in (0, 1)$ only. Named S_p the number of observed mutations at scale p and $S_p(k)$ the number of mutations having exactly k occurrences locally, we can estimate

$$P^{NB}_{sub}(k|p) \simeq \frac{S_p(k)}{S}, \quad k \geq 1. \tag{2.34}$$

Indeed, under the assumption of neutral hypothesis claiming at the nullification of mutation identities and at their equivalent probability to occur in a sample of size p, $P^{NB}_{sub}(k|p)$ can be approximately computed by mean of the heuristic method "favourable events over total events". Then, the complementary is

$$P^{NB}_{sub}(k = 0|p) = 1 - \sum_{k\geq 1} P^{NB}_{sub}(k|p) \simeq 1 - \frac{S_p}{S} \tag{2.35}$$

where the same heuristic reasoning has been made for the last approximation. Substituting Eq.(2.35) into Eq.(2.21), we end up with our estimator

$$\begin{aligned} \hat{S} &= \frac{S_p}{1 - P^{NB}_{sub}(0|p)} \\ &= S_p \cdot \frac{1 - (1 - \xi)^r}{1 - (1 - \hat{\xi}_p)^r}, \end{aligned} \tag{2.36}$$

where ξ is given by Eq.(2.32).

Generalization to mixture of Negative Binomials

When data distribution displays unusual behaviour with different bumps, a mixture of $l \in \mathbb{N}$ Negative Binomials having same ξ but different r_i, $i \in \{1, ..., l\}$ may very well accommodate the empirical RSA. Indeed, the form versatility of a mixture of Negative Binomials, displaying mode bumps and hybrid behaviours (see Fig.2.3), is particularly useful when dealing with data that a single Negative Binomial can not describe properly (see Fig.2.4 taken from Tovo et al. (2017)). Our method works even if we start from a global RSA accommodated by a mixture of $l \in \mathbb{N}$ Negative Binomials. In this case, the estimator is

$$\hat{S} = S_p \frac{\sum_{i=1}^{l} \lambda_i [1 - (1 - \xi)^{r_i}]}{\sum_{i=1}^{l} \lambda_i [1 - (1 - \hat{\xi}_p^{r_i}]}, \tag{2.37}$$

where $\lambda_i \in (0, 1)$, $\sum_{i=1}^{m} \lambda_i = 1$ are the mixture coefficients.
In Eq.(2.37), ξ is obtained from Eq.(2.32) as well, while both $\hat{\xi}_p$, r_i with $i \in \{1, ..., m\}$ are computed through the best fit of the empirical RSA using the theoretical mixture

$$\sum_{i=1}^{m} \lambda_i \cdot c(r_i, \hat{\xi}_p) P(k|r_i, \hat{\xi}_p). \tag{2.38}$$

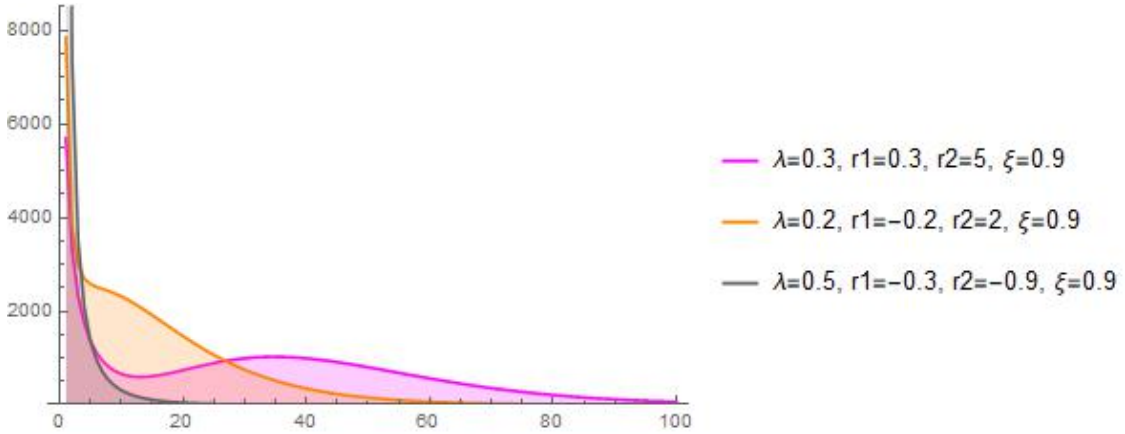


Figure 2.3: SADs produced by a mixture of two Negative Binomials.

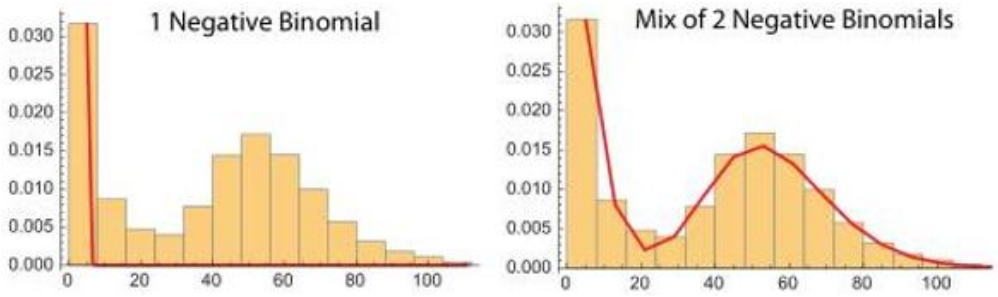


Figure 2.4: Fit with two Negative Binomials.

For our tests in genomics we have assumed global RSA to be a single Negative Binomial since mixtures would not have led to a noticeable improvement in data shape capturing.

2.4.3 Estimator of the number of global mutations for m-extended Negative Binomial method with $r \in (-m, -m+1)$

The derivation of the estimator when starting from a m-extended Negative Binomial RSA follows the same steps as above. Indeed, the same relation holds between the local parameter $\hat{\xi}_p$ and the global parameter ξ. Using the same notations of the previous section, the probability that a mutation occurs k times at the scale p is given by

$$P_{sub}^{m-NB}(k|p) = \frac{S_p(k)}{S}, \qquad k \geq m. \tag{2.39}$$

The complementary is now

$$P_{sub}^{m-NB}(k=0) + \cdots + P_{sub}^{m-NB}(k=m-1|k) = 1 - \sum_{k \geq m} P(k|p) = 1 - \frac{S_p}{S}. \tag{2.40}$$

Combining with the Eq.(2.26), we have

$$\begin{aligned}\hat{S} &= S_p \cdot \frac{\tilde{c}(r,\hat{\xi}_p)}{\tilde{c}(r,\xi)} \\ &= S_p \cdot \frac{1-(1-\xi)^r \cdot \sum_{j=0}^{m-1} \binom{j+r-1}{j} \xi^j}{1-(1-\hat{\xi}_p)^r \cdot \sum_{j=0}^{m-1} \binom{j+r-1}{j} \hat{\xi}_p^j},\end{aligned} \tag{2.41}$$

where ξ is obtained through Eq.(2.32) again.

2.4.4 Estimator of the number of global mutations for 1-normalized Log-Series

The only parameter needed to describe a Log-Series results in two equivalent ways to infer the global number of mutations S when RSA is distributed according to a 1-normalized Log-Series.

The first, easiest, one follows the steps performed for the Negative Binomial (Tovo, 2018). We have

$$\begin{aligned}\hat{S} &= \frac{S_p}{1-P_{sub}^{LS}(k=0|p)} \\ &= S_p \cdot \frac{\log(1-x)}{\log(1-\hat{x}_p)},\end{aligned} \tag{2.42}$$

where the global parameter x is linked to the local one through the formula

$$x = \frac{\hat{x}_p}{p+\hat{x}_p(1-p)} \tag{2.43}$$

derived from the scale invariance property.

The second method requires an estimation for the total abundance N to be applied and it thus results less flexible than the first one. Named N_p the abundance counted at scale $p \in (0,1]$, we have the following relations:

- $S_p = \sum_{k\geq 1} S_p(k) = \sum_{k\geq 1} S \cdot P_{sub}^{LS}(k|p) = \underbrace{S\alpha(x)}_{\tilde{\alpha}} \cdot \sum_{k\geq 1} \frac{(\hat{x}_p)^k}{k} = -\tilde{\alpha}\log(1-\hat{x}_p)$,
- $N_p = \sum_{k\geq 1} kS_p(k) = S\alpha(x)\sum_{k\geq 1}\hat{x}_p = \tilde{\alpha}\frac{\hat{x}_p}{1-\hat{x}_p}$,

where the special series $\sum_{i=1}^{\infty}\frac{y^i}{i} = \log\left(\frac{1}{1-y}\right)$ and $\sum_{i=0}^{\infty} y^i = \frac{1}{1-y}$ have been used, respectively. Manipulating the latter equations, we deal with the following relationship

$$N_p - \tilde{\alpha}\left(\exp\left(\frac{S_p}{\tilde{\alpha}}\right)-1\right) = 0, \tag{2.44}$$

from which, it is possible, from the available data, to get the parameter $\tilde{\alpha}$. Indeed, observe that such a parameter is p-independent so that the above relationship, holding for $p \in (0,1]$, can be used to infer the number of mutations at the global scale. However, we first need an estimation for the unknown parameter $N_{p=1}$ in order to deduce $S_{p=1}$ from Eq.(2.44). From

the scale invariance of a Log-Series distributed RSA, it follows that the mean total abundance $\mathbb{E}(N_p)$ scales linearly with the area p. Indeed,

$$\mathbb{E}(N_p) = \sum_{k\geq 1} kS_p(k) = \sum_{k\geq 1} k\tilde{\alpha}\frac{(\hat{x}_p)^k}{k} = \tilde{\alpha}\frac{\hat{x}_p}{1-\hat{x}_p} = \tilde{\alpha}\frac{px}{1-x} = p\cdot\mathbb{E}(N_{p=1}). \tag{2.45}$$

Setting $S_{p=1} = S$ and $N_{p=1} = N$, it holds consequently that $N = \frac{N_p}{p}$. Substituting such a N into Eq.(2.44), we end with the estimator

$$\hat{S} = \tilde{\alpha}\log\left(1+\frac{N}{\tilde{\alpha}}\right), \qquad \tilde{\alpha} = S\alpha(x). \tag{2.46}$$

2.4.5 Properties of Negative Binomial estimator

In mathematical parlance, an estimator is a statistics used to infer from data the value of an unknown parameter, either finite-dimensional or infinite-dimensional, in a statistical model. Supposed θ the parameter to estimate, an estimator for θ, normally named $\hat{\theta}_n$, where n is the size of the statistical sample, is any function that maps the sample space to a set of sample estimates (Georgii, 2012). Observe that the definition places no restrictions on which statistics can be called estimators. Thus, the attractiveness of different estimators is judged according to their properties, such as unbiasedness, consistency, etc.

Definition 2.2. Let set the bias of $\hat{\theta}_n$ as $B(\hat{\theta}_n) = \mathbb{E}(\hat{\theta}_n) - \theta$, i.e. the distance between the average of the estimates over all possible datasets and the true value of parameter. We say that the estimator $\hat{\theta}_n$ is unbiased for θ if $B(\hat{\theta}_n) = 0$.

Definition 2.3. We say that the estimator $\hat{\theta}_n$ is consistent for θ if the sequence of estimates converges in probability to θ as the number of data points rises to $+\infty$.

Having derived an explicit formula for our estimator of the number of mutations at global scale, we can now check whether above properties are satisfied.

Remark 3. The estimator obtained from a global RSA distributed according to a 1-normalized Negative Binomial of parameters $r > 0$ and ξ,

$$\hat{S} = S_p\cdot\frac{1-(1-\xi)^r}{1-(1-\hat{\xi}_p)^r}, \tag{2.47}$$

is unbiased.

