CSIC
Consejo Superior de Investigaciones Científicas
sysbio
Institute for Integrative Systems Biology
Universitat Politècnica de València
dma
Departamento de Matemática Aplicada

Stochastic dynamic modeling of evolving gene expression programs

Lucas GOIRIZ BELTRÁN

Thesis Advisors:
Dr. Guillermo RODRIGO TÁRREGA
Prof. Dr. J. Alberto CONEJERO CASARES

Valencia, March 2025

Do not trust, verify

— vires in numeris —

Stochastic dynamic modeling of evolving gene expression programs

Lucas Goiriz Beltrán

Abstract

RNA viruses exemplify sophisticated genetic programs endowed with the capacity for change, offering a unique window into the interplay between intricate genomic design and dynamic potential. Building on a large-scale analysis of RNA virus genomic data, this work examines whether the accumulation of genetic modifications adheres to a constant-rate progression or exhibits more complex patterns. Central questions include how the distribution of genetic changes evolves over time, how distinct genetic profiles contribute to observed shifts, and whether these dynamics can be modeled by standard Poisson or Brownian processes. To address these questions, millions of publicly available sequences were processed, focusing on samples collected over an extended time period. Rigorous data filtering, multidimensional scaling, and temporal aggregation of mutational counts were employed to characterize both the mean and the variance in the number of substitutions relative to an early reference genome.

Findings reveal that while the average number of accumulated mutations increases linearly with time, variance in mutation counts follows a more intricate pattern than predicted by the classical molecular clock hypothesis. Specifically, marked accelerations of mutation rates occur when new variants, each carrying distinct genomic profiles, invade the population. Sub- or super-diffusive dynamics emerge in these variant-specific analyses, suggesting that the standard Brownian-motion model, commonly used to approximate neutral evolutionary processes, does not fully capture the underlying complexity. Instead, fractional Brownian motion with a time-dependent diffusion exponent better explains the observed discrepancies between mean and variance in mutation counts. This nuanced perspective is further supported by fluctuations in the ratio of nonsynonymous to synonymous substitutions (dN/dS), indicative of purifying selection with occasional adaptive events as new viral lineages spread.

Beyond refining long-standing molecular clock assumptions, this work contributes a novel analytical framework for understanding how

population-wide dynamics, selection pressures, and transmission bottlenecks drive evolutionary trajectories in rapidly mutating RNA viruses. The integrative methodology, merging large-scale sequence data, variance-based metrics, and advanced modeling, can inform phylodynamic studies of other fast-evolving viruses. By exposing the limitations of purely Poissonian or Brownian models, these findings highlight the need for evolutionary frameworks that incorporate anomalous diffusion processes, profile-specific shifts in rate, and the broader ecological context of viral spread. Consequently, this work offers key insights into RNA virus evolution and underscores the importance of variance-based approaches in detecting critical deviations from conventional models, expanding our theoretical and practical understanding of viral adaptability and emergence.

Modelado dinámico estocástico de programas de expresión génica evolutivos

Lucas Goiriz Beltrán

Resumen

Los virus de ARN ejemplifican programas genéticos sofisticados dotados con capacidad de cambio, ofreciendo una ventana única a la interacción entre un intrincado diseño genómico y un potencial dinámico. Basándose en un análisis a gran escala de datos genómicos de virus de ARN, el presente trabajo examina si la acumulación de modificaciones genéticas se adhiere a una progresión de tasa constante o exhibe patrones más complejos. Las preguntas centrales incluyen cómo evoluciona la distribución de cambios genéticos a lo largo del tiempo, de qué manera perfiles genéticos distintos contribuyen a los cambios observados y si estas dinámicas pueden ser modeladas por procesos estándar de Poisson o Brownianos. Para abordar estas cuestiones, se procesaron millones de secuencias disponibles públicamente, concentrándose en muestras recolectadas durante un período temporal prolongado. Se aplicaron rigurosos filtros de datos, escalado multidimensional y agregación temporal de conteos mutacionales para caracterizar tanto la media como la varianza en el número de sustituciones con respecto a un genoma de referencia inicial.

Los hallazgos revelan que, mientras el número promedio de mutaciones acumuladas aumenta linealmente con el tiempo, la varianza en los conteos de mutaciones sigue un patrón más complejo del que predice la clásica hipótesis del reloj molecular. Específicamente, se observan aceleraciones marcadas en las tasas de mutación cuando nuevas variantes, cada una portadora de perfiles genómicos distintos, invaden la población. En estos análisis específicos de variantes emergen dinámicas sub- o super-difusivas, lo que sugiere que el modelo estándar de movimiento browniano, comúnmente utilizado para aproximar procesos evolutivos neutrales, no captura completamente la complejidad subyacente. En cambio, el movimiento fraccionario browniano con un exponente de difusión dependiente del tiempo explica de manera más adecuada las discrepancias observadas entre la media y la varianza en los conteos de mutaciones. Esta perspectiva matizada se ve respaldada además por las fluctuaciones en la proporción de sustituciones no sinónimas a sinónimas (dN/dS), indicativas de una selección purificadora con eventos

adaptativos ocasionales a medida que se diseminan nuevas líneas virales.

Más allá de refinar las antiguas suposiciones del reloj molecular, este trabajo contribuye con un novedoso marco analítico para comprender cómo las dinámicas a nivel poblacional, las presiones selectivas y los cuellos de botella en la transmisión impulsan las trayectorias evolutivas en virus de ARN de rápida mutación. La metodología integradora, que fusiona datos de secuencias a gran escala, métricas basadas en la varianza y modelos avanzados, puede orientar estudios filodinámicos de otros virus de evolución acelerada. Al exponer las limitaciones de los modelos puramente Poissonianos o Brownianos, estos hallazgos resaltan la necesidad de marcos evolutivos que incorporen procesos de difusión anómala, cambios específicos en la tasa según el perfil y el contexto ecológico más amplio de la propagación viral. En consecuencia, el presente trabajo ofrece importantes aportes sobre la evolución de los virus de ARN y subraya la relevancia de los enfoques basados en la varianza para detectar desviaciones críticas de los modelos convencionales, ampliando así nuestra comprensión teórica y práctica de la adaptabilidad y emergencia viral.

Modelatge dinàmic estocàstic de programes d'expressió gènica evolutius

Lucas Goiriz Beltrán

Resum

Els virus d'ARN exemplifiquen programes genètics sofisticats dotats de la capacitat de canviar, oferint una finestra única sobre la interacció entre un disseny genòmic complex i un potencial dinàmic. Basant-se en una anàlisi a gran escala de dades genòmiques de virus d'ARN, aquest treball examina si l'acumulació de modificacions genètiques s'ajusta a una progressió de taxa constant o bé manifesta patrons més complexos. Les qüestions centrals inclouen com evoluciona la distribució dels canvis genètics al llarg del temps, com els diferents perfils genètics contribueixen als desplaçaments observats i si aquestes dinàmiques poden ser modelades per processos de Poisson o de Brown estàndard. Per abordar aquestes qüestions, es van processar milions de seqüències d'accés públic, amb un focus especial en les mostres recollides durant un període de temps prolongat. Es va dur a terme una rigorosa depuració de dades, escalat multidimensional i agregació temporal dels comptes de mutacions per caracteritzar tant la mitjana com la variància en el nombre de substitucions respecte a un genoma de referència inicial.

Els resultats revelen que, tot i que el nombre mitjà de mutacions acumulades augmenta de manera lineal amb el temps, la variància en els comptes de mutacions segueix un patró més intricada del que prediu la hipòtesi clàssica del rellotge molecular. Concretament, s'observen acceleracions marcades de les taxes de mutació quan noves variants, cadascuna amb perfils genòmics distintius, invadeixen la població. Dins d'aquestes anàlisis específiques per variant emergeixen dinàmiques sub- o super-difusives, la qual cosa suggereix que el model de moviment brownià estàndard, comunament utilitzat per aproximar processos evolutius neutres, no capta completament la complexitat subjacent. En lloc d'això, el moviment fraccionari brownià amb un exponent de difusió dependent del temps explica millor les discrepàncies observades entre la mitjana i la variància dels comptes de mutacions. Aquesta perspectiva matisada es veu reforçada per les fluctuacions en la relació de substitucions nonsinònimes a sinònimes (dN/dS), indicatives d'una selecció purificadora amb esdeveniments adaptatius ocasionals a mesura que es difonen

noves línies vírals.

Més enllà de refinar les assumpcions del rellotge molecular de llarga data, aquest treball aporta un nou marc analític per entendre com les dinàmiques a nivell poblacional, les pressions selectives i els colls de botella en la transmissió impulsen les trajectòries evolutives en virus d'ARN que muten ràpidament. La metodologia integradora, que fusiona dades de seqüències a gran escala, mètriques basades en la variància i modelatge avançat, pot informar estudis filodinàmics d'altres virus d'alta evolució. En exposar les limitacions dels models purament poissonians o brownians, aquests resultats destaquen la necessitat de marcs evolutius que incorporen processos de difusió anòmala, canvis en la taxa específics de cada perfil i el context ecològic més ampli de la propagació viral. En conseqüència, la recerca ofereix coneixements clau sobre l'evolució dels virus d'ARN i subratlla la importància d'enfocs basats en la variància per detectar desviacions crítiques dels models convencionals, ampliant finalment la nostra comprensió teòrica i pràctica de l'adaptabilitat i l'aparició de nous virus.

Contents

Chapter 1

Introduction

Above all, my life is research.
— Margarita Salas

1.1 Mathematical Modeling in Biology

Biological systems are extraordinarily complex, involving multifaceted interactions that span from molecules to entire ecosystems. Because traditional experimental approaches often cannot fully capture this complexity, mathematical modeling has become an indispensable tool for biologists to understand, simplify, and predict system behaviors. These models provide a structured framework for analyzing biological processes, generating hypotheses, and integrating experimental data into a coherent representation of larger systems. By offering quantitative predictions, mathematical models enable the testing of hypotheses under controlled *in silico* conditions, thereby uncovering insights that might be challenging to observe directly [1]. Indeed, many recent biological breakthroughs are rooted in the synergy between modeling, experimental observation, and hypothesis testing [2]. Although models focus on essential features of a system to clarify underlying principles, they rely on assumptions and simplifications that can produce inaccuracies if chosen poorly [3]. Thus, iterative model refinement and experimental validation are vital to ensure that each model accurately represents the biological phenomenon under study.

Mathematical modeling has had a profound impact on deciphering gene regulation, where it has become essential for unraveling the complex mechanisms of transcriptional control and gene expression. Gene regulation

involves a network of interactions between transcription factors, DNA, RNA, and other cellular components, collectively determining whether a gene is activated or silenced. The emergence of high-throughput genomic technologies in the early 21st century generated a substantial amount of data, spurring new advances in modeling [4]. Modeling is especially important in this context because gene expression lies at the intersection of numerous biological processes; even subtle changes in regulatory interactions can drive diseases, phenotypic variation, or evolutionary novelties [4]. By offering a quantitative framework, mathematical models simulate regulatory networks and predict how modifications in specific components influence overall gene expression [1]. These models also clarify how gene expression is coordinated from the single-cell level to entire organisms, yielding a deeper comprehension of development and cellular function. A noteworthy illustration involves modeling gene regulation in early *Drosophila melanogaster* embryogenesis, where spatial expression patterns are explained by reaction-diffusion principles that identify critical interactions [5, 6, 7]. In synthetic biology, mathematical models guide the design of gene circuits that exhibit oscillations [8] or switch-like behaviors [9], demonstrating the broad applications of these approaches.

In epidemiology, mathematical models are instrumental in studying the spread of infectious diseases and assessing the effectiveness of diverse public health interventions [10]. By simulating disease transmission within populations, these models predict outbreaks, gauge the potential outcomes of vaccination strategies, and evaluate containment measures. Over the past few decades, they have become an essential component of public health policy making [11, 12], proving indispensable during crises such as the 2014 Ebola outbreak in West Africa and the recent COVID-19 pandemic. By forecasting disease trajectories, modeling aids both immediate and longer-term policy decisions, including social distancing mandates and lockdown protocols. As infectious diseases continue to present urgent challenges, epidemiological modeling will remain a key instrument for understanding disease dynamics and guiding public health responses [10].

Mathematical modeling has had also a long and influential history in evolutionary biology, where it has been essential for elucidating the mechanisms that govern genetic change. In particular, these models help address questions that are experimentally challenging due to the long timescales characterizing evolutionary processes [3]. They enable the exploration of natural selection, genetic drift, mutation, and migration, all of which collectively determine how genetic variation changes in populations over generations [13, 14]. By incorporating quantitative equations into evolutionary theory, researchers can predict genetic trends, evaluate the forces that drive evolution, and examine the coevolution of species or the emergence

of advantageous mutations [2]. Such modeling further allows the testing of theoretical ideas about evolutionary processes, providing insights into species origins and adaptations that guide future research directions in evolutionary biology. To delve more deeply into how these processes unfold over time, it is crucial to understand evolution as a stochastic phenomenon, where random events play a significant role in shaping genetic outcomes.

1.2 Evolution as a Stochastic Process

Evolution proceeds through four fundamental processes: mutation, natural selection, genetic drift, and recombination, each of which has an inherent random component that lends itself to stochastic models [15, 16]. Mutation introduces new variants at rates typically modeled by Poisson processes, whereas selection favors or disfavors certain variants based on fitness effects. Drift acts as a random sampling event that can fix or eliminate alleles regardless of fitness, especially in small populations, while recombination reorganizes genetic material to produce new haplotype combinations.

Markov models frequently capture these processes by assigning state-dependent transition rates for sequence changes, and continuous-time Markov chains (CTMCs), for instance, quantify the instantaneous substitution rates among nucleotides or amino acids. Traditionally, these models have been used to describe the evolution of DNA sequences. Formally, a Markov chain is a stochastic process $\{X_t, t \geq 0\}$ characterized by the Markov property, which states:

$$P(X_{t+s} = j \mid X_t = i, X_u = x_u, u < t) = P(X_{t+s} = j \mid X_t = i) \tag{1.1}$$

for all states i, j and times $t, s \geq 0$. The process is fully described by its transition probabilities $P_{ij}(t) = P(X_{t+s} = j \mid X_s = i)$, which satisfy the Chapman-Kolmogorov equations:

$$P_{ij}(t+s) = \sum_k P_{ik}(t) P_{kj}(s), \quad \text{for all } t, s \geq 0. \tag{1.2}$$

In a continuous-time context, transitions occur according to rates defined by a rate matrix $Q = \{q_{ij}\}$, where each off-diagonal entry q_{ij} represents the instantaneous rate of moving from state i to state j, and diagonal entries are defined such that $q_{ii} = -\sum_{j \neq i} q_{ij}$. Thus, the transition probability matrix over time t can be computed through matrix exponentiation:

$$P(t) = e^{Qt}. \tag{1.3}$$

To integrate Markov chains into the evolutionary model, we consider the following explanatory scheme:

1. **State Definition**: Define each state of the Markov chain to represent a particular genetic configuration or nucleotide/amino acid sequence.
2. **Transition Rates**: Specify transition rates q_{ij} based on biological mechanisms such as mutation rates, selection coefficients, and drift probabilities.
3. **Rate Matrix Construction**: Assemble the rate matrix Q encapsulating all possible instantaneous transitions between genetic states.
4. **Transition Probabilities Calculation**: Calculate transition probability matrices $P(t)$ using matrix exponentiation to predict evolutionary dynamics over time.
5. **Integration into Phylogenetic or Population Genetic Models**: Employ these computed transition probabilities in models such as Wright-Fisher, Moran, or phylogenetic inference frameworks to estimate divergence times, genetic variability, or ancestral states.

Although Markov models are powerful tools for describing evolutionary processes, their simplest forms make assumptions about rate homogeneity across time and states, and may fail to capture the full complexity of biological systems. For instance, as multiple generations pass, deterministic selection interacts with stochastic mutation and drift, resulting in intricate patterns of divergence. Models such as Wright-Fisher and Moran (including diffusion and coalescent extensions) offer a theoretical foundation for quantifying these random dynamics [17]. Consequently, beneficial mutations can still be lost if chance events override selection in finite populations [18], illustrating how randomness profoundly influences molecular evolution.

Population genetics predominantly examines allele frequency changes within species, relying on models like Wright-Fisher or coalescent theory to incorporate drift, selection, and mutation in a finite population [19]. By tracking the probabilities of fixation or loss over discrete generations, population-genetic analyses elucidate microevolutionary dynamics, including the relative roles of neutrality and adaptation. Phylogenetics, conversely, focuses on substitutions that accumulate between diverging lineages, often modeling sequence evolution with continuous-time Markov chains along the branches of a tree. Here, each branch length reflects the expected number of substitutions per site, and likelihood-based or Bayesian methods assess different phylogenetic hypotheses [20]. Although both fields consider the

same fundamental processes, they differ in timescale and scope: population genetics addresses polymorphisms within populations, while phylogenetics assumes fixed differences between species. Furthermore, recombination complicates phylogenetic inference by creating discordant gene trees, whereas in population genetics it may accelerate adaptation or break up linkage. Recent methodological advances bridge these perspectives, for instance by integrating coalescent theory into species-tree approaches that accommodate incomplete lineage sorting. Ultimately, both population genetics and phylogenetics leverage stochastic models to reconstruct evolutionary history, yet they do so with distinct focuses on the micro versus macro dimensions of biological change. To date, the two fields are complementary and not in competition. A core framework for both fields involves the molecular clock, which uses evolutionary rates to date divergence times across lineages.

1.3 The Concept of the Molecular Clock

The molecular clock hypothesis traces back to the 1960s when Zuckerkandl and Pauling observed a roughly consistent rate of molecular changes and proposed that these changes could be treated as a "clock" for dating evolutionary events [21]. Kimura's Neutral Theory later provided theoretical support by suggesting that the majority of substitutions are selectively neutral and accumulate at a near-constant rate through genetic drift [22]. Early strict clock models treated substitutions as a Poisson process with a uniform rate across all lineages [23], but empirical evidence often contradicted this rigidity. Consequently, relaxed clocks were introduced to permit variable substitution rates across branches within a statistically coherent framework [24].

In a strict molecular clock framework, the accumulation of substitutions is modeled by a Poisson process. Let $N(t)$ denote the number of substitutions that occur along a lineage during a time interval t. Assuming a constant substitution rate μ (substitutions per unit time), the probability that exactly k substitutions occur in time t is given by the Poisson probability mass function:

$$P\left(N(t) = k\right) = \frac{(\mu t)^k e^{-\mu t}}{k!}, \quad k = 0, 1, 2, \ldots$$

Here, the expected number of substitutions is

$$\mathbb{E}[N(t)] = \mu t,$$

and due to the properties of the Poisson distribution, the variance is also

$$\mathbb{V}(N(t)) = \mu t.$$

This formulation encapsulates the idea that both the mean and the variance of the substitution process increase linearly with time. However, empirical studies have noted that observed variance often exceeds the mean (overdispersion), suggesting that factors such as lineage-specific rate heterogeneity, episodic selection, and other evolutionary forces may cause deviations from the strict Poisson assumption [25, 26]. Such observations have motivated the development of relaxed clock models, which allow for rate variation among lineages while maintaining a statistically coherent framework for divergence time estimation.

Estimating these rates and divergence times typically employs likelihood or Bayesian methods, with Bayesian approaches (including MCMC in BEAST) yielding comprehensive posterior distributions of tree topology, rates, and node ages [27]. Nonetheless, current models exhibit inherent limitations: lineage rate heterogeneity can induce overdispersion [25], while temporal rate shifts complicate calibrations that presume uniformity [26]. External factors, episodic selection, and calibration errors further exacerbate these biases, stressing the importance of careful model selection, robust calibrations, and thorough statistical analyses. To address these challenges, refined techniques such as penalized likelihood, correlated and uncorrelated relaxed clocks, life-history trait covariates, and partitioned models for distinct genes have been developed. Although these approaches mitigate biases, they can be computationally demanding when handling extensive datasets. Despite these hurdles, the molecular clock remains invaluable for reconstructing deep evolutionary timelines and guiding phylodynamic studies of fast-evolving pathogens, highlighting both its enduring utility and the necessity for ongoing methodological improvements. In many instances, studies of RNA viruses offer particularly clear examples of these principles.

1.4 Viruses as complex gene expression programs: RNA viruses as a case study

RNA viruses exhibit exceptionally high mutation rates and rapid replication dynamics that render them ideal models for testing molecular clock hypotheses. A virus's genomic architecture is composed of multiple genes, many of which are multifunctional, that form a complex regulatory circuit with numerous interactions with the host cell. In combination with their intricate regulatory circuitry, these mutations set the stage for swift evolutionary divergence. Moreover, their short generation times and large population sizes further facilitate measurable divergence, producing a clear temporal signal in their genome sequences [28]. Consequently, RNA viruses are classified as 'measurably evolving populations", where genetic divergence correlates

directly with elapsed time, enabling precise calibration of molecular clocks [29]. This real-time evolution provides a natural laboratory for studying substitution rate variability, selective pressures, and the underlying dynamics of sequence evolution. As such, researchers can rigorously test and refine theoretical models by directly observing evolutionary processes over short time scales, thereby bridging theoretical frameworks with empirical data. Continuous monitoring of viral genomes not only validates molecular clock assumptions but also enhances our understanding of broader evolutionary and epidemiological phenomena, underscoring the critical role these viruses play in advancing molecular evolutionary research.

In the realm of public health, RNA viruses represent significant threats due to their rapid evolution and widespread transmission, as exemplified by the ongoing SARS-CoV-2 pandemic. High-throughput sequencing technologies have revolutionized viral surveillance by enabling rapid, cost-effective sequencing of whole genomes directly from clinical samples. This surge in genomic data empowers researchers to perform time-resolved phylogenetic analyses and robust mathematical modeling, which are crucial for understanding transmission dynamics and predicting epidemic trends [30, 31]. Phylodynamic approaches integrate evolutionary data to infer key epidemiological parameters such as outbreak origins, transmission rates, and timing of introductions. These models have been instrumental in guiding interventions, ranging from travel restrictions and quarantine measures to vaccine deployment strategies. Specifically, SARS-CoV-2 has been extensively monitored using millions of genome sequences, allowing scientists to track the emergence and spread of variants in near real-time. The integration of genomic and epidemiological metadata further enhances forecasting capabilities, ensuring that public health responses are both timely and data-driven [32, 33, 34].

1.5 Research Gaps and Motivation

Despite significant advances, evolutionary models remain limited by simplifying assumptions that may not fully capture complex biological realities. Classical models often assume independent loci, constant selection pressures, or simplified recombination dynamics, yet empirical studies reveal pervasive epistasis, variable selection, and linkage effects that challenge these assumptions [35, 36]. Refining models to incorporate these factors requires a balance between mathematical tractability and biological realism. Recent approaches leverage stochastic processes, Bayesian inference, and simulation-based methods to enhance predictive accuracy [37, 38]. However, the stochastic frameworks employed in these models typically incorporate

white noise, Gaussian delta-correlated fluctuations, due to its mathematical convenience. While this choice has facilitated substantial advances, it may not capture the full complexity of biological systems. There is growing reason to believe that alternative, more intricate noise models could more accurately represent the stochasticity inherent in evolutionary processes. Furthermore, computational demands grow exponentially, necessitating efficient algorithms and scalable implementations.

Addressing these challenges requires an interdisciplinary approach that integrates evolutionary theory, applied mathematics, and computer science. Advances in Monte Carlo simulations, machine learning, and high-performance computing now enable inference of complex models, but accessibility remains a concern [39]. Open-source software plays a pivotal role in democratizing these tools, ensuring reproducibility and fostering collaboration [40, 41]. Platforms such as BEAST, SLiM, and Approximate Bayesian Computation (ABC) frameworks exemplify how theoretical models can be practically implemented for real-world data analysis. The ongoing challenge lies in developing flexible, scalable, and biologically realistic models that bridge theoretical derivations with empirical data, ultimately refining our understanding of molecular evolution.

References

[1] Banwarth-Kuhn M, Sindi SS (2020) How and why to build a mathematical model: A case study using prion aggregation. *Journal of Biological Chemistry*, 295(15): 5022–5035.

[2] Servedio MR, Brandvain Y, Dhole S, et al (2014) Not just a theory - the utility of mathematical models in evolutionary biology. *PLoS Biology*, 12(12): e1002017.

[3] Brown JM, Thomson RC (2018) Evaluating model performance in evolutionary biology. *Annual Review of Ecology, Evolution, and Systematics*, 49: 95–114.

[4] Ay A, Arnosti DN (2011) Mathematical modeling of gene expression: a guide for the perplexed biologist. *Critical Reviews in Biochemistry and Molecular Biology*, 46(2): 137–151.

[5] Reinitz J, Sharp DH (1995) Mechanism of eve stripe formation. *Mechanisms of Development*, 49(1–2): 133–158.

[6] Jaeger J, *et al.* (2004) Dynamic control of positional information in the early Drosophila embryo. *Nature*, 430(6997): 368–371.

[7] Jaeger J, *et al.* (2004) Dynamical analysis of regulatory interactions in the gap gene system of drosophila melanogaster. *Genetics*, 167(4): 1721–1737.

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

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

[10] Siettos CI, Russo L (2013) Mathematical modeling of infectious disease dynamics. *Virulence*, 4(4): 295–306.

[11] Temime L, Hejblum G, Setbon M, Valleron AJ (2008) The rising impact of mathematical modelling in epidemiology: antibiotic resistance research as a case study. *Epidemiology and Infection*, 136(3): 289–298.

[12] Lozano MA, *et al.* (2021) Open Data Science to Fight COVID-19: Winning the 500k XPRIZE Pandemic Response Challenge. *Machine Learning and Knowledge Discovery in Databases Applied Data Science Track Springer International Publishing*: 384–399.

[13] Johri P, *et al.* (2022) Recommendations for improving statistical inference in population genomics. *PLOS Biology*, 20(5): e3001669.

[14] Rosenberg NA, Stadler T, Steel M (2025) A mathematical theory of evolution: Phylogenetic models dating back 100 Years. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 380(1919): 20230297.

[15] Arenas M (2015) Trends in substitution models of molecular evolution. *Frontiers in Genetics*, 6: 319.

[16] Siepel A (2009) Phylogenomics of primates and their ancestral populations. *Genome Research*, 19(11): 1929–1941.

[17] Groffiths R, Marjoram P (1996) Ancestral inference from samples of DNA sequences with recombination. *Journal of Computational Biology*, 3(4): 479–502.

[18] Haldane JBS (1937) The Effect of Variation on Fitness. *The American Naturalist*, 71(735): 337–349.

[19] Hein J, Schierup M, Wiuf C (2004) Gene genealogies, variation and evolution: a primer in coalescent theory. *Oxford University Press, USA*.

[20] Kosiol C, Goldman N (2011) Markovian and non-Markovian protein sequence evolution: aggregated Markov process models. *Journal of Molecular Biology*, 411(4): 910–923.

[21] Zuckerkandl E, Pauling L (1965) Evolutionary divergence and convergence in proteins. *Evolving Genes and Proteins*: 97–166.

