

Imperial College London

Modelling the Diversity and Persistence of the Human T-Lymphotropic Virus Type-1

Daniel J. Laydon

July 14, 2015

Supervised by Becca Asquith and Charles R. M. Bangham

Submitted in part fulfilment of the requirements for the degree of
Doctor of Philosophy in Mathematical Biology of Imperial College London
and the Diploma of Imperial College London

Copyright Declaration

The copyright of this thesis rests with the author and is made available under a Creative Commons Attribution Non-Commercial No Derivatives licence. Researchers are free to copy, distribute or transmit the thesis on the condition that they attribute it, that they do not use it for commercial purposes and that they do not alter, transform or build upon it. For any reuse or redistribution, researchers must make clear to others the licence terms of this work.

Acknowledgements

I want to express my deepest gratitude to my supervisors Becca Asquith and Charles Bangham, who have sat like patience on a monument waiting for this thesis. I hope that I have not given them cause to regret bestowing a shot at academia upon a glorified bartender with specialised knowledge of only the grape and the grain. They are wonderful teachers, and ideally these pages will reflect a fraction of all they have taught me.

The Department of Immunology has been a wonderful place to work, to learn, and to meet lovely (and interesting) people, all of whom have helped me in one way or another. Anat Melamed deserves particular recognition for her generosity in teaching me sub-basic biology while successfully concealing her incredulity. Viruses are indeed awesome.

I also want to extend my thanks to Vikram Sunkara with whom I enjoyed working and learning immensely. I am in his debt.

I would like to thank The Wellcome Trust and the Medical Research Council UK for funding this work, as well as the hundreds of patients who provided data without which the work would be impossible.

I could not have asked for a kinder, more thoughtful, caring and fun companion than my partner, Agata Duda, whose support has never been less than total during these difficult past years. Word limits prevent me from saying as much as I would like on the subject, and detailed discussion is beyond the scope of the present text.

I would like to dedicate this work to my Mum, the family rock, and to my Dad, possibly the only person outside of mathematics and biology who would have been interested in this thesis. I'm sorry it's late.

Declaration of Originality

All the work presented in this thesis is the author's own unless clearly stated in the statement of collaboration and the relevant sections of the thesis.

Signature

Date

Statement of Collaboration

This work would not have been possible without the work of many others. All of the data on which this thesis depends was obtained by collaborators, and generously provided to me.

All HTLV-1 patient data was collected by the National Centre for Human Retrovirology, St. Mary's Hospital, London, except for the leukaemic patient data which was collected from the Institute for Viral Research, Kyoto University, Japan, and the patient data from asymptomatic carriers of HTLV-1 which was collected in Kumamoto and Kagoshima, Japan. The high-throughput protocol used to generate each HTLV-1 data set was developed by Charles Bangham, Nicolas Gillet, and Anat Melamed. The protocol was performed by Anat Melamed, Nicolas Gillet, Lucy Cook, and Heather Niederer. DNA sequencing was performed by Niall Gormley and his team at Illumina. This data is analysed in Chapters 2, 3, and 5.

Kathleen Sim and J. Simon Kroll of the Department for Paediatrics, Imperial College London provided the 16S *Bifidobacteria* rRNA sequence data. The TCR repertoire data was provided by Sam Darko, Daniel Douek, and David Price of the Vaccine Research Center at the National Institutes of Health in Bethesda, Maryland and of the Institute of Infection and Immunity, Cardiff University School of Medicine. Both of these data sets are analysed in Chapter 2.

Aaron Sim of the Centre for Integrative Systems Biology and Bioinformatics at Imperial College London adapted code for the DivE estimator to make it available as an R package, and contributed substantially to the associated documentation. He further provided valuable input on the DivE estimator.

We are indebted to Vikram Sunkara of the University of Adelaide and the Karlsruhe Institute of Technology for his immeasurable help on the mathematics of stochastic processes. Vikram provided frequent advice, input and expertise on stochastic birth-death processes, and on the hybrid model of within-host HTLV-1 persistence, which are the respective contents of Chapters 4 and 5.

Aileen Rowan provided many helpful insights on features of HTLV-1 biology that are indispensable in mathematical models. Vincent Jansens at Royal Holloway, University of London provided guidance and advice on the stochastic behaviour of biological systems. Lies Boelen, Marjet Elemans and Anat Melamed provided frequent advice on all projects covered in this thesis.

Finally, all of the work presented has been a constant collaboration between myself and my supervisors Becca Asquith and Charles Bangham.

Dissemination

Publications

1. **Laydon D.J.**, Melamed A., Sim A., Gillet N.A., Sim K., Darko S., Kroll, J.S., Douek D.C., Price D.A., Bangham C.R.M. and Asquith B. (2014) *Quantification of HTLV-1 clonality and TCR diversity*. PLOS Comput. Biol. Jun 19;10(6):e1003646
2. Niederer H.A., **Laydon D.J.**, Melamed A., Elemans M., Asquith B., Matsuoka M., Bangham C.R.M. (2014) *HTLV-1 proviral integration sites differ between asymptomatic carriers and patients with HAM/TSP*. J Virol. 11:172.
3. Cook L.B., Melamed A., Niederer H.A., Valganon M., **Laydon D.J.**, Foroni L., Taylor G.P., Matsuoka M., Bangham C.R.M. (2014) *High-throughput analysis of HTLV-1 clonality in adult T-cell leukaemia/lymphoma*. Blood. Jun 19;123(25):3925-31
4. Melamed A., Witkover A.D., **Laydon D.J.**, Brown R., Ladell K., Miners K., Rowan A.G., Gormley N., Price D.A., Taylor G.P., Murphy E.L., Bangham C.R.M. (2014) *Clonality of HTLV-2 in natural infection*. PLOS Pathog. 10(3):e1004006
5. Gillet N.A., Cook L.B., **Laydon D.J.**, Hlela C., Verdonck K., Alvarez C., Gotuzzo E., Clark D., Farr L., Bittencourt A., Asquith B., Taylor G.P., Bangham C.R.M. (2013) *Strongyloidiasis and infective dermatitis alter human T lymphotropic virus-1 clonality in vivo*. PLOS Pathog. 9(4):e1003263.
6. Melamed A., **Laydon D.J.**, Gillet N.A., Tanaka Y., Taylor G.P., Bangham C.R. (2013) *Genome-wide determinants of proviral targeting, clonal abundance and expression in natural HTLV-1 infection*. PLOS Pathog. 9(3):e1003271.
7. Hodson A., **Laydon D.J.**, Bain B.J., Fields P.A., Taylor G.P. (2013) *Pre-morbid human T-lymphotropic virus type I proviral load, rather than percentage of abnormal lymphocytes, is associated with an increased risk of aggressive adult T-cell leukemia/lymphoma*. Haematologica. 98(3):385-8.

8. Taylor G.P., Cook L.B., Demontis M.A., Melamed A., **Laydon D.J.**, Bangham C.R.M. (2014) *Rapid dissemination of human T-lymphotropic virus type 1 during primary infection in transplant recipients*. [submitted]

Presentations

1. Oral Presentation: 16th International Conference on Human Retrovirology HTLV and Related Viruses. Montreal, Canada (June 2013)
2. Poster Presentation: (Travel Award): Viruses, Genes and Cancer Conference. Venice, Italy (October 2012)
3. Poster Presentation: Mathematical Immunology Affinity Group Meeting. Microsoft Research, Cambridge, UK (June 2012)
4. Oral Presentation: HTLV-1 European Research Network (HERN). Dundee, UK (May 2012)
5. Oral Presentation: Wright-Fleming Symposium. Imperial College London (March 2012)

Abstract

Human T-lymphotropic virus type-1 (HTLV-1) infects approximately 10-20 million people worldwide. The virus persists within hosts via *de novo* infection and infected cell proliferation, creating a population structure of multiple clones (infected cell populations with identical genomic proviral integration sites). The number of clones in one host is unknown, and is determined by the rate of *de novo* infection.

Our primary objectives are: i) to estimate HTLV-1 clonal diversity; and ii) to develop a model of HTLV-1 dynamics that can estimate the relative contributions of *de novo* infection and mitotic replication. We use a combination of mathematical modelling, computer simulation and statistical methods to interpret experimental observation.

We develop an estimator (named DivE) to estimate the number of HTLV-1 clones. DivE uses the ecological method of rarefaction, and includes novel model selection criteria. We show that DivE is more accurate than widely-used estimators from population ecology, and we demonstrated that this holds across many systems. Differences between these systems and ecological populations are investigated, and DivE is applied to patients with a range of HTLV-associated diseases.

HTLV-1 clonal abundance varies by several orders of magnitude: quantifying within-host HTLV-1 dynamics requires mathematical modelling at multiple scales. Stochastic processes, important for modelling small populations, are introduced, and we explore properties and approximations of a mass-action birth-death process for biologically realistic species extinction scenarios. We combine ordinary differential equations with stochastic processes in a hybrid model and explore its consequences.

The estimated HTLV-1 clonal diversity is substantially higher than previously thought, which strongly implies higher rates of *de novo* infection. The hybrid model captures known behaviour of HTLV-1, and can be used to infer rates of viral persistence. DivE and the hybrid model are applicable to other biological systems, in particular the study of T and B cell receptor repertoires.

Contents

Copyright Declaration	2
Acknowledgements	3
Declaration of Originality	4
Statement of Collaboration	5
Dissemination	7
Abstract	9
List of Abbreviations	14
List of Figures	18
List of Tables	19
1. Introduction	20
1.1. HTLV-1 Background	20
1.1.1. Epidemiology	20
1.1.2. Associated Diseases	20
1.1.3. Within-Host Persistence	21
1.1.4. Virology and Immunology	23
1.1.5. Preferences of Viral Integration	24
1.1.6. Integration Site Mapping and Quantification Protocol	25
1.1.7. Outline of Thesis	26
2. Quantification of HTLV-1 Clonality and TCR Diversity	29
2.1. Introduction	29
2.2. Materials and Methods	31
2.2.1. Ethics Statement	31
2.2.2. HTLV-1 Data Collection	31
2.2.3. Microbial Data Collection	32
2.2.4. TCR Data Collection	32

2.2.5.	Prochlorococcus Data Collection	32
2.2.6.	DivE Species Richness Estimator	33
2.2.7.	Choice of DivE Candidate Models	35
2.2.8.	Fitting of Models to Data	35
2.2.9.	Estimation of HTLV-1 Population Sizes	36
2.2.10.	Study Design	36
2.2.11.	DivE Frequency Distribution Generation Algorithm	37
2.3.	Results	40
2.3.1.	Comparison of Estimators: Relationship Between Sample Size and Estimated Diversity	40
2.3.2.	Comparison of DivE and Second Order Bias-Corrected Akaike In- formation Criterion (AIC_c)	41
2.3.3.	Comparison of Estimators: Accuracy of Diversity Estimate	45
2.3.4.	Estimate Error as a Function of Data Curvature	47
2.3.5.	Example Application: Estimated Number of HTLV-1 Infected Cell Clones	50
2.3.6.	DivE Uncertainty	52
2.3.7.	Distribution Generation Algorithm	54
2.4.	Discussion	55
3.	Applications of DivE	58
3.1.	Introduction	58
3.2.	Clonal Diversity in Adult T Cell Leukaemia / Lymphoma	58
3.3.	Clonal Diversity in Asymptomatic Carriers	60
3.4.	Relationship Between HTLV-1 and HTLV-2 Clonal Diversity Mirrors Rela- tionship Between Clonal Diversity of $CD4^+$ and $CD8^+$ T Cells in HTLV-1 .	63
3.5.	HTLV-1 Clonality in Early Infection	65
3.6.	Comparison of HTLV-1 Clonality in Blood and Skin	66
3.7.	Discussion	67
4.	Stochastic Modelling: Birth-Death Processes	69
4.1.	Introduction	69
4.2.	Master Equation	71
4.2.1.	Example: Classical Model of Within-Host HIV Infection	72
4.2.2.	Relationship Between ODEs and Expected Value of Stochastic System	73
4.3.	Master Equation: Mass-Action Rate Birth-Death Process	74
4.3.1.	Transition Matrix	77
4.3.2.	Initial Probability Distribution	78
4.3.3.	Matrix Exponential	79

4.4.	Properties of Birth-Death Process Without Source Inflow	81
4.4.1.	Evolution of Extinction Probability	81
4.4.2.	Effects of Truncated State Space	83
4.4.3.	Shape of Probability Distribution $\mathbb{P}(X; t)$ for Large Starting Values x_0	83
4.4.4.	Shape of Probability Distribution $\mathbb{P}(X; t)$ for $x_0 = 1$	87
4.5.	Inclusion of Source Inflow Changes $\mathbb{P}(X; t)$	92
4.6.	Discussion	94
5.	Hybrid Model of HTLV-1 persistence	97
5.1.	Introduction	97
5.2.	Hybrid model	99
5.2.1.	Deterministic Model of HTLV-1 Dynamics	99
5.2.2.	Stochastic Model of Within-Host HTLV-1 Dynamics	101
5.2.3.	Clone State Space and Interactions	103
5.2.4.	Hybrid Model Propagation: Strang Splitting	105
5.2.5.	Infectious Spread r_I	106
5.2.6.	Variance and Covariance of Clones	107
5.2.7.	Clone Movement Between Deterministic and Stochastic Compart- ments	109
5.2.8.	Choice of Density Dependency Parameters H and K	110
5.2.9.	Choice of State Space Upper Limit τ and Frequency F	112
5.2.10.	Modelling the Proportion of Expressing Cells	113
5.3.	Parameter Inference	114
5.3.1.	Parameter Inference with Unseen Clones	116
5.3.2.	Hybrid Model Propagation	117
5.4.	Preliminary Results	117
5.4.1.	System Death with Zero Proliferation	117
5.4.2.	Effect of Proviral Expression on Clone	118
5.4.3.	“Equilibrium” in Hybrid Model	118
5.4.4.	Effect of Infectious Spread Parameter r_I	120
5.4.5.	Effect of Time Step Length	121
5.5.	Identifiability of Rate of Infectious Spread: Proof of Principle	122
5.6.	Discussion	125
5.6.1.	Improvements to Hybrid Model	126
6.	Concluding remarks	128
A.	Appendix A: DivE supporting information	153
A.1.	List of DivE Candidate Models	156

B. Appendix B: Derivation of Expectation, Variance, and Covariance Clone ODEs	163
--	------------

List of Abbreviations

1. ACs - Asymptomatic Carriers of HTLV-1
2. ART - Anti-Retroviral therapy
3. AIC - Akaike Information Criterion.
4. AIC_c - Second Order Bias-corrected Akaike Information Criterion.
5. AIDS - Acquired Immune Deficiency Syndrome
6. ATLL - Adult T Cell Leukaemia Lymphoma
7. CME - Chemical Master Equation
8. CTL - Cytotoxic T-Lymphocyte - $CD8^+$ T cell
9. CTMC - Continuous-Time Markov Chain
10. C_p - Rarefaction/DivE Curvature Parameter
11. DivE - Diversity Estimator
12. DNA - Deoxyribonucleic Acid
13. FSP - Finite State Projection Method
14. HAM/TSP - HTLV-1-Associated Myelopathy / Tropical Spastic Paraparesis
15. HAART - Highly Active Anti-Retroviral Therapy
16. HBZ - HTLV-1 Basic Leucine Zipper Factor
17. HCV - Hepatitis C Virus

18. HIV - Human Immunodeficiency Virus
19. HLA - Human Leukocyte Antigen System
20. HTLV-1 - Human T-Lymphotropic Virus Type-1 (also known as the Human Leukaemia Virus Type-1)
21. HTLV-2 - Human T-Lymphotropic Virus Type-2
22. IDH - Infective Dermatitis Associated with HTLV-1
23. KIR - Killer-cell Immunoglobulin-like Receptor
24. KL-Divergence - Kullback Leibler Divergence
25. LM-PCR - Linker-Mediated Polymerase Chain Reaction
26. LTR - Long Terminal Repeat
27. MHC - Major Histocompatibility Complex
28. MTC - Mother-to-Child Transmission
29. NK Cell - Natural Killer Cell
30. OCI - Oligoclonality Index
31. ODE - Ordinary Differential Equation
32. OTU - Operational Taxonomic Unit
33. PBMC - Peripheral Blood Mononuclear Cell
34. PMF - Probability Mass Function
35. PDF - Probability Distribution Function
36. PCR - Polymerase Chain Reaction
37. PVL - Proviral Load
38. RACE - 5' Rapid Amplification of cDNA Ends

- 39. SDE - Stochastic Differential Equation
- 40. SSA - Stochastic Simulation Algorithm, also known as the Gillespie Algorithm
- 41. StSt - Strongyloides
- 42. TCR - T cell Receptor
- 43. UIS - Unique Integration Site
- 44. VS - Virological Synapse

List of Figures

1.1. HTLV-1 infectious and mitotic spread schematic	22
1.2. HTLV-1 proviral integration site abundance quantification	26
2.1. Outline of DivE Species Richness Estimator	33
2.2. Gini coefficient and Oligoclonality Index (OCI) schematic	38
2.3. Outline of DivE distribution generation algorithm	39
2.4. Comparison of species richness estimators	42
2.5. Comparison of estimators: effect of sample size on estimated diversity. . . .	43
2.6. Estimator bias with sample size not due to small subsamples	44
2.7. Comparison of estimators: accuracy of diversity estimates using TCR data	46
2.8. Ecological estimators underestimate diversity in HTLV-1 infection	48
2.9. Ecological estimators underestimate diversity in HTLV-1 infection	49
2.10. Rarefaction curvature parameter C_p	50
2.11. Test of species richness estimators at different values of curvature parameter C_p using TCR data	51
2.12. Performance of species richness estimators in metagenomic data	52
2.13. Diversity estimates in HTLV-1 infection by estimator	53
2.14. Rarefaction plots from bootstrap samples of HTLV-1, TCR, and Microbial data	53
2.15. Validation of DivE distribution generation algorithm	54
3.1. HTLV-1 proviral load in ACs, ATLL and HAM/TSP patients.	59
3.2. Number of observed clones in ACs, ATLL and HAM/TSP patients.	59
3.3. Relationship between OCI and estimated diversity in ATLL patients.	60
3.4. Comparison of HTLV-1 clonal diversity between ACs and HAM/TSP pa- tients in blood and body.	61
3.5. Relationship between clonal diversity and proviral load in Japanese cohort of ACs and HAM/TSP patients.	62
3.6. Comparison of HTLV-1 and HTLV-2 clonal diversity.	63
3.7. HTLV-1 clonal diversity of infected $CD4^+$ and $CD8^+$ T cells.	64
3.8. DivE estimates in the blood and skin of patients with IDH.	67
4.1. Example non-normal probability distribution function.	70

4.2. Mass action birth-death process flow diagram	75
4.3. Mass action birth-death process flow diagram: truncated state space	77
4.4. Evolution of Extinction Probabilities for range of parameters	82
4.5. Effect of Truncated State Space Ω	84
4.6. Probability distribution evolution: growing species	85
4.7. Probability distribution evolution: shrinking species	85
4.8. Evolution of probability distribution at “equilibrium”	86
4.9. Example PDF evolutions $\mathbb{P}(X; t x_0 = 1)$	88
4.10. The exponential distribution approximates $\mathbb{P}(X; t x_0 = 1)$	89
4.11. Discrepancy between exponential distribution rate parameters $\hat{\lambda}_1$ and $\hat{\lambda}_2$ over time	90
4.12. Kullback-Leibler divergences between $\mathbb{P}(X; t x_0 = 1, \theta_k)$ and exponential distribution approximations	92
4.13. Mass action birth-death process flow diagram with inflow from source (trun- cated state space)	92
4.14. Evolution of Extinction Probabilities with source inflow	93
4.15. Example PDF evolutions $\mathbb{P}(X; t x_0 = 1)$ with source inflow	94
4.16. Kullback-Leibler divergences between $\mathbb{P}(X; t x_0 = 1)$ and normal and ex- ponential distributions with source inflow	95
5.1. Schematic of single clone dynamics	100
5.2. Example clone state space and interactions	104
5.3. Schematic of strang splitting procedure	106
5.4. Schematic of clone promotion and demotion	111
5.5. Proportion of expressing cells over time	115
5.6. Example: expected number of cells and clones over time with zero proliferation	118
5.7. Proviral expression can drive proliferation	119
5.8. Silencing can effect clone death	119
5.9. Example: Number of cells $N(t)$ and clones $S(t)$ at equilibrium, no new infection	120
5.10. Example: Effect of rate of infectious spread parameter r_I	121
5.11. Effect of Time Step Length.	122
5.12. Example Likelihood Function without unseen clones	125

List of Tables

2.1. Comparison of estimates produced by DivE and by weighted, second order Akaike's Information Criterion (AICc)	45
2.2. Comparison of estimator performance for TCR data	46
2.3. Estimator error variation with curvature in TCR data	52
2.4. Performance of DivE frequency distribution generation algorithm	55
3.1. Comparison of clonality of ATLL, HAM/TSP and AC patients	60
4.1. KL-divergence between $\mathbb{P}(X; t \theta_k)$ and reference normal distributions: large x_0	86
A.1. Subsamples used in analysis of relationship between sample size and esti- mated diversity, and in comparison of DivE with AIC_c	153
A.2. DivE species richness estimates for HTLV-1 data	153

1. Introduction

1.1. HTLV-1 Background

1.1.1. Epidemiology

Human T-lymphotropic virus type-1 (HTLV-1) was identified in 1979, and was the first human retrovirus to be discovered [1]. Today it infects an estimated 10 to 20 million people worldwide [2]. The virus has a curious worldwide distribution: though not particularly prevalent, it is endemic in parts of the tropics and subtropics. Southwestern Japan, the Caribbean, sub-saharan Africa and South America are the areas with the highest HTLV-1 prevalence [3, 4], and prevalence increases with age in all of these of regions [4]. HTLV-1 spreads through mother-to-child (MTC) transmission, sexual intercourse and through the use of contaminated blood products [4]. Male-to-female transmission is approximately four times more common than female-to-male transmission [5].

1.1.2. Associated Diseases

While the majority of infected individuals remain lifelong asymptomatic carriers (ACs), the virus is the aetiological agent of the inflammatory disease of the central nervous system HTLV-1-associated myelopathy/tropical spastic paraparesis (HAM/TSP) and Adult T-cell Leukaemia/Lymphoma (ATLL) [6], an aggressive malignancy of mature T-lymphocytes. Symptoms of HAM/TSP can include pain, bladder disturbance, slow progressive spastic paraparesis [7], and weakness of the lower extremities [8].

ATLL has symptoms in common with non-Hodgkin’s lymphoma. Patients typically present heightened thirst, fever, drowsiness, weight loss, and are prone to opportunistic infections [5]. HTLV-1 infected cells in patients with ATLL are known as “flower cells”, T cells where the nucleus has multiple lobules [9]. Chemotherapy initially reduces viral burden, but is usually only effective for approximately six months [5], and patients are now treated with zidovudine and interferon-alfa [10], which can extend life expectancy by as much as two years.

Collectively, these conditions are estimated to affect approximately 5-10% of infected individuals [11]. The proviral load (PVL) is defined by the number of viral copies expressed as a fraction of the number of peripheral blood mononuclear cells (PBMCs). PVL can vary between patients by over four orders of magnitude, and is a known risk factor for the development of HAM/TSP [12] and ATLL [13].

1.1.3. Within-Host Persistence

HTLV-1 persists within host through two primary pathways: *de novo* infection (*infectious spread*) and proliferation of infected cells [14]. Unlike the human immunodeficiency virus type-1 (HIV-1), HTLV-1 does not infect host cells reliably with cell-free virus (virions) [15]. Instead, infectious spread requires cell-to-cell contact between infected and uninfected CD4⁺ T cells [16], and occurs via the formation of a virological synapse (VS) [17, 18], whereupon the genetic material of the virus is reverse transcribed and integrated into the host genome in a different location than that of the infecting cell. The integrated viral DNA is referred to as the *provirus*.

Mitotic spread is the proliferation of a provirus carrying cell. Because mitotic spread utilizes the host's high-fidelity DNA-based DNA polymerase, which is substantially less error-prone than reverse transcriptase, mitotic replication results in relatively little viral genetic sequence variation. With respect to the genomic site of proviral integration (the integration site (IS)), a mitotically replicating cell and its offspring are identical. We refer to a collection of cells with the same IS as a *clone* (Figure 1.1).

HTLV-1 has substantially less genetic sequence variation when compared to other viruses such as HIV-1 [14, 15]. The estimated reverse transcriptase error rate of HTLV-1, while lower than that of HIV-1, is not low enough to fully explain the differences in observed sequence variation [19]. There is therefore a widespread view that mitotic spread accounts for the majority of HTLV-1 proviral load [20], and that infectious spread is the predominant mode of persistence during initial infection, before the host has mounted an effective immune response [17, 20, 21, 22]. Given that anti-retroviral therapy (ART) is currently used as first-line treatment for HTLV-1 infection, and given further that this therapy does not address mitotic spread, determining the ratio of infectious to mitotic spread and investigating whether this value changes either between patients or over the course of infection has significant implications for HTLV-1 treatment.

Because infectious spread integrates the provirus within a new location of the host genome, infectious spread creates new clones. Therefore, an estimate of the number of clones is paramount in quantifying the rate of infectious spread. Estimating clonal diversity has

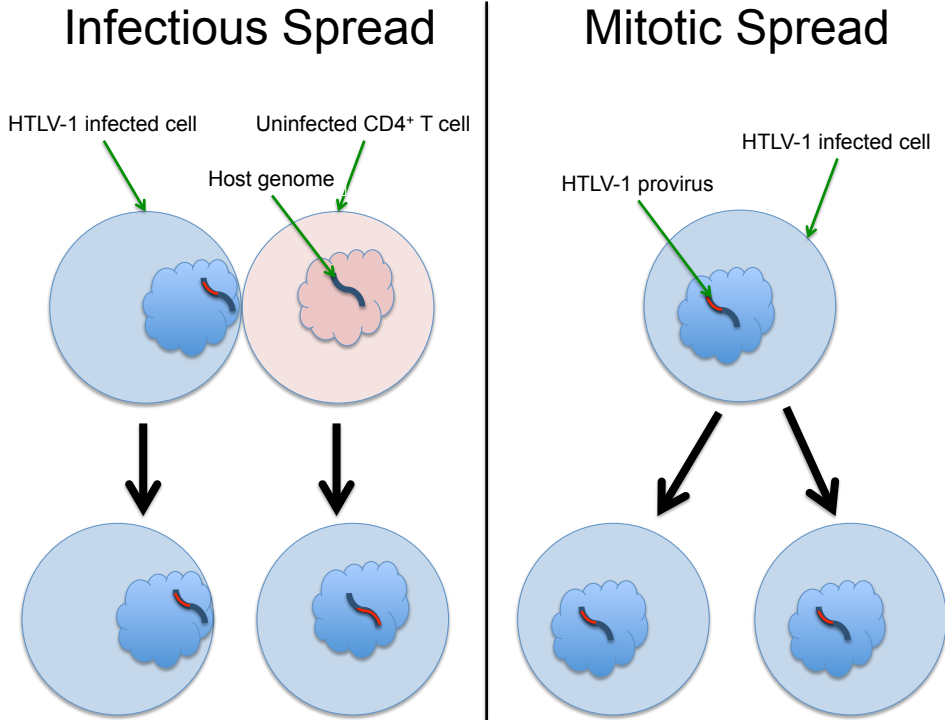


Figure 1.1.: **HTLV-1 infectious and mitotic spread schematic.** Infectious spread (left column): An HTLV-1 infected cell contacts and infects an uninfected CD4⁺ T cell. The HTLV-1 provirus (red) integrates in a different genomic location of the newly infected cell. Mitotic spread (right column): An HTLV-1 infected cell divides, whereupon the provirus resides in the same genomic location in each daughter cell, creating a clone of HTLV-1 infected cells.

been a subject of longstanding interest in HTLV-1, in part due to the belief that the amount of oligoclonal proliferation influences disease status. Estimates have increased over time as experimental techniques have acquired greater precision and sensitivity [20, 22, 23, 24, 25, 26].

In 1992, *Furukawa et al* [23] used a genomic southern blot to measure the clonality of HAM/TSP patients. In each patient, fewer than 10 clones were found from 10 μ g of DNA, all of which were highly expanded. Thus it was inferred that there were of the order of 10¹ clones in the host. In 1995, Eric Wattel's group used a low-throughput linker-mediated polymerase chain reaction (LM-PCR) protocol to measure between 10 and 15 clones from 1-2 μ g of DNA [20, 24], and hypothesized the existence of approximately 10² clones in the host.

In 2011, a high-throughput LM-PCR protocol was developed in Charles Bangham's laboratory (this protocol is summarised in Section 1.1.6). Using this protocol, between 100 and 5000 clones are observed from samples derived from 10 μ g of DNA in a typical AC or

HAM/TSP patient. While ATLL patients have a substantially higher proviral load and far fewer clones, even here there remains wide variation in observed clonal diversity, ranging from a single malignantly infected clone that occupies close to 100% of a patient’s PVL, to as many as 300 clones. It is important to emphasize that HTLV-1 patient data sets, even when derived using a protocol that can reliably measure two orders of magnitude more clones than previous approaches, still only catalogue approximately $\frac{1}{10^8}$ of all infected cells in the body. Further, a substantial proportion of these data sets is occupied by singletons (clones that are observed only once). Therefore current sampling depth does not allow for a complete census, and so the number of clones in one host remains unknown, and must be estimated through theoretical means.

The low sampling depth and high number of low occupancy species encountered in HTLV-1 data are also found in other microbiological and immunological data. T cell data sets often contain several *clonotypes* that are observed only once. Data on the human microbiome have a high number of low frequency operational taxonomic units (OTUs). Another feature in common with HTLV-1 data is the presence of highly expanded clonotypes [27] and high frequency OTUs [28].

It is therefore natural to investigate whether methods to estimate the clonal diversity of HTLV-1 have a wider application in microbiology and immunology, particularly as the number of “classes” is an important quantity in each of these disciplines. In B and T cell immunology, the number of “clonotypes” is associated with the host’s ability to effectively respond to pathogens [27, 29, 30, 31, 32, 33], and studies of the human microbiome have shown an association between bacterial diversity and disease [34, 35].

1.1.4. Virology and Immunology

HTLV-1 is a delta retrovirus with an RNA genome that is approximately 10kb long. Like HIV-1, HTLV-1 encodes for gag, env and pol genes [15]. Though it was previously believed that the virus experiences a long latent period [36], it is now recognized that HTLV-1 establishes a persistent infection whereby the immune system is engaged in a continuous battle with the virus [37]. There is substantial evidence of a strong, sustained, and chronically activated cytotoxic T-lymphocyte (CTL) response to infection [14, 15, 38], in particular a strong positive correlation between PVL and the abundance of HTLV-1-specific CTLs [38]. The CTL response causes PVL to fluctuate and eventually reach a stable equilibrium [14, 15, 39, 40].

Two genes are believed to strongly influence the outcome of infection of HTLV-1 patients. The regulatory protein Tax drives proliferation of infected cells [14], and its expression is

associated with an abundance of HTLV-1-specific CTLs [41]. Spontaneous Tax expression is observed in overnight cell culture *in vitro* [42], and the presence of Tax-specific CTLs suggests frequent expression *in vivo*, although Tax is not usually detected immediately *ex vivo*. While the Tax protein is the immunodominant antigen [5, 41], it is the host response to the protein HTLV-1-basic leucine zipper factor (HBZ) that has the greatest influence on an individual’s risk of developing disease [43, 44]. HBZ is encoded on the complementary strand of the HTLV-1 genome, and is continuously expressed *in vivo* [44]. In HAM/TSP patients, the magnitude of HBZ expression in the host is positively correlated with PVL and the severity of disease, but this association does not hold for the Tax expression [44].

The mechanism underlying the heterogeneity between patients in their PVL and levels of HBZ expression can be explained by the host genotype. The human leukocyte antigen (HLA) system is the major histocompatibility complex (MHC) in humans. HLAs present antigen to the immune system, and fall into three classes. HLA-class I molecules transport fragmented proteins (named *peptides*) from within the host cell to the cell surface; HLA-class II molecules present peptides located in the surrounding vicinity the cell; and HLA-class III genes encode part of the complement system [45]. Individuals differ in their HLA genotype, and the alleles that an individual possesses directly contribute to the strength of immune response that can be mounted to a given pathogen. In the case of HTLV-1, certain HLA alleles differ in their ability to present HBZ peptides to the immune response, and this ability predicts the level of a patient’s proviral load. In particular, possession of the alleles HLA-A*02 and HLA-Cw*08 is positively associated with the probability of remaining asymptomatic, whereas possession of the allele HLA-B*54 is associated with the risk of inflammatory disease. Killer cell immunoglobulin-like receptor (KIR) genotype has been shown to amplify the effects of HLA genotype. Possession of the KIR2DL2 gene enhances both the protective and detrimental HLA class I-restricted anti-viral T cell responses, and predicts final disease status [46]. Further, it is believed that KIR2DL2 enhances T cell-mediated immunity, rather than natural killer (NK) cell mediated immunity [46].

HLA genotype and KIR genotype are important determinants of “CTL quality”, which is the notion that is is not merely the magnitude of the immune response that dictates response to infection, but rather how effective a given cytotoxic T-lymphocyte is in its ability to recognize and destroy a target cell [14, 15, 39, 47, 48].

1.1.5. Preferences of Viral Integration

The proviral integration site for HTLV-1 is not random but instead shows a preference for sites in close proximity (between 1kb and 25kb) to genes and transcriptionally active regions of the host genome [22, 25]. Further, the increased frequency of integration sites

close to transcription units observed *in vitro* is not apparent *in vivo*. After long-term infection, only the most-abundant clones show a similar integration bias near transcription. These observations imply two possibilities: either that there exist selection forces *in vivo*, or that there exist specific features of the genome that favour the expansion and survival of infected cells, or both. Additionally, integration of the provirus in gene-dense regions was found to be associated with higher levels of proviral expression and a greater risk of inflammatory disease [25]. More recently, HTLV-1 has been demonstrated to show a preference for integrating near transcription factor and chromatin remodelling binding sites, notably the cell-cycle regulator and tumor suppressor p53, and the gene expression regulator STAT1 [42]. Patients with ATLL show a similar profile of proviral integration site preference, albeit with some differences [11].

1.1.6. Integration Site Mapping and Quantification Protocol

This thesis makes extensive use of data obtained using a high-throughput LM-PCR protocol, originally developed by *Gillet et al* [22]. The protocol maps the integration site of HTLV-1 infected cells, and quantifies the number of cells in the sample that contain a given integration site. The protocol is summarised here.

DNA is extracted from uncultured PBMCs of HTLV-1 infected patients, randomly fragmented by sonication, and ligated to a partly double-stranded DNA linker. The 3' long-terminal repeat (LTR) of the provirus, together with the surrounding host DNA, are amplified by LM-PCR. A molecule generated by this process is referred to as an *amplicon*. Since fragmentation by sonication is random, a particular combination of a given proviral integration site and a given shear site will be unique. This allows us to distinguish between genuine copies of identical cells, and amplicon duplicates produced by PCR. For example, if there are two amplicons that have the same integration and shear sites, it can be concluded that both amplicons have been derived from a single infected cell (Figure 1.2).

Fragmentation by sonication improves upon previous approaches that use restriction enzymes. The nucleotide sequences split by restriction enzymes are biased to particular locations in the host genome. Because PCR amplification preferentially amplifies short PCR products, proviral integration sites that lie in proximity to restriction enzyme fragmentation sites will be over amplified. Thus the use of restriction enzymes will fail to accurately quantify the frequency of a given integration site. In contrast, shearing the DNA via sonication fragments the DNA randomly, thus removing systematic biases towards particular integration sites.

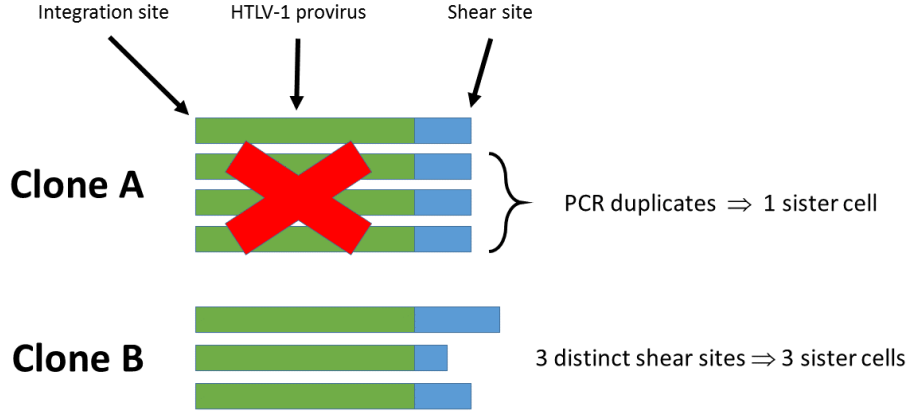


Figure 1.2.: **HTLV-1 proviral integration site abundance quantification.** The combination of a shear site and an integration site allows quantification of HTLV-1 clonal abundance. After PCR amplification, Clone A has 4 amplicons, but only a single shear site; Clone B has three amplicons, but each amplicon has a distinct shear site. Therefore a single sister cell is observed in Clone A, whereas three sister cells are observed in Clone B.

Hence the high-throughput data used derived using the above protocol has three key properties that are essential to estimate and quantify the clonality of HTLV-1. First, PCR duplicates and genuine copies of infected cells can be distinguished. Second, there is no preferential identification of particular integration sites. Finally, the data obtained comprise absolute, and not proportional, counts of infected cells. That is, the frequency of each clone is known, and not merely its *relative abundance* when compared with other clones. Absolute counts are essential in diversity estimation, where the observed frequency of each clone is used to estimate total diversity.

1.1.7. Outline of Thesis

The aim of this thesis is to estimate the clonal diversity of HTLV-1 infection, and to use this information to build a model of the within-host dynamics of HTLV-1 persistence that can infer the ratio of infectious spread to mitotic spread. We interpret empirical observation using statistical methods, computer simulation and mathematical modelling.

Chapter 2

Chapter 2 develops and applies our diversity estimator (named DivE) to HTLV-1 infection. Considering an HTLV-1 clone as a *species* or class of infected cells, estimating the number

of clones in one host is similar to the “unseen species problem”: given the number of species observed in a sample of a population, how many species are there in the total population?

DivE is based on the technique of rarefaction, originally developed in ecology, where the expected number of species is plotted against the number of individuals. Many mathematical models are fitted to multiple subsets of rarefaction data, and we apply novel criteria to assess model performance. We compare estimates of species richness produced with these criteria to those produced using Akaike’s Information Criterion (AIC), an information theoretical method for selection of models. Using observed data and a maximum likelihood modelling approach, we outline an algorithm to estimate the frequency of each species in the wider population, including unseen species.

We investigate the applicability and accuracy of our methods when applied to a range of biological systems where the number of classes of individual is an important, but unknown, population metric. Our methods were applied to virological (HTLV-1 clonality), microbiological (gastrointestinal microbiota), immunological (T cell receptor (TCR) repertoires), and metagenomic (clades of *Prochlorococci* genes) systems. The accuracy of our estimates in each system was compared to that of widely-used non-parametric estimators originally developed in population ecology. We address differences between ecological populations and immunological and microbiological populations, both in the structure of the populations themselves and in collecting data from them.

Chapter 3

Chapter 3 applies DivE to estimate the clonal diversity in a range of patient groups with different manifestation in HTLV-1 associated disease.

We consider patients with clones that have undergone malignant transformation and have progressed to ATLL. Asymptomatic carriers and HAM/TSP patients from two Japanese cohorts are examined, and we compare the clonality of the circulation and the skin in patients with infective dermatitis associated with HTLV-1 (IDH).

We also explore the clonality of human T-lymphotropic virus type-2 (HTLV-2), and contrast this with the clonality of HTLV-1 when restricted to CD4⁺ and CD8⁺ T cells. Finally, we use time series data comprising three immune-suppressed patients who contracted HTLV-1 from contaminated solid organ transplantation. This provides a rare opportunity to examine the clonality of early HTLV-1 infection.

Chapter 4

Chapter 4 introduces techniques in stochastic modelling that we will use to model the dynamics of small HTLV-1 clones. We briefly review existing stochastic methods, originally motivated in chemical kinetics, and which are increasingly applied in mathematical biology. We formally define the *master equation*, and explicitly derive this quantity for a pure birth-death process.

We explore the consequences of a birth-death process that does not contain inflow, and then explore potential approximations that can be made to the probability distribution that describes how a given species obtains a given frequency at a given time, under a variety of parameter values and initial states of the system.

Chapter 5

Here the techniques discussed in Chapter 4 are used to develop a deterministic and stochastic hybrid model of the within-host dynamics of HTLV-1 infection. We outline an idealized model that is not subject to computational runtime and memory constraints. We then discuss approximations that reduce the complexity of the model such that it can be propagated and fitted to empirical data. We describe a maximum likelihood method of parameter estimation that can account for unseen species, and which will ultimately enable inference of the relative contributions of infectious spread and mitotic spread to overall HTLV-1 persistence.

Chapter 6

Concluding remarks and discussion of future work.

2. Quantification of HTLV-1 Clonality and TCR Diversity

2.1. Introduction

How can we estimate the diversity of a population from a sample of that population? In viral infections, the number of viral variants and their population structure inform our understanding of disease pathogenesis, and can suggest treatment strategies [49, 50]. In immunology, the repertoire and population structure of B cell and T cell receptor clonotypes vary with age, and are intimately linked to antimicrobial protective efficacy [30, 32, 33, 51, 52, 53]. In the human microbiome, decreased diversity of the gastrointestinal microbiota is associated with atopy [54], Crohn’s disease and ulcerative colitis [34, 35].

A complete census is usually impossible and so estimators of the number of unseen “species” are required. Here we use the word “species” to denote a class of individuals, such a T cell clonotype, bacterial operational taxonomic unit (OTU) or viral clone. We use the word “individual” to refer to a single T cell sequence read, microbial sequence read, or virus- infected cell. The term “species richness” denotes the number of species in the population under consideration.

Immunological and microbiological data differ in important respects from ecological data. First, in many immunological and microbiological populations, it may be reasonable to assume that species are taxonomically similar, that the spatial distribution of individuals is homogeneous, and that individuals are sampled randomly, independently and with equal probabilities. If made, these simplifying assumptions allow the extrapolation of individual-based rarefaction curves, which depict the expected number of species against the number of individuals sampled [55, 56, 57, 58]. However, the above assumptions are frequently violated in ecological populations [58, 59, 60, 61, 62], where unobserved individuals may differ from observed individuals in their colour, physical size, geographical distribution, movement, variety of habitats and relationship to other species [59], and thus remain unobserved despite substantial subsequent sampling. Second, many common

assumptions about population structure are inappropriate for immunological and microbiological populations, for example that all species have equal frequencies [63, 64, 65], or that the functional form of the population distribution is known [66, 67, 68, 69, 70]. We therefore consider non-parametric estimators.

Non-parametric estimators, such as Chao1 [71], and the abundance-based coverage estimator (ACE) [72], have been proposed. ACE has been suggested to be the best current approach [58, 66, 73] and is widely applied in microbiology and immunology; for example to estimate the diversity of the human gastrointestinal flora [74], human gut metagenome [75], mouse TCR repertoire [76, 77], fungi [78], and the number of HTLV-1 infected cell clones [26]. Although they were originally intended as methods to estimate lower bounds, the Chao1 estimator, and the modified, bias-corrected form Chao1bc [79], have been used to make a point estimate of the number of TCR clonotypes [80, 81], the number of OTUs in hepatitis C virus infection [49], parasite diversity in malaria infection [82] metagenome size [83], the number of integration sites of therapeutic gene therapy vectors [84], soil diversity [85], and again the number of HTLV-1 infected cell clones [22, 26]. In addition to the ACE and the Chao estimator, we also consider two additional non-parametric estimators: the Bootstrap [86] and Good-Turing estimators [87].

Most diversity estimators aim to estimate the species richness in one of two populations of interest: either in the population from which the sample was drawn (e.g. number of microbial species in the gut, given a sample from the gut) or the value where the rarefaction curve saturates (e.g. number of species at the point when further sampling does not yield any new species). These definitions of the population of interest lack flexibility and may be inappropriate or poorly defined for the question in hand. Essentially, given a complete census, either the rarefaction curve will saturate or it will not. If it does not, then an estimate of population size is required to avoid an infinite diversity estimate. Indeed, if some species are represented by a single individual, the rarefaction curve will not saturate. For many microbiological and immunological questions, an estimator that allows the user to specify the size of the population of interest is desirable. For instance, we may wish to know the T cell repertoire diversity of both the blood and the whole body.

The aim of this study was to identify a suitable method for estimating species richness in immunological and microbiological populations. We tested widely-used estimators on samples of microbiological and immunological populations. We found these estimators performed poorly. We therefore developed and validated a new method to estimate species richness and species frequencies.

We used data from three independent sources: (i) viral populations from HTLV-1-infected subjects; (ii) T cell antigen receptor (TCR) clonotype repertoires; and, (iii) infant faecal microbial samples.

Our HTLV-1 data sets consisted of high-throughput data on the abundance of HTLV-1 infected cell clones in 14 HTLV-1 seropositive subjects obtained using the protocol described in [22].

The human gastrointestinal tract contains a densely populated ecosystem of microbes that performs a variety of functions [88]. We obtained high-throughput 16S rRNA sequence data from infant faecal samples. In this study we used observed frequencies of different bacterial operational taxonomic units (OTUs) [28].

T cells are vital to adaptive immunity. The T cell population comprises a diverse repertoire of TCR clonotypes, each defined by the DNA sequence of the expressed TCR. In humans, there are a potential 10^{15} - 10^{20} different TCR clonotypes [89], but the actual number of clonotypes in one person is thought to be far less, and is estimated to be between 10^6 and 10^8 [90, 91]. In this study we used RACE-based data on TCR clonotype abundance. We studied circulating central and effector memory, naïve and total $CD4^+$ and $CD8^+$ T cells.

2.2. Materials and Methods

2.2.1. Ethics Statement

Blood samples were donated by HTLV-1⁺ subjects attending the HTLV-1 clinic at the National Centre for Human Retrovirology (Imperial College Healthcare NHS trust) at St. Mary’s Hospital, London UK, with fully informed written consent. This study was approved by the UK National Research Ethics Service (NRES reference 09/H0606/106). Parents gave full written informed consent for infant faecal sample collection, and all protocols and procedures were approved by the National Research Ethics Service Committee, U.K. (Southampton and South West Hampshire) (ref: 05/Q1702/119). For the TCR data, leukaphereses were performed on healthy donors who provided written informed consent at the National Institutes of Health, USA. The protocol and use of these samples for immunological investigation were approved by the National Institute of Allergy and Infectious Diseases Institutional Review Board.

2.2.2. HTLV-1 Data Collection

Previously reported [22] and new high-throughput data on HTLV-1 clonality provided by our collaborators were analysed. Each HTLV-1 dataset quantifies the abundance of HTLV-1-infected T cell clones. There were 105 datasets, comprising 9 datasets from each

of 10 subjects (three independent samples at each of three time points), and 15 datasets from an additional 4 subjects (Table A.2). All subjects had either HTLV-1-associated myelopathy/tropical spastic paraparesis (HAM/TSP) or were asymptomatic carriers of HTLV-1 (ACs).

2.2.3. Microbial Data Collection

The microbial data was obtained and kindly provided by Kathleen Sim in Simon Kroll’s laboratory at Imperial College. The data were derived from faecal samples obtained from 10 infants. DNA was amplified with two sets of PCR primers, generating 20 datasets [28]. Amplicons of the V3-V5 regions of the 16S rRNA gene were generated by PCR using two sets of universal primers. Sequencing data were generated using the Roche 454 GS Junior platform. Analysis was performed using the QIIME pipeline as described previously [28].

2.2.4. TCR Data Collection

TCR repertoire data was generously provided by our collaborators at the National Institutes of Health in Bethesda Maryland, USA. A total of 16 datasets were collected from two subjects, comprising TCR sequences from four phenotypically defined subsets of CD4⁺ and CD8⁺ T-cells: naïve, central memory (CM), effector memory (EM) and total. After flow cytometric sorting and cell lysis, mRNA was extracted and subjected to a non-nested, template-switch anchored RT-PCR using a 3’ TCRB constant region primer as described previously [92]. This approach allows linear and unbiased amplification of all TCRs irrespective of TRBV or TRBJ gene usage. Paired-end sequencing reactions (each 150bp) were performed using an Illumina HiSeq 2000 sequencer. Raw FASTQ files were annotated using reference TCRB sequences from the ImMunoGeneTics (IMGT) website (<http://www.imgt.org>) and a custom-written Java application. Following annotation, the data were filtered to eliminate potential sequencing and PCR errors.

2.2.5. Prochlorococcus Data Collection

Prochlorococci are vital to energy and nutrient cycling in the oceanic ecosystem, and the genus contains a highly diverse and abundant population of clades. We analysed publicly available metagenomic data describing clades Prochlorococcus. The data were obtained by the Global Ocean Sampling Expedition and contains the frequency of distinct sequence reads of genes of Prochlorococcus clades [93]. Sampling sites, sample collection, library construction, fragment recruitment, and determination of Prochlorococcus abundances are detailed in [93, 94].

2.2.6. DivE Species Richness Estimator

We developed a heuristic approach to estimate species richness, which we named DivE (Diversity Estimator) (Figure 2.1). To calculate the DivE estimator, many models (i.e. functions of the form $y = f(x)$) are fitted to multiple nested subsamples of individual-based rarefaction curves. Each function is fitted to all nested subsamples, and is scored on a set of four criteria. The five best-performing functions are extrapolated and their respective estimates are aggregated to produce the DivE species richness estimate. DivE requires an estimate of population size. If the species richness of a wider population is desired, the same functions are used but extrapolated to a different population size; this is only justified if the two populations are similar in their spatial distribution of individuals.

