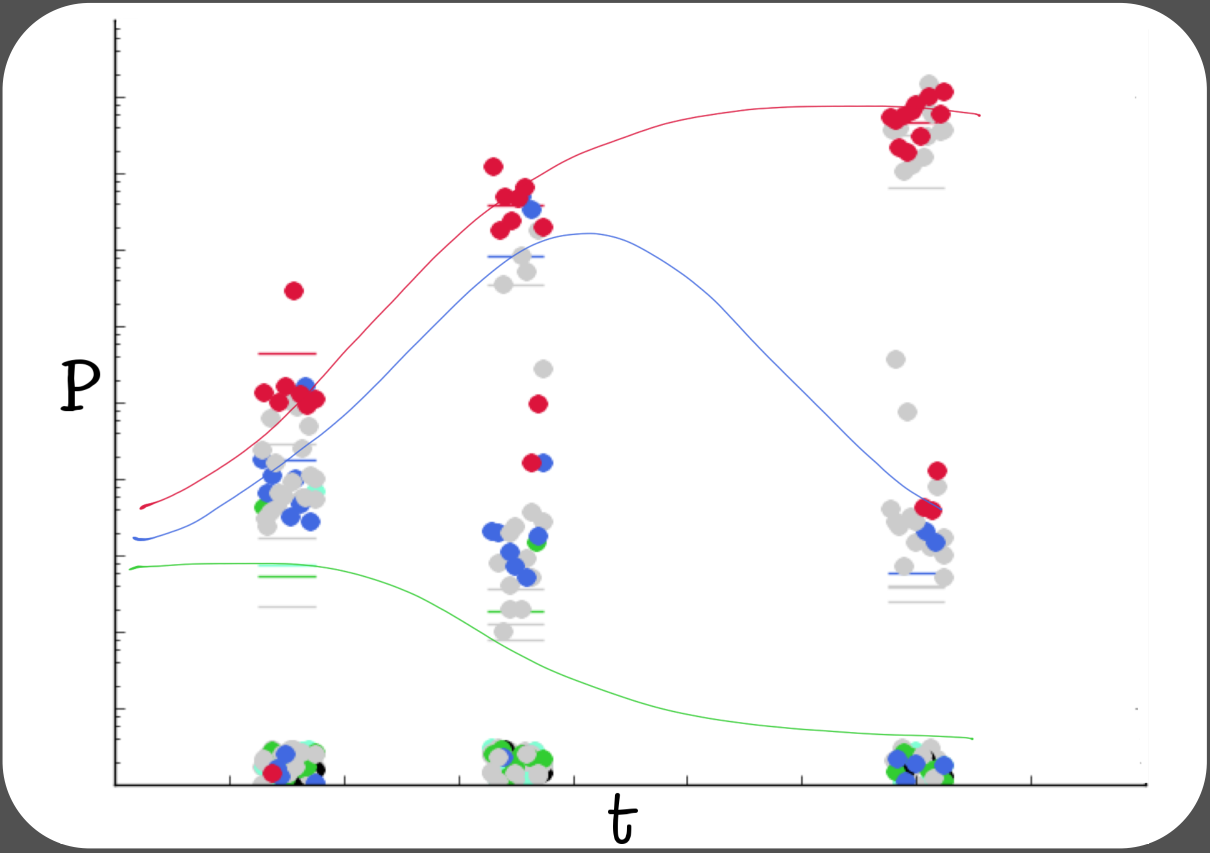
# Model-based inferences in host-pathogen-symbiont interactions

## Implications for the design of experimental and observational studies

Caetano Souto Maior Mendes



Dissertation presented to obtain the Ph.D degree in Biology

Instituto de Tecnologia Química e Biológica António Xavier | Universidade Nova de Lisboa

Oeiras,
February, 2017

# Model-based inferences in host-pathogen-symbinot interactions

## Implications for the design of experimental and observational studies

Caetano Souto Maior Mendes

Dissertation presented to obtain the Ph.D degree in Biology

Instituto de Tecnologia Química e Biológica António Xavier | Universidade Nova de Lisboa

Research work coordinated by:

FUNDAÇÃO CALOUSTE GULBENKIAN
Instituto Gulbenkian de Ciência

Oeiras, February, 2017

UNIVERSIDADE
NOVA
DE LISBOA

-4

supervisor: M Gabriela M Gomes

The author was the beneficiary of a doctoral grant from the AXA Research Fund.

# -3

Dedicated to all the people who actively contributed in any helpful way to the efforts that made this work possible, those who passively did not hamper these same efforts, and not in the least to those who disrupted the work in unfair, irrelevant or unhelpful ways.

*Let us think the unthinkable, let us do the undoable, let us prepare to grapple with the ineffable itself, and see if we may not eff it after all.*

Douglas Adams [Dirk Gently's Holistic Detective Agency]

# -2

*All right, but apart from the sanitation, the medicine, education, wine, public order, irrigation, roads, the fresh-water system, and public health, what have the Romans ever done for us?*

John Cleese [Monty Python's Life of Brian]

# -1 Acknowledgments

First and foremost, I would like to thank my supervisor Gabriela Gomes, who gave the opportunity to start, diverge, and finish with the work that is mostly presented in this thesis. Also indispensable to the development of this work were the many collaborators – whether explicitly acknowledged in the author contributions of the chapters and manuscripts that saw the light of the day so far, or not, as well as smaller interactions that often go unnoticed – in alphabetic order: Claudia Codeço; Claudio Struchiner; David Rasmussen; Denise Kühnert; Flavio Coelho; Gabriel Sylvestre; Luis Teixeira; Paulo Campos; Rafael Maciel de Freitas. I must also acknowledge the current and past members of the group: Bruno Ceña; Catia Bandeiras; Daniela Zwerschke, Delphine Pessoa; Erida Gjini; João Lopes; Sander van Noort; as well the few hundred people that constitute Instituto Gulbenkian de Ciência, except one. The PIBS current and former program directors: Élio Sucena and Thiago Carvalho, the committee for this thesis: Lounès Chikhi and Maria João Amorim. I would like to thank IGC members who were not directly or officially involved in my project, but that nevertheless helped in different ways: Alekos Athanasiadis; Claudine Chaouiya; Joana Loureiro; Joana Sá; Jocelyne Demengeot; Nelson Martins; Paula Duque – and especially colleagues of the Ph.D. programs, and those that were more than just circumstantial friends: Ana Ribeiro, Elvira Lafuente, Vania Silva, Irma Varela, Daniel Damineli, Leila Shirai, Tiago Macedo, Romulo Areal, Vitor Faria, among many others that made this period easy to endure most of the time. To the thesis jury, Professor Ary Hoffmann, Dr. Simon Frost, Dr. Isabel Gordo, and Dr. Ana Nunes, for taking time off their own work to examine mine. Others (from institutions academic or otherwise) I know or had the chance to interact with and helped in whatever meaningful way, be it practical or inspirational*: Alexei Drummond, Emily Jane McTavish, Inger Winkelmann, Kate Langwig, Marc Lipsitch, Mark Thomas, Michael B. Eisen, Michael Turelli, Nick Barton, Olivia Judson, Ricardo Cavalcanti, Roland Regoes, Sebastian Bonhoeffer, Simon Frost, Simon Tavaré, Tanja Stadler, Troy Day, Vinicius Jucá. To others that I may be forgetting, and can understand that it is not possible to remember everyone. To Rosa Souto Maior, my mother, who as a scientist and professor – and therefore overworked for over 35 years – never discouraged me to pursue a career in research; to Lucio Mendes, my father – despite being in the private sector for over 40 years – always encouraged and supported me to go after whatever I wanted. To my friends and family for support, including financial from my parents as well as grandparents, without which maybe it would not be possible to go through unemployment periods before being able to secure funding for a position in research (and sometimes during) – because choosing to do research entails an enormous financial opportunity cost before, during, as well as after the PhD, which is responsible for many people giving up sooner or later and does not seem to be improving. That said, I want to acknowledge the AXA Research Fund for funding 3 years of my research, and the Portuguese Foundation for Science and Technology (FCT), and Fundação Calouste Gulbenkian for additional funding. To the many in practice anonymous people in the open online communities, especially but not limited to Stack Exchange/Overflow, who contribute to open knowledge, without which many details of this work would have taken much longer. To the open source software and programming resources – Python, Debian Linux, Emacs, as well as the LaTeX typesetting template of this thesis – to open access journals, activists and enthusiasts. To people enduring horrible situations due to all sort of neglected academic issues, or taboo in general, who end up depressed, sick or dead. To people I do not know personally, but that took a stand for causes that, while supported by most scientists and activists, still entail an absurd personal risk, such as Alexandra Elbakyan's with Sci-Hub, and Aaron Swartz's progressive efforts to change a ridiculous system that ultimately cost him his life. To open and occupy everything.

*including the quotes in the chapters, where stars indicate paraphrases from what I remember of their words, and take full responsibility for anything stupid or wrong that may have been misquoted by myself

Thesis advisor: M Gabriela M Gomes Caetano Souto Maior Mendes

## Summary

Parasitism has probably existed since the first two organisms were made to interact, and disease has been recognized by humans since long before they realized microscopic organisms could cause it. Despite that realization and the immense scientific progress made in understanding the processes parasites undergo at the many different levels – from molecular to population scales – it is still not trivial to describe how infection happens, and quantifying the underlying processes is essentially an unresolved question. Most scientific work on pathogens is restricted to a single level of organization, and integration is not only rare, but difficult to describe conceptually, and even more so formally or mathematically. Attempts to do so with theoretical-physics-like approach have been hindered by the greater complexity and lack of understanding of biological systems and of appropriate data when compared to physico-chemical systems like quantum mechanical descriptions of atoms and molecular structure, for instance. In the work developed towards the PhD degree, most of which is described in this thesis, some of the gaps in concepts, methods and data incorporation concerning disease transmission, and pathogen proliferation in geral are addressed.

The introduction describes the problem, some of the systems of interest for investigating it, the accepted theoretical basis, data, and other current methods available to tackle the problem. Chapter 1 uses model organism *Drosophila melanogaster*, endosymbiont *Wolbachia*, and *Drosophila* C virus, for dose-response experiments to assess susceptibility to infection in a novel way that allows the estimated parameters to carry to the population level, as well as taking into account heterogeneity in host susceptibility. Chapter 2 explores invasion of a population of insects by *Wolbachia* inducing cytoplasmic incompatibility, and other life-history modulating effects such as fecundity reduction, lifespan reduction, and protection against pathogens – in both a homogeneous and heterogeneous way, as described in the preceding chapter. Chapter 3 investigates the time course of the levels of type 1 dengue virus and *Wolbachia* inside *Aedes aegypti* hosts for different initial challenges by using a dynamic model to describe infection. The final results chapter, number 4, describes population transmission models of dengue virus in a population, by means of forward simulation with parameters in an acceptable range, as well as inference from simulated epidemics using time series of incidence as well as coalescent-based estimation from sequences. The methods are also applied to real data from the city of Rio de Janeiro, obtaining preliminary results for that real setting under one or two-serotype models.

Chapter 5 discusses the results and tries to relate the different levels explored, showing how they shed light on unknown features of pathogen proliferation; it also tries to acknowledge limitations in the methods and structure of the data gathered and produced for the models proposed, why some of the objectives could not be achieved, and what the author of the thesis learned about the entire process of tackling such a big issue.

Thesis advisor: M Gabriela M Gomes Caetano Souto Maior Mendes

## Sumário

O parasitismo provavelmente existe desde que os primeiros dois organismos tiveram de interagir, e as consequências de microparasitas foram reconhecidas por humanos desde muito antes da descoberta dos parasitas em si. Apesar desta descoberta e do imenso progresso feito para o entendimento dos patógenos nos mais diferentes níveis – desde o molecular até a escala da população – a descrição de como uma infecção acontece continua não sendo uma tarefa trivial, e a quantificação dos processos envolvidos é uma questão não resolvida. A maior parte do trabalho científico sobre patógenos fica restrita a um nível de organização, e a integração do conhecimento sobre diferentes níveis não é só rara, mas difícil de descrever conceitualmente, e ainda mais de maneira formal ou matemática. Tentativas de abordar o problema como a física teórica foi tratada no passado esbarram na maior complexidade e falta de entendimento dos sistemas biológicos, e falta de dados adequados em comparação com as descrições de mecânica quântica de átomos ou estrutura molecular. No trabalho desenvolvido para o grau de doutor – a maioria do qual está descrito nesta tese, algumas das lacunas nos conceitos métodos – e incorporação de dados relacionados à transmissão de doenças, e proliferação de patógenos em geral é discutida.

A introdução desta tese descreve o problema, o contexto, e alguns dos sistemas de transmissão de interesse, as bases teóricas aceitas, os métodos, e o tipo de dados disponível para tentar entender infecção e transmissão.. O primeiro capítulo de resultados descreve o uso de *Drosophila melanogaster*, endosimbionte *Wolbachia*, e *Drosophila* C virus para a realização de experimentos de dose-resposta e inferência da suscetibilidade à infecção de uma maneira inédita que permite a utilização dos valores estimados em um contexto popula-

cional, além de levar em consideração a heterogeneidade na suscetibilidade do hospedeiro. A seguir (capítulo 2), um model matemático é utilizado para explorar o impacto de um simbionte que induz incompatibilidade citoplasmática e modula outros aspectos como redução de fecundidade e expectativa de vida, além de proteção contra patógenos – o que é incorporado num contexto homogêneo e heterogêneo, como descrito no capítulo anterior. O penúltimo capítulo de resultados (capítulo 3), investiga a progressão ao longo do tempo dos níveis virais de dengue tipo 1 e *Wolbachia* em mosquitos *Aedes aegypti* infectados com diferentes inóculos do vírus – o que é feito utilizando um modelo dinâmico para descrever os dados observados. O capítulo final dos resultados (capítulo 4) descreve a transmissão de dengue na população, utilizando modelos para um ou mais sorotipos para simulações com parâmetros mais ou menos conhecidos, além de inferência e validação a partir de séries temporais e sequências virais simuladas – os métodos são aplicados a dados reais para a cidade do Rio de Janeiro, obtendo estimativas para transmissão sob modelos com um ou dois sorotipos.

O último capítulo (número 5) discute o conjunto dos resultados e tenta relacioná-los entre os diferentes níveis de organização explorados, e tenta mostrar que questões foram iluminadas por este exercício. No mesmo capítulo são ressaltadas as limitações dos métodos, modelos, dados, e do trabalho como todo, e se explica por que alguns dos objetivos não foram alcançados, além de resumir o que o autor da tese aprendeu sobre o processo completo de tentar investigar um tópico em profundidade.

# Contents

# List of figures

# List of tables

*"Idealism increases in direct proportion to one's distance from the problem."*

John Galsworthy

# 0 Introduction

## 0.1 Mosquitoes, vectorial capacity, and infection

Vectors that transmit human disease are associated to the important concept of vectorial competence: their individual ability to harbor and transmit pathogens, e.g. a mosquito is able to ingest a virus, have it multiply in enough numbers, and reach a tissue where it can be transmitted to next host (Hotez et al. 2016; Lambrechts 2011). Complementing the concept of competence is that of vectorial capacity, which describes not only the individual but the entire of population to sustain disease transmission; for instance, it is important to have a large enough population of mosquitoes and they need to move and bite enough infected humans to keep the circulation of a pathogen – capacity therefore includes population and ecological variables as well as individual properties (Pan et al. 2012).

Although we could go as far as to say that a host is infected if it has pathogens in or on it at any time, this is probably not a useful definition; most likely infection is not trivially defined (Schneider 2011). It is more accurately described as a dynamic process, where a microorganism comes into contact with the host, colonizes a niche (tissue section, organ, or cavity, for instance), and multiplies as any macroscopic species would (Antia et al. 1997). As a species a population of microbes will interact with other species in its environment, and with the environment itself, i.e. the host, being susceptible to their version of the weather and natural disasters. It is unlikely that someone would describe a valley as populated because a few people passed by once, and similarly infection requires a more comprehensive definition.