[22] Bromham L, Penny D (2003) The modern molecular clock. *Nature Reviews Genetics*, 4(3): 216–224.

[23] Takahata N (2007) Molecular clock: An anti-neo-darwinian legacy. *Genetics*, 176(1): 1–6.

[24] Drummond AJ, Ho SYW, Phillips MJ, Rambaut A (2006) Relaxed phylogenetics and dating with confidence. *PLoS Biology*, 4(5): e88.

[25] Gillespie JH (1984) The molecular clock may be an episodic clock. *Proceedings of the National Academy of Sciences*, 81(24): 8009–8013.

[26] Ho SYW, *et al.* (2011) Time-dependent rates of molecular evolution. *Molecular Ecology*, 20(15): 3087–3101.

[27] Kumar S, Hedges SB (2016) Advances in time estimation methods for molecular data. *Molecular Biology and Evolution*, 33(4): 863–869.

[28] Holmes EC (2009) The Evolution and Emergence of RNA Viruses. *Oxford University Press*.

[29] Grenfell BT, *et al.* (2004) Unifying the epidemiological and evolutionary dynamics of pathogens. *Science*, 303(5656): 327–332.

[30] Grubaugh ND, *et al.* (2019) Tracking virus outbreaks in the twenty-first century. *Nature Microbiology*, 4(1): 10–19.

[31] Pollett S, *et al.* (2020) Genomic epidemiology as a public health tool to combat mosquito-borne virus outbreaks. *Journal of Infectious Diseases*, 221(Suppl 3): S308–S318.

[32] Hill V, *et al.* (2021) Progress and challenges in virus genomic epidemiology. *Trends in Parasitology*, 37(12): 1038–1049.

[33] Gire SK, *et al.* (2014) Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. *Science*, 345(6202): 1369–1372.

[34] Luksza M, Lässig M (2014) A predictive fitness model for influenza. *Nature*, 507(7490): 57–61.

[35] Lunzer M, Golding GB, Dean AM (2010) Pervasive cryptic epistasis in molecular evolution. *PLoS Genetics*, 6(10): e1001162.

[36] Neher RA, Shraiman BI (2009) Competition between recombination and epistasis can cause a transition from allele to genotype selection. *Proceedings of the National Academy of Sciences*, 106(16): 6866–6871.

[37] Csilléry K, Blum MGB, Gaggiotti OE, François O (2010) Approximate Bayesian computation (ABC) in practice. *Trends in Ecology & Evolution*, 25(7): 410–418.

[38] Meyer X, Chopard B, Salamin N (2017) Accelerating Bayesian inference for evolutionary biology models. *Bioinformatics*, 33(5): 669–676.

[39] Bryant D (2017) Special Issue: Mathematical and Computational Evolutionary Biology—2015. *Systematic Biology*, 66(1): 1–3.

[40] Bouckaert R, *et al.* (2014) BEAST 2: A software platform for Bayesian evolutionary analysis. *PLoS Computational Biology*, 10(4): e1003537.

[41] Haller BC, Messer PW (2019) SLiM 3: Forward genetic simulations beyond the Wright-Fisher model. *Molecular Biology and Evolution*, 36(3): 632–637.

Chapter 2

Objectives

The evolutionary dynamics of rapidly mutating biological systems involve complex interactions between localized genetic events and broader population-level outcomes, resulting in behaviors that are not well-captured by classical evolutionary models. Quantifying these processes necessitates robust modeling frameworks, a detailed understanding of gene-level heterogeneity, and scalable analytical methods. Therefore, this thesis has the following objectives:

1. Investigate the underlying stochastic processes and localized mutation dynamics in rapidly mutating biological systems, assessing how these micro-level events collectively influence macroscale evolutionary outcomes.

2. Evaluate selective pressures across diverse genomic segments and identify deviations from classical evolutionary models by employing alternative quantitative frameworks to elucidate patterns of adaptation and constraint.

3. Verify whether the molecular clock hypothesis holds in rapidly mutating biological systems or if significant deviations arise due to localized genetic variability.

4. Develop and implement integrated computational and statistical methodologies, coupled with novel large-scale open-source data treatment approaches, to bridge localized evolutionary events with comprehensive genomic-scale dynamics while ensuring reproducibility and scalability.

Chapter 3

A variant-dependent molecular clock with anomalous diffusion models SARS-CoV-2 evolution in humans

Simplicity is prerequisite for reliability.
— Edsger W. Dijkstra

This work has been published in a peer-reviewed journal. See the full citation:

Goiriz L, Ruiz R, Garibo-i-Orts O, Conejero J.A, Rodrigo G. (2023) A variant-dependent molecular clock with anomalous diffusion models SARS-CoV-2 evolution in humans. *Proc. Natl. Acad. Sci. U.S.A.*, 120(30): e2303578120.

In this publication, I performed all the data analysis and figures, except for **Figure 3.2** where OGO made contributions in the data analysis. The results were discussed with RR, JAC and GR. GR designed the research.

3.1 Introduction

Viruses lie at the frontier of living and inert matter, as they lack own metabolism to sustain replication but are subject to Darwinian evolution (*i.e.*, mutation and selection) [1]. As fast evolving biological agents [2], they are ideal substrates from which to learn mechanisms that modulate genetic variation, as well as to test theoretical models of evolution. One important model is the molecular clock hypothesis, introduced by Emile Zuckerkandl and Linus Pauling, which dates back to early times of molecular biology.

Definition 3.1.1 (Molecular Clock Hypothesis (1965))**.** The molecular clock hypothesis states that the rate of evolution of a gene is constant over time. This implies that genes accumulate mutations at a constant rate, providing a molecular basis for measuring evolutionary timescales [3, 4, 5].

Motoo Kimura further developed this theory into a comprehensive framework in known as the neutral theory of molecular evolution.

Definition 3.1.2 (Neutral Theory of Molecular Evolution (1968))**.** The neutral theory of molecular evolution proposes that most evolutionary changes at the molecular level are caused by random genetic drift of selectively neutral mutants [6, 7].

Neutral theory of molecular evolution predicts, in addition, that such clocks are Poissonian stochastic processes (*i.e.*, evolution seen as a Brownian motion with diffusivity such that mean and variance are equal) [7], a concept that will be elaborated upon in **Section 3.4.10**. Despite the results from seminal studies of some viral genes are in tune [8, 9], the molecular clock hypothesis still raises controversy [5], as evolution appears as a highly volatile and vagary stochastic process due to environmental changes, transmission bottlenecks, and recombination and speciation events. Indeed, such a null model can be rejected in numerous cases [10], and overdispersed populations in genetic variation (*i.e.*, with larger variance than mean) seem common across phyla [11]. Nonetheless, without extensive monitoring of evolution in natural conditions for a reasonable period of time, it is difficult to describe the mathematical model underlying such stochastic dynamics.

The emergence (at the end of 2019) and global spread (during 2020) of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has caused the ongoing pandemic [12]. Due to the high impact on public health, huge efforts have been carried out worldwide to sequence the whole viral genome from different patients in real time of pandemic [13, 14, 15, 16, 17, 18, 19]. To date, more than 10 million sequences are available. Importantly, this has allowed tracking introductions and measuring the extent of virus transmission at the community level [13], assessing the effectiveness of containment strategies [14],

realizing the impact of superspreading events [15], identifying novel variants of concern and their key mutations [16], monitoring spatiotemporal invasion dynamics of specific variants [17], disclosing co-infection occurrences with different variants [18], and discovering the circulation of recombinant viruses, which might lead to increased infectiousness or pathogenicity [19]. In addition, computational analyses have shown that SARS-CoV-2 primary evolves under purifying selection in humans, with some sites displaying adaptation [20]. They have also shown that the spectrum of acquired mutations is greatly asymmetric (*viz.*, dominated by C>U and G>U substitutions) [21]. Notwithstanding, a detailed study of the stochastic dynamics underlying such evolutionary motion to assess fundamental scaling laws and conditions for criticality has not been carried out.

In this work, we exploited the unprecedented monitoring of the evolution of a human virus in nature to conduct a study aimed at describing this process as a (non-)Brownian motion, considering the number of acquired mutations as the displacement of the viral particle from the origin (**Fig. 3.1**). For that, several biostatistical analyses over millions of whole genome sequences at the ensemble level were evaluated on the basis of a time-dependent probabilistic mathematical model, without relying on phylogeny. Of note, our results challenge the conventional molecular clock hypothesis by providing new theoretical foundations for viral evolution.

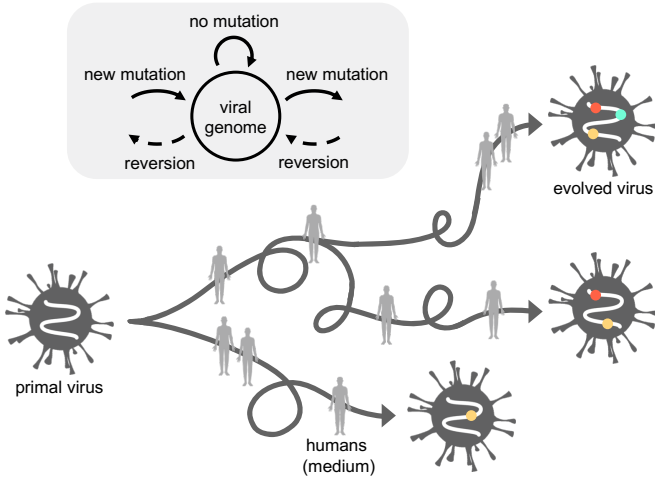

Figure 3.1: Schematics of the evolutionary motion of the virus (viewed as a stochastic process). Inset: associated state-transition diagram.

3.2 Results

We sought to characterize the mean and variance (mean squared displacement) of the overall stochastic process by which the observable viral genome accumulates mutations with time (since the emergence in Wuhan, China). This was modeled in a continuous form by the Langevin equation:

$$\frac{dm(t)}{dt} = \kappa + \xi(t), \tag{3.1}$$

where $m(t)$ is the total number of mutations in the genome at time t, κ the evolution rate (which could be time-dependent), and $\xi(t)$ an integrative noise source whose properties shape the evolutionary motion.

Due to the large number of available SARS-CoV-2 sequences from United Kingdom (UK), our analysis was focused on this country. Using landmark multidimensional scaling (LMDS) [22], we obtained a representation of all available genotypes in a two-dimensional space (**Fig. 3.2**), which served to appreciate the virus evolution as a complex diffusion process. In this polar plot, the radius represented the number of mutations and the angle encompassed the rest of sequence variation.

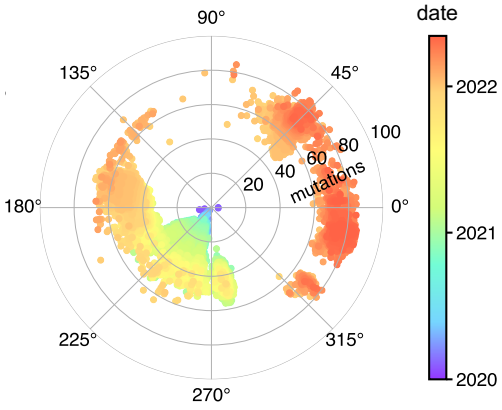

Figure 3.2: 2D projection of all viral sequences colored by date.

To characterize the stochastic process, we first quantified the rate at which the viral genome accumulates a mean number of mutations with time. Considering all types of mutations and discretizing time by weeks (*i.e.*, all sequences available in a week were pooled together), we obtained a macroscopic evolution rate of 0.62 wk^{-1} (Pearson's correlation with no intercept, $P < 10^{-4}$; **Fig. 3.3a**). Substitutions were much more frequent than insertions and deletions (indels). However, at some points (at the end of 2020 and of 2021), an acceleration in the evolution rate was observed, thereby deviating from a

molecular clock model with constant rate. Yet, without phylogenetic inference, this picture just reflected population changes and not strict evolutionary paths. In addition, mutations were classified according to their type (*viz.*, non-coding, synonymous, non-synonymous, and indels), and the ratio between the number of nonsynonymous and synonymous substitutions per site (dN/dS) was estimated (**Fig. 3.3b**), which is a common tool to assess the strength and mode of natural selection [23]. The observed dN/dS signature (fluctuation around 1 over time) suggested evolution under purifying selection of a series of adapted variants.

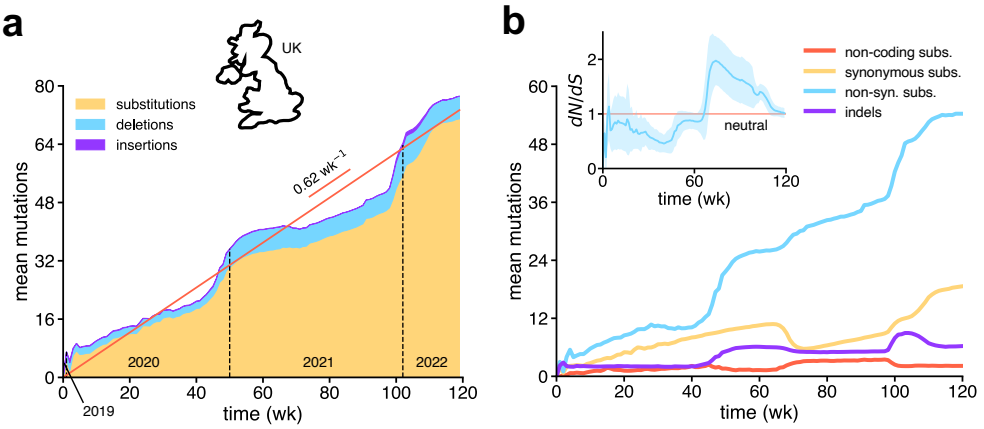

Figure 3.3: a) Time-course of the mean number of accumulated mutations in the viral genome, distinguishing between substitutions, deletions, and insertions. Linear regression over the total shown in red ($R^2 = 0.95$). b) Time-course of the mean number of non-coding substitutions, synonymous substitutions, non-synonymous substitutions, and indels. Inset: dN/dS with time (mean plus/minus standard deviation).

To test whether the accumulation of mutations in SARS-CoV-2 was a Poissonian stochastic process, we also calculated the variance and the dispersion index, understood as the ratio between variance and mean (**Fig. 3.4a**). The study of the variance is often overlooked, despite it is essential to comprehend the evolutionary motion. We found largely sub-Poissonian dynamics (*i.e.*, dispersion index < 1) with two main dispersion bursts at the times at which the evolution rate was accelerated. To inspect the origin of such a dynamic profile, we performed a sequence classification into variants. For simplicity, four variants were considered, *viz.*, Primal, Alpha, Delta, and Omicron.

We realized that the first dispersion burst corresponded to the transition from Primal to Alpha, while the second to the transition from Delta to Omicron (**Fig. 3.4b**). The number of new coronavirus disease 2019

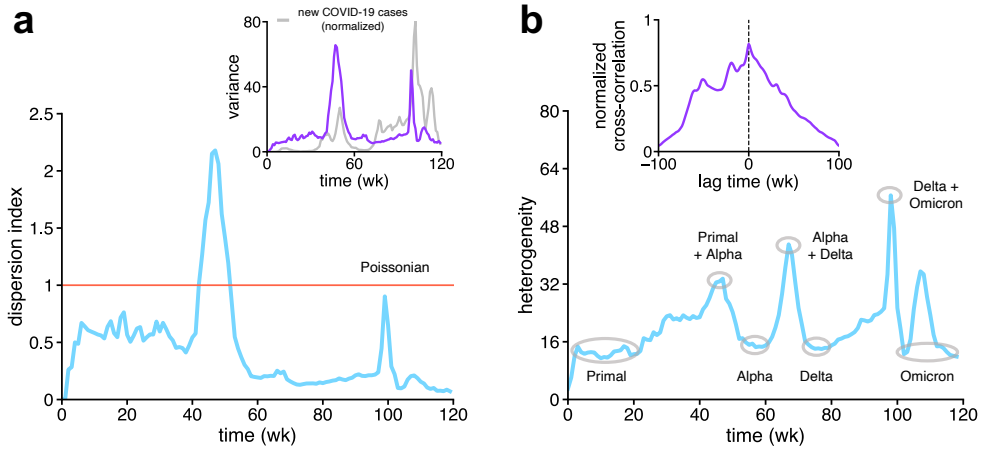

Figure 3.4: a) Time-course of the dispersion index (variance/mean). Inset: variance with time; the normalized number of new COVID-19 cases is superimposed. b) Time-course of the degree of heterogeneity (mean Hamming distance), showing the different stages of the virus population in terms of variants. Inset: normalized cross-correlation between variance and heterogeneity with time.

(COVID-19) cases also correlated with the variance (inset of **Fig. 3.4a**). Representing the distributions of accumulated mutations with time, we disclosed a bimodal behavior during such transitions (**Fig. 3.5a, b**), explaining the increased dispersion. The invading genotypes carried about 15-20 more mutations on average. Moreover, the transition from Alpha to Delta only generated a slight dispersion signal because both variants carried a similar number of mutations. Arguably, outlier SARS-CoV-2 genotypes in the existing distribution at a time led to the emergence of new variants, and the observed accelerations in evolution rate came from the inherent stochasticity of the evolutionary motion followed by rapid, mostly deterministic invasion events once a particular genotype acquired a selective advantage, such as higher transmissibility [17].

Due to the virus population reset caused by the invasion of a new variant, we calculated the time-dependent statistics per variant. The analyses conducted for each variant were independent from each other by considering subsets of properly annotated sequences (*i.e.*, no evolutionary paths between variants were considered). Of note, the evolution rates of Primal, Alpha, and Delta were substantially lower (up to 0.36 wk^{-1}) than the inferred macroscopic value of 0.62 wk^{-1} (**Fig. 3.6a**), in agreement with previous estimates following phylogenetic methods [24]. In the dataset, Omicron was composed of two lineages with sufficient dissimilarity, *viz.*, BA.1 and BA.2 (BA.1 displaced Delta and BA.2 displaced BA.1). Performing a decomposition, we observed that BA.1 evolved faster than BA.2 in UK.

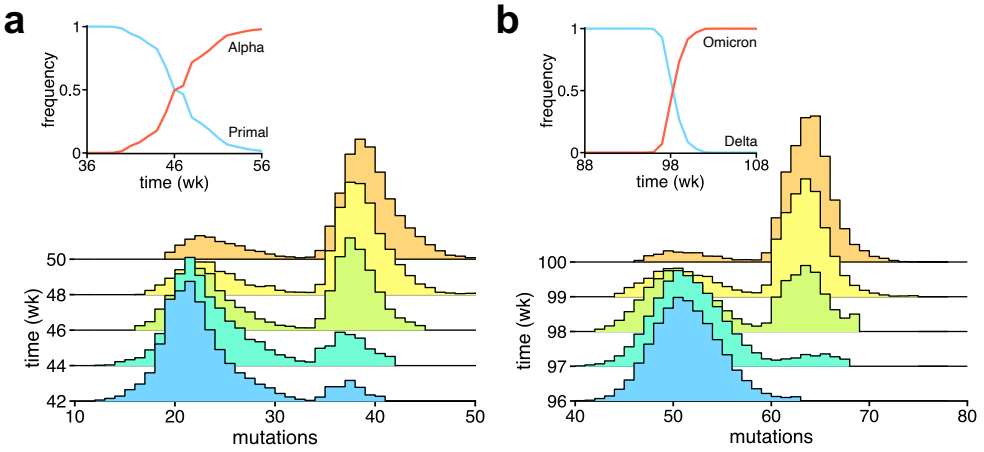

Figure 3.5: Probability-based histograms of the number of accumulated mutations during the transition from Primal to Alpha (a) and Delta to Omicron (b). Insets: population frequency of the variants with time.

Collectively, the mean evolutionary motion was well captured by

$$\mathbb{E}\left[m(t)\right] = \kappa t, \tag{3.2}$$

with a constant variant-dependent rate, considering $\mathbb{E}\left[\xi(t)\right] = 0$ (Pearson's correlations, $P < 10^{-4}$ in all cases). In addition, we found significant nonlinear dependencies of the variance with time in all cases (Fisher-Snedecor's F tests, $P < 10^{-4}$ for Primal, Alpha, and Delta, $P = 0.027$ for Omicron BA.1, and $P = 0.0003$ for Omicron BA.2; **Fig. 3.6b**), which indicated a stochastic behavior with anomalous diffusion [25]. In other words, SARS-CoV-2 underwent a non-Brownian evolutionary motion. This exciting result entailed that the explorations of the genotypic space by the virus to discover new phenotypes at different times were not fully uncorrelated within a clade.

To provide a quantitative picture of the process, we fitted $\mathbb{V}\left[m(t)\right]$ to the general expression Dt^{α}, where D is the diffusion coefficient and α the diffusion exponent. This expression is derived using the approaches shown in **Section 3.4.10** by considering

$$\mathrm{Cov}\left[\xi(t), \xi(s)\right] = \frac{1}{2} D\alpha(\alpha - 1)\left|t - s\right|^{\alpha - 2} \tag{3.3}$$

as the covariance function of the noise source $\xi(t)$. We found subdiffusion ($\alpha = 0.42$, $\alpha = 0.47$, and $\alpha = 0.28$, respectively) in the cases of Primal, Alpha, and Omicron BA.1, while weak superdiffusion ($\alpha = 1.34$) in the case of Delta (Pearson's correlations in log scale, $P < 10^{-4}$ for Primal, Alpha, and

Delta and $P = 0.020$ for Omicron BA.1; **Fig. 3.7a**). Although not plotted, we also found subdiffusion for Omicron BA.2 ($\alpha = 0.37$). The robustness of these results was assessed by bootstrapping, *i.e.*, performing a sampling with replacement of the sequences available each week in the original dataset and recomputing the dynamic profile of the variance. This also allowed dealing with the sequence pseudoreplication issue due to a shared history. Tolerable uncertainties for the diffusion parameters were noticed (inset of **Fig. 3.7a**).

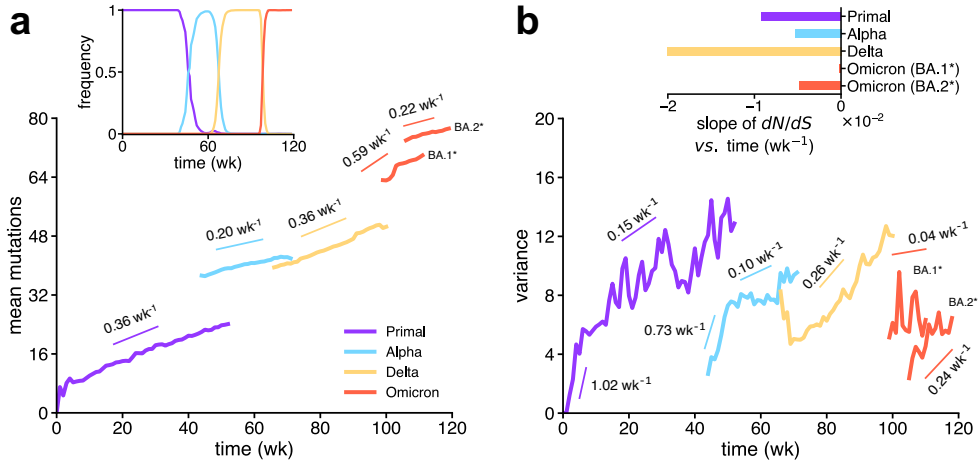

Figure 3.6: a) Time-course of the mean number of accumulated mutations per variant (Omicron decomposed into BA.1 and BA.2). Linear regressions shown in each case ($R^2 \geq 0.90$). Inset: population frequency of the variants with time. b) Time-course of the variance per variant. Piecewise linear regressions shown in each case. Inset: slope of dN/dS with time for each variant obtained by linear regression.

To inspect the origin of anomalous diffusion in evolutionary motion, the rate at which the dN/dS ratio changed with time was analyzed per variant (inset of **Fig. 3.6b**). We observed a decreasing trend in all cases, more pronounced for Delta. This suggested that Delta evolved by accumulating more synonymous mutations per site than the other variants. If these mutations were neutral [26], the evolved genotypes of Primal, Alpha, and Omicron BA.1 would be more constrained as a result of their non-synonymous mutations, thereby explaining, at least in part, the observed subdiffusion patterns. Furthermore, we calculated a reset dispersion index, considering the accumulation of mutations since the appearance of the variant of study (*i.e.*, each time a new variant invades the population, the number of mutations is reset). At long times, we found values in the neighborhood of 1, revealing an asymptotic Poissonian behavior following this metric (**Fig. 3.7b**).

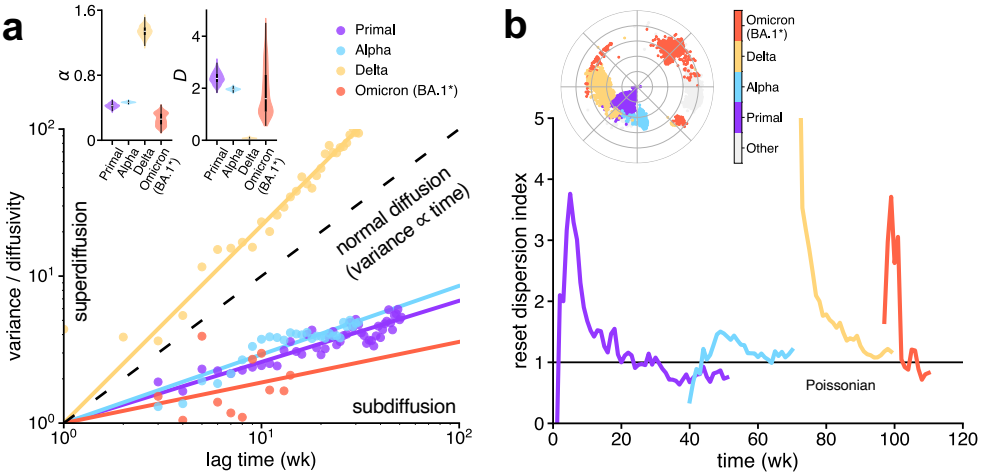

Figure 3.7: a) Representation of the rescaled variance normalized by the diffusivity (D) with time in log scale (points correspond to data; for Primal, Alpha, Delta, and Omicron BA.1, R^2 = 0.87, 0.88, 0.94, and 0.37, respectively, relative to Pearson's correlations in log scale). The slope of the fitted lines (α) defines the type of diffusion ($\alpha > 1$ superdiffusion, $\alpha < 1$ subdiffusion). Shaded areas represent the 95% confidence intervals of the regression lines. Inset: distributions of values for each diffusion parameter (violin plots) obtained by bootstrapping. b) Time-course of the reset dispersion index per variant. Inset: 2D projection of all viral sequences colored by variant (projection as in **Fig. 3.2**). The variant-specific analyses were restricted to the time period in which their population frequency was at least 10%.