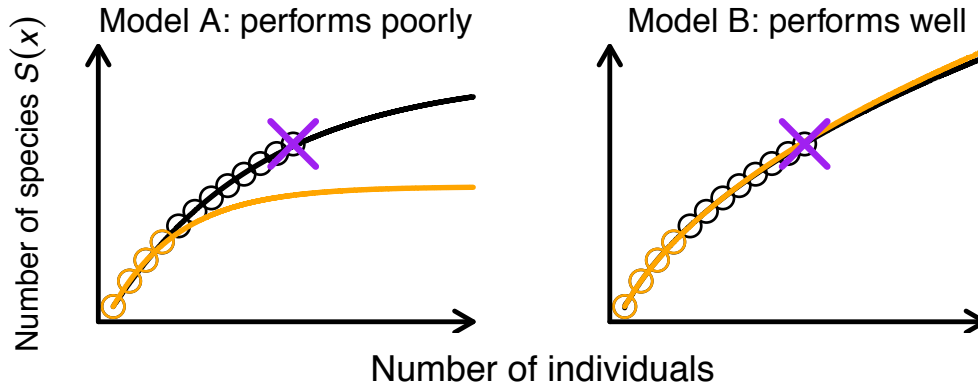


Figure 2.1.: **Outline of DivE Species Richness Estimator.** DivE fits many functions to rarefaction curves (black) and subsamples thereof (orange). Data is denoted by circles; fits by solid lines. Models are scored according to DivE criteria (see Section 2.2.6). The best performing functions are aggregated and extrapolated to estimate species richness. Model A performs poorly as **Accuracy** and **Similarity** criteria are not satisfied. Model B performs well as all criteria are satisfied.

The criteria against which each function fit is scored are:

1. **Discrepancy:** the mean percentage error between data points and prediction of the function.
2. **Accuracy:** the percentage error between the full sample species richness, and the estimate of full sample species richness from a given subsample.
3. **Similarity:** the area between the curve fitted to a subsample and the curve fitted to the full sample, normalized to the area under the curve from the full data, on the interval $[0, N_{obs}]$ where N_{obs} is the size of the full data.

4. **Plausibility:** the predicted number of species must either increase monotonically or plateau and the predicted rate of species accumulation must either decrease or plateau (i.e. for $f(x)$ and $x \geq 1$, where x is the number of individuals, $f'(x) \geq 0$, and $f''(x) \leq 0$).

The rationale behind each criterion is as follows:

1. **Discrepancy:** the function must describe the data to which it was fitted.
2. **Accuracy:** from a subsample, the function should predict the full sample species richness.
3. **Similarity:** an ideal function will produce identical fits from all subsamples. The smaller the area between the function fits, the better the function (Figure 2.1).
4. **Plausibility:** this criterion requires that, as the observed number of individuals increases, the observed number of species does not decrease and the rate of species-accumulation does not increase; the former is biologically impossible and the latter is highly implausible.

Criteria 2., 3. and 4. are independent of the fitting process. That is, they are not constraints by which functions are fitted; instead they are tests of function performance. Note we do not penalise functions for their complexity, for example by using Akaike's Information Criterion (*AIC*). Usually, functions with more parameters can be more easily made to fit data, and thus have a natural advantage that does not correspond to their ability to predict or describe data. However, DivE explicitly tests each function's ability to predict additional data with criteria 2., 3. and 4. and further there is no *a priori* reason why more complicated functions should have an advantage in predicting additional data which played no part in inferring parameter values. In any case, we compare DivE to *AIC* in section 2.3.2.

Each function fit is scored on all four criteria. For criteria 1 to 3, we scored a fit in multiples of empirically chosen precision levels. The precision level for criterion 1 was 0.01%: a score of 1 denotes a function fit where the mean percentage error of the residuals, ϵ , was less than 0.01%; a score of 2 denotes $0.01\% \leq \epsilon \leq 0.02\%$ and so on. Criteria 2 and 3 were similarly scored in multiples of 0.5%. Criterion 4 was implemented by giving a score of 500 to function fits that violated either of its conditions; this value was chosen to exceed the score of any function fit that satisfied this criterion.

The final score for each function is an aggregate of the scores of all function fits across subsamples and criteria, and is calculated as follows. First, the score for each criterion is

defined as the mean of the scores of all subsample fits for that criterion. The final score for each function is the mean of all criteria scores. That is, for a given function, for criterion i (where the criteria are numbered as above) and subsample j , each function has a score D_{ij} . A function’s final score D is given by

$$D = \frac{1}{4} \sum_{i=1}^4 \frac{\sum_{j=1}^n D_{ij}}{n} = \frac{1}{4n} \sum_{i=1}^4 \sum_{j=1}^n D_{ij} \quad (2.1)$$

where n is the number of subsamples (see Table A.1). The DivE species richness estimate is the geometric mean of the estimates provided by the five best-performing, lowest-scoring functions (see section 2.4).

2.2.7. Choice of DivE Candidate Models

A list of 58 candidate functions (see Appendix Section A.1) was chosen from an online repository [95]. Many of these (e.g. logistic, logarithmic, hyperbolic) are widely used in population ecology [55, 96]. Models were fitted by least squares regression using R version 2.14.2 [97] with the package FME [98]. Global fitting was performed using Price’s pseudo-random search algorithm [99] followed by local fitting using the Levenberg-Marquardt algorithm [100].

2.2.8. Fitting of Models to Data

The pseudo-random model fitting algorithm [99] requires that parameter ranges and initial guesses be entered. The runtime incurred in model fitting increases with the size of the parameter space. The need for parameter ranges small enough to yield precise parameter estimates in acceptably short runtimes must be balanced against the need for parameter ranges that adequately encompass appropriate parameter values for data of different scales.

The performance of the “modFit” and “pseudoOptim” functions from the R package “FME”, which are used to estimate model parameters, is sensitive to the choice of initial parameter guesses. Initial guesses were chosen as the midpoint of each parameter range. To obtain better parameter fits, the fitting process was repeated, taking advantage of the fact that data between subsamples (but within patients) is similar. Fitted parameters from a single subsample often provided better initial guesses for a subsequent subsample than the seeds originally given. This iterative process allowed for faster optimization of parameters while keeping parameter ranges relatively large, ultimately improving function fits.

2.2.9. Estimation of HTLV-1 Population Sizes

Population size is a necessary input for species richness estimation when a given subpopulation must be specified, or when a saturating relationship between population size and species richness cannot be assumed. We estimate the number of HTLV-1 infected cells i) in the circulation (N_{blood}), and ii) in the whole body (N_{body}). For the former, we assumed a circulating blood volume of $5L$ and a peripheral blood mononuclear cell (PBMC) count of $3 \times 10^9 L^{-1}$ [45]. HTLV-1 proviral load PVL is defined as the number of viral copies per 100 PBMCs. We assumed that each infected T cell carries a single copy of the HTLV-1 provirus [101]. Thus, we estimated that

$$N_{\text{blood}} = 5 \times 3 \times \frac{PVL}{100} \times 10^9 \quad (2.2)$$

To estimate N_{body} , we assumed that 40% of PBMCs are $CD4^+$ T cells, and that 95% of HTLV-1 $^+$ cells are $CD4^+$ T cells [102]. We therefore multiplied the PVL by $\frac{100}{40}$ to obtain the proportion of infected cells that are $CD4^+$ T cells that are infected, i.e. $PVL_{CD4} = \text{no. copies per 100 } CD4^+ \text{ T cells} = PVL \times \frac{100}{40}$. The total number of lymphocytes in the body is approximately 2×10^{12} , and 50% of lymphocytes are $CD4^+$ T cells. Thus

$$N_{\text{body}} = \frac{PVL}{100} \times \frac{100}{40} \times \frac{1}{0.95} \times 10^{12} \quad (2.3)$$

2.2.10. Study Design

We evaluated DivE and five widely-used non-parametric estimators: the Chao1 bias-corrected estimator (Chao1bc) [79], the abundance-based coverage estimator (ACE) [72], the Bootstrap estimator [86], the Good-Turing estimator [87, 103, 104] and the widely-used negative exponential model [55, 56, 79, 105, 106]. ACE and Chao1 [71], have been suggested as best practice [56, 58, 66, 73, 107] and are widely applied in microbiology and immunology [22, 26, 49, 74, 76, 78, 80, 82, 83, 84, 85]. For ACE, “abundant” species were defined as those with an observed frequency of greater than 10, as recommended in [107].

Due to differences between estimators and between datasets, we conducted multiple, distinct evaluations and validations. We first evaluated, for each estimator, the relationship between estimated diversity and sample size, using the estimates produced from a series of successively smaller, randomly generated *in silico* subsamples of observed data. For the microbial and TCR data respectively, five and six equidistant subsample sizes were chosen from each observed dataset. For the HTLV-1 data, subsample sizes were chosen to be approximately equidistant; however some were removed due to runtime constraints. See

Table A.1 for further details. Second, we measured the accuracy of DivE by comparing the estimated species richness $\hat{S}_{obs}$ at the size of the full dataset N_{obs} from each subsample to the (known) species richness S_{obs} in the full data. Using the same method, we compared DivE to the second order bias-corrected Akaike Information Criterion (AIC_c) [108, 109]. Third, the TCR data have rarefaction curves which plateau. Using smaller subsamples of this data and making the assumption that the species richness of the full data is equal to that of the entire population, we were able to evaluate the accuracies of all estimators together. Finally for 11 of the 14 HTLV-1 patients detailed in Table A.1, three samples were taken at a single time point. For each time point, the three samples were pooled and used as a practical test of DivE’s ability to predict species richness in larger samples.

2.2.11. DivE Frequency Distribution Generation Algorithm

In addition to species richness, we estimated the population frequency distribution. Because of the considerable structural variation between and within immunological and microbiological populations, we developed a general method which does not assume the analytical form of the population structure. This algorithm uses the DivE estimator combined with observed abundances (Figure 2.3). The algorithm was applied to multiple random subsamples of observed data. The estimated distributions were then compared to the full data frequency distribution using two measurements: (i) error, defined as the sum of discrepancies in species frequencies between estimated and observed (full) distributions, divided by the number of individuals in the observed distribution, i.e.

$$\text{error} = \frac{\sum_{i=1}^{\hat{S}_{obs}} |Est_i - Obs_i|}{\sum_{i=1}^{\hat{S}_{obs}} Obs_i} \quad (2.4)$$

and (ii) percentage error between the Gini coefficients of the estimated and observed distributions. The Gini coefficient is an index of dispersion used widely in epidemiology, sociology, biology, and ecology [22, 110] (Figure 2.2).

Consider an observed species frequency distribution where x individuals are dispersed among species $1, \dots, y$, each with observed relative frequencies $p_{(x,1)}, p_{(x,2)}, \dots, p_{(x,y)}$. We define

$$\sum_{i=1}^y p_{(x,i)} = 1 \quad (2.5)$$

That is, all x cells belong to the observed clones $1, \dots, y$, and so 100% of the x cells belong to observed clones.

Now consider the observed frequency distribution’s corresponding species accumulation curve. The final point of this species accumulation data would be (x, y) (Figure 2.3). After

Oligoclonality Index (OCI)/Gini coefficient

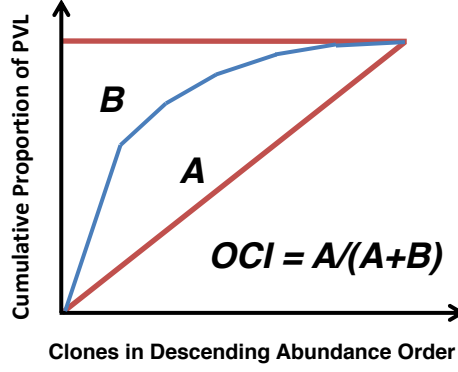


Figure 2.2.: **Gini coefficient and Oligoclonality Index (OCI) schematic.** Blue line shows the cumulative proportion of PVL plotted as a function of the descending rank order of the clone abundance. Red diagonal line (line of equality) denotes that all clones have an equal share of the PVL. The OCI or Gini coefficient is given by $\frac{A}{A+B}$. (Note that the Gini coefficient is often defined with clones in ascending order, in which case the cumulative proportion of PVL will sit below the line of equality.)

fitting the functions returned by DivE to the species accumulation curve, we extrapolate to the point $(x + a, y + 1)$, i.e. until another, unobserved, species is encountered. We determine the proportion of the $x + a$ individuals occupied by the previously observed y species. That is, we wish to calculate

$$p_{max} = \sum_{i=1}^y p_{(x+a,i)} \quad (2.6)$$

Using this quantity, we can calculate the relative frequency of the newly observed species $y + 1$,

$$p_{(x+a,y+1)} = (1 - p_{max}) \quad (2.7)$$

We estimate the most likely number of individuals in the wider population that belong to the species in the observed sample i.e. to species $1, \dots, y$.

Let N be the size of the wider population from which all individuals are drawn *without replacement*, and let r be the number of individuals in the wider population that belong to species $1, \dots, y$. Let d_j be the number of individuals from any of species $1, \dots, y$ drawn on the j^{th} trial. d_j can take values 0 or 1. Then $\mathbb{P}(d_1 = 1) = r/N$, and

$$\mathbb{P}(d_2 = 1 \mid d_1 = 1) = \frac{r-1}{N-1} \quad (2.8)$$

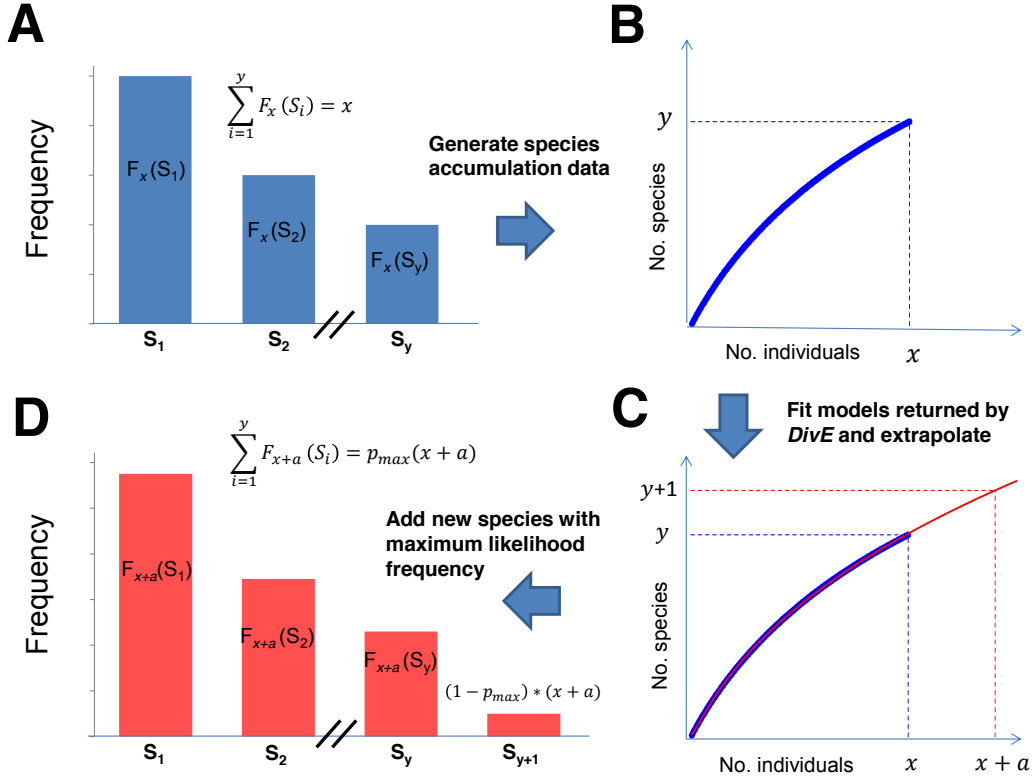


Figure 2.3.: **Outline of DivE distribution generation algorithm.** **A** Truncated species frequency distribution with x individuals distributed among y species. The frequency of species S_i after sampling x individuals is denoted $F_x(S_i)$. **B** Species accumulation data generated from frequency distribution. **C** An aggregate of the best performing functions as returned by DivE is used to extrapolate to point $(x+a, y+1)$, where the next species is predicted. **D** Species S_{y+1} is assigned a frequency of $(1 - p_{max})(x+a)$, where p_{max} is the maximum likelihood proportion of individuals occupied by the y previously observed species. The remaining $p_{max}(x+a)$ individuals are distributed among species $S_1, \dots, S_y$ in proportion to their observed relative frequencies at x . Steps **C** and **D** are repeated until the predicted species richness is reached. See Section 2.2.11 for further details.

More generally,

$$\mathbb{P}(d_j = 1 \mid d_{j-1} = 1 \mid \dots \mid d_1 = 1) = \frac{r - (j - 1)}{N - (j - 1)} \quad (2.9)$$

The likelihood $\mathcal{L}$ of observing $(x + a - 1)$ individuals from species $1, \dots, y$, followed by an individual belonging to a new species is thus

$$\mathcal{L}(r) = \frac{r(r-1)(r-2)\dots(r-(x+a-2))}{N(N-1)(N-2)\dots(N-(x+a-2))} \times \frac{N-r}{N-(x+a-1)} \quad (2.10)$$

$\frac{N-r}{N-(x+a-1)}$ is the probability of drawing an individual that does not belong to species $1, \dots, y$ on the $(x+a)^{th}$ draw (i.e. where $N - (x + a - 1)$ individuals remain) when all previous draws have produced individuals belonging to species $1, \dots, y$. Note we assume that the new species $y + 1$ is observed only once in draws $x + 1, \dots, x + a$. We obtain $\sum_{i=1}^y p_{(x+a,i)}$ by numerically optimizing r_{max} (using the method of Nelder and Mead [111]) such that $\mathcal{L}(r_{max})$ is maximal. We compute $p_{max} = \sum_{i=1}^y p_{(x+a,i)} = \frac{r_{max}}{N}$. Then for $i = 1, \dots, y$,

$$p_{(x+a,i)} = p_{(x,i)} \times p_{max} = p_{(x,i)} \times \frac{r_{max}}{N} \quad (2.11)$$

and

$$p_{(x+a,y+1)} = 1 - p_{max} \quad (2.12)$$

We now have estimates of the frequencies of species $1, \dots, y + 1$ in $x + a$ individuals. We repeat the above process for species $y+2$ (i.e. by extrapolation to the point $(x+a+b, y+2)$) and so on until the predicted number of species in the population is reached (Figure 2.3).

2.3. Results

2.3.1. Comparison of Estimators: Relationship Between Sample Size and Estimated Diversity

Each species richness estimator (Chao1bc [79], Bootstrap [86], ACE [72], Good-Turing [87], the negative-exponential model [55, 56] and DivE) was applied to random subsamples of observed data. We used linear regression to calculate the average proportional increase in estimated diversity as a function of the proportional increase in sample size. Sample size and diversity were normalized respectively to the smallest sample and the estimated diversity at the smallest sample. For example, a “normalized gradient” of 0.5 would mean that, on average, an increase of 10% in sample size would produce a 5% increase in estimated diversity. A value of zero would signify no bias with sample size.

The existing estimators performed poorly when applied to the HTLV-1 and microbial data: estimates systematically increased with sample size. In contrast, DivE produced consistent estimates that showed no obvious relationship with sample size (Figures 2.4 and 2.5). Across subjects and for all methods except DivE, estimates showed significant positive normalized gradients ($p < 0.01$ for every estimator, $n = 14$; two-tailed binomial test) ranging between 0.17 and 0.52 for the HTLV-1 data and 0.3 to 0.45 for the microbial data (Figure 2.5). Conversely, the normalized gradients produced by DivE did not differ significantly from zero ($p = 0.18$, $n = 14$; two-tailed binomial test), and were much smaller (0.0081 and 0.022 for the HTLV-1 and microbial data respectively) (Figure ref:fig:SlopeBeeSwarm). In any specified population there is only one value of species richness, and an accurate estimator will arrive at this value regardless of sample size. An increase in estimate magnitude with sample size implies that estimates of a population's species richness would increase if e.g. greater blood volumes were drawn or technique sensitivity was improved.

The existing estimators were less biased when applied to the TCR data, and estimates were largely consistent. Although the normalized gradients were still significantly positive ($p < 0.0001$ for each estimator except DivE, $n = 16$), their magnitudes were substantially lower than for the HTLV-1 and microbial data. However, existing estimators again increased with sample size for the effector memory (EM) CD8⁺ T cell population from the same subject. These observations can be explained with reference to the TCR rarefaction curves (Figure 2.4). With the exception of the CD8⁺ EM dataset (for which the subsample sizes were considerably smaller), each TCR rarefaction curve reached a plateau, implying that the vast majority of observed clonotypes were encountered early. In contrast, the CD8⁺ EM rarefaction curve did not plateau, suggesting that further sampling would reveal more CD8⁺ EM clonotypes. In common with the microbial and the HTLV-1 datasets, DivE performed well for all TCR datasets, producing consistent results from all subsample sizes. To make sure that the smallest subsamples did not disproportionately contribute to the observed gradients, we repeated the above analysis using only estimates from the largest three subsamples in each patient dataset, which showed almost identical results (Figure 2.6).

2.3.2. Comparison of DivE and Second Order Bias-Corrected Akaike Information Criterion (AIC_c)

The best-performing functions were largely consistent within patients and between subsamples for the microbial and TCR data, although less so for the HTLV-1 data. Ideally, function selection would be consistent across all subsamples. Deviation from this will result

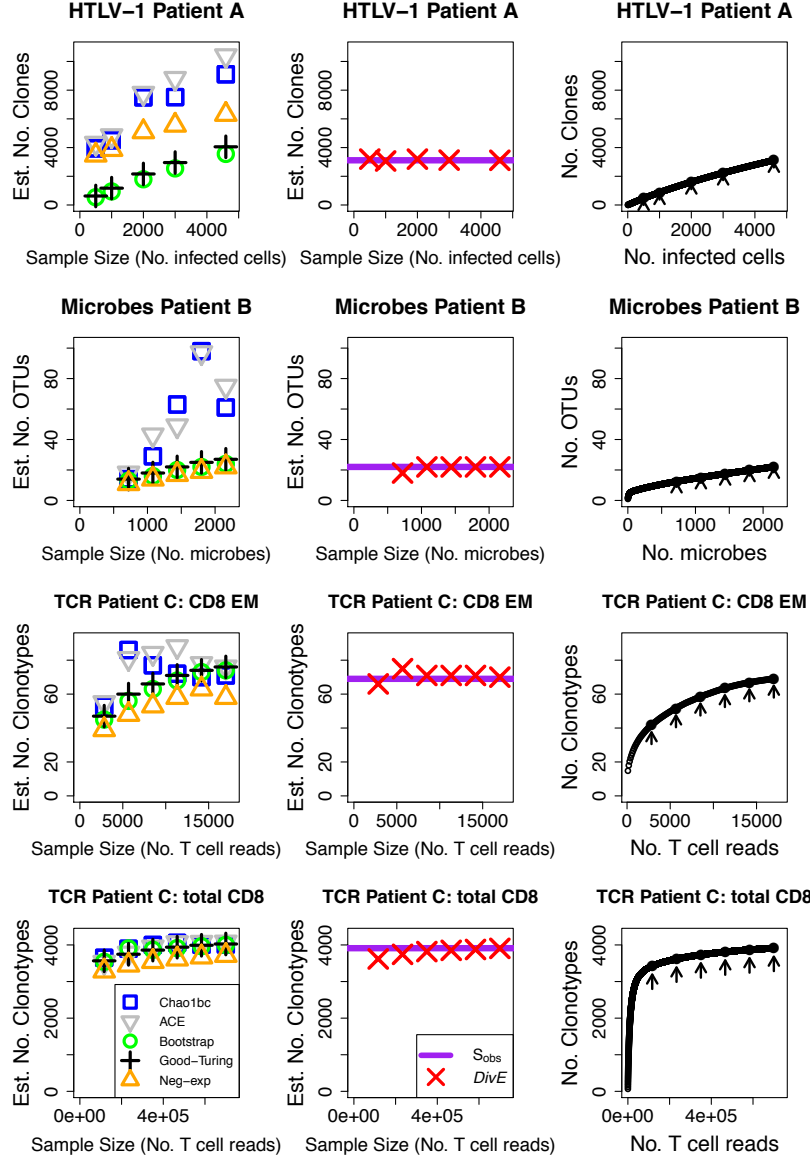


Figure 2.4.: **Comparison of species richness estimators.** **A - D** The Chao1bc (blue), ACE (grey), Bootstrap (green), Good-Turing (black), and negative-exponential estimators (orange) are applied to *in silico* random subsamples of observed data. Examples for HTLV-1, microbial, and TCR data are shown. Estimates systematically increase with sample size in datasets where rarefaction curves do not plateau (e.g. in **I, J, K**). Where rarefaction curves do plateau (e.g. in **L**), estimates are consistent. **E - H** DivE (red) is applied to same subsamples as the other estimators. Performance of DivE was evaluated by comparing the error of estimates ($\hat{S}_{obs}$), to the (known) number of species S_{obs} in the full observed data (purple line), i.e. error = $|S_{obs} - \hat{S}_{obs}|/S_{obs}$. In all datasets, DivE accurately estimates the species richness of the full observed data from subsamples of that data. **I - L** Corresponding HTLV-1, microbial and TCR rarefaction curves: arrows denote the size of the subsample to which each estimator was applied.

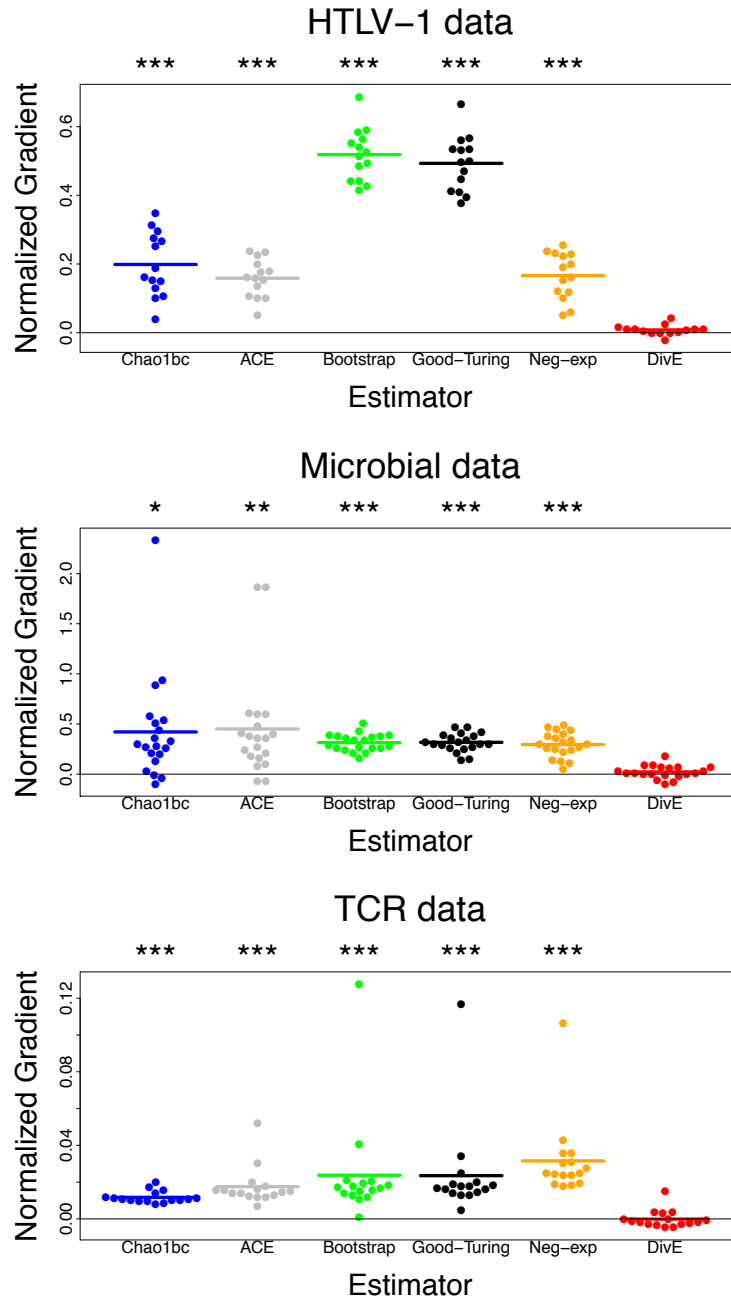


Figure 2.5.: **Comparison of estimators: effect of sample size on estimated diversity.** Normalized gradients measuring proportional increase in estimated diversity against proportional increase in sample size. Normalized gradients (shown for each estimator and each patient data set in Table A.1) were calculated by linear regression. For the HTLV-1 and microbial data, all estimators except DivE show large normalized gradients that are significantly positive. The TCR normalized gradients, though significantly positive, are small and do not show a substantial bias with sample size. *, **, and *** signify $p < 0.01$, $p < 0.001$, and $p < 0.0001$ respectively; two-tailed binomial test ($n = 14, 16, 20$ for the HTLV-1, TCR and microbial data respectively).

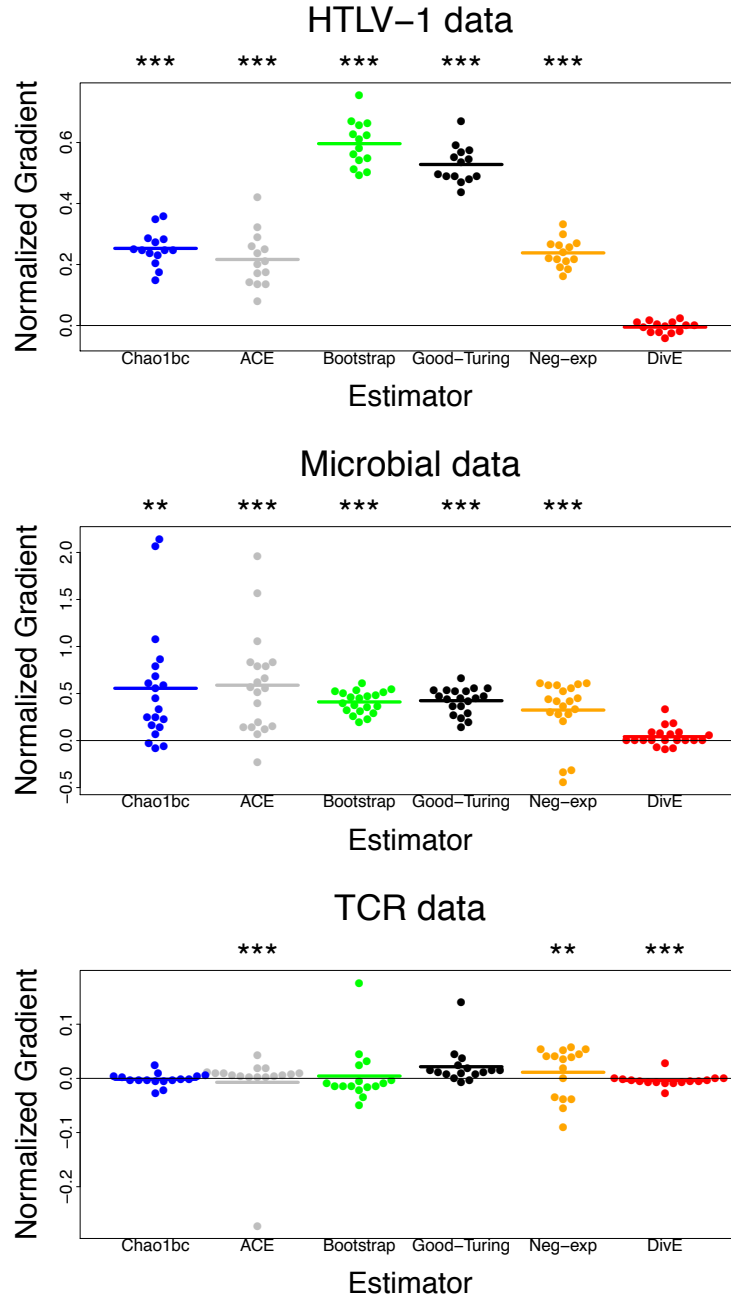


Figure 2.6.: **Estimator bias with sample size not due to small subsamples.** As for Figure 2.5, except that normalized gradients calculated using only largest three subsamples. For the HTLV-1 and microbial data, all estimators except DivE again show large normalized gradients that are significantly positive. The TCR normalized gradients, show no bias with sample size. *, **, and *** signify $p < 0.05$, $p < 0.01$, and $p < 0.001$ respectively; two-tailed binomial test ($n = 14, 16, 20$ for the HTLV-1, TCR and microbial data respectively).

in a discrepancy between S_{obs} and $\hat{S}_{obs}$. This discrepancy is quantified in Figure 2.4 (middle column) and in Table 2.1. To ensure the four criteria provide a useful metric of model performance, we compared DivE to the second order bias-corrected Akaike Information Criterion (AIC_c) [108, 109]. DivE’s mean errors (between the species richness of the full data S_{obs} and $\hat{S}_{obs}$) were 3.3%, 1.0%, and 4.0% for the HTLV-1, TCR and microbial data respectively. These were lower than the corresponding errors of 6.7%, 1.1%, and 7.5%, produced when functions were scored by the AIC_c . This effect was more marked when we considered estimates from small subsamples, defined as those comprising at most 50% of the observed data (Table 2.1). However, the differences between errors were smaller for the TCR data, perhaps also due to the saturating rarefaction curves in these samples.

Table 2.1.: Comparison of estimates produced by DivE and by weighted, second order Akaike’s Information Criterion (AIC_c)

Data source	Method	Small Error ¹	Large sub error ²	Overall ³
Collated	AIC_c	11.3	1.9	5.0
	DivE	5.5	1.4	2.7
HTLV-1	AIC_c	13.1	0.5	6.7
	DivE	6.2	0.5	3.3
TCR	AIC_c	2.4	0.5	1.1
	DivE	1.8	0.6	1.0
Microbial	AIC_c	23.0	3.7	7.5
	DivE	10.4	2.4	4.0

¹ Average percentage error between $\hat{S}_{obs}$ and S_{obs} for small subsamples for each data source. Small subsamples were defined as those $\leq 50\%$ of the size of each patient data set.

² Large subsamples defined as those $> 50\%$ of the size of each patient data set.

³ Average percentage error between $\hat{S}_{obs}$ and S_{obs} across all patient datasets and subsamples for each data source

2.3.3. Comparison of Estimators: Accuracy of Diversity Estimate

When rarefaction curves reach a plateau, we can assume that the value of the plateau is approximately equal to the species richness of the entire population, which the ecological estimators aim to estimate. Thus it is appropriate to evaluate DivE and the existing estimators together using those TCR rarefaction curves which plateau. We took random subsamples of 0.5%, 1%, 2%, 5%, and 10% of the total $CD4^+$ and $CD8^+$ cells for subjects C and E. We then applied each estimator to each subsample and measured its error, defined by

$$\text{error} = \frac{|S_{obs} - \hat{S}_{obs}|}{S_{obs}} \quad (2.13)$$

DivE’s median error was 6.7%, substantially lower than respective median errors of 43.8%, 42.8%, 65.3%, 61.7%, and 50.7% for the Chao1bc, ACE, Bootstrap, Good-Turing and

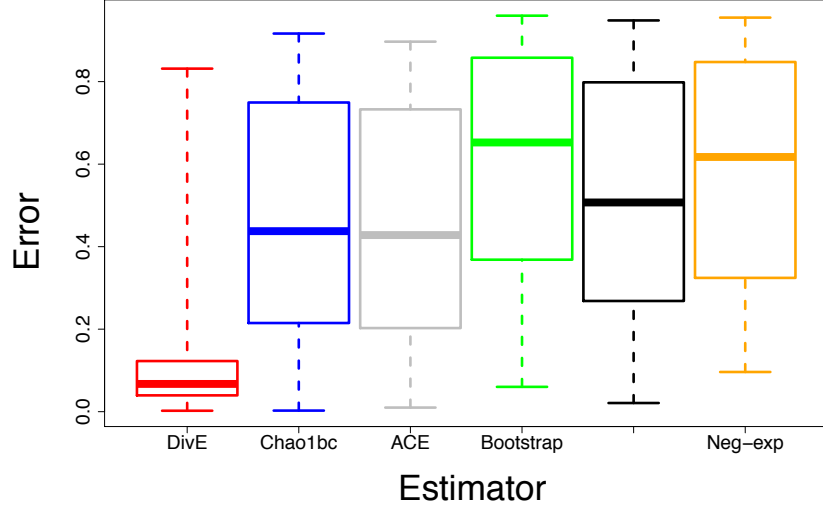


Figure 2.7.: **Comparison of estimators: accuracy of diversity estimates using TCR data.** Random subsamples of 0.5%, 1%, 2%, 5%, and 10% of the total CD4⁺ and CD8⁺ cells for subjects C and E were taken, and each estimator was applied to each subsample. These populations have rarefaction curves that plateau, so making the assumption that the value of the plateau S_{obs} is the diversity of the whole population, the distribution of errors for each estimator error = $\frac{|S_{obs} - \hat{S}_{obs}|}{S_{obs}}$ is shown.

negative exponential estimators ($p < 0.0005$ for each estimator comparison with DivE, $n = 20$; two-tailed binomial test) (Table 2.2, Figure 2.7).

Table 2.2.: Comparison of estimator performance for TCR data.

Estimator	Median Error ¹ (%)	p-value ²
Chao1bc	43.8	0.0004
ACE	42.8	0.0004
Bootstrap	65.3	< 0.0001
Negative-exponential	50.7	< 0.0001
Good-Turing	61.7	< 0.0001
DivE	6.7	NA

¹ Median absolute percentage error between $\hat{S}_{obs}$ and S_{obs}

² p-value of the significance of the differences between the errors of each estimator and DivE ($n = 24$; two-tailed binomial test)

As neither the HTLV-1 nor the microbial data exhibit rarefaction curves that plateau, we cannot apply the same analysis to these datasets. Instead we took advantage of the fact that, for 11 of the 14 HTLV-1 subjects, the data comprised three time points, with three samples drawn at each time point in immediate succession from the subject. For a given

subject and a single time point, the three samples were combined *in silico* to produce a single pooled sample. We compared the observed species richness of the pooled sample to each estimator's estimates from a subsample (Figure 2.8, Figure 2.9). The total blood diversity must be at least as great as that observed by pooling the samples. However, all existing estimators estimate the total diversity to be less than that observed. Based on a single subsample, the Chao1bc, ACE, Bootstrap, Good-Turing and negative exponential estimators respectively estimate medians of 27.0%, 12.7%, 71.1%, 65.5%, and 47.6% fewer clones than observed in the pooled samples ($n = 11$). Since the pooled samples do not saturate, and since the blood contains approximately 10^5 times more infected cells than the pooled sample, the diversity observed in the pooled sample is likely to be a small fraction of the total diversity. Since the existing estimators produce estimates lower than the pooled sample diversity, let alone total blood diversity, this represents a considerable error. We used DivE to produce two estimates: the pooled sample diversity and blood diversity. From the subsamples DivE estimated a median of 2.6×10^3 clones in the pooled samples, a median error of 2.5% ($n = 11$) (Figure 2.8, Figure 2.9). Additionally, DivE estimated 2.8×10^4 clones in the blood, approximately one log higher than the observed pooled sample diversity. Whilst we cannot determine whether or not this is accurate it is at least plausible, considering that it is not less than the diversity of the pooled sample, that the sampling fraction is very small, and that the rarefaction curve has not reached a plateau.

2.3.4. Estimate Error as a Function of Data Curvature

Next we sought to identify conditions under which DivE would be prone to error and should not be applied. When the observed rarefaction curve is linear, the data imply a constant rate of species accumulation, and so provide little information on how quickly the rate of species accumulation will decrease. This is usually indicative of severe under-sampling. We predicted that DivE will fail to give accurate estimates given such a near linear rarefaction curve. We tested this prediction by calculating the error in the DivE estimates as a function of rarefaction data curvature.

The curvature parameter C_p was quantified by the area between the observed rarefaction curve and a linear rarefaction curve, as a fraction of the maximum possible area, which occurs when the rarefaction curve saturates immediately. C_p can take values between 0 and 1, where 1 reflects perfect saturation and 0 reflects a constant rate of species accumulation (Figure 2.10). We took additional samples of 0.1% of the total CD4⁺ and CD8⁺ cells for subjects C and E to obtain lower curvature values.

As expected, at very low curvatures ($0.016 \leq C_p \leq 0.101$), DivE was prone to over-estimation and performed poorly (Figure 2.11), with median error 0.23. However, for

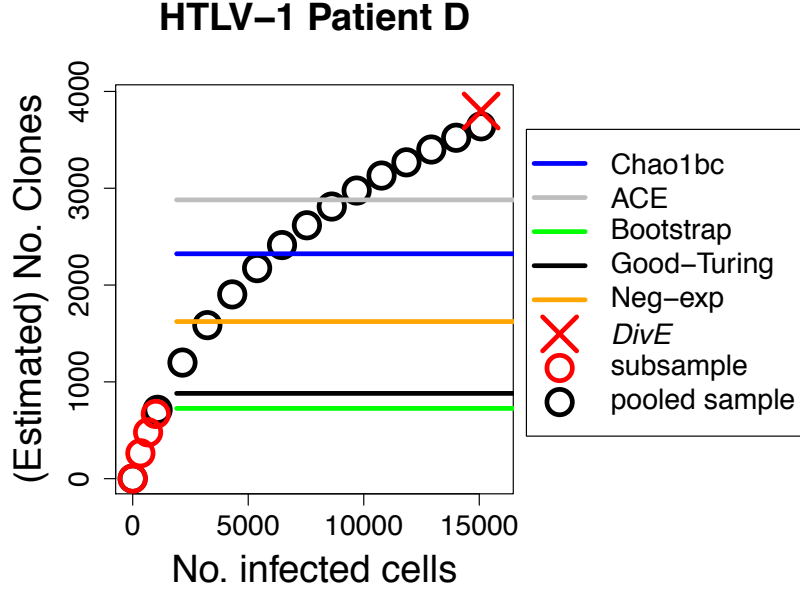


Figure 2.8.: **Ecological estimators underestimate diversity in HTLV-1 infection.**

For HTLV-1 Patient D, three samples are pooled. Rarefaction curves from the pooled sample (black circles) and a subsample (red circles) are shown. Chao1bc, ACE, Bootstrap, Good-Turing and negative exponential estimates (blue, grey, green, black, and orange lines respectively) from the subsample, and DivE estimates (red cross) from the same subsample are plotted. Existing estimators produce a single estimate of diversity, and so their estimates are shown as lines. The diversity in the blood must be at least as great as that observed by pooling the samples. All existing estimators estimate the total diversity to be less than that observed. Given that the observed diversity is likely to be a small fraction of the total diversity this represents a considerable error. We used DivE to produce two estimates: the diversity in the pooled sample (i.e. in 15000 cells, red cross) and the total diversity of the blood. DivE accurately estimates the pooled sample species richness from the subsample, but also predicts higher values of species richness in the blood, consistent with the unseen clones implied by the pooled rarefaction curve. See Figure 2.9 for further examples.

under-sampled populations of intermediate curvature ($0.11 \leq C_p \leq 0.62$) DivE improved markedly (median error = 0.06), and typically outperformed the other estimators (Figure 2.11, Table 2.3). Finally, all estimators perform well when the curvature is high and most of the diversity has been observed (Figure 2.4, panels **D**, **H** and **L**).

We next tested DivE using the *Prochlorococcus* data [93], with multiple subsamples of increasing curvature (as for the TCR data). At low curvatures DivE again performed poorly, but it became more accurate as the curvature increased. For under-sampled popu-

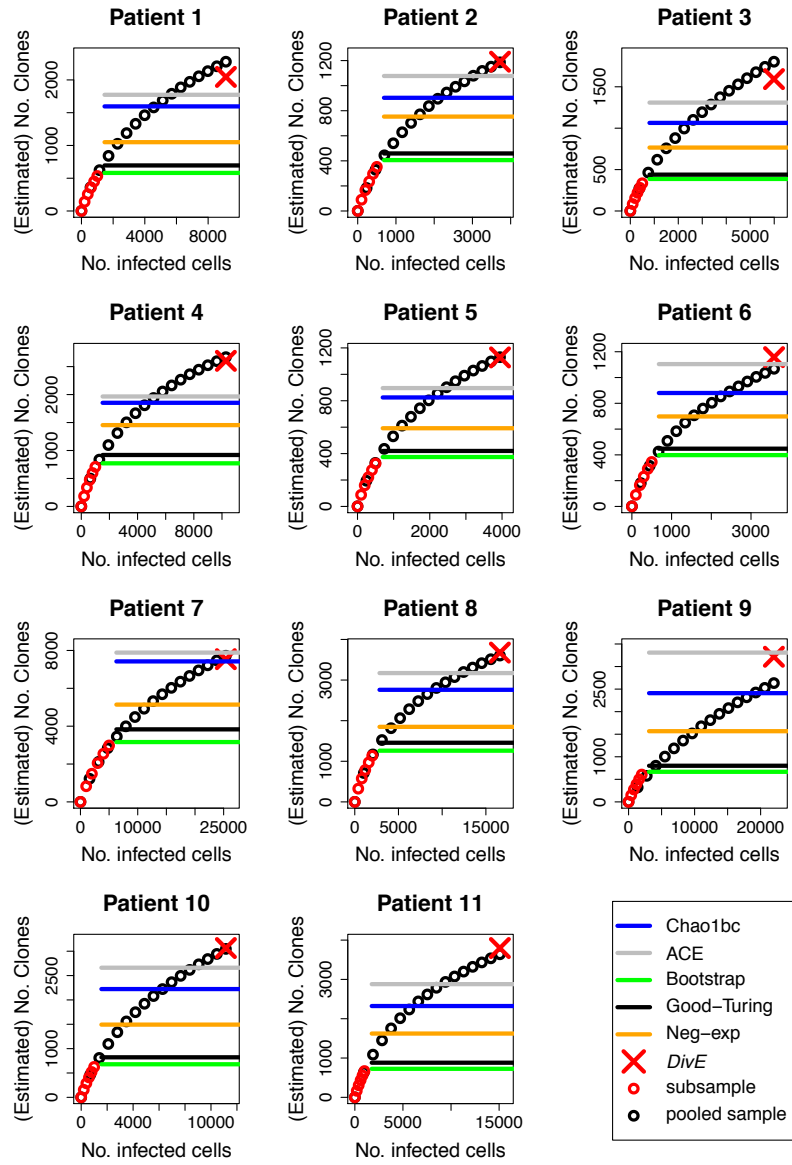


Figure 2.9.: **Ecological estimators underestimate diversity in HTLV-1 infection.**

As for Figure 2.8. Ecological estimators typically estimate fewer clones than observed in pooled sample. In contrast, DivE accurately estimates the pooled sample species richness from a subsample.

lations of intermediate curvature, DivE again outperformed the other estimators, although the differences between the estimator errors were not as dramatic as with the TCR data (Figure 2.12).

Very low curvatures suggest severe under-sampling and researchers should exercise caution

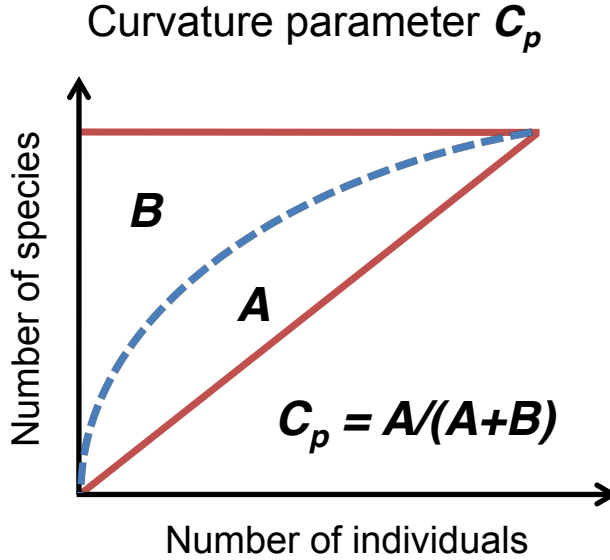


Figure 2.10.: **Rarefaction curvature parameter C_p .** Rarefaction curves (dashed) and lines of constant rate of species-accumulation and perfect saturation (solid) are shown. Areas between the line of constant rate of species-accumulation and the rarefaction curve (A), and between the rarefaction curve and the line of perfect saturation (B) are indicated. Note $C_p = 0$ when the rarefaction curve is linear.

with such data. It is unlikely that any species richness estimator will be accurate or informative in such cases.

2.3.5. Example Application: Estimated Number of HTLV-1 Infected Cell Clones

In both HTLV-1 infection and infection with the related bovine leukaemia virus (BLV), accurate determination of the number of infected cell clones in the host is critical to understanding retroviral dynamics and pathogenesis [25, 112, 113, 114]. As we have seen in Chapter 1, estimates of the number of clones have risen steadily in line with the sensitivity of the methods used. Charles Bangham's laboratory has observed as many as 5000 clones in samples derived from 10 μ g of DNA [11, 22, 115]. Here we make two different estimates of the number of HTLV-1 infected cell clones: (i) in the circulation; and, (ii) in the whole body. See Section 2.2.9 for details of HTLV-1 population size estimation.