In this respect, we have to show the validity of the following equality

$$\mathbb{E}\left[S_p\cdot\frac{1-(1-\xi)^r}{1-(1-\hat{\xi}_p)^r}\right] = S. \tag{2.48}$$

Observe that if the Negative Binomial would have an infinite tail, then the errors in computing the parameters, both at local and global scale, would decrease to 0, allowing us to treat $\frac{1-(1-\xi)^r}{1-(1-\hat{\xi}_p)^r}$ as a constant. In our application to genomics, the Negative Binomial has not an infinite support (the global size of DNA datasets is 46, whereas a 100×100 grid represents the space for the tumor growth in the tumor synthetic datasets). However, in the data analysis

performed, the computational errors are small enough to handle the parameters and the above term as constants.
Then it holds

$$\mathbb{E}\left[S_p \cdot \frac{1-(1-\xi)^r}{1-(1-\hat{\xi}_p)^r}\right] \simeq \frac{1-(1-\xi)^r}{1-(1-\hat{\xi}_p)^r} \cdot \mathbb{E}[S_p] = \frac{c(r,\hat{\xi}_p)}{c(r,\xi)} \sum_{k=1}^{\infty} k\mathbb{P}(S_p = k). \tag{2.49}$$

Now, since the probability for a mutation to be observed, i.e. to have occurrence ≥ 1, is

$$1 - P_{sub}^{NB}(k=0|p) = 1 - \left(1 - \frac{c(r,\xi)}{c(r,\hat{\xi}_p)}\right) = \frac{c(r,\xi)}{c(r,\hat{\xi}_p)}, \tag{2.50}$$

we have that $\mathbb{P}(S_p = k)$, i.e. the probability that exactly k mutations occur at least once at scale p, is given by

$$\mathbb{P}(S_p = k) = \binom{S}{k} \cdot \left(\frac{c(r,\xi)}{c(r,\hat{\xi}_p)}\right)^k \cdot \left(1 - \frac{c(r,\xi)}{c(r,\hat{\xi}_p)}\right)^{S-k} \tag{2.51}$$

It follows that $S_p \sim \text{Bin}\left(S, \frac{c(r,\xi)}{c(r,\hat{\xi}_p)}\right)$. Since the mean of a random variable distributed according to a Binomial of parameters (u, v) is uv, we end with

$$\begin{aligned}\mathbb{E}\left[S_p \cdot \frac{1-(1-\xi)^r}{1-(1-\hat{\xi}_p)^r}\right] &\simeq \frac{c(r,\hat{\xi}_p)}{c(r,\xi)} \cdot \mathbb{E}\left[Bin\left(S, \frac{c(r,\xi)}{c(r,\hat{\xi}_p)}\right)\right] \\ &= \frac{c(r,\hat{\xi}_p)}{c(r,\xi)} \cdot S\frac{c(r,\xi)}{c(r,\hat{xi}_p)} = S.\end{aligned} \tag{2.52}$$

Remark 4. Performing similar steps we can prove that both the estimator obtained from a global RSA distributed according to a m-extended Negative Binomial of parameters $r \in (-m, -m+1)$ and ξ,

$$\hat{S} = S_p \cdot \frac{1-(1-\xi)^r \cdot \sum_{j=0}^{m-1}\binom{j+r-1}{j}\xi^j}{1-(1-\hat{\xi}_p)^r \cdot \sum_{j=0}^{m-1}\binom{j+r-1}{j}\hat{\xi}_p^j}, \tag{2.53}$$

and the one obtained from a global RSA distributed according to a 1-normalized Log-Series of parameter α,

$$S_p \cdot \frac{\log(1-x)}{\log(1-\hat{x}_p)}, \tag{2.54}$$

are unbiased as well.

Remark 5. The estimator obtained from a global RSA distributed according to a 1-normalized Negative Binomial of parameters $r > 0$ and ξ,

$$\hat{S} = S_p \cdot \frac{1-(1-\xi)^r}{1-(1-\hat{\xi}_p)^r}, \tag{2.55}$$

has the property

$$\lim_{p\to 1} Var\left(\hat{S}\right) = 0 \tag{2.56}$$

In this respect, following Remark 3, we can assume that $\frac{1-(1-\xi)^r}{1-(1-\hat{\xi}_p)^r}$ is a constant as well due to the small error size computed on the parameters ξ, $\hat{\xi}_p$ and r. Moreover, we have already showed (see Remark 3) that $S_p \sim \text{Bin}\left(S, \frac{c(r,\xi)}{c(r,\hat{\xi}_p)}\right)$, so that its variance is $S \cdot \frac{c(r,\xi)}{c(r,\hat{\xi}_p)} \cdot \left(1 - \frac{c(r,\xi)}{c(r,\hat{\xi}_p)}\right)$. Thus, it holds

$$VAR\left(S_p \cdot \frac{1-(1-\xi)^r}{1-(1-\hat{\xi}_p)^r}\right) = \left(\frac{c(r,\hat{\xi}_p)}{c(r,\xi)}\right)^2 VAR\left(S_p\right) = S \cdot \frac{c(r,\hat{\xi}_p) - c(r,\xi)}{c(r,\xi)} \tag{2.57}$$

Then, when $p \to 1$ we have $\hat{\xi}_p \to \xi$ and, by continuity, $c(r,\hat{\xi}_p) \to c(r,\xi)$. Our thesis follows directly from the last limit.

Remark 6. Similar steps allow us to prove that both the estimator obtained from a global RSA distributed according to a m-extended Negative Binomial of parameters $r \in (-m, -m+1)$ and ξ,

$$\hat{S} = S_p \cdot \frac{1-(1-\xi)^r \cdot \sum_{j=0}^{m-1} \binom{j+r-1}{j} \xi^j}{1-(1-\hat{\xi}_p)^r \cdot \sum_{j=0}^{m-1} \binom{j+r-1}{j} \hat{\xi}_p^j}, \tag{2.58}$$

and the one obtained from a global RSA distributed according to a 1-normalized Log-Series of parameter α,

$$S_p \cdot \frac{\log(1-x)}{\log(1-\hat{x}_p)}, \tag{2.59}$$

satisfy Eq.(2.56) as well.

2.5 Implementation of test procedure

The analytical method previously exposed has been derived taking into account information on abundances. However, in genomics, data are often available in the form of binary matrices built from presence/absence information and abundances are not readable at first glance. Thus, as stated in Tovo et al. (2019), we need to adapt our computational framework to such a presence/absence context. Hereafter the computational recipe for the inference of the global number of mutations from presence/absence information using the 1-normalized Negative Binomial with $r > 0$ method only is exposed.

- **Initialization.** Let us suppose to have a local scale binary matrix, representing a fraction $p \in (0,1)$ of the global scale matrix, so composed:
  - matrix rows are all loci $s \in \{1, ..., S_p\}$ in which mutations may be observed at local scale,
  - matrix columns c_i with $i \in \{1, ..., M_p\}$, $M_p \geq 2$ are cells of equal size a representing different samples (sequenced individuals, biopsies,...).

  Such a matrix is obtained by associating to each cell c_i a S_p dimensional vector $\Omega(c_i) = \{x_1^i, ..., x_{S_p}^i\}$, with $x_s^i \in \{0,1\}$, $s \in \{1, ..., S_p\}$, $i \in \{1, ..., M_p\}$, whose entries give information on the presence or absence of mutation s in cell c_i, and by joining those vectors into a $S_p \times M_p$ matrix. In other word, x_s^i is set equal to 1 if mutation s occurred in cell c_i, to 0 otherwise.

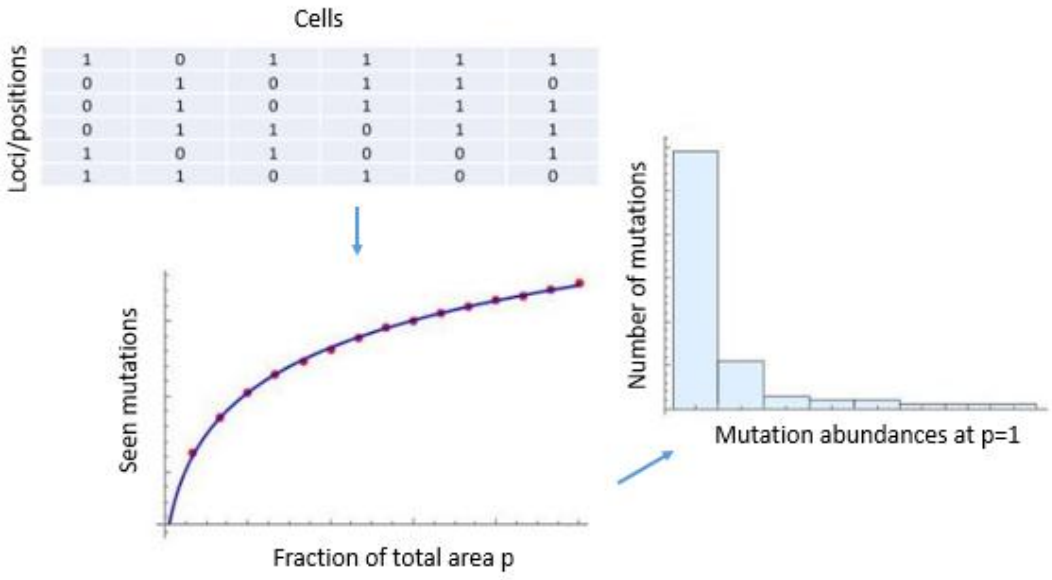


Figure 2.5: Schematic representation of the statistical framework.

- **First step.** Compute the empirical Species-Accumulation Curve in the following way. Under the hypothesis that the global scale has been divided into cells of equal size a and that M_p local cells correspond to a fraction p of the A global ones, compute the average number of observed mutations at sub-sampling scale $p_k = \frac{ka}{A}$, $k \in \{1, ..., M_p\}$, as

$$S_{emp}(p_k) = \frac{1}{\binom{M_p}{k}} \sum_{I \subseteq 1,...,M_p/|I|=k} \sum_{s=1}^{S_p} 1\left(\sum_{i \in I} x_s^i \geq 1\right) \tag{2.60}$$

  where $\mathbb{1}(\epsilon)$ is the indicator function of event ϵ defined as

$$\mathbb{1}(E) = \begin{cases} 1 & \text{if } E \in \epsilon \\ 0 & \text{if } E \notin \epsilon \end{cases} \tag{2.61}$$

  Virtually one has to compute the empirical average of the number of occurred mutations in all subsets of cardinality k with $k \in \{1, ..., M_p\}$. Such a calculus is computational expansive for big size local datasets, thus, in the following simulations, the number of concerned subsets (randomly chosen) is cut to 100.

- **Second step.** Fit the empirical curve with the theoretical one, given by

$$S_{theo}(p_k) = S_p \frac{1 - (1 - U^{NB}(p_k|1, \hat{\xi}_p)^r)}{1 - (1 - \hat{\xi}_p)^r} \tag{2.62}$$

  with

$$U^{NB}(p_k|1, \hat{\xi}_p) = \frac{\hat{\xi}_p}{p_k + \hat{\xi}_p(1 - p_k)}, \tag{2.63}$$

  and get parameters (r, ξ_p) that best match the local empirical SAC.

- **Third step.** Use the formula

$$\xi = \frac{\hat{\xi}_p}{p + \hat{\xi}_p(1-p)}. \tag{2.64}$$

to obtain the global parameter ξ from the local ones and insert ξ, $\hat{\xi}_p$ and r into the expression of the estimator obtained from a global 1-normalized Negative Binomial RSA,

$$\hat{S} = S_p \cdot \frac{1-(1-\xi)^r}{1-(1-\hat{\xi}_p)^r}, \tag{2.65}$$

to infer the number of mutations at the global scale.

Remark 7. The inference using the m-extended Negative Binomial method or the Log-Series method follows the same steps. In those cases, the theoretical curves to fit the empirical ones are

$$S_p \frac{1-(1-U^{m-NB}(p_k|1,\hat{\xi}_p))^r \cdot \sum_{j=0}^{m-1} \binom{j+r-1}{j} U^{m-NB}(p_k|1,\hat{\xi}_p)^j}{1-(1-\hat{\xi}_p)^r \cdot \sum_{j=0}^{m-1} \binom{j+r-1}{j} \hat{\xi}_p^j} \tag{2.66}$$

with

$$U^{m-NB}(p_k|1,\hat{\xi}_p) = \frac{\hat{\xi}_p}{p_k + \hat{\xi}_p(1-p_k)} \tag{2.67}$$

and

$$S_p \frac{\log(1-U^{LS}(p_k|1,\hat{x}_p))}{\log(1-\hat{x}_p)} \tag{2.68}$$

with

$$U^{LS}(p_k|1,\hat{x}_p) = \frac{\hat{x}_p}{p_k + \hat{x}_p(1-p_k)}, \tag{2.69}$$

respectively.

In the following chapters, we will follow this recipe to computationally test our method on two typologies of datasets: one concerning the DNA polymorphisms occurring in the X chromosome of a group of British males and the other regarding mutations occurring in a spatially constrained simulated tumor.