A viral inoculum ingested by a mosquito, for instance, goes into the insect's gut, which presents itself as a strong barrier against crossing into the insect body cavity; once that is surpassed, the viral particles that managed to get through can potentially reach different organs, at which point the pathogen is further faced with immune responses. Whether a particular pathogen challenge results in a population of particles that succeeds in generating a systemic infection depends on the balance between pathogen replication, its ability to bypass passive barriers, the host responses aimed at eliminating the pathogens, and its attempts to evade these attacks. This is a complex, multifactorial system that is guided by nonlinear processes and interactions, which are affected by stochasticity in every step.

If the pathogen is successful, it is able to generate a systemic, persistent infection, and to reach tissues important for further transmission such as salivary glands; if not, it may be completely eliminated or have such low numbers that they are not able to successfully transmit to the next host (Franz 2015).

Under that light, even the more restricted concept of competence is not straightforward to define, and is even harder to quantify; therefore its assessment depends on proxies such as presence of virus after some time, which do not contemplate subtle and complex aspects of infection such as time and dose dependence (Pessoa et al. 2014).

## 0.2 Disease, epidemiology, and health data

Many kinds of epidemiological data reflect different aspects of microbes and their transmission in the population. Although invisible to the naked eye, pathogens have macroscopic manifestations that can be spotted without special tools, notably disease itself.

Records of disease cases are produced every time a diagnostic is made by a doctor; whether that record is properly stored and organized is then a matter of a specific data collection system being in place (Guinovart et al. 2008; Souto-Maior 2011; SINAN). The result of a clinical assessments based on observation of symptoms is most common; in some cases laboratory confirmation of some sort can also be used as a more refined assessment, but although increasing confidence or precision this does not change the nature of the data. The result is then a report of a new case, or persistence of a known case; traditionally, epidemiologists used to think about disease transmission in terms of these quantities: incidence or prevalence of disease.

To be sure, prevalence is an instantaneous measure of the total number or proportion of infected people in the population; like the concentration of a compound in a solution, it is the "density" of infected individuals. Incidence is a rate: the number (or proportion) of people that acquired the disease in some time interval. Measuring prevalence at any point in time relies on being able to count everyone that is carrying the disease then, or at least get an estimate of the proportion of people with the disease – it is harder to think of a system that would repeatedly record that people are still sick, so obtaining this data is probably an active effort. Incidence is arguably easier to measure, because unlike prevalence it relies only on being able to count new cases as they appear, and people will commonly present themselves to a health service when they first get sick. It is expected that not everyone who is sick will see a doctor, so there should be under-reporting of incidence that may be related to severity of the disease, accuracy of the diagnostic, among other factors, and it is usually not trivial to estimate the amount of unreported cases (Gibbons et al. 2014). Because for many purposes both kinds of data are equally useful, whether one or the other is used depends mainly on availability, as well as representativity of the specific disease and also convenience.

A more modern method of observing pathogens indirectly is through the immune responses elicited against them; identifying antibodies specific to a virus or bacterium yields a different kind of data, that is, whether an individual was infected in the past. Arguably, it is a worse kind of data; it says whether an infection happened or not, but not when it happened (unless individuals are periodically assayed for seroconversion, i.e. conversion from absence to presence of antibodies for a specific pathogen). On the other hand, the assay does not rely on testing individuals during the narrow window of time when individuals are sick. It also potentially allows characterizing secondary infections at once (i.e. occurrence of multiple infections can be assessed with a single test). This obviously relies on a laboratorial infrastructure, and sensitive and specific reagents to detect the signatures of an infection and differentiate them from other infections (Domingo et al. 2012).

Pathogen can be observed directly, through a microscope (although not literally in the cases of viruses, for which the closest would instead be an electron microscope); the parasitemia measured in this way is a common way of identifying malaria infections, although probably not practical for large scale, routine detection. Another direct way of detecting a parasite without a microscope is to identify its genetic sequences in a sample through any one of a series of molecular techniques, among which polymerase chain reaction (PCR) is probably the current choice (Saiki et al. 1985), and for some applications sequencing the genome of the pathogen (Heather & Chain 2016).

Observing a parasite directly does not necessarily generate an entirely different kind of data – it may just give greater confidence, which is otherwise no different qualitatively from a clinical diagnosis by an experienced doctor. Nevertheless, different information can be obtained from direct observation or molecular techniques: parasitemia can potentially be quantified and give an additional dimension of the epidemic, e.g. the distribution of parasite loads in the infected individuals; specific strains of pathogens can be ascertain; and the nucleotide sequence can be verified by sequencing methods. There are many conceivable kinds of data about pathogens and epidemics, and the conception of these kinds can be quite arbitrary, for instance: the size distribution of *Plasmodium* parasites; the color of the skin of people infected with hepatitis; the behavioral changes in zombies (Brooks 2006) – whether any of these, as well as the above described types of data are useful depends on the question at hand. Put another way, it depends on what you want to infer from the data.

## 0.3 Dynamic models and theoretical epidemiology

A dynamic model is by definition a function where some variable (let call it $y$) depends on time ($t$), or $y = f(t)$. It can be as simple as a linear combination of the time variable itself, e.g. $y = at + b$ (or a polynomial $y = at^2 + bt + c$), or a nonlinear function such as $y = Ne^{rt}$. Nevertheless, it is important to realize that these explicit formulations are rarely available through an analytical mathematical solution, and numerical methods have to be used instead for many common models.

It is often, if not always, easier to formulate dynamic models in terms of the processes that define the rate of change of one or more variables, that is, by using differential equations on time. For instance, while we known that the explicit equation $y = Ne^{rt}/(1 + kN(e^{rt} - 1))$ defines a logistic growth curve, it is probably more intuitive to describe the model implicitly as $\frac{dy}{dt} = ry - ky^2$, where growth can be seen to depend on the value of $y$ and the parameter $r$, and decrease depending on $k$ and the square of $y$ (therefore $y$ will grow if its value is "small", and eventually saturate when its value becomes "large", assuming $r > k$) Similarly, instead of trying to guess the explicit form of a mathematical model of an epidemic, the basic processes can be specified instead: individuals susceptible to disease ($S$) get infected proportionally to the number of infected individuals ($I$) at some transmission rate $\beta$, that yields a simple equation for a variable corresponding to the proportion of individuals infected: $\frac{dI}{dt} = \beta IS$ (although it may be necessary to specify an equation for the suscpetible individuals variable, which may nevertheless be as simple as realizing that the sum of the two proportions equals unity, i.e. $S + I = 1$, and $\frac{dI}{dt} = \beta I(1 - I)$).

Most of what is recognized as theoretical epidemiology, mathematical epidemiology, or modeling of infectious diseases is based on this simple process of building transmission models (Kermack & McKendrick 1927); it consists essentially on gathering simple facts about disease transmission, and creating a self-consistent conceptual model, which can then be converted into a formal mathematical description of population transmission of a disease. For instance, a human population is first divided into kinds of hosts: susceptible to, infected with, and recovered from infection; how many people hold each status defines the overall state of the system. Then, the processes that increase or decrease the numbers of each kind of host are acknowledged: it was shown how the number of infected individuals increases, and as a consequence the individuals that become infected are no longer susceptible (by definition, at least during the period of infection), so they decrease by the same number; infected individuals may recover at some rate, and could either become susceptible again, or resistant.

The description of these compartments that the hosts can occupy, as well as the processes by which they move between them are a mathless description that underly what are also called compartmental models, and can be easily converted into mathematical equations. The system just described is shown in figure 0.3.1, and is known as the SIR model in its most basic version.

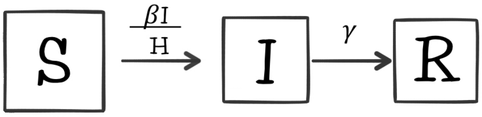


**Figure 0.3.1:** SIR model. Arrows indicate rates at which any individual from one compartment moves to the next.

The SIR model is regarded as a generic model of disease transmission (Anderson & May 1981; Heesterbeek & Roberts 2015); it may still apply to specific diseases that do not have very specific quirks, but otherwise it serves as the backbone of more realistic models. Processes like demography and other characteristics of the population can be integrated to the basic structure: individuals that are born (or migrate) into the population are susceptible, and all individuals die whether or not the causes are related to the disease; absolute numbers (not just proportions) and population size may be used to characterize the compartments; and an endless number of aspects can be added to a model. A straightforward way of incorporating the details above, for instance, is assuming the population has a constant size $H$, maintained by a birth rate $m$ equal to the death rates of each compartment; if transmission is dependent on the total size of the population it may be written as

$\lambda S = \frac{\beta I}{H} S$; everything in the transmission rate term except for the number of susceptibles ($S$) is also known as the force of infection, and is often denoted by $\lambda$.

This modified version is shown in the system of equations 1, and is more useful if the time scales of demography are comparable to that of the depletion of susceptibles in the population:

$$\begin{aligned} \frac{dS}{dt} &= mH - \frac{\beta I}{H} S - mS \\ \frac{dI}{dt} &= \frac{\beta I}{H} S - \gamma I - mI \\ \frac{dR}{dt} &= \gamma I - mR \end{aligned} \tag{1}$$

While model 1 is arguably a more realistic one, there are still endless ways of modifying it to become even more realistic or conform to a specific feature of one or another disease. Common models include alternatives where individuals do not recover, or lose protection after some time; these may be referred to by their compartmental structure – SI and SIRS, respectively – among other simple basic models (Wikipedia).

Estimates for some parameters in the model are sometimes available through observation or experiment, e.g. recovery period for many acute viroses is often 7 days, while others like transmission rate are not easily estimated directly – (but see Gomes et al. 2014; Pessoa et al. 2014). Assuming the compartments are able to correctly describe the state of the population at any point in time (which is no small assumption), and epidemic can be simulated for a given set of parameters in the model.

## 0.4 Analysis of data and statistics

### 0.4.1 Observational studies and statistical tests

Given some piece of data, some features may be glaringly obvious, and no more than looking at it could be necessary to learn something – e.g. the prevalence of HIV has not increased considerably since the early 1980s (WHO); malaria incidence in Mozambique is not constant, but has a seasonal peak in the rainy months (Guinovart et al. 2008); influenza is barely detectable outside of the winter months (van Noort et al. 2012), etc.

Eventually, more subtle observations are likely to require statistical analyses. For instance, the prevalence of tuberculosis in separate locations is almost guaranteed to be different; nevertheless, the differences may be entirely due to chance, and the methodology for comparing this or any other quantities must take into consideration factors such as sample size, and a number of other variables. Study design and analysis are therefore important to describe disease, and a myriad of methods have been developed to do just that, even comprising an entire research area called biostatistics.

### 0.4.2 Least-squares vs. likelihood frequentist inference

Beyond summaries and statistical tests, it may be of interest to infer parameters, which pertain to some model; therefore, in the absence of independent parameter estimates for SIR-like models, they could be estimated from data that represents the output of the system. Inference is also a research area on its own inside statistics; nevertheless, applying the body of knowledge to the estimation of parameters from dynamics models is far from a straightforward task. Even for simpler models, the problem of inferring parameters is not a trivial one, and there is no magic bullet nor one-size-fits-all solution; in the case of linear models it is possible to obtain an analytical solution to the problem of minimizing the sum of the squares of the distances from the model to the data, e.g. the data points to the curve (Weisstein), which is known as the method of *Least-Squares*.

Although the least-squares method is intuitive and has an exact solution, it has some undesirable properties and underlying assumptions; without getting into the technical details, it suffices to say that the method can be shown to be a restricted case of the *Maximum-Likelihood Estimation* (also known for its acronym: *MLE*). Maximum-Likelihood Estimation has the disadvantage of not having a general analytical solution, and therefore relying on optimization algorithms that are not guaranteed to find the best solution; nevertheless, for nonlinear models, the Least-Squares method that solves a linear approximation to the model relies on similar methods and therefore has much of the same caveats.

The concept of "likelihood" is related to specific probabilistic distributions for some data, and can be an elusive one – possibly also due to its flexibility. I will therefore not try to give a formal or comprehensive explanation of *the likelihood* in the abstract; instead I will try to illustrate the advantage of formulations based on likelihood over least-squares and describe it associated to the models in the thesis. One example is a classic coin-toss example: the data for a series of coin tosses can be represented by the number of heads (and conversely of tails) obtained; it is straightforward to compute the likelihood of the data as a binomial process of $N$ coin tosses and $k$ successes (either the number of heads or tails) and probability $p = 0.5$ (if the coin is unbiased, as expected): $\binom{N}{k} p^k (1-p)^{N-k}$. As a conceivable alternative, it is not possible to easily define the distance of the data to a curve (i.e. specify the least squares problem).

Likelihood-based methods are also useful for model comparison when the alternative models are not nested, i.e. one model is a particular case of the other, and benefit from its relationship to concepts of information theory (Akaike 1974) – thus, likelihood-based methods are preferred for a number of reasons, including that they are broader and have useful properties not displayed by other formulations of error models.

### 0.4.3 Bayesian Inference

Orthogonal to the likelihood vs. least squares distinction there is the difference between Bayesian and frequentist statistics, the latter of which most researchers call simply statistics (at least in fields like biology where statistics is used mainly as a tool). While bayesian inference abides to the principles of likelihood based estimation, it arguably extends them in ways that make it more natural, easier to understand, and easier to find out when it is going wrong – which when working with nonlinear dynamics models can be a considerable share of the time – less commonly it is overhyped, quoted as useful outside of statistics, and marketed as a new philosophy (Horgan 2016). The basic principle that underpins everything is a simple statement on conditional probabilities of two variables (Gelman et al. 2013), and was put forward by 18th century priest Thomas Bayes:

$$p(\theta|y) = \frac{p(\theta)p(y|\theta)}{p(y)} \tag{2}$$

If $\theta$ is taken to be the model parameters, and $y$ is the data, the probability of the parameters given the data ($p(\theta|y)$, i.e. the parameter estimates) can be calculated from the probability of the parameters ($p(\theta)$) and probability of the data given the parameters ($p(y|\theta)$, i.e. the likelihood of the data). The estimate of the parameters actually comes in the form of a distribution, the posterior distribution, or simply the *posterior* (as opposed to the distribution of *prior*, i.e. $p(\theta)$). Typically, it is not possible to compute the posterior exactly, and instead it is necessary to use numerical methods to arrive at the distribution; this is often done with Markov Chain Monte Carlo (*MCMC*) implementations. A basic implementation still commonly used was developed decades ago by Metropolis et al. (1953) and Hastings (1970); the relatively intense computational resources required explain the only relatively recent increase in the use of these bayesian methods, which would have been unfeasible without inexpensive powerful computers.

In a nutshell, the method works by starting an iterative chain with some parameter values and proposing new values at each iteration of the chain (based only on the current values, hence the Markovian property), and accepting the proposal using the likelihood function as a criterion to assess improvement – at the end of a number of iterations a large number of samples will form a distribution of the parameter values, and the mean (or median, or mode), as well as the confidence intervals can be computed from the distribution of parameter values. [This is likely to be an unwarrantedly short description of bayesian *MCMC* estimation, but except showing the perceived reasons why it should be favored over other methods for other chapters of the thesis, the technical details are quite mainstream and can be found on textbooks such as Brooks et al. (2012); Gelman et al. (2013).]

A Metropolis-Hastings sampler should be able to make a Markov Chain converge for a large number of iterations, and produce a distribution of samples that are nearly independent, resulting in useful posterior distributions of the model parameters. There are nevertheless many reasons why an *MCMC* implementation could not converge in a reasonable time, or at all. Convergence problems with Markov Chain-based bayesian estimation can nevertheless still be seen as an advantage over maximum-likelihood (and nonlinear least-squares) methods. While *MLE* will spit out an answer it is mostly a black box, and the inner workings of any particular run of the method are essentially unknown; on an *MCMC* chain the lack of convergence is potentially visible, as is the presence of local maxima.

## 0.5 Genealogies, phylogenies, population genetics, and "phylodynamics"

Genealogies have been an important concept in biology since Charles Darwin sketched a tree of life in his Origin of Species (Darwin 1859); although morphological criteria can be and has been used to establish relationships between species, the advent of molecular data, and especially nucleic acid sequencing allowed phylogenetics to be developed into something systematic (Salemi & Vandame 2003). Phylogenies illustrate a kind of data that demands reasonably sophisticated descriptions – it is not possible (not in any way one could easily think of) to describe pairwise clustering as a least-squares problem – and it took some time before practitioners could advance from more empirical neighbor-joining and parsimony methods towards likelihood-based methods (Felsenstein 1981), which converted phylogenetics into a statistically-rigorous, model-based method (as well as almost immediately allowing bayesian methods to be applied to it).