The mean estimated number of clones in the circulation in a single host was 2.9×10^4 . It is unknown whether the population structure of HTLV-1 clones in the blood reflects that

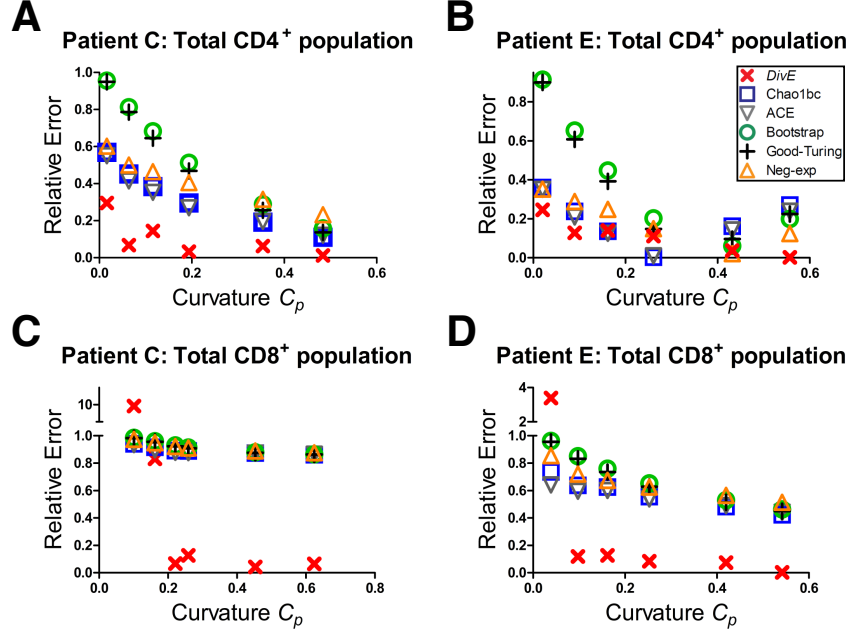


Figure 2.11.: **Test of species richness estimators at different values of curvature parameter C_p using TCR data.** The curvature parameter C_p is plotted against the relative error $= \frac{|S_{obs} - \hat{S}_{obs}|}{S_{obs}}$ of each estimator. Four patient data sets are shown: **A** total $CD4^+$ from patient C; **B** total $CD4^+$ from patient E; **C** total $CD8^+$ from patient C; **D** total $CD8^+$ from patient E. Each point represents an estimate from a subsample of data. Note the plots have different y-axis scales and the y-axes in **C** and **D** are segmented. Broadly, the accuracy of all estimators improves as C_p increases, and this increase is more pronounced for DivE. From $C_p > 0.1$, DivE generally outperforms the existing estimators, but is prone to error at very low values of C_p , when the rarefaction curve implies a near-constant rate of species accumulation.

in solid lymphoid tissue and the spleen. If we assume that these two populations have similar structures, and thus that it is justified to extrapolate to the whole body, we obtain an average of 6.2×10^4 clones, i.e. approximately only twice as many clones, although there are more than 300 times as many infected cells in the body as the blood. These new estimates in the blood and body are approximately 1 and 1.3 logs higher respectively than those calculated using ACE and Chao1bc ($p < 0.0001$, two-tailed paired Mann-Whitney U-test), and more than 2 logs higher than previously published estimates (Figure 2.13) [22, 24, 25, 26].

Table 2.3.: Estimator error variation with curvature in TCR data

Estimator	low ¹ C_p ²	p ³	intermediate C_p	p	overall	p
Chao1bc	56.9	0.45	42.3	0.002	46.8	0.0015
ACE	55.0	0.45	44.2	0.002	46.6	0.0015
Bootstrap	91.4	0.45	53.0	<0.0001	72.1	<0.0001
Neg-exp	60.2	0.45	51.4	0.0002	53.9	0.0003
Good-Turing	90.0	0.45	51.4	<0.0001	69	<0.0001
DivE	23.1	NA	6.5		7.9	NA

¹ Low curvatures C_p in range $0.016 \leq C_p \leq 0.101$, intermediate curvatures in range $0.11 \leq C_p \leq 0.62$.

² Median absolute percentage error between $\hat{S}_{obs}$ and S_{obs} .

³ p-value of the significance of the differences between the errors of DivE and each other estimator, for each curvature range.

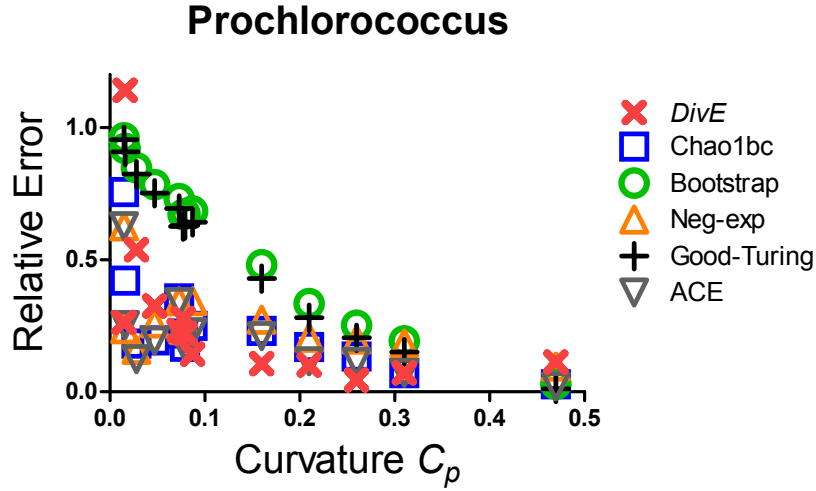


Figure 2.12.: Performance of species richness estimators in metagenomic data.

The curvature parameter C_p is plotted against the error of each estimator. Each point represents an estimate from a sample from the *Prochlorococcus* data. As with the TCR data, DivE typically outperforms the other estimators from $C_p \approx 0.1$ onwards. As predicted, DivE is prone to error at lower values of C_p , but becomes more accurate as C_p increases.

2.3.6. DivE Uncertainty

Because of its heuristic nature, DivE lacks formal statistical confidence intervals. Uncertainty in the estimates produced by DivE has two sources: parameter values in each respective function (within-model variation), and the choice of function (between-model variation). Using standard errors of parameter estimates to calculate confidence intervals ignores uncertainty from function selection. Information theoretic approaches that take account of model selection uncertainty have become increasingly common in ecol-

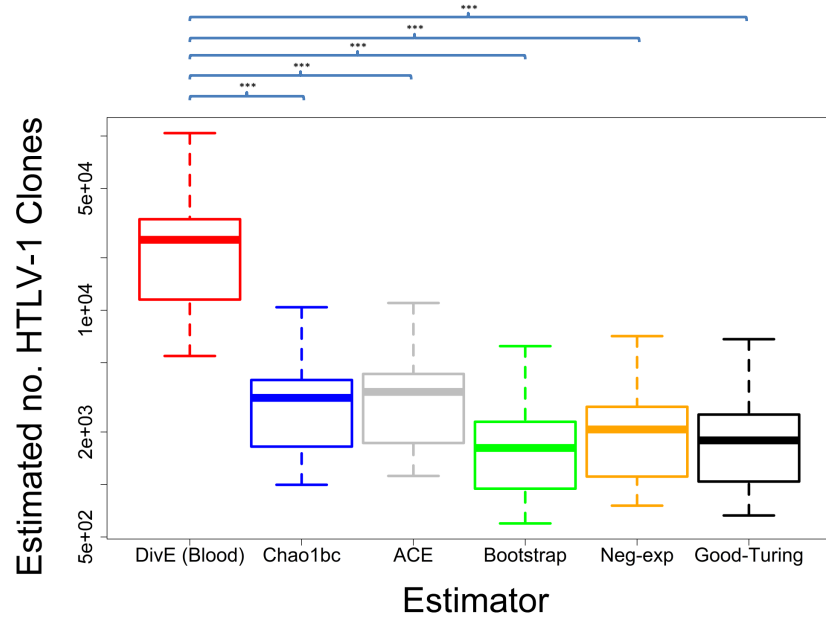


Figure 2.13.: **Diversity estimates in HTLV-1 infection by estimator.** Each estimator was applied to 105 patient datasets, from 14 different HTLV-1⁺ subjects. All subjects either had HAM/TSP or were ACs. *** signifies $p < 0.0001$ respectively; two-tailed Mann-Whitney test.

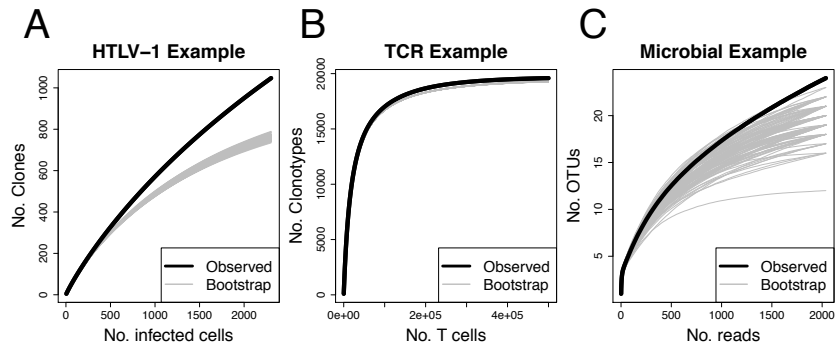


Figure 2.14.: **Rarefaction plots from bootstrap samples of HTLV-1, TCR, and Microbial data.** Rarefaction plots from 100 bootstrap samples (grey) for each of **A** HTLV-1, **B** TCR, and **C** microbial data. The species richness of the bootstrap samples is at most the species richness of the original data (black), and is substantially less in the majority of cases, although this effect is less noticeable with the TCR data.

ogy [116, 117, 118] and elsewhere. There are broadly two approaches: i) computing AIC weights, and ii) repeated resampling and model ranking to determine bootstrap model selection probabilities [109]. However, neither approach is appropriate in our case. We

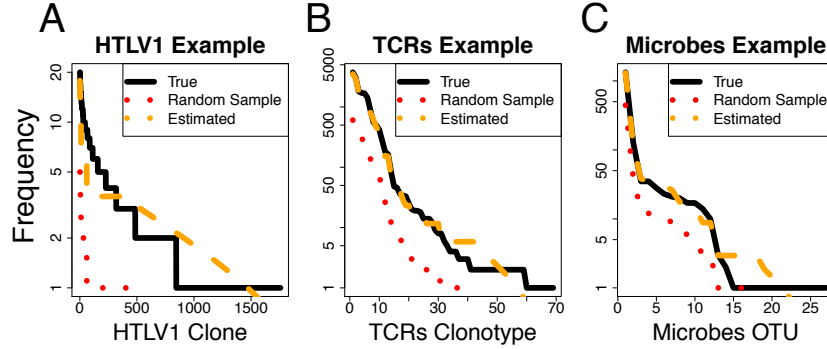


Figure 2.15.: **Validation of Dive distribution generation algorithm.** The Dive distribution generation algorithm (Figure 2.3) was applied to random samples (red dashed) of observed data (black solid). Accuracy was evaluated by comparing the estimated distribution (orange dashed) to the true distribution of the full observed data (black). Examples for HTLV-1 (A), TCR (B) and microbial datasets (C) are shown.

do not rank functions using AIC since this produces less accurate estimates than Dive (Table 2.1), and so we cannot use AIC weights to derive confidence intervals. Further, since there is a systematic bias towards lower species richness in bootstrap samples (Figure 2.14), a similar bias may be introduced in the estimation of bootstrap model selection probabilities, leading in turn to a bias in species richness estimation. Systematic underestimation in bootstrap samples is particular to species richness estimation: this does not highlight a general problem with resampling to quantify model selection uncertainty. As a pragmatic indicator of estimate variability, we use the range of estimates produced by the five best-performing functions; the geometric mean of these five functions is taken as the point estimate (Table A.2).

2.3.7. Distribution Generation Algorithm

The distribution generation algorithm was reasonably accurate for the HTLV-1 data, and considerably more accurate for the TCR and microbial data. The mean error between the estimated and true distributions was 32.1%, 3.7%, and 6.0% for the HTLV-1, TCR and microbial data respectively. The mean error between the estimated and true Gini coefficients was 7.5%, 1.2%, and 1.1% for the HTLV-1, TCR and microbial data respectively (Table 2.4). For the HTLV-1 data, the algorithm underestimated the abundance of the largest clones, but we did not observe this effect in the TCR and microbial data (Figure 2.15).

Table 2.4.: Performance of DivE frequency distribution generation algorithm

Data Source	Mean Error (%) ¹	Mean Gini Error (%) ²
HTLV-1	32.1	7.5
TCR	3.7	1.2
Microbial	6.0	1.1

¹ Mean error across all subjects and all small subsamples, for each data source. Small subsamples were defined as those $\leq 50\%$ of the size of the observed each patient data set. Error defined as the sum of absolute discrepancies between true and estimated frequency distributions, divided by area under true distribution.

² Mean percentage error across all subjects and all small subsamples in the Gini coefficients of the true and estimated distributions.

2.4. Discussion

We wished to estimate species richness in three microbiological and immunological datasets. Initially we used estimators that are reported to perform well in ecology [56, 78, 79, 103, 104, 119]. In the datasets with rarefaction curves that did not plateau, these estimators were biased by sample size. For datasets with rarefaction curves that did plateau, estimates were consistent, but in such cases estimators contribute little information because approximate species richness is already known. Comparable results have been reported elsewhere [56, 60, 105]. By combining data from multiple independent HTLV-1 samples, we showed that these estimators substantially underestimated species richness.

We then developed a new approach, DivE, to estimate species richness and frequency distribution. In our first validation, DivE consistently and accurately estimated the diversity of the observed data from incomplete subsamples of that data. We subsequently determined conditions where DivE would fail and should not be applied. When the rarefaction curvature was low and the data implied a near-constant species-accumulation rate, DivE was prone to overestimation. However, in under-sampled populations of intermediate curvature, DivE substantially improved. The DivE distribution generation algorithm performed with reasonable accuracy (Table 2.4, Figure 2.15). We believe there are two principal reasons why DivE is more accurate than the other estimators. First, functions are selected on their ability to consistently predict additional data. The additional data (i.e. the full sample) has no influence on fitted parameter values, and so DivE not only assesses goodness of fit, but also evaluates the accuracy of the function. Second, DivE does not require *a priori* assumptions regarding the form of the clonotype frequency distribution.

We argue that biologically meaningful and useful estimators should be able to estimate species richness in a specified population. This is not the case with the existing estimators we tested. Any estimator that does not take population size into account effectively

assumes saturation, i.e. that after a certain undefined point, the number of additional individuals does not affect the number of species. However, many populations do not plateau in this way. This is particularly true of immunological populations, where many TCR clonotypes or microbial OTUs are likely to be represented by a single individual. Therefore, while it would be convenient not to require an estimate of population size, any such estimator would lead to false confidence. In contrast, DivE can estimate diversity in any given population size. However, population size estimation can be nontrivial [120, 121, 122]. In spatially homogeneous populations with equiprobable detection of individuals, estimating population size through scaling by area or volume is justifiable e.g. scaling from cells in 50ml of blood to cells in the total blood volume. When population size estimates are unavailable, it is still usually possible to provide meaningful diversity estimates, e.g. the number of microbes per gram of faeces. DivE may also be useful in deciding the depth of sampling required for an adequate census. Deeper sampling may require more DNA sequencing or a larger tissue sample from a patient, and so minimizing sampling depth has financial and ethical benefits. This is not possible with the other estimators we tested.

The HTLV-1 data consisted of absolute species counts, and so we could estimate HTLV-1 diversity. Microbial and TCR datasets were used only for validation as these data consisted of sequence reads and not absolute counts. To the extent that read abundances differ from absolute counts, such data cannot be used to estimate species richness with any abundance-based estimator (e.g. DivE, Chao1bc, and ACE). Over-amplification by PCR may generate a saturating rarefaction curve that is not due to sampling depth, falsely implying that the majority of species have been observed. This can be seen in our TCR data: plateaus were far lower than previously reported diversity estimates [90, 123]. However, absolute counts can often be obtained (e.g. by spiking a sample with a known quantity of identifiable individuals or by barcoding to identify PCR duplicates).

It is unlikely that sequencing error influenced our HTLV-1 diversity estimates, because sequencing error cannot systematically alter proviral integration site mapping. However, species richness estimates from TCR or microbial data are likely to be susceptible to sequencing error. Sequencing error can falsely increase diversity, and this will influence species richness estimates using any estimator; researchers must therefore exercise caution when analysing such data; ideally by preprocessing the data to remove error prior to further analysis. Caution must also be exercised when assuming that the spatial distribution of individuals is uniform. We believe that these assumptions are reasonable for the blood, but skin tissue for example may be more clustered.

DivE is conceptually simple but can be computationally intensive to implement. When applying DivE to a new type of data it is necessary to ascertain which functions perform best. This requires that many functions be fitted to multiple subsamples. If, for a particular

data type, a given set of functions performs consistently well, application becomes much quicker because only these functions need to be fitted, and it is no longer necessary to fit all functions to all subsamples. In our analysis we found that the best performing functions differed between data types. In the HTLV-1 analysis, functions 8, 10, 20, 25, and 45 performed best, contrasting with functions 10, 17, 20, 55, and 56 in the microbial analysis and functions 8, 12, 16, 22, and 23 in the TCR analysis (see A.1 for function analytical forms). Perhaps unsurprisingly, all of the best-performing functions in the TCR analysis saturated. In contrast, in both the HTLV-1 and microbial analysis, some of the best performing functions did not. These results make sense in light of the rarefaction curves displayed in Figure 2.4. In the HTLV-1 analysis, five functions performed consistently well, and we adopted this number of functions as the basis for the DivE estimate. However, since the optimal functions differ between data types, we advocate careful analysis of function scores to decide how many functions should be aggregated. The DivE estimator has been provided as an R package, available at <http://cran.r-project.org/web/packages/DivE/index.html> .

In summary, we have developed and validated a new approach to estimate species richness and distribution that significantly outperformed existing estimators of biodiversity in the datasets we examined.

3. Applications of DivE

3.1. Introduction

In this chapter we present a comparison of the number of clones, as estimated by DivE, in patients with a range of clinical manifestations of HTLV-1 infection. Many of these results have been published in [7, 11, 124, 125]. All data were obtained by our collaborators using the high-throughput protocol outlined in [22], summarised in Section 1.1.6.

3.2. Clonal Diversity in Adult T Cell Leukaemia / Lymphoma

Approximately 5% of HTLV-1 infected individuals develop Adult T cell Leukaemia / Lymphoma (ATLL) [11], an aggressive malignancy of post-thymic T lymphocytes [115, 126]. There is a relatively high incidence of HTLV-1 in mothers of ATLL patients compared with mothers of other diseases caused by HTLV-1 [127], and childhood infection is believed to predispose HTLV-1-infected patients to develop ATLL [126]. Under the Shimoyama classification, ATLL can be subdivided into acute, lymphoma, chronic and smouldering subtypes [9], which differ in their response to treatment and associated prognosis [115].

ATLL patients have a much higher proviral load (PVL) than ACs and HAM/TSP patients (Figure 3.1), but fewer clones are typically observed (Figure 3.2). There is usually a single clone that has undergone malignant transformation and which consequently occupies a large proportion of total PVL. Thus the oligoclonality index (OCI) is substantially higher in a typical ATLL patient than for ACs and HAM/TSPs [22, 11].

We estimated the diversity of 196 patients with ATLL. Population sizes were estimated according to equation (2.2). Samples where $C_p < 0.1$, or where the number of sisters was less than 150 were excluded from the analysis (below 150 sisters renders the subsamples required for DivE unfeasibly small), as were 100% monoclonal samples with only a single clone (a rarefaction curve is useless in the latter case). The mean estimated number

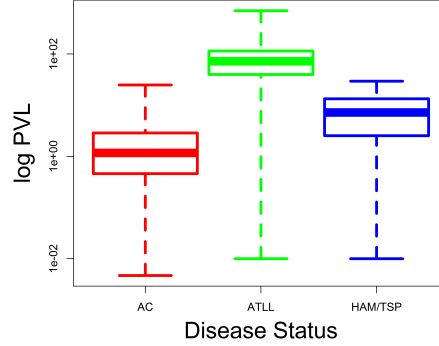


Figure 3.1.: **HTLV-1 proviral load in ACs, ATLL and HAM/TSP patients.** PVL is greater in ATLL patients than it is in HAM/TSP patients, who have higher PVL than ACs ($p < 0.0001$ for all pairwise comparisons, two-tailed Mann-Whitney test). Three ACs with undetectable proviral loads were omitted from plot due to log scale.

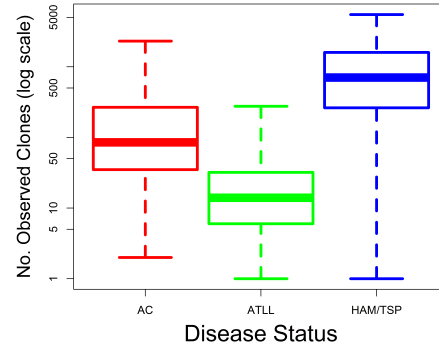


Figure 3.2.: **Number of observed clones in ACs, ATLL and HAM/TSP patients.** There are fewer observed clones in ATLL patients than in HAM/TSP patients, who have higher PVL than ACs ($p < 0.0001$ for all pairwise comparisons, two-tailed Mann-Whitney test).

of clones was 3084, substantially fewer than for ACs and HAM/TSP patients, which is unsurprising given the higher number of observed clones, higher curvature parameter C_p , and OCI (Table 3.1).

ATLL is known to exhibit highly expanded clones and correspondingly high proviral loads. Because highly expanded clones increase the OCI, we would expect that a high OCI would be correlated with a low diversity, as small clones are out-competed by malignant clones. However, it is interesting to note that, within ATLL patients, we see a positive correlation ($p < 0.0001$, two-tailed paired Mann-Whitney test) between estimated diversity and OCI

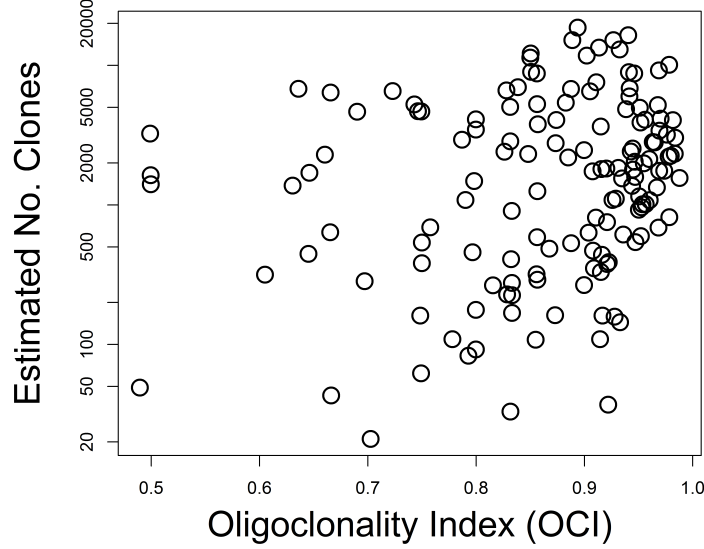


Figure 3.3.: **Relationship between OCI and estimated diversity in ATLL patients.** Estimated diversity (log) strongly positively correlated with OCI ($p < 0.0001$, two-tailed paired Mann-Whitney test)

(Figure 3.3). A high OCI can occur through reduced diversity, but also through the reduced share of PVL among existing clones. We conclude that highly expanded clones suppress the abundance of smaller clones to undetectable levels, thus increasing the OCI.

Table 3.1.: Comparison of clonality of ATLL, HAM/TSP and AC patients

Disease Status	PVL (%)	Observed Clones	C_p	OCI	DivE
ATLL	87.6	32	0.218	0.849	3084
HAM/TSP	8.6	1039	0.152	0.467	30272
AC	2.1	209	0.123	0.501	20679

Mean values of PVL, number of clones observed in sample, Curvature C_p , OCI and DivE given for ATLL, HAM/TSP and AC patients. All differences between groups significant: $p < 0.0001$ for each pairwise comparison of each variable, except HAM/TSP vs. ATLL curvature ($p = 0.004$) and HAM/TSP vs. AC OCI ($p = 0.023$), two-tailed Mann-Whitney test

3.3. Clonal Diversity in Asymptomatic Carriers

Approximately 90% of HTLV-1-infected patients are able to control the virus for decades without developing disease. The mechanisms by which they are able to do so have major implications for the development of treatment. A high proviral load (PVL) is a known

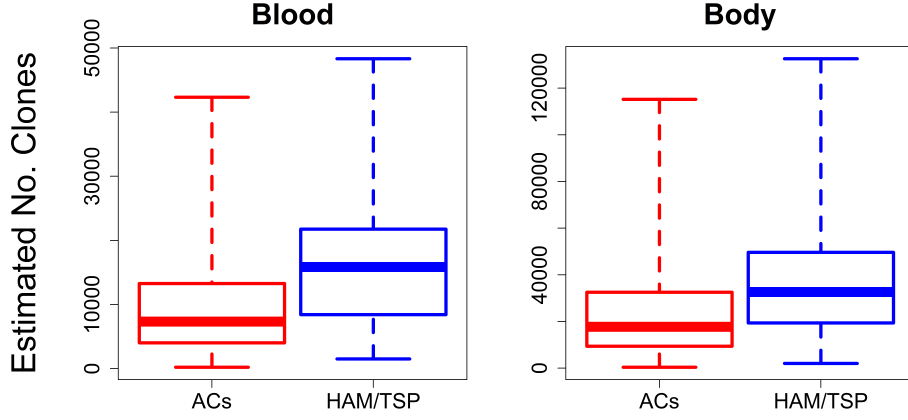


Figure 3.4.: **Comparison of HTLV-1 clonal diversity between ACs and HAM/TSP patients in blood and body.** In both the circulation and the whole body, ACs had significantly fewer clones than patients with HAM/TSP ($p = 0.0001$ and $p = 0.001$ for blood and body respectively; two-tailed Mann-Whitney test)

risk factor of inflammatory disease [22], and PVL in asymptomatic carriers (ACs) is often undetectable.

Undetectable proviral loads can inhibit or preclude a detailed analysis of the clonality of ACs, creating a paradoxical situation whereby the patients who are the most difficult to study are the patients who have the greatest potential to enhance our understanding of HTLV-1 pathogenesis.

We compared the HTLV-1 clonal diversities between ACs and HAM/TSP using data from two cohorts (Figure 3.4). The first cohort consisted of 95 HAM/TSP patients and 68 ACs from Kagoshima, Japan, and the second consisted of 75 ACs from Kumamoto, Japan [7]. Population sizes for blood and body were estimated according to equations (2.2) and (2.3) respectively. Samples where $C_p < 0.1$, or where there were less than 150 infected cells were excluded from the analysis. Therefore the following analysis, as with all analyses of ACs, is biased towards ACs with relatively high proviral loads.

In both the blood and the body, the estimated number of clones was significantly lower in ACs than HAM/TSP patients ($p = 0.0001$ and $p = 0.001$ for blood and body respectively; two-tailed Mann-Whitney test). In ACs, the median number of clones in the blood and body were 7.3×10^3 and 1.7×10^4 respectively, whereas in HAM/TSP patients, the corresponding values were 1.5×10^4 and 3.2×10^4 . In both cases, estimates of clonal diversity showed a strong correlation with PVL (Figure 3.5).

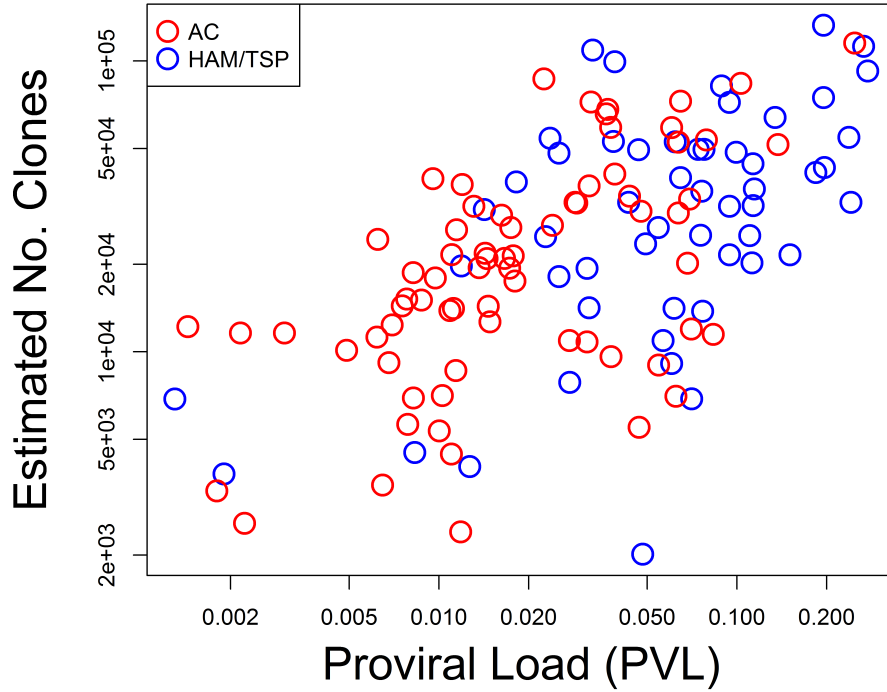


Figure 3.5.: **Relationship between clonal diversity and proviral load in Japanese cohort of ACs and HAM/TSP patients.** Estimated diversity strongly positively correlated with proviral load ($R^2 = 0.247$, $p = 0.0002$ for HAM/TSP and $R^2 = 0.392$, $p < 0.0001$ for AC, linear regression on log scale)

The estimated number of clones in these Japanese cohorts is less than the number of clones estimated in Figure 2.13, and the difference is likely due to the smaller PVL of the Japanese cohort. The mean proviral loads of ACs and HAM/TSP patients in the Japanese cohort was 2.0% and 7.7% respectively, compared to 4.3% and 9.6% in the cohort analysed in Figure 2.13 and Chapter 2.

Fewer clones in ACs strongly implies less infectious spread, which in turn raises questions about the role of infectious spread in disease progression. More precise estimation of the ratio of infectious to mitotic spread in different patients is ongoing (see Chapter 5).

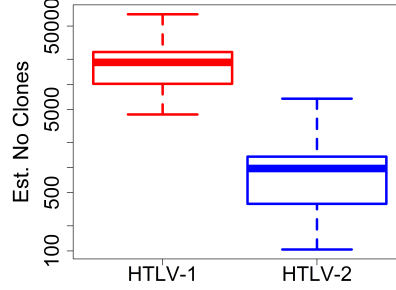


Figure 3.6.: **Comparison of HTLV-1 and HTLV-2 clonal diversity.** DivE estimates are plotted on log scale.

3.4. Relationship Between HTLV-1 and HTLV-2 Clonal Diversity Mirrors Relationship Between Clonal Diversity of $CD4^+$ and $CD8^+$ T Cells in HTLV-1

Human T-lymphotropic virus type-2 (HTLV-2) is a retrovirus that is not known to cause inflammatory disease or malignancy of T cells, unlike the closely related retrovirus HTLV-1. Both viruses establish a persistent lifelong infection, and are transmitted through the sharing of contaminated needles, sexual intercourse and blood transfusion [128]. HTLV-2 does not exhibit the same endemic prevalence of Japanese islands or the Caribbean, but instead is endemic in indigenous populations in Africa [124] and North and South America [129]. The viruses are similar in their genomic structure [130], but they crucially differ in their preferred target host cell [124]. Approximately 95% of HTLV-1 PVL is comprised of $CD4^+$ T cells, with the remaining 5% in $CD8^+$ T cells. Interestingly, this ratio of infected $CD4^+$ T cells to $CD8^+$ T cells is reversed in HTLV-2, where almost 100% of infected cells are $CD8^+$ but infected $CD4^+$ are detected [124].

Here we compare the clonal diversity in the circulation of HTLV-1 and HTLV-2 in 16 HTLV-1⁺ patients (9 ACs, 7 HAM/TSP patients) and 28 HTLV-2⁺ patients. HTLV-2 proviral load is defined in the same way as for HTLV-1, i.e. by the number of viral copies as a fraction of PBMCs, and so population size estimates were again performed using equation (2.2), as for HTLV-1. We then show that the relationship between HTLV-1 and HTLV-2 clonal diversities is similar to the relationship between the clonalities of HTLV-1-infected $CD4^+$ and $CD8^+$ T cells, using another cohort of 6 HAM/TSP patients 6 HTLV-1 ACs.

Figure 3.6 shows DivE estimates of clonal diversity for HTLV-1 and HTLV-2. The mean number of clones for the HTLV-1 patients was 2.0×10^4 , substantially higher than for

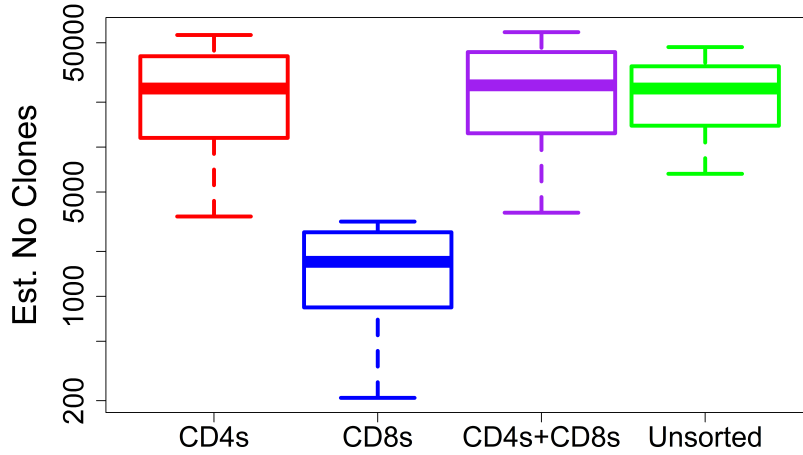


Figure 3.7.: **HTLV-1 clonal diversity of infected $CD4^+$ and $CD8^+$ T cells.** DivE estimates are plotted on log scale. Estimates are from samples where HTLV-1⁺ cells were sorted for $CD4^+$ phenotype (red), $CD8^+$ phenotype (blue), the sum of estimates from these samples (purple), and estimates from samples that were not sorted by T cell phenotype (green). There is broad agreement between the sum of estimates from $CD4^+$ and $CD8^+$ samples, and the estimates of clonal diversity from the unsorted samples.

the mean diversity of HTLV-2 patients at 1253 clones. The difference does not appear to be attributable to the number of infected cells, because the mean values of PVL in this cohort are 6.1% and 9.3% for HTLV-1 and HTLV-2 respectively. Indeed the difference in diversity and the similarity in PVL strongly suggests that mitotic spread makes a greater contribution to HTLV-2 persistence than it does to that of HTLV-1. Since HTLV-2 is not associated with disease, it is therefore possible that the level of infectious spread is a determinant of disease status. If true, a potential mechanism would be that fewer instances of infectious spread will decrease the probability of proviral integration in regions of the genome that are vital to the host, for example in cell-cycle maintenance and replication.

In the second cohort, we estimated the diversity of HTLV-1 clones in $CD4^+$ and $CD8^+$ T cells, as well as total diversity from unsorted samples, using data generously provided by our collaborators *Melamed et al* (manuscript in preparation). $CD4^+$ and $CD8^+$ T cells were magnetically sorted using MACS (Miltenyi), whereupon HTLV-1 integration site data for each population was obtained using the high-throughput protocol described in Section 1.1.6. The data includes the proportions (p_{CD4} and p_{CD8}) of HTLV-1-infected cells that were $CD4^+$ and $CD8^+$ respectively, and so population sizes N_{CD4} and N_{CD8}

were estimated by

$$N_{\text{CD4}} = 5 \times 3 \times \frac{PVL}{100} \times 10^9 \times p_{\text{CD4}} = N_{\text{blood}} \times p_{\text{CD4}} \quad (3.1a)$$

$$N_{\text{CD8}} = 5 \times 3 \times \frac{PVL}{100} \times 10^9 \times p_{\text{CD8}} = N_{\text{blood}} \times p_{\text{CD8}} \quad (3.1b)$$

The mean estimated number of clones was 2.7×10^4 in CD4^+ T cells and 1.7×10^3 in CD8^+ T cells (Figure 3.7). Remarkably, these estimates strongly resemble the clonal diversities of HTLV-1 and HTLV-2. There is also a reasonable agreement between the sum of estimated diversities of CD4^+ and CD8^+ T cells (mean 2.8×10^4) and the estimated diversity from samples that were not sorted for T cell phenotype (mean 2.4×10^4 clones) (Figure 3.7).

These estimates suggest that the preferred target cell, in addition to the genomic location of the proviral integration site, influences the proviral population structure and the pathogenesis of infection. There may be a quality intrinsic to CD8^+ T cells that renders their infection less harmful to the host than infection of their CD4^+ counterparts.

3.5. HTLV-1 Clonality in Early Infection

HTLV-1 infection has a long clinical latency period in patients who go on to develop malignancy or inflammatory disease [113]. Patients do not typically present symptoms until a stable population structure of HTLV-1 clones has already been established, and thus comparatively little is known about early infection. Here we present DivE estimates for three immune-suppressed patients who contracted HTLV-1 after exposure to solid organ transplantation. The circumstances of these patients' infection is highly unusual, and so the following diversity estimates must be considered as a proxy for the clonality of early normal HTLV-1 infection.

Integration site abundance data were available for three timepoints for patient A, four timepoints for patient B, and five timepoints for patient C. It was hoped that these data would enable estimates of clonal diversity over time and inform how the viral population structure is established, but unfortunately after excluding data points where either the number of cells was less than 150, or where $C_p < 0.1$, few data points remained. Population sizes were estimated using Equation (2.2). The date of infection was taken to be the date of transplantation.

We were able to estimate the clonal diversity for all three data points for patient A. The number of clones was 2.3×10^5 at 38 days post infection, which decreased by half to 1.1×10^5 clones at 93 days post infection, and then increased slightly to 1.3×10^5 clones

at 133 days post infection. The dramatic decrease can be attributed to the patient’s PVL, which decreased by more than an order of magnitude in between the first two timepoints.

In patient B, only two timepoints were feasible. Between 50 and 77 days post infection, the number of clones increased from 9.4×10^4 to 1.7×10^5 , again in line with their PVL which increased by a factor of four. For patient C, DivE estimates were only possible for a single time point at 182 days post infection, where the number of clones was 7.1×10^3 .

The mean number of clones over all patients and time points was 1.2×10^5 , substantially higher than for HAM/TSP patients and ACs. The relatively low curvature of the sample may be responsible for overestimating the patients’ clonal diversity. However it is not straightforward how best to reconcile high diversity in the presence of low curvature. Highly diverse populations would indeed produce a rarefaction curve with a lower curvature than would a low diversity population. It may also be true that the number of clones is greater in these patients because of immune suppression as a result of transplantation.

3.6. Comparison of HTLV-1 Clonality in Blood and Skin

Infective dermatitis associated with HTLV-1 (IDH) is the most common manifestation of disease in HTLV-1-infected children [131, 132]. It is a chronic eczema, and there is some evidence that presentation of IDH in patients without neurological symptoms predicts later progression to HAM/TSP [133, 134]. We compare the number of clones in the circulation and in the skin, using a data set with 10 blood samples and 10 skin samples from 6 patients with IDH.

We cannot estimate the clonal diversity of the skin as we can for the blood, because there are complicating factors in population size estimation and in the extent to which samples are reflective of the total skin tissue. Because the distribution of infected cells in the skin is non-uniform, we cannot utilise two important simplifying assumptions. First, that the population is homogeneous (i.e. that samples taken from one location do not systematically differ from samples taken from another). Second, that infected cells are sampled independently, randomly, and with an equal probability of detection. On the contrary, different areas of skin have different concentrations of infected cells, and tissue samples are purposefully chosen from visible skin lesions and thus are highly non-random. Further, while it is safe to assume the random mixing of infected cells in the blood, this assumption is less tenable for the skin. Because there will be a higher concentration of HTLV-1-infected cells in our skin samples than the rest of the skin, we also cannot estimate the population size via scaling as we do in equations (2.2) and (2.3).

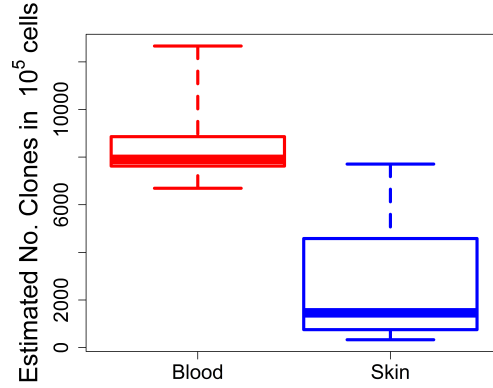


Figure 3.8.: **DivE** estimates in the blood and skin of patients with IDH.

Nevertheless, the clonality of the skin remains of interest, as does its relationship to the clonality of the circulation. Cautioning that the skin estimates are taken from a population where homogeneity cannot be assumed, we estimated the diversity of the blood and skin at a standardised population size of 10^5 infected cells to facilitate comparison. As before, samples where $C_p < 0.1$ or where the number of sisters was less than 150 were excluded from the analysis. The median number of clones was 1458 and 7904 in the skin and blood populations respectively ($p = 0.0006$; two-tailed Mann-Whitney test).

The greater diversity of the blood is perhaps surprising in light of the fact that there is no significant difference between the clonality index or curvatures of the two populations. Further, the difference cannot be attributed to PVL, since an standardised population size that does not depend on PVL was used. It is also revealing that these diversity estimates, while lower than those of ACs and HAM/TSP patients, are based on a substantially smaller population size (10^5 rather than $\sim 10^{11}$ infected cells), suggesting that the bulk of diversity is contained in a relatively small proportion of PVL. This implies that a near complete census could be obtained with blood samples that are large but still feasible.

3.7. Discussion

A difficulty of using DivE is in deciding whether to fit all candidate functions to the data, or whether to use functions that have previously performed well on data that is similar to the population that is to be examined. While the former will produce estimates that are more data-specific, the estimates are rendered less suitable for comparison. That is, are the diversity estimates between two populations different because of the populations, or because of the function chosen?

Using functions that have performed well previously avoids this issue, but is not a viable option where the data sources are markedly different, or where there are strong *a priori* reasons to believe the relationship between population size and diversity may be qualitatively different. Ultimately, because DivE is a heuristic estimator, deciding which approach is most appropriate will depend on the problem at hand.

Two lines of evidence point to infectious spread as a driving determinant of inflammatory disease and malignancy. First, the number of clones is substantially lower in ACs than in HAM/TSP patients, and this has been observed across multiple cohorts. Second, HTLV-2 is much less clonally diverse, but has not been associated with disease or malignancy.

At first, however, the hypothesis that infectious spread correlates with disease status appears difficult to reconcile with the fewer observed and estimated clonal diversities of leukaemic patients. Still, it is highly plausible that greater infectious spread increases the probability of proviral integration in (and disruption of) essential cell-maintenance regions of the host genome. If a malignant clone is established, a polyclonal background (i.e. a population of clones with a low OCI) will eventually change to a monoclonal background as the malignant clone out-competes the other clones. That is, leukaemic patients may initially have a higher diversity that vanishes upon malignant transformation of a single clone.

Alternatively, it may be that it is simply high PVL that causes more clones, and that the per-capita rates of infectious spread do not differ between patients with different diseases. If so, this implies that infectious spread does not play a role in determining disease status. However, this explanation would not account for the differences in observed and estimated clonal diversities of HTLV-1 and HTLV-2, where the proviral loads are similar, and so this explanation is less likely.

In Chapter 5, we develop a model of within-host HTLV-1 dynamics that will infer rates of infectious and mitotic spread. In the next chapter, we outline and develop stochastic mathematical methods required by our within-host model.

4. Stochastic Modelling: Birth-Death Processes

In this chapter, we introduce the concepts and techniques we will use to model the within-host dynamics of HTLV-1 persistence. We define the *master equation*, and provide an explicit derivation of this quantity for a mass-action birth-death process. In particular, we explore approximations that avoid expensive numerical computation of the solution to the master equation. The concepts explored in this chapter will form the basis of the eventual model that is developed and applied in Chapter 5.

4.1. Introduction

Ordinary differential equations (ODEs) are ubiquitous in mathematical modelling. They are used in a wide variety of biological contexts, for example viral infections [38, 135, 136, 137], epidemiology [138, 139, 140], microbiology [141, 142], immunology [136, 143, 144, 145, 146], and genetics [147]. ODEs, often referred to as “reaction rate equations” [148], are deterministic: given a set of initial conditions and details of the ODEs, the state of the system at any time t can be computed, at least numerically.

In many situations, however, deterministic models do not adequately encompass all that we need to know about a system [149, 150, 151], in particular its inherent uncertainty [152]. Broadly, we can view the outcome of a deterministic model as the expected value of the system [153]. This is usually appropriate where the number of particles in the system is large and where random effects can be neglected. However, when there are small particle numbers [154, 155] or when the system is highly variable, we may be interested in a range of possible outcomes, as here the signal to noise ratio is low and random effects have a larger relative effect on the system [156, 157]. In addition, the expected value of the system can actually be misleading. First, the expected value may not be an integer (i.e. ODEs are continuous), and most biological applications of interacting individuals and species require a discrete state space [153]. Rounding will represent a greater error in small values than in large values. Second, and more subtly, the expected value may be highly improbable

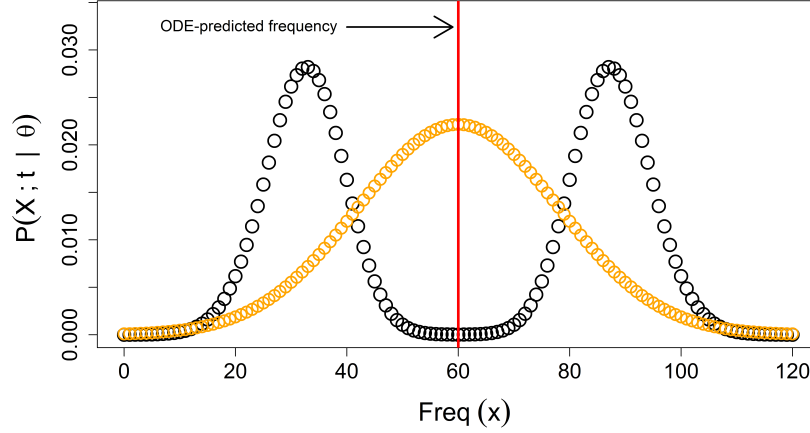


Figure 4.1.: **Example non-normal probability distribution function.** In this bimodal distribution (black), the expected value, which approximately equals the value predicted by an ordinary differential equation, does not correspond to a state or frequency (x) that is actually likely. Here a deterministic model is inappropriate, as is an assumption of normality about the mean (normal distribution given in orange).

when the probability distribution is non-normal, in which case it will no longer correspond to a “typical” value (Figure 4.1).

Stochastic modelling is used to answer the questions: i) how probable is each possible outcome and ii) how do these probabilities change with time? Deterministic modelling can be used to answer these questions, but additional assumptions must be made about the form of the probability distribution, for example that deviations from the predicted value follow a normal distribution [158, 159]. This assumption is not always appropriate, and stochastic modelling is becoming more common in mathematical biology [156, 157]. For example, branching processes have been used to study early HIV infection [160], HIV latency [161], HIV epidemiology [162], and the MMR outbreak [163]. Random walks have been used to model tumour angiogenesis [164], animal movement [165], and chemotaxis [166]. Stochastic differential equations (SDEs) are employed in modelling gene regulatory networks [167], and systems biology [152]. Perhaps the most commonly used stochastic modelling approach is the Stochastic Simulation Algorithm (SSA) of Gillespie [153], which has been used in cell physiology [168], HIV [137, 161], biotechnology [154] gene regulation [155], as well as in its original motivation in chemical kinetics [151, 169]. The Gillespie algorithm is an implicit solution of the Chemical Master Equation (CME), which will be the main focus of this chapter. The CME offers an elegant and complete description [151] of many biological systems (e.g. HIV [161], gene expression [170], intracellular virus

dynamics [156], *E. Coli* [149, 171], and cell division [157, 172]), but it is rarely solved explicitly because it is usually prohibitively large [168, 173, 174]. Branching processes, random walks, the Gillespie algorithm and the master equation provide a description of the inherent uncertainty of the system under examination, albeit with the consequence that their use usually requires substantial computational time [137, 152, 161].