Chapter 3

Inference of unseen variants in DNA single-nucleotide polymorphism datasets

In this chapter we apply the theoretical method we have developed so far to infer the number of unseen variants in real DNA single-nucleotide polymorphisms. These are specific substitutions, affecting one or at most very few nucleotides, that are present to a $> 1\%$ degree within a population. Notice that in these particular datasets the number of variants is indeed known at each scale. Thus, the analysis conducted in this chapter will serve also to test our method. Below, the structure of the datasets and the results obtained from their analysis are presented.

3.1 DNA single-nucleotide polymorphism datasets

DNA polymorphysm simulations have been run on three single-nucleotide variant datasets regarding X chromosome: they do provide global information about mutations' occurrences in three independent regions of X chromosome for 46 British male individuals. Hereafter the three datasets are called data 1, data 2 and data 3 respectively to fast the notations. Datasets information are available in the form of VCF tables, which are binary matrices having all the observed mutated loci of the DNA region as rows and the 46 sampled individuals as columns. When a mutation affects a particular genetic locus of a precise individual, the corresponding matrix cell is filled with 1, otherwise with 0. The VCF tables carry qualitative information on the type of mutation occurred and on the allele frequency as well.

Theoretically one expects that the distribution of the mutations frequencies is approximately given by a power law of exponent $\delta = -1$ (Williams et al., 2018) and, apart from some computational fluctuations, our datasets display such a tendency (see Fig.3.1). However, from (a), (c) and (e) panels of Fig.3.1, we can observe that mutations occurring in the whole ensemble of sequenced individuals have a different behaviour, not properly following the above power law curve: histograms clearly display a final peak, which is a common behaviour highlighted in Williams et al. (2018) as well. Indeed, those mutations (a bit improperly called clonal hereafter) are present from the real beginning in the system, representing somehow a initial condition, whereas we are interested in observing the dynamics from a certain time t forward. Thus, our analysis will not accounted for clonal mutations: corresponding rows will be deleted

		$\frac{10}{46}$	$\frac{15}{46}$	$\frac{20}{46}$	$\frac{25}{46}$	$\frac{30}{46}$	$\frac{35}{46}$	$\frac{40}{46}$
data1	$r \in (-1, 0)$	$-1.28 \cdot 10^{-7}$	-0.00025	-0.0013	-0.005	-0.024	-0.047	-0.071
	$r > 0$	0.194	0.103	0.053	0.014	0.0008	0.0004	0.008
data2	$r \in (-1, 0)$	-0.01	-0.004	-0.004	-0.002	-0.009	-0.041	-0.081
	$r > 0$	0.285	0.277	0.128	0.078	0.027	0.0019	0.0002
data3	$r \in (-1, 0)$	-0.0002	-0.001	-0.117	-0.038	-0.077	-0.105	-0.132
	$r > 0$	0.211	0.1	0.025	0.003	$1.44 \cdot 10^{-7}$	$1.92 \cdot 10^{-7}$	$1.02 \cdot 10^{-7}$

Table 3.1: **Best fitting values for r parameter with Negative Binomial method in the domains r>0 and -1<r<0.** The table shows best fitting values for r when 1-normalized Negative Binomial with $r > 0$ and its extension to $-1 < r < 0$ method are used. Observe that for each dataset such values shrink to 0 from both sides as one may expect from the theory, claiming at a mutational distribution that follows a power law of exponent $\delta = -1$ (Williams et al., 2018).

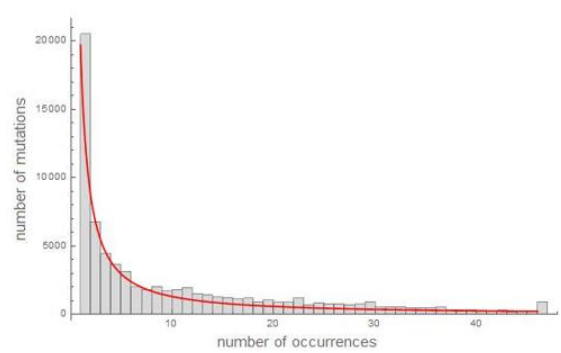
from VCF tables before the start of the computational procedure and our aim will be to infer the number of subclonal mutations (not clonal ones) occurring at the global scale starting from their abundances at many local scales.

Our upscaling method has been tested for increasing local scales $p = \{\frac{10}{46}, \frac{15}{46}, \frac{20}{46}, \frac{25}{46}, \frac{30}{46}, \frac{35}{46}, \frac{40}{46}\}$ using 1-normalized Negative Binomial method with $r > 0$, m-extended Negative Binomial method with $-1 < r < 0$ and Log-Series method with $r \to 0$. Due to the form of the available data, samples have been selected by randomly choosing p columns, i.e. individuals, from the global matrix.
We expect the Log-Series method works more efficiently in this context since relationship $-1 + r = \delta$ (see Prop.2.2.1.) with $\delta \approx -1$ (this is the case) is satisfied for $r \approx 0$.

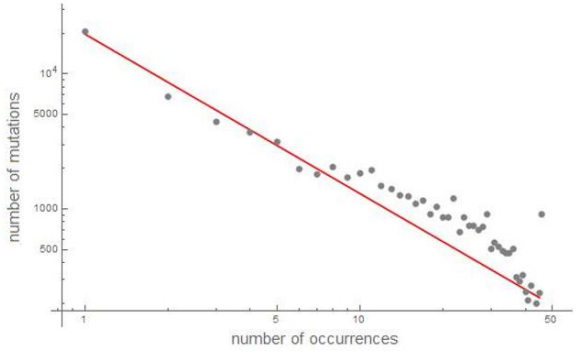
3.1.1 Results

In order to check the stability of the analytical framework for each global dataset of size 46 and at each local scale $p = \{\frac{10}{46}, \frac{15}{46}, \frac{20}{46}, \frac{25}{46}, \frac{30}{46}, \frac{35}{46}, \frac{40}{46}\}$ the upscaling method has been tested on 100 training sets randomly chosen. In particular, simulations have been performed separately for r in the three domains: $r \in (0, \infty)$, $r \in (-1, 0)$, $r \approx 0$.

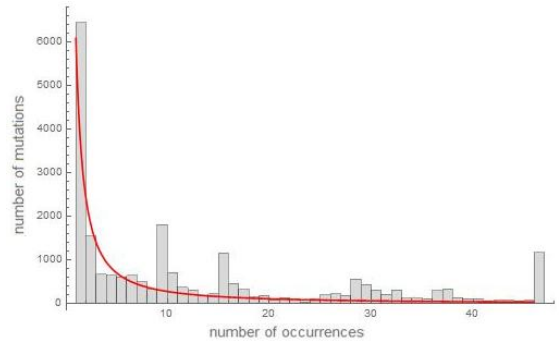
For less than 8% of initial training sets fitting with Negative Binomial with $r > 0$ and with its extension to $-1 < r < 0$ do not perform at all due to the computational limits affecting the algorithm, while no computational problems have occurred with the Log-Series. The only few initial training sets (the number depends on the handled dataset) displaying different behavior have been characterized. They do not represent a good sampling in the sense that they do not reflect the global behavior, since many individuals carrying values consistently far from the average have been selected during random extraction. This unreliable outcome tendency for some samplings gets more noticeable with local scales smaller than $\frac{10}{46}$ where the few available points for the fitting represent a further obstacle to the stability of the method.
However, at each local scale and for the 95%-99% of the training sets for which the fitting with Negative Binomial (both with $r > 0$ and $-1 < r < 0$) works, outputs are consistent with



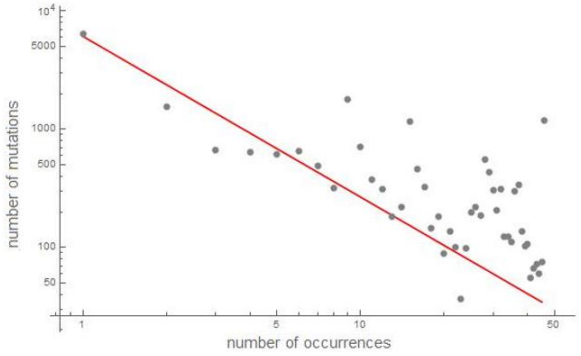
(a) Plot for data 1



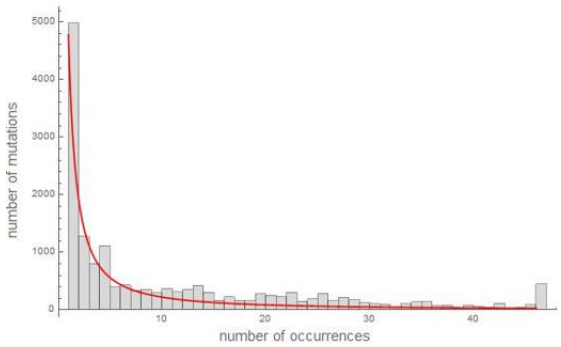
(b) Log-log plot for data 1



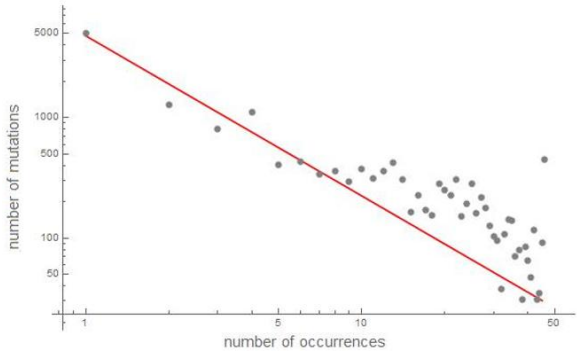
(c) Plot for data 2



(d) Log-log plot for data 2

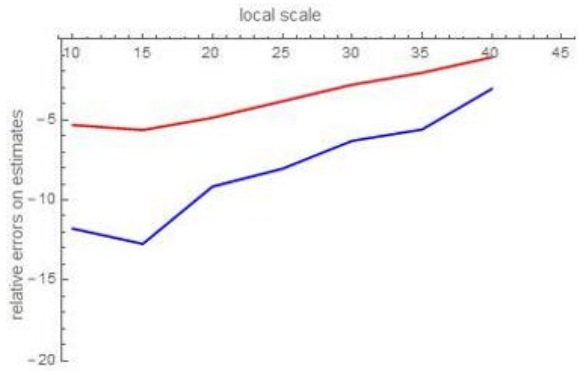


(e) Plot for data 3

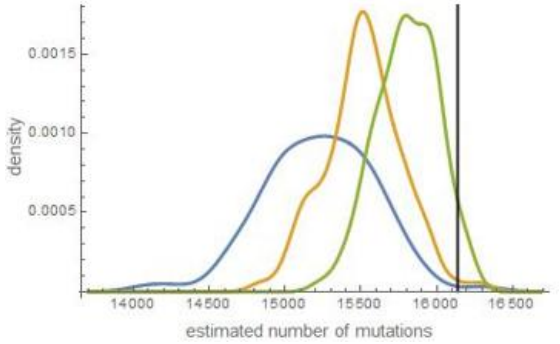


(f) Log-log plot for data 3

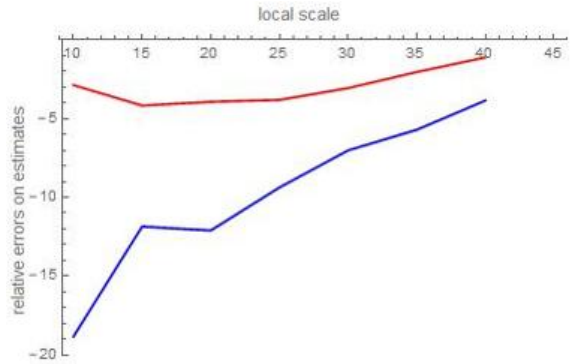
Figure 3.1: **Plot and log-log plot of RSAs at global scale for the three DNA single-nucleotide polymorphism datasets.** Panels on the left hand side show the global RSAs for the three datasets together with the best computational fitting power law having exponent equal to $\delta = -1.18$, $\delta = -1.34$ and $\delta = -1.33$, respectively. Panels on the right hand side display the same information using log-log plots.