In the likelihood-based implementations, a mutation model underlies estimation (with a substitution matrix being assumed and clock rate possibly being estimated) and branch lengths are estimated; otherwise phylogenetics is essentially a clustering method, giving information about the relationship of the samples at hand, but most likely yielding limited information about the population where it came from. Population genetics, on the other hand, is concerned primarily with population dynamics and processes; to that end, the bifurcating properties of genealogies can be used, although the trees themselves are often treated as nuisances (Rosenberg & Nordborg 2002). Bifurcation along time is seen as merging backwards in time, or *coalescence* – the coalescence process or simply "the coalescent" contains important information not only about the individuals sampled but about the entire population where it came from, and it is a powerful tool in population genetics (Charlesworth & Charlesworth 2010).

Entire books have been written as introductions to coalescent theory (Wakeley 2009) or the implementation of general bayesian *MCMC* methods (Drummond & Bouckaert

2015), to which I refer for a more detailed description. The incorporation of nonlinear models like the basic SIR are described by an array of recent publications; I cite the work of Volz (2012) as a self-contained description of arbitrary, implicitly specified (e.g. by differential equations) non-linear population models in the coalescent framework; I also cite earlier contributions by Frost & Volz (2010), and Koelle & Rasmussen (2012), with no ambition of providing a comprehensive list, but instead referring to the five pieces of work which I have personally found most useful.

Finally, I briefly refer to the Wikipedia "viral phylodynamics" entry (as it is sometimes called when it involves nonlinear models and viral sequences), and its original source (Volz et al. 2013) as a primer to the topic. That said, I provide here a very brief description of genealogy-based inference.

Given a tree branching pattern, or an ordered topology with sequences at the tips, the expected time before the first two sequences find a common ancestor is a function of both the birth rate and population size (if the tree nodes are ordered, the first pair of sequences to coalesce is a given information, the actual time is not), as shown by Volz (2012); i.e. the probability that two sequences coalesce is dependent on a population function. It can also be stated intuitively that the probability that a certain number of mutations occur after some time is dependent on a nucleic acid base substitution model (which includes the mutation rate). Therefore, for a population function given by some arbitrary mathematical model (from which the probability of a pair of sequences to coalesce can be calculated), and a substitution model (from which the probability of a discrete number of mutations to occur can also be calculated, e.g. a poisson distribution with parameter equal to the mutation rate), the probability of a specific number of mutations happening before the samples find a common ancestor can be computed (Wakeley 2009).

That can be done for all the coalescence events, i.e. for the whole tree. Therefore, the likelihood that the (compound population plus substitution) model produced the observed sequences can be computed for any given tree. From there, the real-valued parameters (such as processes associated population size or mutation) can be inferred by a method such as *MCMC* estimation, although that assumes that the one tree underlying the model is adequate; additional methods may be necessary to explore tree parameter space, which is comparatively an odd variable type.

So like a model-based inference using incidence data, there is also a likelihood function relating sequence variation to both the tree and population function, and that can be used to describe the phylogenetic relationship but also the population processes.

## 0.6 Thesis outline and objectives

The remaining chapters are structured in the following way: chapter 1 uses a risk analysis kind of model coupled to survival models to estimate dose-independent parameters of susceptibility to viral infection for *Drosophila melanogaster* hosts with and without *Wolbachia* symbionts (Gomes et al. 2014; Pessoa et al. 2014).

Chapter 2 introduces dynamic models of insects, and analyzes the question of invasion of a resident population by a vertically-transmitted symbiont (e.g. *Wolbachia*) that can modulate the host's life history, including protection against horizontally-transmitted pathogens (Souto-Maior et al. 2015a).

Chapter 3 uses a within-host dynamic model to explain the time course data of infection of a *Aedes-DENV* system, inferring parameters related to the infection process, and relating the results to the risk analysis type of models.

Chapter 4 describes population models of disease transmission between mosquitoes and humans, and introduces multiple-serotype models. Forward and reverse modeling results are shown, the latter mostly with simulated data, but also with incidence from the city of Rio de Janeiro, where *DENV* is endemic.

Chapter 5 is a general discussion about how the chapters are related to each other, conclusions, and perspectives.

The general goal of this thesis is the development of a quantitative framework to analyze the introduction and impact of the *Wolbachia* symbiont into *Aedes* mosquito populations as a form of intervention to reduce disease transmission. Two scales are contemplated: the host and population scales. Chapters 2 and 4 are dedicated to population models, with the specific objective of unearthing what can be learned from this approach either through simulation or inference from different kinds of field data. Chapters 1 and 3 describe host-level experiments that give insight into infection of invertebrate hosts. The final chapter discusses how the two scales are related to each other, how the different models could be integrated, what would be the uses of doing so, and which kind of data or experiments would be additionally necessary to do so, as well as what is missing from the results, and what was found to be important after analyzing everything that is contained in this thesis.

# References

1. Akaike, H. 1974 A new look at the statistical model identification. IEEE Trans. Automat. Contr. 19, 716–723. (doi:10.1109/TAC.1974.1100705)

2. Anderson, R. M. & May, R. M. 1981 The Population Dynamics of Microparasites and Their Invertebrate Hosts. Philosophical Transactions of the Royal Society of London B: Biological Sciences 291, 451–524. (doi:10.1098/rstb.1981.0005)

3. Antia, R. & Lipsitch, M. 1997 Mathematical models of parasite responses to host immune defences. Parasitology 115, 155–167.

4. Bandeiras, C., Trovoada, M. J., Gonçalves, L. A., Marinho, C. R. F., Turner, L., Hviid, L., Penha-Gonçalves, C. & Gomes, M. G. M. 2014 Modeling Malaria Infection and Immunity against Variant Surface Antigens in Príncipe Island, West Africa. Plos One 9, e88110. (doi:10.1371/journal.pone.0088110)

5. Brooks, M. 2006 World War Z :an oral history of the zombie war. New York: Three Rivers Press

6. Brooks, S., Gelman, A., Jones, G. & Meng, X. L. 2011 Handbook of Markov Chain Monte Carlo. CRC Press

7. Charlesworth, B. 2010. Elements of evolutionary genetics. Roberts Publishers.

8. Darwin, C. 1859 On the origins of species by means of natural selection. London: Murray

9. Domingo, C., Escadafal, C., Rumer, L., Méndez, J. A., García, P., Sall, A. A., Teichmann, A., Donoso-Mantke, O. & Niedrig, M. 2012 First International External Quality Assessment Study on Molecular and Serological Methods for Yellow Fever Diagnosis. Plos One 7, e36291. (doi:10.1371/journal.pone.0036291)

10. Drummond, A. J. & Bouckaert, R. R. 2015 Bayesian evolutionary analysis with BEAST. Cambridge: Cambridge University Press. (doi:10.1017/cbo9781139095112)

11. Felsenstein, J. 1981 Evolutionary Trees From Gene Frequencies and Quantitative Characters: Finding Maximum Likelihood Estimates. Evolution 35, 1229.

12. Franz, A., Kantor, A., Passarelli, A. & Clem, R. 2015 Tissue Barriers to Arbovirus Infection in Mosquitoes. Viruses 7, 3741–3767. (doi:10.3390/v7072795)

13. Frost, S. D. W. & Volz, E. M. 2010 Viral phylodynamics and the search for an 'effective number of infections'. Philosophical Transactions of the Royal Society of London B: Biological Sciences 365, 1879–1890. (doi:10.1098/rstb.2010.0060)

14. Gelman, A., Carlin, J. B., Stern, H. S. & Rubin, D. B. 2013 Bayesian data analysis, Third Edition. CRC Press. (doi:10.1080/01621459.2014.963405)

15. Gibbons, C. L. et al. 2014 Measuring underreporting and under-ascertainment in infectious disease datasets: a comparison of methods. BMC Public Health 2014 14:1 14, 147. (doi:10.1186/1471-2458-14-147)

16. Gomes, M. G. M., Lipsitch, M., Wargo, A. R., Kurath, G., Rebelo, C., Medley, G. F. & Coutinho, A. 2014 A Missing Dimension in Measures of Vaccination Impacts. PLoS Pathog. 10, e1003849. (doi:10.1371/journal.ppat.1003849)

17. Guinovart, C. et al. 2008 Malaria in rural Mozambique. Part I: Children attending the outpatient clinic. Malaria Journal 2008 7:1 7, 36. (doi:10.1186/1475-2875-7-36)

18. Hastings, W. K. 1970 Monte Carlo Sampling Methods Using Markov Chains and Their Applications. Biometrika 57, 97–109. (doi:10.2307/2334940)

19. Heather, J. M. & Chain, B. 2016 The sequence of sequencers: The history of sequencing DNA. Genomics 107, 1–8. (doi:10.1016/j.ygeno.2015.11.003)

20. Heesterbeek, J. A. P. & Roberts, M. G. 2015 How mathematical epidemiology became a field of biology: a commentary on Anderson and May (1981) 'The population dynamics of microparasites and their invertebrate hosts'. Philosophical Transactions of the Royal Society of London B: Biological Sciences 370, 20140307–20140307. (doi:10.1098/rstb.2014.0307)

21. Horgan, J. 2016 (January) Bayes's Theorem: What's the Big Deal? Scientific American. http://blogs.scientificamerican.com/cross-check/bayes-s-theorem-what-s-the-big-deal/. Accessed 2016-07-20

22. Hotez, P. J., Pecoul, B., Rijal, S., Boehme, C., Aksoy, S., Malecela, M., Tapia-Conyer, R. & Reeder, J. C. 2016 Eliminating the Neglected Tropical Diseases: Translational Science and New Technologies. PLoS neglected tropical diseases 10, e0003895. (doi:10.1371/journal.pntd.0003895)

23. Kermack, W. O. & McKendrick, A. G. 1927 A contribution to the mathematical theory of epidemics. Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences 115, 700–721. (doi:10.1098/rspa.1927.0118)

24. Koelle, K. & Rasmussen, D. A. 2012 Rates of coalescence for common epidemiological models at equilibrium. Journal of The Royal Society Interface 9, 997–1007. (doi:10.1098/rsif.2011.0495)

25. Lambrechts, L. 2011 Quantitative genetics of Aedes aegypti vector competence for dengue viruses: towards a new paradigm? Trends Parasitol. 27, 111–114. (doi:10.1016/j.pt.2010.12.001)

26. Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. 1953 Equation of State Calculations by Fast Computing Machines. The Journal of Chemical Physics 21, 1087–1092. (doi:10.1063/1.1699114)

27. van Noort, S. P., Aguas, R., Ballesteros, S. & Gomes, M. G. M. 2012 The role of weather on the relation between influenza and influenza-like illness. Journal of Theoretical Biology 298, 131–137. (doi:10.1016/j.jtbi.2011.12.020)

28. Pan, J.-Y. et al. 2012 Vector capacity of Anopheles sinensis in malaria outbreak areas of central China. Parasit Vectors 5, 136. (doi:10.1186/1756-3305-5-136)

29. Pessoa, D., Souto-Maior, C., Gjini, E., Lopes, J. S., Ceña, B., Codeço, C. T. & Gomes, M. G. M. 2014 Unveiling Time in Dose-Response Models to Infer Host Susceptibility to Pathogens. PLoS Comput Biol 10, e1003773–9. (doi:10.1371/journal.pcbi.1003773)

30. Rosenberg, N. A. & Nordborg, M. 2002 Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. Nature Reviews Genetics 3, 380–390. (doi:10.1038/nrg795)

31. Saiki, R. K., Scharf, S., Faloona, F., Mullis, K. B. & Horn, G. T. 1985 Enzymatic amplification of b-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. Science, 230, 1350-1354.

32. Salemi, M., & Vandamme, A. M. 2003 The phylogenetic handbook: a practical approach to DNA and protein phylogeny. Cambridge University Press.

33. Schneider, D. S. 2011 Tracing personalized health curves during infections. PLOS Biology 9, e1001158. (doi:10.1371/journal.pbio.1001158)

34. Sistema de Informações de Agravos de Notificação : http://ces.ibge.gov.br/base-de-dados/metadados/ministerio-da-saude/sistema-de-informacoes-de-agravos-de-notificacao-sinan.html. Accessed: 2016-06-23

35. Souto-Maior 2011 (Master's thesis) Patrones de morbilidad en niños menores de 15 años que acuden a los centros de salud de un área rural de Mozambique.

36. Souto-Maior, C., Lopes, J. S., Gjini, E., Struchiner, C. J., Teixeira, L. & M Gomes, M. G. 2015 Heterogeneity in symbiotic effects facilitates Wolbachia establishment in insect populations. Theoretical Ecology 8, 53–65. (doi:10.1007/s12080-014-0235-7)

37. Souto-Maior, C. 2015 Host–Symbiont–Pathogen–Host Interactions: Wolbachia, Vector-Transmitted Human Pathogens, and the Importance of Quantitative Models of Multipartite Coevolution. In Reticulate Evolution, pp. 207–230. Cham: Springer International Publishing. (doi:10.1007/978-3-319-16345-1_8)

38. Volz, E. M., Koelle, K. & Bedford, T. 2013 Viral phylodynamics. PLoS Comput Biol 9, e1002947. (doi:10.1371/journal.pcbi.1002947)

39. Volz, E. M. 2012 Complex Population Dynamics and the Coalescent Under Neutrality. Genetics 190, 187–201. (doi:10.1534/genetics.111.134627)

40. Wakeley J. 2009 Coalescent Theory: An Introduction. Greenwood Village: Roberts & Company Publishers.

41. Weisstein, Eric W. "Least Squares Fitting." From MathWorld–A Wolfram Web Resource. http://mathworld.wolfram.com/LeastSquaresFitting.html. Accessed: 2016-09-28

42. Wikipedia: Biostatistics. https://en.wikipedia.org/wiki/Biostatistics. Accessed: 2016-09-28

43. Wikipedia: Compartmental models in epidemiology. https://en.wikipedia.org/wiki/Compartmental_models_in_epidemiology. Accessed: 2016-09-28

44. World Health Organization: Global epidemic data and statistics. http://www.who.int/hiv/data/global_data/en/. Accessed: 2016-09-28

*"Inference is not the inverse of a hypothesis test."*

Andrew Gelman

# 1

# Unveiling Time in Dose-Response Models to Infer Host Susceptibility to Pathogens

Pessoa, D., Souto-Maior, C., Gjini, E., Lopes, J. S., Ceña, B., Codeço, C. T. & Gomes, M. G. M. 2014 PLoS Comput Biol 10, e1003773–9.

## Author contributions

The supervisor of the thesis, M. Gabriela M. Gomes (MGMG), conceived the experiments and designed the general analyses. The experimental work was performed by Bruno Ceña (BC), who infected the flies and followed their survival daily for the first 40 days, and the author of the thesis, Caetano Souto-Maior (CSM), who followed the flies for the latter 100 days. The main, time-dependent model was developed by Delphine Pessoa (DP), Caetano Souto-Maior (CSM), and Erida Gjini (EG). The main analysis and inference, as well as statistical confidence power tests and experiments, were implemented and performed by Delphine Pessoa. Further and preliminary analyses were performed by Delphine Pessoa (DP), Caetano Souto-Maior (CSM), Erida Gjini (EG), João Solari Lopes (JSL), Cláudia T. Codeço (CTC), Bruno Ceña (BC), and M. Gabriela M. Gomes (MGMG). The manuscript was written and revised by Delphine Pessoa (DP), Caetano Souto-Maior (CSM), Erida Gjini (EG), João Solari Lopes (JSL), Cláudia T. Codeço (CTC), and M. Gabriela M. Gomes (MGMG).

## Abstract

The biological effects of interventions to control infectious diseases typically depend on the intensity of pathogen challenge. As much as the levels of natural pathogen circulation vary over time and geographical location, the development of invariant efficacy measures is of major importance, even if only indirectly inferable. Here a method is introduced to assess host susceptibility to pathogens, and applied to a detailed dataset generated by challenging groups of insect hosts (*Drosophila melanogaster*) with a range of pathogen (Drosophila C Virus) doses and recording survival over time. The experiment was replicated for flies carrying the *Wolbachia* symbiont, which is known to reduce host susceptibility to viral infections. The entire dataset is fitted by a novel quantitative framework that significantly extends classical methods for microbial risk assessment and provides accurate distributions of symbiont-induced protection. More generally, our data- driven modeling procedure provides novel insights for study design and analyses to assess interventions.