4.2. Master Equation

The following has been described previously [173] and is presented here for completeness. Consider a system with a discrete number of individuals, where each individual belongs to exactly one species or type. Assume that there are $S \in \mathbb{N}$ species ($\mathcal{S}_1, \mathcal{S}_2, \dots, \mathcal{S}_S$), and that species $\mathcal{S}_i$ has a frequency $X_i(t) \in \mathbb{N}_0 = \mathbb{N} \cup \{0\}$ at time t . Assume further that all individuals of the same species are identical, and that the position or location of any one individual of any species may be regarded as uniformly distributed (i.e. spatial homogeneity is assumed). Then the vector $X(t) = (X_1(t), X_2(t), \dots, X_S(t))^T \in \mathbb{N}_0^S$ is a random variable that describes the population of the system at time t . Suppose that there are R reactions $\mathcal{R}_1, \dots, \mathcal{R}_R$ that can happen in the system, where a reaction $\mathcal{R}_r$ ($r \in \{1, \dots, R\}$) is given by

$$\mathcal{R}_r : \mathbb{N}_0^S \xrightarrow{c_r} \mathbb{N}_0^S \quad (4.1a)$$

$$\mathcal{R}_r : \sum_{i=1}^S \kappa_{ri}^{\text{IN}} \mathcal{S}_i \xrightarrow{c_r} \sum_{i=1}^S \kappa_{ri}^{\text{OUT}} \mathcal{S}_i \quad (4.1b)$$

$$\mathcal{R}_r : \kappa_{r1}^{\text{IN}} \mathcal{S}_1 + \dots + \kappa_{rS}^{\text{IN}} \mathcal{S}_S \xrightarrow{c_r} \kappa_{r1}^{\text{OUT}} \mathcal{S}_1 + \dots + \kappa_{rS}^{\text{OUT}} \mathcal{S}_S \quad (4.1c)$$

where $\kappa_{ri}^{\text{IN}} \in \mathbb{N}_0$ is the number of particles of the i^{th} species required for the r^{th} reaction, $\kappa_{ri}^{\text{OUT}} \in \mathbb{N}$ is the number of particles of the i^{th} species present after the r^{th} reaction, and $c_r > 0$ is the reaction constant. c_r is closely related to the reaction rate constant, and represents the average probability that a particular combination of $\sum_{i=1}^S \kappa_{ri}^{\text{IN}} \mathcal{S}_i$ individuals will react according to reaction $\mathcal{R}_r$ in an infinitesimal time interval dt [151, 153]. We define the *stoichiometric vector* ν_r as the difference to the state vector $X(t)$ made by reaction r

$$\nu_r = (\kappa_{r1}^{\text{OUT}} - \kappa_{r1}^{\text{IN}}, \dots, \kappa_{rS}^{\text{OUT}} - \kappa_{rS}^{\text{IN}})^T \in \mathbb{Z}^S \quad (4.2)$$

The *propensity function* α_r of reaction $\mathcal{R}_r$ is the reaction constant c_r multiplied by the number of different combinations of individuals required for reaction $\mathcal{R}_r$, and is given by

$$\alpha_r(x) = c_r \prod_{i=1}^S x_i C_{\kappa_{ri}^{\text{IN}}} = c_r \prod_{i=1}^S \binom{x_i}{\kappa_{ri}^{\text{IN}}} = c_r \prod_{i=1}^S \frac{x_i!}{(x_i - \kappa_{ri}^{\text{IN}})! (\kappa_{ri}^{\text{IN}})!} \quad (4.3)$$

for a potential state of the system $x = (x_1, x_2, \dots, x_S)^T \in \mathbb{N}_0^S$ (i.e. x_i is the number of particles of species S_i). The evolution of the state vector $X(t)$ is given by

$$X(t) = x_0 + \sum_{r=1}^R \mathcal{P}_r \left(\int_0^t \alpha_r(X(s)) ds \right) \nu_r \quad (4.4)$$

for $x_0 \in \mathbb{N}_0^S$. Equation (4.4) states that the population $X(t)$ at time t is equal to the initial population x_0 plus the sum over all reactions r of the change ν_r to the population induced by each reaction, multiplied by the (random) number of times the reaction has occurred. The number of times each reaction occurs is given by a Poisson distribution in the time interval $[0, t]$ with expected value of $\int_0^t \alpha_r(X(s)) ds$, i.e a Poisson distribution $\mathcal{P}_r \left(\int_0^t \alpha_r(X(s)) ds \right) \nu_r$.

4.2.1. Example: Classical Model of Within-Host HIV Infection

It is helpful to provide an example to make the preceding section less abstract. We consider a simple model of the within-host dynamics of HIV infection, originally proposed in [40]. Consider populations of uninfected cells x , infected cells y , and cell free virus particles (virions) v . Uninfected cells are created at a constant rate λ , and die at a rate hx . Infected cells are produced at a rate βxv when uninfected cells and virions meet, and die at a rate ay . Finally, virions are produced from infected cells at a rate ky and die at a rate uv . This leads to the following system of ODEs.

$$\dot{x} = \lambda - hx - \beta xv \quad (4.5a)$$

$$\dot{y} = \beta xv - ay \quad (4.5b)$$

$$\dot{v} = ky - uv \quad (4.5c)$$

In this model, there are three “species” ($\mathcal{S}_1 = \mathcal{X}, \mathcal{S}_2 = \mathcal{Y}$, and $\mathcal{S}_3 = \mathcal{V}$) and six reactions, so $S = 3$ and $R = 6$. The reactions are as follows (c.f. Equation (4.1c))

$$\mathcal{R}_1 : * \xrightarrow{\lambda} \mathcal{X} \quad (4.6a)$$

$$\mathcal{R}_2 : \mathcal{X} \xrightarrow{h} * \quad (4.6b)$$

$$\mathcal{R}_3 : \mathcal{X} + \mathcal{V} \xrightarrow{\beta} \mathcal{Y} + \mathcal{V} \quad (4.6c)$$

$$\mathcal{R}_4 : \mathcal{Y} \xrightarrow{a} * \quad (4.6d)$$

$$\mathcal{R}_5 : \mathcal{Y} \xrightarrow{k} \mathcal{Y} + \mathcal{V} \quad (4.6e)$$

$$\mathcal{R}_6 : \mathcal{V} \xrightarrow{u} * \quad (4.6f)$$

where $*$ is defined to mean “no species” (i.e. reaction $\mathcal{R}_1$ creates species $\mathcal{X}$ from source and reactions $\mathcal{R}_4$ and $\mathcal{R}_6$ are destructive). Letting $X(t) = (x(t), y(t), v(t))^T$, the reactions have the following stoichiometric vectors and propensities (c.f. equations (4.2) and (4.3)):

$$\begin{aligned}\nu_1 &= \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \nu_2 = \begin{pmatrix} -1 \\ 0 \\ 0 \end{pmatrix}, \nu_3 = \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix}, \nu_4 = \begin{pmatrix} 0 \\ -1 \\ 0 \end{pmatrix}, \nu_5 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, \nu_6 = \begin{pmatrix} 0 \\ 0 \\ -1 \end{pmatrix}, \\ \alpha_1(X(t)) &= \lambda \times ({}^x C_0 \times {}^y C_0 \times {}^v C_0) = \lambda, \quad \alpha_2(X(t)) = h ({}^x C_1 \times {}^y C_0 \times {}^v C_0) = hx, \\ \alpha_3(X(t)) &= \beta \times ({}^x C_1 \times {}^y C_0 \times {}^v C_1) = \beta xv, \quad \alpha_4(X(t)) = a \times ({}^x C_0 \times {}^y C_1 \times {}^v C_0) = ay, \\ \alpha_5(X(t)) &= k \times ({}^x C_0 \times {}^y C_1 \times {}^v C_0) = ky, \quad \alpha_6(X(t)) = u \times ({}^x C_0 \times {}^y C_0 \times {}^v C_1) = uv\end{aligned}$$

So a stochastic description of the deterministic system (4.5) is given by the random variable $X(t) = (x(t), y(t), z(t))^T$, where

$$\begin{aligned}X(t) &= \begin{pmatrix} x_0 \\ y_0 \\ v_0 \end{pmatrix} + \mathcal{P}_1 \left(\int_0^t \lambda ds \right) \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} + \mathcal{P}_2 \left(\int_0^t hx ds \right) \begin{pmatrix} -1 \\ 0 \\ 0 \end{pmatrix} + \mathcal{P}_3 \left(\int_0^t \beta xy ds \right) \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix} \\ &+ \mathcal{P}_4 \left(\int_0^t ay ds \right) \begin{pmatrix} 0 \\ -1 \\ 0 \end{pmatrix} + \mathcal{P}_5 \left(\int_0^t ky ds \right) \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} + \mathcal{P}_6 \left(\int_0^t uv ds \right) \begin{pmatrix} 0 \\ 0 \\ -1 \end{pmatrix} \quad (4.7)\end{aligned}$$

4.2.2. Relationship Between ODEs and Expected Value of Stochastic System

The stochastic process and the deterministic ODE model are related in the following way. If stochastic effects can be ignored, the population can be described with a deterministic process $z(t) \in \mathbb{N}_0^S$ given by [175, 176]

$$z(t) = z_0 + \sum_{r=1}^R \left(\int_0^t \alpha_r(z(s)) ds \right) \nu_r \quad (4.8)$$

and we derive

$$\frac{d}{dt} [z(t)] = \frac{d}{dt} \left[z_0 + \sum_{r=1}^R \left(\int_0^t \alpha_r(z(s)) ds \right) \nu_r \right] = \sum_{r=1}^R \alpha_r(z(s)) \nu_r \quad (4.9)$$

Note the similarity between Equation (4.8) and Equation (4.4). Taking the time-derivative of the expected value of Equation (4.7) in the above HIV infection example, we have

$$z(\dot{t}) = \begin{pmatrix} \dot{x} \\ \dot{y} \\ \dot{v} \end{pmatrix} = \lambda \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} + hx \begin{pmatrix} -1 \\ 0 \\ 0 \end{pmatrix} + \beta xy \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix} + ay \begin{pmatrix} 0 \\ -1 \\ 0 \end{pmatrix} + ky \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} + uv \begin{pmatrix} 0 \\ 0 \\ -1 \end{pmatrix} \quad (4.10)$$

which is equal to the original system of ODEs in (4.5)

4.3. Master Equation: Mass-Action Rate Birth-Death Process

The probability distribution associated to the random variable $X(t) \in \mathbb{N}_0^S$ in (4.4) is given by $\mathbb{P}(X; t) = \mathbb{P}(X(t) = x | X(0) = x_0)$. Here $\mathbb{P}(X; t)$ is a column vector where each entry is a probability associated to a potential state of the random variable. It can be shown [177, 178, 179] that $\mathbb{P}(X; t)$ the solution to the *Chemical Master Equation* (CME)

$$\frac{\partial \mathbb{P}(X = x; t)}{\partial t} = \sum_{r=1}^R (\alpha_r(x - \nu_r) \mathbb{P}(X = x - \nu_r; t) - \alpha_r(x) \mathbb{P}(X = x; t)) \quad (4.11)$$

We can interpret Equation (4.11) as follows: the rate of change in the probability of the state taking value $X(t) = x$ is the sum over all reactions r of the probability of arriving at state $X(t) = x$ via reaction r (having previously been in state $X(t) = x - \nu_r$) minus the probability of leaving state $X(t) = x$ via reaction r . We explicitly derive the master equation for a mass-action birth-death process of a single species, the properties of which we will explore later in this chapter.

Consider a species $\mathcal{X}$ that can take values $x \in \mathbb{N}_0$, and where individuals of the species can proliferate and die respectively with per-capita rates π and δ . In this case the state vector is $X(t) \in \mathbb{N}_0$, and a “state” of the system corresponds to a “frequency” of the species (we use both terms interchangeably). We assume that proliferation and death occur without contact between individuals, so we have mass-action propensities $\alpha_1(X(t) = x) = \pi x$ and $\alpha_2(X(t) = x) = \delta x$, where $X(t)$ is the random variable associated with the frequency of species $\mathcal{X}$ at time t . We have the following reactions



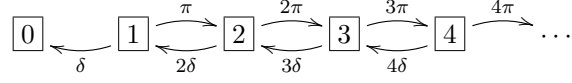


Figure 4.2.: **Mass action birth-death process flow diagram.** The potential states (in this case frequencies) the species can take are shown, along with the propensity of each reaction at each state. Note that there is no upper bound of species frequency, and there is no source inflow from $\boxed{0}$ to $\boxed{1}$.

The dynamics of this birth-death process are summarised in Figure 4.2. What is the probability $\mathbb{P}(X(t) = x)$ that species $\mathcal{X}$ takes frequency x at time t ?

Consider a time interval Δt , small enough such that at most a single reaction can take place in the interval $[t, t + \Delta t]$. Species $\mathcal{X}$ can thus take frequency $x \in \mathbb{N}$ at time $t + \Delta t$ only when it was in one of three states at time t : i) $X(t) = x$; ii) $X(t) = x - 1$; iii) $X(t) = x + 1$, that is: i) if nothing happened; ii) if an individual proliferated; or iii) if an individual died. Hence

$$\begin{aligned} \mathbb{P}(X = x; t + \Delta t) = & \mathbb{P}(X = x; t + \Delta t | X = x; t) \times \mathbb{P}(X = x; t) \\ & + \mathbb{P}(X = x; t + \Delta t | X = x - 1; t) \times \mathbb{P}(X = x - 1; t) \\ & + \mathbb{P}(X = x; t + \Delta t | X = x + 1; t) \times \mathbb{P}(X = x + 1; t) \end{aligned} \quad (4.13)$$

We can rewrite Equation (4.13) informally as

$$\begin{aligned} \mathbb{P}(X = x; t + \Delta t) = & \mathbb{P}(\text{NOTHING from } \boxed{x}) \times \mathbb{P}(X = x; t) \\ & + \mathbb{P}(\text{BIRTH from } \boxed{x - 1}) \times \mathbb{P}(X = x - 1; t) \\ & + \mathbb{P}(\text{DEATH from } \boxed{x + 1}) \times \mathbb{P}(X = x + 1; t) \end{aligned} \quad (4.14)$$

We make use of the following property,

$$\mathbb{P}(X = y; t + \Delta t | X = x; t) = \mathbb{P}(X = y; \Delta t | X = x; 0) \quad (4.15)$$

for any $x, y \in \mathbb{N}_0$. This property states that the probability of an event is constant over time. Consider the probability mass function f_{Pois} of the Poisson distribution. The probability $f_{\text{Pois}}(k; \lambda)$ of k events occurring at an average rate of λ is given by

$$f_{\text{Pois}}(k; \lambda) = \frac{e^{-\lambda} \lambda^k}{k!} \quad (4.16)$$

The expected number of births from state $\boxed{x}$ in the interval $[0, \Delta t]$ is given by $\int_0^{\Delta t} \pi x dt = \pi x \Delta t$ (note that here x is a constant). Hence our Poisson process describing k births from

state $\boxed{x}$ in time interval Δt becomes

$$f_{\text{Pois}}(k; \Delta t | \pi x) = \frac{e^{-\pi x \Delta t} (\pi x \Delta t)^k}{k!} \quad (4.17)$$

So the probability of there being zero births from state $\boxed{x}$ in $[t, t + \Delta t]$ is given by

$$\mathbb{P}(\text{No births from } \boxed{x}) = \frac{e^{-\pi x \Delta t} (\pi x \Delta t)^0}{0!} = e^{-\pi x \Delta t} = 1 - \pi x \Delta t + \mathcal{O}((\Delta t)^2) \quad (4.18)$$

Similarly we have,

$$\mathbb{P}(\text{No deaths from } \boxed{x}) = 1 - \delta x \Delta t + \mathcal{O}((\Delta t)^2) \quad (4.19)$$

$$\mathbb{P}(\text{1 birth from } \boxed{x-1}) = (x-1)\pi \Delta t + \mathcal{O}((\Delta t)^2) \quad (4.20)$$

$$\mathbb{P}(\text{1 death from } \boxed{x+1}) = (x+1)\delta \Delta t + \mathcal{O}((\Delta t)^2) \quad (4.21)$$

Combining equations (4.18) and (4.19), we obtain

$$\begin{aligned} \mathbb{P}(\text{NOTHING from } \boxed{x}) &= \mathbb{P}(\text{No births from } \boxed{x}) \times \mathbb{P}(\text{No deaths from } \boxed{x}) \\ &= 1 - (\pi + \delta)x \Delta t + \mathcal{O}((\Delta t)^2) \end{aligned} \quad (4.22)$$

so Equation (4.13) becomes

$$\begin{aligned} \mathbb{P}(X = x; t + \Delta t) &= [1 - (\pi + \delta)x \Delta t] \times \mathbb{P}(X = x; t) \\ &\quad + [(x-1)\pi \Delta t] \times \mathbb{P}(X = x-1; t) \\ &\quad + [(x+1)\delta \Delta t] \times \mathbb{P}(X = x+1; t) \\ &\quad + \mathcal{O}((\Delta t)^2) \end{aligned} \quad (4.23)$$

Hence

$$\begin{aligned} \frac{\mathbb{P}(X = x; t + \Delta t) - \mathbb{P}(X = x; t)}{\Delta t} &= -x(\pi + \delta) \times \mathbb{P}(X = x; t) \\ &\quad + (x-1)\pi \times \mathbb{P}(X = x-1; t) \\ &\quad + (x+1)\delta \times \mathbb{P}(X = x+1; t) \\ &\quad + \mathcal{O}((\Delta t)) \end{aligned} \quad (4.24)$$

Letting $\Delta t \rightarrow 0$, we obtain the master equation of our birth-death process

$$\begin{aligned} \frac{\partial (\mathbb{P}(X = x; t))}{\partial t} &= -x(\pi + \delta) \times \mathbb{P}(X = x; t) \\ &\quad + (x-1)\pi \times \mathbb{P}(X = x-1; t) \\ &\quad + (x+1)\delta \times \mathbb{P}(X = x+1; t) \end{aligned} \quad (4.25)$$

The following hold:

1. $\sum_i A_{ij} = 0 \quad \forall j$
2. $A_{ij} \geq 0 \quad \forall i \neq j$
3. $A_{ij} \leq 0 \quad \forall i = j$ and
4. $A_{ii} = -\sum_{j \neq i} A_{ij}$

Note that the species can “bounce” from the upper bound $\boxed{\tau}$ back to $\boxed{\tau - 1}$, but cannot escape from $\boxed{0}$ to $\boxed{1}$. The transition matrix (sometimes named a stochastic matrix, the “state companion matrix” in control theory [181], or the “infinitesimal generator of a continuous-time Markov chain (CTMC)” [182]) gives the reaction rates that can occur between all states of the system in a single infinitesimal step Δt . For example, if reaction $\mathcal{R}_2$ occurs when the system is at state $\boxed{3}$, the systems will move to state $\boxed{2}$, and so $A_{32} = 3\delta$. For $i \neq j$, the propensity of a transition from state j to state i in k infinitesimal steps is given by $(A^k)_{ij}$. For $i = j$, A_{ij} can be interpreted as the rate of change of probability from a given state i to any other state j .

In our system above, A does not have any time-dependent factors, and so Equation (4.28) has solution

$$\mathbb{P}(X; t) = e^{At} \mathbb{P}_0 \quad (4.30)$$

where $\mathbb{P}_0 = \mathbb{P}(X; 0)$ is the *initial probability distribution* and e^{At} is the *matrix exponential* [182].

4.3.2. Initial Probability Distribution

The initial probability distribution $P_0 \in \mathbb{R}^\tau$ [182] can be used to enter a fixed initial value, in which case we have for initial value $X(0) = x_0$ and $x \in \{0, \dots, \tau\}$

$$\mathbb{P}_0 = \delta_{x, x_0} \quad (4.31)$$

where δ_{x, x_0} [183] is the Kronecker delta such that [183]

$$\delta_{x, x_0} = \begin{cases} 1, & \text{if } x = x_0 \\ 0, & \text{if } x \neq x_0 \end{cases} \quad (4.32)$$

However, P_0 does not have to be restricted to a fixed (deterministic) initial value in a stochastic system. Any discrete probability distribution on the state space (in our case

$\{0, \dots, \tau\}$) is acceptable, and can be used to model the uncertainty of the initial state of the system. For example, in the fixed case where $x_0 = 2$, or when we are uncertain over the input, we may have respectively

$$\mathbb{P}_0 = \begin{pmatrix} \boxed{0} \\ \boxed{1} \\ \boxed{2} \\ \boxed{3} \\ \vdots \\ \boxed{\tau} \end{pmatrix} \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \quad \text{and} \quad \mathbb{P}_0 = \begin{pmatrix} \boxed{1} \\ \boxed{2} \\ \boxed{3} \\ \boxed{4} \\ \vdots \\ \boxed{\tau} \end{pmatrix} \begin{pmatrix} 0.3 \\ 0.5 \\ 0.1 \\ 0.05 \\ \vdots \\ 0.00001 \end{pmatrix} \quad (4.33)$$

4.3.3. Matrix Exponential

The matrix exponential is defined as follows. If $M \in \mathbb{C}$ is an $m \times m$ square matrix $m \in \mathbb{N}$, then the matrix exponential is given by

$$\exp(M) = e^M = \sum_{k=0}^{\infty} \frac{1}{k!} M^k \quad (4.34)$$

The quantity M^k is always defined since M is finite and square. Intuitively, e^{At} denotes the probability of starting at state j and finishing at state i in the time interval $[0, t]$, and so we have

$$\sum_i^{\tau} (e^{At})_{ij} = 1 \quad \forall j \in \{1, \dots, \tau\} \quad (4.35a)$$

$$0 \leq (e^{At})_{ij} \leq 1 \quad \forall i, j \in \{1, \dots, \tau\} \quad (4.35b)$$

For $a, b \in \mathbb{C}$, we have

$$e^0 = I_m \quad (4.36a)$$

$$e^{aM+bM} = e^{aM} e^{bM} \quad (4.36b)$$

where I_m is the $m \times m$ identity matrix defined as

$$(I_m)_{ij} = \delta_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases} \quad (4.37)$$

Equation (4.36a) can be reconciled with the intuitive interpretation of e^{At} . If $t = 0$, then $e^{At} = e^0 = I_{\tau+1} = \delta_{ij}$. That is, after zero time has elapsed, the probability of starting at

state i and remaining at state i is 1, and the probability of moving to any other state $j \neq i$ is 0. In our example (Figure 4.3), indexing state 1 as the state $\boxed{0}$, we have $(e^{At})_{11} = 1$ and $(e^{At})_{i1} = 0 \quad \forall i \neq 1$ because a species cannot proliferate once it is extinct and we do not have a source or inflow, i.e. there is no flow out of state $\boxed{0}$ to any other state.

Computing the matrix exponential is straightforward in principle, but considerably more difficult in practice when the dimension of the M is large [184]. Because of the role of the matrix exponential in solving linear differential equations, finding reliable and efficient approximations of the calculation is an ongoing area of research [181, 184, 185, 186]. Compounding this problem, the probability distribution is often desired for multiple time points (i.e. to obtain the evolution of the probability distribution). Fortunately however, the probability distribution for a series of time points can be calculated recursively using a single matrix exponential. More precisely, if the propensities do not change with time, for a given time step s , we can compute the matrix exponential e^{As} , and then recursively calculate the probability distribution at any time that is a multiple of s . Using Equation (4.36b), for $k \in \mathbb{N}_0$, we have the following recurrence relation:

$$\mathbb{P}(X = x; t = s) = e^{As} P_0 \quad (4.38a)$$

$$\mathbb{P}(X = x; t = ks) = e^{As} \mathbb{P}(X = x; t = (k-1)s) \quad (4.38b)$$

In long form, we have

$$\mathbb{P}(X = x; t = ks) = e^{Aks} P_0 = (e^{As})^k P_0 = e^{As} \mathbb{P}(X = x; t = (k-1)s) \quad (4.39)$$

This is important in that it means the computationally expensive [184] matrix exponential e^{As} only needs to be calculated once, and then recursively multiplied k times to P_0 . While k may be large, for our purposes matrix multiplication k times incurs significantly less runtime than calculating the matrix exponential k times.

In the following section we explore novel consequences of the birth-death process described in Figure 4.3 for a range of proliferation and death parameters. We examine potential approximations to $\mathbb{P}(X = x; t)$ that can save substantial runtime without distorting the probability distribution. It is important to note that, in contrast to other birth-death processes examined [173, 174, 187, 188] there is no source inflow from $\boxed{0}$ to $\boxed{1}$ in our birth-death process. That is, we do not have a reaction $\mathcal{R}_j$ such that

$$\mathcal{R}_j : * \xrightarrow{\lambda} \mathcal{X} \quad (4.40)$$

like there is in, for example, reaction $\mathcal{R}_1$ (4.6a) of the HIV example. In many situations,

it does not make biological sense for there to be a constant inflow into the species. For example a species cannot proliferate once it has become extinct.

In our simulations, we use the `expm` function in the R package `matrix` [97, 189] to calculate the matrix exponential, which uses the Padé approximation.

4.4. Properties of Birth-Death Process Without Source Inflow

4.4.1. Evolution of Extinction Probability

The master equation can explicitly model the probability that a given species takes frequency or state $\boxed{0}$ at time t . That is, we can calculate the probability of extinction as a function of time. This would not be possible using an ODE to model a birth-death process described in Figure 4.3. If a given species has a higher death rate than proliferation rate, we would intuitively expect the species to eventually become extinct. In such an ODE model, the predicted frequency of the species will decrease to an asymptote of 0, but will never actually reach this value. We model the evolution of the extinction probability using a range of values of proliferation π , death δ and starting frequency x_0 (Figure 4.4).

Four parameter sets $\theta_k = \{\pi_k, \delta_k\}$, $k \in \{1, 2, 3, 4\}$ were randomly chosen from a uniform distribution $\mathcal{U}(0, 1)$. For each parameter set, two starting frequencies x_0 were randomly chosen from a uniform distribution $\mathcal{U}(1, \tau/3)$, where $\tau = 500$ and $\tau/3$ was chosen so that the boundary τ would not adversely affect results (see Section 4.4.2). In all cases, the extinction probability at time t , $\mathbb{P}(X = 0; t) = \mathbb{P}(Ext; t)$ was strictly increasing and approached an asymptote (Figure 4.4). We define $\mathbb{P}(Ext)^*$ to be the asymptote of the extinction probability, and this value is determined by the interplay between the values of π , δ and x_0 . By inspection, based on our numerical simulations, the value of the asymptote appears to be given by the function

$$\mathbb{P}(Ext)^* = h(\pi, \delta, x_0) = \min \left\{ \left(\frac{\delta}{\pi} \right)^{x_0}, 1 \right\} \quad (4.41)$$

We tested this prediction using the randomly chosen parameters and starting values above. In all cases, $h(\pi, \delta, x_0)$ correctly predicted the value of the asymptote (Figure 4.4).

Because there is no flow from state $\boxed{0}$ to any other state, the probability of extinction must increase over time. While this makes sense, it can be counterintuitive in certain situations. If we consider a growing species (i.e. $\pi > \delta$) that starts from a relatively high frequency

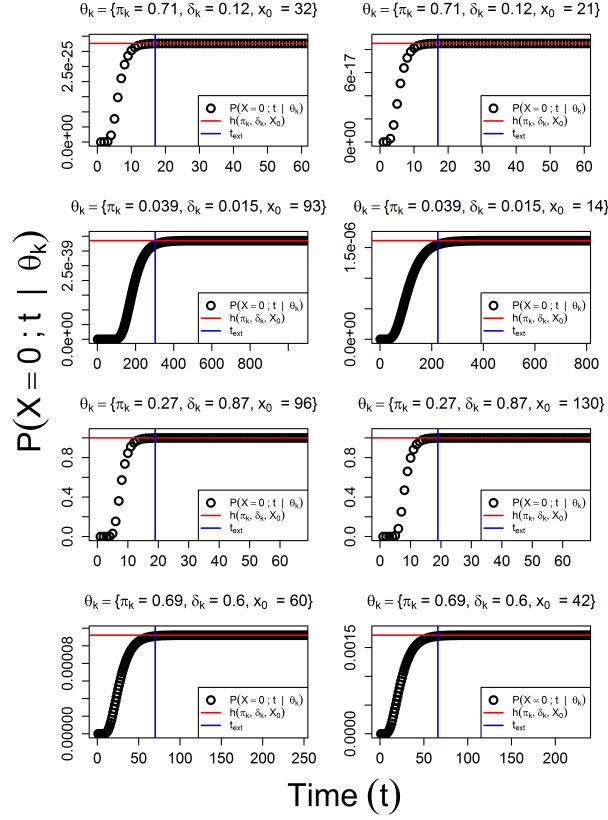


Figure 4.4.: **Evolution of Extinction Probabilities for range of parameters.** All plots show a strictly increasing extinction probability over time, eventually approaching an asymptote (red line), given in Equation (4.41). The vertical blue line denotes the minimum time t_{ext} such that $(\mathbb{P}(X = 0; t_{ext}) - \mathbb{P}(X = 0; t_{ext} - 1)) / \mathbb{P}(X = 0; t_{ext} - 1) < 1/1000$.

x_0 , a deterministic formulation would suggest that as the species frequency increases, the probability of extinction decreases. However, when we consider the stochasticity of the system, it is simply that the *expected frequency* increases.

If we again index state $\boxed{0}$ as state number 1, we have for $x \in \Omega$, $\mathbb{P}(X = 0; t) = \mathbb{P}(X = x; t)_{11}$, i.e. the first value in the $m \times 1$ state probability vector. Then by Equation (4.38),

$$\mathbb{P}(X = x; t + \Delta t)_{11} = \sum_{j=1}^{\tau} (e^{A\Delta t})_{1j} \times \mathbb{P}(X = x; t)_{j1} \quad (4.42)$$

$$\mathbb{P}(X = x; t + \Delta t)_{11} = \left((e^{A\Delta t})_{11} \times \mathbb{P}(X = x; t)_{11} \right) + \left(\sum_{j=2}^{\tau} (e^{A\Delta t})_{1j} \times \mathbb{P}(X = x; t)_{j1} \right) \quad (4.43)$$

Since $(e^{A\Delta t})_{11} = 1$, (i.e. remaining at state $\boxed{0}$ having started at state $\boxed{0}$ is certain for all $t \geq 0$), and since $(e^{A\Delta t})_{1j} \geq 0$ and $\mathbb{P}(X = x; t)_{j1} \geq 0$, we have

$$\mathbb{P}(X = x; t + \Delta t)_{11} \geq \left(\mathbb{P}(X = x; t)_{11} \right) \quad (4.44)$$

Subtracting $\mathbb{P}(X = x; t)_{11}$ from both sides and dividing by $\Delta t > 0$, we have

$$\frac{\mathbb{P}(X = x; t + \Delta t)_{11} - \mathbb{P}(X = x; t)_{11}}{\Delta t} \geq 0 \quad (4.45)$$

Letting $\Delta t \rightarrow 0$ gives

$$\frac{d\mathbb{P}(Ext; t)}{dt} \geq 0 \quad (4.46)$$

This does not hold for birth-death processes where there is a non-zero source flow from $\boxed{0}$ to $\boxed{1}$, as then $(e^{A\Delta t})_{11} \neq 1$.

4.4.2. Effects of Truncated State Space

Under the birth-death process model given in Figure 4.2, given a set of growing parameters (i.e. $\pi > \delta$), there is no upper limit on the frequency that a species can take. However, to make the problem computationally tractable (and to make the master equation solvable using the matrix exponential), we must impose an upper limit τ and truncate the infinite state space $\Omega = \mathbb{N}_0$ to the reduced state space $\Omega_\tau = \{0, \dots, \tau\}$ as in Figure 4.3. Ideally τ is chosen so that the probability of the species taking frequency greater than τ is negligible [150], i.e. so that $\mathbb{P}(X \geq \tau; t) \approx 0$. For example, if $x_0 = \tau - 1$ and $\pi > \delta$, a significant amount of probability mass will build up at $X = \tau$. Clearly this is the result of truncating the state space, and not because of any underlying biological behaviour we wish to examine (Figure 4.5).

4.4.3. Shape of Probability Distribution $\mathbb{P}(X; t)$ for Large Starting Values x_0

Figures 4.6, 4.7, and 4.8 show the the probability distribution evolution for growing ($\pi > \delta$), shrinking ($\pi < \delta$), and “equilibrium” ($\pi = \delta$) parameters. In order to examine the form of $\mathbb{P}(X; t)$ when it is not distorted by the use of a truncated state space, large values of τ were chosen with starting frequencies $x_0 = \tau/2$. In the growing and shrinking cases, $\tau = 10^4$, and in the equilibrium case $\tau = 1.5 \times 10^4$. Both of these values require substantial computational time to calculate a single matrix exponential (approximately 7 and 24 hours

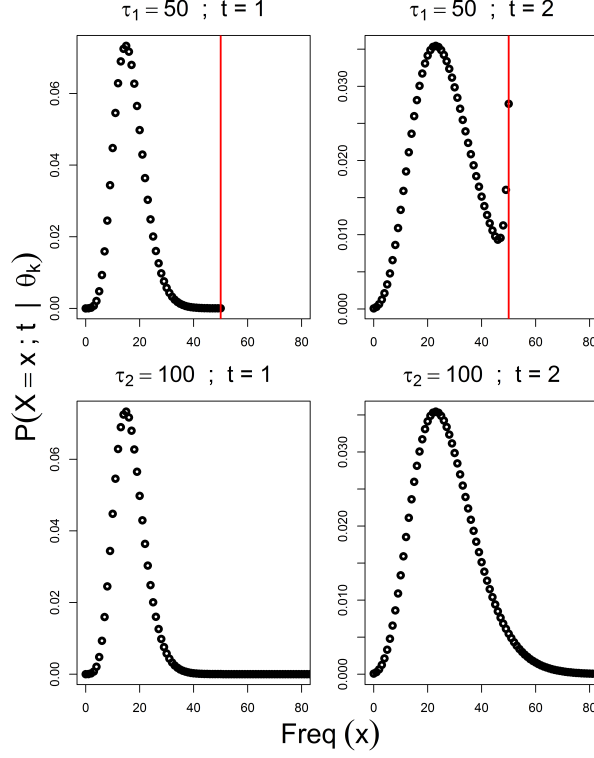


Figure 4.5.: **Effect of Truncated State Space Ω .** A growing species (i.e. $\pi > \delta$) is shown. In the first row, $\tau = 50$ is not large enough to prevent the probability distribution being distorted in the time frame $t \in \{1, 2\}$. In the second row, $\tau = 100$ appears large enough so that $\mathbb{P}(X; t)$ not significantly distorted. In each case, $\theta_k = \{\pi = 1, \delta = 0.5, x_0 = 10\}$. In first row, value of boundary τ shown in red.

for $\tau = 10^4$ for $\tau = 1.5 \times 10^4$ respectively) making them impractical for parameter fitting where many parameters must be tested and many matrix exponentials calculated.

In all cases, the form of each distribution $\mathbb{P}(X; t | \theta_k)$ strongly resembled a Gaussian distribution, and we assessed their similarity using the Kullback-Leibler divergence. For two discrete probability distributions P and Q , the Kullback-Leibler (KL) divergence of Q from P is defined by

$$D_{KL}(P||Q) = \sum_i P(i) \log \left(\frac{P(i)}{Q(i)} \right) \quad (4.47)$$

The KL-divergence is a measure of the loss of information when P is approximated by Q [109, 118]. $D_{KL}(P||Q) = 0$ if and only if $P = Q$, i.e. zero information is lost if P is approximated by itself. Note that in general, $D_{KL}(P||Q) \neq D_{KL}(Q||P)$.

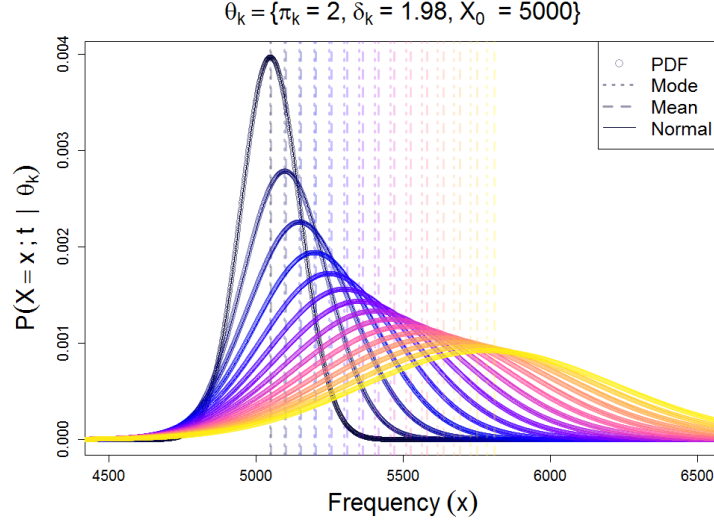


Figure 4.6.: **Evolution of probability distribution given “growth” parameters** $2 = \pi < \delta = 1.98$. $\mathbb{P}(X;t)$ is plotted for multiple values of t (circles). Later timepoints are represented by lighter colours. In all timepoints, the mean (dashed vertical lines) is greater than the mode (dotted vertical lines), and the discrepancy between these two values increases over time, as does the variance of each PDF. A normal distribution (solid lines) $\mathcal{N}(\mu(t), \sigma(t)^2)$ is plotted, and is an excellent fit to each $\mathbb{P}(X;t)$.

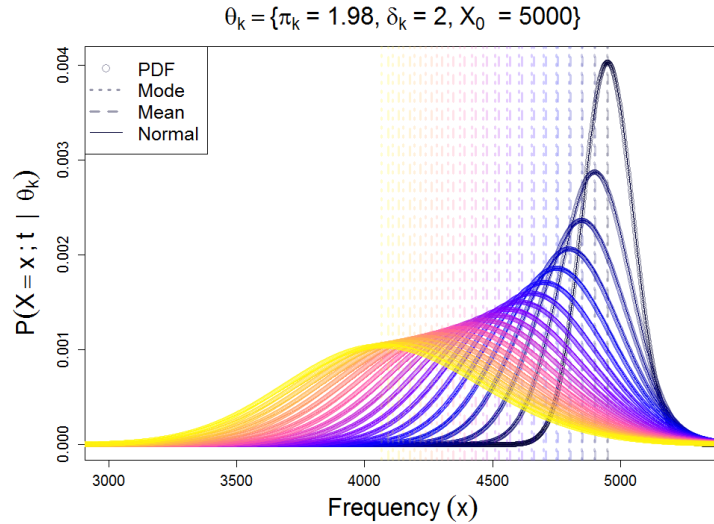


Figure 4.7.: **Evolution of probability distribution given “shrinking” parameters** $1.98 = \pi < \delta = 2$. As for Figure 4.6.

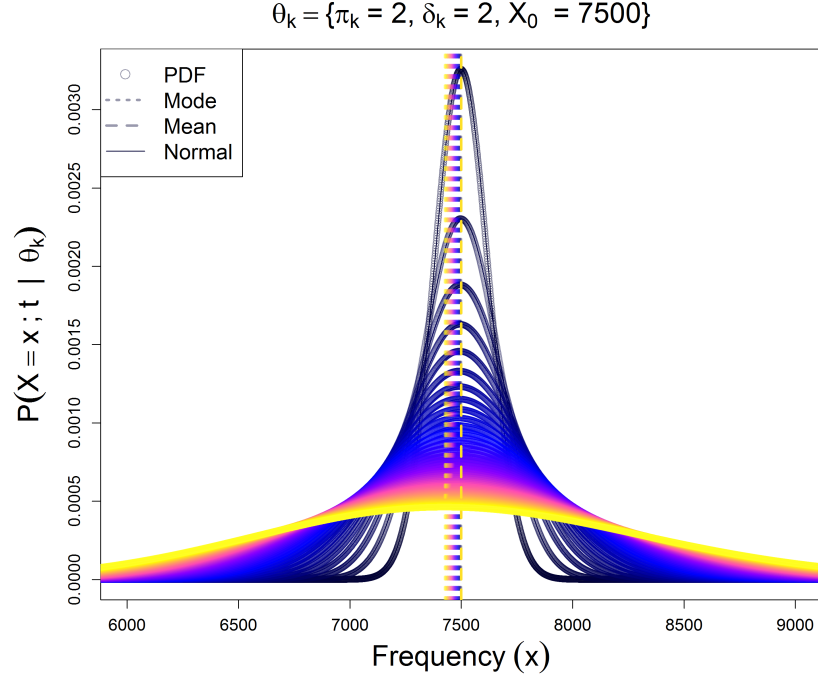


Figure 4.8.: **Evolution of probability distribution given “equilibrium” parameters** $\pi = \delta = 2$. As for Figures 4.6 and 4.7. Though proliferation and death rates are equal, the distribution moves left over time. This is most visible for the mode, but the mean also moves left to a small extent. This is unlikely to be due to the choice of $\tau = 15000$, as the $x_0 = 7500$ is suitably far away.

Table 4.1.: KL-divergence between $\mathbb{P}(X; t | \theta_k)$ and reference normal distributions: large x_0

θ_k	$D_{KL}(\mathbb{P}(X; t) \mathcal{N}(\mu(t), \sigma(t)^2))^{\dagger}$
Growing ($\pi > \delta, \tau = 10^4$)	0.0002
Shrinking ($\pi < \delta, \tau = 10^4$)	0.0003
Equilibrium ($\pi = \delta, \tau = 1.5 \times 10^4$)	0.0003

[†] Highest Kullback-Leibler divergence between $\mathbb{P}(X; t)$ and reference normal distributions $\mathcal{N}(\mu(t), \sigma(t)^2)$ for all time points.

$D_{KL}(\mathbb{P}(X; t) || \mathcal{N}(\mu(t), \sigma(t)^2))$ was calculated to compare each $\mathbb{P}(X; t)$ to its corresponding reference normal distribution $\mathcal{N}(\mu(t), \sigma(t)^2)$. Reference normal distributions were calculated using expectations $\mu(t) = \mathbb{E}[\mathbb{P}(X; t)]$ and variances $\sigma(t)^2 = \text{Var}[\mathbb{P}(X; t)]$. Figures 4.6, 4.7, and 4.8 show $\mathbb{P}(X; t | \theta_k)$, together with each $\mathcal{N}(\mu(t), \sigma(t)^2)$ for multiple t values. The reference normal distributions are an excellent fit to each $\mathbb{P}(X; t)$, and the values of $D_{KL}(P || Q) \neq D_{KL}(Q || P)$ are low (see Figures 4.6, 4.7, and 4.8 and Table 4.1).

For each parameter set, the variance of $\mathbb{P}(X; t)$ increases over time, reflecting the increased

uncertainty that arises as the system is propagated. Further, for each parameter set tested, the mode is less than the expected value, indicating that the Gaussian characterisation of $\mathbb{P}(X; t)$ remains an approximation, albeit an acceptable one. This effect has been observed previously [148], and is particularly curious when we consider the “equilibrium case” when $\pi = \delta = 2$. Under a deterministic framework, equal proliferation and death rates would result in zero change to the system. If a set of ordinary differential equations are used to approximate the expected value of the system (or indeed the entire system), as in Section 4.2.2, this will overestimate the expected value as the “sinkhole” effect of the absorbing state $\boxed{0}$ will not be considered. While this does not result in large practical differences in our system within the time-frame we have modelled, it does change our view of the system: the stochastic formulation of equilibrium will eventually result in extinction; the deterministic formulation will not.

The strictly increasing extinction probability accounts for the decreasing value of the expectation. Clearly, as $\mathbb{P}(X = 0; t)$ increases, $\mathbb{P}(X \neq 0; t)$ decreases, and so when $\pi = \delta$ there is less probability mass contributing to non-zero values of the expectation summation. However this decrease is very small, and it remains the case that approximating stochastic systems with ODEs renders the modelling of many biological systems tractable. The fact that the entire probability distribution, and not just the expected value moves to the left (Figure 4.8) is an interesting feature of the stochastic system.

4.4.4. Shape of Probability Distribution $\mathbb{P}(X; t)$ for $x_0 = 1$

The Gaussian-like distributions we observe when given large starting values x_0 do not appear when $x_0 = 1$. Instead, the plots closely resemble exponential distributions (Figure 4.9). We now examine the extent to which $\mathbb{P}(X; t | x_0 = 1)$ can be approximated by an exponential distribution. Because the exponential distribution can be fully described by a single parameter λ , such an approximation could save substantial computational time.

The exponential distribution has probability density function given by

$$f(x; \lambda) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases} \quad (4.48)$$

Taking the natural logarithm gives for $x \geq 0$,

$$\begin{aligned} F(x; \lambda) &= \log(f(x; \lambda)) = \log(\lambda e^{-\lambda x}) \\ &= \log(\lambda) + \log(e^{-\lambda x}) \\ &= \log(\lambda) - \lambda x \end{aligned} \quad (4.49)$$

$$\theta_1 = \{\pi_1 = 0.5, \delta_1 = 0.1, x_0 = 1\} \quad \theta_2 = \{\pi_2 = 0.4, \delta_2 = 0.6, x_0 = 1\}$$

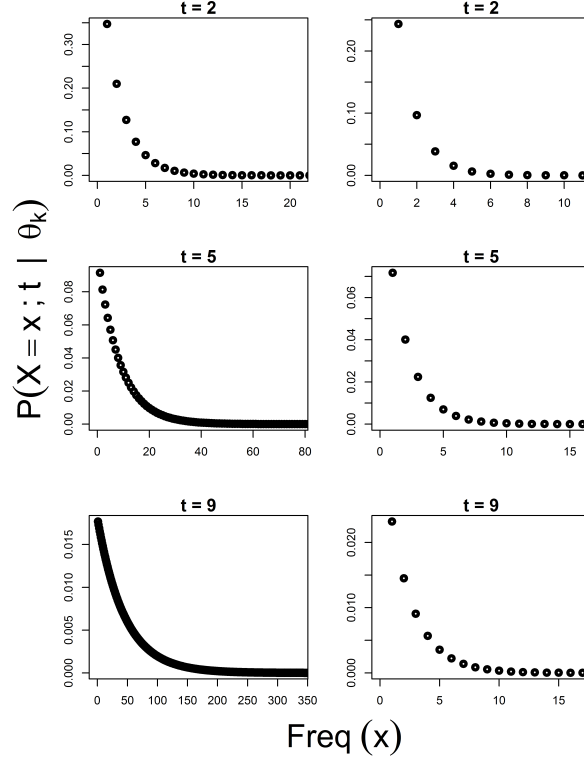


Figure 4.9.: **Example PDF evolutions** $\mathbb{P}(X; t | x_0 = 1)$. For example parameter sets θ_1 and θ_2 $\mathbb{P}(X; t | x_0 = 1)$, $\mathbb{P}(X; t | x_0 = 1)$ resembles an exponential distribution for all time points. Note that the extinction probability $\mathbb{P}(X = 0; t | x_0 = 1)$ is not shown.

Because $F(x; \lambda)$ is linear with intercept $\log(\lambda)$ and slope $-\lambda$, we can take the natural log of $\mathbb{P}(X; t | x_0 = 1)$, and use linear regression to determine whether an exponential distribution with rate λ (and expectation $1/\lambda$) is an acceptable approximation of $\mathbb{P}(X; t | x_0 = 1)$. That is, we regress using the function

$$Y = aX + b \tag{4.50}$$

where $Y = \log(\mathbb{P}(X; t | x_0 = 1))$.

Figure 4.10 shows a series of probability distributions $\mathbb{P}(X; t | x_0 = 1)$ for parameter sets $\theta_1 = \{\pi_1 = 2, \delta_1 = 1.98\}$ and $\theta_2 = \{\pi_2 = 1.98, \delta_2 = 2\}$, both when plotted on a logarithmic scale and not. When plotted on the log scale, $\mathbb{P}(X; t | x_0 = 1)$ is strikingly linear. Running a linear regression on $\log(\mathbb{P}(X; t | x_0 = 1))$ for $t \in \{0.5, 1, 1.5, \dots, 19.5, 20\}$, the minimum value of the coefficient of determination R^2 was greater than 0.9999 for each θ_k , indicating that the log-linear model proposed in Equation (4.50) is an excellent fit to $\mathbb{P}(X; t | x_0 =$

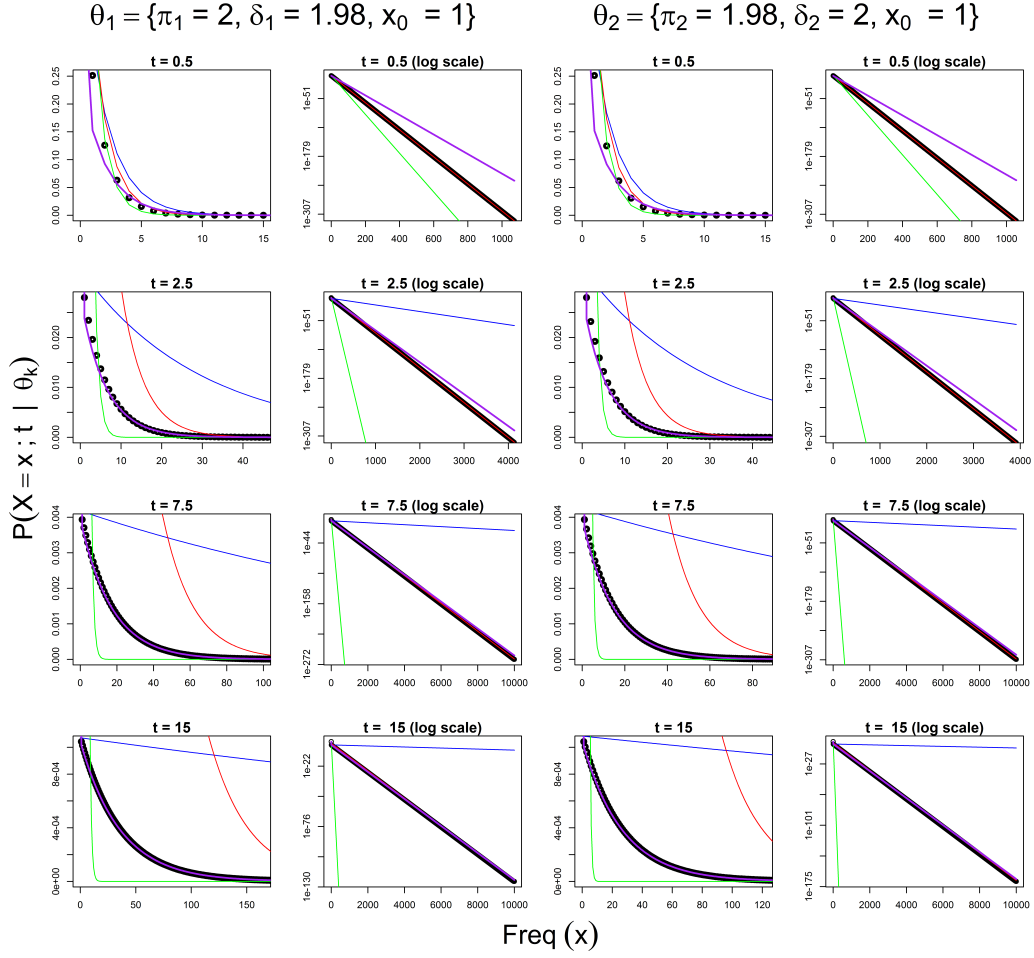


Figure 4.10.: **The exponential distribution approximates $\mathbb{P}(X; t | x_0 = 1)$.** PDF progressions for two example parameter sets θ_1 (leftmost two columns) and θ_2 (rightmost two columns) and four time points $t \in \{0.5, 2.5, 7.5, 15\}$ (rows) are shown (black dots). For each time point and θ_k , the PDF is plotted on different scales: one magnified to the non-negligible probability mass for all states except $\boxed{0}$ (first and third column); and one plotted on the log scale (second and fourth columns). In each plot, four exponential distributions with different rate parameters $(\lambda_i)_{i=1}^4$ are plotted: $\hat{\lambda}_1 = -a$ (red); $\hat{\lambda}_2 = e^b$ (blue); $\hat{\lambda}_3 = 1/\mathbb{E}[\mathbb{P}(X; t | x_0 = 1)]$ (green); $\hat{\lambda}_4 = 1/\mathbb{E}[\mathbb{P}_\varphi(X; t | x_0 = 1)]$ (purple). Purple distributions give best fit to $\mathbb{P}(X; t | x_0 = 1)$. See Section 4.4.4 for further details.