3.3 Discussion

The observation of patterns of anomalous diffusion in biology has opened new avenues of research [25]. Intriguingly, recent studies in which the physical movement of single SARS-CoV-2 virions was monitored throughout the infectious cycle highlighted transient and variant-dependent directionality and confinement outside and inside the cell [27, 28], indicating deviation from a pure Brownian motion. Here, we have presented a new application domain in evolution. Of note, we uncovered that a probabilistic model with constant variant-dependent evolution rate and nonlinear mutational variance with time explained the SARS-CoV-2 evolutionary motion in humans during the first 120 weeks of pandemic in UK. This model might be used to refine phylodynamic approaches aimed at understanding the spread and adaptation of the virus. As shown, canonical descriptions based on the Poisson distribution do not accurately capture the observed dispersion at all time points.

These findings can also be situated within classical evolutionary theory, particularly Fisher's geometric model (FGM), which provides a quantitative

framework to describe how organisms adapt in high-dimensional trait spaces.

Definition 3.3.1 (Fisher's Geometric Model)**.** Fisher's Geometric Model (FGM) conceptualizes the phenotype of an organism as a point $x \in \mathbb{R}^n$ in an n-dimensional Euclidean trait space, where each dimension represents an independent quantitative trait under stabilizing selection. The phenotypic optimum, corresponding to maximal fitness, is located at the origin of this space. The fitness of a phenotype x is given by:

$$W(x) = W_0 \exp\left(-\frac{\|x\|^2}{2\sigma^2}\right), \tag{3.4}$$

where $W_0 > 0$ denotes the maximal attainable fitness, σ^2 controls the strength of stabilizing selection (i.e., the curvature of the fitness landscape), and $\|x\|^2$ is the squared Euclidean distance from the phenotypic optimum. Mutational effects are modeled as random, isotropic displacements in phenotype space. The model predicts that the probability of a mutation being beneficial decreases as the phenotype approaches the optimum and increases with trait-space dimensionality n.

Mutations correspond to random displacements Δx in phenotype space, drawn from a symmetric distribution. The likelihood that a mutation increases fitness decreases with proximity to the optimum and increases with dimensionality n, due to the geometry of high-dimensional spaces. As a result, FGM predicts a declining rate of adaptation as populations approach a fitness peak [29], along with a predominance of small-effect beneficial mutations. These properties make it a useful null model for interpreting adaptive dynamics in evolving populations [30].

In light of FGM, our findings offer both congruences and challenges. The subdiffusive mutational dynamics observed for Alpha and Omicron BA.1 variants are consistent with populations exploring a constrained phenotypic region near a fitness optimum, where evolutionary motion becomes increasingly restricted, an outcome that aligns with FGM's expectation of diminishing returns on adaptation. In contrast, the weak superdiffusive behavior seen in Delta reflects a more expansive phenotypic trajectory, possibly indicating that this variant originated from a population still far from the optimum and with greater adaptive potential. However, the punctuated emergence of variants with large mutational jumps, often involving dozens of substitutions, suggests evolutionary leaps not easily reconciled with FGM's assumption of small-effect mutations and a static fitness landscape [31]. These deviations may stem from shifts in host immunity, transmission dynamics, or selective pressures that effectively relocate the optimum over time. Furthermore, FGM's assumption of isotropic, additive mutational

effects is likely violated in the context of SARS-CoV-2, where epistasis and context-dependent interactions are pervasive.

While FGM provides a valuable conceptual scaffold for interpreting our results within classical evolutionary theory, it falls short of fully capturing the complexity of SARS-CoV-2 dynamics. A more complete model would require generalizations that account for moving or rugged fitness landscapes, pervasive epistasis, and non-additive mutational effects, factors increasingly recognized as central to viral evolution. In parallel it is worth to note the bias in this type of studies caused by the fact that most of the sequenced viral genomes came from symptomatic infected people. Another issue is the imbalance in sequencing effort among countries, which prevents performing comprehensive analyses at a global scale. Further studies are required to assess the potential impact of movement and contact restrictions, vaccination, and self-diagnostic testing on the observed dynamic patterns. Overall, we anticipate deep implications of our data-driven results for future evolutionary and genomic studies, especially when dealing with fast evolving biological agents such as viruses.

3.4 Materials and Methods

3.4.1 Whole-genome sequencing data

The nucleotide sequences of the SARS-CoV-2 genomes used in this study were retrieved from the GISAID database (`https://www.gisaid.org`). As of May 2022, 10791877 sequences and the corresponding metadata were downloaded. Our analysis was restricted to the data from UK, which consists of 2735543 sequences.

3.4.2 Pairwise sequence alignments

The nucleotide sequences of the SARS-CoV-2 genomes were aligned against a reference genome by means of Multiple Alignment using Fast Fourier Transform (MAFFT) [32]. The results were collected in Clustal format. In this work, the reference sequence (root) was `hCoV-19/Wuhan/IVDC-HB-01/2019` (`EPI_ISL_402119`), which has 100% identity with the GeneBank reference genome (`NC_045512.2`) as shown by a Clustal Omega [33] alignment.

3.4.3 Construction of a functional dataset

For each sequence, the number of mutations (substitutions, insertions, and deletions) with respect to the reference SARS-CoV-2 genome were counted. This information was retrieved from the MAFFT output alignments. In

addition, the sequence collection dates and Pangolin lineages were retrieved from the metadata. Next, the sequences with unreliable recorded dates, whose unresolved base content surpassed 1% (proportion of `Ns`), or whose size was below 25 kb were discarded, as they were considered of low quality. Duplicated entries in the dataset were also removed. Furthermore, sequences isolated from non-human hosts were discarded. Then, the variant names were assigned where applicable based on the Pangolin lineage annotation. The Pangolin lineage-to-variant mapping was performed with information available at the Cov-lineages initiative (https://cov-lineages.org). All sequences dated earlier than 21 February 2021 were annotated as Primal variant (*i.e.*, the original SARS-CoV-2 variant from Wuhan). The sequences were grouped by weeks. Finally, for each week, mutation outliers were filtered out to avoid artifacts in the calculated statistical parameters. These outliers could originate from incorrect date annotations, aberrant evolutionary trajectories, or sudden point introduction [34, 35]. For each week, if the number of sequences was greater than 20, a Generalized Extreme Studentized Deviate (GESD) many-outlier procedure was applied [36]. Otherwise, a filtering based on interquartile ranges was performed (*i.e.*, the upper/lower bound was equal to the first/third quartile point plus/minus the interquartile range).

All data analyses were performed in Python using the libraries Pandas (https://pandas.pydata.org), NumPy (https://numpy.org), SciPy (https://scipy.org), Scikit-learn (https://scikit-learn.org), and Biopython (https://biopython.org).

3.4.4 Landmark multidimensional scaling

Landmark multidimensional scaling (LMDS) is a variation of classical multidimensional scaling (MDS) [37] used to analyze and visualize dissimilarities between items based on a set of pairwise distance measures. The technique uses a small number of landmark items to compute their pairwise distances and estimate the distances between the remaining items, which are then mapped into a low-dimensional space using MDS [22]. LMDS presents several advantages over traditional MDS, including reduced computational complexity, scalability, and flexibility, making it an appropriate tool for analyzing large and complex datasets.

LMDS and principal component analysis (PCA) are both techniques used for dimensionality reduction, but they have some fundamental differences in their goals and methods. In contrast to LMDS, PCA is used to identify the underlying structure in a dataset by finding the principal components that explain the most variance in the data, capturing this way the most important patterns in the data [38]. The interpretation of the coordinates in the projection space of both techniques is different as well, with PCA

representing the patterns of variation and MDS representing the similarities and dissimilarities among the data points.

To obtain a representation of all available sequences in a two-dimensional (2D) space, a procedure based on landmark multidimensional scaling (LMDS) was followed [22]. For that, the Hamming distance between any two sequences was calculated (given by the number of mutations that separate each other). Metric axioms (minimality, symmetry, and triangle inequality) hold for the Hamming distance, so LMDS can be applied. The following sequences were used as landmarks:

hCoV-19/England/CAMC-C42AEA/2020	hCoV-19/England/LSPA-2E824E8/2021
hCoV-19/England/PHEC-YYBI3UW/2021	hCoV-19/Scotland/NORT-YBF4CD/2021
hCoV-19/England/LSPA-3DC3179/2022	hCoV-19/England/ALDP-3A3CE1D/2022
hCoV-19/England/ALDP-2E0DCFC/2021	hCoV-19/Wales/PHWC-PYDUBM/2021
hCoV-19/England/QEUH-F8AA01/2021	hCoV-19/Scotland/QEUH-9AD0C0/2020

Table 3.1: Sequence identifiers of landmarks employed during LMDS analysis.

Polar coordinates were used to project the sequences in a 2D space. For the sake of interpretability, the radius was directly the total number of mutations from root and the angle was obtained from the coordinates generated by LMDS. It is worth noting that the position of two sequences in the projection plane will be determined by their similarity according to the Hamming distance. Consequently, if two sequences have the same number of mutations but these mutations are different, they will be located in different regions of the plane (i.e., their angles will be different and the radius will be the same).

3.4.5 Calculation of statistical parameters

For each set of SARS-CoV-2 sequences in a week, the mean and variance of the number of accumulated mutations (including substitutions and indels) were computed. For this computation, only the number of mutations was considered, not their type (*i.e.*, different sequences with the same number of mutations counted the same). Then, the dispersion index was calculated, defined as the ratio between variance and mean. In addition, the mean Hamming distance between all sequence pairs in a week was computed to realize the extent of sequence heterogeneity. To compute the normalized cross-correlation between mutational variance and sequence heterogeneity, the `correlate` function from NumPy was used, having previously divided the statistical parameters by their norm (using the `linalg.norm` function). Finally, the number of COVID-19 cases in UK was retrieved from the database Our World in Data (https://ourworldindata.org/). The weekly number

of new cases was computed. Probability-based histograms of the total number of mutations were obtained with the NumPy `histogram` function (`density=True`).

The calculation of mutational mean and variance was also performed per variant. This was done for the major lineages Primal, Alpha, Delta, and Omicron [39], considering only the time period in which the variant represented at least the 10% of the population. This limit was applied to avoid artifacts in the calculated statistical parameters due to a low number of sequences. In the case of Omicron, the calculation was performed for the sublineages BA.1 and BA.2 because of their great difference in mutations [40]. For each variant, a reset dispersion index was also calculated, defined as the ratio between variance and the mean number of accumulated mutations since the first appearance of the variant in the population (*i.e.*, each time a new variant invades the population, the number of mutations is reset). To some extent, this is in tune with the definition of a founder genotype for each clade from which to start counting as done in ref. [41].

Specifically, if there are N_k sequences in the k^{th} week, the mean number of accumulated mutations in that week, denoted by $\mathbb{E}[m_k]$, is calculated as follows:

$$\mathbb{E}[m_k] = \frac{1}{N_k} \sum_{i=1}^{N_k} m_{k,i} \quad , \tag{3.5}$$

where $m_{k,i}$ is the number of mutations of the i^{th} sequence in the k^{th} week. And the unbiased variance, denoted by $\mathbb{V}[m_k]$, is calculated as

$$\mathbb{V}[m_k] = \frac{1}{N_k - 1} \sum_{i=1}^{N_k} \left(m_{k,i} - \mathbb{E}[m_k]\right)^2 . \tag{3.6}$$

To perform the calculations for a particular variant v, only the sequences annotated as such were considered (note that one sequence is linked at most to one particular variant). If there are $N_{v,k}$ sequences in the k^{th} week for variant v, the mean number of accumulated mutations in that week, denoted by $\mathbb{E}[m_{v,k}]$ is calculated as

$$\mathbb{E}[m_{v,k}] = \frac{1}{N_{v,k}} \sum_{i=1}^{N_{v,k}} m_{v,k,i} \quad , \tag{3.7}$$

where $m_{v,k,i}$ is the number of mutations of the i^{th} sequence in the k^{th} week annotated as variant v. And the variance, denoted by $\mathbb{V}[m_{v,k}]$, is calculated as

$$\mathbb{V}\left[m_{v,k}\right] = \frac{1}{N_{v_k} - 1} \sum_{i=1}^{N_{v,k}} \left(m_{v,k,i} - \mathbb{E}\left[m_{v,k}\right]\right)^2 . \tag{3.8}$$

Given that each sequence either belongs to a single variant, or to none, as a corollary of the above definitions the following relation holds:

$$N_k = \sum_{v \in V} N_{v,k} + N_{\emptyset,k}, \tag{3.9}$$

where V is the set of variants and $N_{\emptyset,k}$ denotes the number of sequences that are not linked to any variant of V in the k^{th} week.

To assess the dispersion dynamics of the mutation distribution, the dispersion index (ρ) is used.

Definition 3.4.1 (Dispersion Index)**.** Given a stochastic process $X(t)$ with finite mean $\mathbb{E}\left[X(t)\right]$ and variance $\mathbb{V}\left[X(t)\right]$, the Dispersion Index $\rho(t)$, also known as the Fano Factor, is a measure of the relative variability of the process and is defined as

$$\rho(t) = \frac{\mathbb{V}\left[X(t)\right]}{\mathbb{E}\left[X(t)\right]}$$

Following **Definition 3.4.1**, the dispersion index in the k^{th} week (ρ_k) is calculated as

$$\rho_k = \frac{\mathbb{V}\left[m_{v,k}\right]}{\mathbb{E}\left[m_{v,k}\right]}, \tag{3.10}$$

and the reset dispersion index in the k^{th} week (ρ_k^{reset}) as

$$\rho_k = \frac{\mathbb{V}\left[m_{v,k}\right]}{\mathbb{E}\left[m_{v,k}\right] - \mu_v}, \tag{3.11}$$

where μ_v is the mean number of mutations in the week in which $\mathbb{V}\left[m_{v,k}\right]$ is minimal. This condition corresponded to the first appearance of the variant of study in the population for Primal (first instance 26 Jan 2020), Alpha (first instance 25 Oct 2020), and Omicron BA.1 (first instance 21 Nov 2021), according to our functional dataset. In the case of Delta, however, the date of minimal variance did not coincide with the date of first instance, but rather with the moment at which the AY.4 lineage became dominant (16 May 2021) after a first period of time of selection within the Delta population, so this date was used for the mutation count reset.

3.4.6 Global *vs.* variant-based analysis

The global analysis was carried out considering all available sequences from our functional dataset (*i.e.*, pooling together the sequences even if they corresponded to different variants to compute the mean and variance for each week). This analysis served to appreciate the overall evolutionary trajectory by which the observable viral genome accumulates mutations with time in UK. That is, it allowed having a bird's eye perspective. More in detail, we could calculate a macroscopic evolution rate by linear regression between mean and time, and we could evaluate the dispersion dynamics of the resulting mutation distribution by representing the variance/mean ratio with time.

By contrast, the variant-based analysis was carried out considering only the sequences corresponding to a given variant, according to the annotation. This was done for Primal, Alpha, Delta, and Omicron (distinguishing also between the BA.1 and BA.2 lineages). Our analysis showed alternation of periods of evolution at lower rates and bursts of dispersion due to invasion events. We acknowledge that previous work already showed changes in the evolution rate with time in the particular case of SARS-CoV-2 [24, 42]. However, because our study was not based on phylogeny, we were able to process all available sequences in the database, gaining accuracy. In addition, and most importantly, because the dynamic profile of the variance was also analyzed, we were able to disclose anomalous diffusion patterns. This is a notable result that may contribute to change our understanding of virus evolution.

3.4.7 Comparison with phylogenetic methods

While in this work we are interested in how a viral population evolves, phylogenetic approaches mainly focus on genotypic differences in order to reconstruct evolutionary paths. Phylogenetic methods have been applied to produce estimates of the SARS-CoV-2 evolution rate, reporting values of 0.3-0.4 wk^{-1} during the first year of pandemic [24, 43]. These values are in tune with our calculations in the case of Primal. This congruence suggested us the formation with time of a sufficiently heterogenous viral population. Indeed, using the metric of heterogeneity, once a variant was fixed, the divergence between two arbitrary sequences of the population was about 11-15 mutations.

Further phylogenetic inferences have pointed out a transient increase of the evolution rate concomitant with the emergence of new invading variants (e.g., Alpha) [42]. However, by restricting the study to the sequences within the clade, the evolution rate did not appear to increase but rather to be maintained or even reduced (e.g., in the cases of Alpha and Delta). According to our analysis and also others in the field following non-phylogenetic approaches [41], Alpha evolved a bit slower than Primal and Delta did at a similar rate.

Despite the uncertainty associated with the emergence of new variants [34], viral population- based studies are useful to understand the mutation-selection dynamics.

3.4.8 Categorization of mutations

For each viral sequence present in the functional dataset, the set of substitutions with respect to root were broken down into several categories: non-coding substitutions (*i.e.*, substitutions that fall on non-coding regions of the genome), synonymous substitutions (*i.e.*, substitutions that fall on coding regions but do not trigger amino acid changes), and non-synonymous substitutions (*i.e.*, substitutions that trigger amino acid changes in coding regions). Insertions and deletions were counted into a unique variable called indels. Finally, the weekly mean and variance in terms of number of non-coding substitutions, synonymous substitutions, non-synonymous substitutions, and indels were computed.

3.4.9 Estimation of natural selection signatures

The ratio between the number of nonsynonymous and synonimous substitutions per site (dN/dS) was used to realize the sense of natural selection [23] during SARS-CoV-2 evolution, as it is a suitable statistical parameter to estimate the balance between positive (adaptive), neutral, and negative (purifying) selection acting on a set of protein-coding genes [44]. A simple method was employed to estimate the dN/dS ratio for each viral sequence [45], assuming that:

i The total length of the protein-coding genes was constant (equal to that of the reference genome).

ii The four nucleotides had equal frequencies.

iii The substitution events were random.

First, the total number of synonymous (S) and non-synonymous (N) sites was estimated. Given that the probability of maintaining the same amino acid sequence is 5% if the substitution occurs at the first position of the codon, 0% if it occurs at the second position, and 72% if it occurs at the third position, it turns out that $S \approx (0.05+0.72)R$ and $N \approx 3(R-S)$, where R is the total length in base pairs of the protein-coding genes. Then, the proportion of synonymous (p_S) and non-synonymous (p_N) substitutions per site were computed. Second, these proportions were corrected to account for multiple potential changes at the same site. The genetic distance of synonymous (d_S) and non-synonymous (d_N) substitutions per site was estimated using the Jukes-Cantor formula.

Definition 3.4.2 (Jukes-Cantor Formula)**.** Assuming equal base frequencies and equal probability of substition occurrence between any pair of nucleotides, the genetic distance can be approximated by

$$d = -\frac{3}{4} \ln \left(1 - \frac{4}{3}p\right)$$

where d is the estimated number of substitutions per site, and p is the observed proportion of nucleotide sites at which the two sequences differ [46].

For the particular cases of synonymous (d_S) and non-synonymous (d_N) substitutions, the expression becomes

$$\begin{cases} d_S = -\frac{3}{4} \ln \left(1 - \frac{4}{3}p_S\right) \\ d_N = -\frac{3}{4} \ln \left(1 - \frac{4}{3}p_N\right) \end{cases}$$

where p_S and p_N are the observed proportions of synonymous and non-synonymous nucleotide differences, respectively.

3.4.10 Mathematical modeling of evolutionary motion

To motivate the development of a novel molecular clock model, we will briefly explore how a Poisson point process can be employed to model DNA sequence evolution. We then extend the Poisson point process into a continuous stochastic process, and finally present a model that captures the anomalous diffusion patterns observed in the evolutionary motion of SARS-CoV-2.

Evolution as a Poisson point process

The Poisson distribution is commonly used to model the occurrence of infrequent events within a fixed time or space interval. In the context of genetic mutations during DNA replication, each generation (defined as a replicative cycle) can introduce changes or substitutions in the DNA sequence due to various factors like errors induced by the DNA polymerase, radiation, or the presence of chemicals. Since the likelihood of a mutation at a specific position in the DNA sequence is assumed to be small, constant, and independent between generations (as supported by experimental evidence), the number of mutations in a lineage over n generations can be accurately described using the Poisson distribution.

Let u be the rate of mutations per generation, and n the number of generations. In this scenario, the number of mutations that occur in a lineage during these n generations follows a Poisson distribution with a mean value of un. In addition, if each generation takes the same amount of time, the number of mutations in the lineage during a specific time period t can be described as a homogeneous Poisson point process, denoted as $\{N(t), t \geq 0\}$,

where $N(t)$ represents the total number of mutations that have taken place up to (and including) time t. Consequently, the probability of observing exactly n mutations, denoted as $N(t) = n$ at time t, is given by

$$\Pr(N(t) = n) = \frac{e^{-\kappa t} (\kappa t)^n}{n!}, \tag{3.12}$$

where κ the rate of substitutions for a given unit of time. Importantly, **Equation 3.12** implies that the number of mutations in a lineage at time $t = 0$ is 0 and that the increments of the process are independent.

For further developments, it is convenient to compute the moment generating function, $M_{N(t)}(s)$, of the Poisson process:

$$\begin{aligned} M_{N(t)}(s) &= \mathbb{E}\left[e^{sN(t)}\right] \qquad (3.13)\\ &= \sum_{n=0}^{\infty} e^{sn} \frac{e^{-\kappa t} (\kappa t)^n}{n!} \\ &= e^{-\kappa t} \sum_{n=0}^{\infty} \frac{(\kappa t e^s)^n}{n!} \\ &= e^{-\kappa t} e^{\kappa t e^s} \\ &= e^{\kappa t (e^s - 1)} \qquad (3.14) \end{aligned}$$

By means of **Equation 3.14** it is straightforward to demonstrate that the mean and variance of the process are both given by κt:

$$\begin{aligned} \mathbb{E}[N(t)] &= \frac{\partial}{\partial s} M_{N(t)}(s)\Big|_{s=0} \qquad (3.15)\\ &= \left[\kappa t e^{\kappa t (e^s - 1) + s}\right]_{s=0} \\ &= \kappa t \qquad (3.16) \\ \mathbb{V}[N(t)] &= \mathbb{E}\left[(N(t) - \mathbb{E}[N(t)])^2\right] \qquad (3.17)\\ &= \mathbb{E}\left[N^2(t) - 2N(t)\mathbb{E}[N(t)] + \mathbb{E}[N(t)]^2\right] \\ &= \mathbb{E}\left[N^2(t)\right] - \mathbb{E}[N(t)]^2 \\ &= \frac{\partial^2}{\partial s^2} M_{N(t)}(s)\Big|_{s=0} - (\kappa t)^2 \\ &= \left[\kappa t (\kappa t e^s + 1) e^{\kappa t (e^s - 1) + s}\right]_{s=0} - (\kappa t)^2 \\ &= (\kappa t)^2 + \kappa t - (\kappa t)^2 \\ &= \kappa t \qquad (3.18) \end{aligned}$$

As a corollary, and recalling **Definition 3.4.1**, it is possible to assess that the process' dispersion index $\rho_{N(t)}$ is equal to 1.

Evolution approximated as a continuous stochastic process

Similar to how the Poisson distribution can be approximated by a Gaussian distribution through the central limit theorem, a Poisson point process can be approximated by a Wiener process. The Wiener process, also known as Brownian motion, is a continuous-time stochastic process characterized by independent and stationary increments. It is usually represented as $\{W(t), t \geq 0\}$, where $W(t)$ is a random variable representing the displacement of a particle at time t, its increments follow a normal distribution with a mean $\mathbb{E}[W(t)] = 0$ and, if it's the standard Wiener process, a covariance function $\mathrm{Cov}[W(t), W(s)] = \min(t, s)$. The Wiener process is widely used as a model for random fluctuations in various physical systems.

Therefore, the number of mutations during DNA replication can be thought of a kind of *random motion*, which we call *evolutionary motion*, in the abstract space of all possible DNA sequences and it is defined as the following Langevin stochastic differential equation

$$\frac{dm(t)}{dt} = \kappa + \sqrt{\kappa}\zeta(t), \tag{3.19}$$

where $\zeta(t)$ is a Gaussian white noise characterized by $\mathbb{E}[\zeta(t)] = 0$ and covariance function $\mathrm{Cov}[\zeta(t), \zeta(s)] = \delta(t - s)$. Note that $\zeta(t)$ is defined as the formal derivative of the standard Wiener process $W(t)$, an assertion which has to be handled with caution because the Wiener process is nowhere differentiable with probability 1, therefore the Langevin formalism needs to be interpreted in a distributional sense. **Equation 3.19** can be solved analytically:

$$\begin{aligned} \frac{dm(t)}{dt} &= \kappa + \sqrt{\kappa}\zeta(t) \\ m(t) &= m(0) + \kappa t + \sqrt{\kappa}\int_0^t \zeta(s)ds \end{aligned} \tag{3.20}$$

Equation 3.20 can be further simplified given the number of mutations at time $t = 0$ is 0:

$$\begin{aligned} m(t) &= m(0) + \kappa t + \sqrt{\kappa}\int_0^t \zeta(s)ds \\ &= \kappa t + \sqrt{\kappa}\int_0^t \zeta(s)ds \end{aligned} \tag{3.21}$$

Note that the reformulation given at **Equation 3.19**, which results in the solution at **Equation 3.21**, maintains the original Poisson process' mean and variance:

$$\begin{aligned}
\mathbb{E}\left[m(t)\right] &= \mathbb{E}\left[\kappa t + \sqrt{\kappa}\int_0^t \zeta(s)ds\right] && (3.22)\\
&= \kappa t + \sqrt{\kappa}\mathbb{E}\left[\int_0^t \zeta(s)ds\right] && (3.23)\\
&= \kappa t + \sqrt{\kappa}\int_0^t \mathbb{E}\left[\zeta(s)\right]ds && (3.24)\\
&= \kappa t && (3.25)
\end{aligned}$$

$$\begin{aligned}
\mathbb{V}\left[m(t)\right] &= \mathbb{E}\left[\left(m(t) - \mathbb{E}\left[m(t)\right]\right)^2\right] && (3.26)\\
&= \mathbb{E}\left[\left(\kappa t + \sqrt{\kappa}\int_0^t \zeta(s)ds - \kappa t\right)^2\right]\\
&= \kappa\mathbb{E}\left[\left(\int_0^t \zeta(s)ds\right)\left(\int_0^t \zeta(u)du\right)\right] && (3.27)\\
&= \kappa\mathbb{E}\left[\int_0^t\int_0^t \zeta(s)\zeta(u)dsdu\right] && (3.28)\\
&= \kappa\int_0^t\int_0^t \mathbb{E}\left[\zeta(s)\zeta(u)\right]dsdu && (3.29)\\
&= \kappa\int_0^t\int_0^t \mathrm{Cov}\left[\zeta(s),\zeta(u)\right]dsdu && (3.30)\\
&= \kappa\int_0^t\int_0^t \delta(s-u)dsdu && (3.31)\\
&= \kappa\int_0^t 1du\\
&= \kappa t && (3.32)
\end{aligned}$$

Note that the steps that involve **Equations 3.23, 3.24, 3.27, 3.28** and **3.29** are consequence of Fubini's theorem, which allows the interchange of the order of integration, as the integrals are known to be absolutely convergent and the fact that the expectation is a linear operator. The step that involves **Equations 3.29** and **3.30** are consequence of the definition of covariance. The step that involves **Equations 3.30** and **3.31** is a consequence of the definition of the Dirac delta functional. *While the same result can be obtained*

through Itô's formalism, we will continue using Langevin's formalism due to its particular relevance in Biology.