## Author Summary

While control options for plant, animal, and human pathogens are emerging rapidly, reliable assessment of the effect of interventions in biological systems presents many challenges. A major question is how to connect laboratory experiments and measurements with the relevant process in natural settings, where hosts are subject to pathogen exposures that vary in time and geographical location. With this aim, measures of protection that are invariant under varying exposure intensity need to be developed and integrated with mathematical models. In this article, we introduce a method to assess host susceptibility to pathogens, and apply it to survival of *Drosophila melanogaster* challenged with different doses of *Drosophila* C virus. By replicating the procedure in groups of flies that carry the symbiont *Wolbachia*, we are able to estimate how the viral protection induced by this intracellular bacterium is distributed in the host population. Our results disentangle host infection status from observed mortality, accounting naturally for time since exposure. The multiple-dose design proposed challenges traditional study designs to assess interventions.

## 1.1 Introduction

Hosts exposed to disease-causing agents respond in accordance to the challenge dose. Therefore dose-response curves contain information about disease processes that can be extracted by suitable analytic frameworks. Early examples concerning microbial risk assessment include counting lesions caused by tobacco mosaic virus on plant leaves [1], as well as human responders to experimental challenge with polio viruses [2], *Vibrio cholerae* [3] and *Streptococcus pneumoniae* [4], for escalating challenge doses. Dose-response models have been in use for analyses and extrapolation of experimental datasets [5].

Models that account for the sigmoidal shape in log-linear scale of the typical dose-response curve have been derived mechanistically, based on the assumption that each individual pathogen has a probability of infection independent of others, the so-called independent action hypothesis [6]. This results in a one-parameter exponential-function model [7]. The frequent observation of shallower-than-exponential, or overdispersed, relationships has then prompted the implementation of heterogeneity in the probability of infection of individual hosts [8–10].

In the 1960s, Furumoto and Mickey [9] developed a dose-response model that could accommodate both shallow and steep increases in the response by considering the probability of infection of individual hosts described by a Beta-distribution. Although a mechanistic justification for this specific distribution has not been given, the model has been widely applied in microbial risk assessment due to its ability to outperform the simple exponential model [5].

Susceptibility distributions other than Beta have also been considered and are more commonly used in frailty models adopted in survival analysis [11], where the data consist of survivor counts over time in host groups that are constantly subject to a hazard [12,13]. These frailty models appeared in the 1980s and have since been adapted to infection hazards, where surviving signifies remaining uninfected [14–16]. While most informative when the exposure is continued or repeated over time, these formalisms would be inadequate for estimating distributions of susceptibility to infection from instantaneous challenge protocols.

The importance of accounting for time between challenge and observable toxicity responses to pathogens or other agents has been recognized. Recent models in ecotoxicology [17,18], consider explicit kinetics within exposed organisms. Also in microbial risk analysis, previous studies [19,20] have included time postinoculation as an additional parameter in classic dose-response models, although using an approach that conceptually allows for a different susceptibility distribution at each time point. Here we present a schema to infer a distribution of host susceptibilities to infection that holds consistently across dose and time. We introduce an experimental design and inference framework that enables such inferences by analyzing simultaneously a collection of survival curves, each representing a different challenge dose. The resulting Beta distributions are compared against those obtained by classic dose-response models based on single day measurements.

Recent evidence for symbiotic interactions that reduce host susceptibility to pathogens has stimulated the development of quantitative frameworks to assess the levels of individual and population protection attributable to specific symbionts. The intracellular bacterium *Wolbachia*, found among many arthropod species including *Drosophila melanogaster*, is one such symbiont [21,22]. To analyze the protection conferred by *Wolbachia* to *D. melanogaster*, we apply our inference framework simultaneously to two sets of time-dependent dose-response data: in one set the flies carry the symbiont bacterium *Wolbachia* (Wolb+); while in the other they do not (Wolb$^{-}$). In this instance we extract the Beta distribution that best describes individual protection attributable to *Wolbachia*, as well as population statistics valid across entire dose ranges.

## 1.2 Methods

### 1.2.1 Survival data

We used virus free *D. melanogaster* lines with DrosDel w$^{1118}$ background, with or without the endogenous *Wolbachia* strain wMelCS [21,23,24]. Flies were reared in standard food at 25uC. To assure that potential for heterogeneities are minimized by the experimental procedure, we used fifty 3–6 days old adult males per group, 10 per replicate and 5 replicates. To study the response to viral infection, we anesthetized with CO$_2$ and pricked flies with different doses of *Drosophila* C virus (*DCV*). We used tenfold serial dilutions – from 10$^1$0 TCID$_{50}/ml$ to 10$^4$ TCID$_{50}/ml$ – in Tris-HCl buffer, pH 7.5. Controls were pricked with buffer solution only. We used the pricking protocol described in [24], produced and titrated virus as in [21]. After pricking, we kept flies at 18°C and checked daily survival until day 80 and twice a week until the end of the experiment. Food was changed every 5 days. We summarized the data in 16 dose-response curves (8 per group, including control) from day 0 after treatment until day 139.

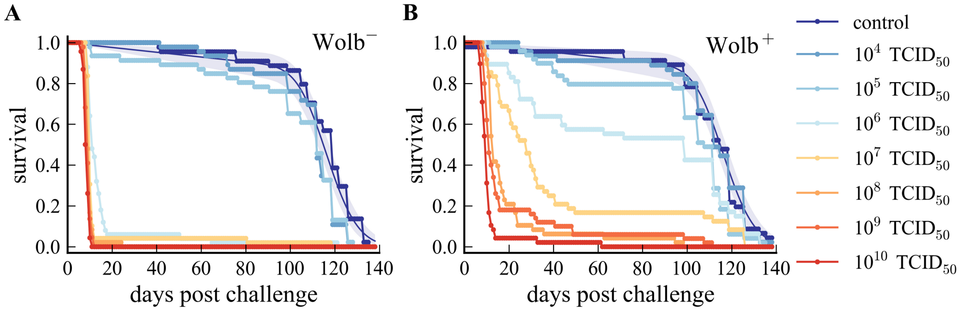


**Figure 1.2.1: Figure 1. Survival curves for Wolb$^-$ (A) and Wolb$^+$ (B) groups of** *D. melanogaster.* Dots represent experimental data. Dark blue curves show the model fit to the survival of control flies. Shaded areas represents 95% CI (credible intervals).

### 1.2.2 Dose-response model

Starting from established models, we refine the occurrence of mortality from infection, i.e. the *response*, as a function of the concentration of infectious units given to hosts, i.e. the dose. We present a step-by-step derivation of descriptions that integrate dimensions that are usually treated separately as well as the motivations for doing so.

Assuming independent action of infectious units, each unit has probability $p$ of causing an infection, while for $d$ infectious units infection occurs with a probability described by *Binomial(d, p)*. Given further considerations about the distribution of infectious units in a homogeneous solution (see [9] for a complete derivation of the expression), the number of units causing infection can be described by a Poisson distribution, resulting in the exponential dose-response model [7], that describes the probability of infection in a host challenged with pathogen dose $d$:

$$\pi_{hom} = 1 - e^{-pd} \tag{1.1}$$

This most basic formulation is hereafter referred to as the homogeneous dose-response model.

Furumoto and Mickey [9] expanded this formulation by allowing the probability of infection to be described by a parametric distribution, specifically the Beta distribution. To facilitate normalization across datasets, here we maintain the probability p fixed across individual hosts (as in [25]), and introduce a multiplicative parameter, the susceptibility factor $0 < x < 1$, to describe any natural or induced effect that decreases susceptibility. We assume that susceptibility to infection is Beta-distributed so as to describe the variation of susceptibility in the host population. Thus, we obtain the probability phet that a host contracts infection as

$$\pi_{het} = 1 - \int_0^1 e^{-xpd} q(x) dx \tag{1.2}$$

where $q(x) = x^{a-1}(1-x)^{b-1}/B(a,b)$ and $B$ is the Beta function. We refer to this formulation as the heterogeneous dose-response model.

At last we introduce a small parameter e to account for a small probability of ineffective challenge, such that $M \sim Binomial(n, (1-\varepsilon)\pi)$ is the random variable representing the number of infected hosts, in a group of n hosts challenged with a given dose. Assuming that an ineffectively challenged host behaves like a control host with regard to death rates, the probability that m hosts are dead a number of days after challenge is then

$$P(M=m) = \binom{n}{m}[(1-\varepsilon)\pi]^m[1-(1-\varepsilon)\pi]^{n-m} \tag{1.3}$$

where $\pi$ is either $\pi_{hom}$ (1) or $\pi_{het}$ (2) depending on which dose-response model is adopted.

The parameters to be estimated for this dose-response model are the maximum probability of infection per infectious unit $(p)$, the shape parameters for the Beta distribution that describes the susceptibility factor $(a, b)$, and the probability of ineffective challenge $(e)$.

These models require a choice of how many days post-challenge cumulative mortality should be measured, which is difficult to establish for host-pathogen systems where times to death from infection or other causes overlap significantly. To overcome this difficulty, we develop a model that integrates an explicit representation of time to death with the dose-response process for infection just described. It should, however, be noted that time is introduced with the main purpose of enabling the use of survival curves to obtain robust estimates for probabilities of infection given different challenge intensities and consistently infer susceptibility to infection. From this perspective, parameters defined from now on should be regarded as auxiliary and will be implemented as simply as possible.

### 1.2.3 Time-dependent model for control group

We first consider a survival model for a control group of flies pricked with buffer solution only (no $DCV$), subject to two hazards: $h_0$ , an age-dependent death hazard rate; and $h_k$ , a background age-independent death hazard rate. The overall death hazard rate for uninfected hosts is therefore

$$h_U(t) = h_0(t) + h_\kappa(t) \tag{1.4}$$

Denoting TU the random variable representing time to death of control hosts, we have

$$P(T_U = t) = P(T_\kappa = t)P(T_0 > t) + P(T_0 = t)P(T_\kappa > t) \tag{1.5}$$

where $T_0$ and $T_\kappa$ are the times to death from $h_0$ and $h_k$, respectively. Their corresponding distributions are assumed to be $T_0 \sim Gamma(M_0, s_0)$ and $T_\kappa \sim Uniform(1/\kappa)$ , where $\kappa$ is the background mortality rate, $m_0$ is the mean time to death, and $s_0$ is the shape parameter for the Gamma distribution of day of death from aging.

### 1.2.4 Time-dependent dose-response model

Hosts challenged with pathogen can become infected or remain uninfected and this infection status is hidden. If uninfected, they are subject to the age-dependent hazard rate that affects control hosts, hU ; if infected, they are subject to an infection hazard rate, h1, and the age-independent background mortality. Thus the overall hazard rate of infected hosts is

$$h_I(t) = h_1(t) + h_\kappa(t) \tag{1.6}$$

Now let $I \sim Binomia(n, (1 - \varepsilon)\pi)$ be the random variable representing the number of hosts infected by challenge with a given pathogen dose. Then the probability that $i$ hosts are infected after n hosts were challenged is

$$P(I = i) = \binom{n}{i}[(1 - \varepsilon)\pi]^i[1 - (1 - \varepsilon)\pi]^{n-i} \tag{1.7}$$

where $\pi$ is either $\pi_{hom}$ (1) or $\pi_{het}$ (2) depending on which dose-response model is adopted.

Let $T$ be the random variable representing the time to death of hosts challenged by a given pathogen dose. The probability density of observing a death event at time t given that i hosts are infected is

$$P(T = t|I = i) = \frac{n - i}{n}P(T_U = t) + \frac{i}{n}P(T_I = t) \tag{1.8}$$

where $T_I$ denotes the distribution of time to death of infected hosts, given by

$$P(T_I = t) = P(T_\kappa = t)P(T_1 > t) + P(T_1 = t)P(T_\kappa > t) \tag{1.9}$$

and $T_1$ is the distribution of times to death from the infection hazard rate $h_1$. This distribution is assumed to follow $T_1 \sim Gamma(m_1, s_1)$, where $m_1$ is the mean time to death of infected hosts, and $s_1$ is the shape parameter for the Gamma distribution of day of death from infection.

In setting the priors for parameter estimation we note that background mortality is small and therefore $\kappa$ is kept small by setting $1/k$ to be much greater than the last day of the experiment. To enforce that deaths due to infection occur earlier than deaths due to aging, we constrain the mean time to infection death to be lower than old-age death, i.e. $m_1 < m_0$ , and the probability of dying before the end of the study to be greater for infected hosts, i.e. $P(T_0 \leq t_{max}) \leq P(T_1 \leq t_{max})$, where $t_{max}$ is the last day of the experiment.

To construct the likelihood to be maximized by the parameter estimation procedure, we let $D_j$ be the random variable denoting the day fly $j$ died and $S$ the random number of survivors up to $t_{max}$. Then the likelihood of observing the actual number of survivors s and the times of death $d = [d_1, ..., d_{n-s}]$, for a given dose is

$$P(S = s, D = d) = \sum_{i=1}^{n} P(S = s, D = d | I = i) P(I = i) \tag{1.10}$$

$$= \sum_{i=1}^{n} \left[ P(T > tmax | I = i) \prod_{j=1}^{n-s} P(d_j - 1 < T < d_j | I = i) \right] P(I = i) \tag{1.11}$$

Since the observations for each dose are independent, taking the product of the likelihoods over the different doses yields the global expression for the likelihood of the entire dataset.

In this time-dependent dose-response model, the parameters to be estimated are the maximum probability of infection per infectious unit ($p$) used for normalization purposes, the Beta distribution shape parameters to describe variation in susceptibility factor $(a, b)$, the parameters that control death due to aging $(m_0, s_0)$, infection $(m_1, s_1)$, and background mortality ($\kappa$), as well as probability of ineffective challenge ($\varepsilon$). Parameters $\kappa$ and $\varepsilon$ are typically small and were introduced to improve performance of the likelihood.

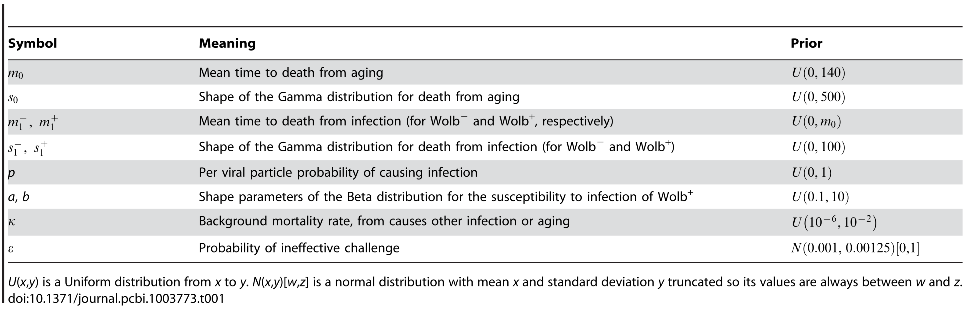
### 1.2.5 Parameter estimation

Model parameters were estimated using Markov chain Monte Carlo sampling implemented with the PyMC package [26] (code available from [27]). The prior distributions considered are listed in Table 1. Initial values were chosen so as to start with a non-zero likelihood. Using Metropolis-Hastings algorithm, we ran two separate chains for 252,000 iterations. The first 27,000 iterations were discarded. The recording interval was set to 250 so that the autocorrelation between samples was negligible. Convergence was assessed by inspection of the trace plots. All analyses were performed on the pooled samples from the two replicate chains.

## 1.3 Results

Groups of *Wolbachia*-negative (Wolb$^-$) and positive (Wolb+) *D. melanogaster* flies were challenged with a range of *DCV* doses and survival curves were traced as shown in Figure 1. This dataset was analyzed by applying the models introduced in Methods.