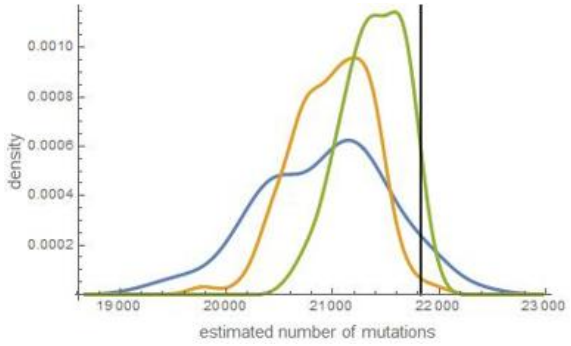
(a) Relative errors for data 1



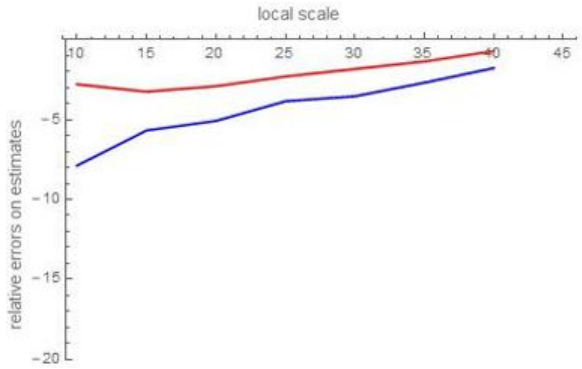
(b) Estimates histograms at different scales for data 1



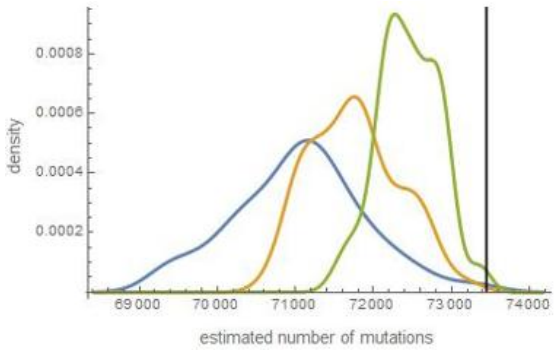
(c) Relative errors for data 2



(d) Estimates histograms at different scales for data 2



(e) Relative errors for data 3



(f) Estimates histograms at different scales for data 3

Figure 3.2: **Average estimates and errors for the three DNA single-nucleotide polymorphism datasets.** Panels on the left hand side show the average relative error (red line) and the average maximum error (blue line) as local scale increases, while those on the right display how estimates are distributed at different local scales - $\frac{15}{46}$ (lightblue curve), $\frac{25}{46}$ (orange curve), $\frac{35}{46}$ (green curve) - for all three datasets. Note that for local scales greater than $\frac{15}{46}$, an asymptotic and predictable behavior arises with both errors decreasing constantly and smooth histograms narrowing the correct number to predict (black vertical line).

			$\frac{10}{46}$	$\frac{15}{46}$	$\frac{20}{46}$	$\frac{25}{46}$	$\frac{30}{46}$
data1	S=16139	S_{pred}	15279.6	15229.1	15354	15518.3	15685.4
		err%	-5.32	-5.63	-4.86	-3.84	-2.81
data2	S=21826	S_{pred}	21200.6	20916.2	20968.9	20995.5	21158.9
		err%	-2.86	-4.16	-3.92	-3.8	-3.05
data3	S=73447	S_{pred}	71400.4	71060.6	71315.6	71758.9	72107.2
		err%	-2.78	-3.24	-2.9	-2.29	-1.82

Table 3.2: **Average Log-Series estimates and errors at different local scales for DNA polymorphism datasets.** The table shows the average estimates of the global number of mutations, S, and the relative errors at increasing local scales, obtained with the application of the Log-Series method to the three DNA single-nucleotide polymorphism datasets.

expectation. Indeed, r parameter that best fits empirical curve to the theoretical one shrinks to 0 from both sides. When $r > 0$ the best fitting parameter fluctuates closing to 0^+ for each training sets and at each local scale, whereas when the domain for r is $(-1, 0)$ the best r parameter narrows 0^-, not monotonically as well (see Tab.3.1). The Log-Series estimator

works well with an average relative error below 5% for all datasets and a maximum relative error that lies between $7\% - 15\%$ at smaller scale. Asymptotically, it holds that

- the average relative errors decrease,
- the estimation biases decrease

in all tests as Fig.3.2 shows. In particular the estimates dispersion reduces as the local scale increases: the shape of the histogram produced by the 100 training sets estimates gets sharper and thinner and narrows the correct number to estimate, meaning that the probability of randomly selecting a training set that gives back a good final estimate becomes higher with the local scale size.

Chapter 4

Inference of unseen variants in synthetic tumor datasets

The major interest for us is to test the validity of our analytical framework in inferring the global number of variants in a tumor from local information. However, in cancer field one generally works with really few biopsies that rarely reach the number of ten per tumor/patient, a too small size to test our framework with profit. Indeed, to check the legitimacy of our method we need to have information on mutation occurrences at several local scales and at the global one, so that we can compute the final error and test the reliability of the method. Thus the ten available biopsies, i.e. the maximum information we have, should be regarded as the global scale and the upscaling method would start from a sub-sampling of these biopsies. It follows that the sample, i.e. the local scale, would be compounded by a integer $\ll 10$ number of biopsies that would lead to an empirical SAC composed by too few points for a reliable fitting. Consequently, to not tackle these computational problems, we have been forced to work with synthetic datasets, whose size we could control. We have produced the artificial data for our tests by exploiting a spatially constrained stochastic model for tumor growth (Chkhaidze et al., 2019). Such a model simulates both the spatial evolution of a single cell tumor and the multi-region sequencing data derived from spatial sampling of a neoplasm. In what follows, we will introduce at first the stochastic model used to generate the data following what reported in Chkhaidze et al. (2019), then our focus will shift to the presentation of the results obtained from the simulations on those synthetic datasets.

4.1 A spatially constraint tumor growth model

4.1.1 Stochastic framework and computational steps

To produce data we have exploited a stochastic spatial cellular automaton model for tumour growth which incorporates cell division, cell death, clonal selection and random mutations into the dynamics (Chkhaidze et al., 2019). Here, the spatial progression of the tumor from a single initial cancer cell on a lattice is monitored, tracking the evolutionary trajectory for each cell of the tumor. For a realistic scenario, different spatial constrains are accounted in the simulations, in particular a boundary-driven growth can be parameterized. This latter wants cells close to tumor periphery only to be allowed to proliferate; those in the tumor centre are supposed to be unable to divide because of the lack of empty space around them. The size

effect of such a constrain is controlled by mean of the parameter $a \in (0, 1)$, which represents the fraction of the growing tumor radius where cells can proliferate. That is, $a \to 0^+$ creates a tumor periphery of a small width with external cells only allowed to divide, giving rise to a polynomial progression. When $a \to 1^-$ the periphery width is close to 100%, meaning that every cell is able to create its own space for proliferation, so that an exponential growth arises.

In the simulations, the spatial tumorgenesis, accounting for all events that may occur, is modelled according to a Gillespie algorithm. This latter is a classical probabilistic tool used in stochastic simulations to obtain at each run a statistically correct trajectory, i.e. a possible probability mass function solution, of a stochastic master equation. In the case of our interest, the algorithm works as follows.

- **Initialization.**
  - Fix the dimension of a 2D lattice which defines all possible spatial directions for the tumor growth and place a single cancer cell, having b and d as birth and death rates respectively, in one of lattice central spots.
  - Define a set of times at which driver mutations carrying proliferative advantages will occur.

Algorithm for tumor expansion will work by generating new daughter cells along with new passenger and driven mutations affecting them and suitably positioning the cells in the lattice, once per spot, until a pre-fixed desire simulation time T is reached.

- **First event.** The first event that will occur, which can be the division or the death of the unique cell present in the lattice at the beginning, is the one with the smallest clock value between those of the two plausible events. To compute these values, for each event we draw an exponential random variable with mean equal to the event rate and we sample it. Then, we compare the cell birth time, t_{b1}, and the cell death time, t_{d1}, in order to determine the winning event.
  - If $t_{d_1} < t_{b_1}$, first-cell death occurs and the cell position in the lattice is freed. In such a case, the current simulation ends since the only one cell present in the lattice is driven to death, i.e. remove, and a new simulation can start.
  - If $t_{b_1} < t_{d_1}$, first-cell division occurs, giving birth to two daughter cells that have to settle on the grid. One replaces its parent cell in the grid spot, while the second is positioned in one of the eight empty von Neumann neighbouring spots of the parent cell (see Fig.4.1).

    Once the two daughter cells have been placed in the lattice, new passenger mutations can be added to both of them. The quantity of such mutations is determine by drawing a Poisson variable of mean μ.

    Finally driver mutations, whose occurring times have been fixed from the beginning, need to be accounted. If one of those scheduled times is minor than t_{b_1}, i.e. a driver mutation occurs before the end of the first-cell division event, then a daughter cell is provided with a proliferative advantage (higher birth rate or lower

death rate) quantified through the s parameter. In particular, s is such that

$$1 + s = \frac{birth_{mutant} - death_{mutant}}{birth_{normal} - death_{normal}}$$

To end the first event step, the simulation time, which starts from 0, is increased by the winning event clock value.

Same computational steps have to be performed until the end of the simulation. A bit more attention needs to be paid as simulation clock runs and more cells are present in the lattice. Indeed, while at first event step all the eight von Neumann neighbouring spots of the divided cell are empty because only one cell inhabits the lattice, at n event step all those spots may be occupied and a heuristic method has to be found to place the second daughter cell in a suitable position of the lattice.
To be clear, let us describe in details what the n^{th} event may be. Suppose that after the $(n-1)^{th}$ event the simulation clock signs the time $\hat{T} < T$ and that there are N cells currently in the lattice. Each of them has its own birth and death rate depending on the number of driver mutations that the corresponding parent cell has transmitted.

- N^{th} **event.** To determine the n^{th} event compute the time of each plausible event (division or death) for each cell in the lattice by drawing an exponential variable having the cell event transition probability as mean and compare the obtained times. Select the smallest time, $\bar{T}$, among those computed: the cell dealing with that time will undergone the winning event.

  - If such a time regards a death event, then n^{th} event is a cell death with the cell corresponding to the winning time remove from the lattice.

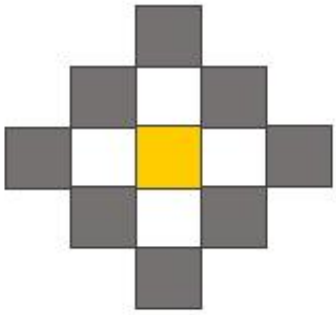
  - If $\bar{T}$ concerns a division event, then n^{th} event will be the division of the cell whose birth event time has won. For the first daughter cell no obstacles arise in its placing on the lattice since it will occupy the spot of the parent cell. For the second daughter no free von Neumann spots may be available if the lattice is already high inhabited. In such a case, a direction among the eight defined by the von Neumann spots is randomly selected and all cells along that direction are pushed further to create an empty spot. If along the chosen direction the spots are all inhabited until the grid boundary so that the push can not happen, another direction is selected (see Chkhaidze et al. (2019) for details). After new cells positioning, both passenger and driver mutations are accounted in the same manner of above. In particular, a driver mutation may affect a daughter cell only if a scheduled time for drivers lies in the temporal interval $(\hat{T}, \hat{T} + \bar{T})$.

4.1.2 Parameters calibration

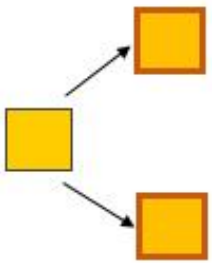
Once the values of all the parameters - b, d, a, s, μ - have been fixed, we have seen that Gillespie algorithm works simulating how a tumor may spatially evolve. However, there is no theoretical hint on which values the parameters should be set equal to in the model: their joint distribution $p(\phi)$ must be computed a posteriori. In order to obtain such a distribution, we have to repeat the following steps for a sufficient number of times.

- From a priori distributions, select a parameters set ϕ.