As a corollary, the corresponding dispersion index $\rho_{m(t)}$ remains equal to 1 as expected, since the Wiener process is a continuous-time approximation of the Poisson process. One concern that arises from this reformulation is that the number of mutations $m(t)$ is no longer an integer, hence the model may not seem suitable for counting mutations in a lineage during a specific time period t. However, this issue can be easily solved by applying a rounding function to $m(t)$ whenever it is necessary to obtain an integer value.

Anomalous Diffusion

In the preceding section we demonstrated that, according to the molecular clock hypothesis, the number of mutations occurring in a lineage during a specific time period t can be described as a Brownian motion (**Equation 3.21**) exhibiting a mean and variance equal to κt (**Equations 3.25** and **3.32**), where κ represents the rate of substitutions for a given unit of time, akin to a microscopic particle moving in a fluid as a consequence of thermal forces.

However, it is well known that the diffusion of microscopic particles in a fluid does not always conform to Brownian motion. In fact, the diffusion of particles in a fluid can be classified into three main categories depending on their mean squared displacement (MSD; also understood as the variance of the stochastic process governing the motion): normal diffusion, subdiffusion, and superdiffusion. Under normal diffusion, the MSD of the particle is proportional to t, while under subdiffusion and superdiffusion the MSD of the particle is proportional to t^α, where α is known as the diffusion exponent, with $\alpha < 1$ for the former case and $\alpha > 1$ for the latter [47].

Similarly to a microscopic particle moving in a fluid, the number of mutations in a lineage during a specific time period t may not conform to a Brownian motion, as described by several studies observing overdispersed and underdispersed populations. Therefore, it is reasonable to consider that the number of mutations in a lineage during a specific time period t may exhibit anomalous diffusion.

Evolution as a fractional Brownian motion

Multiple stochastic definitions of anomalous diffusion exist, and it is usually left to the researcher to use the one that best fits their problem. In this work, fractional Brownian motion (fBm) is used as a model for anomalous diffusion due to its simple yet powerful mathematical properties.

Definition 3.4.3 (Fractional Brownian Motion)**.** A fractional Brownian motion (fBm; denoted as $W_\alpha(t)$) is a continuous-time stochastic process that

generalizes classical Brownian motion (Wiener process). It is characterized by stationary increments, mean $\mathbb{E}[W_\alpha(t)] = 0$, and a covariance function of the form $\mathrm{Cov}[W_\alpha(t), W_\alpha(s)] = \frac{1}{2}(t^\alpha + s^\alpha - |t-s|^\alpha)$, where $\alpha \in (0,2)$ is the diffusion exponent which determines the degree of long-term dependence of the process. It is related to the Hurst exponent H by $\alpha = 2H$.

It is straightforward to see that the Wiener process is a special case of the fBm, as the covariance function of the Wiener process is recovered when $\alpha = 1$. To reformulate the number of mutations in a lineage during a specific time period t as a fBm, we will modify the Langevin stochastic differential equation shown in **Equation 3.19**:

$$\frac{dm(t)}{dt} = \kappa + \sqrt{\kappa}\eta(t), \tag{3.33}$$

where $\eta(t)$ is an appropriate noise source characterized by $\mathbb{E}[\eta(t)] = 0$ and a covariance function such that $\mathrm{Cov}[W_\alpha(t), W_\alpha(s)] = \int_0^t \int_0^s \mathrm{Cov}[\eta(u), \eta(v)]\, dudv$. Therefore, we interpret $\eta(t)$ as the generalized derivative of the fBm $W_\alpha(t)$, specifically we define

$$\eta(t) := \frac{dW_\alpha(t)}{dt}$$

in the distributional sense. Since $W_\alpha(t)$ is not differentiable in the classical sense when $\alpha \neq 1$, the above operation is understood via the action of tempered distributions on test functions in $\mathcal{S}(\mathbb{R})$ (the Schwartz space).

Definition 3.4.4 (Schwartz space)**.** The Schwartz space $\mathcal{S}(\mathbb{R})$ is the space of all infinitely differentiable functions $f : \mathbb{R} \to \mathbb{R}$ such that for all $k, l \in \mathbb{N}$ the function $x^k \frac{d^l}{dx^l} f(x)$ is rapidly decreasing, i.e., $\sup_{x\in\mathbb{R}} \left| x^k \frac{d^l}{dx^l} f(x) \right| < \infty$.

By differentiating the covariance function of $W_\alpha(t)$ twice, with respect to t and s, we obtain the covariance function of $\eta(t)$. It is straightforward to see that $\frac{\partial^2}{\partial t \partial s} \mathrm{Cov}[W_\alpha(t), W_\alpha(s)]$ is equivalent to computing $-\frac{1}{2}\frac{\partial^2}{\partial t \partial s}|t-s|^\alpha$ given that the terms t^α and s^α end up vanishing during the differentiation. To ease subsequent computations, we define the following change of variables:

$$\begin{cases} x &= t - s \\ \partial_t &= \partial_x \\ \partial_s &= -\partial_x \end{cases}$$

therefore

$$-\frac{1}{2}\frac{\partial^2}{\partial t \partial s}|t-s|^\alpha = -\frac{1}{2}\frac{\partial^2}{\partial t \partial s}|x|^\alpha$$
$$= \frac{1}{2}\frac{d^2}{dx^2}|x|^\alpha$$

again, understood in the distributional sense. Let $g(x) = |x|^\alpha$ and $\psi(x) \in \mathcal{S}(\mathbb{R})$ be a test function, the second distributional derivative of $g(x)$ is defined as

$$\langle g'', \psi \rangle = \int_{-\infty}^{\infty} g''(x)\psi(x)dx \tag{3.34}$$

Using integration by parts and the property that $\psi(x)$ is rapidly decreasing, we obtain

$$\begin{aligned}
\langle g'', \psi \rangle &= \int_{-\infty}^{\infty} g''(x)\psi(x)dx \\
&= \left[g'(x)\psi(x)\right]_{-\infty}^{\infty} - \int_{-\infty}^{\infty} g'(x)\psi'(x)dx \\
&= 0 - \int_{-\infty}^{\infty} g'(x)\psi'(x)dx \\
&= -\left[g(x)\psi''(x)\right]_{-\infty}^{\infty} + \int_{-\infty}^{\infty} g(x)\psi''(x)dx \\
&= \int_{-\infty}^{\infty} g(x)\psi''(x)dx
\end{aligned} \tag{3.35}$$

By using the definition $g(x)$ and the integral's linearity, we can work back from **Equation 3.35** an expression that parallels **Equation 3.34**:

$$\begin{aligned}
\int_{-\infty}^{\infty} g(x)\psi''(x)dx &= \int_{-\infty}^{\infty} |x|^\alpha \psi''(x)dx \\
&= \int_{-\infty}^{0} (-x)^\alpha \psi''(x)dx + \int_{0}^{\infty} x^\alpha \psi''(x)dx \\
&= \left[(-x)^\alpha \psi'(x)\right]_{-\infty}^{0} + \int_{-\infty}^{0} \alpha(-x)^{\alpha-1}\psi'(x)dx \\
&\quad + \left[-x^\alpha \psi'(x)\right]_{0}^{\infty} - \int_{0}^{\infty} \alpha x^{\alpha-1}\psi'(x)dx \\
&= \int_{-\infty}^{0} \alpha(-x)^{\alpha-1}\psi'(x)dx - \int_{0}^{\infty} \alpha x^{\alpha-1}\psi'(x)dx
\end{aligned}$$

$$
\begin{aligned}
&= \left[\alpha(-x)^{\alpha-1}\psi(x)\right]_{-\infty}^{0} + \int_{-\infty}^{0} \alpha(\alpha-1)(-x)^{\alpha-2}\psi(x)dx \\
&\quad + \left[\alpha x^{\alpha-1}\psi(x)\right]_{0}^{\infty} + \int_{0}^{\infty} \alpha(\alpha-1)x^{\alpha-2}\psi(x)dx \\
&= \int_{-\infty}^{0} \alpha(\alpha-1)(-x)^{\alpha-2}\psi(x)dx + \int_{0}^{\infty} \alpha(\alpha-1)x^{\alpha-2}\psi(x)dx \\
&= \int_{-\infty}^{\infty} \alpha(\alpha-1)\left|x\right|^{\alpha-2}\psi(x)dx
\end{aligned}
$$

Resulting in the following equivalence:

$$
\left\langle g'', \psi \right\rangle = \int_{-\infty}^{\infty} g''(x)\psi(x)dx = \int_{-\infty}^{\infty} \alpha(\alpha-1)\left|x\right|^{\alpha-2}\psi(x)dx
$$

So the distributional second derivative of $g(x)$ is equivalent to $\alpha(\alpha - 1)\left|x\right|^{\alpha-2}$, which results in the following expression for the covariance function of $\eta(t)$:

$$
\mathrm{Cov}\left[\eta(t), \eta(s)\right] = \frac{1}{2}\alpha(\alpha-1)\left|t-s\right|^{\alpha-2} \tag{3.36}
$$

This expression is a well defined generalized function (distribution) on $\mathbb{R}^2 \backslash \{t = s\}$, where the set of points $\{t = s\}$ represent a singularity. However, this singularity is integrable in the sense of distributions, so the integral $\mathrm{Cov}\left[W_\alpha(t), W_\alpha(s)\right] = \int_0^t \int_0^s \mathrm{Cov}\left[\eta(u), \eta(v)\right] dudv$ remains well defined. This definition allows for the computation of the appropriate mean and variance of the process (MSD of the evolutionary motion):

$$
\begin{aligned}
\mathbb{E}\left[m(t)\right] &= \mathbb{E}\left[\kappa t + \sqrt{\kappa}\int_0^t \eta(s)ds\right] && (3.37)\\
&= \kappa t && (3.38)
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{V}\left[m(t)\right] &= \mathbb{E}\left[\left(m(t) - \mathbb{E}\left[m(t)\right]\right)^2\right] && (3.39)\\
&= \kappa\mathbb{E}\left[\left(\int_0^t \eta(s)ds\right)^2\right] \\
&= \kappa\int_0^t\int_0^t \mathbb{E}\left[\eta(s)\eta(u)\right] dsdu && (3.40)
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{\alpha\kappa}{2}(\alpha-1)\int_0^t\int_0^t |s-u|^{\alpha-2}\,ds du \\
&= \frac{\alpha\kappa}{2}(\alpha-1)\int_0^t\left[\int_0^u (u-s)^{\alpha-2}\,ds + \int_u^t (s-u)^{\alpha-2}\,ds\right] du \\
&= \frac{\alpha\kappa}{2}\int_0^t \left[s^{\alpha-1} + (t-s)^{\alpha-1}\right] du \\
&= \kappa t^{\alpha}
\end{aligned} \tag{3.41}
$$

Therefore, by using fBm as a model for anomalous diffusion, the number of mutations in a lineage during a specific time period t can be described as a stochastic process with a mean and variance equal to κt (in line with the molecular clock hypothesis [4]) and κt^{α}, respectively. Therefore, the analysis of $\mathbb{V}[m(t)]$ with time is instrumental to assess the nature of the stochastic movement. Previous evolutionary studies of viruses mainly focused on the mean behavior [10, 24, 41, 42, 43], i.e., evaluating the relationship $\mathbb{E}[m(t)] = \kappa t$, so our study is pertinent due to the completeness achieved.

Particularly, our model for the evolution of SARS-CoV-2 follows a more generalized approach, as it decouples the mean mutation rate from the diffusion coefficient (the latter also being κ in **Equation 3.41**). Being $m(t)$ the number of accumulated mutations in the viral genome at time t, the stochastic differential equation introduced in **Section 3.2**

$$
\frac{dm(t)}{dt} = \kappa + \xi(t) \tag{3.1}
$$

governs the dynamics of the system. Here κ is the evolution rate and $\xi(t)$ is an integrative noise source whose statistical properties match a fBm [48]:

$$
\begin{cases}
\mathbb{E}[\xi(t)] = 0 \\
\mathrm{Cov}[\xi(t),\xi(s)] = \frac{1}{2}D\alpha(\alpha-1)|t-s|^{\alpha-2}
\end{cases}. \tag{3.42}
$$

In this formulation, D is the diffusion coefficient and α the diffusion exponent. Again, the solution for the mean evolutionary motion is

$$
\mathbb{E}[m(t)] = \kappa t, \tag{3.43}
$$

which is compatible to the molecular clock hypothesis [4]. Linear regressions were performed between the calculated mean number of mutations in the viral sequences and time using the `LinearRegression` function from Scikit-learn. This was done for the global data and also for the major lineages Primal, Alpha, Delta, and Omicron (BA.1 and BA.2). Similar to the general model, the solution for the variance is

$$
\mathbb{V}[m(t)] = Dt^{\alpha}, \tag{3.44}
$$

which is compatible with the fBm model, as the MSD is then proportional to a power of time. However, this solution comes from a fixed initial condition (*i.e.*, no variability at $t = 0$). The calculated variance from the viral sequences was fitted to the general expression

$$\mathbb{V}[m(t)] = \sigma_0^2 + Dt^\alpha, \tag{3.45}$$

where σ_0^2 is a parameter that accounts for the initial variance in the population (σ_0^2 was directly computed from the set of sequences). Nonlinear regressions were performed between $\mathbb{V}[m(t)] - \sigma_0^2$ and time using the `curve_fit` function from SciPy. This was done for the global data and also for the major lineages Primal, Alpha, Delta, and Omicron (BA.1). In the case of Delta, the variance analysis was restricted to the AY.4 sublineage, which was the dominant in UK after a first period of time in which other sublineages coexisted (the fixation of the AY.4 sublineage led to a decrease in variance).

3.4.11 Statistical significance of the diffusion parameters

Bootstrapping was applied to assess the robustness of the estimations of D and α. This approach resamples the original dataset with replacement to generate new bootstrap datasets. We can then fit the same model to each of these bootstrap datasets, obtaining a distribution of model parameters. This distribution can be used to estimate the variability of the model parameters and to evaluate the robustness of the model to small changes in the original dataset [49]. In this work, a random sampling with replacement of the sequences was performed each week. The sampling size was defined as the 50% of the total number of sequences available in each week in the original dataset (*i.e.*, if there are 100 sequences in a week, the bootstrap sample size for that week is 50, what is called subsampling). This was done for all 120 weeks in an independent manner. We chose a sample size that was large enough to capture the key characteristics of the original dataset, but small enough to make the bootstrap procedure computationally feasible and robust to observations with a disproportionate impact on the results. With the new bootstrap dataset, the mean and variance were calculated. This procedure was repeated 1000 times. As a result, a distribution of values for each diffusion parameter was obtained, having performed 1000 independent regressions. This was done for the major variants Primal, Alpha, Delta, and Omicron (BA.1).

References

[1] Koonin E, Dolja V (2013) A virocentric perspective on the evolution of life. *Curr Opin Virol*, 3: 536–557.

[2] Drake J, Charlesworth B, Charlesworth D, Crow J (1998) Rates of spontaneous mutation. *Genetics*, 148: 1667–1686.

[3] Zuckerkandl E, Pauling L (1965) Evolutionary divergence and convergence in proteins. *Evolving Genes and Proteins*: 97–166.

[4] Ayala F (1999) Molecular clock mirages. *BioEssays*, 21: 71–75.

[5] Kumar S (2005) Molecular clocks: four decades of evolution. *Nat Rev Genet*, 6: 654–662.

[6] Kimura M (1968) Evolutionary rate at the molecular level. *Nature*, 217(5129): 624–626.

[7] Kimura M (1987) Molecular evolutionary clock and the neutral theory. *J Mol Evol*, 26: 24–33.

[8] Gojobori T, Moriyama E, Kimura M (1990) Molecular clock of viral evolution, and the neutral theory. *Proc Natl Acad Sci USA*, 87: 10015–10018.

[9] Leitner T, Albert J (1999) The molecular clock of HIV-1 unveiled through analysis of a known transmission history. *Proc Natl Acad Sci USA*, 96: 10752–10757.

[10] Jenkins G, Rambaut A, Pybus O, Holmes E (2002) Rates of molecular evolution in RNA viruses: a quantitative phylogenetic analysis. *J Mol Evol*, 54: 156–165.

[11] Bedford T, Wapinski I, Hartl D (2008) Overdispersion of the molecular clock varies between yeast, Drosophila and mammals. *Genetics*, 179: 977–984.

[12] Li J, Lai S, Gao G, Shi W (2021) The emergence, genomic diversity and global spread of SARS-CoV-2. *Nature*, 600: 408–418.

[13] Bedford T, *et al.* (2020) Cryptic transmission of SARS-CoV-2 in Washington state. *Science*, 370: 571–575.

[14] López M, *et al.* (2021) The first wave of the COVID-19 epidemic in Spain was associated with early introductions and fast spread of a dominating genetic variant. *Nat Genet*, 53: 1405–1414.

[15] Lemieux J, *et al.* (2021) Phylogenetic analysis of SARS-CoV-2 in Boston highlights the impact of superspreading events. *Science*, 371: eabe3261.

[16] Tegally H, *et al.* (2021) Detection of a SARS-CoV-2 variant of concern in South Africa. *Nature*, 592: 438–443.

[17] Kraemer M, *et al.* (2021) Spatiotemporal invasion dynamics of SARS-CoV-2 lineage B.1.1.7 emergence. *Science*, 373: 889–895.

[18] Rockett R, *et al.* (2022) Co-infection with SARS-CoV-2 Omicron and Delta variants revealed by genomic surveillance. *Nat Commun*, 13: 2745.

[19] Jackson B, *et al.* (2021) Generation and transmission of interlineage recombinants in the SARS-CoV-2 pandemic. *Cell*, 184: 5179–5188.

[20] Rochman N, *et al.* (2021) Ongoing global and regional adaptive evolution of SARS-CoV-2. *Proc Natl Acad Sci USA*, 118: e2104241118.

[21] Yi K, *et al.* (2021) Mutational spectrum of SARS-CoV-2 during the global pandemic. *Exp Mol Med*, 53: 1229–1237.

[22] de Silva V, Tenenbaum J, Global versus local methods in nonlinear dimensionality reduction. In: *Advances in Neural Information Processing Systems* (2003), 721–728.

[23] Nielsen R (2005) Molecular signatures of natural selection. *Annual Review of Genetics*, 39(1): 197–218, doi:10.1146/annurev.genet.39.073003.112420.

[24] Ghafari M, *et al.* (2022) Purifying selection determines the short-term time dependency of evolutionary rates in SARS-CoV-2 and pH1N1 influenza. *Mol Biol Evol*, 39: msac009.

[25] Manzo C, Garcia-Parajo M (2015) A review of progress in single particle tracking: from methods to biophysical insights. *Rep Prog Phys*, 78: 124601.

[26] De Maio N, *et al.* (2021) Mutation rates and selection on synonymous mutations in SARS-CoV-2. *Genome Biol Evol*, 13: evab087.

[27] Christie S, *et al.* (2022) Single-virus tracking reveals variant SARS-CoV-2 spike proteins induce ACE2-independent membrane interactions. *Sci Adv*, 8: eabo3977.

[28] Kreutzberger A, *et al.* (2022) SARS-CoV-2 requires acidic pH to infect cells. *Proc Natl Acad Sci USA*, 119: e2209514119.

[29] Couce A, Tenaillon OA (2015) The rule of declining adaptability in microbial evolution experiments. *Frontiers in Genetics*, 6: 99, doi:10.3389/fgene.2015.00099.

[30] Tenaillon O (2014) The utility of Fisher's geometric model in evolutionary genetics. *Annual Review of Ecology, Evolution, and Systematics*, 45: 179–201, doi:10.1146/annurev-ecolsys-120213-091846.

[31] Miller CR, Joyce P, Wichman HA (2011) Mutational effects and population dynamics during viral adaptation challenge current models. *Genetics*, 187(1): 185–202, doi:10.1534/genetics.110.121400.

[32] Katoh K, Misawa K, Kuma K, Miyata T (2002) MAFFT: A novel method for rapid multiple sequence alignment based on Fast Fourier transform. *Nucleic Acids Res*, 30: 3059–3066.

[33] Sievers F, *et al.* (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology*, 7(1), doi:10.1038/msb.2011.75.

[34] Hill V, *et al.* (2022) The origins and molecular evolution of SARS-CoV-2 lineage B.1.1.7 in the UK. *Virus Evol*, 8: veac080.

[35] Michaelsen T, *et al.* (2022) Introduction and transmission of SARS-CoV-2 lineage B.1.1.7, Alpha variant, in Denmark. *Genome Med*, 14: 47.

[36] Rosner B (1983) Percentage Points for a Generalized ESD Many-Outlier Procedure. *Technometrics*, 25: 165–172.

[37] Mead A (1992) Review of the development of multidimensional scaling methods. *J R Stat Soc D*, 41: 27–39.

[38] Jolliffe I, Cadima J (2016) Principal component analysis: A review and recent developments. *Philos Trans R A*, 374: 20150202.

[39] da Costa C, de Freitas C, Alves C, Lameira J (2022) Assessment of mutations on RBD in the Spike protein of SARS-CoV-2 Alpha, Delta and Omicron variants. *Sci Rep*, 12: 8540.

[40] Kumar S, Karuppanan K, Subramaniam G (2022) Omicron (BA.1) and sub-variants (BA.1.1, BA.2, and BA.3) of SARS-CoV-2 spike infectivity and pathogenicity: A comparative sequence and structural-based computational assessment. *J Med Virol*, 94: 4780–4791.

[41] Neher R (2022) Contributions of adaptation and purifying selection to SARS-CoV-2 evolution. *Virus Evol*, 8: veac113.

[42] Tay J, Porter A, Wirth W, Duchene S (2022) The emergence of SARS-CoV-2 variants of concern is driven by acceleration of the substitution rate. *Mol Biol Evol*, 39: msac013.

[43] Wang S, *et al.* (2022) Molecular evolutionary characteristics of SARS-CoV-2 emerging in the United States. *Med Virol*, 94: 310–317.

[44] Kryazhimskiy S, Plotkin J (2008) The population genetics of dN/dS. *PLoS Genet*, 4: e1000304.

[45] Nei M, Gojobori T (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol*, 3: 418–426.

[46] Jukes TH, Cantor CR (1969) Evolution of protein molecules. *Mammalian Protein Metabolism*: 21–132, doi:10.1016/b978-1-4832-3211-9.50009-7.

[47] Muñoz-Gil G, *et al.* (2021) Objective comparison of methods to decode anomalous diffusion. *Nature Communications*, 12(1), doi:10.1038/s41467-021-26320-w.

[48] Kursawe J, Schulz J, Metzler R (2013) Transient aging in fractional Brownian and Langevin-equation motion. *Phys Rev E*, 88: 062124.

[49] Efron B (1979) Bootstrap methods: Another look at the jackknife. *Ann Statist*, 7: 1–26.

Chapter 4

PyEvoMotion: a software to perform the temporal statistical analysis of genome evolution

One accurate measurement is worth a thousand expert opinions.
— Grace Hopper

This work has been sent to a peer-reviewed journal.

Goiriz L, Rodrigo G. (2025) PyEvoMotion: a software to perform the temporal statistical analysis of genome evolution. *Under review.*

In this publication, I performed all the software development, testing and documentation. The results were discussed with GR. GR designed the research.

4.1 Introduction

The study of molecular evolution is a central topic in biology. The molecular clock hypothesis assumes that genes accumulate mutations at a constant rate over time [1]. Moreover, under the consideration that most of the accumulated mutations are neutral, the Poisson distribution models the expected variability. The molecular clock hypothesis has become a cornerstone of modern phylogenetic techniques, which are now standard for studying the evolutionary relationships between species and organisms [2].

It has been shown, however, that the simple molecular clock model fails to universally recapitulate evolutionary trajectories. Observations revealed that in some cases mutations do not accumulate at a constant rate [3]. This led to the development of relaxed molecular clocks, in which the rates of mutation accumulation are not uniform across lineages [4]. Although these clocks have proven to be more accurate in certain cases, they still face difficulties to model, for instance, overdispersed populations [5]. A proper analysis of the time-dependent distribution of the number of mutations in the population is necessary to understand and eventually predict the evolutionary trajectories that take place in nature.

Although previous studies have attempted to abstract molecular evolution as a type of diffusion process in the sequence space [1, 6], little attention has been given to the form of the underlying stochastic process. In our previous work, we showed that non-Brownian evolutionary motions occurred within the lineages of a virus, leading to non-Poissonian distributions [7]. Here, we present PyEvoMotion, a Python tool aimed to infer a generalized molecular clock model upon bulk genomic data analysis, featuring a command-line interface and enough modularity for integration into larger Python pipelines. PyEvoMotion is intended to complement traditional phylogenetic analyses.

Traditional phylogenetic methods, while powerful, face computational limitations when applied to large datasets. Indeed, analyzing more than 10^4 sequences becomes impractical due to the exponential complexity of reconstructing evolutionary trees [8]. To overcome this bottleneck, statistical approaches provide a viable alternative [7, 9]. These methods simplify the representation of evolutionary relationships by focusing on patterns of population genetics rather than exhaustive tree reconstruction based on genetic variation. PyEvoMotion leverages stochastic mathematical modelling to assess evolutionary trends, aiming to process datasets orders of magnitude larger than those typically analyzed. This capability is essential for handling the unprecedented volume of genomic data generated by high-throughput sequencing efforts [10].

4.2 Implementation

4.2.1 Data processing

The general workflow of PyEvoMotion is illustrated in Figure 4.1. This tool requires two essential input files: a `.fasta` file containing nucleic acid sequences and a `.tsv` file with the corresponding metadata. Users can customize their analyses by specifying several parameters and filters.

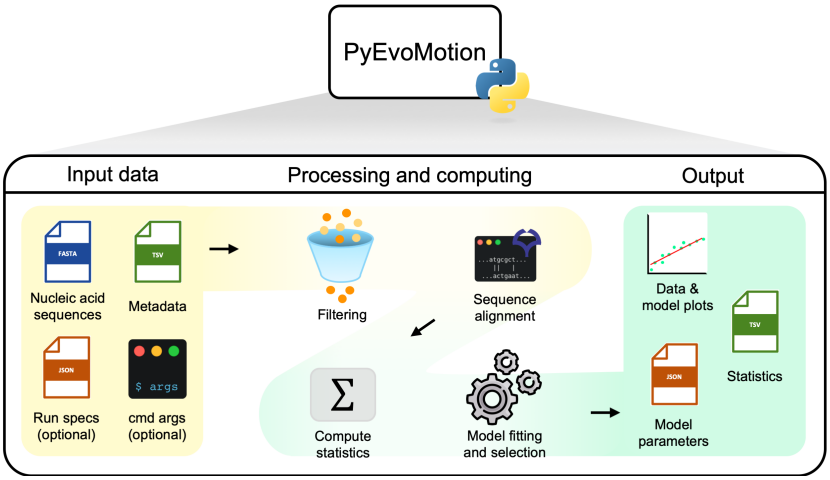

Figure 4.1: Overview of PyEvoMotion. Mandatory input data include nucleic acid sequences (in `.fasta` format) and their corresponding metadata (in `.tsv` format). The metadata must include collection dates, as these are essential for model fitting. Output files include dynamic data representation plots and statistical parameters.