**Table 1.3.1:** Model parameters and their corresponding prior distributions.



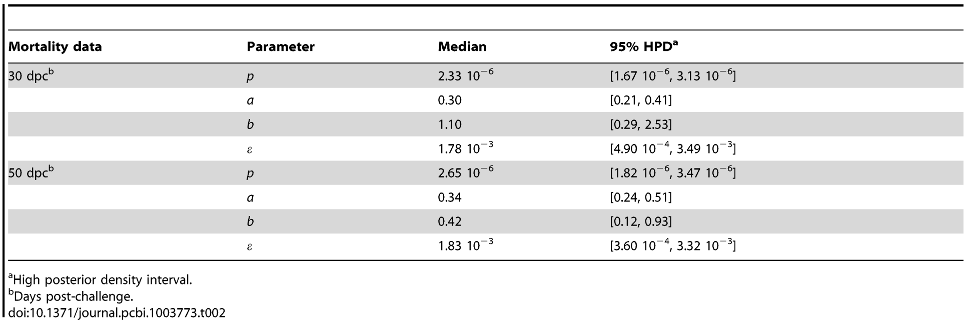
| Symbol | Meaning | Prior |
|---|---|---|
| $m_0$ | Mean time to death from aging | $U(0, 140)$ |
| $s_0$ | Shape of the Gamma distribution for death from aging | $U(0, 500)$ |
| $m_1^-$, $m_1^+$ | Mean time to death from infection (for Wolb$^-$ and Wolb$^+$, respectively) | $U(0, m_0)$ |
| $s_1^-$, $s_1^+$ | Shape of the Gamma distribution for death from infection (for Wolb$^-$ and Wolb$^+$) | $U(0, 100)$ |
| *p* | Per viral particle probability of causing infection | $U(0, 1)$ |
| *a*, *b* | Shape parameters of the Beta distribution for the susceptibility to infection of Wolb$^+$ | $U(0.1, 10)$ |
| $\kappa$ | Background mortality rate, from causes other infection or aging | $U(10^{-6}, 10^{-2})$ |
| $\varepsilon$ | Probability of ineffective challenge | $N(0.001, 0.00125)[0,1]$ |

*U*(*x*,*y*) is a Uniform distribution from *x* to *y*. *N*(*x*,*y*)[*w*,*z*] is a normal distribution with mean *x* and standard deviation *y* truncated so its values are always between *w* and *z*.
doi:10.1371/journal.pcbi.1003773.t001

### 1.3.1 Susceptibility distribution from selected day mortality

To emphasize the importance of day selection to infer distributions of susceptibility to infection by classic dose-response models [5] we have applied these procedures to mortality data observed by two specific days (30 and 50). Parameter estimates from these models are listed in Table 2. The model fits to the mortality data at the selected days are shown in Figure 2, as well as the associated distribution of Wolb+ susceptibilities and the posterior samples for the Beta distribution shape parameters. For simplicity we have adopted the homogeneous model for Wolb$^{-}$ and focus on comparing susceptibility distributions of Wolb+ inferred at different days. Mean protection conferred by *Wolbachia* in this illustration is estimated as 79% and 56%, based on mortality measurements at day 30 and 50, respectively. Moreover, the distributions have fundamentally different shapes, with the appearance of a high susceptibility group as 11 time progresses. This sensitivity to the day by which mortality data are collected is a concern that raises the need to disentangle infection status from the associated time-dependent mortality. In the following sections, infection and mortality are estimated explicitly using the integrated time-dependent model described in Methods. The procedure is illustrated in Figure 3.

**Table 1.3.2:** Estimated parameters by applying dose-response models to selected day mortality.



| Mortality data | Parameter | Median | 95% HPD[a] |
|---|---|---|---|
| 30 dpc[b] | $p$ | $2.33\ 10^{-6}$ | $[1.67\ 10^{-6}, 3.13\ 10^{-6}]$ |
| | $a$ | 0.30 | [0.21, 0.41] |
| | $b$ | 1.10 | [0.29, 2.53] |
| | $\varepsilon$ | $1.78\ 10^{-3}$ | $[4.90\ 10^{-4}, 3.49\ 10^{-3}]$ |
| 50 dpc[b] | $p$ | $2.65\ 10^{-6}$ | $[1.82\ 10^{-6}, 3.47\ 10^{-6}]$ |
| | $a$ | 0.34 | [0.24, 0.51] |
| | $b$ | 0.42 | [0.12, 0.93] |
| | $\varepsilon$ | $1.83\ 10^{-3}$ | $[3.60\ 10^{-4}, 3.32\ 10^{-3}]$ |

[a] High posterior density interval.

[b] Days post-challenge.

doi:10.1371/journal.pcbi.1003773.t002

### 1.3.2 Aging and background mortality

Control curves from Wolb⁻ and Wolb+ flies pricked with buffer solution (no *DCV*) were compared with the Kaplan-Meier method using the log-rank test and no significant difference was found (with a p-value of 0.47). By fitting the uninfected time-dependent model (4-6) to the control survival curves (Figure 1) we estimated the parameters describing aging ($m_0, s_0$) and background ($\kappa$) mortality (Table 3).

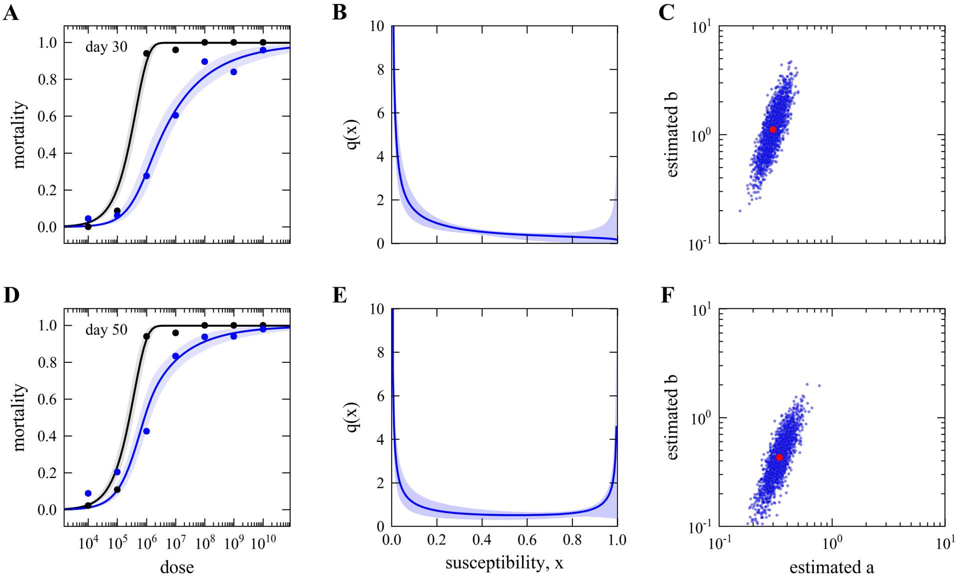


**Figure 1.3.1: Figure 2. Dose-response curves and susceptibility distributions inferred from mortality measurements 30 and 50 days post- challenge.** Dose-responses models adopted here are the standard formulations (1–3). A,D, Curves represent the fitted dose-response model to mortality on selected day post-challenge (dots), for Wolb⁻ (black) and Wolb$^+$ (blue). Shaded areas represent the 95% CI. B,E, Distribution of susceptibility to infection in Wolb$^+$. The posterior median distribution is the curve and the shaded area is the 95% CI. C,F, Posterior samples of the Beta-distribution shape parameters describing Wolb$^+$ susceptibility in blue. Red dot mark the median of the respective distributions. The homogeneous model was adopted for Wolb⁻.

### 1.3.3 Susceptibility distribution from survival curves

For each group of flies (Wolb$^-$ and Wolb+), the time-dependent dose-response model constructed in Methods was fitted simultaneously to the entire dataset of survival curves (one for each *DCV* challenge dose), fixing across doses the distribution of times to death from infection $(m_I, s_I)$ and aging $(m_0, s_0)$, while estimating the susceptibility parameters $(p, a, b)$ that govern the dependence of response on challenge dose according to the adopted dose-response model. The estimated parameter values are listed in Table 4. The deviance information criterion (DIC) [27] favored the homogeneous model for the Wolb$^-$ group and the heterogeneous model for Wolb+ (Text S1). Mean time to death from infection is 9 and 14 days in the Wolb$^-$ and Wolb+ groups, respectively. The variance in time to death from infection is lower for Wolb$^-$, with a standard deviation of 2 days, compared to 6 days in the Wolb+. Figure 4 compares fitted with observed survival curves.

The fitted dose-response curves that result from this analysis are shown in Figure 5A, while the inferred distribution of Wolb+ susceptibilities normalized by the Wolb$^-$ measure is displayed in Figure 5B and the corresponding posterior distribution of the Beta shape parameters is in Figure 5C. Given the homogeneity in the Wolb$^-$ group, the distribution of susceptibility in Wolb+ provides a direct indication of how antiviral protection conferred by *Wolbachia* is distributed among its carriers. Typically defined as 1−RR , where RR is the risk reduction attributed to the susceptibility modifier (*Wolbachia* in this case), we determine the mean protection conferred by the symbiont to its host as 85% (with a 95% HPD of 60-93%).

### 1.3.4 Comparison with selected day mortality

To assess the best possible performance of classic methods [5] in the inference of susceptibility distributions (for Wolb+ in the case) we must have previously reduced the set of survival curves to a set of effectively infected proportions - one entry per challenge dose. To search for a range of days in which absolute mortality might provide an approximate indication of infection, we compare the estimated proportions effectively infected by each challenge dose with the mortality proportion measured at each day. Using a normalized

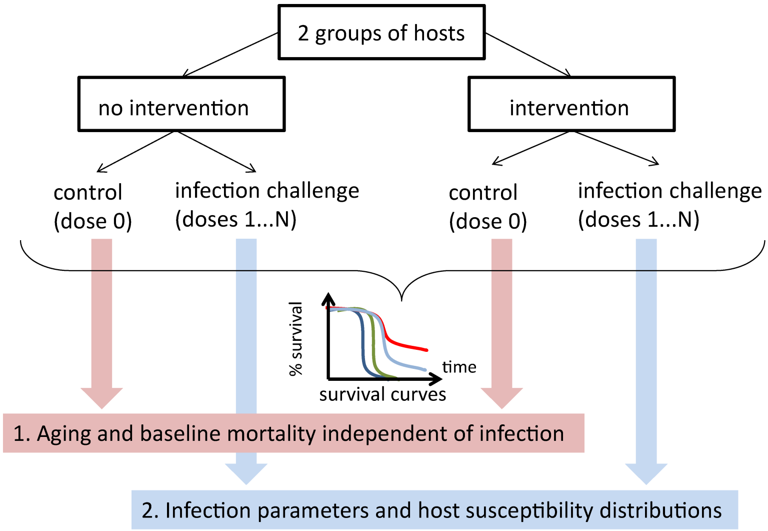


**Figure 1.3.2: Figure 3. Schematic illustration of the proposed experimental design and inference procedure.**

Euclidean distance between these two measures, a day- selection score is provided by the red curve in Figure 6. We identify day 30 as optimal and 17-46 as the interval of days in which the score is at least 95% of the optimal. Reassuringly, the optimal day appears to coincide with the saturation of infection-induced mortality (see position of vertical dash-dotted gray line in relation to the Gamma distributions). We now recall Figure 2 and Table 2 for the inferences based on day 30 mortality data to confirm that classic dose-response models can in principle infer susceptibility distributions that are consistent with those obtained under our extended model (Figure 5). A major issue, however, is that results are sensitive to a day-selection criterion that relies on having previously carried out the entire procedure. The appearance of a high susceptibility group in distributions inferred at later days are an artifact due to the accumulation of background mortality that should be factored out. These results highlight the importance of adequately representing the time

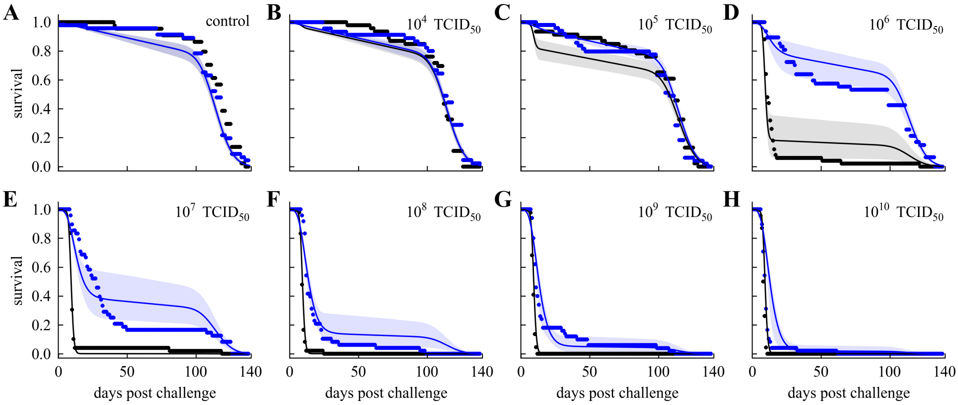


**Figure 1.3.3: Fit of time-dependent dose-response model to survival curves.** Black and blue dots are the observed proportions surviving over time for Wolb$^-$ and Wolb$^+$ groups, respectively. The curve is the fitted mean posterior survival over time and the shaded area is the 95% CI. Fifty flies per group were pricked with: A, buffer solution (shown for comparison but not used on this analysis); and B, $10^4$; C, $10^5$; D, $10^6$; E, $10^7$; F, $10^8$; G, $10^9$; H, $10^1 0$ TCID$_{50}$ DCV.

dimension in the analysis.

## 1.4 Discussion

Dose-response models have become standard quantitative frameworks in microbial risk assessment. Less recognized is their ability to estimate host trait distributions. Here we illustrate the concept by extracting host susceptibility distributions from mortality measured as a function of pathogen challenge dose, but similar procedures can be developed for measures of infection or infectiousness (instead of mortality), and can be made a function of other environmental variables such as temperature or humidity (instead of dose). Understanding how to detach host trait distributions from environmental variables is crucial for the formulation of measures that can be transported between laboratory and natural conditions [28].

We address this problem with an experimental design and inference framework that enables the estimation of distributions of host susceptibility to infection by analyzing simul-

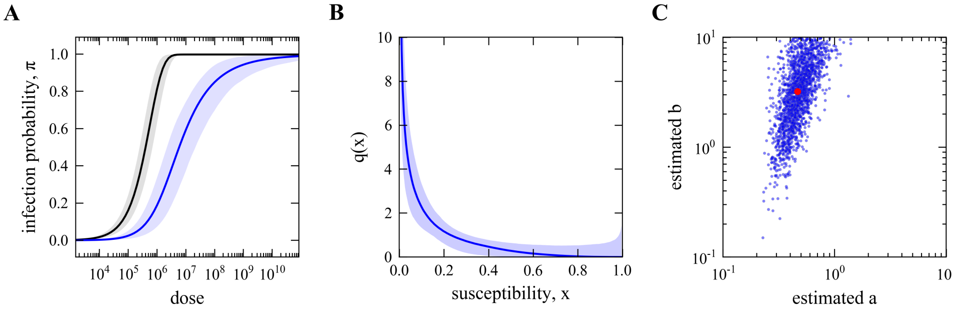
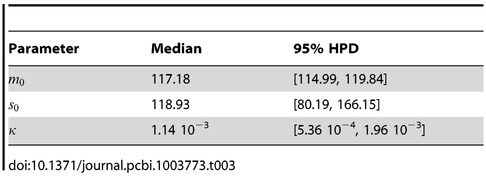


**Figure 1.3.4: Figure 5. Dose-response curves and susceptibility distributions inferred from survival curves**. A, Curves represent the estimated dose- response relationships from fitting the model described in Methods to survival over time, for Wolb⁻ (black) and Wolb$^+$ (blue). Shaded areas represent the 95% CI. B, Distribution of susceptibility to infection in Wolb$^+$. The posterior median distribution is the curve and the shaded area is the 95% CI. C, Posterior samples of the Beta-distribution shape parameters describing Wolb$^+$ susceptibility in blue. Red dot marks the median of distribution.