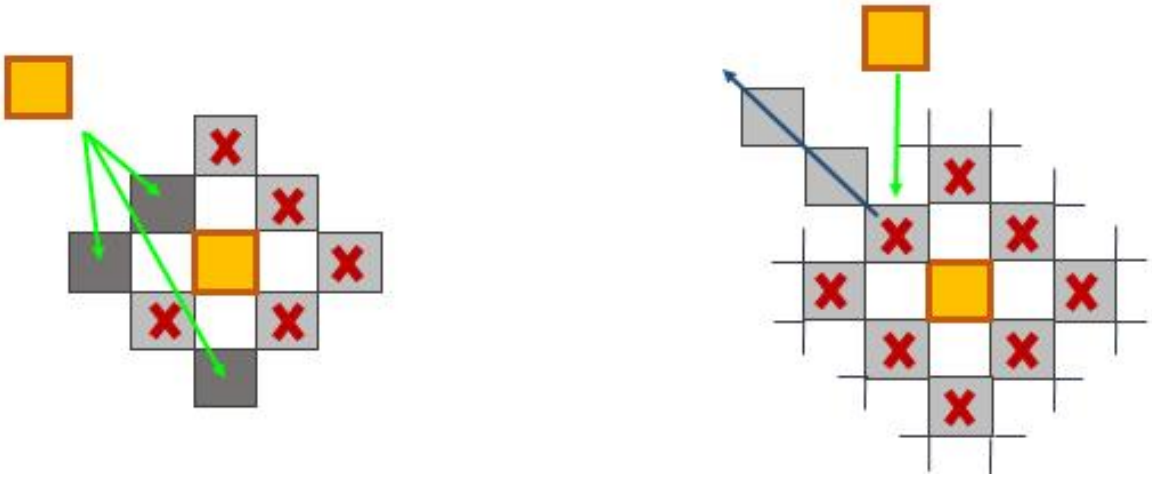
- Insert such a set into the model and run it in order to generate the corresponding output.
- Evaluate the distance between the Gillespie output and the output of a target model, i.e. a predefined set of summary statistic firstly evaluated empirically. If the distance is less than a fixed threshold, then parameters set ϕ can be accepted, otherwise it must be get rid of.
- Rank the acceptable parameter sets and select the joint distribution from those sets for future reliable simulations.



(a) Von Neumann neighborhood (dark gray squares) for the yellow spot in a 2d square grid



(b) Cell division event with parent cell giving born to two identical daughter cells (orange-bounded yellow squares)



(c) Placing of the daughter cells in suitable spots of the grid depending on the occupancy (red-crossed light gray squares) of parent cell's von Neumann spots

Figure 4.1: **Schematic representation of the stochastic model algorithm for cell division events.** In the top panel the eight 2d von Neumann spots (dark gray ones) for the yellow cell (x_0, y_0) are represented. When a cell division event occurs, two daughter cells, carrying the same genetic heritage of the parent cell, are generated (b). No difficulties in placing them on the grid arise when at least one von Neumann spot of the parent cell is free (first figure of (c)): a daughter cell replaces the parent cell and the other occupies one of the parent cell's von Neumann free spots. If none of these latter are empty, the cells in the grid are pushed in a randomly chosen direction until a free spot for the second daughter cell is found (second figure of (c)).

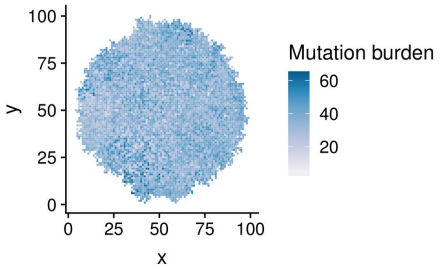
4.2 Tests on synthetic data

Once a good set of parameters ϕ has been selected and the model has run until the desired simulation time T, the output contains enough information to test our method. Indeed, algorithm steps provide with the exact position on the grid of each cell (x-y coordinates) and the mutations they carry. Thus, mutation abundances are available.

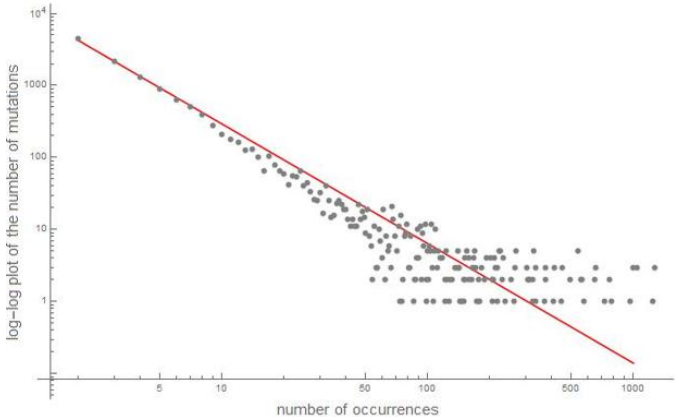
We have tested our upscaling framework on four datasets, provided us by the Institute of Cancer Research- Bioinformatics Section, simulating both exponential (simulation 1 and simulation 2) and polynomial (simulation 3 and simulation 4) tumour growth. Due to the form of available data, we have opted to define local scales by sampling among occurrences. Tests at increasing local scales $p \in \{\frac{1}{100}, \frac{10}{100}, \frac{20}{100}, \frac{30}{100}, \frac{40}{100}, \frac{50}{100}, \frac{60}{100}, \frac{70}{100}, \frac{80}{100}, \frac{90}{100}\}$ have been conducted.

It has been proved that assuming a well-mixed population and an exponential neutral growth results in a variant allele frequency distribution, i.e. the distribution of the mutation frequencies, that goes like a power-law of exponent 2. If parameters for a polynomial neutral growth are set, a power-law with an exponent > 2 arises. From $-1 + r = \delta$ relationship, it follows that the best r to capture data behaviour is negative. Thus, we need to verify which interval $(-m, -m + 1)$ it belongs to in order to use the correct m-extended Negative Binomial method in our tests. Due to the r scale invariance, we can deduce suitable $m \in \mathbb{N} \setminus \{0\}$ by

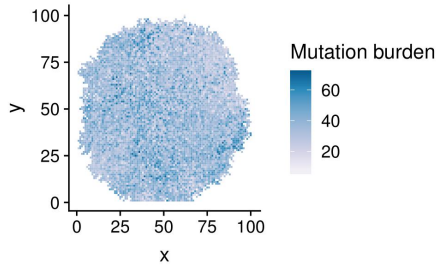
computing the δ exponent occurring at the global scale. Neglecting perturbations that may affect distribution at small n and noise effect, it results $m = 1$ for each dataset.



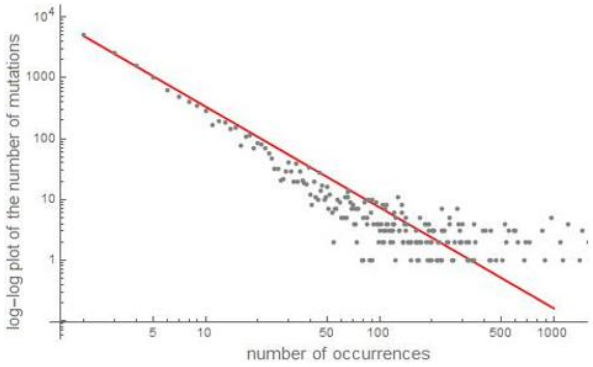
(a) Spatial evolution for simulation 1



(b) Log-log plot for simulation 1



(c) Spatial evolution for simulation 2

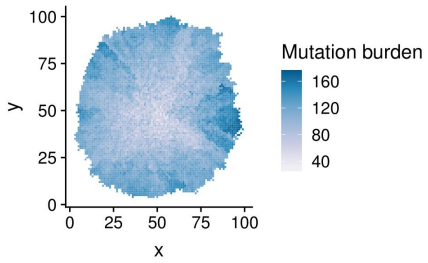


(d) Log-log plot for simulation 2

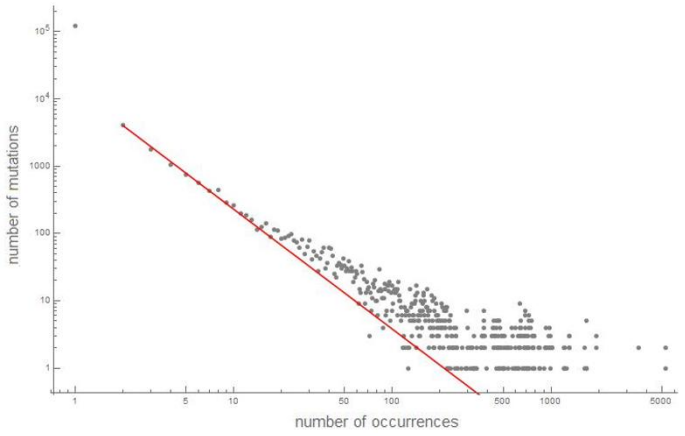
Figure 4.2: **Tumor growth representation and log-log plot of global RSA for exponential growth datasets.** Panels on the left hand side represent the spatial tumor evolution on 100 $\times$ 100 grid at the end of the simulations with the mutational burden for each cell. For an exponential tumor growth such a burden is homogeneous, with no peaks in some regions because each cell has the same probability of proliferating and thus of acquiring new mutations. Corresponding panels on the right show the global distribution of the mutation occurrences and the best fitting power-law in a log-log plot. The best exponents δ for the central section of distribution are $\approx$ -1.66 in both cases.

4.2.1 Results

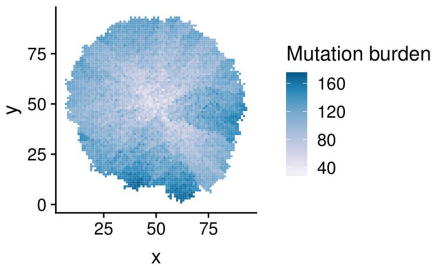
We have tested our 1-extended Negative Binomial method on the four datasets mentioned - simulation 1, simulation 2, simulation 3 and simulation 4 -, considering local scales $p \in \{\frac{1}{100}, \frac{10}{100}, \frac{20}{100}, \frac{30}{100}, \frac{40}{100}, \frac{50}{100}, \frac{60}{100}, \frac{70}{100}, \frac{80}{100}, \frac{90}{100}\}$ and running 100 simulations each time in order



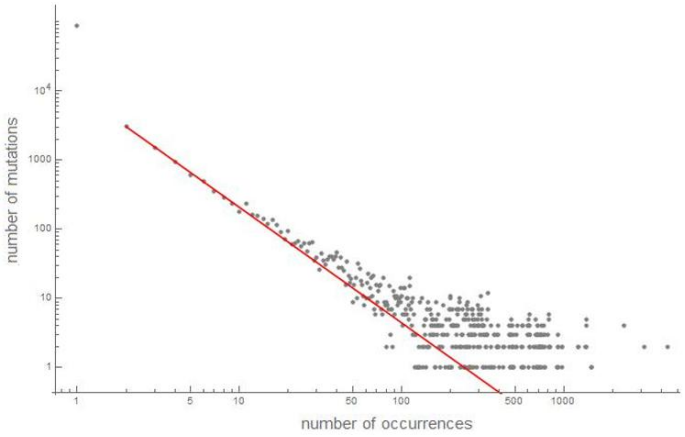
(a) Spatial evolution for simulation 3



(b) Log-log plot for simulation 3



(c) Spatial evolution for simulation 4



(d) Log-log plot for simulation 4

Figure 4.3: **Tumor growth representation and log-log plot of global RSA for polynomial growth datasets.** Panels on the left hand side represent the spatial tumor evolution on 100 × 100 grid at the end of the simulations with the mutational burden for each cell. For a polynomial growth mutations are accumulated along the directions in which the cells are allowed to proliferate, giving rise to thicker regions (dark blue) near the growing tumor edge. Corresponding panels on the right show the global distribution of the mutation occurrences and the best fitting power-law in a log-log plot. Moving top-down, the best exponents δ for the central section of distribution are $\approx$ -1.78 and $\approx$ -1.67, respectively.

			$\frac{5}{100}$	$\frac{10}{100}$	$\frac{20}{100}$	$\frac{30}{100}$	$\frac{40}{100}$	$\frac{50}{100}$
simulation2	S=26569	S_{pred}	27094	27866.3	28580.1	28532.3	28330.5	28035.2
		err%	1.97	4.88	7.56	7.38	6.63	5.51
simulation4	S=30485	S_{pred}	31640.4	32961.4	33183.3	32991.6	32633.3	32270
		err%	3.78	8.12	8.85	8.22	7.04	5.85

Table 4.1: **Average Negative Binomial estimates for -1<r<0 and relative errors at different local scales for simulated tumor datasets with exponential growth.** The table shows the average estimates of the global number of mutations, S, with the Negative Binomial method having $-1 < r < 0$ and the relative errors at smaller local scales for each synthetic tumor datasets displaying an exponential growth.