To begin, the temporal granularity of the analysis can be adjusted by defining the time intervals for grouping sequences and calculating statistics. By default, this interval is set to 7 d. Additionally, data filtering options are available to enhance the quality and specificity of the analysis. For instance, the length filter excludes sequences that do not meet a minimum length threshold, thereby removing low-quality genomes (unresolved bases set to a maximum of 1% `N`). The genome position filter allows users to restrict the analysis to specific genomic regions, which is particularly useful for examining genes or genetic clusters of interest. A date range filter further refines the dataset by limiting the analysis to sequences collected within a specified timeframe.

The tool also enables users to select the types of mutations to include in the analysis. Options include: `total` (aggregating all mutations without distinction), `substitutions`, and `indels` (a combined category of insertions and deletions). These three analyses can be done at once with the option

`all`. Filters based on metadata values provide additional flexibility, enabling users to focus on sequences that meet specific criteria in their non-molecular attributes.

After parsing the sequence data, the reference sequence is extracted, defined as the first entry in the `.fasta` file. Following the pre-processing step, each sequence is aligned to the reference sequence using the MAFFT algorithm [11]. Below is an excerpt from PyEvoMotion's `parser.py` source file showing how MAFFT is invoked as a subprocess via the `_run_mafft()` class method, which directly communicates with the MAFFT executable via binary pipes. The `generate_alignment()` class method takes two strings, invokes `_run_mafft()`, retrieves the aligned output binary data, and parses it via Biopython's `AlignIO`, ensuring precise, high-throughput alignment of genomic data prior to mutation detection:

Listing 4.1: Calling MAFFT for sequence alignment.

```
@classmethod
def generate_alignment(cls, seq1: str, seq2: str) -> MultipleSeqAlignment:
"""
Generate a multiple sequence alignment of the input sequences using ``MAFFT``.

:param seq1: The first sequence to be aligned.
:type seq1: str
:param seq2: The second sequence to be aligned.
:type seq2: str
:return: The aligned sequences.
:rtype: ``MultipleSeqAlignment``
"""

id_1 = seq1.id
id_2 = seq2.id

if seq1.id == seq2.id:
    id_1 += "_ref"

return AlignIO.read(
    StringIO(cls._run_mafft({
        id_1: seq1.seq,
        id_2: seq2.seq
    })),
    "fasta"
)

@staticmethod
def _run_mafft(seqs_dict: dict[str,str], outformat: str = "fasta") -> str:
    """
    This function runs the MAFFT multiple sequence alignment tool on the input sequences.

    It raises an exception if the return code is not 0 (i.e. there was an error running MAFFT).

    :param seqs_dict: A dictionary containing the sequences to be aligned. The keys are the sequence
                      names and the values are the sequences.
    :type seqs_dict: dict[str,str]
    :param outformat: The output format of the alignment. Default is fasta.
    :type outformat: str
    :return: The aligned sequences as parsed from stdout. If the output format is clustal, it returns
             the alignment in clustal format; otherwise, it returns the alignment in fasta format.
    :rtype: str
    """

    cmd = ["mafft"]
    template_format = ">{}\n{}\n"

    if outformat == "clustal":
        cmd.extend(["--clustalout", "-"])

    elif outformat != "fasta":
        print(f"Unknown output format: {outformat}. Defaulting to fasta.")
```

```
    cmd.append("-")

    input_data = bytes(
        "".join(
            template_format.format(name, seq)
            for name, seq in seqs_dict.items()
        ),
        "utf-8"
    )

    ps = Popen(
        cmd,
        stdin=PIPE,
        stdout=PIPE,
        stderr=PIPE,
        shell=False
    )
    ps.stdin.write(input_data)
    ps.stdin.close()

    err = ps.stderr.read().decode("utf-8")
    out = ps.stdout.read().decode("utf-8")

    if (ps.returncode != 0) and not(ps.returncode is None):
        raise Exception(
            f"Error running MAFFT:\nStdout:\n{out}\n\nStderr:\n{err}\nReturn code: {ps.returncode}"
        )

    return out
```

Mutation events are identified from the sequence alignments and filtered based on the user-defined mutation types and genomic regions of interest. The model follows the methodology presented in Chapter 3, employing a stochastic differential equation that describing the accumulation of mutations over time. The simpler model assumes Gaussian white noise, leading to a scenario where the average number of mutations grows linearly and the variability scales directly with time, resembling Brownian motion. An alternative, more complex model introduces time-dependent noise characterized by a diffusion exponent, resulting in a fractional Brownian motion where the variance scales nonlinearly; both models converge when the diffusion exponent is 1, aligning with the neutral theory of molecular evolution. Statistical analyses are then conducted on the filtered mutation data for each time interval specified, computing mean and (*unbiased*) variance as

$$\mu_k = \frac{1}{N_k} \sum_{i=1}^{N_k} m_{k,i}, \tag{4.1}$$

$$s_k^2 = \frac{1}{N_k - 1} \sum_{i=1}^{N_k} (m_{k,i} - \mu_k)^2 \tag{4.2}$$

where $m_{k,i}$ represents the number of mutations observed in the i^{th} sequence during the k^{th} time interval, while N_k denotes the total number of sequences within that interval. Consequently, μ_k and s_k^2 correspond to the mean and variance of mutations in the k^{th} time interval, respectively. These statistical measures serve as the basis for fitting a molecular clock model. The

`compute_stats()` method in PyEvoMotion's `core.py` source file illustrates how μ_k and s_k^2 are simultaneously computed for each time interval k.

Listing 4.2: Calculation of mean and variance of mutation counts per time interval.

```
def compute_stats(self,
    DT: str,
    origin: str,
    mutation_kind: str = "all"
) -> pd.DataFrame:
    """
    Compute the length, mean and variance of the data.

    It computes the mean and variance of the data for the specified mutation kind (or kinds)
    in the specified datetime interval and origin.

    :param DT: The string datetime interval that will govern the grouping.
    :type DT: str
    :param origin: The string datetime that will be the origin of the grouping.
    :type origin: str
    :param mutation_kind: The kind of mutation to compute the statistics for. Has to be one of
                          ``all``, ``total``, ``substitutions``, ``insertions``, ``deletions``
                          or ``indels``. Default is ``all``.
    :return: The statistics of the data.
    :rtype: ``pd.DataFrame``
    """

    grouped = self.date_grouper(self.data, DT, origin)
    levels = [
        f"number of {x}"
        for x in self._mutation_type_switch(mutation_kind)
    ]

    return pd.concat(
        (
            pd.DataFrame(self._invoke_method(grouped[levels], method))
            .rename(
                columns=lambda col: f"{method} {col}"
                if method != "size" else "size"
            )
            for method in ("mean", "var", "size")
        ),
        axis=1
    ).reset_index(level=['date'])
```

Note that the `date_grouper()` method is a helper function that groups the data by date using `pandas`' `Grouper` class.

Furthermore, PyEvoMotion offers several configurable run-specific parameters to enhance usability and reproducibility. Users can opt to visualize the output data directly, export the plots in PDF format, save the run parameters as a `.json` file for future reference, or initialize a run using a pre-existing `.json` file. These features ensure that analyses are both customizable and reproducible, catering to diverse research needs.

4.2.2 Model selection

PyEvoMotion estimates the parameters for both models (i.e., κ and D for the null model and κ, D, and α for the challenging model) and then performs a statistical test to select the best option. For that, the calculated values of mean and variance of mutations at each time are represented, and curves are fitted.

κ is directly the slope of the line fitted to the mean of mutations with time (this is the same for both models). The variance of mutations needs to be

rescaled before fitting because the stochastic processes assume a start from the origin. Then, the initial variance is subtracted to all values ($s_k^2 - s_0^2$) and time is shifted so that $t_0 = 0$. In the case of the null model, D is the slope of the intercept-free line fitted to the rescaled variance with time. In the case of the challenging model, a nonlinear regression with a power law relationship is used to obtain the values of D and α.

The fitting of the variance determines the choice of the molecular clock model, which is accomplished by an F-test on the residuals of the fits using

$$F = \frac{\frac{\mathrm{RSS}_1 - \mathrm{RSS}_2}{p_2 - p_1}}{\frac{\mathrm{RSS}_2}{n - p_2}}, \tag{4.3}$$

where RSS_1 and RSS_2 are the residual sum of squares for the null and challenging models, respectively, p_1 and p_2 are the number of parameters for the null and challenging models, respectively (i.e., $p_1 = 1$ and $p_2 = 2$), and n is the number of data points. The F-test is performed at a significance level of 0.05. The snippet below, taken from PyEvoMotion's `base.py` source file, shows how the `adjust_model()` class method implements linear (null) and a power-law (challenging) fittings to the data, and how an F-test is applied to compare the models:

Listing 4.3: Linear (null) and power-law (challenging) model fitting to the data, and F-test procedure for model comparison.

```
@classmethod
def adjust_model(cls,
    x: pd.Series,
    y: pd.Series,
    name: str = None
) -> dict[str, any]:
    """Adjust a model to the data.

    :param x: The features. It is a single pandas Series.
    :type x: pd.Series
    :param y: The target. It is a single pandas Series.
    :type y: pd.Series
    :param name: The name of the data. Default is ``None``.
    :type name: str
    :return: A dictionary with the model.
    :rtype: ``dict[str, any]``
    :raises ValueError: If the dataset is empty or full of NaN values. This may occur if the grouped
                        data contains only one entry per group, indicating that the variance cannot
                        be computed.
    """

    x,y = cls._remove_nan(x, y)

    # Raises an error if the dataset is empty at this point
    if (x.size == 0) or (y.size == 0):
        cls._check_dataset_is_not_empty(
            pd.DataFrame(),
            "Perhaps NaN appeared on all entries. Check if the grouped data contains only one
            entry per group, as this may cause NaN values when computing the variance."
        )

    # Not fitting the intercept in model 1 because data is passed scaled to the minimum
    model1 = cls.linear_regression(x, y, fit_intercept=False)
    model2 = cls.power_law_fit(x, y)

    _, p = cls.F_test(model1, model2, y)

    if p < 0.05:
```

```
        model = model2
    else:
        model = model1

    if name:
        return {name: model}
    else:
        return model
```

The function returns a `dict` containing the the best-fit parameters and the chosen model.

4.2.3 Modularity

PyEvoMotion includes a command line interface designed for Unix-based systems. Given that most bioinformatic analyses consist of larger workflows, PyEvoMotion provides its outputs in standard formats such as `.tsv` and `.json`, which can be easily integrated into existing pipelines. The tool comes also available as a Python module, allowing users to incorporate its functionality and helper utilities into their own Python-based workflows with ease.

Interoperability limitations are minimal but not negligible, as PyEvoMotion relies heavily on MAFFT for sequence alignment. The absence of a proper foreign function interface (FFI) between PyEvoMotion and MAFFT necessitates calling the latter as a subprocess, creating a performance bottleneck. Future versions might mitigate this limitation by introducing a more Python-friendly interface.

The incorporation of alternative mathematical models is possible with little effort in PyEvoMotion due to its modular architecture. Moreover, extended versions of the tool might automatically identify different lineages when analyzing long-time datasets and use piecewise models for the mean and variance of mutations.

4.3 Validation

The utility of PyEvoMotion was validated with a real dataset containing whole-genome sequences of the SARS-CoV-2 Alpha variant from the GISAID database [12]. The sequences were divided into two groups based on their country of origin: the United Kingdom (UK) and the United States of America (USA). For each group, we randomly sampled 9999 sequences, kept the samples collected between October 2020 and August 2021, and analyzed the number of accumulated mutations over time with respect to the NCBI reference sequence NC_045512.2. All calculations (including data parsing, filtering, sequence alignments, and model fitting) were achieved in less than 1 h in a personal computer, showing the potential scalability of the approach.

Figure 4.2 shows the results. In the case of UK, the inferred evolution rate was $\hat{\kappa} = 0.19$ wk^{-1}, and the challenging model was the best option, inferring a diffusion coefficient of $\hat{D} = 1.82$ and a diffusion exponent of $\hat{\alpha} = 0.43$. In the case of USA, the inferred evolution rate was $\hat{\kappa} = 0.32$ wk^{-1}, and the challenging model was also selected, with inferred diffusion parameters of $\hat{D} = 0.69$ and $\hat{\alpha} = 0.67$.

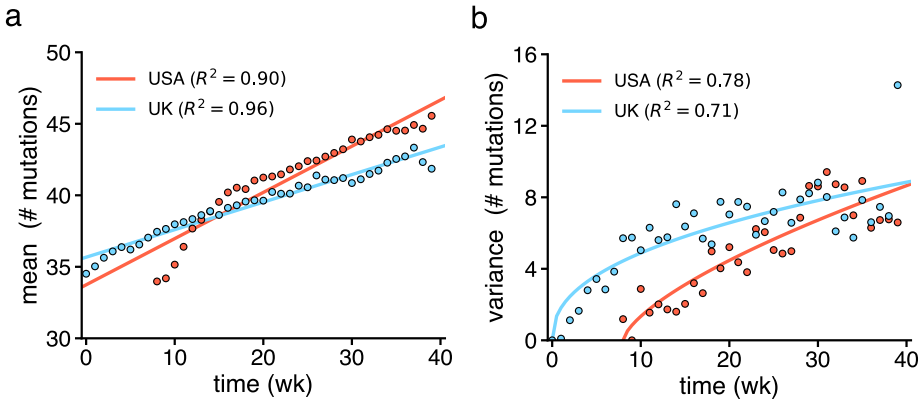

Figure 4.2: Mutational mean and variance over time in the SARS-CoV-2 Alpha variant genomes from UK and USA. a) Mean number of accumulated mutations. b) Scaled variance of the number of accumulated mutations. Points correspond to calculated values from the sequence dataset and lines to inferred molecular clock models.

4.4 Conclusions

Here, we present a high-throughput data-processing, open-source, user-friendly software, called PyEvoMotion, to study evolutionary motions under a statistical perspective provided a collection of genomic sequences. PyEvoMotion is designed to be flexible and customizable, offering a wide range of options for data analysis. Such statistical analysis is complementary to phylogenetic tree reconstructions and molecular assays that measure the impact of key mutations [13].

Nonetheless, our work presents some limitations. In the models, the evolution rate is assumed constant, despite it can vary with time if lineages with higher fitness emerge and even be non-linear if adaptation is the dominant process [14]. This would require applying the date filter to limit the analysis to a subset of sequences, as we did to obtain the results shown in Figure 4.2. Moreover, this statistical approach fails to provide meaningful insight if the collection of sequences is not sufficiently large and does not span in time.

In addition to virus evolution, PyEvoMotion might be used to study the tempo and mode of accumulation of mutations in bacteria [14] or in cancer cells [15]. Understanding the dynamics of these rapidly evolving biological entities might have biomedical implications.

Data availability

The open source software is available on GitHub at https://github.com/luksgrin/PyEvoMotion and on SourceForge at https://sourceforge.net/projects/pyevomotion. Genomic data used in the validation were extracted from the GISAID database (https://www.gisaid.org) and are available on SourceForge.

References

[1] Kimura M (1987) Molecular evolutionary clock and the neutral theory. *J Mol Evol*, 26(1–2): 24–33.

[2] Kumar S (2005) Molecular clocks: Four decades of evolution. *Nat Rev Genet*, 6(8): 654–662.

[3] Ayala F (1997) Vagaries of the molecular clock. *Proc Natl Acad Sci USA*, 94(15): 7776–7783.

[4] Drummond A, Ho S, Phillips M, Rambaut A (2006) Relaxed phylogenetics and dating with confidence. *PLoS Biol*, 4(5): e88.

[5] Bedford T, Wapinski I, Hartl D (2008) Overdispersion of the molecular clock varies between yeast, drosophila and mammals. *Genetics*, 179(2): 977–984.

[6] Huynen M, Stadler P, Fontana W (1996) Smoothness within ruggedness: The role of neutrality in adaptation. *Proc Natl Acad Sci USA*, 93(1): 397–401.

[7] Goiriz L, *et al.* (2023) A variant-dependent molecular clock with anomalous diffusion models SARS-CoV-2 evolution in humans. *Proc Natl Acad Sci USA*, 120(30): e2303578120.

[8] Chor B, Tuller T (2005) Maximum likelihood of evolutionary trees: hardness and approximation. *Bioinformatics*, 21(S1): i97–i106.

[9] Obermeyer F, *et al.* (2022) Analysis of 6.4 million SARS-CoV-2 genomes identifies mutations associated with fitness. *Science*, 376(6599): 1327–1332.

[10] Oude Munnink B, *et al.* (2020) Rapid SARS-CoV-2 whole-genome sequencing and analysis for informed public health decision-making in the Netherlands. *Nat Med*, 26(9): 1405–1410.

[11] Katoh K, Standley D (2013) MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol Biol Evol*, 30(4): 772–780.

[12] Khare S, *et al.* (2021) GISAID's role in pandemic response. *China CDC Weekly*, 3(49): 1049–1051.

[13] Mlcochova P, *et al.* (2021) SARS-CoV-2 B.1.617.2 Delta variant replication and immune evasion. *Nature*, 599(7883): 114–119.

[14] Tenaillon O, *et al.* (2016) Tempo and mode of genome evolution in a 50,000-generation experiment. *Nature*, 536(7615): 165–170.

[15] Borgsmuller N, *et al.* (2023) Single-cell phylogenies reveal changes in the evolutionary rate within cancer and healthy tissues. *Cell Genomics*, 3(9): 100380.

Chapter 5

Deciphering microscopic drivers of viral genome-scale molecular clock dynamics

Nothing in biology makes sense except in the light of evolution.
— Theodosius Dobzhansky

This work has been sent to a peer-reviewed journal.

Goiriz L, Rodrigo G. (2025) Deciphering microscopic drivers of viral genome-scale molecular clock dynamics. *Under review.*

In this work, I performed all the data analysis and figures. The results were discussed with GR. GR designed the research.

5.1 Introduction

Viruses, and in particular RNA viruses, with their short generation times and high mutation rates, offer a unique window into the dynamics of evolution [1]. Unlike in many organisms, the mutation accumulation process in viruses can be observed in real time, making them ideal systems for studying the mechanisms that drive genetic variation. By analyzing viral genomes, we can test evolutionary theories, such as neutral theory [2] or lethal mutagenesis [3], we can track how mutations spread through populations [4, 5], which results in a critical tool in epidemiology to propose suitable interventions, and we can ultimately assess how selection, drift, and recombination shape genetic diversity [6, 7]. However, viruses do not evolve uniformly across their genomes. They contain several genes coding for different types of proteins (e.g., structural and non-structural), each subject to singular selection pressures that shape its evolution rate and variability [8]. For example, genes encoding instrumental proteins to generate progeny [e.g., capsid or RNA-dependent RNA polymerase (RdRp)] tend to evolve more slowly due to strong purifying selection, while genes encoding proteins involved in direct interactions with host cell elements (e.g., glycoproteins with receptor-binding domains) are usually subject to positive selection [9]. In this regard, further research is needed to determine the precise contribution of each gene, which may additionally change with time, to the evolution of a virus as a whole entity.

Evolutionary processes are often described using stochastic models that assume regularity in the accumulation of genetic changes. The molecular clock hypothesis is a traditional theoretical framework derived from neutral theory that has provided foundational insights into diversification by relying on the assumption that mutations accumulate at a relatively constant rate over time and that the extent of variability in a population follows a Poisson distribution [10, 11]. Despite the wide application of this model, virus evolution can deviate from its predictions due to factors such as transmission bottlenecks, host immune pressures (e.g., vaccination), and environmental changes [12]. In this regard, relaxed molecular clock descriptions have been developed to allow evolution rates to vary among phylogenetic tree branches, thereby accommodating the potential effects of external factors [13]. In these models, however, variation is always driven by delta-correlated (white) Gaussian noise to end with Poissonian statistics. We can then consider that each gene from a virus evolves at a given rate displaying a distinctive level of variation and that the integration of the different dynamics results into the observed evolutionary movement of the virus [14]. Disentangling the overlooked statistical properties that characterize such individual dynamics is important for a fundamental understanding of the process and requires further introspective analysis of

high-throughput sequencing data.

The global public health crisis provoked by the emergence of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2; [15]) prompted the implementation of genomic surveillance programs. It is therefore appealing to focus the study on this RNA virus to take advantage of the resulting database with millions of whole genome sequences. Notably, such efforts facilitated the detection of novel variants of concern [5], as well as the spatiotemporal monitoring of the impact of introduction and superspreading events [4, 16]. They have also allowed the quantification of the interplay between adaptation, neutrality, and purifying selection during virus evolution in nature at an unprecedented resolution [14, 17]. In particular, it has been shown that rapid bursts of adaptation were followed by phases of slow evolution in which purifying selection predominated. Moreover, we have pointed out that each variant evolved at a different rate and that the observed genetic variability is compatible with a stochastic anomalous diffusion process (viz., with time-dependent Gaussian noise) distinctively parameterized in each case [18]. This indicates deviations from Poissonian statistics and highlights evolutionary constraints associated with each genotype. Notwithstanding, the mechanisms underlying this anomalous diffusion behavior remain unknown.

In this work, we analyze the gene-level mutational dynamics of SARS-CoV-2 to determine how microscopic contributions shape macroscopic evolutionary trends. For that, we focus on the different within-variant evolution phases and quantify for each gene the process of mutation accumulation and the generation of variability. More specifically, we calculate the mean number of accumulated mutations, the mutational variance of the population, and the frequency of each mutation to provide a comprehensive evolutionary picture for each of the 27 genes of SARS-CoV-2 [19] (ORF3b and ORF9c were discarded in our study). We aim to identify whether specific genes show deviations from the classical molecular clock predictions and how these local effects drive the anomalous diffusion behavior reported at the genome-scale level. Unraveling these complexities might also enhance our ability to predict virus evolution [20], with implications in epidemiology. To effectively process a massive number of genomic sequences, we here follow a statistical approach [21, 18], examining population genetic compositions over time rather than reconstructing phylogenetic trees from genetic variation. Conventional phylogenetic methods [22] show great utility to track pathogen evolution in public health crisis scenarios in order to e.g. identify mutations that increase fitness and to trace transmission routes (especially when mobility data are aggregated; [4]). However, they do not directly follow the genetic shaping of the population through characteristic statistical parameters (moments, frequencies) and highly depend on proper prior selection; in addition to struggling with scalability due to an excessive computational

demand, thereby becoming unpractical or requiring approximations when dealing with a vast number of sequences [23]. Through this top-down approach from large empirical observations, we seek to refine probabilistic models of viral genome evolution under the molecular clock prism and further appreciate the complexity of the mutation-selection process in infectious agents circulating in nature.

5.2 Results

We implemented a simple-method computational pipeline to process the massive number of available SARS-CoV-2 whole genome sequences and extract time-dependent statistical parameters characterizing the evolutionary trajectory of the virus. We mainly focused our analysis on the United Kingdom (UK) during the first 120 weeks of pandemic, because a strong genomic surveillance program was implemented in this geographic region at that time [24]. After curating the available dataset to remove poor-quality, non-date-annotated, and outlier items, we obtained 1.87 million high-quality genome sequences. With respect to the reference sequence from Wuhan (China), we identified 108,347 different mutations distributed throughout the virus genome during those 120 weeks, including substitutions, insertions, and deletions (indels). During this period, new variants of concern emerged (Alpha, Delta, and Omicron), which invaded the population one after another as a consequence of acquiring distinct mutations that enhanced their transmissibility and immune evasion capabilities [25].

A temporal representation of the mean number of accumulated mutations by gene pointed out to PLpro (encoding a protease for polyprotein processing), RdRp (encoding the replicase), S (encoding the spike protein for viral entry), and N (encoding the nucleocapsid) as the most evolved genes since the pandemic origin (**Fig. 5.1a**). ORF9b also appeared as evolved as it overlaps with N [19]. With the emergence of Omicron, S separated from the rest of genes in terms of divergence. Other genes started to evolve with the emergence of variants, such as ORF8 with Alpha, nsp4, nsp6, and ORF7a with Delta, and M and ORF6 with Omicron. In addition, a representation of the dispersion index of the virus population by gene, calculated as the ratio between the mutational variance and mean [26], yielded a mosaic-like picture (**Fig. 5.1b**).

Non-monotonous variations in dispersion index indicate complex evolutionary trajectories, beyond deviations from Poissonian statistics. The highest genetic dispersion occurred in S with the invasion of Omicron BA.1, in agreement with multiple introductions of different sublineages in the UK [27]. Despite having significant evolution rate, RdRp continuously displayed reduced levels of dispersion, which could be a signature of purifying selection

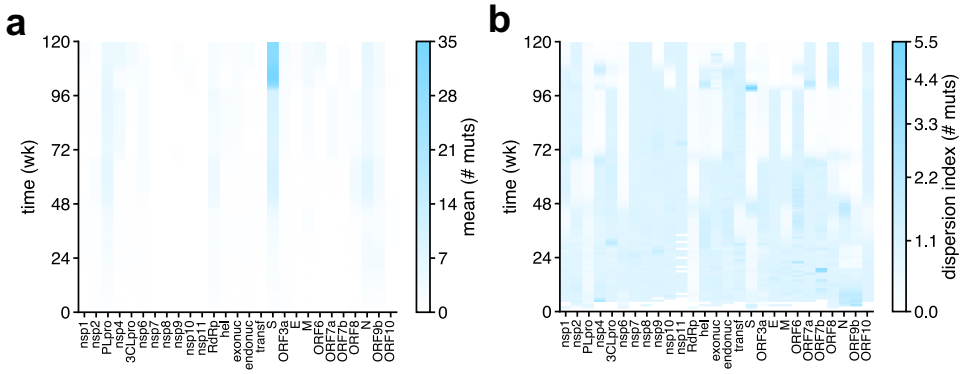

Figure 5.1: a) Time-course of the mean number of accumulated mutations in each gene of the SARS-CoV-2 genome, including substitutions and indels (with respect to the reference sequence from Wuhan; data from the UK). Heatmap colored using a linear scale. b) Time-course of the dispersion index (variance-to-mean ratio) of accumulated mutations in each gene of the SARS-CoV-2 genome. Heatmap colored using an inverse hyperbolic sine scale.

in order to maintain an efficient viral replication. Moreover, during the period in which Delta was the dominant variant, the dispersion in several genes was substantially low, reflecting a founder effect.