**Table 1.3.3:** Estimated parameters governing time to death from causes other than DCV infection.



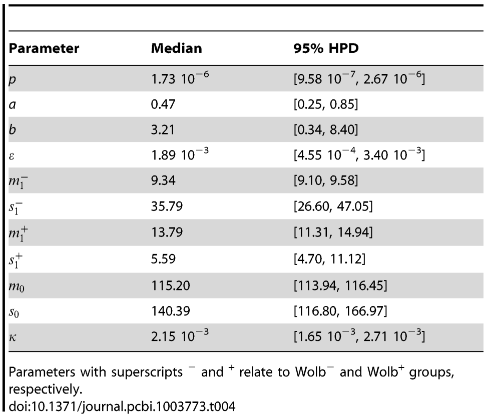
| Parameter | Median | 95% HPD |
|---|---|---|
| $m_0$ | 117.18 | [114.99, 119.84] |
| $s_0$ | 118.93 | [80.19, 166.15] |
| $\kappa$ | 1.14 $10^{-3}$ | [5.36 $10^{-4}$, 1.96 $10^{-3}$] |

doi:10.1371/journal.pcbi.1003773.t003

taneously a collection of survival curves, each representing a different challenge dose (Figure 3). The procedure is illustrated on a specifically collected dataset where two distinct groups of hosts (*D. melanogaster*) were experimentally challenged by viruses (*DCV*): one group consists of isogenic flies where no significant variability in susceptibility to infection is found; and another with the same genetic background but now carrying the symbiont bacterium *Wolbachia* known to reduce susceptibility to *DCV* [21;22].

Our inferences indicate that *Wolbachia* confers on average 85% *DCV* protection to *D. melanogaster* under the specified laboratory conditions, and suggest significant variability in this effect. This variance in susceptibility is induced by the symbiont, since model selec-

**Table 1.3.4:** Parameters governing estimated number infected per dose of DCV challenge and time to death from infection using time-dependent dose-response models described in Methods.



| Parameter | Median | 95% HPD |
|---|---|---|
| $p$ | $1.73\ 10^{-6}$ | $[9.58\ 10^{-7}, 2.67\ 10^{-6}]$ |
| $a$ | 0.47 | [0.25, 0.85] |
| $b$ | 3.21 | [0.34, 8.40] |
| $\varepsilon$ | $1.89\ 10^{-3}$ | $[4.55\ 10^{-4}, 3.40\ 10^{-3}]$ |
| $m_1^-$ | 9.34 | [9.10, 9.58] |
| $s_1^-$ | 35.79 | [26.60, 47.05] |
| $m_1^+$ | 13.79 | [11.31, 14.94] |
| $s_1^+$ | 5.59 | [4.70, 11.12] |
| $m_0$ | 115.20 | [113.94, 116.45] |
| $s_0$ | 140.39 | [116.80, 166.97] |
| $\kappa$ | $2.15\ 10^{-3}$ | $[1.65\ 10^{-3}, 2.71\ 10^{-3}]$ |

Parameters with superscripts $^-$ and $^+$ relate to Wolb$^-$ and Wolb$^+$ groups, respectively.

doi:10.1371/journal.pcbi.1003773.t004

tion criteria did not support heterogeneity in the susceptibility of flies not carrying *Wolbachia*. Since the *Drosophila* and *Wolbachia* populations used in this study are isogenic, the heterogeneity in susceptibility of *Wolbachia*-carrying flies uncovered here indicates variation in the host-microorganism interaction that lacks a genetic basis. A simple hypothesis is that variance in *Wolbachia* levels at the individual host level leads to variance in resistance to viruses. Although several lines of evidence support this hypothesis [29-32], further experiments are required to discriminate whether heterogeneity in resistance is directly linked to variance in *Wolbachia* levels or, alternatively, a result of another environmental/physiological variance that is only expressed in the presence of *Wolbachia*.

Previous estimates of protection were based on survival analysis or viral titres in a dose-specific manner [21;22;24]. To our knowledge, the experimental design and analysis presented here provides the first estimation of protection in way that is detached from challenge dose. Future developments might consider: estimation of alternative distributions to compare with the shapes suggested by the Beta family; extension of the adopted experimental design to measure responses other that mortality; and move towards host populations and environmental conditions that are closer to natural systems.

The parameters estimated here should not be seen as isolated from the relevant 15 ecological context. On the contrary, they are intended as a first step to inform the construction of ecological and epidemiological models where *Wolbachia*, other symbionts, or interventions that modify host susceptibility to infection, are introduced to induce desired transitions in populations. The introduction of *Wolbachia* into *Aedes aegypti* and other arthropod vectors is being considered as a promising strategy to control dengue and other infectious diseases of humans (see [33] and references therein). The inference frameworks presented can be readily adapted to provide accurate quantification of *Wolbachia*-induced protection and integrated in population models of public health importance.

The challenge of considering the time dependence of processes leading to observable ecotoxicity responses has also been addressed in toxicology where the so-called General Unified Model of Survival (GUTS) has been proposed [18]. These models simulate the time-course of external and internal processes leading to toxic effects on organisms to generate an output that can be fitted to mortality over time. While those studies tend prioritize the mechanistic descriptions of the toxicokinetic and toxicodynamic processes that damage the organisms, we have chosen to adopt a phenomenological approach and focus on the inference and interpretation of how susceptibility to infection is distributed in a population.

In epidemiological systems, the baseline transmission intensity is often not directly measurable but indirectly inferred in a model-based manner. Dose-response models, on the other hand, can account for experimentally controlled patterns of exposure [34;35]. Variation in host susceptibility to pathogens is one component of both classes of systems that mostly influences estimates of intervention impacts [28]. Therefore, building on the methods developed here furthers our potential to accurately evaluate the burden of infectious diseases and design effective interventions.

## 1.5 Acknowledgements

We thank Luis Teixeira for help with the experimental work, and Jorge Carneiro for valuable discussions concerning statistical models.

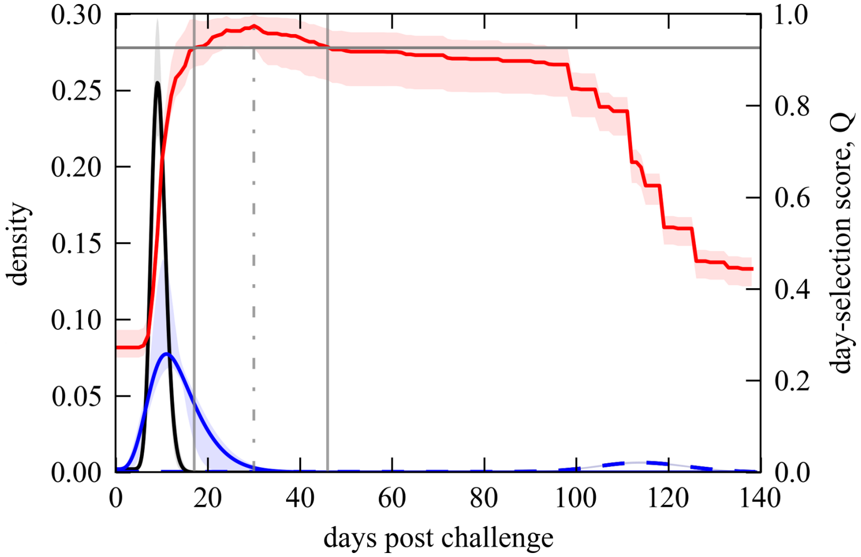


**Figure 1.4.1: Figure 6. Selection of optimal days to collect mortality measurements for traditional dose-response models.** The red line traces a score for how well mortality at any given day represents infection estimated by the time-dependent model (referto axis on the right). The score is given by $Q = 1 - \sqrt{[\sum_{j=1}^{\Delta}(y_j^- - x_j^-)^2 + (y_j^+ - x_j^+)^2]/2\Delta}$ where $\Delta$ denotes the number of doses in the dataset, $x_j - (x_j^+)$ represents the proportion infected in the Wolb- (Wolb$^+$) group subject to DCV dose $j$, and $y_j^-(t)(y_j^+(t))$ the observed mortality proportion over time in the Wolb- (Wolb$^+$) group subject to DCV dose j. Gray vertical lines mark the optimal day to measure mortality for dose-response models (day 30, dash-dotted line) and the limits of the acceptable range (days 17 and 46). Dashed lines represent the Gamma distributions that describe old-age mortality, and black (blue) full curves refer to the Gamma distributions that describe infection-induced mortality in Wolb- (Wolb$^+$) (refer to axis on the left). Curves are the mean posterior probabilities and shaded areas represent the

# References

1. Holmes, F. O. 1929 Local lesions in tobacco mosaic. Botanical Gazette 87, 39–55. (doi:10.1086/333923)

2. Henry, J. L., Jaikaran, E. S., Davies, J. R., Tomlinson, A. J., Mason, P. J., Barnes, J. M., Beale, A. J. 1966 A study of poliovaccination in infancy: excretion following challenge with live virus by children given killed or living poliovaccine. The Journal of Hygiene 64, 105–120.

3. Hornick, R. B., Music, S. I., Wenzel, R., Cash, R., Libonati, J. P., Snyder, M. J., Woodward, T. E. 1971 The Broad Street pump revisited: response of volunteers to ingested cholera vibrios. Bulletin of the New York Academy of Medicine 47, 1181–1191.

4. Ferreira, D. M. et al. 2013 Controlled Human Infection and Rechallenge with Streptococcus pneumoniae Reveals the Protective Efficacy of Carriage in Healthy Adults. American Journal of Respiratory and Critical Care Medicine 187, 855–864. (doi:10.1164/rccm.201212-2277OC)

5. Haas CN, Rose JB, Gerba CP (1999) Quantitative Microbial Risk Assessment. New York: John Wiley & Sons, Inc.

6. Druett, H. A. 1952 Bacterial Invasion. Nature 170, 288–288. (doi:10.1038/170288a0)

7. Teunis, P. F., Havelaar, A. H. 2000 The Beta Poisson dose-response model is not a single-hit model. Risk Anal. 20, 513–520.

8. Kleczkowski, A. 1950 Interpreting relationships between the concentrations of plant viruses and numbers of local lesions. J. Gen. Microbiol. 4, 53–69. (doi:10.1099/00221287-4-1-53)

9. Furumoto, W. A., Mickey, R. 1967 A mathematical model for the infectivity-dilution curve of tobacco mosaic virus: theoretical considerations. Virology 32, 224–233. (doi:10.1016/0042-6822(67)90272-3)

10. Furumoto, W. A., Mickey, R. 1967 A mathematical model for the infectivity-dilution curve of tobacco mosaic virus: Experimental tests. Virology 32, 216–223. (doi:10.1016/0042-6822(67)90271-1)

11. Therneau TM, Grambsch PM (2000) Modeling Survival Data: Extending the Cox Model, Springer-Verlag.

12. Hougaard, P. 1986 A class of multivariate failure time distributions. Biometrika 73, 671–678. (doi:10.1093/biomet/73.3.671)

13. Aalen OO (1988) Heterogeneity in survival analysis. Statistics in Medicine 7: 1121–1137.

14. Halloran, M. E., Longini, I. M., Struchiner, C. J. 1996 Estimability and Interpretation of Vaccine Efficacy Using Frailty Mixing Models. American Journal of Epidemiology 144, 83–97. (doi:10.1093/oxfordjournals.aje.a008858)

15. Longini, I. M., Halloran, M. E. 1996 A Frailty Mixture Model for Estimating Vaccine Efficacy. Applied Statistics 45, 165. (doi:10.2307/2986152)

16. Ben-Ami, F., Regoes, R. R., Ebert, D. 2008 A quantitative test of the relationship between parasite dose and infection probability across different host-parasite combinations. Proceedings of the Royal Society B: Biological Sciences 275, 853–859. (doi:10.1098/rspb.2007.1544)

17. Baas, J., Jager, T., Kooijman, B. 2010 Understanding toxicity as processes in time. Science of The Total Environment 408, 3735–3739. (doi:10.1016/j.scitotenv.2009.10.066)

18. Jager, T., Albert, C., Preuss, T. G., Ashauer, R. 2011 General Unified Threshold Model of Survival - a Toxicokinetic-Toxicodynamic Framework for Ecotoxicology. Environ. Sci. Technol. 45, 2529–2540. (doi:10.1021/es103092a)

19. Huang, Y., Haas, C. N. 2009 Time-Dose-Response Models for Microbial Risk Assessment. Risk Analysis 29, 648–661. (doi:10.1111/j.1539-6924.2008.01195.x)

20. Toth, D. J. A. et al. 2013 Quantitative Models of the Dose-Response and Time Course of Inhalational Anthrax in Humans. PLoS Pathog. 9, e1003555. (doi:10.1371/journal.ppat.1003555)

21. Teixeira, L., Ferreira, Á., Ashburner, M. 2008 The Bacterial Symbiont Wolbachia Induces Resistance to RNA Viral Infections in Drosophila melanogaster. PLOS Biology 6, e1000002–11. (doi:10.1371/journal.pbio.1000002)

22. Hedges, L. M., Brownlie, J. C., O'Neill, S. L. 2008 Wolbachia and virus protection in insects. Science (doi:10.1126/science.1125694))

23. Ryder E, Blows F, Ashburner M, Bautista-Llacer R, Coulson D, et al 2004 The DrosDel collection: a set of P-element insertions for generating custom chromosomal aberrations in Drosophila melanogaster. Genetics 167, 797–813.

24. Chrostek, E., Marialva, M. S. P., Esteves, S. S., Weinert, L. A., Martinez, J., Jiggins, F. M., Teixeira, L. 2013 Wolbachia Variants Induce Differential Protection to Viruses in Drosophila melanogaster: A Phenotypic and Phylogenomic Analysis. PLoS genetics 9, e1003896–22. (doi:10.1371/journal.pgen.1003896)

25. van der Werf, W., Hemerik, L., Vlak, J. M., Zwart, M. P. 2011 Heterogeneous Host Susceptibility Enhances Prevalence of Mixed-Genotype Micro-Parasite Infections. PLoS Comput Biol 7, e1002097. (doi:10.1371/journal.pcbi.1002097)

26. Patil, A., Huard, D., Fonnesbeck, C. J. 2010 PyMC: Bayesian Stochastic Modelling in Python. Journal of statistical software 35, 1–81.

27. Pessoa D (2014) DISE Dose-Invariant Susceptibility Estimator. Database: Gitgub. https://github.com/dpessoaIGC/Dose-Invariant-Susceptibility- Estimator.

28. Spiegelhalter, D. J., Best, N. G., Carlin, B. P., Van Der Linde, A. 2002 Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 64, 583–639. (doi:10.1111/1467-9868.00353)

29. Gomes, M. G. M., Lipsitch, M., Wargo, A. R., Kurath, G., Rebelo, C., Medley, G. F., Coutinho, A. 2014 A Missing Dimension in Measures of Vaccination Impacts. PLoS Pathog. 10, e1003849. (doi:10.1371/journal.ppat.1003849)

30. Souto-Maior, C., Lopes, J. S., Gjini, E., Struchiner, C. J., Teixeira, L., M Gomes, M. G. 2015 Heterogeneity in symbiotic effects facilitates Wolbachia establishment in insect populations. Theoretical Ecology 8, 53–65. (doi:10.1007/s12080-014-0235-7)

31. Osborne, S. E., Leong, Y. S., O'Neill, S. L., Johnson, K. N. 2009 Variation in Antiviral Protection Mediated by Different Wolbachia Strains in Drosophila simulans. PLoS Pathog. 5, e1000656–9. (doi:10.1371/journal.ppat.1000656)

32. Frentiu, F. D., Robinson, J., Young, P. R., McGraw, E. A., O'Neill, S. L. 2010 Wolbachia -Mediated Resistance to Dengue Virus Infection and Death at the Cellular Level. Plos One 5, e13398. (doi:10.1371/journal.pone.0013398)

33. Lu, P., Bian, G., Pan, X., Xi, Z. 2012 Wolbachia Induces Density-Dependent Inhibition to Dengue Virus in Mosquito Cells. PLoS neglected tropical diseases 6, e1754. (doi:10.1371/journal.pntd.0001754)

34. Osborne, S. E., Iturbe-Ormaetxe, I. 2012 Antiviral protection and the importance of Wolbachia density and tissue tropism in Drosophila simulans. Applied and …

35. Iturbe-Ormaetxe, I., Walker, T., O' Neill, S. L. 2011 Wolbachia and the biological control of mosquito-borne disease. Nature Publishing Group 12, 508–518. (doi:10.1038/embor.2011.84)

36. Pujol, J. M., Eisenberg, J. E., Haas, C. N., Koopman, J. S. 2009 The Effect of Ongoing Exposure Dynamics in Dose Response Relationships. PLoS Comput Biol 5, e1000399–12. (doi:10.1371/journal.pcbi.1000399)

37. Mayer, B. T., Koopman, J. S., Ionides, E. L., Pujol, J. M., Eisenberg, J. N. S. 2011 A dynamic dose-response model to account for exposure patterns in risk assessment: a case study in inhalation anthrax. Journal of the Royal Society, Interface / the Royal Society 8, 506–517. (doi:10.1098/rsif.2010.0491)

*This [solving mathematical models of Wolbachia spread] is as much fun as you can have with your clothes on.**

Michael Turelli

# 2

# Heterogeneity in symbiotic effects facilitates Wolbachia establishment in insect populations

Souto-Maior, C., Lopes, J. S., Gjini, E., Struchiner, C. J., Teixeira, L., & M Gomes, M. G. 2015 Theoretical Ecology 8, 53–65.