			$\frac{30}{100}$	$\frac{40}{100}$	$\frac{50}{100}$	$\frac{60}{100}$	$\frac{70}{100}$	$\frac{80}{100}$
simulation1	S=136284	S_{pred}	143206	146847	147228	144725	141934	139727
		err%	5.07	7.75	8.03	6.19	4.14	2.52
simulation3	S=101138	S_{pred}	102967	106803	107504	106934	105703	103946
		err%	1.8	5.6	6.29	5.73	4.51	2.77

Table 4.2: **Average Negative Binomial estimates for -1<r<0 and relative errors at different local scales for simulated tumor datasets with polynomial growth.** The table shows the average estimates of the global number of mutations, S, with the Negative Binomial method having $-1 < r < 0$ and the relative errors at different local scales for each synthetic tumor datasets displaying an exponential growth.

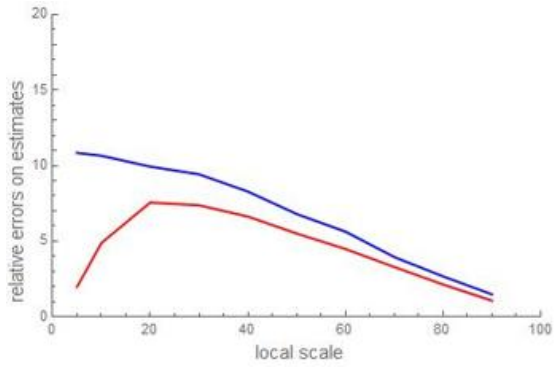
to check the stability of the method.

Due to the best fitting power-law exponent values, we would expected $r \approx -0.78$ for simulation 1, ≈ -0.66 for simulation 2, ≈ -0.67 for simulation 3 and ≈ -0.66 for simulation 4 in order to satisfy $-1+r = \delta$. Instead, best fitting of $S_{emp}(p_k)$ to $S_{theo}(p_k)$ gives back $r \to 0^-$ for each dataset. We suggest this computational discrepancy with theory is due to a misreading of the cut off effect, i.e. of $\hat{\xi}_p$ size. Indeed, we aim to fit an empirical curve based on a power law to a theoretical one based on a Negative Binomial which we have proved, at least asymptotically, to be a power law with an exponential cut off. Fitting output would match our analytical framework if cut off effect was null, i.e. $\hat{\xi}_p \approx 1$. However, the adopted fitting algorithm works balancing r and $\hat{\xi}_p$ size effects and this results in an overestimation of $\hat{\xi}_p$ effect. Despite of such a computational misreading, estimates are good. Comments on estimate goodness are split up according to the growth typology that datasets display.

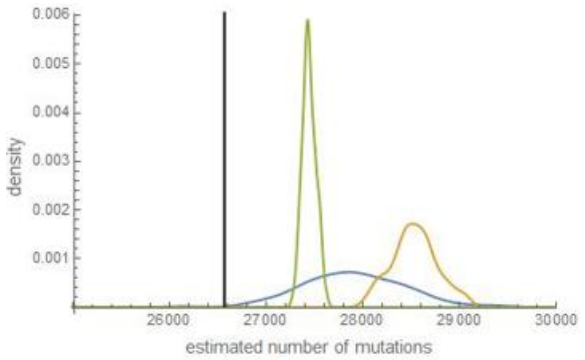
- For exponential growth datasets - simulation 1 and simulation 2 - average error on estimates is $< 9\%$ at each scale. Asymptotically, the error curves have the decreasing shapes one would expect from the theory. Indeed, as initial knowledge increases, errors and estimate dispersion decrease with estimate histograms getting sharper and narrowing the correct number to predict. Again, it means that as p increases, the probability

to randomly pick an initial training set that estimates well gets higher. Smaller scales display a counterintuitive tendency with an average error smaller than the one occurring at subsequent larger scales. Such a behaviour is due to the higher grade of bias (noticeable maximum error and flatter estimate histogram) that the estimates at these scales undergo. It follows that the effect of the overestimating training sets are balanced by that of the underestimating ones and this leads to a small error in average but also to a small probability of randomly picking a good training set.

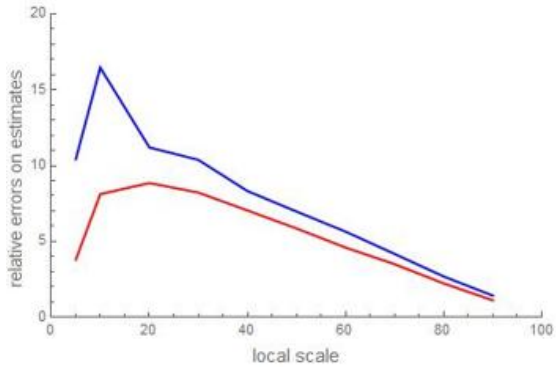
- For polynomial growth datasets - simulation 3 and simulation 4 - outputs are slightly more unstable. When starting local scales belong to $\{\frac{50}{100}, \frac{60}{100}, \frac{70}{100}, \frac{80}{100}, \frac{90}{100}$ we recognize a decreasing and promising behaviour with both average and maximum errors less than 10%. Method results to be not reliable at smallest scales when an error of $\approx 50\%$ may occur for each dataset. We guess such a worst output could be due to the m-interval we have chosen for the extended Negative Binomial. Indeed, the complete polynomial growth datasets displayed a ≈ 4-exponent power law behaviour that we have read as ≈ 1.78-exponent one by neglecting the noise and the initial perturbation. This could have led to an incorrect distribution normalization. Moreover, observe that here underestimates and overestimates are both present up to $p = \frac{50}{100}$, meaning that at those scales the method is a bit unstable. It needs a higher level of initial knowledge to perform well.



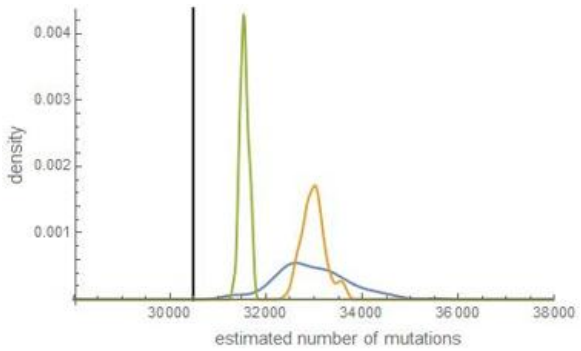
(a) Average errors for simulation 1



(b) Estimates histograms at different scales for simulation 1

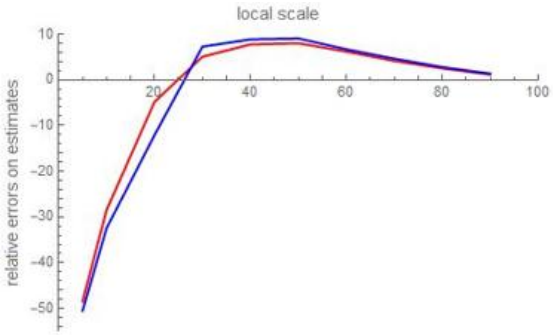


(c) Average errors for simulation 2

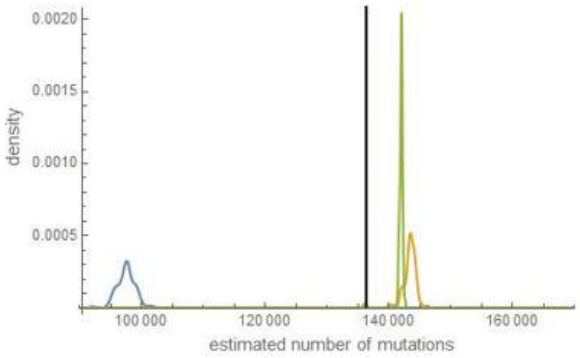


(d) Estimates histograms at different scales for simulation 2

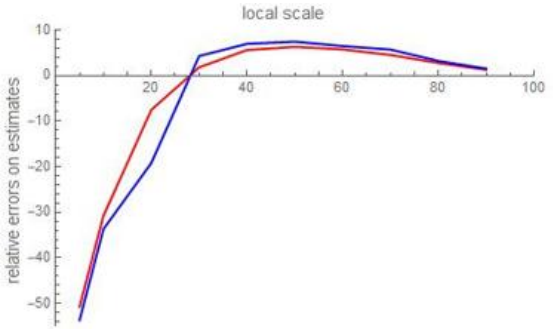
Figure 4.4: **Errors and estimates for simulated tumor exponential growth datasets.** Panels on the left hand side show average relative error (red line) and average maximum error (blue line) as local sampling increases while those on the right provide a graphical representation of estimates distribution at different local scales - $\frac{10}{100}$ (lightblue curve), $\frac{30}{100}$ (orange curve), $\frac{70}{100}$ (green curve). Asymptotically both errors decrease constantly and smooth histograms narrow the correct number to predict (black vertical line).



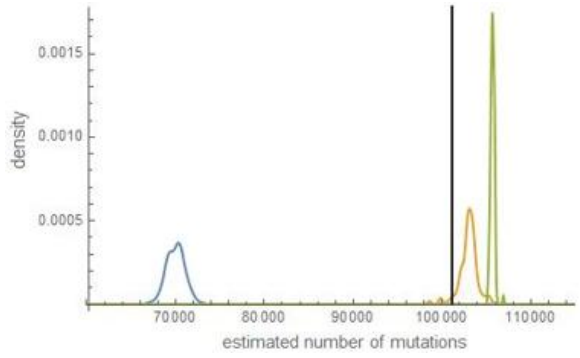
(a) Average errors for simulation 3



(b) Estimates histograms at different scales for simulation 3



(c) Average errors for simulation 4



(d) Estimates histograms at different scales for simulation 4

Figure 4.5: **Errors and estimates for simulated tumor polynomial growth datasets.** Panels on the left hand side show the average relative error (red line) and the average maximum error (blue line) at different local scales. Panels on the right display the estimates distribution at $p = \frac{10}{100}$ (lightblue curve), $\frac{30}{100}$ (orange curve), $\frac{70}{100}$ (green curve). At larger scales $\left(\geq \frac{50}{100}\right)$ an expected decreasing behaviour arises, while for smaller scales an unstable tendency with both underestimates and overestimates characterizes our results.

Conclusions

To summarize, in this thesis we have adapted a statistical framework, firstly developed for ecological purposes, to genomics to infer global information on mutation occurrences and distribution from a limited number of local samples. Based on Negative Binomial distribution and its properties, our method can count on two strengths:

- Flexibility → depending on distribution parameter values, Negative Binomial, together with its Log-Series special case and its extended version, can reproduce many SADs shapes. Internal modes, power-laws with exponential cut off and their mixtures are among SAD behaviours that can be easily accommodated;
- Scale-invariance → Negative Binomial, Log-Series, m-extended Negative Binomial and their mixtures are all self-similar under binomial sampling with parameters at two different scales linked each other through a computable formula.

Moreover, our analytical framework ends with an estimator carrying all the attracting properties it should: consistency and unbiasedness. Thus, it can be profitably used for predictions. Performed tests on both DNA polymorphisms and simulated spatial tumor growth datasets have given back promising outputs with relative errors on estimates less than 10% at each scale. Asymptotically, as the sample size increases, our method estimates tend to be concentrated around the correct number to predict and both average and maximum relative errors decrease constantly.

The statistical method we have proposed in the present thesis could be theoretically stretched out and other simulations could be conducted to address interlinked questions.
The identification of clonal mutations is crucial in medicine because therapies have as main target the most spread genetic alterations within a tumor, i.e. the clonal ones. Thus, we would be interested in being able of quantify the chance for a mutation to be present in almost all cancer cells starting from its local occurrence to provide a probabilistic framework for those targeting settings.

Another future perspective of our method could be the characterization of the local scale size needed to obtain a desired final error on the estimate of the number of global mutations. Fixed an initial threshold, we would like to understand a priori which size of the local scale gives back an error minor than the threshold. We guess that such an aim could be pursued by performing a sufficient high number of tests at different local scales and deducing an empirical categorization from them. Moreover, the computational tests on spatial tumor growth datasets have revealed that final relative errors are below 10% in most cases, but also that errors display a fluctuating behaviour with both overestimates and underestimates

before a predictable asymptotic trend. Thus, another interesting issue to tackle could be the definition of a confidence interval for our estimator, capable of signaling us if underestimation or overestimation will result at defined local scales. This statistical problem could be very difficult to solve theoretically, computational tests may help us in providing a plausible interval.