Overall, SARS-CoV-2 evolved through adaptation in a punctuated manner by accumulating non-synonymous mutations (**Fig. 5.2a**). However, while the emergence and invasion of Alpha and Omicron BA.1 entailed bursts of non-synonymous mutations (at weeks 44 and 99, respectively), with Delta the number of non-synonymous substitutions was roughly maintained and the synonymous substitutions significantly decreased (at week 66). The dispersion index peaks (> 1) clearly supported this view of the transitions (**Fig. 5.2b**; see also the dispersion index calculated for genes encoding structural and non-structural proteins in **Fig. 5.2c**). This is in tune with an independent emergence of Alpha and Delta in late 2020, noting that Delta first spread in India and then was introduced in the UK. Inferences of the evolution rates by gene and mutation type revealed the inclination of S, E, N, ORF7a, and exonuc to incorporate non-synonymous changes (a signature suggesting adaptation). For example, mutations H655Y and T95I in S, T9I in E, and P13L in N have been highly ranked in terms of their positive effect on viral fitness [21]. Inferences also showed that nsp9 and nsp10 preferentially accumulated synonymous mutations (a signature of purifying selection), and that nsp6 and ORF8 preferentially accumulated indels (**Fig. 5.2d**). It has been proposed that ORF8 knockout contributes to increase viral fitness, although this may depend on the genetic background [28].

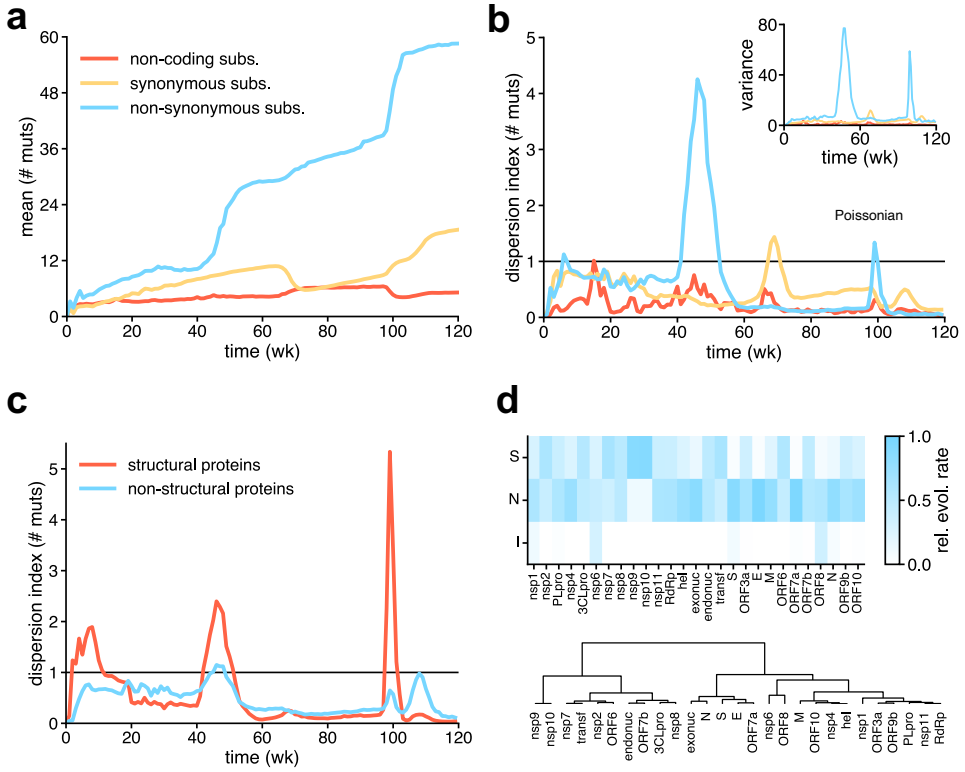

Figure 5.2: a) Time-course of the mean number of non-coding, synonymous, and non-synonymous substitutions. b) Time-course of the dispersion index of non-coding, synonymous, and non-synonymous substitutions. Inset, variance of substitutions with time. c) Time-course of the dispersion index in genes encoding structural and non-structural proteins. d) Inferred relative evolution rate for each gene and mutation type [synonymous substitutions (S), non-synonymous substitutions (N), and indels (I)]. Bottom, hierarchical clustering of genes.

We also analyzed the number of accumulated mutations in each gene from datasets of other countries [viz., the United States of America (USA), Spain, India, South Africa, and Japan]. Comparing the inferred evolution rates, we found similar dynamics overall (**Fig. 5.3a**). Furthermore, we observed an interesting relationship between mutational mean and variance at particular times strictly dominated by a single variant in the UK. Considering the absolute number of mutations in each gene, Poissonian statistics hold in the case of low mutational load, being sub-Poissonian otherwise. Nevertheless, all late-stage mutational distributions get closer to the Poissonian regime by accounting for the founder effect (**Fig. 5.3b**). Together, these results highlight the non-homogeneous genetic modulation of the virus and encourage further

investigation at the variant level to completely decipher the evolutionary role of each gene.

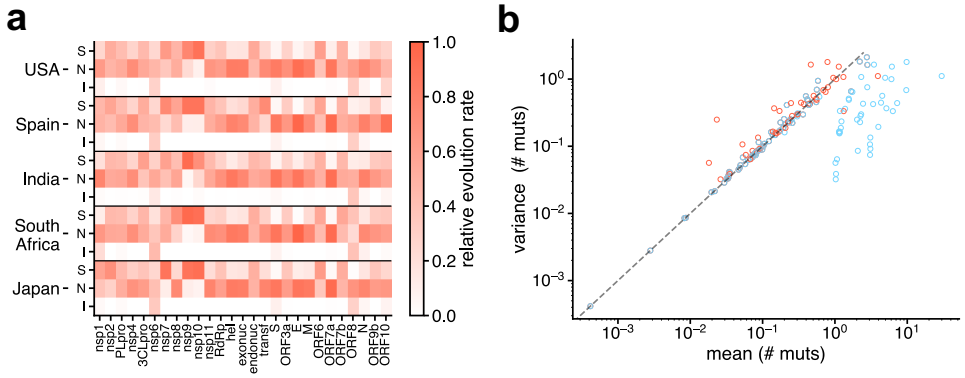

Figure 5.3: a) Inferred relative evolution rate for each gene and mutation type in different countries. b) Scatter plot between mean and variance in each gene at different weeks (1, 30, 60, 90, and 119). Dashed line corresponds to equal variance and mean (i.e., Poissonian regime).

According to our previous work, the whole genome of the virus accrued mutations at a nearly linear rate when particularizing for a given variant after emergence [18], as predicted by the traditional molecular clock model. However, by analyzing the mean mutational load of each gene, we observed widely different dynamics (note that to ensure high statistical power, we restricted the analysis to the time window in which each variant represented at least 10% of the population). While linear trends could be observed in various cases, we also found genes displaying highly non-linear mutation accumulation and even genes with non-monotonous and decreasing trends (**Fig. 5.4**). The analysis of the genetic variability observed in the populations of the distinct variants also revealed intriguing microscopic aspects of the evolutionary movement of the virus, noting that an empirical model with a power law relationship between time and variance better explains the genomic data [18]. As in the case of the mean, non-linear, non-monotonous, and decreasing trends were found when computing the variance of the number of mutations for the different genes (**Fig. 5.5**).

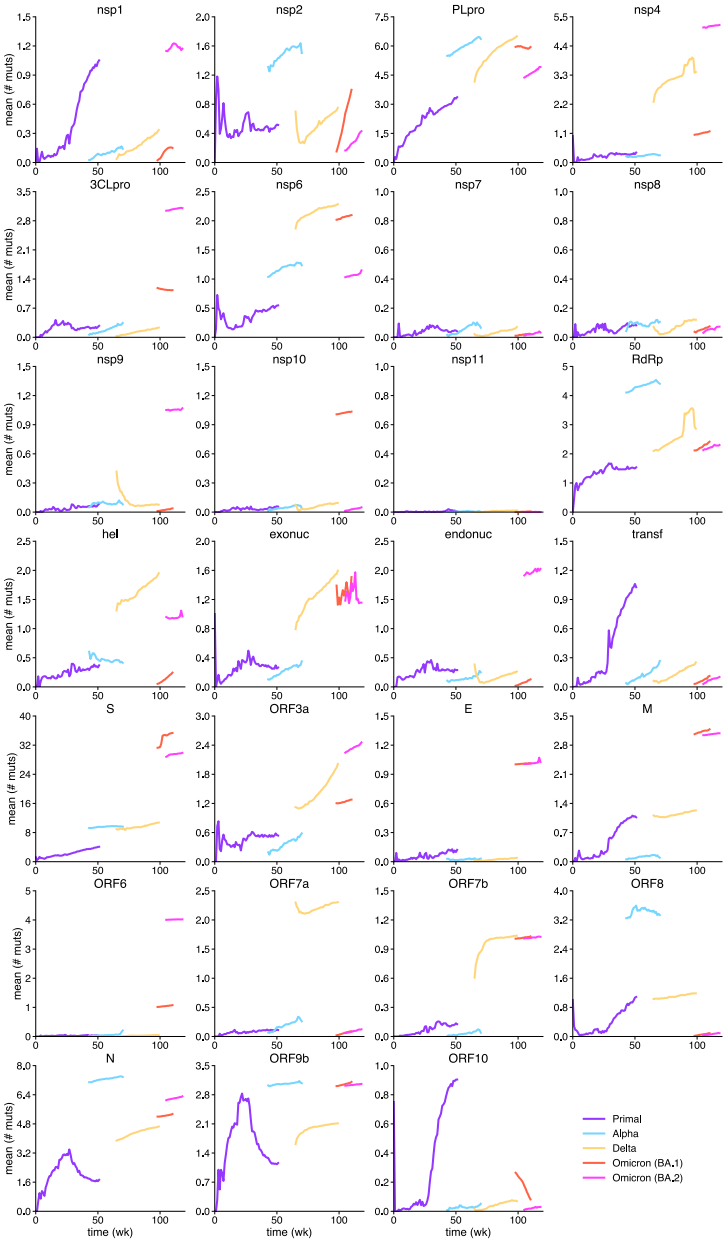

Figure 5.4: Gene-by-gene analysis of the variant-specific evolutionary trajectory of SARS-CoV-2. Time-course of the mean number of accumulated mutations in each gene of the SARS-CoV-2 genome and for each variant (Primal, Alpha, Delta, and Omicron BA.1 and BA.2). Mutations included substitutions and indels (with respect to the reference sequence from Wuhan; data from the UK). Analysis restricted to the time period in which the population frequencies of the variants were at least 10%.

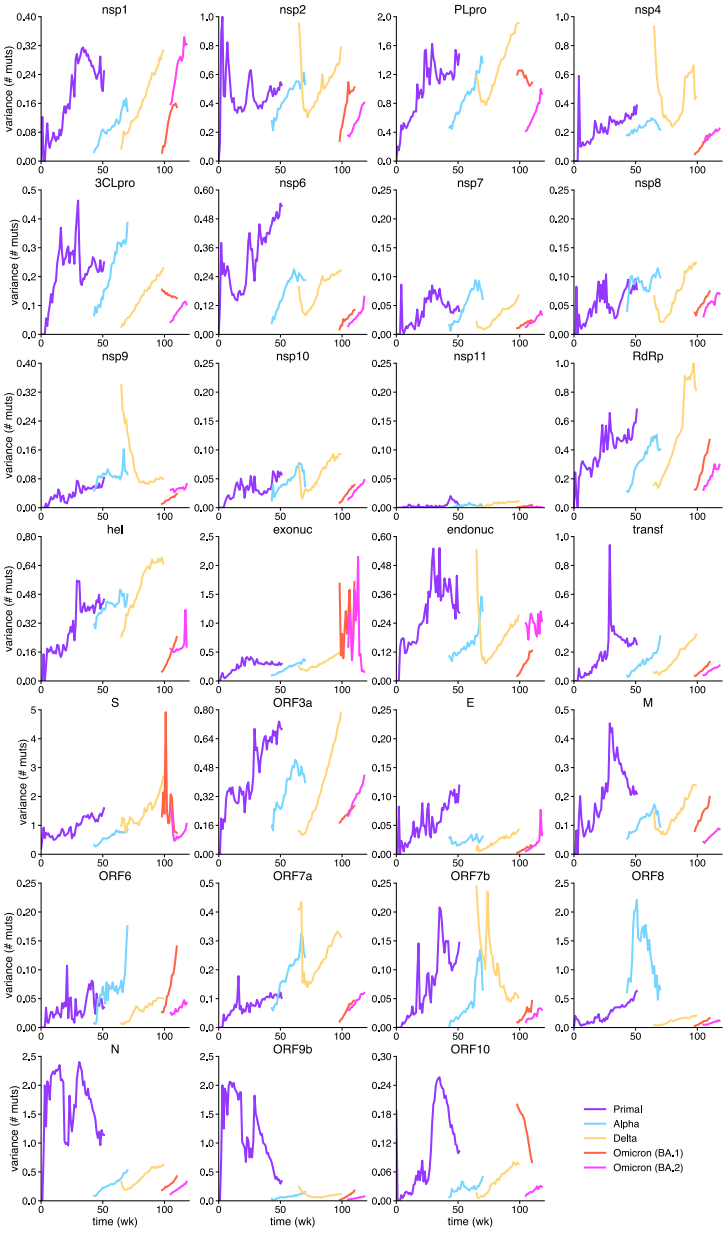

Figure 5.5: Gene-by-gene analysis of the variant-specific evolutionary diffusion of SARS-CoV-2. Time-course of the variance of accumulated mutations in each gene of the SARS-CoV-2 genome and for each variant (Primal, Alpha, Delta, and Omicron BA.1 and BA.2). Mutations included substitutions and indels (with respect to the reference sequence from Wuhan; data from the UK). Analysis restricted to the time period in which the population frequencies of the variants were at least 10%.

5.3 Conclusions

Here, we have examined the microscopic factors that drive the genome-scale molecular clock dynamics of SARS-CoV-2, capitalizing on the unprecedented quantity of high-resolution sequencing data. By dissecting the virus' genome into its constituent genes, we uncovered strong non-uniformity in the patterns of mutation accumulation, highlighting the mosaic-like nature of SARS-CoV-2 evolution. Although global, genome-wide analyses reveal an overall quasi-linear accumulation of mutations at the variant level, consistent with certain predictions of the molecular clock framework, gene-level investigations reveal that specific loci often deviate from canonical Poissonian behaviors. Critically, some genes displayed highly non-monotonous or even decreasing trends in both mean mutational load and variance, underscoring the importance of local selection pressures, functional constraints, and founder effects in shaping evolutionary outcomes.

Our findings show that genes encoding essential viral functions, such as those involved in replication (RdRp) or structural integrity (N, S, and M), exhibit distinct evolutionary trajectories, reflecting a delicate balance between purifying selection and bursts of adaptive mutations. This dynamic is further evidenced by the shifts in dispersion indices through time: while some genes remain under tightly constrained evolution due to critical functional requirements (e.g., RdRp), others, particularly those facing strong interactions with host immune elements, undergo episodic bursts of adaptation (e.g., S). For instance, the non-monotonous or even decreasing mutation trajectories observed in certain genes likely result from transient selective sweeps or founder events that reset local genetic diversity, illustrating how chance transmission bottlenecks can temporarily invert or slow evolutionary trends at specific loci [17]. Importantly, our approach leverages a statistical "top-down" perspective on the data and circumvents the computational hurdles frequently associated with large-scale phylogenetic reconstructions, thereby enabling the detection of subtle gene-level signatures of sub- or super-Poissonian mutation dynamics.

In sum, the gene-by-gene exploration presented here deepens our understanding of viral population diversification and its relationship to the classical molecular clock hypothesis. While certain global properties of SARS-CoV-2 evolution can be adequately captured by simple stochastic models, the local deviations at specific genes illustrate that the true evolutionary landscape is more intricate, non-uniform and punctuated: periods of quasi-clock-like, gradual change are interspersed with bursts of accelerated evolution when selective pressures shift [29].Future research that integrates functional assays, structural studies, and comparative genomics across closely related coronaviruses will help elucidate the mechanisms behind

these locus-specific dynamics and further enhance our ability to forecast evolutionary trajectories relevant to public health interventions.

5.4 Materials and Methods

5.4.1 Genomic data

The nucleotide sequences of the SARS-CoV-2 genomes used in this study were retrieved from the GISAID database [30]. As of May 2022, more than 10 million sequences and the corresponding metadata were downloaded. In this work, only sequences originating from the UK, the USA, Spain, India, South Africa, and Japan were included in the analysis due to their overall quality and to have a broad geographic coverage (the main analysis, however, was done with the data from the UK). The nucleotide sequences of the SARS-CoV-2 genomes were aligned against the NCBI reference genome `NC_045512.2` (`hCoV-19/Wuhan/IVDC-HB-01/2019`, GISAID accession `EPI_ISL_402119`) using Multiple Alignment using Fast Fourier Transform (MAFFT; [31]). The results were collected in Clustal format. For each sequence in the dataset, the number of mutations (substitutions and indels) with respect to the reference genome were counted. This information was retrieved from the MAFFT output alignments. Additionally, the sequence collection dates and Pangolin lineages (to assign the variant names) were retrieved from the metadata. Next, a filtering step was applied to discard sequences without recorded dates, sequences of poor quality (e.g., whose size was below 25 kb), sequences isolated from non-human hosts, duplicated entries, and outlier sequences (which could originate from incorrect date annotations, aberrant evolutionary trajectories, or sudden point introductions; [18]).

5.4.2 Statistical calculations

The genome sequences of the resulting dataset were grouped by weeks. For each week, the mean number of accumulated mutations of the viral population, the variance of the accumulated mutations, and the dispersion index (defined as the ratio between variance and mean) were computed. Protein-coding regions were translated to calculate the degree of heterogeneity in the population as the mean Hamming distance between all amino acid sequence pairs, as well as the frequency of each different mutation (non-synonymous substitutions and indels). Moreover, the mean dN/dS ratio was calculated to get an estimate of natural selection, relying on the method proposed by [32].

The dynamic analysis was performed by accounting for the different types of mutations accumulated in the whole genome and in each gene of the

virus (27 genes were considered: nsp1, nsp2, PLpro, nsp4, 3CLpro, nsp6, nsp7, nsp8, nsp9, nsp10, nsp11, RdRp, hel, exonuc, endonuc, transf, S, ORF3a, E, M, ORF6, ORF7a, ORF7b, ORF8, N, ORF9b, and ORF10, discarding ORF3b and ORF9c; [19]). Variant-specific analyses were also performed in which only the sequences assigned to a given variant were considered [14, 18]. Here, Primal (the primary lineage from the reference sequence), Alpha, Delta, and Omicron (BA.1 and BA.2) were the variants of study. Only time periods during which a given variant represented at least 10% of the population were considered, thereby ensuring reliable statistical calculations. Mutations were classified by type (substitutions and indels), and among substitutions we distinguished non-coding substitutions (i.e., substitutions occurring in non-coding regions), synonymous substitutions, and non-synonymous substitutions. Mutations were also classified as those affecting genes coding for structural (S, E, M, and N) and non-structural proteins. The evolution rate for each gene and mutation type was estimated as the slope of the best-fit linear regression model applied to the temporal count of mutations (synonymous and non-synonymous substitutions and indels). To facilitate comparisons across genes regarding the mutation types preferentially accumulated over time, these slopes were normalized (viz., dividing by the sum of evolution rates for each mutation type). All data analyses were performed in Python using the libraries Pandas (https://pandas.pydata.org), NumPy (https://numpy.org), SciPy (https://scipy.org), Scikit-learn (https://scikit-learn.org), and Biopython (https://biopython.org) [33, 34].

References

[1] Elena SF, Sanjuán R (2007) Virus evolution: insights from an experimental approach. *Annu Rev Ecol Evol Syst*, 38: 27–52.

[2] Gojobori T, Moriyama EN, Kimura M (1990) Molecular clock of viral evolution, and the neutral theory. *Proc Natl Acad Sci USA*, 87: 10015–10018.

[3] Crotty S, Cameron CE, Andino R (2001) RNA virus error catastrophe: direct molecular test by using ribavirin. *Proc Natl Acad Sci USA*, 98: 6895–6900.

[4] Kraemer MUG, *et al.* (2021) Spatiotemporal invasion dynamics of SARS-CoV-2 lineage B.1.1.7 emergence. *Science*, 373: 889–895.

[5] Viana R, *et al.* (2022) Rapid epidemic expansion of the SARS-CoV-2 Omicron variant in southern Africa. *Nature*, 603: 679–686.

[6] Nelson MI, *et al.* (2006) Stochastic processes are key determinants of short-term evolution in influenza a virus. *PLoS Pathog*, 2: e125.

[7] Tully DC, Fares MA (2009) Shifts in the selection-drift balance drive the evolution and epidemiology of foot-and-mouth disease virus. *J Virol*, 83: 781–790.

[8] Gray RR, *et al.* (2011) The mode and tempo of hepatitis C virus evolution within and among hosts. *BMC Evol Biol*, 11: 131.

[9] Kistler KE, Huddleston J, Bedford T (2022) Rapid and parallel adaptive mutations in spike S1 drive clade success in SARS-CoV-2. *Cell Host Microbe*, 30: 545–555.

[10] Kimura M (1987) Molecular evolutionary clock and the neutral theory. *J Mol Evol*, 26: 24–33.

[11] Kumar S (2005) Molecular clocks: four decades of evolution. *Nat Rev Genet*, 6: 654–662.

[12] Jenkins GM, Rambaut A, Pybus OG, Holmes EC (2002) Rates of molecular evolution in RNA viruses: a quantitative phylogenetic analysis. *J Mol Evol*, 54: 156–165.

[13] Drummond AJ, Ho SYW, Phillips MJ, Rambaut A (2006) Relaxed phylogenetics and dating with confidence. *PLoS Biol*, 4: e88.

[14] Neher RA (2022) Contributions of adaptation and purifying selection to SARS-CoV-2 evolution. *Virus Evol*, 8: veac113.

[15] Wu F, *et al.* (2020) A new coronavirus associated with human respiratory disease in China. *Nature*, 579: 265–269.

[16] Lemieux JE, *et al.* (2021) Phylogenetic analysis of SARS-CoV-2 in Boston highlights the impact of superspreading events. *Science*, 371: eabe3261.

[17] Markov PV, *et al.* (2023) The evolution of SARS-CoV-2. *Nat Rev Microbiol*, 21: 361–379.

[18] Goiriz L, *et al.* (2023) A variant-dependent molecular clock with anomalous diffusion models SARS-CoV-2 evolution in humans. *Proc Natl Acad Sci USA*, 120: e2303578120.

[19] Yang H, Rao Z (2021) Structural biology of SARS-CoV-2 and implications for therapeutic development. *Nat Rev Microbiol*, 19: 685–700.

[20] Rouzine IM, Rodrigo A, Coffin JM (2001) Transition between stochastic evolution and deterministic evolution in the presence of selection: general theory and application to virology. *Microbiol Mol Biol Rev*, 65: 151–185.

[21] Obermeyer F, *et al.* (2022) Analysis of 6.4 million SARS-CoV-2 genomes identifies mutations associated with fitness. *Science*, 376: 1327–1332.

[22] Bouckaert R, *et al.* (2014) BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS Comput Biol*, 10: e1003537.

[23] De Maio N, *et al.* (2023) Maximum likelihood pandemic-scale phylogenetics. *Nat Genet*, 55: 746–752.

[24] COVID-19 Genomics UK (COG-UK) consortium (2020) An integrated national scale SARS-CoV-2 genomic surveillance network. *Lancet Microbe*, 1: e99–e100.

[25] Tao K, *et al.* (2021) The biological and clinical significance of emerging SARS-CoV-2 variants. *Nat Rev Genet*, 22: 757–773.

[26] Gillespie JH (1984) The molecular clock may be an episodic clock. *Proc Natl Acad Sci USA*, 81: 8009–8013.

[27] Tsui JLH, *et al.* (2023) Genomic assessment of invasion dynamics of SARS-CoV-2 Omicron BA.1. *Science*, 381: 336–343.

[28] Wagner C, *et al.* (2024) Positive selection underlies repeated knockout of ORF8 in SARS-CoV-2 evolution. *Nat Commun*, 15: 3207.

[29] Tay JH, Porter AF, Wirth W, Duchene S (2022) The emergence of SARS-COV-2 variants of concern is driven by acceleration of the substitution rate. *Molecular Biology and Evolution*, 39(2), doi:10.1093/molbev/msac013.

[30] Khare S, *et al.* (2021) GISAID's role in pandemic response. *China CDC Weekly*, 3: 1049–1051.

[31] Katoh K, Misawa K, Kuma Ki, Miyata T (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res*, 30: 3059–3066.

[32] Goldman N, Yang Z (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol Biol Evol*, 11: 725–736.

[33] McKinney W (2012) Python for data analysis: Data wrangling with Pandas, NumPy, and IPython. *O'Reilly Media*.

[34] Cock PJA, *et al.* (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25: 1422–1423.

Addendum: sRNA-induced synthetic translational bursting in bacteria

All the effects of Nature are only the mathematical consequences of a small number of immutable laws.
— Pierre-Simon Laplace

This addendum complements the central theme of the thesis, which delves deeply into the macroscopic evolutionary processes in viruses and their stochastic underpinnings, by shifting the focus to the microscopic origins of biological noise. The stochastic nature of evolution is driven by complex interactions among population dynamics, molecular mechanisms, and environmental fluctuations. Within this broader framework, gene expression represents a fundamental, intrinsically noisy process whose randomness propagates upward to influence larger-scale phenomena such as evolution. Here, the addendum examines a specific molecular mechanism: translational bursting in gene expression, using synthetic sRNA systems in *E. coli* as a model. The examination of the mechanisms by which these systems induce and regulate cellular bursts offers a glimpse into the complex molecular mechanisms underlying biological noise and provides valuable insights into how molecular stochasticity propagates into evolutionary dynamics. This supports the contention that a multiscale perspective on randomness is essential for understanding and harnessing the intricacies of biological function.

To date, this work has not been published in a peer-reviewed journal.

Further reading on related works to which I contributed, and whose methodologies are based on similar principles, can be found in the following references:

Dolcemascolo R, **Goiriz L**, Montagud-Martínez R, Rodrigo G. (2022) Gene regulation by a protein translation factor at the single-cell level. *PLoS Comput. Biol.*, 18(5): e1010087.

Dolcemascolo R, Heras-Hernández M, **Goiriz L**, Montagud-Martínez R, Requena-Menéndez A, Ruiz R, Pérez-Ràfols A, Higuera-Rodríguez R.A, Pérez-Ropero G, Vranken W.F, Martelli T, Kaiser W, Buijs J, Rodrigo G. (2024) Repurposing the mammalian RNA-binding protein Musashi-1 as an allosteric translation repressor in bacteria. *eLife*, 12: RP91777.

Dolcemascolo R, Ruiz R, Baldanta S, **Goiriz L**, Heras-Hernández M, Montagud-Martínez R, Rodrigo G. (2024) Probing the orthogonality and robustness of the mammalian RNA-binding protein Musashi-1 in Escherichia coli. *J. Biol. Eng.* 18(52).

Introduction

Gene expression is inherently stochastic, resulting in considerable cell-to-cell variability even among clonal bacterial populations [1, 2]. One major source of this variability is bursting, where transcription or translation occurs in episodic bouts rather than continuously. Single-cell studies in *E. coli* have shown that mRNA synthesis happens in random pulses [3], and that protein production occurs in sporadic bursts [4]. These findings validate theoretical predictions that rare events, such as a promoter switching to an active state, can drive phenotypic heterogeneity, a behavior long described by stochastic models like the random telegraph model [5]. Moreover, the intrinsic noise from low copy-number molecules leads to mRNA and protein distributions that deviate from simple Poisson statistics [6], which is crucial for both understanding genetic circuits and designing robust biotechnological applications [7].