$1 | \theta_k), k = \{1, 2\}$.

However, simply because a distribution is log-linear does not mean that it is exponential. If each $\mathbb{P}(X; t | x_0 = 1)$ was a perfect exponential distribution, then Equation (4.49) implies that both the intercept and the slope of the line fitted to the natural log of $\mathbb{P}(X; t | x_0 = 1)$

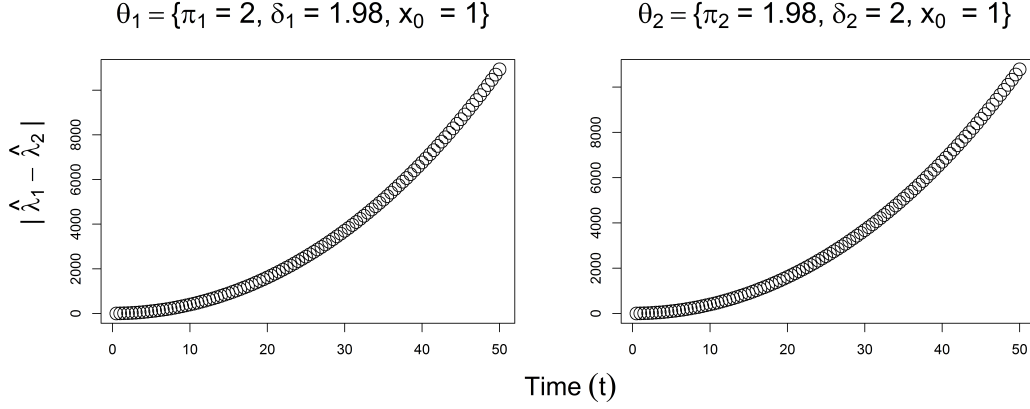


Figure 4.11.: **Discrepancy between exponential distribution rate parameters $\hat{\lambda}_1$ and $\hat{\lambda}_2$ over time.** $\hat{\lambda}_1 = -a$, the slope of $\log(\mathbb{P}(X; t | x_0 = 1))$ and $\hat{\lambda}_2 = e^b$, derived from the intercept of $\log(\mathbb{P}(X; t | x_0 = 1))$ (Equation (4.50)) are initially similar, but diverge over time.

could be used to give an equivalent estimate of the rate λ and expectation $1/\lambda$. That is, $\hat{\lambda}_1 = -a$ should be equal to $\hat{\lambda}_2 = e^b$. As is shown in Figure 4.11, while $\hat{\lambda}_1$ is approximately equal to $\hat{\lambda}_2$ at early time points, they diverge substantially thereafter. Indeed, while $\hat{\lambda}_1$ is an excellent fit to the distribution on the log scale, it does not approximate the distribution well on the left-hand side of the plots. Since this is where the bulk of the probability mass lies, neither $\text{Exp}(\hat{\lambda}_1)$ nor $\text{Exp}(\hat{\lambda}_2)$ are acceptable approximations of $\mathbb{P}(X; t | x_0 = 1)$.

It is possible that the discrepancy between $\hat{\lambda}_1$ and $\hat{\lambda}_2$ and the inadequacy of $\text{Exp}(\hat{\lambda}_1)$ and $\text{Exp}(\hat{\lambda}_2)$ is due to the distorting effects of the absorbing state $\boxed{0}$ and extinction probability. From Figure 4.10, we can see that the log-linearity breaks down at the extinction probability in later time points. Thus the value of the intercept ($\log(\hat{\lambda}_2)$) systematically increases the magnitude of the slope ($\hat{\lambda}_1$). This has the effect of underestimating the extinction probability while overestimating the probability of every other state.

It is therefore natural to consider how to approximate the probability distribution of non-zero states. We make use of the *conditional distribution* [190] $\mathbb{P}_N(X; t)$ as

$$\mathbb{P}_N(X; t) = \varphi(t) \times \mathbb{P}(X = x; t) \quad (4.51)$$

for $x \geq 1$, where

$$\varphi(t) = \frac{1}{1 - \mathbb{P}(X = 0; t)} \quad (4.52)$$

$\mathbb{P}_N(X; t)$ is simply the probability distribution $\mathbb{P}(X; t)$ without the extinction probability but which is then normalised so that it sums to 1. The expectation $\mu_N(t)$ of the conditional

distribution is given by

$$\begin{aligned}
\mu_N(t) &= \sum_{i=1}^{\tau} i \times \mathbb{P}_N(X = i; t) \\
&= \sum_{i=1}^{\tau} i \times (\varphi(t) \times \mathbb{P}(X = i; t)) \\
&= \sum_{i=0}^{\tau} i \times (\varphi(t) \times \mathbb{P}(X = i; t)) \\
&= \varphi(t) \times \mu(t)
\end{aligned} \tag{4.53}$$

because the extinction probability adds zero to the expectation $\mu(t)$.

The purple lines in Figure 4.10 are given by $\mathbb{P}_D(X; t)$, given by

$$\mathbb{P}_D(X; t) = \begin{cases} \frac{1}{\varphi(t)} \times \text{Exp}(\frac{1}{\mu_N(t)}) & \text{if } x \geq 1 \\ \mathbb{P}(X = 0; t) & \text{if } x = 0 \end{cases} \tag{4.54}$$

That is, $\mathbb{P}_D(X; t)$ is an exponential distribution with expectation $\mu_N(t)$ that includes $\mathbb{P}(X = 0; t)$, and so must be scaled down by a factor of $(1 - \mathbb{P}(X = 0; t))$ or $\frac{1}{\varphi_N(t)}$. $\mathbb{P}_D(X; t)$ provides an excellent fit to the distributions $\mathbb{P}(X; t)$, particularly at later time points. The fit is less impressive at earlier time points towards the right of the distribution, but since this represents an absolute discrepancy of less than 10^{-100} , this is an acceptable error.

Exponential distributions with rate $1/\mu(t)$ were also used to estimate $\mathbb{P}(X; t)$ (Figure 4.10, green lines), but these did not provide an acceptable fit. Figure 4.12 show Kullback-Leibler divergences over time between $\mathbb{P}(X; t | x_0 = 1, \theta_k)$ and each proposed exponential distribution approximation. All approximations become less accurate over time, except for $\mathbb{P}_D(X; t)$, which eventually reaches a divergence of less than 10^{-9} , again indicating an excellent approximation.

While it is interesting that $\mathbb{P}(X; t | x_0 = 1)$ can be closely approximated by an exponential distribution for the parameters we have chosen, this is of dubious value if $\mathbb{P}(X; t | x_0 = 1)$ must be calculated explicitly in order to parameterise that distribution. Fortunately, our simulations indicate that the extinction probability $\mathbb{P}(Ext; t)$ can be calculated accurately using a much smaller value of τ and state space, since if only the extinction probability is desired, other values of the PDF can be distorted. Further, the expected value can be approximately estimated using ODEs, and so $\mathbb{P}(X; t)$ and $\mathbb{P}_N(X; t)$ can be calculated using only the $\mathbb{P}(Ext; t)$ and the ODEs. Finally, at later time points, our repeated numerical simulations suggest we need only use Equation (4.41) to calculate the extinction probability.

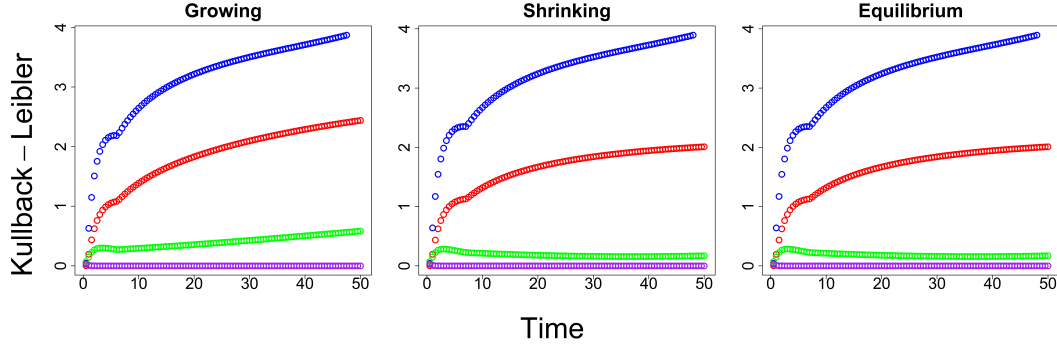


Figure 4.12.: **Kullback-Leibler divergences between $\mathbb{P}(X; t | x_0 = 1, \theta_k)$ and exponential distribution approximations.** Kullback-Leibler divergences between $\mathbb{P}(X; t | x_0 = 1, \theta_k)$ and $\text{Exp}(\hat{\lambda}_l)$, $l \in \{1, \dots, 4\}$ are plotted over time for growing ($\theta_1 = \{\pi = 2, \delta = 1.98\}$), shrinking ($\theta_2 = \{\pi = 1.98, \delta = 2\}$), and equilibrium ($\theta_3 = \{\pi = 2, \delta = 2\}$) parameters. Definitions of each $\hat{\lambda}_l$ are as for Figure 4.11, i.e. $\hat{\lambda}_1 = -a$ (red); $\hat{\lambda}_2 = e^b$ (blue); $\hat{\lambda}_3 = 1/\mathbb{E}[\mathbb{P}(X; t | x_0 = 1)]$ (green); $\hat{\lambda}_4 = 1/\mathbb{E}(\mathbb{P}_\varphi(X; t | x_0 = 1))$ (purple). Consistent with Figure 4.11, purple distributions give best fit to $\mathbb{P}(X; t | x_0 = 1)$. See Section 4.4.4 for further details on each exponential distribution.

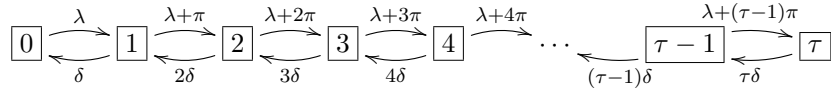


Figure 4.13.: **Mass action birth-death process flow diagram with inflow from source (truncated state space).** As for Figure 4.3, but with inclusion of constant rate λ of inflow from source.

4.5. Inclusion of Source Inflow Changes $\mathbb{P}(X; t)$

Many biological systems consider inflow from a source, or a reaction $\mathcal{R} : * \xrightarrow{\lambda} \mathcal{X}$ that takes a species of frequency $\boxed{0}$ to $\boxed{1}$. For example, considering the two “species” of mRNA and protein, source inflow would come from DNA transcription. Or considering a population of T cells, source inflow comes from the thymus. However, this formulation is inappropriate for many systems, for example the extinction of animal species. Indeed at the level of a TCR clonotype, we cannot assume a constant rate of creation *for that particular clonotype*.

The inclusion of the source inflow fundamentally changes the probability distribution $\mathbb{P}(X; t)$. Figure 4.14 shows the evolution of the extinction probability for a range of randomly chosen parameters π and δ and with source λ . In contrast to Figure 4.4, the extinction probability is no longer strictly increasing. Further, the form of $\mathbb{P}(X; t | x_0 = 1)$ can no longer be approximated by an exponential distribution, but under certain scenarios can be approximated by a Gaussian distribution.

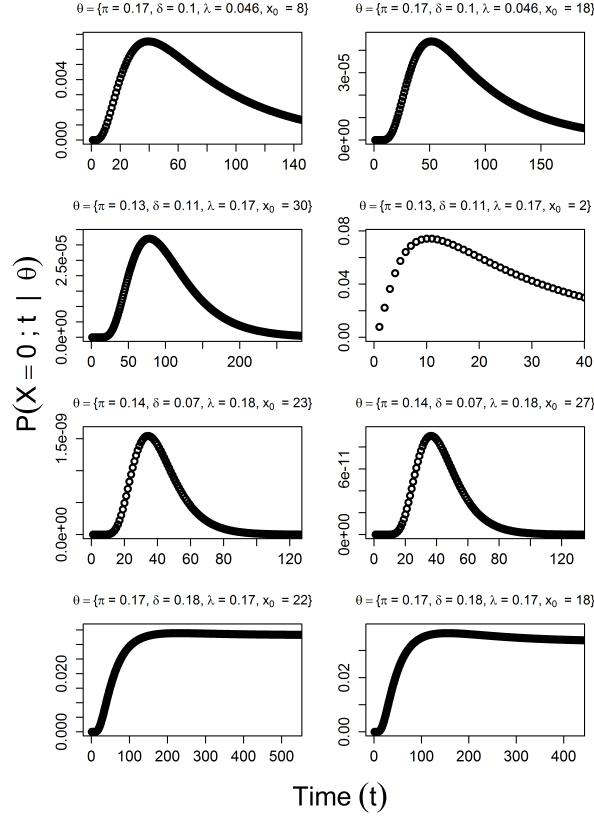


Figure 4.14.: **Evolution of Extinction Probabilities with source inflow.** Inclusion of source inflow from $\boxed{0}$ to $\boxed{1}$ results in different properties of $\mathbb{P}(X;t)$. The extinction probability is no longer increasing.

We again used the Kullback-Leibler divergence to compare $\mathbb{P}(X;t|x_0 = 1)$ to reference Gaussian distributions $\mathcal{N}(\mu(t), \sigma(t)^2)$, and to exponential distributions $\text{Exp}(1/\mu(t))$. For two sets of parameters $\theta_1 = \{\pi_1 = 0.02, \delta_1 = 0.015, \lambda_1 = 2\}$ and $\theta_2 = \{\pi_2 = 0.015, \delta_2 = 0.02, \lambda_2 = 1\}$, Figure 4.16 shows $D_{KL}(\mathbb{P}(X;t|x_0 = 1) || \mathcal{N}(\mu(t), \sigma(t)^2))$ and $D_{KL}(\mathbb{P}(X;t|x_0 = 1) || \text{Exp}(1/\mu(t)))$, i.e. the information lost by approximating $\mathbb{P}(X;t|x_0 = 1)$ by normal and exponential distributions, and how the information loss varies over time. For both parameter sets, $\mathbb{P}(X;t|x_0 = 1)$, was better approximated by a normal distribution than an exponential distribution, and this approximation improved (i.e. the KL-divergence decreased) over time.

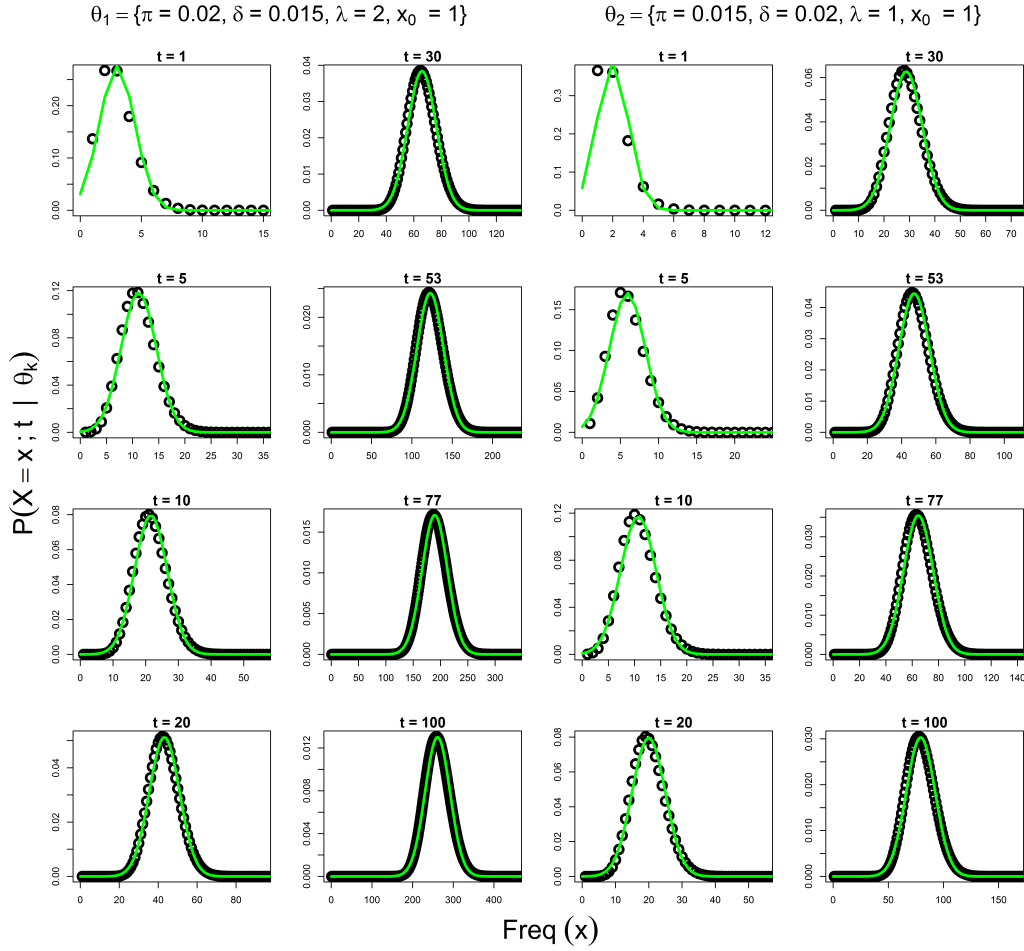


Figure 4.15.: **Example PDF evolutions** $\mathbb{P}(X; t | x_0 = 1)$ **with source inflow**. Inclusion of source inflow changes $\mathbb{P}(X; t | x_0 = 1)$ so that it closely resembles a Gaussian distribution (plotted in green lines).

4.6. Discussion

In this chapter, we gave a brief overview of stochastic processes and their increased use in modelling biological systems. We discussed the master equation and gave its explicit derivation of a simple mass-action birth-death process of one species. We explored consequences of this process without a source inflow, and found that for high starting frequencies, the probability distribution could be accurately approximated using a Gaussian distribution. We further explored how the probability distribution evolves when the species is modelled from a starting frequency $x_0 = 1$. In this case, a Gaussian distribution no longer gave an accurate description of $\mathbb{P}(X; t)$, which was instead well approximated by a simple manipulation of an exponential distribution. This is useful because of the amount

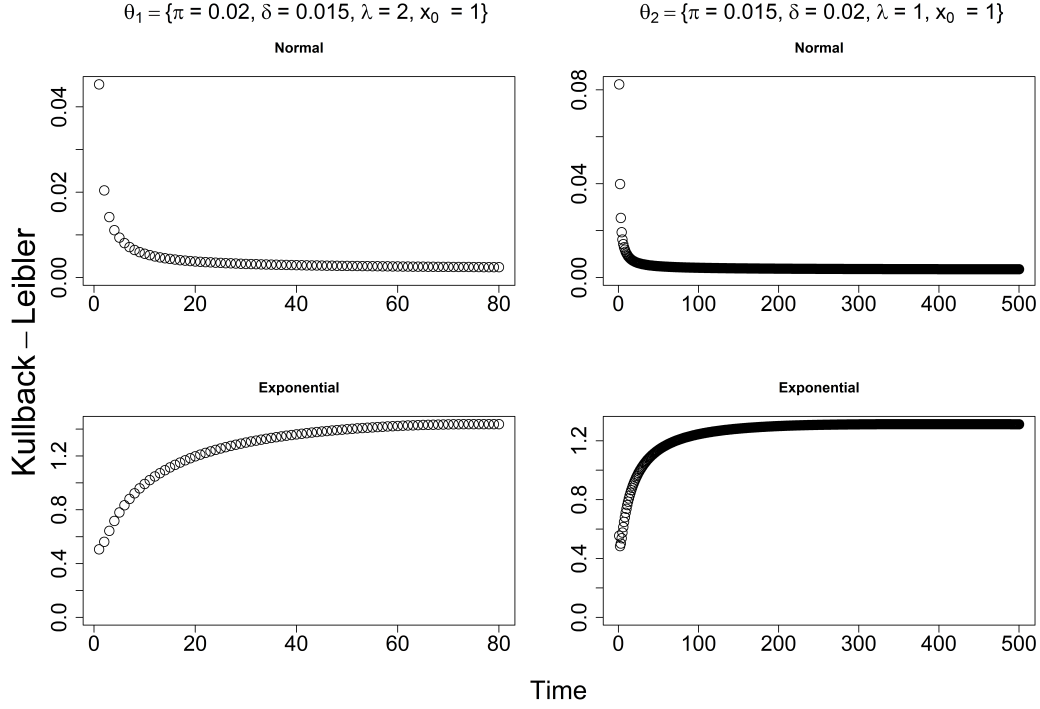


Figure 4.16.: **Kullback-Leibler divergences between $\mathbb{P}(X; t | x_0 = 1)$ and normal and exponential distributions with source inflow.** Kullback-Leibler divergences plotted for between $\mathbb{P}(X; t | x_0 = 1)$ and normal distribution $\mathcal{N}(\mu(t), \sigma(t)^2)$ (top row) and between $\mathbb{P}(X; t | x_0 = 1)$ and exponential distributions $\text{Exp}(1/\mu(t))$ (bottom row). Over time, the normal distributions become more similar and exponential distributions become less similar to $\mathbb{P}(X; t | x_0 = 1)$, for parameter sets θ_1 where $\pi > \delta$ (left column) and θ_2 where $\pi < \delta$. At all times, exponential distributions give poor approximation to $\mathbb{P}(X; t | x_0 = 1)$.

of computational power that can be saved. Computing the solution of the master equation can be prohibitively time-consuming for large truncated state spaces Ω_τ . By approximating using the manipulated exponential distribution, only the expectation (approximately calculated using ODEs) and the extinction probability need to be calculated.

A major shortcoming of the preceding analysis is that there are substantial time constraints when explicitly solving the master equation (Equation (4.30)). To model a state space Ω_τ where $\tau = 10^4$ incurs a runtime of approximately 7 hours for a single matrix exponential (using R function “expm” from the package “matrix”). While this is feasible when exploring the behaviour of a species given a limited set of parameter values, it is not conducive to parameter fitting, where the probability distributions from many parameter sets in a potentially large parameter space must be compared to empirical data.

Furthermore, a truncated state space of 10^4 will often not be biologically realistic. For example, the DivE distribution algorithm conservatively predicts that a single HTLV-1⁺ T cell clone can reach a frequency of 10^{10} infected cells. This makes direct solution of the master equation impossible, both in terms of runtime and memory. The transition matrix of a mass-action birth-death process where $\tau = 10^{10}$ will have $(\tau+1)^2 = (10^{10} + 1)^2 \approx 10^{20}$ entries. While only the minority of these entries (specifically $3\tau - 1$, as each state can only move to its immediate neighbours, except state $\boxed{\tau}$ which can only decrease and $\boxed{0}$ which is an absorbing state) will be non-zero, all terms in the matrix exponential except the first column will be non-zero. Assuming that each non-zero entry is a single precision floating point that uses 4 bytes, the matrix exponential would use $10^{20} - 10^{10} \times 4 \approx 4 \times 10^{20}$ bytes which is approximately 400,000 petabytes. Clearly this is not possible with current computational power.

The finite state projection method (FSP) aims to address the difficulty of an intractably large state space [183, 191, 192]. The FSP truncates the state space in a different way than we have. The FSP does not consider a pre-determined minimum and maximum frequency, but instead iteratively expands a truncated state space until the probability that the system takes values outside this state space is less than an arbitrary error level ϵ (runtime-permitting). The limits on the state space change as the system is propagated.

However, the FSP has disadvantages. $\mathbb{P}(X; t)$ has to be calculated at each time step, and this cannot be expedited. For our purposes, repeated iterations of the FSP, while faster than matrix exponential calculation, is much slower than simple matrix multiplication with a single, pre-calculated matrix exponential. Like the Gillespie algorithm, the runtime of the FSP increases markedly with simulation duration, whereas a matrix exponential does not. Further, the truncated state space is likely to substantially increase as the system is run.

While it has been argued that stochastic modelling is necessary even for large species [137], it is generally accepted that deterministic modelling is acceptable for biological processes that occur on a large scale. For systems with dynamics that occur on small and large scales, a hybrid model of deterministic and stochastic systems can be employed. In the next chapter, we discuss a hybrid model of the within-host dynamics of HTLV-1 infection that examines mechanisms of viral persistence.

5. Hybrid Model of HTLV-1 persistence

5.1. Introduction

Like all retroviruses, human T-Lymphotropic virus type-1 (HTLV-1) persists within host via two mechanisms. Infectious spread (i.e. *de novo* infection) occurs through cell-cell contact between infected and uninfected CD4⁺ T cells, whereupon the virus will integrate in a different location of the target cell genome than its progenitor. In mitotic spread, infected cells proliferate to produce identical sister cells with the same genomic site of proviral integration. The interplay between infectious and mitotic spread results in multiple, distinct clones, where a clone is defined as a population of identically infected cells, each with a common integration site [20, 22, 125].

The relative contributions of infectious spread and mitotic spread are unknown, and this ratio will directly determine the efficacy of different drug regimens. Although no effective treatment for HTLV-1 infection has yet been identified [193], anti-retroviral therapy (ART) is believed to reduce infectious spread [21] because of its ability to drastically reduce reverse transcription in HIV infection [194, 195]. Similarly, immuno-suppressants such as Ciclosporin will inhibit T cell proliferation and thus may be effective in reducing mitotic spread [196, p. 163].

It is not possible to directly observe ongoing within-host HTLV-1 persistence in real time, and a number of factors make it difficult to estimate the ratio of infectious to mitotic spread from existing data. Since infectious spread establishes new clones, the number of clones is closely related to the rate of infectious spread. However, estimating the number of clones in the host from blood samples is nontrivial (see Chapter 2). Further, even if the number of clones is known, the frequency of each clone, which informs the rate of mitotic spread, is unknown. Finally, even if a complete census of the number of clones and their frequencies was available, this would not in itself determine the ratio, as the turnover rate would be unknown.

Despite the difficulties in directly observing or estimating quantitatively the ratio of infectious to mitotic spread, the prevailing view is that mitotic spread accounts for the majority

of HTLV-1 persistence [14, 20, 24, 197], and that infectious spread is thought to be negligible after the initial phase of infection [21, 24, 198]. This belief is supported by three main observations. First, it was thought that there were relatively few HTLV-1 clones in one host [20, 22, 23, 24, 25, 26, 113, 114]. Second, it has been reported that HTLV-1 has relatively little sequence variation when compared with other viruses including HIV, Hepatitis C Virus (HCV) and influenza A [14, 15, 199, 200, 201, 202, 203]. Since the host DNA-based DNA polymerase used in cell proliferation (mitotic spread) is substantially less error-prone than the viral RNA-based DNA polymerase used in reverse transcription (infectious spread), lack of sequence variation would imply that mitotic spread is the most common form of spread. Third, many HTLV-1⁺ clones have been observed at multiple time points separated by several years [22, 23], and it is all but impossible that repeated proviral integration (infectious spread) at the same integration site would occur frequently enough to maintain a long-lived clone.

However, the preceding three observations do not necessarily imply that infectious spread is negligible [14], particularly when we consider the number of clones. First, estimates of the number of clones have increased over time [20, 22, 23, 24, 25, 26], and DivE estimates that there are approximately 10^4 clones in ACs and HAM/TSP patients [11, 125, 204]. Second, a relatively small amount of sequence variation is unsurprising if only a few clones are observed, and this is compounded if multiple cells of the same clone are sampled. Finally, the repeated observation of several clones over many years implies the presence of mitotic spread, but not an absence of infectious spread. Indeed, many new clones are detected in subsequent samples, and so it is not necessary to surmise that all clones were established during initial infection. Further, the observation of a temporary but dramatic reduction in PVL upon treating HAM/TSP patients with the reverse transcriptase inhibitor lamivudine [21] strongly implies that infectious spread remains important in HTLV-1 persistence.

There has been renewed interest in the role that infectious spread may play in HTLV-1 associated disease. *Melamed et al* [124] compared the clonality of HTLV-1 to HTLV-2 and found that the oligoclonality index (OCI) of HTLV-2 clones was higher than that of HTLV-1, implying a higher relative contribution of mitotic spread in HTLV-2 persistence. Since HTLV-2 is not known to cause disease, it is conceivable that infectious spread may be a driving determinant of disease. In addition, as we have seen in Section 3.3, there are fewer clones in asymptomatic carriers, which is consistent with the hypothesis that infectious spread plays a role in progression to inflammatory disease.

Even when taking recent estimates of clonal diversity into account, there is still good reason to believe that mitotic spread is predominant, as 10^4 clones (created by infectious spread) must be compared to approximately 10^{11} infected cells (maintained by mitotic spread). However, this does not consider clones that are continuously created and killed

by the immune response and natural death. In this chapter, we develop a deterministic and stochastic hybrid model of HTLV-1 dynamics to directly model the relative contributions of infectious spread and mitotic spread.

5.2. Hybrid model

We first outline a model that does not consider runtime and memory constraints. We then discuss approximations that reduce the complexity of the model such that it can be run in an acceptably short time frame.

5.2.1. Deterministic Model of HTLV-1 Dynamics

We consider a system with $S(t) \subset \mathbb{N}$ clones at time t , where each clone is indexed by a natural number $i \in \mathbb{N}$. In a given instant, an infected cell will either express some or all viral proteins (we term this state *expressing*) or it will not express viral proteins (we term this state *latent*). Proliferation and death rates will be different between expressing and latent cells, and so we consider the expressing and latent compartments of the clone separately. By expressing viral proteins or by silencing the expression of viral proteins, the cell can switch between expressing and latent compartments.

We have the following system of ODEs for each clone (Figure 5.1).

$$\dot{x}_i = \frac{\pi x_i}{K + N(t)} - \delta x_i + \varepsilon_i y_i - s_i x_i \quad (5.1a)$$

$$\dot{y}_i = \frac{p y_i}{K + N(t)} - d y_i - \varepsilon_i y_i + s_i x_i \quad (5.1b)$$

where $N(t) = \sum_{j=1}^{S(t)} x_j(t) + y_j(t)$ is the total number of infected cells summed over all clones alive at time t ; π and p are the proliferation rates of expressing and latent cells respectively, δ and d are the death rates of expressing and latent cells respectively, and ε_i and s_i are the respective clone-specific rates of viral expression and silencing, and K is the density dependency parameter (see Section 5.2.8).

The above ODEs model the dynamics within a clone, but do not incorporate infectious spread or describe the number of clones. The number of clones $S(t)$ at time t is given by

$$S(t) = S(0) + \int_0^t r_I \left[\sum_{i \in S(s)} x_i(s) \right] ds - \int_0^t \sum_{i \in S(s)} \mathcal{D}_i(s) ds \quad (5.2)$$

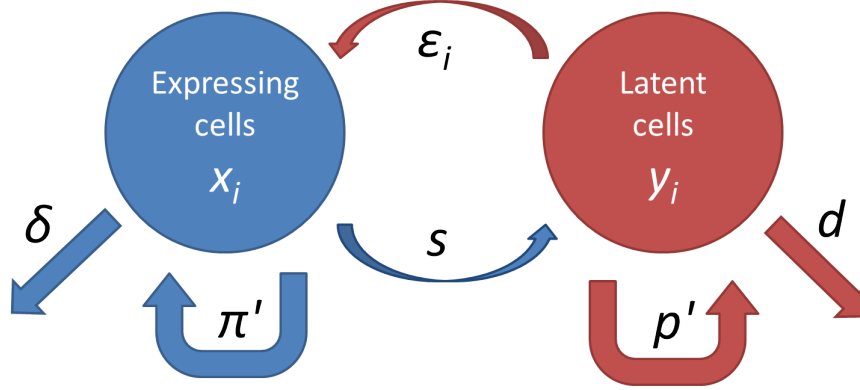


Figure 5.1.: **Schematic of single clone dynamics.** A single HTLV-1 clone of infected cells is divided into expressing (blue) and latent (red) compartments. Expressing and latent cells can proliferate and die, or switch between compartments. Density-dependent proliferation rates $\pi'(t)$ and $p'(t)$ are defined in Equation (5.4).

where r_I is the rate of infectious spread. It is assumed that only expressing cells can infect uninfected cells. $\mathcal{D}_i(t)$ is the extinction probability of clone i at time t , although it does not have a simple functional form. We discuss $\mathcal{D}_i(t)$ in Section 5.2.5.

In Chapter 4, we explored situations where deterministic modelling is inappropriate. First, deterministic models do not account for random effects [156]. Second, they are continuous (i.e. not discrete) [153]. Third, their values approximately correspond to the expectation of a system [175, 205]. Finally, if an attempt is made to estimate uncertainty, it is often assumed that deviations from the expectation are normally distributed [158, 159]. However, for species that are low in copy number (in our case small HTLV-1⁺ T cell clones), random effects need to be considered [154, 155, 157], rounding to integers will present a greater error than for large clones, the expected value is unlikely to satisfactorily encompass the behaviour of the clone over time, and the assumption of normality is frequently violated. This is particularly true when clone extinction is likely (see Section 4.4.4). Because small clones contain the majority of information on infectious spread, we must reformulate our deterministic model to a discrete stochastic model of clone dynamics. Instead of modelling a single state (expected value) of the clone at a given time, we must model many potential states of the clone and their corresponding probabilities. Crucially, we must do this without restricting the shape that the probability distribution may take over time by specifying a particular analytical form.

5.2.2. Stochastic Model of Within-Host HTLV-1 Dynamics

Under a stochastic framework, we describe within host HTLV-1 dynamics with a set of reactions and their corresponding *propensities*, at a time when the entire system of clones (but not necessarily each individual clone) has reached steady state. The dynamics of an individual clone can be summarised via the following seven reactions: latent cells can proliferate, die, or express viral proteins; expressing cells can proliferate, die, be silenced, or infect uninfected cells (Figures 5.1 and 1.1). Thus the total number of possible reactions R at time t is $R = 7S(t)$. Let $X(t) = ((X_i(t), Y_i(t))_{i \in S(t)})^T$ be the state vector at time t of the expressing and latent compartments of all clones. $X(t)$ is a random variable in $\mathbb{N}_0^{S(t)}$ that consists of the random variables $X_i, Y_i \in \mathbb{N}_0$ of the frequencies $x_i(t)$ and $y_i(t)$ of “species” $\mathcal{X}_i$ and $\mathcal{Y}_i$. $X(t)$ is the set of expressing and latent cells of clone i , ($i = 1, \dots, S(t)$), and is of length $2S(t)$.

For a single clone i , we have the following reactions:

$$\mathcal{R}_{7(i-1)+1} : \mathcal{X}_i \xrightarrow{\pi'(t)} 2\mathcal{X}_i \quad (5.3a)$$

$$\mathcal{R}_{7(i-1)+2} : \mathcal{X}_i \xrightarrow{\delta} * \quad (5.3b)$$

$$\mathcal{R}_{7(i-1)+3} : \mathcal{X}_i \xrightarrow{s} \mathcal{Y}_i \quad (5.3c)$$

$$\mathcal{R}_{7(i-1)+4} : \mathcal{Y}_i \xrightarrow{p(t)} 2\mathcal{Y}_i \quad (5.3d)$$

$$\mathcal{R}_{7(i-1)+5} : \mathcal{Y}_i \xrightarrow{d} * \quad (5.3e)$$

$$\mathcal{R}_{7(i-1)+6} : \mathcal{Y}_i \xrightarrow{\varepsilon_i} \mathcal{X}_i \quad (5.3f)$$

$$\mathcal{R}_{7(i-1)+7} : \mathcal{X}_i \xrightarrow{r_I} \mathcal{X}_i + \mathcal{X}_{S(t)+1} \quad (5.3g)$$

where

$$\pi'(t) = \frac{\pi}{K + N(t)} \quad (5.4a)$$

$$p'(t) = \frac{p}{K + N(t)} \quad (5.4b)$$

In our model, the above reactions are monomolecular (that is, the reactions do not depend on contact between species or clones). This is an approximation, but an acceptable one since HTLV-1 proviral load remains stable over many years [206], and so the number of uninfected cells stays more or less constant. Therefore, using Equation (4.3), the reactions

have the following propensities.

$$\alpha_{7(i-1)+1}(X(t)) = \frac{\pi x_i}{K + N(t)} \quad (5.5a)$$

$$\alpha_{7(i-1)+2}(X(t)) = \delta x_i \quad (5.5b)$$

$$\alpha_{7(i-1)+3}(X(t)) = s x_i \quad (5.5c)$$

$$\alpha_{7(i-1)+4}(X(t)) = \frac{p y_i}{K + N(t)} \quad (5.5d)$$

$$\alpha_{7(i-1)+5}(X(t)) = d y_i \quad (5.5e)$$

$$\alpha_{7(i-1)+6}(X(t)) = \varepsilon_i y_i \quad (5.5f)$$

$$\alpha_{7(i-1)+7}(X(t)) = r_I x_j \quad (5.5g)$$

with the following stoichiometric vectors:

$$\nu_{7(i-1)+1} = \begin{pmatrix} \boxed{1} \\ \boxed{2} \\ \vdots \\ \boxed{2i-1} \\ \boxed{2i} \\ \vdots \\ \boxed{2S(t)-1} \\ \boxed{2S(t)} \end{pmatrix} \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix}, \quad \nu_{7(i-1)+2} = \begin{pmatrix} \boxed{1} \\ \boxed{2} \\ \vdots \\ \boxed{2i-1} \\ \boxed{2i} \\ \vdots \\ \boxed{2S(t)-1} \\ \boxed{2S(t)} \end{pmatrix} \begin{pmatrix} 0 \\ 0 \\ \vdots \\ -1 \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix}, \quad \nu_{7(i-1)+3} = \begin{pmatrix} \boxed{1} \\ \boxed{2} \\ \vdots \\ \boxed{2i-1} \\ \boxed{2i} \\ \vdots \\ \boxed{2S(t)-1} \\ \boxed{2S(t)} \end{pmatrix} \begin{pmatrix} 0 \\ 0 \\ \vdots \\ -1 \\ 1 \\ \vdots \\ 0 \\ 0 \end{pmatrix},$$

$$\nu_{7(i-1)+4} = \begin{pmatrix} \boxed{1} \\ \boxed{2} \\ \vdots \\ \boxed{2i-1} \\ \boxed{2i} \\ \vdots \\ \boxed{2S(t)-1} \\ \boxed{2S(t)} \end{pmatrix} \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \\ \vdots \\ 0 \\ 0 \end{pmatrix}, \quad \nu_{7(i-1)+5} = \begin{pmatrix} \boxed{1} \\ \boxed{2} \\ \vdots \\ \boxed{2i-1} \\ \boxed{2i} \\ \vdots \\ \boxed{2S(t)-1} \\ \boxed{2S(t)} \end{pmatrix} \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ -1 \\ \vdots \\ 0 \\ 0 \end{pmatrix}, \quad \nu_{7(i-1)+6} = \begin{pmatrix} \boxed{1} \\ \boxed{2} \\ \vdots \\ \boxed{2i-1} \\ \boxed{2i} \\ \vdots \\ \boxed{2S(t)-1} \\ \boxed{2S(t)} \end{pmatrix} \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \\ -1 \\ \vdots \\ 0 \\ 0 \end{pmatrix}$$

and

$$\nu_{7(i-1)+7} = \begin{pmatrix} \boxed{1} \\ \boxed{2} \\ \vdots \\ \boxed{2i-1} \\ \boxed{2i} \\ \vdots \\ \boxed{2S(t)-1} \\ \boxed{2S(t)} \\ \boxed{2S(t)+1} \\ \boxed{2S(t)+2} \end{pmatrix} \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}$$

5.2.3. Clone State Space and Interactions

The “curse of dimensionality” [174, 205] refers to the exponentially increasing number of states that must be modelled as more species of the system are introduced. As we have seen in Chapter 4, for a single clone $Z(t) \in \mathbb{N}_0$ that is not divided into expressing and latent states, the state space (i.e. the set of all states the system can take) is given by $\Omega = \{z : z \in \{0, \dots, \tau\}\}$, where τ is the maximum frequency of the clone and number of states is given by $|\Omega| = \tau + 1$. If we divide the clone into expressing and latent cells, where each can reach a frequency of τ , the state space becomes $\Omega = \{(x, y) : x, y \in \{0, \dots, \tau\}\}$, and $|\Omega| = (\tau + 1)^2$.

If all clones interact and are modelled using a single master equation, assuming all clones have the same frequency range, the size of the state space will increase to $(\tau + 1)^{2S(t)}$. Thus the complexity and runtime of the model increases exponentially with the number of clones. In contrast, when each clone is modelled in isolation, the number of clones affects complexity and runtime linearly. Therefore to make the model computationally tractable, we model each clone individually, which involves further approximations.

With the exception of infectious spread, each *reaction* in Equation (5.3) occurs within a clone and does not affect other clones, and infectious spread does not affect any pre-existing clones. However, the proliferation *propensities* (equations (5.5a) and (5.5d)) of a clone depend on the combined frequencies of all other clones, which would seem to necessitate modelling all clones in a single state space. There are two approximations we can make to avoid this problem. First, we could “linearise” proliferation rates such that Equations (5.5a) and (5.5d) become $\alpha_{7(i-1)+1}(X(t)) = \pi'(t)x_i$, and $\alpha_{7(i-1)+4}(X(t)) = p'(t)y_i$, where

$$\pi'(t) = \frac{\pi}{K + N(\lfloor t \rfloor)} \quad (5.6)$$

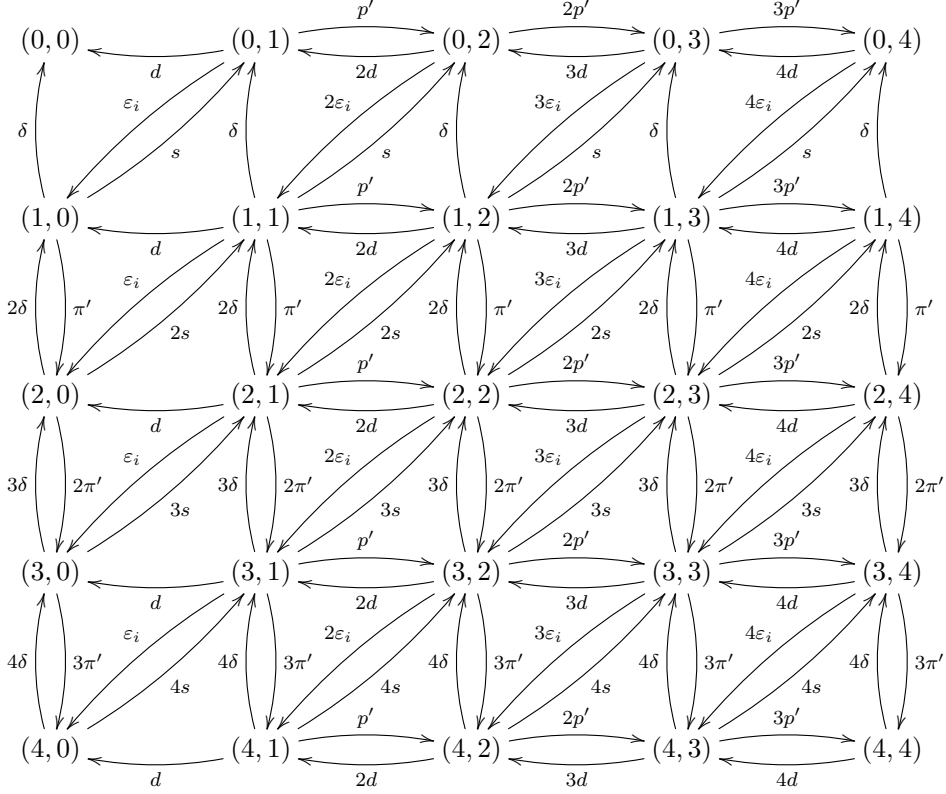


Figure 5.2.: Example clone state space and interactions for upper bound $\tau = 4$. Mappings between states are described by mass action terms. Here π' and p' are “aggregate” proliferation rates that are constant within a time step. See Section 5.2.3 for further details.

and

$$p'(t) = \frac{p}{K + N(\lfloor t \rfloor)} \quad (5.7)$$

are step-functions that remain constant within a time step and are updated between time steps ($\lfloor t \rfloor$ is the “floor” function that returns the integer smaller than t). Since the number of infected cells $N(t) = \sum_{j=1}^{S(t)} x_j + y_j$ is unlikely to change substantially within time steps, this is an acceptable approximation. An example state space of a clone where $\tau = 4$, and the interactions between possible clone states is given in Figure 5.2.

However, changing proliferation values between time steps requires the the matrix exponential be recalculated at every time step, and thus this will not reduce runtime to a sufficient extent. The second approximation then is to assume that $N(t)$ is constant. This will not be valid in early infection, but is acceptable during chronic infection when the

system is at equilibrium, as the HTLV-1 proviral load (PVL) remains stable for many years [22, 206]. Assuming the number of infected cells is constant, the propensities (5.5a) and (5.5d) become

$$\alpha_{7(i-1)+1}(X(t)) = \frac{\pi x_i}{H + x_i + y_i} \quad (5.8)$$

and

$$\alpha_{7(i-1)+4}(X(t)) = \frac{p y_i}{H + x_i + y_i} \quad (5.9)$$

and we can model each clone individually with $H = K + N(t^*)$.

Modelling each clone individually in this way incurs an error because the propensities for proliferation are dependent only upon the initial state of the system, and do not change with the growth or decay of the other clones. However, this approximation allows for a single matrix exponential to be calculated for each clone, and for the evolution of the clone probability distribution to be calculated using recursive property of the matrix exponential (see Section 4.3.3). We discuss parameters H and K in Section 5.2.8.

5.2.4. Hybrid Model Propagation: Strang Splitting

We have seen previously that using the matrix exponential to calculate the solution to the master equation requires the state space Ω_τ to be finite (Section 4.3.1). Here the upper bound τ refers to the maximum frequency that either the expressing or latent cells can take, and thus the maximum frequency of a clone is 2τ (Figure 5.2). If A_{Ω_τ} is the transition matrix of the state space $\Omega_\tau = \{(x, y) : x, y \in \{0, \dots, \tau\}\}$ then A_{Ω_τ} is a $|\Omega_\tau|$ by $|\Omega_\tau|$ matrix and so the number of entries in A_{Ω_τ} is $(\tau + 1)^4$. The DivE distribution generation algorithm (Section 2.2.11) estimates that the frequency of a given clone can range up to size 10^{10} , and so it is not possible to model all clones stochastically, even with our simplifying approach that models each clone independently.

Fortunately, for clones above a given size F (see Section 5.2.9), the distorting effects of the extinction probability are negligible. Thus for large clones we can assume that the probability distribution will be well described using only the expectation, variance, and covariance of the expressing and latent compartments [176]. These quantities can be modelled deterministically using a system of ODEs [207], which we discuss in Section 5.2.6.

We thus partition our system of the within-host dynamics of HTLV-1 infection into two constituent systems: a deterministic system $D(t)$ modelled by a system of ODEs, and a stochastic system $S(t)$ modelled by multiple master equations. We propagate these systems concurrently as a *hybrid model*, using the method of “Strang splitting” [208] (Figure 5.3). The Strang splitting method is as follows (formulation below taken from [173]):

1. Choose Duration and length of time step h
2. Initialize
3. Run $D(t)$ for half time step, keep $S(t)$ constant
4. Run $S(t)$ for a whole time step, keep $D(t)$ constant
5. Run $D(t)$ for next half time step, keep $S(t)$ constant
6. Repeat until duration reached

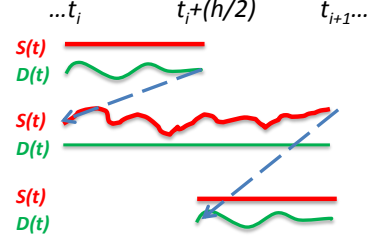


Figure 5.3.: **Schematic of strang splitting procedure.** Dashed arrows indicate that values from deterministic system are used as initial conditions for stochastic system and vice-versa.

1. For the duration of the simulation $t_{end} > 0$, choose $N_t \in \mathbb{N}$. Let $h = t_{end}/N_t$ and $t_n = nh \quad \forall n = 0, \dots, N_t$.
2. Set $n = 0$ and set initial conditions.
3. Half time step in deterministic $D(t)$: Solve deterministic process in time interval $[t_n, t_n + h/2]$ and keep $S(t)$ constant.
4. Full time step in stochastic $S(t)$: Solve stochastic process (i.e. clone-specific master equations) in time interval $[t_n, t_{n+1}]$ and keep $D(t_n + h/2)$ constant.
5. Next half time step in deterministic system $D(t)$: Solve deterministic process in time interval $[t_n + h/2, t_{n+1}]$ and keep $S(t)$ constant.
6. If $n < N_t$, set $n \leftarrow n + 1$ and return to step 3.