## Author contributions

The supervisor of the thesis, M. Gabriela M. Gomes, conceived the work. The main models were developed and implemented by the author of the thesis, Caetano Souto-Maior; the continuous-distribution model was implemented by João Solari Lopes. Early implementations and analyses were performed by the author of the thesis, Caetano Souto-Maior, by Erida Gjini, and by João S. Lopes. All shown simulations and analyses in the main text were performed by the author of the thesis, Caetano Souto-Maior. The stability analysis appendix was written by the supervisor of the thesis, M. Gabriela M. Gomes. The manuscript was written by the author of the thesis, Caetano Souto-Maior, the supervisor of the thesis, M. Gabriela M. Gomes, by Erida Gjini, Joao S. Lopes, Luis Teixeira, and Claudio J. Struchiner.

ABSTRACT

Facultative vertically transmitted bacterial symbionts often manipulate host reproductive biology to facilitate their persistence. *Wolbachia* is one such symbiont where frequency-dependent reproductive benefits are opposed by frequency-independent fitness costs leading to bistable dynamics. Introduction of carriers does not assure invasion unless the initial frequency is above a threshold determined by the balance of costs and benefits. Recent laboratory experiments have uncovered that *Wolbachia* also protects their hosts from horizontally transmitted pathogens. The expected consequence for this phenotype in natural environments is to lower the invasion threshold by a factor that increases with the extent of pathogen exposure. Here we introduce a series of mathematical models to address how resistance to pathogens affects *Wolbachia* invasion. First, under homogeneous population assumptions, we obtain an analytical expression for the invasion threshold in terms of pathogen exposure, and find a new regime where symbiont releases result in elimination of the entire host population provided pathogen abundance is high. Second, we distribute Wolbachia effects such that some carriers are totally resistant and others not at all, and explore how this manifests at different challenge intensities, to conclude that heterogeneity further lowers the threshold for Wolbachia invasion and increases system resilience by reducing the odds of population collapse.

## 2.1 Introduction

*Wolbachia* are vertically transmitted, obligatory intracellular bacteria present in a great number of species of arthropods and nematodes. In insects, they often manipulate reproduction of their hosts and thus assure persistence. One such reproductive manipulation phenotype prevents successful reproduction between male *Wolbachia*-carriers and female non-carriers due to cytoplasmic incompatibility (CI), confering a fitness advantage to the carrier population (Werren *et al.* 2008), which is stronger when *Wolbachia* is more prevalent. Such positive frequency-dependence induces a strong Allee effect when *Wolbachia* is costly to the host (Taylor and Hastings 2005), such that the resident population approaches elimination or remains nearly unchanged depending on whether the initial frequency of carriers is above or below a threshold (Caspari and Watson 1959) that can be expressed as a function of the various *Wolbachia*-associated effects (Hoffmann *et al.* 1990, 1986).

The use of *Wolbachia* to manipulate the life history of disease vectors has been proposed, and more recently *Wolbachia* has been shown to confer protection against viral infections in *Drosophila melanogaster* (Hedges *et al.* 2008; Teixeira *et al.* 2008), and against mosquito-borne pathogens of humans in Aedes and Anopheles species (Bian *et al.* 2013; Blagrove et al 2012; Moreira *et al.* 2009). Trial releases have been carried out in populations of *Ae. aegypti* in Australia (Hoffmann *et al.* 2011), as a pilot intervention for interrupting endemic dengue virus transmission (Maciel-de-Freitas *et al.* 2012; McGraw and O'Neill 2013). The perspective of successfully introducing carriers into natural mosquito populations where *Wolbachia* is absent, and the variety of strains to be chosen from, makes it of practical interest to estimate the thresholds for invasion of *Wolbachia*-carriers.

The invasion threshold has been the center of discussion concerning the spread of *Wolbachia* in both population-genetic (Turelli and Hoffmann 1991) and population-ecological models (Hancock *et al.* 2011) if symbionts are believed to be parasitic. In the case of mutualistic associations of *Wolbachia* or other symbionts there is no threshold, and they should be able to diffuse spatially and spread (Himler *et al.* 2011), with dynamics similar to that of an advantageous gene (Barton and Turelli 2001; Fisher 1937). *Wolbachia* strains inducing fitness costs, however, may face an invasion threshold that constrains spatial spread from a local introduction (Barton and Turelli 2001). In a finite population, the level of the threshold also affects the probability of fixation given a rare introduction (Jansen *et al.* 2008) to an extent that is modulated by environmental factors. In the case of strains artificially introduced into *Ae. aegypti* mosquitoes (which are not natural hosts to *Wolbachia*), for instance, fitness costs were identified in laboratory experiments and must be contemplated in the assessment of planned releases of larger populations of carriers (Yeap *et al.* 2011).

The idea that vertically-transmitted parasites may interfere with host susceptibility to pathogens, which in turn affects the prevalence and persistence of the former in a host population has been previously proposed (Jones *et al.* 2007; Lively *et al.* 2005), and extended to CI-inducing parasites, such as *Wolbachia* (Fenton *et al.* 2011), carried by many host species (Hilgenboecker *et al.* 2008). The intuitive idea is that protection against virulent pathogens would compensate for the costs of carrying the parasite, allowing its persistence. Despite this general qualitative feature, the ability of CI-inducing parasites to actually invade a resident population is determined by a threshold frequency, which depends explicitly on fitness costs and benefits (Carrington *et al.* 2011; Turelli 2010; Turelli and Hoffmann 1995), but is likely to be entangled with ecological parameters. Such threshold frequency has not been formally described, nor has the influence of host ecology been thoroughly considered. Furthermore, all ecological models of *Wolbachia* assume homogeneity in the effects of the vertically-transmitted parasite, but variation in costs or benefits can further affect these descriptions.

In this work we provide analytical results of systems where a CI-inducing parasite (hereafter, *Wolbachia*) may induce fecundity and longevity costs while partially protecting its hosts from a generalist pathogen ensemble. We derive an exact expression for the invasion threshold under these conditions, and show how it relates to the quantity predicted by population genetics. We emphasize the importance of specifying the traits affected by fitness costs, as their weights on the invasion threshold depend on interactions with ecological and environmental processes. We describe the temporal dynamics of local invasion. Finally, we consider heterogeneity in pathogen protection. For a simple all-or-nothing implementation we derive analytical model solutions, which motivate the introduction of a more realistic distribution estimated from experimental data and a discussion of the repercussions of heterogeneity on invasion and final size of the *Wolbachia*-carrying population.

## 2.2 Model Foundations

Our models build on continuous-time implementations of single-species models (Bellows 1981; Hassell *et al.* 1976) also used more recently for tripartite interactions involving a host, a vertically-transmitted parasite, and a horizontally-transmitted pathogen (Fenton *et al.* 2011; Jones *et al.* 2007). These models consider a single stage of the host, namely the adult stage, but see Crain *et al.* (2011); Hancock *et al.* (2011) for age-structured models. The frameworks expand on population-genetic models by explicitly including demographic processes, such as birth and natural mortality rates. The specification of these processes allows *Wolbachia*-mediated effects to manifest explicitly on the relevant ecological parameters, i.e. fecundity costs reduce birth rates, while longevity costs increase baseline natural mortality exclusively, enabling different forces of selection to be separately analysed. We construct and analyse a series of models with incremental attention on the implementation of host susceptibility to pathogens. Definitions of all parameters, adopted values, and references are given in Table 1.

**Table 2.2.1:** Parameters, values and references

| Definition | Value | Reference |
|---|---|---|
| $a$, Birth rate of insect host | 10 | Jones *et al.* (2007) |
| $b$, Density-independent death rate of insect host | 1 | Jones *et al.* (2007) |
| $k$, Density-dependent death constant of insect host | 0.01 | Jones *et al.* (2007) |
| $s_h$, Proportion of inviable offspring in incompatible crosses | 0.9 | Blagrove et al (2012)<br>Yeap *et al.* (2011) |
| $s_f$, Relative fecundity reduction of *Wolbachia* carriers | 0; 0.6 | Yeap *et al.* (2011)<br>Walker *et al.* (2011) |
| $s_l$, Relative lifespan reduction of *Wolbachia* carriers | 0; 0.1; 0.4 | This Study;<br>Yeap *et al.* (2011);<br>Walker *et al.* (2011) |
| $\lambda$, Infection-induced mortality | 0 – 10 | This study |
| $\sigma$, $\overline{\sigma}$, Mean relative susceptibility of *Wolbachia* carriers | 0.4; 0.21 | This study |

### 2.2.1 Total pathogen protection

In the special case where *Wolbachia* confers total protection against pathogens, the model is given by the following system of differential equations:

$$\begin{aligned}\frac{dU}{dt} &= aU\left[\frac{U+W(1-s_h)}{N}\right]-[b+f(N)]U-\lambda U,\\ \frac{dW}{dt} &= aW(1-s_f)-\left[\frac{b}{1-s_l}+f(N)\right]W,\end{aligned} \tag{2.1}$$

where $U$ is the number of females not carrying *Wolbachia*, or non-carriers, $W$ is the female *Wolbachia*-carrying population, and $N$ is the total female population, therefore $N=U+W$. The male/female ratio is assumed constant and equal to unity as in Fenton *et al.* (2011).

In the absence of pathogens, the net rate of change in the $U$ population is given by a birth term, $aU[U+W(1-s_h)]/N$, where $a$ is the birth rate and $s_h$ is the proportion of offspring killed in crosses of $W$ males with $U$ females, and a death term, $-[b+f(N)]U$, where $b$ and $f(N)$ are the density-independent and density-dependent rates, respectivelly. Similarly, the rate of change in the $W$ population is given by a birth term, $aW(1-s_f)$, where $s_f$ is a fecundity cost that affects the birth rate $a$, and a death term, $-[b/(1-s_l)+f(N)]W$, which has the same interpretation as that in the $U$ population, except that the density-independent component is modified by a life-shortening fitness cost $s_l$.

In introducing pathogen effects we do not take explicit account of pathogen-infected hosts in the dynamics, and $-\lambda U$ models the infection-induced mortality at a constant rate $\lambda$, following Schmid-Hempel (1998). As equilibria are seldom attained in natural systems, we envisage a situation where new strains or new pathogen species are constantly introduced ensuring that manipulation of the target host species has a negligible effect on the prevalence of pathogens. Notice that, mathematically, this implementation is equivalent to simply increasing the death rate of $U$ hosts by $\lambda$ and, in principle, composite parameters could be defined to describe death rates specific to $U$ and $W$ hosts so that the number of parameters would be reduced by one. To ease interpretation, however, we maintain the three-parameter formulation for the death rates. In the Supporting Text S1 we analytically solve a model where pathogen-infected hosts are explicitly represented, as in Jones *et al.* (2007) and Fenton *et al.* (2011), and show that the results agree qualitatively with those presented here.

The parameters specifically associated to *Wolbachia* so far are: the intensity of the cytoplasmic incompatibility, $s_h$, corresponding to the proportion of offspring killed in crosses of *Wolbachia*-carrying males with non-*Wolbachia*-carrying females; the fecundity cost, $s_f$ , representing the combined effects of reduced number of eggs laid by carriers and their reduced viability, that is, the reduction in progeny of the $W$ population by effects of *Wolbachia* on the fecundity of the host; and the longevity cost, $s_l$, that models any increase in the death rate that may be induced by the symbiont, that is, life-shortening effects.

This model assumes that *Wolbachia* is perfectly transmitted from mother to progeny. In Supporting Text S2 we extend the formulation and demonstrate that the results of interest are robust to small imperfections in *Wolbachia* transmission.

When $\lambda$ is zero the system is bistable for realistic parameter values and equivalent to that proposed by Fenton *et al.* (2011) in the absence of pathogens. Under pathogen-induced mortality (henceforth simply force of infection) greater than zero, that is, $\lambda > 0$ , the system retains the features displayed by the pathogen-free model, including bistability.

Besides the origin, $(U, W) = (0, 0)$, the system accommodates three non-trivial equilibria: pre-invasion, $(U_{pre}, W_{pre})$, and post-invasion, $(U_{pos}, W_{pos})$, both stable, and an intermediate saddle point, $(U_{uns}, W_{uns})$, which is unstable. Expressions from local stability analyses are provided in Appendix A and numerical illustrations in Figure 2.2.1, for the special case $f(N) = kN$. This linear function is used throughout to represent density-dependent mortality, while comments on sensitivity of the results to this choice are provided where appropriate.

Equilibrium proportions of carriers are defined as $p_i = W_i/N_i$, where $N_i = U_i + W_i$, for $i = pre, pos, uns$. In analogy to population genetics, absence of *Wolbachia* has $p_{pre} = 0$, and the intermediate, unstable equilibrium defines the invasion threshold, $p_{uns} = \hat{p}$. Above this threshold, *Wolbachia* is expected to increase in frequency and invade, so we expect $p_{pos} = 1$, although under slightly imperfect vertical transmission of *Wolbachia*, a residual $U$ population may be maintained (see Supporting Text S2), and under sufficiently high force of infection, *Wolbachia* invasion may cause elimination of the entire insect population if protection is partial ($N_{pos} = 0$, see partial-protection implementations below).

### 2.2.2 Pre-invasion equilibrium

In the absence of *Wolbachia*, the equilibrium of system (2.1) is reached with the total insect density, $N_{pre}$, given by the sum of $U_{pre} = f^{-1}(a - b - \lambda)$ and $W_{pre} = 0$; therefore, this stable equilibrium population depends on the net growth, the difference between birth and death rates, with carrying capacity of the population regulated by the inverse of the density-dependence function, $f^{-1}$, so that the total population is defined for a suitable choice of $f$ and parameters adequate to the species of interest (Bellows 1981; Hassell *et al.* 1976; Maynard-Smith and Slatkin 1973).

The pre-invasion equilibium refers to an insect population that is susceptible to pathogens and therefore decreases with the force of infection, $\lambda$, being able to persist as long as pathogen pressure remains below a critical intensity:

$$\lambda < a - b, \tag{2.2}$$

simply meaning that killing by pathogens must remain less than net growth by birth and death processes.

### 2.2.3 Post-invasion equilibrium

The system admits an alternative stable equilibrium when the non-carrier population is absent and only *Wolbachia*-carrying individuals are present, which is expected if carriers invade successfully. In this case, the total population post-invasion, $N_{pos}$, is given by the sum of $U_{pos} = 0$ and $W_{pos} = f^{-1}(a[1 - s_f] - b[1 - s_l]^{-1})$, where $a(1 - s_f)$ is the birth rate of carriers affected by fecundity cost $s_f$, and $b(1 - s_l)^{-1}$ is the death rate affected by longevity cost $s_l$; therefore, as with the non-carriers, equilibrium is a function of net growth.