Moreover, the promising results on both the analyzed typologies of datasets have legitimated the usage of our method in oncology. However, its application to real tumor datasets is a bit more delicate because of the small size of those datasets and of the not neglectable correlation factor among mutations. In Chapter 5 we have justified the computational impossibility to regard the available datasets on human biopsies as global scales to test our upscaling method within a comfort framework in which we do have quantitative knowledge at any scales. However, the legitimacy that our tests have provided to our method allows to use those available biopsies as local scale in order to predict the mutation occurrences at the whole tumor scale (whose quantitative information are unknown). In this case, great attention must be paid to the correlation degree that the initial, available biopsies have. Indeed, if a training set contains two or more biopsies highly correlated, then the initial quantitative knowledge decreases, i.e. those biopsies do no provide new information, and method would work with an actually smaller scale leading to an higher error on estimates in general. Thus, a characterization of training sets is needed to make sub-sampling capable of capturing the real mutation profile of the tumor. We guess that studying cancer phylogenetic trees may provide this characterization: if some biopsies belong to the same tree branch, then they are affected by a not negligible correlation and not reliable estimates may be given back by training sets containing them.

Appendix

Below the R and Mathematica codes used for our tests are reported.

- **R code for empirical SAC when DNA single-nucleotide polymorphism datasets are analyzed.** Remind that here sub-sampling consists on randomly selecting columns and that the number of analyzed sub-sampling combinations to obtain an average behaviour is cut to 100.

```
# Import the VCF table and drop the qualitative information to obtain the
    binary matrix on presence/absence information on mutations at global
    scale.
import_dataset = read.table("dataset name.vcf")
global_binarymatrix = import_dataset[,10:55]

# Compute the number of clonal mutations by summing by rows and comparing
    the abundance of each mutation with the number of columns. Subtract the
     clonals from the global number of mutations , i.e. the row length of
    the global matrix, to get the number of mutations to predict.
global_occurrences_summation = rowSums(global_binarymatrix)
number_of_clonals = 0
for (i in 1:length(global_occurrences_summation))
{if (global_occurrences_summation[i]==ncol(global_binarymatrix)
  {number_of_clonals = number_of_clonals+1}}
number_to_estimate = length(global_occurrences_summation)-number_of_
    clonals

# Chose the size of the local scale $(\frac{k}{46}$ and initialize the
    test matrix (empirical_SAC_k). It is a 100x2k matrix that would be
    filled by inserting the k empirical SAC points obtained from each
    simulation and the k random columns composing the local scale at
    subsequent matrix rows.
k = sampling size
empirical_SAC_k = matrix(0,nrow=100,ncol=2*k)

# For each simulation, i.e. for each of the 100 local scale considered,
    randomly sample k columns from the 46 available and extract the
    corresponding local scale.
for (a in 1:100)
{columns_selection = sample(1:ncol(global_binarymatrix),k)
localmatrix_with_clonals = global_binarymatrix[,columns_selection]

# Compute the clonal mutations at local scale by summing by rows and
    comparing each abundance total with the k size of the local scale.
    Delete from the local matrix the rows corresponding to clonal mutations
    .
localclonal_mutations = c()
```

```
localclonals_sum = rowSums(localmatrix_with_clonals)
for (s in 1:nrow(global_binarymatrix))
{if (localclonals_sum[s]==k)
{localclonal_mutations = c(localclonal_mutations,s)}}
loc_binarymatrix = localmatrix_with_clonals[-localclonal_mutations,]

# Compute the empirical SAC by sub-sampling the local scale at any
    intermedium column size scale {1, 2, ..., k}. For each of this latter
    scale, extract all the possible subsets having cardinality equal to the
     current sub-sampling column number if they are <100 (sub-sampling of
    size 1, k-1, k), otherwise (sub-sampling of size 2, ..., k-2) extract
    100 subsets of the suitable cardinality by randomly sampling the
    columns from the local matrix. For each subset of any sub-samples
    compute the number of mutations present in the subset by summing by
    subset rows (rowSums) and account only for non zero totals with length
    (...[...!=0]) command. Fill the (current simulation number row, sub-
    sampling column size column) cell of the empirical_SAC_k with the
    average of the number of mutations present in the 100 subsets having
    the same cardinality.
first_SAC_point = rep(0,k)
for (b in 1:k)
{first_SAC_point[b] = sum(loc_binarymatrix[,b])}
empirical_SAC_k[a,1]= sum(first_SAC_point)/k
internalSACpoints = rep(0,100)
for (j in 2:(k-2))
{for (z in 1:100)
{random_subsampling = sample(1:k,j)
subsamplingdata = loc_binarymatrix[,random_subsampling]
subsampling_occurrences = rowSums(subsamplingdata)
internal_SACpoints[z] = length(subsampling_occurrences[subsampling_
    occurrences!=0])
}
  empirical_SAC_k[a,j] = sum(internal_SACpoints)/100
}
penultimate_SACpoint = rep(0,k)
for (c in 1:k)
{current_subsampling = loc_binarymatrix[,-c]
mutation_occurrences = rowSums(current_subsampling)
penultimate_SACpoint[c] = length(mutation_occurrences[mutation_occurrences
    !=0])
}
empirical_SAC_k[a,(k-1)] = sum(penultimate_SACpoint)/k
locscale_occurrences = rowSums(loc_binarymatrix)
last_SACpoint = length(locscale_occurrences[locscale_occurrences!=0])*1
empirical_SAC_k[a,k] = last_SACpoint

# For each simulation track the k column randomly extract from the global
    matrix to compound the initial local matrix by filling the last k
    columns of the empirical_SAC_k with the indexes of the extracted
    columns.
for (d in 1:k)
{empirical_SAC_k[a,k+d] = columns_selection[d]}
}
```

- **Mathematica code for empirical SAC when spatial constrained tumor growth datasets are analyzed.**
  Here local scale is defined by randomly sampling among mutation occurrences. Note that simulations with increasing local scales $p = \frac{k}{100}$, $k \in \{1, 10, 20, ..., 80, 90\}$ have been conducted separately to lighten the computational load. Moreover, for a better visualization of computational steps at each local scale sub-samplings are treated separately as well.

```
(*Import the mutation frequencies occurring in the whole tumor cells
    spatial simulation and compute the number of mutations to predict.*)
mutationglobalfrequencies=Import["mutation frequencies file.txt path","
    List"];
numbertopredict=Length[mutationglobalfrequencies];

(*Create the vector of global abundances (samplingvector) by numerically
    labelling all the mutation present at global scale and repeating the
    corresponding mutation index for a number of times equal to the
    mutation abundance.*)
samplingvector=Flatten[Table[
   Table[i, {y, 1, mutationglobalfrequencies[[i]]}], {i, 1,
    Length[mutationglobalfrequencies]}]];

(*Initialize the sub-sampling tables corresponding to an initial knowledge
     of 1%, 10%, 20%, 30%, ..., 100% of the global information and the
    empirical SAC tables for each of the 100 simulations to 0. *)
1percentpoint = Table[0, {i, 1, 100}];
10percentpoint = Table[0, {i, 1, 100}];
20percentpoint = Table[0, {i, 1, 100}];
30percentpoint = Table[0, {i, 1, 100}];
40percentpoint = Table[0, {i, 1, 100}];
50percentpoint = Table[0, {i, 1, 100}];
60percentpoint = Table[0, {i, 1, 100}];
70percentpoint = Table[0, {i, 1, 100}];
80percentpoint = Table[0, {i, 1, 100}];
90percentpoint = Table[0, {i, 1, 100}];
100percentpoint = Table[0, {i, 1, 100}];

1percentSACpoints = Table[0, {i, 1, 100}];
10percentSACpoints = Table[0, {i, 1, 100}];
20percentSACpoints = Table[0, {i, 1, 100}];
30percentSACpoints = Table[0, {i, 1, 100}];
40percentSACpoints = Table[0, {i, 1, 100}];
50percentSACpoints = Table[0, {i, 1, 100}];
60percentSACpoints = Table[0, {i, 1, 100}];
70percentSACpoints = Table[0, {i, 1, 100}];
80percentSACpoints = Table[0, {i, 1, 100}];
90percentSACpoints = Table[0, {i, 1, 100}];
100percentSACpoints = Table[0, {i, 1, 100}];

(*For each of the 100 simulation, randomly extract the k% of the vector of
     global abundances (samplingvector) to get the local scale (
    ksizesampling) corresponding to a k% of initial knowledge.*)
For[j = 0, j < 100, j++;
 ksizesampling =
  RandomSample[samplingvector, Round[k*Length[samplingvector]/100]];
```

```
(*For each sub-sampling corresponding to a portion equal to the 1%, 10%,
    20%, ... 100% of the local scale, randomly extract for 100 times a % of
     ksizesampling equal to the current sub-sampling scale and delete from
    such a sub-vector the mutations present <= 1 times (mutations whose
    index is repeated less than 2 times) using DeleteCases. Count with
    BinCounts all the different mutation indexes present in the local
    abundances vector and fill the corresponding %percentdata vector at the
     position given by the current extraction number with that number. Fill
     the corresponding %percentSACpoint with the average of %percentadata (
    having length equal to the number of extraction = 100) at each
    simulation position.*)
For[i = 0, i < 100, i++;
 subsampling1 =
  RandomSample[ksizesampling, Round[1*Length[ksizesampling]/100]];
 1percentinitialdata =
  BinCounts[
   subsampling1, {-0.5, Length[mutationglobalfrequencies] + 0.5, 1}];
 delete1 = DeleteCases[1percentdata, 0];
 1percentdata =
  BinCounts[delete1, {0.5, Max[1percentinitialdata] + 0.5, 1}];
 1percentpoint[[i]] = Total[1percentdata];
1percentSACpoints[[j]] = Mean[1percentpoint];

For[i = 0, i < 100, i++;
 subsampling10 =
  RandomSample[ksizesampling, Round[10*Length[ksizesampling]/100]];
 10percentinitialdata =
  BinCounts[
   subsampling10, {-0.5, Length[mutationglobalfrequencies] + 0.5, 1}];
 delete10 = DeleteCases[10percentdata, 0];
 10percentdata =
  BinCounts[delete10, {0.5, Max[10percentinitialdata] + 0.5, 1}];
 10percentpoint[[i]] = Total[10percentdata];
10percentSACpoints[[j]] = Mean[10percentpoint];

For[i = 0, i < 100, i++;
 subsampling20 =
  RandomSample[ksizesampling, Round[20*Length[ksizesampling]/100]];   20
      percentinitialdata =
  BinCounts[
   subsampling20, {-0.5, Length[mutationglobalfrequencies] + 0.5, 1}];
 delete20 = DeleteCases[20percentdata, 0];
 20percentdata =
  BinCounts[delete20, {0.5, Max[20percentinitialdata] + 0.5, 1}];
 20percentpoint[[i]] = Total[20percentdata];
20percentSACpoints[[j]] = Mean[20percentpoint];

For[i = 0, i < 100, i++;
 subsampling30 =
  RandomSample[ksizesampling, Round[30*Length[ksizesampling]/100]];
 30percentinitialdata =
  BinCounts[
   subsampling30, {-0.5, Length[mutationglobalfrequencies] + 0.5, 1}];
 delete30 = DeleteCases[30percentdata, 0];
 30percentdata =
  BinCounts[delete30, {0.5, Max[30percentinitialdata] + 0.5, 1}];
 30percentpoint[[i]] = Total[30percentdata];
```

```
30percentSACpoints[[j]] = Mean[30percentpoint];

For[i = 0, i < 100, i++;
 subsampling40 =
  RandomSample[ksizesampling, Round[40*Length[ksizesampling]/100]];
 40percentinitialdata =
  BinCounts[
   subsampling40, {-0.5, Length[mutationglobalfrequencies] + 0.5, 1}];
 delete40 = DeleteCases[40percentdata, 0];
 40percentdata =
  BinCounts[delete40, {0.5, Max[40percentinitialdata] + 0.5, 1}];
 40percentpoint[[i]] = Total[40percentdata]];
40percentSACpoints[[j]] = Mean[40percentpoint];

For[i = 0, i < 100, i++;
 subsampling50 =
  RandomSample[ksizesampling, Round[50*Length[ksizesampling]/100]];
 50percentinitialdata =
  BinCounts[
   subsampling50, {-0.5, Length[mutationglobalfrequencies] + 0.5, 1}];
 delete50 = DeleteCases[50percentdata, 0];
 50percentdata =
  BinCounts[delete50, {0.5, Max[50percentinitialdata] + 0.5, 1}];
 50percentpoint[[i]] = Total[50percentdata]];
50percentSACpoints[[j]] = Mean[50percentpoint];

For[i = 0, i < 100, i++;
 subsampling60 =
  RandomSample[ksizesampling, Round[60*Length[ksizesampling]/100]];
 60percentinitialdata =
  BinCounts[
   subsampling60, {-0.5, Length[mutationglobalfrequencies] + 0.5, 1}];
 delete60 = DeleteCases[60percentdata, 0];
 60percentdata =
  BinCounts[delete60, {0.5, Max[60percentinitialdata] + 0.5, 1}];
 60percentpoint[[i]] = Total[60percentdata]];
60percentSACpoints[[j]] = Mean[60percentpoint];

For[i = 0, i < 100, i++;
 subsampling70 =
  RandomSample[ksizesampling, Round[70*Length[ksizesampling]/100]];
 70percentinitialdata =
  BinCounts[
   subsampling70, {-0.5, Length[mutationglobalfrequencies] + 0.5, 1}];
 delete70 = DeleteCases[70percentdata, 0];
 70percentdata =
  BinCounts[delete70, {0.5, Max[70percentinitialdata] + 0.5, 1}];
 70percentpoint[[i]] = Total[70percentdata]];
70percentSACpoints[[j]] = Mean[70percentpoint];

For[i = 0, i < 100, i++;
 subsampling80 =
  RandomSample[ksizesampling, Round[80*Length[ksizesampling]/100]];
 80percentinitialdata =
  BinCounts[
   subsampling80, {-0.5, Length[mutationglobalfrequencies] + 0.5, 1}];
 delete80 = DeleteCases[80percentdata, 0];
```

```
  80percentdata =
   BinCounts[delete80, {0.5, Max[80percentinitialdata] + 0.5, 1}];
  80percentpoint[[i]] = Total[80percentdata];
 80percentSACpoints[[j]] = Mean[80percentpoint];

 For[i = 0, i < 100, i++;
  subsampling90 =
   RandomSample[ksizesampling, Round[90*Length[ksizesampling]/100]];
  90percentinitialdata =
   BinCounts[
    subsampling90, {-0.5, Length[mutationglobalfrequencies] + 0.5, 1}];
  delete90 = DeleteCases[90percentdata, 0];
  90percentdata =
   BinCounts[delete90, {0.5, Max[90percentinitialdata] + 0.5, 1}];
  90percentpoint[[i]] = Total[90percentdata];
 90percentSACpoints[[j]] = Mean[90percentpoint];

 For[i = 0, i < 100, i++;
  subsampling100 =
   RandomSample[ksizesampling, Round[100*Length[ksizesampling]/100]];
  100percentinitialdata =
   BinCounts[
    subsampling100, {-0.5, Length[mutationglobalfrequencies] + 0.5, 1}];
  delete100 = DeleteCases[100percentdata, 0];
  100percentdata =
   BinCounts[delete100, {0.5, Max[100percentinitialdata] + 0.5, 1}];
  100percentpoint[[i]] = Total[100percentdata];
 100percentSACpoints[[j]] = Mean[100percentpoint];
]

(*Create a matrix by joining the simulation ouput %percentSACpoints and
    transpose it in order to have a final empiricalkscaleSACs matrix of
    dimension 100x11 having as rows the mutation occurrences at the sub-
    sampling  1%, 10%, 20%, ..., 100% of the initial local abundance matrix
     for 100 different local matrices randomly determined.*)
empiricalkscaleSACs=Transpose[1percentSACpoints, 10percentSACpoints, 20
    percentSACpoints, 30percentSACpoints, 40percentSACpoints, 50
    percentSACpoints, 60percentSACpoints, 70percentSACpoints, 80
    percentSACpoints, 90percentSACpoints, 100percentSACpoints]
```