In addition to the error incurred by modelling part of the stochastic system deterministically, there is an error associated in not propagating both systems exactly simultaneously, and this error is a function of the length of the time step h [173]. Modelling the whole system for a single time step consists of steps 3 to 5. In our system, between time steps, some deterministically modelled clones become small enough to necessitate stochastic modelling, whereas some stochastically modelled clones may grow to the extent that they need to be modelled deterministically (see section 5.2.7). We choose our time step to be equal to one day.

5.2.5. Infectious Spread r_I

Equation (5.2) models the number of clones at time t . Though it is a continuous process, we approximate the number of clones created in the interval at time $[t_n, t_n + h/2]$ by

$S_{new}(t_n + h/2)$ where

$$S_{new}(t_n + h/2) = (h/2) \times r_I \sum_{i=1}^{S(t_n)} x_i(t_n + h/2) \quad (5.10)$$

New clones are assumed to start in state $(x_{S(t_n)+k}, y_{S(t_n)+k}) = (0, 1)$, for $k \in \{1, \dots, S_{new}(t_n + h/2)\}$. Note that $r_I \sum_{i=1}^{S(t_n)} x_i(t_n + h/2)$ does not model the probability that a *particular* integration site will be infected, i.e. that a particular, as yet un-established clone will go from state $\boxed{(0, 0)}$ to state $\boxed{(0, 1)}$. It is also assumed that all new clones are created at the beginning of a time step. The error associated with this assumption is proportional to the length of the time step h and inversely proportional to the lifetime of a clone. Because our time step is equal to one day, this will not present a significant source of error. Equation (5.10) is deterministic, which is reasonable since the mass action term $\sum_{i=1}^{S(t_n)} x_i(t_n + h/2)$ will always be large.

An interesting feature of our model is the ability to estimate the expected number of clones at time t (see Equation (5.2)), which is not straightforward when using a purely deterministic formulation. The birth of new clones is a simple mass-action term that is easily described deterministically. Because death occurs at the level of the cell, and not the clone, it is less straightforward to estimate the number of *clones* that will be killed given a number of cells killed. For example, if we know that 10^3 cells are killed, this may result in the death of 0 clones (i.e. if all dead cells are from one sufficiently large clone) or as many as 10^3 clones (i.e. where each clone was a singleton).

In contrast, our model allows the extinction probability of a single clone to be calculated for any time step. Therefore, the expected number of clones that die in the interval $[t, t + \Delta t]$ is given by

$$\int_t^{t+\Delta t} \sum_{i \in S(s)} 1 \times \mathcal{D}_i(s) ds \quad (5.11)$$

where $\mathcal{D}_i(t) = \mathbb{P}((X_i, Y_i) = (0, 0) ; t)$ is the extinction probability of clone i at time t .

5.2.6. Variance and Covariance of Clones

Under a stochastic paradigm, the ODEs in equation (5.1) approximately correspond to the expected frequencies of the expressing and latent compartments of a given clone i . Because the propensities of our model (equation (5.5)) have degrees of at most 1, we can derive another set of ODEs to model the variance and covariance [207] of a given species

i. With our notation, the moments are given by

$$\frac{dm_i}{dt} = \sum_{r=1}^R \nu_r^i \left(\alpha_r(X(t)) + \sum_{k,l} \frac{\partial^2 \alpha_r(m)}{\partial x_k \partial x_l} \frac{C_{kl}}{2!} \right) \quad (5.12a)$$

$$\frac{dC_{ij}}{dt} = \sum_{r=1}^R \left(\nu_r^i \sum_k \frac{\partial \alpha_r(X(t))}{\partial x_k} \frac{C_{kj}}{1!} + \nu_r^j \sum_l \frac{\partial \alpha_r(X(t))}{\partial x_l} \frac{C_{il}}{1!} \right) + \sum_{r=1}^R \nu_r^{[i,j]} \left(\alpha_r(X(t)) + \sum_{k,l} \frac{\partial^2 \alpha_r(X(t))}{\partial x_k \partial x_l} \frac{C_{kl}}{2!} \right) \quad (5.12b)$$

where m_i is the expected value of species i , C_{ij} is the covariance between species i and species j , ν_r^i is the i^{th} component of ν_r , the stoichiometric vector of reaction r , and $\nu_r^{[i,j]} = \nu_r^i \nu_r^j$. Note that the subscripts i and j in the above equations refer to the i^{th} and j^{th} entry of the state vector $X(t)$. Because we divide clones into two compartments (expressing and latent), $X(t)$ has $2S(t)$ entries. Re-indexing so that the subscript i refers to the i^{th} clone ($i \in \{1, \dots, S(t)\}$), we introduce the following shorthand. Let

$$x_i := \mathbb{E}[X_i] \quad (5.13a)$$

$$y_i := \mathbb{E}[Y_i] \quad (5.13b)$$

$$V_{x_i} := \text{Var}[X_i] \quad (5.13c)$$

$$V_{y_i} := \text{Var}[Y_i] \quad (5.13d)$$

$$C_{x_i, y_i} := \text{Cov}[X_i, Y_i] \quad (5.13e)$$

To relate this to Equation (5.12), we have $V_{x_i} = V_{2i-1} = C_{(2i-1), (2i-1)} = \text{Cov}[X_i, X_i] = \text{Var}[X_i]$, $V_{y_i} = V_{2i} = C_{(2i), (2i)} = \text{Cov}[Y_i, Y_i] = \text{Var}[Y_i]$, and $C_{x_i, y_i} = C_{(2i-1), (2i)} = \text{Cov}[X_i, Y_i]$.

We approximate by assuming that the frequency of a given clone is independent of any other clone, i.e. that for $i \neq j$, $\text{Cov}[X_i, X_j] = \text{Cov}[Y_i, Y_j] = \text{Cov}[X_i, Y_j] = 0$. This assumption is valid at the level of the individual clone, in that $\text{Cov}[X_i, Y_j] \approx 0$ for $i \neq j$. However $\sum_{k=1}^{S(t)} C_{ik}$ may not be negligible. Nevertheless, to substantially reduce the dimension of the system of ODEs, we must assume that clone frequencies are approximately independent of each other. With these assumptions, it can be shown that we have the following system of ordinary differential equations to describe the expectations, variances of covariance of the expressing and latent compartments x_i and y_i of clone i . See Appendix B for further

details.

$$\frac{dx_i}{dt} = \frac{\pi x_i}{K + N(t)} - \delta x_i + \varepsilon_i y_i - s_i x_i \quad (5.14a)$$

$$\frac{dy_i}{dt} = \frac{p y_i}{K + N(t)} - d y_i - \varepsilon_i y_i + s_i x_i \quad (5.14b)$$

$$\begin{aligned} \frac{dV_{x_i}}{dt} = & \left(\frac{\pi(K + N(t)) - \pi x_i}{(K + N(t))^2} - \delta - s_i \right) 2V_{x_i} + 2\varepsilon_i C_{x_i, y_i} \\ & + \left(\frac{\pi}{(K + N(t))^2} + \delta + s_i \right) x_i + \varepsilon_i y_i \end{aligned} \quad (5.14c)$$

$$\begin{aligned} \frac{dV_{y_i}}{dt} = & \left(\frac{p(K + N(t)) - p y_i}{(K + N(t))^2} - d - \varepsilon_i \right) 2V_{y_i} + 2s_i C_{x_i, y_i} \\ & + \left(\frac{p}{(K + N(t))^2} + d + \varepsilon_i \right) y_i + s_i x_i \end{aligned} \quad (5.14d)$$

$$\begin{aligned} \frac{dC_{x_i, y_i}}{dt} = & \left(\frac{(\pi + p)(K + N(t)) - \pi x_i - p y_i}{(K + N(t))^2} - \delta - s_i - d - \varepsilon_i \right) C_{x_i, y_i} \\ & + s_i V_{x_i} + \varepsilon_i V_{y_i} - s_i x_i - \varepsilon_i y_i \end{aligned} \quad (5.14e)$$

We also require the variance of $x_i + y_i$, and so we compute $V_{x_i + y_i} = V_{x_i} + V_{y_i} + 2C_{x_i, y_i}$.

5.2.7. Clone Movement Between Deterministic and Stochastic Compartments

Recall that F is the frequency above which we can safely approximate the stochastic behaviour of a clone as a deterministic process. See Section 5.2.9 for a discussion on how F is determined. When the expected value of the size of a deterministically modelled clone $x_i(t) + y_i(t)$ clone drops below F , it is modelled stochastically and values $x_i(t)$ and $y_i(t)$ are rounded to the nearest integer. If the state $(x_i(t), y_i(t)) \in \Omega_\tau$ is indexed by κ , then the initial distribution $\mathbb{P}_0$ is given by the Kronecker delta $\delta_{x, \kappa}$, as in Equation (4.31). Note this requires $F < \tau$.

We “promote” a stochastically modelled clone to a deterministically modelled clone at time t if either:

$$\mathbb{P}(X_i + Y_i = 2\tau; t) - \mathbb{P}(X_i + Y_i = (2\tau - 1); t) > \epsilon \quad (5.15a)$$

$$\mathbb{P}(X_i = \tau; t) - \mathbb{P}(X_i = (\tau - 1); t) > \epsilon \quad (5.15b)$$

$$\mathbb{P}(Y_i = \tau; t) - \mathbb{P}(Y_i = (\tau - 1); t) > \epsilon \quad (5.15c)$$

where $\epsilon > 0$. These conditions guard against the upper limit τ distorting the probability distribution. If the state space is unconstrained and the distribution is unimodal, then we

would ordinarily expect for $a > b > \mathbb{E}[X_i + Y_i]$,

$$\mathbb{P}(X_i + Y_i = a; t) < \mathbb{P}(X_i + Y_i = b; t) \quad (5.16)$$

However, using a truncated state space Ω_τ causes probability mass to build up at $X_i + Y_i = 2\tau$ (Section 4.4.2). To prevent a distortion of the expectation, we promote when $\mathbb{P}(X_i + Y_i = 2\tau; t) - \mathbb{P}(X_i + Y_i = (2\tau - 1); t) > \epsilon$. In practice ϵ must be greater than zero as extremely small build ups of probability mass happen before extinction probability nears $\mathbb{P}(Ext)^*$. We set $\epsilon = 1/1000$, as this value allows for feasibly small values of τ that do not overly distort the probability distribution, and that allows $\mathbb{P}(Ext; t)$ to get close to its asymptote.

When a clone is promoted, we again make use of the conditional distribution $\mathbb{P}_N(X(t); t)$ (Section 4.4.4). This is defined for the two-dimensional case as it is for the one-dimensional case, i.e.

$$\mathbb{P}_N(X(t) = \omega; t) = \varphi(t) \times \mathbb{P}(X(t) = \omega; t) \quad (5.17)$$

for $(0, 0) \neq \omega \in \Omega_\tau$, and where

$$\varphi(t) = \frac{1}{1 - \mathbb{P}((X_i + Y_i) = 0; t)} = \frac{1}{1 - \mathbb{P}(X(t) = (0, 0); t)} \quad (5.18)$$

“Inflated moments” $\mathbb{E}[\mathbb{P}_N(X(t); t)]$ and $\text{Var}[\mathbb{P}_N(X(t); t)]$ can be calculated from $\mathbb{P}_N(X(t); t)$ in the usual way, and it is these quantities that are entered into the system of ODEs given in Equation (5.14) (Figure 5.4).

This allows us to deterministically model clones with an expectation of a sufficient size, even if their extinction probability is non-negligible. However, the extinction probability of each clone must be recorded. We do this in the following way. When the duration of simulation t_{end} is reached, the extinction probability must be reinstated. Thus the distribution that is modelled around the expectation and variance of the clone must be scaled down accordingly.

5.2.8. Choice of Density Dependency Parameters H and K

Due to competition for resources and lymphocyte capacity, per-capita rates of cell proliferation vary with the total number of cells in the population. Cells will be comparatively free to proliferate when there are fewer cells with which to compete. Proliferation rates for expressing and latent of HTLV-1⁺ cells have been estimated previously [209]. In our model, these rates correspond to $\pi'(t)$ and $p'(t)$ (Equation (5.4)), and so values of K , π

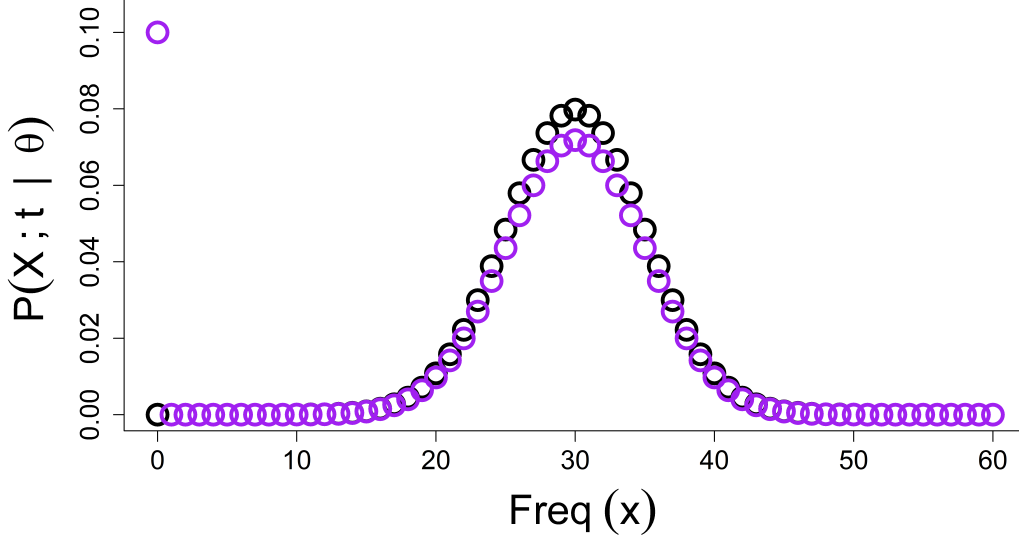


Figure 5.4.: **Schematic of clone promotion and demotion.** An example $\mathbb{P}(X(t); t)$ is shown (purple), along with its corresponding conditional distribution $\mathbb{P}_N(X(t); t)$. Moments of $\mathbb{P}_N(X(t); t)$ are then modelled deterministically using Equation (5.14).

and p must be estimated. Since we are modelling the system at an equilibrium state, we set the death rates of expressing and latent cells to be equal to the proliferation rates.

To determine the carrying capacity parameter K and the proliferation rates π and p , we consider a simple model of cell proliferation. Let x be the number of circulating CD4^+ cells, and let z be the number of HTLV-1 infected cells (note here we do not divide infected cells into clones, or into expressing and latent compartments). We assume proliferation rates of latently infected cells are equal to proliferation rates of uninfected cells. The functions $f(x)$ and $g(z)$ of the proliferation rates of latent and expressing HTLV-1⁺ cells are given by

$$f(x) = \frac{p}{K + x} \quad (5.19a)$$

$$g(z) = \frac{\pi}{K + z} \quad (5.19b)$$

where K determines the extent to which the rate of proliferation changes with population size. $K > 0$ is the frequency at which the proliferation rate is halved. Between them, functions $f(x)$ and $g(z)$ have three parameters (K , π and p). To determine K and p , we need two data points, for which we use published estimates of proliferation rates [209, 210] (values of $f(x)$), together with population size estimates (values of x). Estimates of K and

p will be more accurate if data points cover a wide range of CD4^+ T cell population sizes. Therefore, we use data from ART-untreated HIV patients who are lymphopenic [210] as well as data from HTLV-1 infection [209]. The number N_{CD4} of CD4^+ T cells in the body of a non-lymphopenic patient is estimated as 10^{12} (see Section 2.2.9). We estimate the corresponding number in lymphopenic patients N_{lymph} using the average ratio ρ of normal CD4 counts per μL to CD4 counts in HIV-infected, ART-untreated patients, where data is again taken from [210]. Therefore N_{lymph} is given by

$$N_{\text{lymph}} = \rho \times N_{\text{CD4}} = 6.6 \times 10^{11} \quad (5.20)$$

Let N_{body} be the number of HTLV-1-infected cells in the body. Let r_{lymph} be the CD4^+ T cell proliferation rate in lymphopenic patients. Let r_{CD4} be the CD4^+ T cell proliferation rate in healthy patients (we will use this value as a proxy for the proliferation rate of HTLV-1 latently infected cells). Finally, let $r_{\text{HTLV-1}}$ be the proliferation rate of HTLV-1 infected cells that express viral proteins. Then we solve the following system of equations for p , π and K .

$$r_{\text{lymph}} = \frac{p}{K + N_{\text{lymph}}} \quad (5.21a)$$

$$r_{\text{CD4}} = \frac{p}{K + N_{\text{CD4}}} \quad (5.21b)$$

$$r_{\text{HTLV-1}} = \frac{\pi}{K + N_{\text{body}}} \quad (5.21c)$$

5.2.9. Choice of State Space Upper Limit τ and Frequency F

As we have seen in Section 4.4.2, the need to set a value of τ that is low enough to be computationally feasible must be balanced against the need to not infringe upon the trajectory of a growing clone. Here we outline a general procedure for calculating τ , as well as the frequency F above which it is appropriate to model a species deterministically, where parameters are unknown.

Calculate F

Let Θ be our parameter space. Choose the parameter set $\theta_D \in \Theta$ that most favours death of individuals (in our case this would be the maximum possible values for δ and d , and the lowest possible values of π and p). For an extinction probability $\epsilon_1 > 0$, choose $F \in \mathbb{N}$ to be the highest value such that

$$\mathbb{P}(X = 0; t = h \mid \theta_D, x_0 = F) > \epsilon_1 \quad (5.22)$$

(where $\mathbb{P}(X = 0; t) = \mathbb{P}(X(t) = (0, 0); t)$ in the case of an HTLV-1⁺ clone that is divided into expressing and latent compartments). This definition states that deterministically-modelled clones must be modelled stochastically as soon as their frequency is low enough for extinction to be probable in the next time step of length h . This will be valid for any parameter set $\theta_k \in \Theta$ because of the choice of θ_D .

Calculate τ

Choose the parameter set $\theta_G \in \Theta$ that most favours growth of individuals. Let t_{ext} be the minimum value such that time taken such that

$$\left(\frac{|\mathbb{P}(\text{Ext}; t_{\text{ext}}) - \mathbb{P}(\text{Ext}; t_{\text{ext}} - h)|}{\mathbb{P}(\text{Ext}; t_{\text{ext}} - h)} \right) < \epsilon_2 \quad (5.23)$$

i.e. t_{ext} is the time taken for the extinction probability to change by less than $\epsilon_2 > 0$ and thus be near its asymptote. We choose τ to be the smallest value such that

$$(\mathbb{P}(X + Y = 2\tau; t_{\text{ext}} | \theta_G, (x_0, y_0) = (a, b)) - \mathbb{P}(X + Y = 2\tau - 1; t_{\text{ext}} | \theta_G, (x_0, y_0) = (a, b))) < \epsilon_3 \quad (5.24)$$

for $\epsilon_3 > 0$, for all a, b such that $a + b \leq F$. Note that the choice of time step h contributes to the definitions of τ and F . Having shorter and more numerous time steps takes less time than having longer and fewer time steps, because runtime scales linearly with the number of time steps, but exponentially with the size of the state space. Choosing $\epsilon_1 = 1/1000$, $\epsilon_2 = 0.05$, and $\epsilon_3 = 1/100$, we obtain $\tau = 21$ and $F = 4$ infected cells.

5.2.10. Modelling the Proportion of Expressing Cells

Our HTLV-1 data provide the abundance of clones, but not their respective proportions of expressing and latent cells. We now estimate the proportion of infected cells $\rho_i(t)$ that express viral proteins at time t for each clone i , given the clone rates of proviral expression and silencing. We will apply these proportions to our data to create initial conditions for the expressing and latent compartment of each clone. The following derivation was made by Vikram Sunkara. Let

$$\rho_i(t) = \frac{x_i(t)}{x_i(t) + y_i(t)} \quad (5.25)$$

By the product rule we have

$$\frac{d\rho_i(t)}{dt} = \frac{dx_i(t)}{dt} \frac{1}{x_i(t) + y_i(t)} + x_i(t) \frac{1}{(x_i(t) + y_i(t))^2} \left(\frac{dx_i(t)}{dt} + \frac{dy_i(t)}{dt} \right). \quad (5.26)$$

We examine each term of the above equation separately. Using Equation (5.1a) we have

$$\begin{aligned} \frac{dx_i(t)}{dt} \frac{1}{x_i(t) + y_i(t)} &= \frac{\pi x_i(t)}{(x_i(t) + y_i(t))(K + \sum_{j=1}^{S(t)} x_j(t) + y_j(t))} - \frac{\delta x_i(t)}{x_i(t) + y_i(t)} + \frac{\varepsilon y_i(t)}{x_i(t) + y_i(t)} - \frac{s_i x_i(t)}{x_i(t) + y_i(t)} \\ &= \frac{\pi \rho_i(t)}{K + \sum_{j=1}^{S(t)} x_j(t) + y_j(t)} - \delta \rho_i(t) + \varepsilon_i(1 - \rho_i(t)) - s_i \rho_i(t). \end{aligned} \quad (5.27)$$

We derive an expression for $\left(\frac{dx_i(t)}{dt} + \frac{dy_i(t)}{dt}\right)$ by adding Equations (5.1a) and (5.1b) to obtain

$$\left(\frac{dx_i(t)}{dt} + \frac{dy_i(t)}{dt}\right) = \frac{\pi x_i(t)}{K + \sum_{j=1}^{S(t)} x_j(t) + y_j(t)} - \delta x(t) + \frac{p_i y(t)}{K + \sum_{j=1}^{S(t)} x_j(t) + y_j(t)} - dy_i(t). \quad (5.28)$$

We rewrite $x_i(t) \frac{1}{(x_i(t) + y_i(t))^2} \left(\frac{dx_i(t)}{dt} + \frac{dy_i(t)}{dt}\right)$ as,

$$\rho_i \frac{1}{x_i(t) + y_i(t)} \left(\frac{dx_i(t)}{dt} + \frac{dy_i(t)}{dt}\right).$$

By substituting equation (5.28) into the equation above we get the expression

$$\rho_i(t) \frac{1}{x_i(t) + y_i(t)} \left(\frac{dx_i(t)}{dt} + \frac{dy_i(t)}{dt}\right) = \frac{\rho_i(t)^2 \pi}{K + \sum_{j=1}^{S(t)} x_j(t) + y_j(t)} - \delta \rho_i(t)^2 + \frac{\rho_i(t)(1 - \rho_i(t))p}{K + \sum_{j=1}^{S(t)} x_j(t) + y_j(t)} - \rho_i(t)(1 - \rho_i(t))d. \quad (5.29)$$

We now have an ODE for $\rho_i(t)$. Substituting Equations (5.29) and (5.27) into (5.26) gives

$$\frac{d\rho_i(t)}{dt} = (\rho_i(t) - \rho_i(t)^2) \left(\frac{\pi - p}{K + \sum_{j=1}^{S(t)} x_j(t) + y_j(t)} + (d + \delta) \right) - \rho_i(t)(\varepsilon_i + s_i) + \varepsilon_i \quad (5.30)$$

For a given parameter set, we can predict the evolution the expected proportion of expressing cells over time.

We use Equation (5.30) to estimate initial proportions of the expressing and latent compartments of a clone. Combining all clones together and defining $\varepsilon = \sum_{i=1}^{S(t)} \varepsilon_i$ we obtain $\rho(t) = \sum_{i=1}^{S(t)} \rho_i(t)$. Our numerical simulations show that Equation (5.30) converges to a steady state ρ^* (Figure 5.5), and we calculate

$$x_i(0) = \rho^* \times (x_i(0) + y_i(0)) \quad , \quad y_i(0) = (1 - \rho^*) \times (x_i(0) + y_i(0)) \quad (5.31)$$

5.3. Parameter Inference

We first discuss how to fit the parameters of a hybrid model in an ideal case where there are no unseen clones.

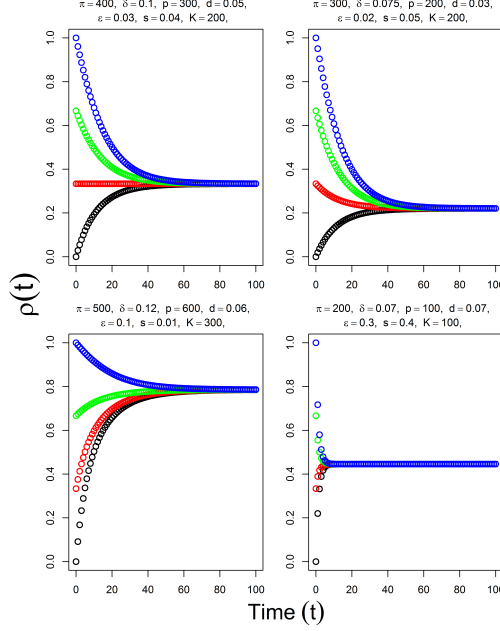


Figure 5.5.: **Proportion of expressing cells over time.** $\rho(t)$ is modelled for four parameter sets, with four initial conditions ($\rho(0) \in \{0, 0.33, 0.67, 1\}$) each. In all cases, $\rho(t)$ reaches a steady state ρ^* .

Let $D_{t_0}, D_{t_1}, \dots, D_{t_M}$ be data from timepoints $t_0, \dots, t_M$, where each $D_{t_j} = \{d_{1,t_j}, \dots, d_{S_j,t_j}\}$ is a collection of clone frequencies. Then, if $Y(t) = ((X_i(t) + Y_i(t))_{i \in S(t)})^T$ is the state vector of each clone's total frequency (expressing plus latent cells), then the likelihood $\mathcal{L}(\theta_k)$ of a set of parameters $\theta_k = \{r_I, \pi, p, \delta, d, s, (\varepsilon_i)_{i \in S(t)}, K\}$ is

$$\mathcal{L}(\theta_k) = \prod_{j=1}^M \mathbb{P}(Y(t) = D_{t_j} ; t_j | \theta_k) \quad (5.32)$$

Assuming independence of clone frequencies and timepoints, this becomes

$$\mathcal{L}(\theta_k) = \prod_{j=1}^M \prod_{i=1}^{S_j} \mathbb{P}((X_i + Y_i) = d_{i,t_j} ; t_j | \theta_k) \quad (5.33)$$

where S_j is the number of clones at time t_j . The parameter choice k_{max} such that $\mathcal{L}(\theta_k)$ is maximal will be our final parameter estimate. We will optimize using the R package FME [98].

5.3.1. Parameter Inference with Unseen Clones

For simplicity, assume that there are only two data points D_1 and D_2 from timepoints t_1 and t_2 . There are two categories of clone data to which we will fit our hybrid model:

1. Clones observed in both D_1 and D_2 . Let $\mathcal{A}$ denote the set of all clones in this category.
2. Clones observed at one timepoint but not the other, and clones not observed at either timepoint. Unobserved clones may be new (i.e. established by infectious spread between time t_0 and t_1) or old. Let $\mathcal{B}$ denote the set of all clones in this category.

Note that $\mathcal{A} \cup \mathcal{B}$ is the set of all modelled clones. We construct the likelihood function differently for each set, so that

$$\mathcal{L}(\theta_k) = \mathcal{L}(\mathcal{A}) \times \mathcal{L}(\mathcal{B}) \quad (5.34)$$

The first category is calculated in exactly the same way as the ideal case above. Since clones in this category are observed at both time points, we can order one of the data sets so that clone i at timepoint 1 has the same integration site as clone i at timepoint 2, we have

$$\mathcal{L}(\mathcal{A}) = \prod_{i=1}^{|\mathcal{A}|} \mathbb{P}((X_i + Y_i) = d_{i,t_j} ; t_j | \theta_k) \quad (5.35)$$

Constructing the likelihood $\mathcal{L}(\mathcal{B})$ of the second category is less straightforward. Because these are unseen clones in both timepoints, we cannot order the data sets so that clones with identical indices have identical integration sites (and are in fact the same clone). Instead, we ask how probable it is that *any* of the clones in $\mathcal{B}$ modelled from timepoint 1 produced *any* of the clones in $\mathcal{B}$ observed in timepoint 2.

Suppose that there are h clones in $\mathcal{B}$ modelled from time t_1 and s clones in $\mathcal{B}$ observed at t_2 . We first assume that there are more modelled clones than observed clones that fall into the second category (i.e. $h > s$). This is very likely to be true for realistic values of infectious spread r_I . This is because, strictly speaking, stochastically modelled clones do not die, but instead have increasingly high probabilities of extinction. If $(d_j)_{j=1}^s$ are the frequencies of clones in $\mathcal{B}$ observed at t_2 , then define $d_{s+1}, \dots, d_h = 0$. If $h < s$, then we can create random variables $X_{h+1}, Y_{h+1}, \dots, X_s, Y_s$ that take frequency 0 with probability 1 and all other frequencies with probability 0 (i.e. each $X_j, Y_j = \delta_{x,0}$). We define, for $d = \max\{s, h\}$,

$$\mathcal{L}(\mathcal{B}) = \sum_{\sigma \in P} \prod_{i=1}^d \mathbb{P}((X_i + Y_i) = D_{\sigma(i)} | \theta_k) \quad (5.36)$$

where P is the set of all permutations on $\{1, \dots, d\}$. Note that if $h < s$, $\mathcal{L}(\mathcal{B}) = 0$, because $\mathbb{P}((X_{h+l} + Y_{h+l}) = d_j) = 0 \ \forall j, \forall l \in \{1, \dots, (s - h)\}$. In practice, this rules out values of infectious spread r_I that are impossibly low. If $h < s$, then there are more clones at timepoint 2 than clones have been created and killed from timepoint 1.

5.3.2. Hybrid Model Propagation

The presence of unseen clones in our data requires that the clone frequencies to which we fit our hybrid model must be estimated, and so we use the DivE distribution generation algorithm (Section 2.2.11). We model the dynamics of clones in the body, and not the blood, because this allows us to model extinction. If zero cells of a particular clone are observed (or estimated using the algorithm) in the blood, this does not necessarily imply that the clone is extinct, as cells from the clone could be in the solid lymphoid tissue. We model clones in the body to avoid this difficulty.

The system described in Equation (5.14) has $5S(t)$ ordinary differential equations, where $S(t)$ is the number of clones at time t . Since the $S(t)$ can exceed 10^5 , having five equations (values of $x_i, y_i, V_{x_i}, V_{y_i}$ and C_{x_i, y_i}) per clone is not feasible. Instead we group clones into categories based on the order of magnitude of their abundance. We group stochastic clones based on their starting frequency. In addition, we allow new clones to have different rates of proviral expression to reflect the heterogeneity of the behaviour of clones with different integration sites.

5.4. Preliminary Results

We now present preliminary results of our hybrid model. Our two outcomes of interest are the total number of infected cells, and the expected number of clones over time (Section 5.2.5). We investigate the behaviour of the hybrid when applied to distributions as estimated by the DivE distribution generation algorithm. Parameters were chosen to explore the validity of the model.

5.4.1. System Death with Zero Proliferation

Unsurprisingly, the system dies when proliferation rates in expressing and latent compartments are set equal to zero (Figure 5.6). What is more interesting is that the expected number of cells decreases far more quickly than does the expected number of clones. This is consistent with our knowledge of the HTLV-1 clonal population structure: large clones constitute the bulk of proviral load; the majority of clones are of low abundance.

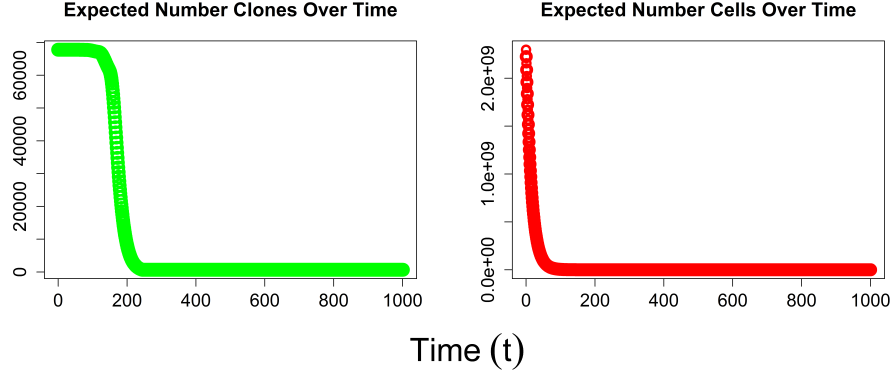


Figure 5.6.: **Example: expected number of cells and clones over time with zero proliferation.** $[\theta_k = \{\pi = p = 0, \delta = 0.1, d = 0.01, \varepsilon = 0.1, s = 0.03, K = 2.2 \times 10^3, r_I = 10^{-6}\}]$ For example patient A, both the number of clones $S(t)$ and the number of cells $N(t)$ decrease to zero, but $N(t)$ decreases far more sharply than $S(t)$.

5.4.2. Effect of Proviral Expression on Clone

The extent of proviral expression can dictate the fate of a clone. Figures 5.7 and 5.8 show clone probability distributions where with $\varepsilon > s$ and $\varepsilon < s$ respectively. Between each figure, π, p, δ are d are unchanged, with $\pi > \delta$ and $p < d$. Hence the only factor in determining the growth or death of a clone is are the ratio between ε and s . Although these are parameters that we have chosen and that are not inferred from empirical data, it is reassuring that in principle the model can account for the reported effects of proviral expression on clonal abundance [22, 42].

5.4.3. “Equilibrium” in Hybrid Model

Figure 5.9 shows the number of clones and the number of cells given a set of parameters $\theta = \{\pi = p = 2.8 \times 10^{10}, \delta = d = 0.2, K = 1000\}$ (resulting in $\pi'(t) = p'(t) = \delta = d = 0.2$), and with $r_I = 0$. Under a purely deterministic formulation of our model, these parameters would result in zero change to the system over time. However, as we have seen in Section 4.4.3, the combination of equal proliferation and death rates with the absorbing state $(0, 0)$ results instead in the slow attrition of the clone, due to the lack of source inflow and due to the strictly increasing extinction probability. The increased probability of extinction feeds into both the expected number of clones $S(t)$ and the number of infected cells $N(t)$. Because $r_I = 0$, no new clones are born, and so both $S(t)$ and $N(t)$ decrease over time, although $N(t)$ decreases by only a relatively small amount.

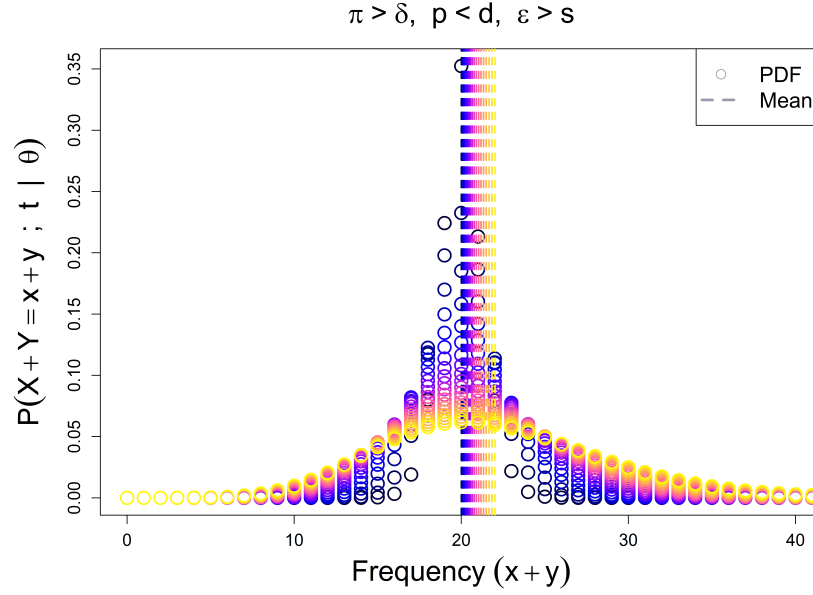


Figure 5.7.: **Proviral expression can drive proliferation.** $\mathbb{P}(X;t)$ is plotted for multiple values of t (circles). Later timepoints are represented by lighter colours. The expectation $\mathbb{E}[\mathbb{P}(X;t)]$ is shown in dashed vertical lines. In the expressing compartment $\pi > \delta$, and in the latent compartment, $p < d$. For $\varepsilon > s$, the expectation increases over time.

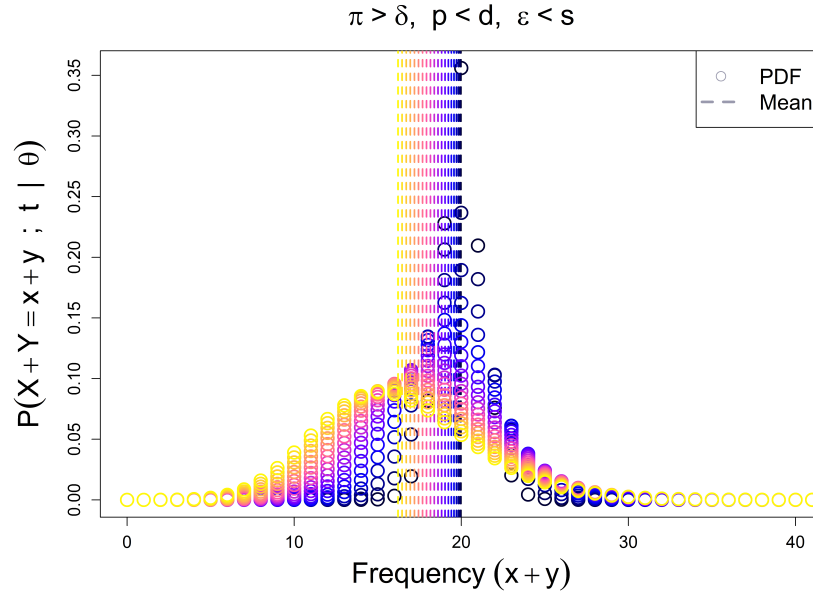


Figure 5.8.: **Silencing can effect clone death.** As for Figure 5.7, but with $\varepsilon < s$. Here the higher silencing rate than expression rate results in clone shrinking.

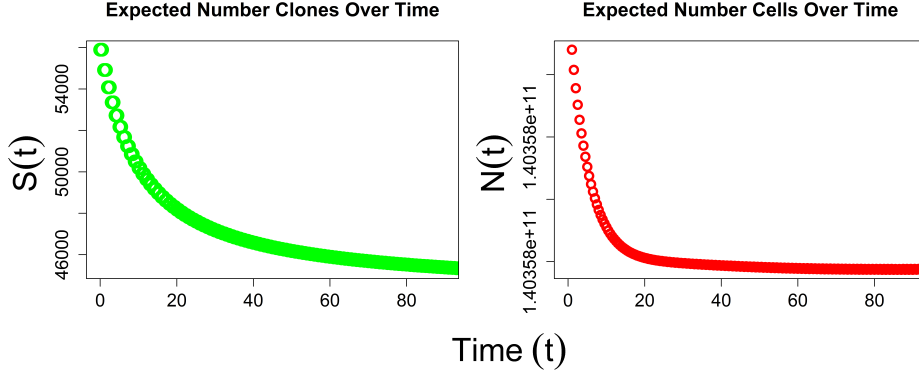


Figure 5.9.: **Number of cells $N(t)$ and clones $S(t)$ at equilibrium, no new infection.** While the system is at “equilibrium” ($\pi'(t) = \delta$ and $p'(t) = d$) the expected number of clones falls as the stochastic clones die away due to the strictly increasing extinction probability and because $r_I = 0$. Thus the number of infected falls too by a relatively small amount.

5.4.4. Effect of Infectious Spread Parameter r_I

Figure 5.10 the evolution of the system for example patient C with parameters $r_I = 10^{-6}$, and where net proliferation rates $\pi'(t)$ and $p'(t)$ are initially greater than death rates $\delta = d = 0.1$ (with $\pi'(0) = p'(0) = 0.2$). Consistent with intuitive expectation, the density dependency of the system confines the existing population of cells to reach a steady state. At very early timepoints, there is a slight but perceptible decrease in the expected number of cells. This is due to the fact that, at time $t = 0$, all clones (stochastically modelled or otherwise) have an extinction probability of zero. Therefore, the instant that the system is propagated, the expected number of clones decreases. Because $r_I > 0$, the number of clones increases over time, even where the number of infected cells reaches a steady state. This is due to the use of a single matrix exponential, the justification for which being that the number of infected cells, and thus the aggregate proliferation rates $\pi'(t)$ and $p(t)$ does not change substantially over time. However, in the above example, for large t_l , $\pi'(t_l) \approx \delta$ and $p(t_l) = d$, but this is not reflected in the stochastically modelled clones. While this problem does not arise when the system is propagated from equilibrium. the above results highlight the importance of updating the matrix exponential between time steps when the system is not at equilibrium.

The model behaves as we would expect, and can produce qualitatively different results depending on the parameters given. Crucially, the model is responsive to the changing the rate of infectious spread r_I , which suggests that the model will be useful in inferring rates of infectious spread with patient data.

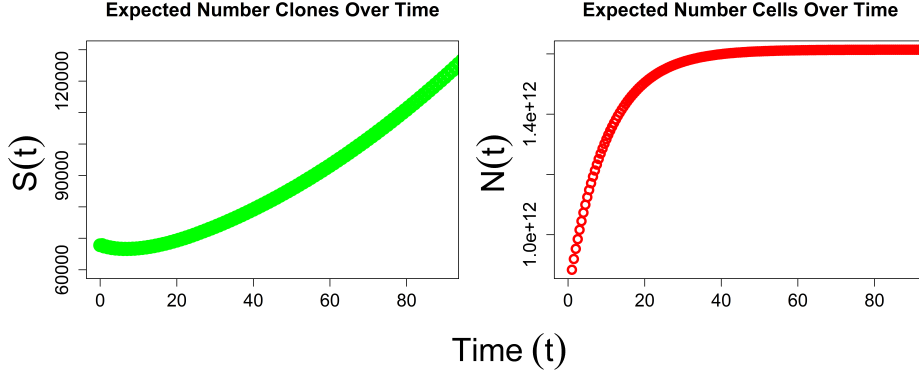


Figure 5.10.: **Example: Effect of rate of infectious spread parameter r_I .** For example patient C the system “grows” ($\pi'(0) = p'(0) = 0.2$, $\delta = d = 0.1$), but the density dependency of the model stops unchecked growth. While the system as a whole, the infectious spread parameter $r_I = 10^{-6}$ results in an increasing expected number of clones $S(t)$ over time.

5.4.5. Effect of Time Step Length

In the preceding sections, the time step length h was chosen to be one day. However, because the deterministic and stochastic components of the hybrid model are not propagated exactly simultaneously (see section 5.2.4), we incur a splitting error, which varies in proportion to h . To ensure that the error is sufficiently small, we run the model for multiple sets of parameters (representing “growth”, “equilibrium” and “death”) and multiple time step lengths. In each instance, we record a number of observables, and h is chosen such that decreasing the time step further does not affect the observables. We choose as our observables the number of clones at a given duration of the hybrid (400 days); and the minimum number of clones over the model run.

In Figure 5.11, each observable is plotted as a function of the inverse of h . Each plot saturates, and the growth parameters require the least stringent (i.e. the longest) time step before saturation. The equilibrium and death parameters require shorter and a greater number of time steps. Since the expected number of clones is affected by h , the choice of h will affect the fitted value of the rate of infectious spread r_I , and the plots suggest that a value of $h = 1/30$ of a day will result in a sufficiently small Strang splitting error.

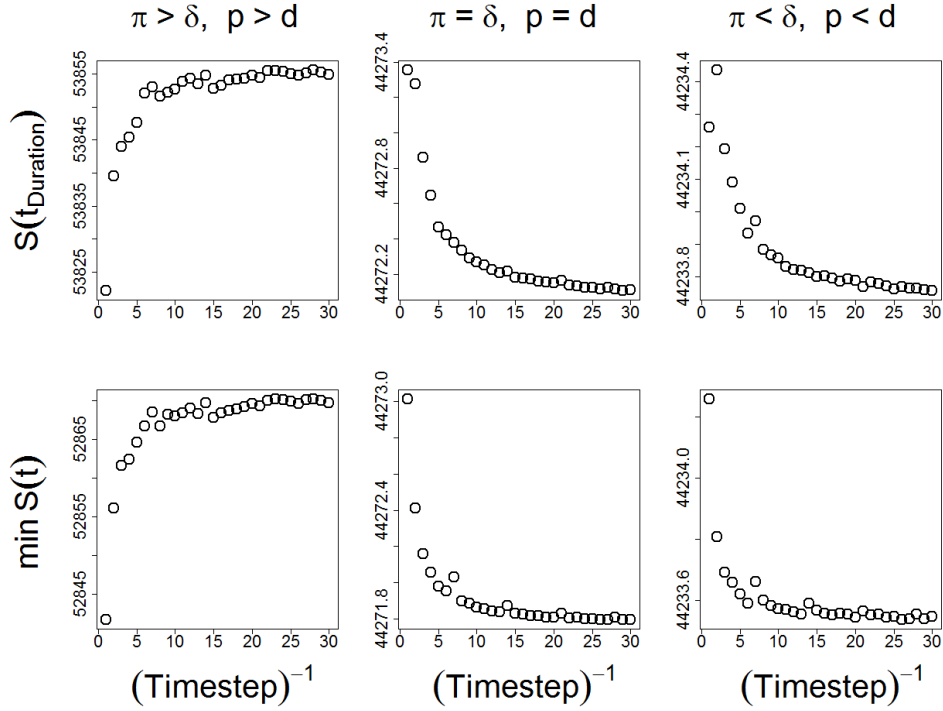


Figure 5.11.: **Effect of Time Step Length.** The effect of changing the length of time step h is examined with reference to: the expected number of clones at a given time (400 days) and the minimum number of clones throughout the model run. Analysis performed for three parameter sets: $\theta_G = \{r_I = 2 \times 10^{-11}, \pi'(0) = p'(0) = 0.2, \delta = d = 0.1\}$ (left column - "growth"); $\theta_E = \{r_I = 2 \times 10^{-11}, \pi'(0) = p'(0) = \delta = d = 0.2\}$ (middle column - "equilibrium"); $\theta_D = \{r_I = 2 \times 10^{-11}, \pi'(0) = p'(0) = 0.2, \delta = d = 0.4\}$ (right column - "death").

5.5. Identifiability of Rate of Infectious Spread: Proof of Principle

The following method was produced by Vikram Sunkara, with input from myself on HTLV-1 biology and simplifying assumptions that can be made.

We now present a proof of principle that the rate of infectious spread r_I can be identified under idealised conditions. We assume that the population is at equilibrium, that the proportion of infected cells that express viral proteins is content, and that there are no unseen clones in any data points.

Jahnke et al [188] gave analytical solutions to the chemical master equation for systems with a highly-restrictive set of monomolecular reactions. These reactions include: i) con-

stant source inflow and ii) mass-action death. Importantly, the analytical solutions do not allow mass-action birth.

We cannot use this analytical solution to describe probability distribution associated with the birth and death of *infected cells* within a clone, because here both birth and death are mass-action reactions, and because there is no source inflow from $\boxed{0}$ to $\boxed{1}$.

However, we can make approximations such that the two types of reactions listed above describe the birth and death of *infected clones*. If the number of expressing cells $x = \sum_{j=1}^{S(t)} x_i$ is constant, then the rate of birth of new clones is given by $\lambda = r_I \times x$, which is also constant.

Clone death will not usually be a mass-action reaction. That is, there is not a constant death rate per clone, since different clones are of different sizes and thus will take differing times to die. For this reason, we require that the system is at equilibrium. At equilibrium, we can more safely assume all clones that die have only recently been born. This enables us to surmise that the only clones that die start from a frequency of one. If we further assume that per-cell rates of viral expression and death do not differ between clones, then all clones have the same death rate β . We now define a stochastic process $Z(t)$ that approximates the number of *new* clones at time t . We have

$$Z(t) := \mathcal{P}(\lambda) - \mathcal{P}\left(\int_0^t \beta Z(s) ds\right) \quad (5.37)$$

The clone death rate β is estimated as the reciprocal of the estimated time to die, given that clones are new and thus started from state $(x_i, y_i) = (0, 1)$. We have

$$\beta = \frac{1}{\mathbb{E}(T : (x = 0, y = 1) \rightarrow (x = 0, y = 0) \mid \theta_k = \{\pi_k, \delta_k, p_k, d_k, \varepsilon_k, s_k, K_k\}, N^*)} \quad (5.38)$$

where $N^* = N(t^*)$ is the total number of infected cells at time t^* when the system has reached a steady state. The expected time to die is calculated from results given in [211].

Equation (5.37) is the master equation that has the following corresponding probability distribution. Using results from [188], it can be shown that the probability of observing $S \in \mathbb{N}_0$ new clones at time t is given by

$$\mathbb{P}(Z(t) = S \mid Z(0) = 0) = \left(\frac{\psi(t)^S e^{-\psi(t)}}{S!} \right) \quad (5.39)$$

where

$$\psi(t) := \frac{\lambda(1 - e^{-\beta t})}{\beta} \quad (5.40)$$

Since $\lambda = r_I$, we can estimate r_I using this probability distribution, together with a maximum likelihood approach so long as the following hold: i) the total number of infected

cells is known; and ii) if the number of new clones S at time t is known. It is the latter condition that requires us to have a complete census. Knowing the number of clones (as estimated by DivE or any other method) is insufficient: we must know which clones are new. We will use the following procedure to determine whether the rate of infectious spread is identifiable.