While transcriptional bursting, driven by promoter activation kinetics, is well established, recent studies have focused on post-transcriptional bursting (fluctuations arising after mRNA is synthesized). In bacteria, small regulatory RNAs (sRNAs), which are approximately 50-200 nucleotides long, modulate target mRNA translation and stability [8, 9]. In *E. coli*, many sRNAs bind near the 5' untranslated region of mRNAs, often with the assistance of the Hfq chaperone, to block ribosome binding and promote degradation. Unlike protein transcription factors, sRNAs act stoichiometrically, leading to an all-or-none effect on individual mRNAs: a transcript is either sequestered in an inactive sRNA-bound complex (translation "OFF") or remains free for translation (translation "ON"). This binary regulation can result in bursts of protein production when sRNA levels drop or become saturated [10].

Mathematical modeling has been key to understanding post-transcriptional bursting. Approaches such as stochastic differential equations and chemical master equation formulations have been employed to capture the dynamics of sRNA-mRNA interactions. Models analogous to the promoter telegraph model describe two states, "ON" (mRNA free of sRNA) and "OFF" (mRNA bound by sRNA), and can replicate the bursty behavior observed in experiments. The stochastic simulation algorithm, along with analytical methods like linear noise approximation, has clarified how parameters such as binding affinities, transcription and degradation rates determine burst size and frequency. Empirical studies, including single-cell fluorescence and smFISH, have compared these model predictions with experimental data, validating the role of sRNA regulation in modulating gene expression noise [3, 7, 10, 11].

Synthetic biology further leverages these insights to both test and harness post-transcriptional regulation. For instance, CRISPR interference (CRISPRi) employs a catalytically inactive Cas9 and guide RNA to reversibly modulate gene activity, mimicking natural on/off switching [12]. Similarly,

synthetic sRNA systems, designed with tailored antisense regions, reproduce natural sRNA mechanisms to control translation [13]. Additionally, toehold switch RNAs, engineered mRNA structures that block translation until triggered by a specific RNA, demonstrate how RNA-only systems can be used to construct controllable gene circuits [14, 15]. These synthetic tools provide powerful platforms to investigate the fundamental principles of biological noise from molecular scale up to its propagation to the evolutionary scale.

Here, we harness a set of mutants of the RAJ11 construct, which is is a synthetic riboregulatory system designed to control gene expression at the post-transcriptional level by exploiting base-pairing interactions between a small regulatory RNA (sRNA) and the 5' untranslated region (5' UTR) of a target mRNA [16], to empirically assess the phenomenon of post-transcriptional bursting. In this design, the ribosome-binding site on the mRNA is initially sequestered by an intramolecular hairpin, preventing efficient translation. When the sRNA binds to a complementary "toehold" region within that hairpin, it causes a conformational rearrangement that liberates the ribosome-binding site and activates translation. We examined multiple variants of this system, including the wild-type RAJ11 design and several engineered mutants with altered sRNA-mRNA binding affinities, to systematically probe how the strength and dynamics of sRNA-mediated repression influence translational bursting. All constructs share the same overall architecture, allowing direct comparisons. Overall, this system serves as a minimal and predictable framework for studying how RNA–RNA interactions, influenced by thermodynamic parameters and sequence accessibility, modulate protein output.

Results

Deconvolution of Expression Distributions

To accurately quantify the protein output distribution, we deconvolved cellular autofluorescence from the total fluorescence measurements to isolate the true protein expression signals from background noise. Single-cell fluorescence measurements of our synthetic sRNA constructs yield unimodal but highly skewed intensity distributions of protein per cell (**Fig. 5.6**) due to the burst-like nature of protein expression. Without removing the background contribution, these heavy-tailed distributions would be convolved with extrinsic noise from cellular autofluorescence, potentially obscuring their true shape and overestimating cell-to-cell variability that is not due to actual gene expression bursts. By mathematically deconvolving the measured fluorescence data using a control autofluorescence profile, we eliminate this artifact and recover the genuine distribution of protein expression levels. This

procedure ensures that analyses of translational bursting (e.g., determining burst sizes and frequencies) reflect actual biological variability rather than measurement noise, thereby allowing accurate quantification of the noise characteristics inherent to the sRNA-regulated system.

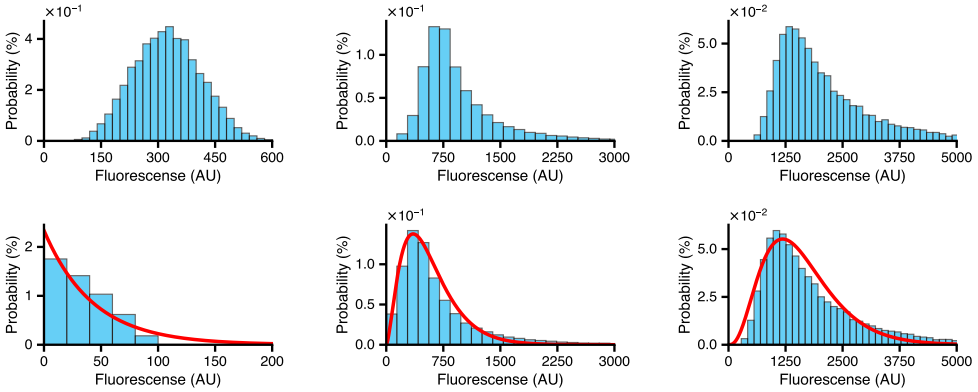

Figure 5.6: Histograms of the measured fluorescence signals (top row) and their deconvolution results (bottom row) for three RAJ11 constructs (`RAJ11M`, `RAJ11_10`, and `RAJ11_100`). Bars represent the empirical probability distribution of the fluorescence intensities, whereas the red curves in the bottom row denote the fitted gamma distributions based on the estimated shape and scale parameters.

We observed that the wild-type sRNA-regulated construct produces bursts of protein that lead to a heavy-tailed distribution. The deconvolved distributions represent the true protein expression variability attributable to the RAJ11 construct. All conditions yielded unimodal but highly skewed distributions of protein per cell. We observed that the wild-type sRNA-regulated construct produces bursts of protein that lead to a heavy-tailed distribution.

Key findings are that all sRNA-regulated cases produce right-skewed, bursty protein distributions, and weaker sRNA–mRNA interactions lead to higher noise (greater cell-to-cell variability) for a comparable mean output (**Fig. 5.6**, **Fig. 5.7**). These results quantitatively demonstrate post-transcriptional bursting in a real genetic system, in agreement with theoretical expectations [11]. Moreover, by fitting the distributions, we confirm that gamma-like forms appropriately describe the protein count variability, as seen previously in bursty gene expression systems. The wild-type sRNA effectively reduces relative noise and skewness compared to the weaker mutants, highlighting an intrinsic noise-control mechanism in the post-transcriptional bursting regime.

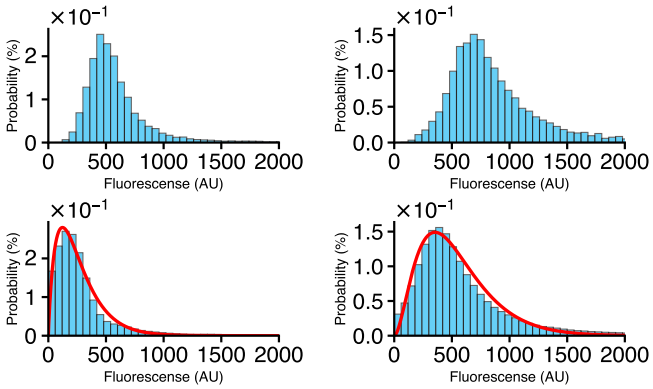

Figure 5.7: Histograms of the measured fluorescence signals (top row) and their deconvolution results (bottom row) for `RAJ11_100_m26` (left) and `RAJ11_100_m32` (right). Bars depict the empirical probability distributions of the fluorescence intensities, while the red curves in the bottom panels represent fitted gamma distributions with the estimated shape and scale parameters.

Discussion

Mechanism of Post-Transcriptional Bursting

Our preliminary findings suggest that that small RNAs can induce bursts of protein production by toggling translation on and off. In the RAJ11 system, the sRNA binding to the 5'UTR acts as a molecular switch: when bound, it exposes the Shine-Dalgarno sequence and triggers a period of active translation (the "on" state). Each binding event can generate a burst of protein until the complex dissociates or the mRNA is deactivated, at which point translation returns to "off". This mechanism is analogous to transcriptional bursting (where promoter states drive bursts of mRNA production) but occurs at the post-transcriptional level, downstream of mRNA synthesis. The derived burst frequency ($\sim k_{\text{on}}\langle s \rangle$) and burst size ($\sim k_p/k_{\text{off}}$) provide a useful intuitive picture: a higher sRNA concentration or binding rate increases how often bursts occur, while a slower unbinding (or inactivation) rate prolongs each translational burst, yielding more protein per burst.

Our incomplete analytical treatment (using Langevin equations and a telegraph process) is yet to reproduce these burst characteristics, consistent with previous stochastic models of gene expression bursts. Importantly, the post-transcriptional bursts we observe are a direct consequence of molecule-to-molecule variability at the RNA level (sRNA–mRNA interactions), rather than promoter state changes, which is consistent with the concept introduced by [11]: bursts of protein synthesis can arise "beyond" transcription, through dynamic RNA-level regulation.

In essence, an mRNA that is otherwise translationally silent can, upon sRNA binding, stochastically produce a flurry of protein, thereby increasing cell-to-cell heterogeneity in protein levels. Such bursts were evident in our experiments as heavy-tailed protein distributions. The right-skewed gamma distributions fitted to our data (with shape parameters <4) are indicative of infrequent large bursts and corroborate the idea that translational activation is episodic. These distributional signatures mirror those of classical transcriptional bursting – for example, the negative binomial distributions reported for noisy gene expression in bacteria – highlighting a unifying principle: bursting is a fundamental outcome of on/off dynamic regulation, whether at the promoter or post-transcriptional level.

Tuning of Expression Noise by sRNA–mRNA Interactions

By comparing wild-type and mutant sRNAs, it is hinted that the kinetic parameters of sRNA binding (and unbinding) directly influence the magnitude of noise in protein expression. All variants produced similar mean protein levels in our conditions (owing to compensatory adjustments in burst frequency/duration), yet their noise profiles differed markedly. The wild-type sRNA, which has a strong binding affinity (large negative ΔG of binding), yielded a protein count distribution with lower normalized variance (CV) and moderate skewness. In contrast, the weakened-binding mutant (with seed mismatches) exhibited nearly Poisson-level noise and higher relative variance, even though the average protein output remained comparable. This trend suggests that strong sRNA binding "smooths out" protein production, making translation more constitutive at the single-cell level.

These observations resonate with studies of microRNA-based regulation in eukaryotes, where introducing an RNA-mediated repression can either buffer or amplify noise depending on kinetic regime [17]. In our bacterial case, the sRNA serves as an activator of translation, but the effect on noise is analogous: robust regulation (fast binding) buffers noise, whereas weaker regulation approaches the unregulated scenario of random translation initiation with full noise.

Notably, [18] observed in an *E. coli* sRNA network that noise is strongly influenced by sRNA-mediated mRNA decay and complex formation rates, which is consistent with our findings that altering sRNA–mRNA interaction strength modulates noise. Thus, our results bridge genotype (RNA sequence mutations) to phenotype (noise in protein levels): small changes in the seed region free energy (on the order of a few kcal/mol) can move the system from a relatively "buffered" expression regime to a highly stochastic one, without significantly changing mean expression. This tunability is of great interest for synthetic biology, as it offers a knob to adjust phenotypic variability

independently of mean output [10, 19].

Moreover, the ability to tune noise via RNA regulators supports the idea that cells could evolve or select for different noise levels by mutations in non-coding regions – adding an evolutionary dimension to the role of post-transcriptional regulation.

Comparison to Transcriptional Bursting and Biological Implications

It is insightful to compare the bursting we observe post-transcriptionally with the well-characterized phenomenon of transcriptional bursting. Transcriptional bursts (arising from promoters switching between inactive and active states) also produce heavy-tailed mRNA distributions and super-Poissonian noise [20]. In our post-transcriptional bursting, the ultimate protein distributions can be mathematically similar (e.g. both can often be described by negative binomial statistics), but the mechanistic origin is different: instead of promoter chromatin or transcription factor binding kinetics, it is the RNA-RNA interaction kinetics that set the burst parameters. One practical implication is in the timing and integration of cellular decisions. Transcriptional bursting is often implicated in fate determination events by providing occasional high transcript pulses. Post-transcriptional bursting could likewise generate heterogeneity in protein levels from uniformly transcribed mRNAs, thereby enabling phenotypic diversification without changes at the DNA level. This may be particularly relevant in stress responses or developmental contexts where small RNAs are known to play critical regulatory roles. Our results underscore that the RNA layer adds an additional opportunity for noise modulation in gene expression.

The distinctive feature of post-transcriptional bursts is that they can be rapidly tunable by introducing or modulating a regulatory RNA. In contrast, altering transcriptional burst kinetics often requires promoter mutations or changing transcription factor dynamics. From an evolutionary perspective, an organism might employ sRNA regulators to finetune expression noise on short evolutionary timescales (since RNA sequences can mutate to adjust binding affinity) while keeping the mean expression of a gene constant. This could be advantageous for bet-hedging strategies in fluctuating environments, or for coordinating multi-gene networks where only certain genes' expression noise needs adjustment. In summary, our study demonstrates that post-transcriptional bursting is a distinct but analogous phenomenon to transcriptional bursting, broadening the paradigm of how stochastic gene expression can arise. It highlights the often under-appreciated role of RNA interactions in generating phenotypic diversity, supporting the view that gene regulation at the RNA level is as significant as DNA-level regulation in shaping

cellular heterogeneity.

Future investigations could explore the interplay of transcriptional and post-transcriptional bursts occurring together, and how cells orchestrate these layers of regulation to achieve desired expression dynamics [21]. Our findings not only contribute to the fundamental understanding of gene expression noise but also provide design principles for constructing synthetic gene circuits with controllable noise via sRNA regulators.

Materials and Methods

RAJ11 Construction Configuration

The wild-type RAJ11 configuration deesigned on [16] comprises two core genetic elements: an sRNA module specifically designed to bind a complementary stretch of nucleotides (the "toehold") on the 5' UTR of the target mRNA, and a 5' UTR whose native secondary structure sequesters the ribosome-binding site (RBS). In the absence of the sRNA, this 5' UTR hairpin prevents ribosomes from accessing the RBS, resulting in minimal basal translation. The sRNA was computationally optimized in [16] to expose a highly accessible seed region that can quickly and stably hybridize to the 5' UTR. Once bound, the intramolecular hairpin in the mRNA is destabilized, making the RBS accessible for translation initiation. This design achieves tight cis-repression of the downstream GFP gene, followed by robust trans-activation only when the RAJ11 sRNA is present.

In addition to the wild-type RAJ11 configuration (referred to here as `RAJ11M`), we generated two key mutants, `RAJ11_m26` and `RAJ11_m32`, each introducing specific nucleotide substitutions that either modify the seed region of the sRNA or its complementary segment on the 5' UTR. For instance, `RAJ11_m26` carries selected base changes in the sRNA seed to evaluate how disruptions in base-pair complementarity affect binding kinetics, whereas `RAJ11_m32` alters nucleotides within the 5' UTR to test changes in RBS accessibility under partial or complete mismatch scenarios. By comparing the dynamic range of GFP expression across these constructs, we aimed to determine the extent to which each mutation weakens sRNA–mRNA pairing. This systematic set of mutants was instrumental for validating our thermodynamic predictions step by step, as it allowed us to link specific sequence changes to measurable shifts in translational activation.

Furthermore, we placed the `RAJ11M` configuration in both a high-copy-number plasmid (`RAJ11M`) and a low-copy-number plasmid (`RAJ11M_P15`), and tested each under varying levels of anhydrotetracycline (ATC) induction, specifically 0 ng/mL (`RAJ11M`), 10 ng/mL (`RAJ11_10`), and 100 ng/mL (`RAJ11_100`). We then introduced the `RAJ11_m26` and `RAJ11_m32`

mutants with 100 ng/mL of ATC (`RAJ11_100_m26` and `RAJ11_100_m32`). By comparing the resulting GFP outputs across these plasmid backgrounds and induction levels, we aimed to evaluate how increased sRNA transcription affects riboregulation in both wild-type and mutant variants.

Mathematical Model of sRNA-Mediated Gene Expression

We adopted the stochastic modeling framework described in [11] to describe a gene post-transcriptionally regulated by a small RNA. The system consists of an mRNA (x) and its protein product (y). Transcription of mRNA occurs with rate k_m, c is the copy number of the gene, and translation of protein occurs only when the mRNA is in an active state (ribosome-binding site accessible) with rate k_p. The sRNA binding toggles this state. We formalized this using two coupled Langevin equations (Ito stochastic differential equations) for $x(t)$ and $y(t)$:

$$\frac{dx}{dt} = ck_m - \gamma_m x + \sqrt{ck_m + \gamma_m x}\xi_m(t) \tag{5.1}$$

$$\frac{dy}{dt} = k_p\zeta(t) - \gamma_p y + \sqrt{k_p, \zeta(t) + \gamma_p y}\xi_p(t) \tag{5.2}$$

where γ_m and γ_p are the degradation rates of mRNA and protein, respectively (so $1/\gamma_m$, $1/\gamma_p$ are their average lifetimes). The terms $\xi_m(t)$ and $\xi_p(t)$ represent independent Gaussian white noise processes (with $\langle\xi_i(t)\rangle = 0$ and $\langle\xi_i(t)\xi_j(t')\rangle = \delta_{ij}\delta(t-t')$) accounting for intrinsic stochastic fluctuations.

The steady state solution for $\mathbb{E}[x(t)]$ is given by

$$\begin{aligned} 0 &= \mathbb{E}\left[ck_m - \gamma_m x + \sqrt{ck_m + \gamma_m x}\xi_m(t)\right] \\ 0 &= ck_m - \gamma_m \mathbb{E}[x] \\ \mathbb{E}[x(t)] &= \frac{ck_m}{\gamma_m} \end{aligned}$$

We can only get this expression through 3 important assumptions:

1. The noise term is zero on average (given by the white noise process).
2. The amplitude of the stochastic process and the stochastic process are uncorrelated, hence the expectation of the product is the product of the expectations.
3. We are using the mean-field approximation, i.e. the amplitude of the noise is finite, hence it disappears when multiplying by 0.

Now that we have the expression for $\langle x \rangle$, we can substitute it back into the original equation by means of the mean-field approximation ($x = \langle x \rangle + \Delta x$):

$$\begin{aligned}
\frac{d\left(\mathbb{E}[x] + \Delta x(t)\right)}{dt} &= ck_m - \gamma_m \left(\mathbb{E}[x] + \Delta x(t)\right) + \sqrt{ck_m + \gamma_m \left(\mathbb{E}[x] + \Delta x(t)\right)}\xi_m(t) \\
\frac{d\Delta x(t)}{dt} &= ck_m - \gamma_m \mathbb{E}[x] - \gamma_m \Delta x(t) + \sqrt{ck_m + \gamma_m \mathbb{E}[x]}\xi_m(t) \\
\frac{d\Delta x(t)}{dt} &= -\gamma_m \Delta x(t) + \sqrt{2ck_m}\xi_m(t)
\end{aligned}$$

To solve the above equation, we will use the Fourier transform of the equation. We'll denote the Fourier transform of f as $\hat{f}$, equivalently as $\mathcal{F}\{f\}$. In this analysis, we assume that $\Delta x(t)$ is interpreted within the framework of tempered distributions, specifically belonging to the dual space $\mathcal{S}'(\mathbb{R})$ of the Schwartz space. This assumption guarantees that the Fourier transform is well-defined, even though white noise is not square-integrable in the classical sense.

$$\begin{aligned}
\mathcal{F}\left\{\frac{d\Delta x(t)}{dt}\right\} &= \mathcal{F}\left\{-\gamma_m \Delta x(t) + \sqrt{2ck_m}\xi_m(t)\right\} \\
i\omega\hat{\Delta x}(\omega) - \Delta x_0 &= -\gamma_m \hat{\Delta x}(\omega) + \sqrt{2ck_m}\hat{\xi}_m(\omega) \\
\hat{\Delta x}(\omega) &= \left(\frac{\sqrt{2ck_m}}{i\omega + \gamma_m}\right)\hat{\xi}_m(\omega)
\end{aligned}$$

This shows that the transfer function from noise to mRNA fluctuations is given by $G_x(i\omega) = \frac{\sqrt{2ck_m}}{\gamma_m + i\omega}$. Note that we have proceeded with the assumption that the initial condition is zero ($\Delta x_0 = 0$). The next step of the derivation involves using the convolution theorem.

Theorem (Convolution theorem)**.** Let f and g be two functions with Fourier transforms $\hat{f}$ and $\hat{g}$, respectively. Then the convolution of f and g, denoted as $f * g$ is given by:

$$f(t) * g(t) = \mathcal{F}^{-1}\left\{\hat{f}(\omega)\hat{g}(\omega)\right\}$$

In simpler terms, the Fourier transform of the convolution of two functions equals the product of the Fourier transforms of each of the functions. In our particular case, we have

$$\hat{\Delta x}(\omega) = \hat{g}(\omega)\hat{\xi}_m(\omega)$$

where $\hat{g}(\omega) = \left(\frac{\sqrt{2ck_m}}{i\omega + \gamma_m}\right)$. Therefore,

$$\Delta x(t) = \int_0^t \xi_m(\tau) g(t-\tau) d\tau$$

It is then necessary to compute the inverse Fourier transform of $\hat{g}(\omega)$. Our primary objectives are to verify the convergence of the integral obtained via the inverse Fourier transform and to explicitly address the behavior for $t < 0$. In this context, we note that for $t < 0$ the solution is defined to be zero, corresponding to the general form $H(t)e^{-\alpha t}$, where $H(t)$ is the Heaviside step function. Since the convolution integral is evaluated over the interval $(0, t)$, the case for $t < 0$ is naturally excluded. The inverse Fourier transform of $\hat{g}(\omega)$ is given by:

$$\begin{aligned} g(t) &= \mathcal{F}^{-1}\left\{\hat{g}(\omega)\right\} \\ &= \left(\sqrt{2ck_m}\right) \mathcal{F}^{-1}\left\{\frac{1}{i\omega + \gamma_m}\right\} \\ &= \left(\sqrt{2ck_m}\right) \frac{1}{2\pi} \int_{-\infty}^{+\infty} \frac{1}{i\omega + \gamma_m} e^{i\omega t} d\omega \end{aligned}$$

Let's define $g(z) = \frac{1}{iz+\gamma_m}$, with $z \in \mathbb{C}$. We will introduce the following theorems:

Theorem. Suppose $f(z)$ is a function of $z \in \mathbb{C}$ and $f(z)$ is defined in the upper-half plane. If there is an $a > 1$ and $M > 0$ such that

$$|f(z)| < \frac{M}{|z|^a}$$

for a large $|z|$, then

$$\lim_{R\to\infty} \int_{C_R} f(z) dz = 0$$

Where C_R is the path (in this case an arch) that goes from R in the real axis and goes back to $-R$ through the upper-half plane. This theorem is for functions that decay faster than $\frac{1}{z}$.

Theorem. Suppose $f(z)$ is a function of $z \in \mathbb{C}$ and $f(z)$ is defined in the upper-half plane. If there is $M > 0$ such that

$$|f(z)| < \frac{M}{|z|}$$

for a large $|z|$, then

$$\lim_{x_1\to\infty, x_2\to\infty} \int_{C_1+C_2+C_3} f(z)e^{iaz}dz = 0$$

Where $C_1 + C_2 + C_3$ is a rectangular path that goes from point x_1 in the real axis up to the upper-half plane until height $(x_1 + x_2)i$, then goes to the left until $-x_2$ and goes down to the real axis. This theorem is for functions that decay like $\frac{1}{z}$.

For our particular case, we'll use the previous theorem (as we can see that $g(z)$ decays like $\frac{1}{z}$). So, we need to find a value of $M > 0$ so that the bound holds:

$$\begin{aligned}
|g(z)| &= \left|\frac{1}{iz+\gamma_m}\right| \\
&= \frac{1}{|iz+\gamma_m|} \\
&= \frac{1}{\sqrt{(\mathrm{Re}(z))^2 + (\gamma_m - \mathrm{Im}(z))^2}} \\
&= \frac{1}{\sqrt{\gamma_m^2 - 2\gamma_m \mathrm{Im}(z) + \mathrm{Re}^2(z) + \mathrm{Im}^2(z)}} \\
&= \frac{1}{\sqrt{\gamma_m^2 - 2\gamma_m \mathrm{Im}(z) + |z|^2}}
\end{aligned}$$

for a large $|z|$, we can see that $|g(z)| \approx \frac{1}{|z|}$, so we can use the theorem as any $M > 1$ is sufficient to make the bound hold:

$$\begin{aligned}
|g(z)| &= \frac{1}{\sqrt{\gamma_m^2 - 2\gamma_m \mathrm{Im}(z) + |z|^2}} \\
&\approx \frac{1}{|z|} < \frac{M}{|z|}
\end{aligned}$$

With this condition met, we can use the standard contour $C_1+C_2+C_3+C_4$ which forms a rectangle that goes from point x_1 in the real axis up to the upper-half plane until height $(x_1 + x_2)i$, then goes to the left until $-x_2$, then goes down to the real axis and finally goes back to our starting point. Note then that if we take the limits $x_1 \to \infty$ and $x_2 \to \infty$ on contour C_4, we have our original integral (the one from the inverse Fourier transform):

$$\lim_{x_1\to\infty,x_2\to\infty}\int_{C_4} g(z)e^{iaz}dz = \lim_{x_1\to\infty,x_2\to\infty}\int_{-x_2}^{x^1}\frac{e^{izt}}{iz+\gamma_m}dz$$

Furthermore, given the bound $M > 1$ we know that

$$\lim_{x_1\to\infty,x_2\to\infty}\int_{C_1+C_2+C_3} g(z)e^{iaz}dz = 0$$

Therefore, the entire contour integral over $C_1 + C_2 + C_3 + C_4$:

$$\lim_{x_1\to\infty,x_2\to\infty}\int_{C_1+C_2+C_3+C_4} g(z)e^{iaz}dz = 0 + \lim_{x_1\to\infty,x_2\to\infty}\int_{C_4}\frac{e^{izt}}{iz+\gamma_m}dz$$

So, the inverse Fourier transform is given by the contour integral over $C_1 + C_2 + C_3 + C_4$. We can then use the residue theorem to compute the integral.