- **Mathematica code for fitting and estimates.** Set empiricalSACk = empiricalSACk = empiricalkscaleSACs of above codes, we have

```
(*Define the functions for the computable formula bridging the RSA local
    parameters to those at global scale, for the theoretical SAC and for
    the 1-Negative Binomial estimator.*)
\[Xi]p[samplep_, \[Xi]_] := (samplep*\[Xi])/(1 - \[Xi]*(1 - samplep))
SACtheorical[samplep_, r_, \[Xi]_, S_] :=
 S*(1 - (1 - \[Xi]p[samplep, \[Xi]])^r)/(1 - (1 - \[Xi])^r)
\[Xi]global[\[Xi]_, p_] := \[Xi]/(p + \[Xi]*(1 - p))
Spredicted[r_, \[Xi]_, p_, Sp_] :=
 Sp*(1 - (1 - \[Xi]global[\[Xi], p])^r)/(1 - (1 - \[Xi])^r)

(*Initialize the tables for the best parameters fitting and the one for
    the estimates to 0.*)
Rk = Table[0, {i, 1, 100}];
CSIk = Table[0, {i, 1, 100}];
SPREDk = Table[0, {i, 1, 100}];

(* Fit the empirical SAC curve (points curve in bidimensional plane having
     sub-sampling size  as x axis and average number of occurred mutations,
     i.e. empriricalSACk, as y axis) to the theoretical one using
    NonlinearModelFit tool to get the best local parameters. Observe that
    such a command requires the expression for both theoretical and
    empirical SACs, the parameters domain and some initial parameter values
     to start the search.*)
Monitor[For[i = 0, i < 100, i++;
  subsetS = empiricalSACk[[i]];
  scaleS = Range[1, Length[empiricalSACk[[1]]]/11 // N;
  dataS = Table[{scaleS[[j]], subsetS[[j]]}, {j, 1, 11}] ;
  FittingSAC =
   NonlinearModelFit[
    dataS, {SACtheorical[samplep, r, \[Xi],
      empiricalSACk[[i, -1]]], -1 < r < 0 &&
      0 < \[Xi] < 1}, {{r, initialvalueforR}, {\[Xi], initialvalueforCSI
          }}, samplep];

(*The outputs of NonlinearModelFit, i.e. best fitting r and \[Xi]
    parameters, are insert into the corresponding parameter tables at the
    current simulation position and used as input by the estimator function
    . This latter requires the two local parameters, the size of the local
    scale and the number of mutation occurring at local scale (last row
    points of empriricalSACk which correspond to a sub-sampling of 100% of
    the local scale) to give back the predictions.*)
  Rk[[i]] = FittingSAC["ParameterTableEntries"][[1, 1]];
  CSIk[[i]] = FittingSAC["ParameterTableEntries"][[2, 1]];
  SPREDk[[i]] = Spredicted[Rk[[i]], CSIk[[i]], k/100, subsetS[[-1]]];
  ], i]
```

Bibliography

Philipp M. Altrock, Lin L. Liu, and Franziska Michor. The mathematics of cancer: integrating quantitative models. *Nature Reviews Cancer 15*, 2015.

Sandro Azaele, Samir Suweis, Jacopo Grilli, Igor Volkov, Jayanth R. Banavar, and Amos Maritan. Statistical mechanics of ecological systems: Neutral theory and beyond. 2016. doi: 10.1103/RevModPhys.88.035003.

Niko Beerenwinkel, Tibor Antal, David Dingli, Arne Traulsen, Kenneth W. Kinzler, Victor E. Velculescu, Bert Vogelstein, and Martin A. Nowak. The hallmarks of cancer. *PLOS Computational Biology*, 2007. doi: 10.1371/journal.pcbi.0030225.

Helen M. Byrne. Dissecting cancer through mathematics: from the cell to the animal model. *Nature Reviews Cancer 10*, 2010. doi: https://doi.org/10.1038/nrc2808.

Ketevan Chkhaidze, Timon Heide, Benjamin Werner, Marc J Williams, Weini Huang, Giulio Caravagna, Trevor A Graham, and Andrea Sottoriva. Spatially constrained tumour growth affects the patterns of clonal selection and neutral drift in cancer genomic data. *bioRxiv*, 2019. doi: 10.1101/544536. URL `https://www.biorxiv.org/content/early/2019/02/11/544536`.

Richard Durrett. *Probability Models for DNA Sequence Evolution.* second edition, 2010.

Rick Durrett and Simon Levin. Stochastic spatial models: A user's guide to ecological applications. *Philosophical Transactions of the Royal Society of London, Series B*, 343, 1994. doi: 10.1098/rstb.1994.0028.

Philippe Flajolet and Robert Robert Sedgewick. *Analytic Combinatorics.* Cambridge University Press, 2009. ISBN 978-0521898065.

Hans-Otto Georgii. *Stochastics: Introduction to Probability and Statistics.* De Gruyter, 2012.

Patsy Haccou, Peter Jagers, and Vladimir A. Vatutin. *Branching Processes - Variation, growth and extinction of populations.* Cambridge: Cambridge Univer- sity Press, 2005. ISBN 978-0-521-83220-5.

Douglas Hanahan and Robert A. Weinberg. The hallmarks of cancer. *Cell 100*, 2000. doi: 10.1016/S0092-8674(00)81683-9.

Douglas Hanahan and Robert A. Weinberg. Hallmarks of cancer: the next generation. *Cell 5*, 2011. doi: https://doi.org/10.1016/j.cell.2011.02.013.

Stephen P. Hubbell. *The Unified Neutral Theory of Biodiversity and Biogeography*. Princeton University Press, 2001. URL http://www.jstor.org/stable/j.ctt7rj8w.

Irina Kareva. What can ecology teach us about cancer? *Translational oncology 4.5*, 2011.

Motoo Kimura. Dna and the neutral theory. 1986. doi: 10.1098/rstb.1986.0012.

Caterina A. M. La Porta and Stefano Zapperi. *The Physics of Cancer*. Cambridge University Press, 2017.

Robert H. MacArthur and Edward O. Wilson. *The Theory of Island Biogeography*. Princeton University Press, 1967. URL http://www.jstor.org/stable/j.ctt19cc1t2.

Anna Tovo. *Mathematical modelling and statistics of biodiversity*. Tesi di Dottorato, Università degli Studi di Padova, 2018.

Anna Tovo, Samir Suweis, Marco Formentin, Marco Favretti, Igor Volkov, Jayanth R. Banavar, Sandro Azaele, and Amos Maritan. Upscaling species richness and abundances in tropical forests. *Science Advances*, 2017. doi: 10.1126/sciadv.1701438. URL https://advances.sciencemag.org/content/3/10/e1701438.

Anna Tovo, Marco Formentin, Samir Suweis, Samuele Stivanello, Sandro Azaele, and Amos Maritan. Inferring macro-ecological patterns from local species' occurrences. *bioRxiv*, 2019. doi: 10.1101/387456. URL https://www.biorxiv.org/content/early/2019/01/10/387456.

Igor Volkov, Jayanth R. Banavar, Stephen P. Hubbell, and Amos Maritan. Neutral theory and relative species abundance in ecology. *Nature, 1035-1037*, 2013.

Mark J. Williams, Benjamin Werner, Timon Heide, Christina Curtis, Chris P. Barnes, Andrea Sottoriva, and Trevor A. Graham. Quantification of subclonal selection in cancer from bulk sequencing data. *Nature Genetics, volume 50*, 2018. URL https://doi.org/10.1038/s41588-018-0128-6.

Aknowledgements

This work has been a long journey that could have not be possible without the contribution of many people to whom goes my sincerest gratitude.

First of all, I would like to express my deep appreciation to my supervisor, Dr. Marco Formentin, for all the support he provided me during this period, for the patience he had in reviewing this thesis and for the enthusiasm he constantly put in the project. I would like to thank my co-supervisors Dr. Anna Tovo for the precious help in facing the computational challenges occurred in the thesis and Dr. Stefano Lise for having given me the chance to carry out my master project at the Institute of Cancer Research and for having been an encouraging guide, always available for discussions and explanations. I could not have asked for better supervisors! A thank goes also to all those people at the research center in London who have welcomed me so warmly and provided me with useful suggestions for my professional future.

Besides my academic guides, I would like to thanks all my friends for their support and presence. Ila and Genny, thank you for having been my trust university fellows from the real beginning: girls, we really did it, we have really come to an end. I guess we can finally rename our Whatsapp group chat now! Marti and Jessi thank you for all the laughs and the light-heartedness we had: Torre Archimede has never looked boring or stressful with you! A heartfelt thank goes to my old "valdagnesi" friends: Andrew, Giorgia, Roby, Meg (and all the others), it could sound a bit foregone, but simply thank you for always being there for me.

Finally, I would like to thank my adorable family for loving and supporting me so much. Thank you for having been through all my worries and my complaints (poor guys) and for having shared and celebrated any single goal I pursued. In particular, I would like to thank you, Marta, my little partner in crime. I can not count the times you have encouraged and helped me in this last period (especially when "Ops, something went wrong" notices had occurred). Thank you for being my first supporter! And last but not least, Ale: I would like to thank you for filling my life of calm and optimism and for driving force to improve myself day after day.