1. Choose parameters $\theta = \{\pi, \delta, p, d, \varepsilon, s, K\}$ and infectious spread rate r_I .
2. Choose initial clone frequencies.
3. Run the stochastic simulation algorithm [153] with parameters θ and r_I .
4. Take M “samples” from the stochastic simulation algorithm. In each sample D_j , record the time t_j the sample was taken, the total number of infected cells N_j , and the number of new clones Z_j .
5. Using sampled data, estimate via least squares regression parameters $\hat{\theta}$.
6. Using $\hat{\theta}$, calculate $\hat{\beta}$.
7. Estimate parameters $\hat{\lambda}$ and thus $\hat{r}_I$ by maximizing the likelihood function $\mathcal{L}(\hat{r}_I, \hat{\theta})$ given below and assessing the accuracy of $\hat{\lambda}$ as an estimate of λ .

The likelihood of $\mathcal{L}(\hat{r}_I, \hat{\theta})$ is given by

$$\mathcal{L}(\hat{r}_I, \hat{\theta}) = \mathcal{L}(\hat{\lambda}, \hat{\beta}) = \prod_{j=1}^M \mathbb{P}(Z(t) = Z_j \mid \hat{\lambda}, \hat{\beta}) \quad (5.41)$$

The $\hat{\lambda}_{\max}$ that maximizes $\mathcal{L}(\hat{r}_I, \hat{\theta}) = \mathcal{L}(\hat{\lambda}, \hat{\beta})$ is our final estimate of λ . $\hat{r}_I$ is then calculated as

$$\hat{r}_I = \frac{\lambda}{x}$$

where x is estimated as $x = \rho^* N_j$. Thus the estimate of r_I will only be constant if the population is at equilibrium where N_j constant.

Figure 5.12 shows the likelihood of different $\hat{r}_I$ values for a given set of inferred parameters $\hat{\theta}$. The value of $\hat{r}_I$ that maximizes the likelihood is remarkably close to the parameter on which the Gillespie realisation were based.

While the assumptions on which this method is based are unrealistic (we will not have access to as many data points, nor a full clone census at any data point), the preceding analysis demonstrates that the rate of infectious spread is at least identifiable given perfect data, and so serves as an important proof of principle. We are developing the method to account for unseen clones.

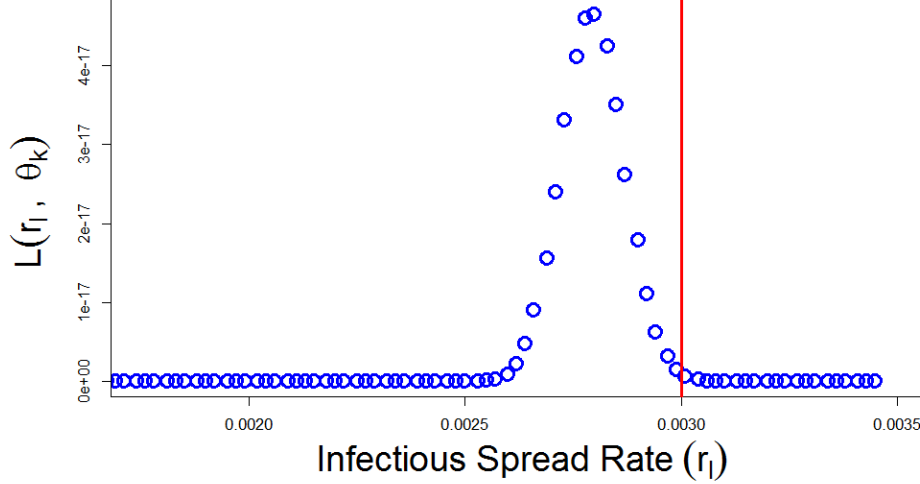


Figure 5.12.: **Example Likelihood Function without unseen clones.** $\mathcal{L}(\hat{r}_I, \theta_k)$ is shown for different values of $\hat{r}_I$ using the procedure outlined in Section 5.5. The value of r_I used to create the *in silico* data from which $\mathcal{L}(\hat{r}_I, \theta_k)$ is calculated is shown in red. Value of $\hat{r}_I$ which maximizes $\mathcal{L}(\hat{r}_I, \theta_k)$ close to original r_I .

5.6. Discussion

In this chapter, we developed a model of the dynamics of within-host HTLV-1 persistence. The model divides the total population of HTLV-1-infected cells first into clones, and second into expressing and latent compartments of each clone. We first gave a purely deterministic formulation of the model, described by a series of ordinary differential equations (ODEs). Because the ODEs are both continuous and deterministic, we presented the model under a stochastic formulation to capture the uncertainty in the fate of a given clone. However, because of the large number of clones and the very high frequencies some clones can take, it is not feasible to model the entire system stochastically, even when the system is substantially simplified by modelling each clone individually.

We therefore explored how to model the system of HTLV-1 clones in a hybrid model of deterministic and stochastic parts that influence each other and are modelled concurrently. Any deterministic modelling here is an approximation of the “true” stochastic process, but in our system, the error incurred in modelling a clone deterministically varies inversely with its frequency. That is, the larger the clone, the more acceptable it is to model the clone deterministically.

Using results from *Engblom et al* [207], our deterministic approximations of the stochastic

system employ a system of ODEs that describe the first and second moments of the probability distribution of each clone. The set of ODEs that describe the expectation of each clone correspond to the original deterministic model.

Clones can radically change in size [11], most notably when small clones invade the population after undergoing malignant transformation. Thus our model has to allow for clones moving between the stochastic and deterministic compartments, and we defined conditions for this movement that were based on the probability of extinction of a given clone.

Preliminary results of the model are encouraging. The model can replicate known features of HTLV-1 infection, as well as more general features of stochastic processes that we explored in Chapter 4. The most important feature of the model is that different values of infectious spread r_I produce noticeably different outcomes. This means that it is highly likely that patient data can be used to infer the rate of infectious spread.

We also outlined an alternative model that aims to show that the rate of infectious spread is at least identifiable in principle, given a complete census of clones and their frequencies at every time point for which we have data. We are investigating extensions to the model that will allow us to infer r_I under less restrictive conditions, for example to allow clones to have different expression rates or where a complete census is unavailable. It will be more difficult to extend the model so that the assumption of equilibrium is not required, as this would imply that the birth rate of clones is no longer constant.

5.6.1. Improvements to Hybrid Model

The ability to model the entire system of all HTLV-1 clones using a single master equation would alleviate many of the difficulties encountered in modelling within-host dynamics of HTLV-1. It would nullify concerns regarding: when it is acceptable to model a clone deterministically; what value of τ would be an acceptable upper bound for a single clone; how to approximate density-dependency in the stochastically modelled clones, and how the extinction probability of a given clone should be incorporated in a hybrid model. However, it is extremely unlikely that computational power will increase to the extent that the use of a single stochastic system would be feasible in the near future.

A more realistic goal is to have a larger value of τ . This would allow us to better estimate the extinction probability and thus achieve more accurate estimates of r_I , and would further make the conditions for promoting a stochastically modelled clone more stringent (Section 5.2.7). Thus a smaller error associated with hybrid modelling would be incurred. It may also be possible to update proliferation rates between time-steps, and this would

more accurately capture density dependency, at the cost of requiring the calculation of the matrix exponential at each time step. Additionally, it would be interesting to investigate whether the assumption of independence of clone frequencies (i.e. that $\text{Cov}[X_i, Y_j] = 0, i \neq j$) is valid, and this could be investigated by including additional covariance terms in the system of ODEs given in Equations (5.14).

Rather than use the per-capita death rate of infected cells δ as a fixed parameter, including δ as a free parameter would help the model to better explain the wide variation between individuals in their proviral load. It would further capture the heterogeneity of an individuals CD8⁺ T cell response to the virus [43]. The reason that we have not done so is because we are keen to reduce the parameter space by as much as possible in pursuit of the rate of infectious spread, although this remains a priority for our future work.

6. Concluding remarks

There is extensive, though incomplete, knowledge of the qualitative behaviour of viral persistence. However, a quantitative understanding is still in its infancy. Indeed it can be argued that it is only relatively recently that a realistic attempt at quantitative understanding has been possible, and this is primarily due to the increased availability and precision of high-throughput sequencing data.

Without good data, perhaps the most that could be said for mathematical models (besides their elegance) is their ability to explicitly and precisely explore the consequences of assumptions. With better data, however, comes the ability to truly compare theory with observation, and so refine it with empirical underpinning. Further, in the absence of perfect or complete data, it is uncontroversial to say that mathematical modelling is and will continue to be an integral part of biology.

In HTLV-1 infection, it has long been known that the virus persists in many distinct clones, which are the aggregate of many years of *de novo* viral replication and infected cell proliferation. The primary motivation for estimating clonal diversity or rates of viral persistence more quantitatively is of course to inform drug development. Another reason though, is that when the idiosyncrasies of each problem are stripped away, they are at their core quite general.

Estimating the number of HTLV-1 clones is at least in principle the same problem as estimating biodiversity, or any problem where the number of *classes* contains important information about a population or process of interest and must be estimated. Estimating rates of HTLV-1 infectious and mitotic spread in the presence of an immune response is similar to estimating rates of “creation”, “copying” and “death”. Therefore, in addition to the problems themselves, the tools and techniques required to adequately address them are also valuable.

Our first aim was to estimate the number of HTLV-1 clones in a single host. Because of the problem’s similarity to the “unseen species problem” in population ecology, we first assessed the suitability of widely-used non-parametric estimators of species richness. We found that each estimator rose systematically with sample size when applied to *in*

silico subsamples, much as estimates of clonal diversity have risen over time as method sensitivity has improved. This provided scant assurance that the diversity estimate from a full data set, which is in essence a subsample of the infected cells in the whole body, would not also increase if the data set was larger. That is, a larger volume of blood would falsely imply a higher diversity.

Perhaps one reason for these methods' widespread use in disciplines outside of their original setting is their apparent simplicity. Each method is straightforward to use, but this belies a complex derivation that makes it difficult to critique on theoretical grounds. Further, given the nature of the problem (we do not know how many species there are) estimator accuracy is difficult to measure, and so it is unsurprising that within immunology and microbiology, adoption of these estimators has been fairly uncritical.

In response to the estimators' poor performance with HTLV-1 data, we developed our own diversity estimator, DivE. The essential idea of this method is to incorporate a failsafe against the systematic bias with sample size that we observed with the previous estimators. DivE fits simple mathematical models to rarefaction data (which plots the estimated diversity against the number of individuals). Importantly, DivE also fits each model to subsamples of that data in order to assess which model is most able to produce consistent estimates while still describing the data to which it was fitted (a constant function would be consistent, but wrong). We found that DivE did not systematically increase with sample size. Furthermore, because DivE requires an input of population size, we were able to assess its accuracy as well as its consistency, and here too it performed well.

DivE can be applied beyond HTLV-1 clonality. We used data provided by our collaborators on T cell receptor repertoires and human microbiota to compare DivE to the ecological estimators. While DivE performed well, the ecological estimators displayed the same bias with sample size that we had observed earlier in the HTLV-1 data, with an interesting exception. When observed rarefaction curves reached a plateau, ecological estimates were largely consistent. However in this case, the saturating curve suggests that further sampling would not yield significantly greater diversity. Thus our analysis suggests that the ecological estimators are consistent only when they are redundant.

A key limitation of DivE became apparent using the TCR data: DivE cannot produce accurate estimates for severely undersampled data. Though this is arguably no different than the principle of "garbage in, garbage out", we attempted to better define under-sampling with the curvature parameter. Our analysis provided guidelines for when it is permissible to use DivE. The requirement of a population size input can be viewed either as a positive or a negative attribute. While we argue such an input should be an essential component of any diversity estimator, it is often non-trivial to estimate and so it would be convenient not to have to estimate this quantity.

Our second aim was to produce a mathematical model of the within-host dynamics of HTLV-1 that describes both infectious and mitotic spread. We aimed to model the evolution of clones at vastly different scale, without ignoring indispensable features of the virus' behaviour. Random effects, a discrete state space, and extinction are all important features when modelling small clones. Because small clones contain the majority of information on infectious spread, we must employ stochastic modelling.

Stochastic processes are a large field and there are many ways in which to model the uncertainty of a particular population. Some of these methods are based on the underlying deterministic process (the signal) but incorporate a noise term. Another method, the Gillespie's Stochastic Simulation Algorithm, randomly chooses the timing of changes ("reactions") of the system, and which change occurs.

While each of these processes might capture the inherent uncertainty we require of our HTLV-1 clones, they have a drawback in that, for identical parameters, different results may be achieved between successive simulations (or realisations). We therefore elected to use the chemical master equation (CME), which describes the rate of change of the probability distribution associated with state of a given system. This is a description of a stochastic process that is not itself susceptible to stochasticity.

We considered birth-death process where extinction, and not merely the possibility that a species temporarily takes a frequency of zero, is possible. In our analyses of such birth-death process, we found that extinction fundamentally changes the nature of the probability distribution. Because numerically solving the CME is time-consuming, we explored approximations that can be made.

When the starting frequency of a species is low, we found that manipulations of the exponential distribution gave excellent representations of the evolving probability distribution, to such a high extent that is worth investigating how this approximation holds when investigated with proper mathematical rigour, most immediately in the case of multiple interacting species. We further found the less surprising result that the Gaussian distribution approximates the probability distribution when starting frequencies are sufficiently high. It will be important to more precisely quantify the conditions under which either approximation is appropriate.

We applied the techniques described above to build a deterministic and stochastic hybrid model. While it is simple to assert that large clones can be modelled deterministically and small clones stochastically, it is difficult in practice to specify how large is large enough. Further, defining how parts of the system (in our case species) can move between the deterministic and stochastic compartments is non-trivial. Ultimately, the solution to both problems is subjective, and we chose to address each problem in terms of the extinction

probability, because our previous analysis had shown this to be the fundamental factor preventing a normal approximation to the evolving probability distribution. Thus “large” clones are those that have a negligible probability of extinction within a given time step, and stochastic clones are moved to the deterministic compartment when there is no change to the extinction probability.

An appealing quality of our hybrid model is that it is not susceptible to distortion by sampling variation, and thus the relationship between assumption and consequence is preserved. Preliminary results suggest that the model captures fundamental dynamics of HTLV-1 infection. If proliferation values are zero, then unsurprisingly the system dies away. Interestingly though, the expected number of clones decreases far more slowly than the expected number of cells, consistent with experimental observation whereby most clones contain few cells, and a few clones contain most cells.

The hybrid model also displays markedly different patterns in the expected number of clones over time for different values of the rate of infectious spread, strongly implying that it will be possible to infer this rate from empirical patient data. If so, this would represent a substantial advance in our knowledge of HTLV-1 persistence.

We are investigating how our hybrid model could be applied to investigate early HTLV-1 infection, and the amendments that this would require. For example, the assumption of equilibrium would no longer be appropriate, and this would mean that rates of proliferation would no longer be constant, as we have assumed for stochastic clones.

Outside of HTLV-1, there are several potential applications of our work, in particular modelling the human T cell receptor (TCR) repertoire. The mechanisms by which the immune system is reconstituted after immune suppression or transplantation are poorly understood. Drawing parallels between this system and HTLV-1 infectious and mitotic spread, we would like to investigate the extent to which re-constitution occurs through the generation of new TCR clonotypes, or through the increased expansion of existing clonotypes.

In population genetics, our model could be employed to estimate the rates of mutation and copying of genes throughout generations, and could be used to model invasion and eventual fixation of a particular gene. A very interesting application would be to measure the extent of PCR error [212]. One can argue that PCR is essentially a birth-death process with the presence of errors, each of which create a new “species” of amplicon or PCR product. Quantifying the error-rate would perhaps facilitate correction of it. Finally, HIV-1 also exhibits proviral integration, and so one would expect that the system of HIV clones would behave in a similar manner to HTLV-1 clones, raising implications for HIV-1

latency. We would expect then that the ratio of infectious to mitotic spread in HTLV-1 would be reversed in HIV-1.

It is important to stress that even if the analogy of copying, creating and death can only be extended so far, and if the biological processes and models of these process are very different, then it is still possible that the machinery of our hybrid, whereby deterministic compartments interact with stochastic compartments, will be of use. For example, the movement of species between deterministic and stochastic compartments is defined in terms of the probability distribution and its interaction with the upper limits of the state space, not the model itself. This means that our analysis of interacting compartments is potentially as applicable as the actual model.

This work was only possible thanks to frequent collaboration between ourselves and many experimentalists, who generously shared their data and their time. I hope that the transaction has not been zero-sum. According to Blaise Pascal, “As we speak of poetical beauty, so ought we to speak of mathematical beauty and medical beauty” [213]. I cannot argue with the great man, although it is probably wise not to invite judgement on whether the preceding pages warrant such description.

Bibliography

- [1] R. C. Gallo, “History of the discoveries of the first human retroviruses: HTLV-1 and HTLV-2,” *Oncogene*, vol. 24, no. 39, pp. 5926–5930, 2005.
- [2] F. A. Proietti, A. B. F. Carneiro-Proietti, B. C. Catalan-Soares, and E. L. Murphy, “Global epidemiology of HTLV-1 infection and associated diseases,” *Oncogene*, vol. 24, no. 39, pp. 6058–6068, 2005.
- [3] A. Gessain and R. Mahieux, “Epidemiology, origin and genetic diversity of HTLV-1 retrovirus and STLV-1 simian affiliated retrovirus,” *Bulletin de la Societe de pathologie exotique (1990)*, vol. 93, no. 3, pp. 163–171, 2000.
- [4] A. Gessain and O. Cassar, “Epidemiological aspects and world distribution of HTLV-1 infection,” *Frontiers in microbiology*, vol. 3, 2012.
- [5] C. R. Bangham, “HTLV-1 infections,” *Journal of clinical pathology*, vol. 53, no. 8, pp. 581–586, 2000.
- [6] Y. Satou and M. Matsuoka, “Virological and immunological mechanisms in the pathogenesis of human T-cell leukemia virus type 1,” *Reviews in medical virology*, vol. 23, no. 5, pp. 269–280, 2013.
- [7] H. A. Niederer, D. J. Laydon, A. Melamed, M. Elemans, B. Asquith, M. Matsuoka, and C. R. Bangham, “HTLV-1 proviral integration sites differ between asymptomatic carriers and patients with HAM/TSP,” *Virology journal*, vol. 11, no. 1, p. 172, 2014.
- [8] Y. Iwasaki, “Pathology of chronic myelopathy associated with HTLV-I infection (HAM/TSP),” *Journal of the neurological sciences*, vol. 96, no. 1, pp. 103–123, 1990.
- [9] A. Hodson, D. J. Laydon, B. J. Bain, P. A. Fields, and G. P. Taylor, “Pre-morbid human T-lymphotropic virus type I proviral load, rather than percentage of abnormal lymphocytes, is associated with an increased risk of aggressive adult T-cell leukemia/lymphoma,” *Haematologica*, vol. 98, no. 3, pp. 385–388, 2013.

- [10] P. S. Gill, W. Harrington Jr, M. H. Kaplan, R. C. Ribeiro, J. M. Bennett, H. A. Liebman, M. Bernstein-Singer, B. M. Espina, L. Cabral, S. Allen, *et al.*, “Treatment of adult T-cell leukemia–lymphoma with a combination of interferon alfa and zidovudine,” *New England Journal of Medicine*, vol. 332, no. 26, pp. 1744–1748, 1995.
- [11] L. B. Cook, A. Melamed, H. Niederer, M. Valganon, D. J. Laydon, L. Foroni, G. P. Taylor, M. Matsuoka, and C. R. M. Bangham, “The role of HTLV-1 clonality, proviral structure and genomic integration site in adult T cell leukemia/lymphoma,” *Blood*, 2014.
- [12] M. Nagai, K. Usuku, W. Matsumoto, D. Kodama, N. Takenouchi, T. Moritoyo, S. Hashiguchi, M. Ichinose, C. R. Bangham, S. Izumo, *et al.*, “Analysis of HTLV-1 proviral load in 202 HAM/TSP patients and 243 asymptomatic HTLV-1 carriers: high proviral load strongly predisposes to HAM/TSP,” *Journal of neurovirology*, vol. 4, no. 6, pp. 586–593, 1998.
- [13] A. Okayama, S. Stuver, M. Matsuoka, J. Ishizaki, G.-i. Tanaka, Y. Kubuki, N. Mueller, C.-c. Hsieh, N. Tachibana, and H. Tsubouchi, “Role of HTLV-1 proviral DNA load and clonality in the development of adult T-cell leukemia/lymphoma in asymptomatic carriers,” *International Journal of Cancer*, vol. 110, no. 4, pp. 621–625, 2004.
- [14] D. Wodarz, M. A. Nowak, and C. R. Bangham, “The dynamics of HTLV-1 and the CTL response,” *Immunology Today*, vol. 20, no. 5, pp. 220 – 227, 1999.
- [15] C. R. M. Bangham, “The immune control and cell-to-cell spread of human T-lymphotropic virus type 1,” *Journal of General Virology*, vol. 84, no. 12, pp. 3177–3189, 2003.
- [16] K. Okochi and H. Sato, “Transmission of ATL (HTLV-1) through blood transfusion,” *Retroviruses in human lymphoma/leukemia. Japan Scientific Society Press, Tokyo, Japan*, pp. 129–135, 1985.
- [17] T. Igakura, J. C. Stinchcombe, P. K. C. Goon, G. P. Taylor, J. N. Weber, G. M. Griffiths, Y. Tanaka, M. Osame, and C. R. M. Bangham, “Spread of HTLV-1 between lymphocytes by virus-induced polarization of the cytoskeleton,” *Science*, vol. 299, no. 5613, pp. 1713–1716, 2003.
- [18] M. Nejmeddine and C. R. Bangham, “The HTLV-1 virological synapse,” *Viruses*, vol. 2, no. 7, pp. 1427–1447, 2010.

- [19] L. M. Mansky, "In vivo analysis of human T-cell leukemia virus type 1 reverse transcription accuracy," *Journal of Virology*, vol. 74, no. 20, pp. 9525–9531, 2000.
- [20] E. Wattel, J. P. Vartanian, C. Pannetier, and S. Wain-Hobson, "Clonal expansion of human T-cell leukemia virus type I-infected cells in asymptomatic and symptomatic carriers without malignancy," *Journal of Virology*, vol. 69, no. 5, pp. 2863–8, 1995.
- [21] G. P. Taylor, S. E. Hall, S. Navarrete, C. A. Michie, R. Davis, A. D. Witkover, M. Rossor, M. A. Nowak, P. Rudge, E. Matutes, C. R. M. Bangham, and J. N. Weber, "Effect of lamivudine on human T-cell leukemia virus type 1 (HTLV-1) DNA copy number, T-cell phenotype, and anti-tax cytotoxic T-cell frequency in patients with HTLV-1-associated myelopathy," *Journal of Virology*, vol. 73, no. 12, pp. 10289–10295, 1999.
- [22] N. A. Gillet, N. Malani, A. Melamed, N. Gormley, R. Carter, D. Bentley, C. Berry, F. D. Bushman, G. P. Taylor, and C. R. Bangham, "The host genomic environment of the provirus determines the abundance of HTLV-1-infected T cell clones," *Blood*, vol. 117, no. 11, pp. 3113–3122, 2011.
- [23] Y. Furukawa, J. Fujisawa, M. Osame, M. Toita, S. Sonoda, R. Kubota, S. Ijichi, and M. Yoshida, "Frequent clonal proliferation of human T-cell leukemia virus type 1 (HTLV-1)-infected T cells in HTLV-1-associated myelopathy (HAM-TSP)," *Blood*, vol. 80, no. 4, pp. 1012–1016, 1992.
- [24] E. Wattel, M. Cavrois, A. Gessain, and S. Wain-Hobson, "Clonal expansion of infected cells: a way of life for HTLV-1," *JAIDS Journal of Acquired Immune Deficiency Syndromes*, vol. 13, pp. S92–S99, 1996.
- [25] K. N. Meekings, J. Leipzig, F. D. Bushman, G. P. Taylor, and C. R. Bangham, "HTLV-1 integration into transcriptionally active genomic regions is associated with proviral expression and with HAM/TSP," *PLoS Pathog*, vol. 4, no. 3, p. e1000027, 2008.
- [26] C. C. Berry, N. A. Gillet, A. Melamed, N. Gormley, C. R. Bangham, and F. D. Bushman, "Estimating abundances of retroviral insertion sites from DNA fragment length data," *Bioinformatics*, vol. 28, no. 6, pp. 755–762, 2012.
- [27] D. Mendoza, C. Royce, L. E. Ruff, D. R. Ambrozak, M. F. Quigley, T. Dang, V. Venturi, D. A. Price, D. C. Douek, S. A. Migueles, *et al.*, "HLA B* 5701+ long-term nonprogressors/elite controllers are not distinguished from progressors by the clonal composition of HIV-specific CD8+ T-cells," *Journal of virology*, pp. JVI–06982, 2012.

- [28] K. Sim, M. J. Cox, H. Wopereis, R. Martin, J. Knol, M.-S. Li, W. O. Cookson, M. F. Moffatt, and J. S. Kroll, “Improved detection of Bifidobacteria with optimised 16S rRNA-gene based pyrosequencing,” *PLoS One*, vol. 7, no. 3, p. e32543, 2012.
- [29] D. K. Dunn-Walters and A. A. Ademokun, “B cell repertoire and ageing,” *Current opinion in immunology*, vol. 22, no. 4, pp. 514–520, 2010.
- [30] A. Ademokun, Y.-C. Wu, V. Martin, R. Mitra, U. Sack, H. Baxendale, D. Kipling, and D. K. Dunn-Walters, “Vaccination-induced changes in human B-cell repertoire and pneumococcal IgM and IgA antibody at different ages,” *Aging cell*, vol. 10, no. 6, pp. 922–930, 2011.
- [31] S. D. Boyd, Y. Liu, C. Wang, V. Martin, and D. K. Dunn-Walters, “Human lymphocyte repertoires in ageing,” *Current Opinion in Immunology*, vol. 25, no. 4, pp. 511 – 515, 2013. Host pathogens / Immune senescence.
- [32] I. Messaoudi, J. A. G. Patino, R. Dyal, J. LeMaout, and J. Nikolich-Zugich, “Direct link between MHC polymorphism, T cell avidity, and diversity in immune defence,” *Science*, vol. 298, no. 5599, pp. 1797–1800, 2002.
- [33] M. P. Davenport, D. A. Price, and A. J. McMichael, “The T cell repertoire in infection and vaccination: implications for control of persistent viruses,” *Current opinion in immunology*, vol. 19, no. 3, pp. 294–300, 2007.
- [34] S. Ott, M. Musfeldt, D. Wenderoth, J. Hampe, O. Brant, U. Fölsch, K. Timmis, and S. Schreiber, “Reduction in diversity of the colonic mucosa associated bacterial microflora in patients with active inflammatory bowel disease,” *Gut*, vol. 53, no. 5, pp. 685–693, 2004.
- [35] P. Seksik, L. Rigottier-Gois, G. Gramet, M. Sutren, P. Pochart, P. Marteau, R. Jian, and J. Dore, “Alterations of the dominant faecal bacterial groups in patients with Crohn’s disease of the colon,” *Gut*, vol. 52, no. 2, pp. 237–242, 2003.
- [36] M. Hidaka, J. Inoue, M. Yoshida, and M. Seiki, “Post-transcriptional regulator (rex) of HTLV-1 initiates expression of viral structural proteins but suppresses expression of regulatory proteins,” *The EMBO journal*, vol. 7, no. 2, p. 519, 1988.
- [37] C. R. Bangham, “The immune response to HTLV-I,” *Current Opinion in Immunology*, vol. 12, no. 4, pp. 397 – 402, 2000.
- [38] M. Nowak and R. M. May, *Virus Dynamics: Mathematical Principles of Immunology and Virology: Mathematical Principles of Immunology and Virology*. Oxford university press, 2000.

- [39] C. R. Bangham and M. Osame, "Cellular immune response to HTLV-1," *Oncogene*, vol. 24, no. 39, pp. 6035–6046, 2005.
- [40] M. A. Nowak and C. R. M. Bangham, "Population dynamics of immune responses to persistent viruses," *Science*, vol. 272, no. 5258, pp. 74–79, 1996.
- [41] P. K. Goon, A. Biancardi, N. Fast, T. Igakura, E. Hanon, A. J. Mosley, B. Asquith, K. G. Gould, S. Marshall, G. P. Taylor, *et al.*, "Human cell lymphotropic virus (HTLV) type-I-specific CD8+ cells: Frequency and immunodominance hierarchy," *Journal of Infectious Diseases*, vol. 189, no. 12, pp. 2294–2298, 2004.
- [42] A. Melamed, D. J. Laydon, N. A. Gillet, Y. Tanaka, G. P. Taylor, and C. R. M. Bangham, "Genome-wide determinants of proviral targeting, clonal abundance and expression in natural HTLV-1 infection," *PLoS Pathog*, vol. 9, p. e1003271, 03 2013.
- [43] A. MacNamara, A. Rowan, S. Hilburn, U. Kadolsky, H. Fujiwara, K. Suemori, M. Yasukawa, G. Taylor, C. R. M. Bangham, and B. Asquith, "HLA class I binding of HBZ determines outcome in HTLV-1 infection," *PLoS Pathog*, vol. 6, p. e1001117, 09 2010.
- [44] M. Saito, T. Matsuzaki, Y. Satou, J.-i. Yasunaga, K. Saito, K. Arimura, M. Matsuoka, Y. Ohara, *et al.*, "In vivo expression of the HBZ gene of HTLV-1 correlates with proviral load, inflammatory markers and disease severity in HTLV-1 associated myelopathy/tropical spastic paraparesis (HAM/TSP)," *Retrovirology*, vol. 6, no. 19, pp. 720–725, 2009.
- [45] S. M. Lewis, B. J. Bain, I. Bates, and J. V. Dacie, *Dacie and Lewis practical haematology*. Elsevier Health Sciences, 2006.
- [46] N.-K. Seich al Basatena, A. MacNamara, A. M. Vine, C. L. Thio, J. Astemborski, K. Usuku, M. Osame, G. D. Kirk, S. M. Donfield, J. J. Goedert, C. R. Bangham, M. Carrington, S. I. Khakoo, and B. Asquith, "KIR2DL2 enhances protective and detrimental HLA class I-mediated immunity in chronic viral infection," *PLoS Pathog*, vol. 7, p. e1002270, 10 2011.
- [47] T. Kattan, A. MacNamara, A. G. Rowan, H. Nose, A. J. Mosley, Y. Tanaka, G. P. Taylor, B. Asquith, and C. R. Bangham, "The avidity and lytic efficiency of the CTL response to HTLV-1," *The Journal of Immunology*, vol. 182, no. 9, pp. 5723–5729, 2009.
- [48] C. R. M. Bangham, "CTL quality and the control of human retroviral infections," *European Journal of Immunology*, vol. 39, no. 7, pp. 1700–1712, 2009.

- [49] G. P. Wang, S. A. Sherrill-Mix, K.-M. Chang, C. Quince, and F. D. Bushman, “Hepatitis C virus transmission bottlenecks analyzed by deep sequencing,” *Journal of virology*, vol. 84, no. 12, pp. 6218–6228, 2010.
- [50] B. N. Bimber, B. J. Burwitz, S. O’Connor, A. Detmer, E. Gostick, S. M. Lank, D. A. Price, A. Hughes, and D. O’Connor, “Ultradeep pyrosequencing detects complex patterns of CD8+ T-lymphocyte escape in simian immunodeficiency virus-infected macaques,” *Journal of virology*, vol. 83, no. 16, pp. 8247–8253, 2009.
- [51] C.-A. Siegrist and R. Aspinall, “B-cell responses to vaccination at the extremes of age,” *Nature Reviews Immunology*, vol. 9, no. 3, pp. 185–194, 2009.
- [52] E. J. Yager, M. Ahmed, K. Lanzer, T. D. Randall, D. L. Woodland, and M. A. Blackman, “Age-associated decline in T cell repertoire diversity leads to holes in the repertoire and impaired immunity to influenza virus,” *The Journal of experimental medicine*, vol. 205, no. 3, pp. 711–723, 2008.
- [53] H. Chen, Z. M. Ndhlovu, D. Liu, L. C. Porter, J. W. Fang, S. Darko, M. A. Brockman, T. Miura, Z. L. Brumme, A. Schneidewind, *et al.*, “TCR clonotypes modulate the protective effect of HLA class I molecules in HIV-1 infection,” *Nature immunology*, vol. 13, no. 7, pp. 691–700, 2012.
- [54] M. Wang, C. Karlsson, C. Olsson, I. Adlerberth, A. E. Wold, D. P. Strachan, P. M. Martriacardi, N. Åberg, M. R. Perkin, S. Tripodi, *et al.*, “Reduced diversity in the early fecal microbiota of infants with atopic eczema,” *Journal of Allergy and Clinical Immunology*, vol. 121, no. 1, pp. 129–134, 2008.
- [55] S. M. Scheiner, “Six types of species-area curves,” *Global Ecology and Biogeography*, vol. 12, no. 6, pp. 441–447, 2003.
- [56] R. K. Colwell and J. A. Coddington, “Estimating terrestrial biodiversity through extrapolation,” *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, vol. 345, no. 1311, pp. 101–118, 1994.
- [57] R. K. Colwell, A. Chao, N. J. Gotelli, S.-Y. Lin, C. X. Mao, R. L. Chazdon, and J. T. Longino, “Models and estimators linking individual-based and sample-based rarefaction, extrapolation and comparison of assemblages,” *Journal of Plant Ecology*, vol. 5, no. 1, pp. 3–21, 2012.
- [58] N. J. Gotelli and R. K. Colwell, “Estimating species richness,” *Biological diversity: frontiers in measurement and assessment*, pp. 39–54, 2011.

- [59] R. M. May *et al.*, “How many species are there on earth?,” *Science(Washington)*, vol. 241, no. 4872, pp. 1441–1449, 1988.
- [60] S.-H. Hong, J. Bunge, S.-O. Jeon, and S. S. Epstein, “Predicting microbial species richness,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 1, pp. 117–122, 2006.
- [61] J. C. Tipper, “Rarefaction and rarefaction-the use and abuse of a method in paleoecology,” *Paleobiology*, pp. 423–434, 1979.
- [62] E. W. Fager, “Diversity: a sampling study,” *American Naturalist*, pp. 293–310, 1972.
- [63] R. C. Lewontin and T. Prout, “Estimation of the number of different classes in a population,” *Biometrics*, vol. 12, no. 2, pp. 211–223, 1956.
- [64] J. Darroch, “The multiple-recapture census: I. estimation of a closed population,” *Biometrika*, pp. 343–359, 1958.
- [65] B. C. Arnold and R. J. Beaver, “Estimation of the number of classes in a population,” *Biometrical journal*, vol. 30, no. 4, pp. 413–424, 1988.
- [66] J. Bunge and M. Fitzpatrick, “Estimating the number of species: a review,” *Journal of the American Statistical Association*, vol. 88, no. 421, pp. 364–373, 1993.
- [67] L. Holst, “Some asymptotic results for incomplete multinomial or Poisson samples,” *Scandinavian Journal of Statistics*, pp. 243–246, 1981.
- [68] V. Kalinin, “Functionals related to the Poisson distribution and statistical structure of a text,” *Articles on Mathematical Statistics and the Theory of Probability*, pp. 202–220, 1965.
- [69] D. R. McNeil, “Estimating an author’s vocabulary,” *Journal of the American Statistical Association*, vol. 68, no. 341, pp. 92–96, 1973.
- [70] G. E. Willmot, “The Poisson-inverse Gaussian distribution as an alternative to the negative binomial,” *Scandinavian Actuarial Journal*, vol. 1987, no. 3-4, pp. 113–127, 1987.
- [71] A. Chao, “Nonparametric estimation of the number of classes in a population,” *Scandinavian Journal of statistics*, pp. 265–270, 1984.
- [72] A. Chao and S.-M. Lee, “Estimating the number of classes via sample coverage,” *Journal of the American statistical Association*, vol. 87, no. 417, pp. 210–217, 1992.

- [73] A. Chao, P. C. Li, S. Agatha, and W. Foissner, "A statistical approach to estimate soil ciliate diversity and distribution based on data from five continents," *Oikos*, vol. 114, no. 3, pp. 479–493, 2006.
- [74] P. B. Eckburg, E. M. Bik, C. N. Bernstein, E. Purdom, L. Dethlefsen, M. Sargent, S. R. Gill, K. E. Nelson, and D. A. Relman, "Diversity of the human intestinal microbial flora," *Science*, vol. 308, no. 5728, pp. 1635–1638, 2005.
- [75] J. Qin, R. Li, J. Raes, M. Arumugam, K. S. Burgdorf, C. Manichanh, T. Nielsen, N. Pons, F. Levenez, T. Yamada, *et al.*, "A human gut microbial gene catalogue established by metagenomic sequencing," *Nature*, vol. 464, no. 7285, pp. 59–65, 2010.
- [76] J. Wong, D. Mathis, and C. Benoist, "TCR-based lineage tracing: no evidence for conversion of conventional into regulatory T cells in response to a natural self-antigen in pancreatic islets," *The Journal of experimental medicine*, vol. 204, no. 9, pp. 2039–2045, 2007.
- [77] R. Pacholczyk, H. Ignatowicz, P. Kraj, and L. Ignatowicz, "Origin and T cell receptor diversity of Foxp3+ CD4+CD25+ T cells," *Immunity*, vol. 25, no. 2, pp. 249–259, 2006.
- [78] M. Unterseher, M. Schnittler, C. Dormann, and A. Sickert, "Application of species richness estimators for the assessment of fungal diversity," *FEMS Microbiology Letters*, vol. 282, no. 2, pp. 205–213, 2008.
- [79] A. Chao, "Species estimation and applications," in *Encyclopedia of Statistical Sciences* (B. V. N. Balakrishnan, C.B. Read, ed.), Wiley Online Library, 2005.
- [80] N. L. La Gruta, W. T. Rothwell, T. Cukalac, N. G. Swan, S. A. Valkenburg, K. Kedzierska, P. G. Thomas, P. C. Doherty, S. J. Turner, *et al.*, "Primary CTL response magnitude in mice is determined by the extent of naive T cell recruitment and subsequent clonal expansion," *The Journal of clinical investigation*, vol. 120, no. 6, pp. 1885–1894, 2010.
- [81] M. Shugay, D. A. Bolotin, E. V. Putintseva, M. V. Pogorelyy, I. Z. Mamedov, and D. M. Chudakov, "Huge overlap of individual TCR beta repertoires," *Frontiers in immunology*, vol. 4, 2013.
- [82] J. A. Bailey, T. Mvalo, N. Aragam, M. Weiser, S. Congdon, D. Kamwendo, F. Martinson, I. Hoffman, S. R. Meshnick, and J. J. Juliano, "Use of massively parallel

- pyrosequencing to evaluate the diversity of and selection on *Plasmodium falciparum* csp T-cell epitopes in Lilongwe, Malawi,” *Journal of Infectious Diseases*, vol. 206, no. 4, pp. 580–587, 2012.
- [83] T. Frisli, T. Haverkamp, K. Jakobsen, N. C. Stenseth, and K. Rudi, “Estimation of metagenome size and structure in an experimental soil microbiota from low coverage next-generation sequence data,” *Journal of applied microbiology*, vol. 114, no. 1, pp. 141–151, 2013.
- [84] G. P. Wang, A. Garrigue, A. Ciuffi, K. Ronen, J. Leipzig, C. Berry, C. Lagresle-Peyrou, F. Benjelloun, S. Hacein-Bey-Abina, A. Fischer, *et al.*, “DNA bar coding and pyrosequencing to analyze adverse events in therapeutic gene transfer,” *Nucleic acids research*, vol. 36, no. 9, pp. e49–e49, 2008.
- [85] P. D. Schloss and J. Handelsman, “Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness,” *Applied and environmental microbiology*, vol. 71, no. 3, pp. 1501–1506, 2005.
- [86] E. P. Smith and G. van Belle, “Nonparametric estimation of species richness,” *Biometrics*, pp. 119–129, 1984.
- [87] I. J. Good, “The population frequencies of species and the estimation of population parameters,” *Biometrika*, vol. 40, no. 3-4, pp. 237–264, 1953.
- [88] C. Palmer, E. M. Bik, D. B. DiGiulio, D. A. Relman, and P. O. Brown, “Development of the human infant intestinal microbiota,” *PLoS biology*, vol. 5, no. 7, p. e177, 2007.
- [89] J. J. Miles, D. C. Douek, and D. A. Price, “Bias in the $\alpha\beta$ T-cell repertoire: implications for disease pathogenesis and vaccination,” *Immunology and cell biology*, vol. 89, no. 3, pp. 375–387, 2011.
- [90] T. P. Arstila, A. Casrouge, V. Baron, J. Even, J. Kanellopoulos, and P. Kourilsky, “A direct estimate of the human $\alpha\beta$ T cell receptor diversity,” *Science*, vol. 286, no. 5441, pp. 958–961, 1999.
- [91] H. S. Robins, P. V. Campregher, S. K. Srivastava, A. Wacher, C. J. Turtle, O. Kahsai, S. R. Riddell, E. H. Warren, and C. S. Carlson, “Comprehensive assessment of t-cell receptor β -chain diversity in $\alpha\beta$ t cells,” *Blood*, vol. 114, no. 19, pp. 4099–4107, 2009.
- [92] D. A. Price, S. M. West, M. R. Betts, L. E. Ruff, J. M. Brenchley, D. R. Ambrozak, Y. Edghill-Smith, M. J. Kuroda, D. Bogdan, K. Kunstman, *et al.*, “T cell receptor recognition motifs govern immune escape patterns in acute SIV infection,” *Immunity*, vol. 21, no. 6, pp. 793–803, 2004.

- [93] D. B. Rusch, A. C. Martiny, C. L. Dupont, A. L. Halpern, and J. C. Venter, “Characterization of *Prochlorococcus* clades from iron-depleted oceanic regions,” *Proceedings of the National Academy of Sciences*, vol. 107, no. 37, pp. 16184–16189, 2010.
- [94] D. B. Rusch, A. L. Halpern, G. Sutton, K. B. Heidelberg, S. Williamson, S. Yooseph, D. Wu, J. A. Eisen, J. M. Hoffman, K. Remington, *et al.*, “The Sorcerer II global ocean sampling expedition: northwest Atlantic through eastern tropical Pacific,” *PLoS biology*, vol. 5, no. 3, p. e77, 2007.
- [95] J. Phillips, “Zunzun.com: Online curve fitting and surface fitting web site,” June 2012.
- [96] C. Flather, “Fitting species-accumulation functions and assessing regional land use impacts on avian diversity,” *Journal of Biogeography*, vol. 23, no. 2, pp. 155–168, 1996.
- [97] R Development Core Team, *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2010. ISBN 3-900051-07-0.
- [98] K. Soetaert and T. Petzoldt, “Inverse modelling, sensitivity and monte carlo analysis in R using package FME,” *Journal of Statistical Software*, vol. 33, no. 3, pp. 1–28, 2010.
- [99] W. L. Price, “A controlled random search procedure for global optimisation,” *The Computer Journal*, vol. 20, no. 4, pp. 367–370, 1977.
- [100] J. J. Moré, “The Levenberg-Marquardt algorithm: implementation and theory,” in *Numerical analysis*, pp. 105–116, Springer, 1978.
- [101] L. B. Cook, A. G. Rowan, A. Melamed, G. P. Taylor, and C. R. Bangham, “HTLV-1-infected T cells contain a single integrated provirus in natural infection,” *Blood*, vol. 120, no. 17, pp. 3488–3490, 2012.
- [102] J. Richardson, A. Edwards, J. Cruickshank, P. Rudge, and A. Dalglish, “In vivo cellular tropism of human T-cell leukemia virus type 1,” *Journal of Virology*, vol. 64, no. 11, pp. 5682–5687, 1990.
- [103] T.-J. Shen, A. Chao, and C.-F. Lin, “Predicting the number of new species in further taxonomic sampling,” *Ecology*, vol. 84, no. 3, pp. 798–804, 2003.

- [104] W. W. Esty, “The efficiency of Good’s nonparametric coverage estimator,” *The Annals of Statistics*, pp. 1257–1260, 1986.
- [105] J. B. Hughes, J. J. Hellmann, T. H. Ricketts, and B. J. Bohannon, “Counting the uncountable: statistical approaches to estimating microbial diversity,” *Applied and Environmental Microbiology*, vol. 67, no. 10, pp. 4399–4406, 2001.
- [106] J. Sobern M and J. Llorente B, “The use of species accumulation functions for the prediction of species richness,” *Conservation biology*, vol. 7, no. 3, pp. 480–488, 1993.
- [107] J. Bunge, “Statistical estimation of uncultivated microbial diversity,” in *Uncultivated Microorganisms*, pp. 160–178, Springer, 2009.
- [108] C. M. Hurvich and C.-L. Tsai, “Regression and time series model selection in small samples,” *Biometrika*, vol. 76, no. 2, pp. 297–307, 1989.
- [109] K. P. Burnham and D. R. Anderson, *Model selection and multimodel inference: a practical information-theoretic approach*. Springer, 2002.
- [110] C. Gini, *Sulla misura della concentrazione e della variabilita dei caratteri*. Premiate officine grafiche C. Ferrari, 1914.
- [111] J. A. Nelder and R. Mead, “A simplex method for function minimization,” *The computer journal*, vol. 7, no. 4, pp. 308–313, 1965.
- [112] A.-F. Florins, N. Gillet, B. Asquith, M. Boxus, C. Burteaux, J.-C. Twizere, P. Urbain, F. Vandermeers, C. Debaq, M. T. Sanchez-Alcaraz, *et al.*, “Cell dynamics and immune response to BLV infection: a unifying model,” *Frontiers in bioscience: a journal and virtual library*, vol. 12, 2007.
- [113] A.-S. Gabet, F. Mortreux, A. Talarmin, Y. Plumelle, I. Leclercq, A. Leroy, A. Gessain, E. Clity, M. Joubert, and E. Wattel, “High circulating proviral load with oligoclonal expansion of HTLV-1 bearing T cells in HTLV-1 carriers with strongyloidiasis,” *Oncogene*, vol. 19, no. 43, pp. 4954–4960, 2000.
- [114] M. Cavois, S. Wain-Hobson, A. Gessain, Y. Plumelle, and E. Wattel, “Adult T-cell leukemia/lymphoma on a background of clonally expanding human T-cell leukemia virus type-1-positive cells,” *Blood*, vol. 88, no. 12, pp. 4646–4650, 1996.
- [115] C. R. Bangham, L. B. Cook, and A. Melamed, “HTLV-1 clonality in adult T-cell leukaemia and non-malignant HTLV-1 infection,” *Seminars in Cancer Biology*, vol. 26, no. 0, pp. 89 – 98, 2014. Emerging topics in human tumor virology.

- [116] J. B. Johnson and K. S. Omland, “Model selection in ecology and evolution,” *Trends in ecology & evolution*, vol. 19, no. 2, pp. 101–108, 2004.
- [117] P. A. Stephens, S. W. Buskirk, G. D. Hayward, and C. Martinez Del Rio, “Information theory and hypothesis testing: a call for pluralism,” *Journal of Applied Ecology*, vol. 42, no. 1, pp. 4–12, 2005.
- [118] K. P. Burnham and D. R. Anderson, “Kullback-Leibler information as a basis for strong inference in ecological studies,” *Wildlife research*, vol. 28, no. 2, pp. 111–119, 2001.
- [119] J. Hortal, P. A. Borges, and C. Gaspar, “Evaluating the performance of species richness estimators: sensitivity to sample grain size,” *Journal of Animal Ecology*, vol. 75, no. 1, pp. 274–287, 2006.
- [120] A. Chao, “Estimating the population size for capture-recapture data with unequal catchability,” *Biometrics*, pp. 783–791, 1987.
- [121] K. P. Burnham and W. S. Overton, “Robust estimation of population size when capture probabilities vary among animals,” *Ecology*, vol. 60, no. 5, pp. 927–936, 1979.
- [122] A. Chao, “Estimating population size for sparse data in capture-recapture experiments,” *Biometrics*, pp. 427–438, 1989.
- [123] K. Naylor, G. Li, A. N. Vallejo, W.-W. Lee, K. Koetz, E. Bryl, J. Witkowski, J. Fulbright, C. M. Weyand, and J. J. Goronzy, “The influence of age on T cell generation and TCR diversity,” *The Journal of Immunology*, vol. 174, no. 11, pp. 7446–7452, 2005.
- [124] A. Melamed, A. D. Witkover, D. J. Laydon, R. Brown, K. Ladell, K. Miners, A. G. Rowan, N. Gormley, D. A. Price, G. P. Taylor, E. L. Murphy, and C. R. M. Bangham, “Clonality of HTLV-2 in natural infection,” *PLoS Pathog*, vol. 10, p. e1004006, 03 2014.
- [125] D. J. Laydon, A. Melamed, A. Sim, N. A. Gillet, K. Sim, S. Darko, J. S. Kroll, D. C. Douek, D. A. Price, C. R. M. Bangham, and B. Asquith, “Quantification of HTLV-1 clonality and TCR diversity,” *PLoS Comput Biol*, vol. 10, p. e1003646, 06 2014.
- [126] A. Hodson, S. Crichton, S. Montoto, N. Mir, E. Matutes, K. Cwynarski, T. Kumaran, K. M. Ardeshtna, A. Pagliuca, G. P. Taylor, *et al.*, “Use of zidovudine and interferon

- alfa with chemotherapy improves survival in both acute and lymphoma subtypes of adult T-cell leukemia/lymphoma,” *Journal of Clinical Oncology*, vol. 29, no. 35, pp. 4696–4701, 2011.
- [127] B. Hanchard, “Adult T-cell leukemia/lymphoma in Jamaica: 1986-1995,” *JAIDS Journal of Acquired Immune Deficiency Syndromes*, vol. 13, pp. S20–S25, 1996.
 - [128] A. L. Bittencourt, “Vertical transmission of HTLV-I/II: a review,” *Revista do Instituto de Medicina Tropical de São Paulo*, vol. 40, no. 4, pp. 245–251, 1998.
 - [129] D. F. Roucoux and E. L. Murphy, “The epidemiology and disease outcomes of human T-lymphotropic virus type II,” *AIDS Rev*, vol. 6, no. 3, pp. 144–154, 2004.
 - [130] G. Feuer and P. L. Green, “Comparative biology of human T-cell lymphotropic virus type 1 (HTLV-1) and HTLV-2,” *Oncogene*, vol. 24, no. 39, pp. 5996–6004, 2005.
 - [131] M. d. F. S. P. de Oliveira, P. L. Fatal, J. R. L. Primo, J. L. S. da Silva, E. d. S. Batista, L. Farr, and A. L. Bittencourt, “Infective dermatitis associated with human T-cell lymphotropic virus type 1: Evaluation of 42 cases observed in Bahia, Brazil,” *Clinical Infectious Diseases*, vol. 54, no. 12, pp. 1714–1719, 2012.
 - [132] N.-K. McGill, J. Vyas, T. Shimauchi, Y. Tokura, and V. Piguet, “HTLV-1-associated infective dermatitis: updates on the pathogenesis,” *Experimental Dermatology*, vol. 21, no. 11, pp. 815–821, 2012.
 - [133] A. L. Bittencourt and M. d. F. P. d. Oliveira, “Cutaneous manifestations associated with HTLV-1 infection,” *International Journal of Dermatology*, vol. 49, no. 10, pp. 1099–1110, 2010.
 - [134] R. Okajima, A. C. P. Oliveira, J. Smid, J. Casseb, and J. A. Sanches, Jr., “High prevalence of skin disorders among HTLV-1 infected individuals independent of clinical status,” *PLoS Negl Trop Dis*, vol. 7, p. e2546, 11 2013.
 - [135] A. Perelson and P. Nelson, “Mathematical analysis of HIV-1 dynamics in vivo,” *SIAM Review*, vol. 41, no. 1, pp. 3–44, 1999.
 - [136] B. Asquith, C. T. T. Edwards, M. Lipsitch, and A. R. McLean, “Inefficient cytotoxic T lymphocytemediated killing of HIV-1infected cells in vivo,” *PLoS Biol*, vol. 4, p. e90, 03 2006.
 - [137] E. L. Read, A. A. Tovo-Dwyer, and A. K. Chakraborty, “Stochastic effects are important in intrahost HIV evolution even when viral loads are high,” *Proceedings of the National Academy of Sciences*, vol. 109, no. 48, pp. 19727–19732, 2012.