Theorem (Cauchy's Residue Theorem)**.** Let $f(z)$ be a function that is analytic in a region D except for a finite number of isolated singularities. For any simple closed curve C that is positively oriented and lies in D:

$$\oint_C f(z)dz = 2\pi i\sum_{k=1}^{n}\text{Res}(f, z_k)$$

where z_k are the singularities of $f(z)$ in D and $\text{Res}(f, z_k)$ denotes the residue of f at z_k.

The residue of a function $f(z)$ at a point z_0 is given by:

$$\text{Res}(f, z_0) = \lim_{z\to z_0}(z - z_0)f(z)$$

In our case, we have a simple pole at $z = \gamma_m i$, as the denominator of $g(z)$ is zero at $z = \gamma_m i$. Since $\gamma_m > 0$, we have that the pole is in the upper-half plane. In addition, the pole is a simple pole, we can compute the residue through the definition:

$$\begin{aligned}
\text{Res}(g, \gamma_m i) &= \lim_{z\to\gamma_m i}(z - \gamma_m i)\left(\frac{e^{izt}}{iz+\gamma_m}\right)\\
&= e^{-\gamma_m t}\lim_{z\to\gamma_m i}\left(\frac{z-\gamma_m i}{iz+\gamma_m}\right)\\
&= \frac{e^{-\gamma_m t}}{i}\lim_{z\to\gamma_m i}\left(\frac{iz+\gamma_m}{iz+\gamma_m}\right) = \frac{e^{-\gamma_m t}}{i} = -ie^{-\gamma_m t}
\end{aligned}$$

Now we can use Cauchy's Residue Theorem to compute the integral:

$$\begin{aligned}\int_{C_4} \frac{e^{izt}}{iz+\gamma_m} dz &= 2\pi i \text{Res}(g, \gamma_m i) \\ &= 2\pi i \left(-ie^{-\gamma_m t}\right) \\ &= 2\pi e^{-\gamma_m t}\end{aligned}$$

Returning to the inverse Fourier transform, we have:

$$\begin{aligned}g(t) &= \left(\sqrt{2ck_m}\right)\left(\frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{e^{i\omega t}}{i\omega+\gamma_m} d\omega\right) \\ &= \sqrt{2ck_m} e^{-\gamma_m t}\end{aligned}$$

Therefore, the general solution for $\Delta x(t)$ is given by:

$$\begin{aligned}\Delta x(t) &= \int_0^t \xi_m(\tau) g(t-\tau) d\tau \\ &= \sqrt{2ck_m} \int_0^t \xi_m(\tau) e^{-\gamma_m(t-\tau)} d\tau\end{aligned}$$

Besides the above, we can also compute the spectral density of $\Delta x(t)$, which we'll denote as $S_{\Delta x}$. The spectral density is defined as the squared absolute value of the Fourier transform of the process. So for our process, we have that

$$\begin{aligned}S_{\Delta x} &= \left|\mathcal{F}\left\{\Delta x\right\}\right| = \left|\hat{\Delta x}\right|^2 \\ &= \left|\sqrt{2ck_m}\left(\frac{1}{i\omega+\gamma_m}\right)\hat{\xi}_m(\omega)\right|^2 \\ &= \left(\frac{2ck_m}{\omega^2+\gamma_m^2}\right)\left|\hat{\xi}_m(\omega)\right|^2\end{aligned}$$

Do note that one needs to be extra careful when using these definitions, because the Fourier transform of white noise process is defined only as a kind of limit of smooth processes. Fortunately, as long as we only work with linear systems this definition indeed works. And it provides a useful tool for determining covariance functions of stochastic differential equations.

A great property is that, since the Fourier transform of white noise is a constant (imagine a white noise signal in the time domain; it appears as a

random fluctuation where the amplitude changes rapidly and unpredictably. When the Fourier transform of this signal is taken to analyze its frequency components, it can be seen that there is no preference for any specific frequency; all frequencies are equally likely to contribute to the noise), the spectral density of the process is also a constant.

Therefore

$$\begin{aligned} S_{\Delta x} &= \left(\frac{2ck_m}{\omega^2 + \gamma_m^2} \right) \left| \hat{\xi}_m(\omega) \right|^2 \\ &= \frac{2ck_m}{\omega^2 + \gamma_m^2} \end{aligned}$$

Next we need to introduce the Wiener-Khinchin theorem.

Theorem (Wiener-Khinchin)**.** For a wide-sense stationary random process, the power spectral density $S_x(\omega)$ and the autocorrelation function $C_x(t)$ form a Fourier transform pair:

$$S_x(\omega) = \mathcal{F}\{C_x(t)\} \quad \text{and} \quad C_x(t) = \mathcal{F}^{-1}\{S_x(\omega)\}$$

where $\mathcal{F}$ denotes the Fourier transform and $\mathcal{F}^{-1}$ its inverse.

Let's denote the autocorrelation function of $\Delta x(t)$ as $C_{\Delta x}$, so by means of the Wiener-Khinchin theorem, we have that

$$\begin{aligned} C_{\Delta x} &= \mathcal{F}^{-1}\left\{S_{\Delta x}\right\} \\ &= \mathcal{F}^{-1}\left\{ \frac{2ck_m}{\omega^2 + \gamma_m^2} \right\} \\ &= \frac{2ck_m}{2\gamma_m} \mathcal{F}^{-1}\left\{ \frac{2\gamma_m}{\omega^2 + \gamma_m^2} \right\} \\ &= \frac{ck_m}{\gamma_m} e^{-\gamma|\Delta t|} \end{aligned}$$

Note that the inverse Fourier transform here was computed using the property of the Fourier transform being a linear operator, and using the known result that $\mathcal{F}^{-1}\left\{\frac{2\alpha}{\alpha^2+\omega^2}\right\} = e^{\alpha|t|}$

With these definitions, it is left to compute the statistical properties of the protein $y(t)$, which is left unsolved as this work is yet to be published.

Data Collection and Analysis

For experimental validation, *E. coli* cells carrying the RAJ11 construct were analyzed by flow cytometry. The construct consisted of a reporter gene with the cis-RAJ11 5'UTR and a constitutive promoter, and sRNA (trans-RAJ11) expressed from a regulated promoter. We collected single-cell fluorescence distributions for: (1) no sRNA (sRNA promoter off), (2) wild-type sRNA (induced to a certain level, labeled "100" which corresponds to a saturating inducer concentration), and (3) two sRNA mutants (with the same induction level). Three biological replicate samples were measured for each condition. To isolate the true protein signal distribution, we performed a deconvolution to subtract autofluorescence. The autofluorescence (background fluorescence of cells with no reporter) was measured and its distribution $b(x)$ obtained as a histogram. The raw fluorescence distribution with the reporter (which includes both background and reporter signal) is denoted $a(x)$. We treated the relationship as a convolution $a(x) = (b * c)(x)$, where $c(x)$ is the sought distribution of true reporter signal.

We implemented the deconvolution in Python by constructing the convolution matrix B for b and solving the linear system $B, \hat{c} = a$ for $\hat{c}$ (discrete approximation of the convolution equation). Because this inverse problem is ill-posed and B is non-square, we obtained a least-squares solution with Tikhonov regularization (ridge regression) to enforce smoothness and non-negativity. Specifically, we solved $(B^T B + \lambda I)\hat{c} = B^T a$ with regularization parameter $\lambda = 1$ (chosen empirically to suppress high-frequency noise in $\hat{c}$ without over-smoothing). The solution $\hat{c}(x)$ was taken as the deconvolved probability density of cellular fluorescence due to the reporter. We validated the deconvolution by reconvolving $\hat{c}$ with b to obtain $\hat{a}$ and comparing it to the measured a. The coefficient of determination R^2 between a and $\hat{a}$ was above 0.95 in all cases, indicating a good fit. We also computed a t-statistic to confirm that $\hat{c}$ significantly improved the fit over a null model.

The deconvolved distributions $\hat{c}(x)$ for each condition were then analyzed to extract quantitative descriptors of bursting. We fit each distribution with a gamma distribution, $P(x) = x^{k-1}e^{-x/\theta}/[\Gamma(k)\theta^k]$, using maximum likelihood (via SciPy's `fit` function). This provided the shape parameter k and scale θ for each condition, along with estimates of uncertainty. The fits were visually excellent and yielded $R^2 > 0.95$ in each case. From the theoretical distribution and from the raw data, we calculated the mean, variance, noise (defined as variance divided by mean squared, CV^2), and skewness for each condition. These higher moments characterize bursty distributions: a Poisson distribution has $\mathrm{CV}^2 = 1$ and skewness $= 1$ (if treated as counts); higher values indicate more burstiness/variability. We found skewness strictly higher than 1 for all cases, indicating that the distributions are indeed bursty. Nonetheless,

further analysis of the data is needed to confirm the validity of the theoretical model.

Conclusions

This work takes a glipse into the pivotal role of sRNA-induced post-transcriptional regulation in shaping gene expression variability, looking to complement the well-characterized phenomenon of transcriptional bursting. Drawing on our modeling framework, single-cell fluorescence data, and the systematic exploration of wild-type vs. mutant sRNA constructs, we seek to demonstrate that sRNA-mediated translational on/off switching can yield robust bursts of protein production and generate heavy-tailed protein distributions. We envision these results to add depth to our understanding of multiscale biological noise: while transcriptional events can encode bursts at the mRNA level, the additional sRNA layer provides another point at which stochastic fluctuations may be amplified, modulated, or buffered.

These fluctuations, in turn, do contribute to a hierarchical cascade of noise propagation, where molecular-level stochasticity accumulates and integrates across multiple regulatory layers to ultimately shape macroscopic phenotypic variation. This multi-level aggregation of noise sources parallels other biological phenomena where microscopic randomness gives rise to emergent stochastic behaviors at larger scales – from protein conformational dynamics to evolutionary trajectories. Indeed, just as the superposition of multiple noise sources in gene expression can lead to non-Poissonian statistics and complex temporal patterns, the integration of various molecular-level mutational processes across different genes can collectively drive anomalous diffusion-like behaviors in evolutionary dynamics. Understanding these connections between microscopic stochasticity and macroscopic phenomenology is crucial for developing accurate quantitative models of biological variation across scales, and is yet to be fully uncovered.

References

[1] Elowitz MB, Levine AJ, Siggia ED, Swain PS (2002) Stochastic gene expression in a single cell. *Science*, 297(5584): 1183–1186.

[2] Swain PS, Elowitz MB, Siggia ED (2002) Intrinsic and extrinsic contributions to stochasticity in gene expression. *Proceedings of the National Academy of Sciences*, 99(20): 12795–12800.

[3] Golding I, Paulsson J, Zawilski SM, Cox EC (2005) Real-time kinetics of gene activity in individual bacteria. *Cell*, 123(6): 1025–1036.

[4] Yu J, *et al.* (2006) Probing gene expression in live cells, one protein molecule at a time. *Science*, 311(5767): 1600–1603.

[5] McAdams HH, Arkin A (1997) Stochastic mechanisms in gene expression. *Proceedings of the National Academy of Sciences*, 94(3): 814–819.

[6] Paulsson J (2004) Summing up the noise in gene networks. *Nature*, 427(6973): 415–418.

[7] Taniguchi Y, Choi PJ, Li GW, et al (2010) Quantifying *E. coli* proteome and transcriptome with single-molecule sensitivity. *Science*, 329(5991): 533–538.

[8] Gottesman S (2004) The small RNA regulators of *Escherichia coli*: roles and mechanisms. *Annual Review of Microbiology*, 58: 303–328.

[9] Waters LS, Storz G (2009) Regulatory RNAs in bacteria. *Cell*, 136(4): 615–628.

[10] Levine E, Zhang Z, Kuhlman T, Hwa T (2007) Quantitative characteristics of gene regulation by small RNA. *PLoS Biology*, 5(9): e229.

[11] Rodrigo G (2018) Post-transcriptional bursting in genes regulated by small RNA molecules. *Physical Review E*, 97(3): 032401.

[12] Qi LS, Larson BM, Gilbert LA, et al (2013) Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. *Cell*, 152(5): 1173–1183.

[13] Sharma V, Yamamura A, Yokobayashi Y (2012) Engineering artificial small RNAs for conditional gene silencing in *Escherichia coli*. *ACS Synthetic Biology*, 1(1): 6–13.

[14] Green AA, Silver PA, Collins JJ, Yin P (2014) Toehold switches: de-novo-designed regulators of gene expression. *Cell*, 159(4): 925–939.

[15] Qian L, Winfree E (2011) Scaling up digital circuit computation with DNA strand displacement cascades. *Science*, 332(6034): 1196–1201.

[16] Rodrigo G, *et al.* (2015) Exploring the Dynamics and Mutational Landscape of Riboregulation with a Minimal Synthetic Circuit in Living Cells. *Biophysical Journal*, 109(5): 1070–1076, doi:10.1016/j.bpj.2015.07.033.

[17] Schmiedel JM, *et al.* (2015) MicroRNA control of protein expression noise. *Science*, 348(6230): 128–132.

[18] Arbel-Goren R, *et al.* (2016) Transcript degradation and noise of small RNA-controlled genes in a switch activated network in Escherichia coli. *Nucleic Acids Research*, 44(14): 6707–6720.

[19] Mutalik VK, *et al.* (2012) Rationally designed families of orthogonal RNA regulators of translation. *Nature Chemical Biology*, 8(5): 447–454.

[20] Ozbudak EM, *et al.* (2002) Regulation of noise in the expression of a single gene. *Nature Genetics*, 31(1): 69–73.

[21] Tanenbaum ME, *et al.* (2014) A protein-tagging system for signal amplification in gene expression and fluorescence imaging. *Cell*, 159(3): 635–646.

Chapter 6

General Discussion

People think that mathematics is complicated.
Mathematics is the simple bit, it's the stuff we can understand.
It's cats that are complicated.
— John H. Conway

This work set out to examine how stochastic dynamics shape both viral genome evolution and post-transcriptional regulatory programs. By uniting large-scale sequence analysis with mechanistic modeling, it challenged long-held assumptions of constant molecular clocks and purely transcription-focused gene regulatory frameworks. We explored RNA virus genomes as prime examples of dynamic molecular automata, investigating how mutations accumulate over time and whether these changes align with the classic neutral theory [1, 2]. In addition, we briefly explored how bursty translational regulation as a source of noise that can create distinct gene expression states using synthetic gene circuits in bacteria as case study [3, 4]. Together, these projects advanced a core hypothesis: biological systems often deviate from simple, uniform models, and statistical-based methods are essential for fully understanding their complexity. This work emphasizes that evolution and gene regulation are inherently stochastic processes, with constraints, and shifting contexts shaping their trajectories.

Key findings highlight the importance of anomalous diffusion in viral genome evolution. Using SARS-CoV-2 data, this work uncovers that mutation accumulation deviates markedly from a simple constant-rate (Poissonian) process. While the average number of substitutions increased roughly linearly with time (consistent with the classical molecular clock hypothesis in a broad sense), the variance in mutation counts was substantially higher or lower than expected at different periods. In particular, distinct variant-dependent

patterns emerged: when new variants of concern arose and swept through the population, the evolutionary dynamics exhibited anomalous diffusion behavior. For example, the early Wuhan strain (Primal) and Alpha and Omicron variants showed subdiffusive dynamics – mutations accumulated more slowly and incrementally, as if evolution was temporarily "slowed down" – whereas the Delta variant showed superdiffusive behavior, with an accelerated pace of mutation accrual. This means that instead of a uniform random walk through sequence space, SARS-CoV-2 evolution alternated between periods of constrained, slower exploration and bursts of rapid change. Such findings are unprecedented in viral molecular evolution: they indicate that the virus' spread through sequence space cannot be captured by a single diffusion rate. Notably, these patterns corresponded to epidemiological transitions; bursts of genetic change coincided with the replacement of one variant by another (e.g., the rapid emergence of Alpha, then Delta, then Omicron). In these phases, the molecular clock "ticks" faster or slower depending on the variant, violating the assumption of a strictly constant rate. This result is strongly supported by our variance-based analysis and is further evidenced by fluctuations in the ratio of nonsynonymous to synonymous mutations (dN/dS) over time, which indicated predominantly purifying selection punctuated by episodes of adaptive evolution during variant sweeps. Contextualizing these findings, we note that the classical molecular clock concept (where genetic divergence grows linearly with time under neutral drift) has long been challenged by evidence of rate variation and overdispersion [5]. Our observations of variant-specific rate shifts provide a concrete example of such overdispersion in a real-world dataset, echoing earlier suggestions that viral evolutionary rates can depend on the timescale and conditions of measurement [5]. Traditional relaxed-clock phylogenetic models allow rates to vary across lineages [6], but the fractional Brownian motion framework employed here offers a more explicit characterization of the anomalous diffusion we detected, capturing the memory and heterogeneity in substitution processes.

The notion of anomalous diffusion itself is well-established in other realms of biology: for instance, macromolecules in cells often diffuse subdiffusively due to crowding [7, 8]. Observing analogous subdiffusive and superdiffusive regimes in viral genome evolution is a novel insight of this thesis, effectively importing concepts from statistical physics into evolutionary biology. It suggests that viral populations exploring sequence space can behave like particles in a complex medium, sometimes experiencing constraints (perhaps due to fitness landscapes or transmission bottlenecks) and occasionally taking long jumps (e.g., after a fitness leap or immune escape mutation).

Additionally, the brief exploration of bursty translational regulation as a source of noise at intracellular level ties into broader literature showing that

noise and variability are fundamental in gene expression and can be harnessed by regulatory designs [9]. Therefore, both intracellular processes and viral evolution deviate from simplistic models in predictable ways: anomalous diffusion and episodic rate changes emerge as unifying themes that connect our findings to foundational principles in biophysics and evolutionary theory [1, 5]. Despite these advances, there are important methodological and analytical limitations to acknowledge.

First, the sampling of viral genomic data is inherently biased. The sequences analyzed (e.g., SARS-CoV-2 genomes from global databases) are an opportunistic sample, skewed by factors such as surveillance intensity, geographic sampling disparities, and sequencing protocols over the pandemic's course. This means certain epochs or regions are over-represented while others (e.g., early spreading in low-surveillance areas) might be under-sampled, potentially influencing the observed mutation variance. Such bias is a common concern in phylodynamic studies and could lead to misestimation of diffusion parameters if, for instance, bursts of evolution coincided with periods of intensified sampling. Although we mitigated this by stringent data filtering and aggregation over time windows, a truly unbiased sample is unattainable and remains a caveat.

Second, there are assumptions in the modeling approach that constrain interpretation. The use of a fractional Brownian motion model with a single diffusion exponent exponent per viral variant, while capturing sub- or superdiffusive trends, simplifies the complex biological reality. In practice, multiple processes: continuous natural selection, changing transmission dynamics, and recombination events operate simultaneously, as we observed when diving deeper into a gene-by-gene analysis. Our model primarily attributes deviation from clock-like behavior to a diffusion exponent, effectively lumping together various causes into a single parameter. For example, purifying selection tends to remove deleterious mutations, which can lead to an apparent slowdown in divergence (subdiffusion) over time, whereas adaptive bursts (e.g., a new beneficial mutation sweeping) can cause sudden jumps (contributing to superdiffusion); our approach captures these in a phenomenological way but does not explicitly parameterize selection. Similarly, recombination (which is known to distort phylogenetic signals and clock estimates) was not explicitly modeled here [10], although there are widely known cases of viral recombination in the scenario of a co-infection. While SARS-CoV-2's recombinant lineages were extremely infrequent during the studied period, any unnoticed recombination could violate model assumptions of a tree-like evolution and affect substitution counts.

Third, generalizability of the findings warrants caution. The viral evolution analysis was specific to SARS-CoV-2 during a pandemic scenario; RNA viruses with different life histories or mutation rates (e.g., measles virus, or a DNA

virus like herpesvirus) might not exhibit the same diffusion characteristics. The anomalous diffusion observed might partly reflect the unique immune and epidemiological dynamics of COVID-19 (such as rapid global spread and strong variant sweeps). Applying our model to other pathogens may require adjustments, and one should be careful not to over-interpret the quantitative values (like the exact diffusion exponents) beyond this system.

Finally, there are limitations in the analytical methods themselves. Estimating evolutionary variance over time required binning sequences by sampling date, an approach that smooths over some detail; different bin sizes or outlier removal criteria could slightly change the measured degree of diffusion anomaly. We chose parameters based on robustness checks, but these choices involve trade-offs (e.g., temporal resolution vs. noise in variance estimates).

Future research could explore these concepts across different viral families and more complex organisms. The anomalous diffusion framework could refine molecular epidemiological tools by identifying when a population transitions from one variant-dominated phase to another [5]. Integrating temporal, spatial, and phenotypic data would improve our grasp of how adaptive lineages outcompete others under shifting selection pressures. Such efforts stand to enhance our predictive power, whether in forecasting viral spread or in engineering robust gene circuits. By tackling these open questions, this work's contributions can guide ongoing inquiries into the stochastic underpinnings of molecular evolution and improve both theoretical models and practical applications in biology.

References

[1] Kimura M (1968) Evolutionary rate at the molecular level. *Nature*, 217(5129): 624–626.

[2] Zuckerkandl E, Pauling L (1965) Evolutionary divergence and convergence in proteins. *Evolving Genes and Proteins*: 97–166.

[3] Elowitz MB, Levine AJ, Siggia ED, Swain PS (2002) Stochastic gene expression in a single cell. *Science*, 297(5584): 1183–1186.

[4] Dolcemascolo R, Goiriz L, Montagud-Martínez R, Rodrigo G (2022) Gene regulation by a protein translation factor at the single-cell level. *PLoS Computational Biology*, 18(5): e1010087.

[5] Ho SYW, *et al.* (2011) Time-dependent rates of molecular evolution. *Molecular Ecology*, 20(15): 3087–3101.

[6] Drummond AJ, Ho SYW, Phillips MJ, Rambaut A (2006) Relaxed phylogenetics and dating with confidence. *PLoS Biology*, 4(5): e88.

[7] Weiss M, Elsner M, Kartberg F, Nilsson T (2004) Anomalous subdiffusion is a measure for cytoplasmic crowding in living cells. *Biophysical Journal*, 87(5): 3518–3524.

[8] Höfling F, Franosch T (2013) Anomalous transport in the crowded world of biological cells. *Reports on Progress in Physics*, 76(4): 046602.

[9] Paulsson J (2004) Summing up the noise in gene networks. *Nature*, 427(6973): 415–418.

[10] Holmes EC, Pybus OG, Harvey PH (1999) The molecular population dynamics of HIV-1. *The Evolution of HIV*: 177–207.

Chapter 7

General Conclusions

This PhD dissertation was intended to expand our knowledge on the evolutionary dynamics of rapidly mutating biological agents, with focus on SARS-CoV-2 as a model system for studying molecular evolution under natural conditions. By integrating mathematical modeling, stochastic processes, and large-scale genomic data analysis, we aimed to challenge traditional assumptions of the molecular clock and propose a more flexible, variant-dependent framework. The work further contributes to the development of more realistic models of viral evolution, offering insights into mutation rates, selective pressures, and diffusion patterns in genomic space. In particular, this thesis has reached the following main conclusions:

1. RNA Virus evolution can follow variant-dependent, non-Brownian dynamics best described by a stochastic model incorporating fractional Brownian noise, where anomalous diffusion exponents quantify deviations from the classical molecular clock through subdiffusive or superdiffusive mutation rates.

2. Variant-specific analyses reveal distinct mutation accumulation trajectories characterized by bursts in genetic variance during emergence and replacement events, yet converging toward asymptotic Poissonian behavior as captured by a reset-based dispersion index.

3. Mutation accumulation in RNA viruses exhibits strong gene-level heterogeneity, deviating from the uniform patterns predicted by classical molecular clock models.

4. Genes under functional constraint exhibit reduced variance and dispersion indicative of purifying selection, genes involved in host interaction undergo episodic adaptive bursts, marked by high dispersion indices and non-monotonic mutation dynamics.

5. A top-down statistical analysis enables detection of localized evolutionary dynamics, revealing how gene-specific contributions shape genome-scale anomalous diffusion.

6. Non-phylogenetic, population-level approaches offer scalable alternatives for modeling large viral datasets, underscoring the importance of implementing and distributing these novel models as open-source software to ensure broad accessibility and reproducibility.

Acknowledgements

This thesis is the result of a long journey filled with constant changes and, ultimately, growth. Starting in late 2020 during a time of social isolation and uncertainty, and concluding in a period of hope, reconnection, and new beginnings, this journey would not have been the same without the support of many people. Some of you I stumbled upon by chance, while others I have known for a long time.

First and foremost, I would like to thank my supervisor, Dr. Guillermo Rodrigo, for his guidance, support, and patience (which I have put to the test on more than one occasion). Since the beginning of my academic career during my Bachelor's thesis in 2018, you have not only been a constant source of inspiration as a scientist but also a mentor and a friend. I would consider myself very lucky if I ever become a fraction of the scientist and person you are.

Secondly, I would like to thank J. Alberto Conejero for his constant support within the university. Although your availability has decreased over the years, I am grateful that you have always found time to help me with my doubts and concerns, even if it meant having phone calls during our commutes home.

I would like to thank my lab mates, with whom I have shared countless hours of work, laughter, and frustration. Special gratitude goes to those of you I have had the pleasure to work with more closely, such as Roser, Raúl, Roswitha, and María. This thesis is as much yours as it is mine.

I would like to thank my Investmat mentors and friends, Antoni and Christian, who adopted this *biotechnologist* as a fellow mathematician and who have given me all the mathematical foundations I needed to survive Investmat, which ultimately led me to this thesis. Perhaps next time there is a conference in Cullera, I can teach you a thing or two about elliptic curves.

Special thanks to Nicolás, who, despite the distance, has always been there to share stories about our misadventures, even when those calls had to happen in the middle of the night. Valencia eagerly awaits your return.

I would also like to thank my Biotech friends, both those who stayed in Valencia - Adrián, Ramón, and Arcadio - for their support and for always being there to hang out every second Thursday or so, even if it meant having

to listen to me talk about mathematics, and those who unfortunately did not stay - Jorge and Iñigo. Despite the physical distance between some of us now, I am glad that our bonds remain as strong as they were during our time together from 2015 to 2019 when all six of us shared countless memories.

I would like to thank my friends from L'Alfàs for allowing me to escape from the city and the university every now and then, even if some of you have already left to pursue your dreams. Thank you, Roberto, Adrián, David, Álvaro, and the rest of the gang.

I would also like to thank the Secureum, Spearbit, and webtrES communities for accepting me as one of their own and for enabling me to learn the intricacies of computer science, cybersecurity, and software development. Your friendship and shared experiences have profoundly shaped this thesis. Special thanks to Rajeev from Secureum, who has been a constant source of inspiration, and to my Spanish web3 colleagues Pablo, Eugenio, and Víctor, with whom I spent countless nights hacking, building, and breaking protocols, interspersed with dancing adventures around the world. I am also deeply grateful to the Spearbit/Cantina team, especially Hari, Alex, and Noah, who not only gave me the opportunity to contribute to the core team but also, unknowingly, mentored me in various soft skills.

Finally, to my parents, who were calling me "Dr." right after I started my PhD journey. Maybe academia wasn't that bad after all, was it?