- [138] G. P. Garnett, “An introduction to mathematical models in sexually transmitted disease epidemiology,” *Sexually Transmitted Infections*, vol. 78, no. 1, pp. 7–12, 2002.
- [139] A. M. Foss, P. T. Vickerman, Z. Chalabi, P. Mayaud, M. Alary, and C. H. Watts, “Dynamic modeling of herpes simplex virus type-2 (HSV-2) transmission: issues in structural uncertainty,” *Bulletin of mathematical biology*, vol. 71, no. 3, pp. 720–749, 2009.
- [140] S. M. Goodreau, L. P. Goicochea, and J. Sanchez, “Sexual role and transmission of HIV type 1 among men who have sex with men, in peru,” *Journal of Infectious Diseases*, vol. 191, no. Supplement 1, pp. S147–S158, 2005.
- [141] J. Schluter and K. R. Foster, “The evolution of mutualism in gut microbiota via host epithelial selection,” *PLoS Biol*, vol. 10, p. e1001424, 11 2012.
- [142] S. Marino, N. T. Baxter, G. B. Huffnagle, J. F. Petrosino, and P. D. Schloss, “Mathematical modeling of primary succession of murine intestinal microbiota,” *Proceedings of the National Academy of Sciences*, vol. 111, no. 1, pp. 439–444, 2014.
- [143] S. Bonhoeffer, R. M. May, G. M. Shaw, and M. A. Nowak, “Virus dynamics and drug therapy,” *Proceedings of the National Academy of Sciences*, vol. 94, no. 13, pp. 6971–6976, 1997.
- [144] A. S. Perelson, D. E. Kirschner, and R. D. Boer, “Dynamics of HIV infection of CD4+ T cells,” *Mathematical Biosciences*, vol. 114, no. 1, pp. 81 – 125, 1993.
- [145] B. Asquith, Y. Zhang, A. J. Mosley, C. M. de Lara, D. L. Wallace, A. Worth, L. Kaftantzi, K. Meekings, G. E. Griffin, Y. Tanaka, D. F. Tough, P. C. Beverley, G. P. Taylor, D. C. Macallan, and C. R. M. Bangham, “In vivo T lymphocyte dynamics in humans and the impact of human T-lymphotropic virus 1 infection,” *Proceedings of the National Academy of Sciences*, vol. 104, no. 19, pp. 8035–8040, 2007.
- [146] D. Macallan, B. Asquith, A. Irvine, D. Wallace, A. Worth, H. Ghattas, Y. Zhang, G. Griffin, D. Tough, and P. Beverley, “Measurement and modeling of human T cell kinetics,” *European Journal of Immunology*, vol. 33, no. 8, pp. 2316–2326, 2003.
- [147] A. J. Stewart, R. M. Seymour, A. Pomiankowski, and M. Reuter, “Under-dominance constrains the evolution of negative autoregulation in diploids,” *PLoS Comput Biol*, vol. 9, p. e1002992, 03 2013.

- [148] M. S. Samoilov and A. P. Arkin, “Deviant effects in molecular reaction pathways,” *Nature biotechnology*, vol. 24, no. 10, pp. 1235–1240, 2006.
- [149] A. Arkin, J. Ross, and H. H. McAdams, “Stochastic kinetic analysis of developmental pathway bifurcation in phage λ -infected *Escherichia coli* cells,” *Genetics*, vol. 149, no. 4, pp. 1633–1648, 1998.
- [150] T. Jahnke and W. Huisinga, “A dynamical low-rank approach to the Chemical Master Equation,” *Bulletin of Mathematical Biology*, vol. 70, no. 8, pp. 2283–2302, 2008.
- [151] D. T. Gillespie, “Exact stochastic simulation of coupled chemical reactions,” *The Journal of Physical Chemistry*, vol. 81, no. 25, pp. 2340–2361, 1977.
- [152] D. J. Wilkinson, “Stochastic modelling for quantitative description of heterogeneous biological systems,” *Nature Reviews Genetics*, vol. 10, no. 2, pp. 122–133, 2009.
- [153] D. T. Gillespie, “A general method for numerically simulating the stochastic time evolution of coupled chemical reactions,” *Journal of Computational Physics*, vol. 22, no. 4, pp. 403 – 434, 1976.
- [154] L. Salwinski and D. Eisenberg, “In silico simulation of biological network dynamics,” *Nature biotechnology*, vol. 22, no. 8, pp. 1017–1019, 2004.
- [155] D. Bratsun, D. Volfson, L. S. Tsimring, and J. Hasty, “Delay-induced stochastic oscillations in gene regulation,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 41, pp. 14593–14598, 2005.
- [156] R. Srivastava, L. You, J. Summers, and J. Yin, “Stochastic vs. deterministic modeling of intracellular viral kinetics,” *Journal of Theoretical Biology*, vol. 218, no. 3, pp. 309 – 321, 2002.
- [157] R. Steuer, “Effects of stochasticity in models of the cell cycle: from quantized cycle times to noise-induced oscillations,” *Journal of Theoretical Biology*, vol. 228, no. 3, pp. 293 – 301, 2004.
- [158] M. Peifer and J. Timmer, “Parameter estimation in ordinary differential equations for biochemical processes using the method of multiple shooting,” *Systems Biology, IET*, vol. 1, no. 2, pp. 78–88, 2007.
- [159] J. Vanlier, C. Tiemann, P. Hilbers, and N. van Riel, “Parameter uncertainty in biochemical models described by ordinary differential equations,” *Mathematical Biosciences*, vol. 246, no. 2, pp. 305 – 314, 2013.

- [160] D. Wick and S. G. Self, “Early HIV infection in vivo: branching-process model for studying timing of immune responses and drug therapy,” *Mathematical Biosciences*, vol. 165, no. 2, pp. 115 – 134, 2000.
- [161] J. M. Conway and D. Coombs, “A stochastic model of latently infected cell reactivation and viral blip generation in treated HIV patients,” *PLoS Comput Biol*, vol. 7, p. e1002033, 04 2011.
- [162] H. Knolle, “A discrete branching process model for the spread of HIV via steady sexual partnerships,” *Journal of Mathematical Biology*, vol. 48, no. 4, pp. 423–443, 2004.
- [163] P. Haccou, P. Jagers, and V. A. Vatutin, *Branching processes: variation, growth, and extinction of populations*. No. 5, Cambridge University Press, 2005.
- [164] B. Sleeman and I. Wallis, “Tumour induced angiogenesis as a reinforced random walk: Modelling capillary network formation without endothelial cell proliferation,” *Mathematical and Computer Modelling*, vol. 36, no. 3, pp. 339 – 358, 2002.
- [165] H.-i. Wu, B.-L. Li, T. A. Springer, and W. H. Neill, “Modelling animal movement as a persistent random walk in two dimensions: expected magnitude of net displacement,” *Ecological Modelling*, vol. 132, no. 1, pp. 115–124, 2000.
- [166] E. A. Codling, M. J. Plank, and S. Benhamou, “Random walk models in biology,” *Journal of the Royal Society Interface*, vol. 5, no. 25, pp. 813–834, 2008.
- [167] T. Tian, K. Burrage, P. M. Burrage, and M. Carletti, “Stochastic delay differential equations for genetic regulatory networks,” *Journal of Computational and Applied Mathematics*, vol. 205, no. 2, pp. 696 – 707, 2007. Special issue on evolutionary problems.
- [168] C. V. Rao, D. M. Wolf, and A. P. Arkin, “Control, exploitation and tolerance of intracellular noise,” *Nature*, vol. 420, no. 6912, pp. 231–237, 2002.
- [169] D. T. Gillespie, “Stochastic simulation of chemical kinetics,” *Annu. Rev. Phys. Chem.*, vol. 58, pp. 35–55, 2007.
- [170] H. McAdams and A. Arkin, “Stochastic mechanisms in gene expression,” *Proceedings of the National Academy of Sciences*, vol. 94, no. 3, pp. 814–819, 1997.
- [171] P. Guptasarma, “Does replication-induced transcription regulate synthesis of the myriad low copy number proteins of Escherichia coli?,” *Bioessays*, vol. 17, no. 11, pp. 987–997, 1995.

- [172] W. Alt and J. J. Tyson, “A stochastic model of cell division (with application to fission yeast),” *Mathematical Biosciences*, vol. 84, no. 2, pp. 159 – 187, 1987.
- [173] T. Jahnke and M. Kreim, “Error bound for piecewise deterministic processes modeling stochastic reaction systems,” *Multiscale Modeling & Simulation*, vol. 10, no. 4, pp. 1119–1147, 2012.
- [174] V. Kazeev, M. Khammash, M. Nip, and C. Schwab, “Direct solution of the Chemical Master Equation using quantized tensor trains,” *PLoS Comput Biol*, vol. 10, p. e1003359, 03 2014.
- [175] T. G. Kurtz, “The relationship between stochastic and deterministic models for chemical reactions,” *The Journal of Chemical Physics*, vol. 57, no. 7, pp. 2976–2978, 1972.
- [176] C. Gardiner, *Handbook of Stochastic Methods for Physics, Chemistry, and the Natural Sciences*. Springer complexity, Springer, 2004.
- [177] N. G. Van Kampen, *Stochastic processes in physics and chemistry*, vol. 1. Elsevier, 1992.
- [178] D. T. Gillespie, “A rigorous derivation of the chemical master equation,” *Physica A: Statistical Mechanics and its Applications*, vol. 188, no. 13, pp. 404 – 425, 1992.
- [179] T. Jahnke, “On reduced models for the chemical master equation,” *Multiscale Modeling & Simulation*, vol. 9, no. 4, pp. 1646–1676, 2011.
- [180] M. Hegland, C. Burden, L. Santoso, S. MacNamara, and H. Booth, “A solver for the stochastic master equation applied to gene regulatory networks,” *Journal of Computational and Applied Mathematics*, vol. 205, no. 2, pp. 708 – 724, 2007. Special issue on evolutionary problems.
- [181] C. Moler and C. Van Loan, “Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later,” *SIAM Review*, vol. 45, no. 1, pp. 3–49, 2003.
- [182] W. J. Stewart, *Introduction to the numerical solution of Markov chains*, vol. 41. Princeton University Press Princeton, 1994.
- [183] S. Peleš, B. Munsky, and M. Khammash, “Reduction and solution of the chemical master equation using time scale separation and finite state projection,” *The Journal of chemical physics*, vol. 125, no. 20, p. 204104, 2006.

- [184] C. Moler and C. Van Loan, “Nineteen dubious ways to compute the exponential of a matrix,” *SIAM Review*, vol. 20, no. 4, pp. 801–836, 1978.
- [185] N. Higham, “The scaling and squaring method for the matrix exponential revisited,” *SIAM Review*, vol. 51, no. 4, pp. 747–764, 2009.
- [186] L. Lopez and V. Simoncini, “Analysis of projection methods for rational function approximation to the matrix exponential,” *SIAM Journal on Numerical Analysis*, vol. 44, no. 2, pp. 613–635, 2006.
- [187] M. Hegland, A. Hellander, and P. Lötstedt, “Sparse grids and hybrid methods for the chemical master equation,” *BIT Numerical Mathematics*, vol. 48, no. 2, pp. 265–283, 2008.
- [188] T. Jahnke and W. Huisinga, “Solving the chemical master equation for monomolecular reaction systems analytically,” *Journal of mathematical biology*, vol. 54, no. 1, pp. 1–26, 2007.
- [189] D. Bates and M. Maechler, *Matrix: Sparse and Dense Matrix Classes and Methods*, 2014. R package version 1.1-2-2.
- [190] E. A. Vandoorn and P. Schrijner, “Ratio limits and limiting conditional distributions for discrete-time birth-death processes,” *Journal of mathematical analysis and applications*, vol. 190, no. 1, pp. 263–284, 1995.
- [191] B. Munsky and M. Khammash, “The finite state projection approach for the analysis of stochastic noise in gene networks,” *Automatic Control, IEEE Transactions on*, vol. 53, no. Special Issue, pp. 201–214, 2008.
- [192] V. Sunkara and M. Hegland, “An optimal finite state projection method,” *Procedia Computer Science*, vol. 1, no. 1, pp. 1579–1586, 2010.
- [193] A. Machuca, B. Rods, and V. Soriano, “The effect of antiretroviral therapy on HTLV-1 infection,” *Virus Research*, vol. 78, no. 12, pp. 93 – 100, 2001.
- [194] F. Brun-Vezinet, C. Boucher, C. Loveday, D. Descamps, V. Fauveau, J. Izopet, D. Jeffries, S. Kaye, C. Krzyanowski, A. Nunn, *et al.*, “HIV-1 viral load, phenotype, and resistance in a subset of drug-naïve participants from the Delta trial,” *The Lancet*, vol. 350, no. 9083, pp. 983–990, 1997.
- [195] S. Broder, “The development of antiretroviral therapy and its impact on the HIV-1/AIDS pandemic,” *Antiviral Research*, vol. 85, no. 1, pp. 1 – 18, 2010. Twenty-five Years of Antiretroviral Drug Development: Progress and Prospects.

- [196] H. J. Kaminski, *Myasthenia gravis and related disorders*. Springer, 2009.
- [197] G.-i. Tanaka, A. Okayama, T. Watanabe, S. Aizawa, S. Stuver, N. Mueller, C.-C. Hsieh, and H. Tsubouchi, “The clonal expansion of human T lymphotropic virus type 1 infected T cells: A comparison between seroconverters and long-term carriers,” *Journal of Infectious Diseases*, vol. 191, no. 7, pp. 1140–1147, 2005.
- [198] N. Gillet, A. Florins, M. Boxus, C. Burteau, A. Nigro, F. Vandermeers, H. Balon, A.-B. Bouzar, J. Defoiche, A. Burny, *et al.*, “Mechanisms of leukemogenesis induced by bovine leukemia virus: prospects for novel anti-retroviral therapies in human,” *Retrovirology*, vol. 4, no. 1, p. 18, 2007.
- [199] E. Paine, J. Garcia, T. C. Philpott, G. Shaw, and L. Ratner, “Limited sequence variation in human T-lymphotropic virus type 1 isolates from north american and african patients,” *Virology*, vol. 182, no. 1, pp. 111 – 123, 1991.
- [200] J. Overbaugh and C. R. Bangham, “Selection forces and constraints on retroviral sequence variation,” *Science*, vol. 292, no. 5519, pp. 1106–1109, 2001.
- [201] Y. Ina and T. Gojobori, “Molecular evolution of human T-cell leukemia virus,” *Journal of molecular evolution*, vol. 31, no. 6, pp. 493–499, 1990.
- [202] M. Martell, J. I. Esteban, J. Quer, J. Genesc, A. Weiner, R. Esteban, J. Guardia, and J. Gmez, “Hepatitis C virus (HCV) circulates as a population of different but closely related genomes: quasispecies nature of HCV genome distribution.,” *Journal of Virology*, vol. 66, no. 5, pp. 3225–3229, 1992.
- [203] M. Lauck, M. V. Alvarado-Mora, E. A. Becker, D. Bhattacharya, R. Striker, A. L. Hughes, F. J. Carrilho, D. H. O’Connor, and J. R. R. Pinho, “Analysis of hepatitis C virus intrahost diversity across the coding region by ultradeep pyrosequencing,” *Journal of Virology*, vol. 86, no. 7, pp. 3952–3960, 2012.
- [204] N. A. Gillet, L. Cook, D. J. Laydon, C. Hlela, K. Verdonck, C. Alvarez, E. Gotuzzo, D. Clark, L. Farr, A. Bittencourt, B. Asquith, G. P. Taylor, and C. R. M. Bangham, “Strongyloidiasis and infective dermatitis alter human T lymphotropic virus-1 clonality in vivo,” *PLoS Pathog*, vol. 9, p. e1003263, 04 2013.
- [205] S. Engblom, “Spectral approximation of solutions to the chemical master equation,” *Journal of Computational and Applied Mathematics*, vol. 229, no. 1, pp. 208 – 221, 2009.

- [206] D. Sasaki, Y. Doi, H. Hasegawa, K. Yanagihara, K. Tsukasaki, M. Iwanaga, Y. Yamada, T. Watanabe, and S. Kamihira, “High human T cell leukemia virus type-1(htlv-1) provirus load in patients with HTLV-1 carriers complicated with HTLV-1-unrelated disorders,” *Virology Journal*, vol. 7, no. 1, p. 81, 2010.
- [207] S. Engblom, “Computing the moments of high dimensional solutions of the master equation,” *Applied Mathematics and Computation*, vol. 180, no. 2, pp. 498–515, 2006.
- [208] G. Strang, “On the construction and comparison of difference schemes,” *SIAM Journal on Numerical Analysis*, vol. 5, no. 3, pp. 506–517, 1968.
- [209] M. Elemans, A. Florins, L. Willems, and B. Asquith, “Rates of CTL killing in persistent viral infection in vivo,” *PLoS Comput Biol*, vol. 10, p. e1003534, 04 2014.
- [210] K. Ladell, M. K. Hellerstein, D. Cesar, R. Busch, D. Boban, and J. M. McCune, “Central memory CD8+ T cells appear to have a shorter lifespan and reduced abundance as a function of HIV disease progression,” *The Journal of Immunology*, vol. 180, no. 12, pp. 7907–7918, 2008.
- [211] T. House, J. V. Ross, and D. Sirl, “How big is an outbreak likely to be? Methods for epidemic final-size calculation,” *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, vol. 469, no. 2150, 2012.
- [212] K. Best, T. Oakes, J. M. Heather, J. Shawe-Taylor, and B. Chain, “Sequence and primer independent stochastic heterogeneity in PCR amplification efficiency revealed by single molecule barcoding,” *bioRxiv*, p. 011411, 2014.
- [213] B. Pascal and E. Havet, *Pensées*. Dezobry et E. Magdeleine, 1852.

A. Appendix A: DivE supporting information

Table A.1.: Subsamples used in analysis of relationship between sample size and estimated diversity, and in comparison of DivE with AIC_c

Data Source	Patient	Population	Subsample Sizes	Subsample Percentages
HTLV-1	H1	NA	500, 1000, 2000, 3000, 4596	10.9, 21.8, 43.5, 65.3, 100
HTLV-1	H2	NA	500, 1000, 2000, 3000, 4210	11.9, 23.8, 47.5, 71.3, 100
HTLV-1	H3	NA	500, 1000, 2000, 4000, 4967	10.1, 20.1, 40.3, 80.5, 100
HTLV-1	H4	NA	500, 1000, 2000, 3000, 3992	12.5, 25.1, 50.1, 75.2, 100
HTLV-1	H5	NA	500, 1000, 1623	30.8, 61.6, 100
HTLV-1	H6	NA	500, 1000, 2000, 2002	25, 50, 99.9, 100
HTLV-1	H7	NA	500, 1000, 2000, 3000, 3406	14.7, 29.4, 58.7, 88.1, 100
HTLV-1	H8	NA	500, 1000, 1693	29.5, 59.1, 100
HTLV-1	H9	NA	500, 1000, 1127	44.4, 88.7, 100
HTLV-1	H10	NA	500, 1000, 5000, 9000, 12962	3.86, 7.71, 38.6, 69.4, 100
HTLV-1	H11	NA	500, 1000, 2000, 4000, 5388	9.28, 18.6, 37.1, 74.2, 100
HTLV-1	H12	NA	500, 1000, 2000, 4000, 5007	9.99, 20, 39.9, 79.9, 100
HTLV-1	H13	NA	500, 1000, 2000, 3000, 3716	13.5, 26.9, 53.8, 80.7, 100
HTLV-1	H14	NA	500, 1000, 2000, 3000, 3907	12.8, 25.6, 51.2, 76.8, 100
TCRs	T1	CD4 Total	166667, 333333, 5e+05, 666667, 833333, 1e+06	16.7, 33.3, 50, 66.7, 83.3, 100
TCRs	T2	CD4 Total	333333, 666667, 1e+06, 1333333, 1666667, 2e+06	16.7, 33.3, 50, 66.7, 83.3, 100
TCRs	T1	CD8 Total	116667, 233333, 350000, 466667, 583333, 7e+05	16.7, 33.3, 50, 66.7, 83.3, 100
TCRs	T2	CD8 Total	166667, 333333, 5e+05, 666667, 833333, 1e+06	16.7, 33.3, 50, 66.7, 83.3, 100
TCRs	T1	CD4 Naïve	83333, 166667, 250000, 333333, 416667, 5e+05	16.7, 33.3, 50, 66.7, 83.3, 100
TCRs	T1	CD4 CM	83333, 166667, 250000, 333333, 416667, 5e+05	16.7, 33.3, 50, 66.7, 83.3, 100
TCRs	T1	CD4 EM	21667, 43333, 65000, 86667, 108333, 130000	16.7, 33.3, 50, 66.7, 83.3, 100
TCRs	T1	CD8 Naïve	8167, 16333, 24500, 32667, 40833, 49000	16.7, 33.3, 50, 66.7, 83.3, 100
TCRs	T1	CD8 CM	12667, 25333, 38000, 50667, 63333, 76000	16.7, 33.3, 50, 66.7, 83.3, 100
TCRs	T1	CD8 EM	2833, 5667, 8500, 11333, 14167, 17000	16.7, 33.3, 50, 66.7, 83.3, 100
TCRs	T2	CD4 Naïve	107833, 215667, 323500, 431333, 539167, 647000	16.7, 33.3, 50, 66.7, 83.3, 100
TCRs	T2	CD4 EM	35000, 70000, 105000, 140000, 175000, 210000	16.7, 33.3, 50, 66.7, 83.3, 100
TCRs	T2	CD4 CM	85500, 171000, 256500, 342000, 427500, 513000	16.7, 33.3, 50, 66.7, 83.3, 100
TCRs	T2	CD8 Naïve	42667, 85333, 128000, 170667, 213333, 256000	16.7, 33.3, 50, 66.7, 83.3, 100
TCRs	T2	CD8 EM	6783, 13567, 20350, 27133, 33917, 40700	16.7, 33.3, 50, 66.7, 83.3, 100
TCRs	T2	CD8 CM	5667, 11333, 17000, 22667, 28333, 34000	16.7, 33.3, 50, 66.7, 83.3, 100
Microbes	M1	bP1	427, 640, 853, 1067, 1280	33.4, 50, 66.6, 83.4, 100
Microbes	M2	bP10	689, 1034, 1378, 1722, 2067	33.3, 50, 66.7, 83.3, 100
Microbes	M3	bP2	685, 1028, 1371, 1713, 2056	33.3, 50, 66.7, 83.3, 100
Microbes	M4	bP3	720, 1080, 1439, 1799, 2159	33.3, 50, 66.7, 83.3, 100
Microbes	M5	bP4	679, 1018, 1358, 1698, 2037	33.3, 50, 66.7, 83.4, 100
Microbes	M6	bP5	687, 1031, 1375, 1718, 2062	33.3, 50, 66.7, 83.3, 100
Microbes	M7	bP6	820, 1230, 1639, 2049, 2459	33.3, 50, 66.7, 83.3, 100
Microbes	M8	bP7	774, 1161, 1548, 1935, 2322	33.3, 50, 66.7, 83.3, 100
Microbes	M9	bP8	786, 1178, 1571, 1964, 2357	33.3, 50, 66.7, 83.3, 100
Microbes	M10	bP9	789, 1184, 1579, 1973, 2368	33.3, 50, 66.7, 83.3, 100
Microbes	M1	uP1	381, 571, 761, 952, 1142	33.4, 50, 66.6, 83.4, 100
Microbes	M2	uP10	522, 782, 1043, 1304, 1565	33.4, 50, 66.6, 83.3, 100
Microbes	M3	uP2	567, 850, 1133, 1417, 1700	33.4, 50, 66.6, 83.4, 100
Microbes	M4	uP3	503, 755, 1007, 1258, 1510	33.3, 50, 66.7, 83.3, 100
Microbes	M5	uP4	568, 852, 1135, 1419, 1703	33.4, 50, 66.6, 83.3, 100
Microbes	M6	uP5	640, 960, 1281, 1601, 1921	33.3, 50, 66.7, 83.3, 100
Microbes	M7	uP6	698, 1046, 1395, 1744, 2093	33.3, 50, 66.7, 83.3, 100
Microbes	M8	uP7	589, 884, 1179, 1473, 1768	33.3, 50, 66.7, 83.3, 100
Microbes	M9	uP8	531, 797, 1063, 1328, 1594	33.3, 50, 66.7, 83.3, 100
Microbes	M10	uP9	605, 907, 1209, 1512, 1814	33.4, 50, 66.6, 83.4, 100

Table A.2.: DivE species richness estimates for HTLV-1 data

Patient	Timepoint	DivE	DivE lower bound	DivE upper bound
1	1	2.50E+04	3.20E+03	1.70E+06
1	1	2.70E+04	3.50E+03	2.00E+06

1	1	2.40E+04	3.10E+03	1.60E+06
1	2	2.50E+04	3.40E+03	1.30E+06
1	2	2.60E+04	3.70E+03	1.20E+06
1	2	1.80E+04	3.00E+03	6.30E+05
1	3	2.40E+04	3.50E+03	1.10E+06
1	4	1.80E+04	3.20E+03	5.00E+05
1	4	2.50E+04	3.00E+03	2.00E+06
1	4	2.00E+04	3.00E+03	7.50E+05
2	1	1.00E+04	2.00E+03	1.90E+05
2	1	8.80E+03	1.80E+03	1.50E+05
2	1	1.10E+04	1.80E+03	3.00E+05
2	2	1.10E+04	1.80E+03	3.40E+05
2	2	9.50E+03	2.00E+03	1.50E+05
2	2	1.20E+04	2.00E+03	3.30E+05
2	3	6.40E+03	1.60E+03	1.20E+05
2	3	9.00E+03	1.60E+03	2.50E+05
2	3	9.00E+03	1.60E+03	2.30E+05
3	1	2.60E+04	3.40E+03	1.30E+06
3	1	3.30E+04	3.60E+03	2.00E+06
3	1	3.20E+04	3.80E+03	1.50E+06
3	2	3.10E+04	3.90E+03	1.50E+06
3	2	2.80E+04	3.80E+03	1.10E+06
3	2	2.50E+04	3.30E+03	1.50E+06
3	3	1.90E+04	2.70E+03	9.70E+05
3	3	2.40E+04	3.00E+03	1.20E+06
3	3	2.10E+04	2.80E+03	9.80E+05
4	1	2.00E+04	3.60E+03	3.90E+05
4	1	1.90E+04	3.50E+03	3.70E+05
4	1	1.70E+04	3.50E+03	2.50E+05
4	2	2.00E+04	3.80E+03	3.70E+05
4	2	1.90E+04	4.30E+03	2.30E+05
4	2	2.40E+04	4.10E+03	5.50E+05
4	3	2.10E+04	3.90E+03	4.40E+05
4	3	1.70E+04	3.60E+03	2.20E+05
4	3	2.30E+04	4.00E+03	5.60E+05
5	1	8.90E+03	1.60E+03	2.10E+05
5	1	1.10E+04	1.90E+03	2.70E+05
5	1	9.80E+03	1.80E+03	2.20E+05
5	2	7.30E+03	1.60E+03	1.10E+05
5	2	8.20E+03	1.70E+03	1.40E+05
5	2	7.40E+03	1.60E+03	1.70E+05
5	3	8.70E+03	1.60E+03	2.60E+05
5	3	8.10E+03	1.50E+03	2.00E+05
5	3	7.20E+03	1.50E+03	1.70E+05
6	1	1.10E+04	2.10E+03	2.20E+05
6	1	8.80E+03	1.80E+03	1.40E+05
6	1	1.10E+04	2.10E+03	1.90E+05
6	2	7.40E+03	1.80E+03	6.70E+04
6	2	1.10E+04	2.00E+03	1.80E+05

6	2	8.60E+03	1.70E+03	1.10E+05
6	3	5.70E+03	1.50E+03	4.70E+04
6	3	9.00E+03	1.70E+03	1.90E+05
6	3	5.50E+03	1.50E+03	3.80E+04
7	1	7.70E+04	9.70E+03	4.70E+06
7	1	7.70E+04	9.20E+03	5.10E+06
7	1	8.10E+04	1.00E+04	5.10E+06
7	2	7.10E+04	9.90E+03	3.20E+06
7	2	6.70E+04	9.90E+03	1.90E+06
7	2	7.40E+04	9.00E+03	4.50E+06
7	3	6.80E+04	1.00E+04	2.30E+06
7	3	6.50E+04	1.00E+04	2.30E+06
7	3	7.80E+04	9.70E+03	4.70E+06
8	1	3.50E+04	5.00E+03	1.40E+06
8	1	3.00E+04	5.20E+03	7.40E+05
8	1	3.00E+04	5.80E+03	5.20E+05
8	2	2.70E+04	5.20E+03	5.30E+05
8	2	2.30E+04	5.50E+03	2.40E+05
8	2	2.30E+04	5.80E+03	2.10E+05
8	3	2.80E+04	4.70E+03	7.20E+05
8	3	3.10E+04	4.50E+03	1.30E+06
8	3	2.60E+04	4.60E+03	6.80E+05
9	1	5.70E+04	4.80E+03	1.30E+07
9	1	5.90E+04	6.30E+03	5.30E+06
9	1	5.50E+04	6.20E+03	4.70E+06
9	2	6.80E+04	6.10E+03	9.90E+06
9	2	4.90E+04	7.30E+03	1.90E+06
9	2	6.70E+04	6.40E+03	7.70E+06
9	3	6.60E+04	4.50E+03	2.70E+07
9	3	5.30E+04	4.30E+03	1.00E+07
9	3	5.50E+04	4.60E+03	1.00E+07
10	1	3.30E+04	4.50E+03	1.80E+06
10	1	3.90E+04	4.80E+03	2.50E+06
10	1	3.80E+04	4.70E+03	2.50E+06
10	2	3.00E+04	5.00E+03	8.80E+05
10	2	3.20E+04	5.30E+03	8.20E+05
10	2	3.30E+04	4.90E+03	1.20E+06
10	3	3.00E+04	4.00E+03	1.60E+06
10	3	3.20E+04	4.80E+03	1.20E+06
10	3	2.80E+04	4.20E+03	9.90E+05
11	1	2.20E+04	3.70E+03	6.90E+05
11	1	3.10E+04	4.30E+03	1.40E+06
11	1	3.60E+04	5.10E+03	1.60E+06
11	2	2.70E+04	4.70E+03	7.20E+05
11	2	2.60E+04	5.30E+03	4.00E+05
11	2	3.40E+04	4.80E+03	1.50E+06
11	3	2.80E+04	4.90E+03	6.00E+05
11	3	3.20E+04	5.10E+03	9.80E+05
11	3	2.80E+04	5.10E+03	5.60E+05

12	1	1.00E+05	9.80E+03	8.30E+06
13	1	2.00E+04	2.90E+03	6.80E+05
13	2	1.60E+04	3.50E+03	1.90E+05
14	1	3.70E+04	3.70E+03	4.60E+06
14	2	2.60E+04	4.30E+03	8.10E+05

A.1. List of DivE Candidate Models

1. Logistic

$$y = \frac{a_1}{a_2 + x^{-a_3}}$$

2. Negative exponential

$$y = \frac{a_1}{a_2} (1 - e^{-a_2 x})$$

3. Logarithmic

$$y = a_1 \log(a_2 x + 1)$$

4. Quadratic Logarithmic

$$y = a_1 \log(a_3 x + 1) + a_2 \log(a_3 x + 1)^2$$

5. Quartic Logarithmic

$$y = a_1 \log(a_5 x + 1) + a_2 \log(a_5 x + 1)^2 + a_3 \log(a_5 x + 1)^3 + a_4 \log(a_5 x + 1)^4$$

6. Quintic Logarithmic

$$y = a_1 \log(a_6 x + 1) + a_2 \log(a_6 x + 1)^2 + a_3 \log(a_6 x + 1)^3 + a_4 \log(a_6 x + 1)^4 + a_5 \log(a_6 x + 1)^5$$

7. NIST Kirby

$$y = \frac{a_1 + a_2 x + a_3 x^2}{x^2 + a_2 x}$$

8. NIST MGH09

$$y = \frac{a_1(x^2 + a_2 x)}{x^2 + a_3 x + a_4}$$

9. Hyperbolic G

$$y = \frac{a_1 x}{a_2 + x} + \frac{a_3 x}{a_4 + x}$$

10. Cubic Logarithmic

$$y = a_1 \log(a_4 x + 1) + a_2 \log(a_4 x + 1)^2 + a_3 \log(a_4 x + 1)^3$$

11. Gunary

$$y = \frac{x}{a_1 + a_2 + a_3 x^{0.5}}$$

12. Cellular Conductance

$$y = \frac{a_1}{1 + e^{\frac{x-a_2}{a_3}}} + \frac{a_4}{1 + e^{\frac{x-a_5}{a_5}}}$$

13. NIST Thurber

$$y = \frac{a_1 + a_2 x + a_3 x^2 + a_4 x^3}{1 + a_5 x + a_6 x^2 + a_7 x^3}$$

14. Lorentzian Modified Peak F

$$y = \frac{a_1}{a_2 + \left(\frac{x-a_3}{a_4}\right)^{a_5}}$$

15. Pseudo Voight peak modified

$$y = a_1 \left[a_4 \left(\frac{1}{1 + \frac{x-a_2}{a_3}} \right)^{a_5} + (1 - a_4) e^{-0.5 \left(\frac{x-a_2}{a_3} \right)^{a_6}} \right]$$

16. Five-parameter logistic with offset

$$y = a_1 + \frac{a_2 - a_1}{\left(1 + \left(\frac{x}{a_3} \right)^{a_4} \right)^{a_5}}$$

17. Triple exponential

$$y = a_1 e^{a_2 x} + a_3 e^{a_4 x} + a_5 e^{a_6 x} + a_7$$

18. NIST Bennett 5

$$y = a_1 (a_2 + x)^{\frac{-1}{a_3}} + a_4$$

19. BioScience B

$$y = a_1 \left(1 - \left(1 + \left(\frac{x}{a_2} \right)^{a_3} \right)^{-a_4} \right) + a_5$$

20. High-low affinity double

$$y = \frac{a_1 a_2 x}{1 + a_2 x} + \frac{a_3 a_4 x}{1 + a_4 x}$$

21. Logistic B with offset

$$y = \frac{a_1}{1 + \left(\frac{x}{a_2} \right)^{a_3}} + a_4$$

22. Hyperbolic logistic

$$y = \frac{a_1 x^{a_2}}{a_3 + x^{a_2}}$$

23. Hill

$$y = \frac{a_1 x^{a_2}}{a_3^{a_2} + x^{a_2}}$$

24. Log-normal peak with offset

$$y = a_1 \exp \left(-0.5 \left(\frac{\log(x+1) - a_2}{a_3} \right)^2 \right) + a_4$$

25. Inverse Bleasdale

$$y = \frac{x}{\left((a_1 + a_2 x)^{\frac{-1}{a_3}} \right)}$$

26. Double exponential

$$y = a_1 e^{a_2 x} + a_3 e^{a_4 x} + a_5$$

27. Polytrope transform with offset

$$y = \frac{a_1}{(a_3 x + a_4)^{a_2}} + a_5$$

28. Generalized product accumulation

$$y = \frac{a_1(a_2 - x)}{(a_3 + (a_2 - x))} + a_4(a_2 - x) + a_5$$

29. Generalized substrate depletion

$$y = \frac{a_1 x}{a_2 + x} - a_3 x - a_4$$

30. Weibull peak

$$y = a_1 \exp \left(-0.5 \left(\frac{\log \left(\frac{x}{a_2} \right)}{a_3} \right)^2 \right)$$

31. Gaussian peak modified with offset

$$y = a_1 \exp \left(-0.5 \left(\frac{x - a_2}{a_3} \right)^{a_4} \right) + a_5$$

32. Morgan-Mercer-Flodin (MMF) with offset

$$y = \frac{a_1 a_2 + a_3 x^{a_4}}{a_2 + x^{a_4}} + a_5$$

33. Log-normal peak modified with offset

$$y = a_1 \exp \left(-0.5 \left(\frac{(\log(x + 1) - a_2)}{a_3} \right)^{a_4} \right) + a_5$$

34. Weibull

$$y = a_1 - a_2 \exp(-a_3 x^{a_4})$$

35. Weibull peak modified with offset

$$y = a_1 - a_2 \exp \left(-0.5 \left(\frac{\log \left(\frac{x}{a_2} \right)}{a_3} \right)^{a_4} \right) + a_5$$

36. General polyfunctional 2 logs

$$y = a_1 x^{a_2} + a_3 x^{a_4} + a_5 \log(x + 1)^{a_6} + a_7 \log(x + 1) x^{a_8}$$

37. General polyfunctional

$$y = a_1 x^{a_2} + a_3 x^{a_4} + a_5 x^{a_6} + a_7 x^{a_8}$$

38. General polyfunctional log

$$y = a_1 x^{a_2} + a_3 x^{a_4} + a_5 x^{a_6} + a_7 \log(x+1) x^{a_8}$$

39. Bradley transform

$$y = a_1 \log(-a_2 \log(a_3 x + a_4))$$

40. Lorentzian modified peak C

$$y = \frac{a_1}{a_2 + (x - a_3)^{a_4}}$$

41. Janoschek growth modified

$$y = a_1(a_1 - a_4)(1 - \exp(-a_2 x^{a_3}))$$

42. General polyfunctional atan

$$y = a_1 x^{a_2} + a_3 x^{a_4} + a_5 x^{a_6} + a_7 \operatorname{atan}(x)^{a_8}$$

43. Simple equation 16

$$y = a_1 x^{(a_2 + a_3 \log(x))}$$

44. Simple equation 30

$$y = a_1 x^{a_2 x^{a_3}}$$

45. Simple equation 26 with offset

$$y = \frac{a_1}{(1 + a_2 x^{a_3})^2} + a_4$$

46. Simple equation 40 with offset

$$y = a_1 \log(x + a_2)^{a_3} + a_4$$

47. Harris

$$y = \frac{1}{(a_1 + a_2 x^{a_3})}$$

48. General polyModelal tanh

$$y = a_1x^{a_2} + a_3x^{a_4} + a_5x^{a_6} + a_7\tanh(x)^{a_8}$$

49. General polyfunctional negative exponential

$$y = a_1x^{a_2} + a_3x^{a_4} + a_5x^{a_6} + a_7e^{-a_8x}$$

50. General polyfunctional sin

$$y = a_1x^{a_2} + a_3x^{a_4} + a_5x^{a_6} + a_7\sin(x)^{a_8}$$

51. General polyfunctional log negative exponential

$$y = a_1x^{a_2} + a_3x^{a_4} + a_5e^{-a_6x} + a_7\log(x+1)x^{a_8}$$

52. General polyfunctional log tanh

$$y = a_1x^{a_2} + a_3x^{a_4} + a_5\tanh(x)^{a_6} + a_7\log(x+1)x^{a_8}$$

53. General polyfunctional log atan

$$y = a_1x^{a_2} + a_3x^{a_4} + a_5\operatorname{atan}(x)^{a_6} + a_7\log(x+1)x^{a_8}$$

54. Arrhenius law stretched

$$y = a_1 \exp\left(-\frac{a_2}{x}\right)^{a_3}$$

55. Quadratic logarithmic variable shape

$$y = a_1 \log(a_3x+1) + a_2 \log(a_4x+1)^2$$

56. Cubic logarithmic variable shape

$$y = a_1 \log(a_4x+1) + a_2 \log(a_5x+1)^2 + a_3 \log(a_6x+1)^3$$

57. Quartic logarithmic variable shape

$$y = a_1 \log(a_5x + 1) + a_2 \log(a_6x + 1)^2 + a_3 \log(a_7x + 1)^3 + a_4 \log(a_8x + 1)^4$$

58. Quintic logarithmic variable shape

$$y = a_1 \log(a_6x+1)+a_2 \log(a_7x+1)^2+a_3 \log(a_8x+1)^3+a_4 \log(a_9x+1)^4+a_5 \log(a_{10}x+1)^5$$

B. Appendix B: Derivation of Expectation, Variance, and Covariance Clone ODEs

Under a stochastic paradigm, the ODEs in equation (5.1) approximately correspond to the expected frequencies of the expressing and latent compartments of a given clone i . Because the propensities of our model (equation (5.5)) have degrees of at most 1, we can derive another set of ODEs to model the variance and covariance [207] of a given species i . With our notation, the moments are given by

$$\frac{dm_i}{dt} = \sum_{r=1}^R \nu_r^i \left(\alpha_r(X(t)) + \sum_{k,l} \frac{\partial^2 \alpha_r(m)}{\partial x_k \partial x_l} \frac{C_{kl}}{2!} \right) \quad (\text{B.1a})$$

$$\frac{dC_{ij}}{dt} = \sum_{r=1}^R \left(\nu_r^i \sum_k \frac{\partial \alpha_r(X(t))}{\partial x_k} \frac{C_{kj}}{1!} + \nu_r^j \sum_l \frac{\partial \alpha_r(X(t))}{\partial x_l} \frac{C_{il}}{1!} \right) + \sum_{r=1}^R \nu_r^{[i,j]} \left(\alpha_r(X(t)) + \sum_{k,l} \frac{\partial^2 \alpha_r(X(t))}{\partial x_k \partial x_l} \frac{C_{kl}}{2!} \right) \quad (\text{B.1b})$$

where m_i is the expected value of species i , C_{ij} is the covariance between species i and species j , ν_r^i is the i^{th} component of ν_r , the stoichiometric vector of reaction r , and $\nu_r^{[i,j]} = \nu_r^i \nu_r^j$. Note that the subscripts i and j in the above equations refer to the i^{th} and j^{th} entry of the state vector $X(t)$. Because we divide clones into two compartments (expressing and latent), $X(t)$ has $2S(t)$ entries. Re-indexing so that the subscript i refers to the i^{th} clone ($i \in \{1, \dots, S(t)\}$), we introduce the following shorthand. Let

$$x_i := \mathbb{E}[X_i] \quad (\text{B.2a})$$

$$y_i := \mathbb{E}[Y_i] \quad (\text{B.2b})$$

$$V_{x_i} := \text{Var}[X_i] \quad (\text{B.2c})$$

$$V_{y_i} := \text{Var}[Y_i] \quad (\text{B.2d})$$

$$C_{x_i, y_i} := \text{Cov}[X_i, Y_i] \quad (\text{B.2e})$$

To relate this to Equation (B.1), we have $V_{x_i} = V_{2i-1} = C_{(2i-1), (2i-1)} = \text{Cov}[X_i, X_i] = \text{Var}[X_i]$, $V_{y_i} = V_{2i} = C_{(2i), (2i)} = \text{Cov}[Y_i, Y_i] = \text{Var}[Y_i]$, and $C_{x_i, y_i} = C_{(2i-1), (2i)} = \text{Cov}[X_i, Y_i]$.

We approximate by assuming that the frequency of a given clone is independent of any other clone, i.e. that for $i \neq j$, $\text{Cov}[X_i, X_j] = \text{Cov}[Y_i, Y_j] = \text{Cov}[X_i, Y_j] = 0$. This assumption is valid at the level of the individual clone, in that $\text{Cov}[X_i, Y_j] \approx 0$ for $i \neq j$. However $\sum_{k=1}^{S(t)} C_{ik}$ may not be negligible. Nevertheless, to substantially reduce the dimension of the system of ODEs, we must assume that clone frequencies are approximately independent of each other. With these assumptions, it can be shown that we have the following system of ordinary differential equations to describe the expectations, variances of covariance of the expressing and latent compartments x_i and y_i of clone i . See Appendix B for further details.

We now list all partial derivatives of our system. We have the following first order partial derivatives

$$\frac{\partial \alpha_{7(i-1)+1}(m)}{\partial x_i} = \frac{\pi(K + N(t)) - \pi x_i}{(K + N(t))^2}, \quad \frac{\partial \alpha_{7(i-1)+2}(m)}{\partial x_i} = \delta, \quad \frac{\partial \alpha_{7(i-1)+3}(m)}{\partial x_i} = s_i,$$

and

$$\frac{\partial \alpha_{7(i-1)+4}(m)}{\partial y_i} = \frac{p(K + N(t)) - p y_i}{(K + N(t))^2}, \quad \frac{\partial \alpha_{7(i-1)+5}(m)}{\partial y_i} = d, \quad \frac{\partial \alpha_{7(i-1)+6}(m)}{\partial y_i} = \varepsilon_i$$

For second order partial derivatives, we have

$$\frac{\partial^2 \alpha_{7(i-1)+a}(m)}{\partial x_k \partial y_l} = 0 \quad \forall i, k, l, \text{ for } a \in \{2, 3, 5, 6\}$$

i.e. for all non-proliferation propensities. We can ignore the second order partial derivatives of the proliferation propensity functions $\alpha_{7(i-1)+1}(m)$ and $\alpha_{7(i-1)+4}(m)$, as

$$\frac{\partial^2 \alpha_{7(i-1)+b}(m)}{\partial x_k \partial y_l} = \mathcal{O}(-3) \approx 0 \quad \forall i, k, l, \text{ for } b \in \{1, 4\}$$

and

$$\frac{\partial^2 \alpha_{7(i-1)+1}(m)}{\partial x_i^2} \approx \frac{\partial^2 \alpha_{7(i-1)+4}(m)}{\partial y_i^2} = \mathcal{O}(-3) \approx 0$$

Thus we have

$$\frac{dx_i}{dt} = \frac{\pi x_i}{K + N(t)} - \delta x_i + \varepsilon_i y_i - s_i x_i \quad (\text{B.3a})$$

$$\frac{dy_i}{dt} = \frac{p y_i}{K + N(t)} - d y_i - \varepsilon_i y_i + s_i x_i \quad (\text{B.3b})$$

$$\begin{aligned} \frac{dV_{x_i}}{dt} &= \left(\frac{\pi(K + N(t)) - \pi x_i}{(K + N(t))^2} - \delta - s_i \right) 2V_{x_i} + 2\varepsilon_i C_{x_i, y_i} \\ &\quad + \left(\frac{\pi}{(K + N(t))^2} + \delta + s_i \right) x_i + \varepsilon_i y_i \end{aligned} \quad (\text{B.3c})$$

$$\begin{aligned} \frac{dV_{y_i}}{dt} &= \left(\frac{p(K + N(t)) - p y_i}{(K + N(t))^2} - d - \varepsilon_i \right) 2V_{y_i} + 2s_i C_{x_i, y_i} \\ &\quad + \left(\frac{p}{(K + N(t))^2} + d + \varepsilon_i \right) y_i + s_i x_i \end{aligned} \quad (\text{B.3d})$$

$$\begin{aligned} \frac{dC_{x_i, y_i}}{dt} &= \left(\frac{(\pi + p)(K + N(t)) - \pi x_i - p y_i}{(K + N(t))^2} - \delta - s_i - d - \varepsilon_i \right) C_{x_i, y_i} \\ &\quad + s_i V_{x_i} + \varepsilon_i V_{y_i} - s_i x_i - \varepsilon_i y_i \end{aligned} \quad (\text{B.3e})$$

We also require the variance of $x_i + y_i$, and so we compute $V_{x_i + y_i} = V_{x_i} + V_{y_i} + 2C_{x_i, y_i}$.