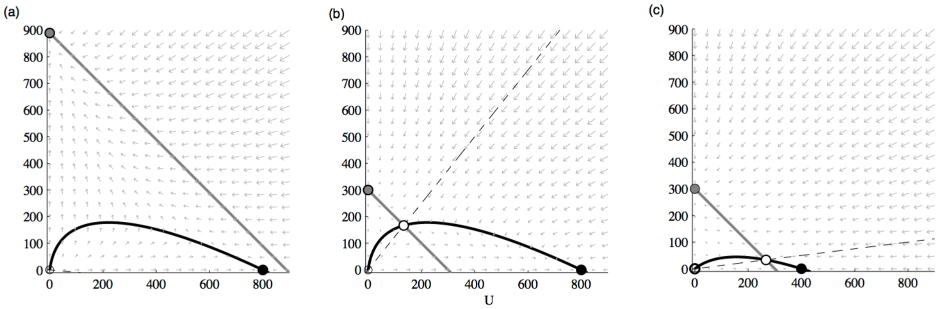


**Figure 2.2.1:** Phase plane for the total resistance model (eqs. 2.1) and the possible equilibria. Parameter values are $a = 10$, $b = 1$, $k = 0.01$, $s_h = 0.9$; and $s_f = 0$, $s_l = 0.1$, $\lambda = 1$ (a); $s_f = 0.6$, $s_l = 0$, $\lambda = 1$ (b); $s_f = 0.6$, $s_l = 0$, $\lambda = 5$ (c). Solid black lines are nullclines for the $U$ population, which may have more than one solution, and gray lines are nullclines for the $W$ population. Dashed lines are invariant for the flow and separate the basins of attraction for $U$-only and $W$-only steady states, defining the threshold for *Wolbachia* invasion.

### 2.2.4 Invasion threshold

An additional, unstable equilibrium, which we denote as $(U_{uns}, W_{uns})$ suggests the existence of a threshold initial frequency of *Wolbachia* carriers, $\hat{p} = W_{uns}/(U_{uns} + W_{uns})$, above which invasion is expected:

$$\hat{p} = \frac{s_f}{s_h} + \frac{b}{as_h}\left(\frac{s_l}{1 - s_l}\right) - \frac{\lambda}{as_h}, \tag{2.3}$$

or zero, whichever is higher. Indeed, a straightforward calculation confirms that the straight line that traverses the origin and the unstable equilibrium is invariant for the flow given in (2.1) and therefore separates the basins of attraction of the pre-invasion and post-invasion equilibria. This establishes $\hat{p}$ as the invasion threshold. See the phase planes in Figure 2.2.1.

Population genetics predicts $\hat{p} = s_f/s_h$ given the assumption of perfect *Wolbachia* transmission (Turelli and Hoffmann 1991). The same result is obtained here in the absence of horizontally transmitted pathogens ($\lambda = 0$) and longevity costs associated to *Wolbachia* ($s_l = 0$).

Expression (2.3) shows how the threshold of invasion depends on demographic parameters in the presence of longevity costs, extending previous studies, where ecological parameters are either absent (Turelli 2010) or costs apply to the net growth of the population (Hancock *et al.* 2011), and not to each process separately. By having CI and fecundity costs affect birth rates specifically, and longevity costs affect death rates only, the unstable equilibrium depends explicitly on these demographic parameters. In the case of fecundity costs, since both this and CI act by reducing births, the component of the threshold induced by this balance (first term on the right-hand side of (2.3)) is independent of host ecology; however, in the presence of longevity costs, the relevant ecological process is death rate, while CI must still act on birth rates only, so this threshold term depends on the death/birth ratio, or $b/a$, as seen in the second term. By a similar argument, the component of the threshold due to *Wolbachia* protection against pathogens depends on how important infection-induced mortality is, which appears as the $\lambda/a$ ratio.

Two aspects are worth noting from this analysis: first, the threshold of invasion is independent of the density-dependent mortality, that is, it is not necessary to specify a function $f$ to arrive at expression (2.3) as long as it affects $U$ and $W$ populations equally; and second, each *Wolbachia*-mediated effect results in an independently interpretable term in the expression for the threshold. This result holds for the model with partial protection, below, and the model including imperfect transmission (Supporting Text S2).

## 2.3 Implementations of partial protection

In the more realistic scenario where *Wolbachia* induces only partial protection against pathogens, a relaxed negative term, $-\sigma\lambda W$, is introduced in the equation for the rate of change of *Wolbachia*-carrying insects, where $\sigma$ is a factor between zero and one that represents the relative susceptibility of $W$ when compared to the $U$ population: a $\sigma$ value of zero means *Wolbachia* carriers have no susceptibility to infection, as described above, and unity means they are equally susceptible as the non-carriers.

### 2.3.1 Homogeneous model

Under the assumption that protection is homogeneous, all *Wolbachia*-carrying individuals have the same relative susceptibility to infection $\sigma$, and the model is given by the following equations:

$$\begin{aligned}\frac{dU}{dt} &= aU\left[\frac{U + W(1 - s_h)}{N}\right] - [b + kN]U - \lambda U, \\ \frac{dW}{dt} &= aW(1 - s_f) - \left[\frac{b}{1 - s_l} + kN\right] W - \sigma\lambda W.\end{aligned} \tag{2.4}$$

The introduction of the $\sigma\lambda W$ term does not qualitatively alter the equilibria of the system.

An invading *Wolbachia*-carrying population with relative susceptibility $\sigma$ is expected to reach equilibrium densities now dependent on the force of infection, such that $N_{pos}$ is the sum of $U_{pos}$ (which in this particular case is equal to zero) and $W_{pos} = k^{-1}(a[1 - s_f] - b[1 - s_l]^{-1} - \sigma\lambda)$. We therefore note that the balance between fecundity and longevity costs imposed, $s_f$ and $s_l$, and the fitness benefits *Wolbachia* confers, in terms of reduced susceptibility to pathogens, $\sigma$, will determine how the post-invasion equilibrium population size compares to the pre-invasion. In particular, for a monotonically increasing density-dependence function, such as $f(N) = kN$, when $as_f + bs_l(1 - s_l)^{-1} < (1 - \sigma)\lambda$, namely for relatively smaller costs than benefits, we expect invading *Wolbachia* carriers, $W$, to establish a larger population than the original $U$ population, while the reverse is expected otherwise.

As in the case where *Wolbachia* confers total protection from pathogens, also under partial protection there is a reduction in the invasion threshold with the force of infection, although more moderate:

$$\hat{p} = \frac{s_f}{s_h} + \frac{b}{as_h}\left(\frac{s_l}{1 - s_l}\right) - \frac{\lambda(1 - \sigma)}{as_h}, \tag{2.5}$$

or zero, whichever is higher. Figure 2.3.1 depicts the threshold and equilibrium populations of *Wolbachia*-carriers after invasion as functions of the force of infection for an implementation of partial protection ($\sigma = 0.4$), in comparison with no protection ($\sigma = 1$), and the previously described extreme of total protection ($\sigma = 0$). In the presence of protection, pathogen exposure lowers the threshold for *Wolbachia* invasion and exclusion of the resident population, establishing a post-invasion population of size $N_{pos} = k^{-1}(a[1 - s_f] - b[1 - s_l]^{-1} - \sigma\lambda)$ as long as pathogen pressure remains bound:

$$\lambda < \frac{1}{\sigma}\left[a(1 - s_f) - \frac{b}{1 - s_l}\right]. \tag{2.6}$$

Comparison with the condition for persistence of the pre-invasion population (eq. 2.2) reveals that, depending on the level of protection and costs induced by *Wolbachia*, there may be a range in the force of infection where the introduction of *Wolbachia* may eliminate a previously established insect population, with extinction of both carriers and non-carriers (range of $\lambda$ where $N_{pos} = 0$ and $N_{pre} > 0$ in Figure 2.3.1(b)). Intuitively, CI enables a population with lower fitness to replace the resident population. Once replacement occurs we are left with a population of reduced size which may even vanish if pathogen pressure is high enough. Although different choices for the density-dependence function and other assumptions modify these descriptions, the reported phenomena persist within reasonable bounds.

While population-genetic models predict low invasion thresholds if the fitness costs are small, compared to strains that impose higher fecundity and lifespan costs, and are therefore considered to be more parasitic, the added effect of protection against pathogens must be taken into account in a broader picture. More costly strains could provide an overall greater fitness if they provide enough protection against natural enemies to balance the fecundity costs. In terms of the invasion threshold this is described by $\hat{p}$ (eq. 2.5); the quantitative effect of protection against pathogens is explored by plotting the expression as a function of force of infection, $\lambda$, for a range of values, and susceptibility, $\sigma$, from complete to zero susceptibility. The resulting surfaces are shown in Figure 2.3.2 for fitness costs approximating those estimated for *w*Mel (a), and *w*MelPop-CLA (b) strains introduced into *Ae. aegypti* (Walker *et al.* 2011; Yeap *et al.* 2011). Ecological parameters are simply

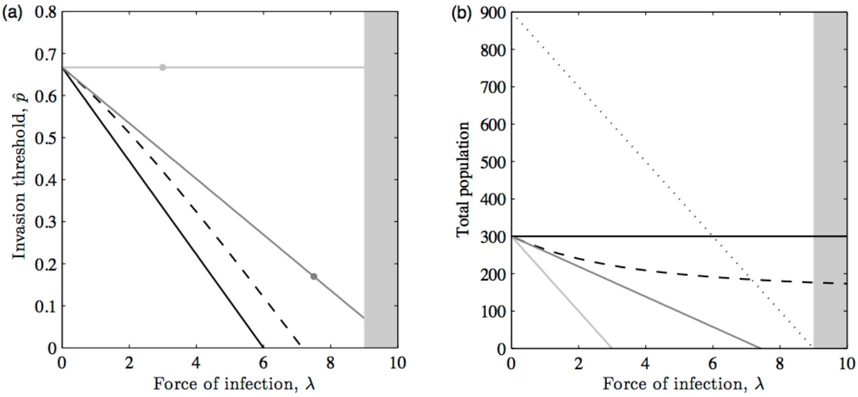


**Figure 2.3.1:** Invasion threshold under total pathogen protection (model 2.1, solid black line), as well as homogeneous (model 2.4, dark grey line), and all-or-nothing (model 2.7, dashed line) partial protection models ($\overline{\sigma} = 0.40$), and no protection (light grey line), as a function of the force of pathogen infection $\lambda$ (a). Dots mark a transition from a regime where *Wolbachia* invasion progresses to fixation (left) to another where elimination of the entire insect population occurs (right). Total insect population after *Wolbachia* invasion ($N_{pos}$) under the four models (line styles as before) and total insect population in the absence of *Wolbachia* ($N_{pre}$) (dotted line) (b). Shaded areas mark a regime where the force of infection precludes persistence of the non-carrier population. Parameter values are $a = 10$, $b = 1$, $k = 0.01$, $s_f = 0.6$, $s_l = 0$, $s_h = 0.9$.

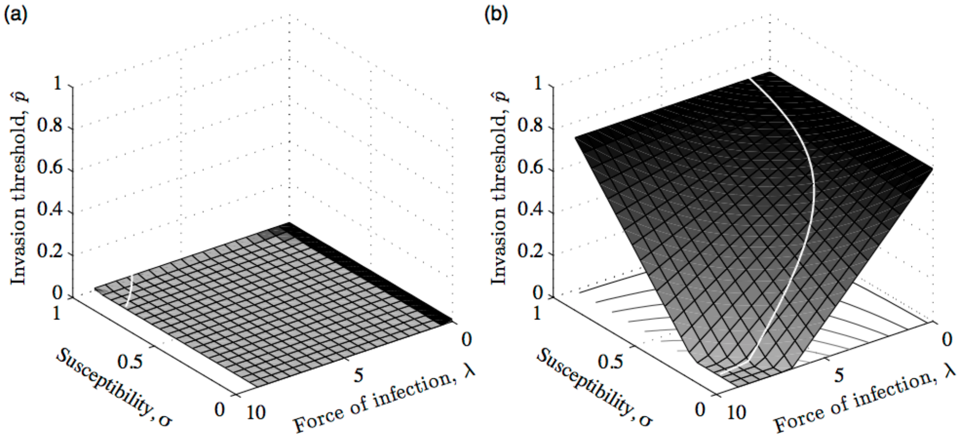


**Figure 2.3.2:** Threshold frequency for invasion by *Wolbachia*-carrying insects as a function of both force of pathogen infection and relative level of resistance under the homogeneous model (eqs. 2.4). White lines mark a transition from a regime where *Wolbachia* invasion progresses to fixation (right) to another where elimination of the entire insect population occurs (left). Parameters values are $a = 10$, $b = 1$, $k = 0.01$, $s_h = 0.9$ ; and $s_f = 0$, $s_l = 0.1$ (a); $s_f = 0.6$, $s_l = 0.4$, (b)

those of previous studies where species are not specified (Fenton *et al.* 2011; Jones *et al.* 2007), so the surfaces are not intended as quantitative predictions for *Ae. aegypti*. The white line divides the threshold surface into a region where invasion drives the *Wolbachia* population to fixation (right) and another where the entire insect population collapses as a result (left), with the elimination of the originally resident population in both cases. Noticeably, invasion by a less costly strain would be easily attainable by the introduction of a small number of *Wolbachia*-carrying individuals, $\hat{p}$ being zero for most of the parameter space. In the case of a strain associated with high fecundity costs and life-shortening, the threshold is high in the absence of pathogens or protection, as expected, however this is considerably lowered by a combination of these two factors. Indeed, there is a region of low susceptibility and high force of infection where the threshold is removed ($\hat{p} = 0$).

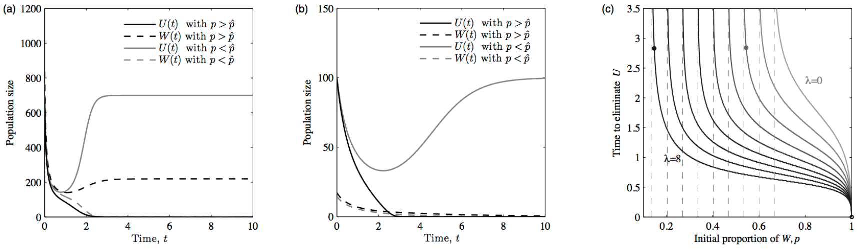


**Figure 2.3.3:** Simulations (model 2.4) illustrating the threshold phenomenon and expected times to invasion of a non-carrier population in equilibrium. A population of *Wolbachia* carriers (grey) attempts invasion of a non-carrier population (black) with initial frequency slightly below (solid lines) or above (dashed lines) the threshold for the elimination of non-carriers. The force of infection $\lambda = 2$ conducts carriers to fixation when $p > \hat{p} \approx 0.54$ (a); $\lambda = 8$ results in elimination of the entire insect population when $p > \hat{p} \approx 0.15$ (b). (c) Time elapsed from time of introduction of *Wolbachia* carriers until elimination of non-carriers for a range of initial frequency of carriers $p$, and a set of values for the force of pathogen infection $\lambda$ equal to 0, 1, 2, 3, 4, 5, 6, 7 and 8. Dots in (c) mark the initial frequencies relative to scenarios in (a) and (b). Parameter values are as in Figure 1.

### 2.3.2 Dynamics of *Wolbachia* invasion

The analytical threshold $\hat{p}$ tells us whether *Wolbachia*-carriers at a given initial frequency invade and eliminate a resident population of non-carriers; it does not, however, provide direct information on the timescale to reach post-invasion equilibrium. We therefore proceed to use simulations, which consist of numerical integration of the system of differential equations, to describe temporal features of the system.

Figure 2.3.3 shows simulations of model (2.4). Panels (a) and (b) illustrate scenarios where invasion leads to *Wolbachia* fixation ($\lambda = 2$) and collapse of the entire insect population ($\lambda = 8$), respectively. Because simulations were performed assuming $b = 1$ for the density-independent component of the death rate, the time unit can be interpreted as the average generation time. Panel 3(c) provides a summary of expected times to elimination of the resident, in terms of the initial proportion of *Wolbachia*-carriers, letting the force of infection take a set of values between 0 and 8. Dynamics are modeled deterministically on real variables, and we set as a criterion that elimination occurs when the population (in this case $U$) falls below 0.5. Each curve describes a trend from instantaneous (when $p$ approaches 1) to infinite (when $p$ asymptotically approaches the threshold $\hat{p}$) time for elimination of the $U$ population.

### 2.3.3 Heterogeneity in *Wolbachia* effects

#### All-or-nothing protection

Variation in *Wolbachia* effects has been found in association with genetic factors (Chrostek *et al.* 2013; Osborne *et al.* 2009), although environmental and developmental factors are also likely to be involved. To perform a theoretical exploration of heterogeneity we assume that protection is distributed in an all-or-nothing manner, adopting terminology from vaccine studies (Smith *et al.* 1984), resulting in a mean factor $\overline{\sigma}$ of 0.4, as in the homogeneous case, but with coefficient of variation of 1.22.

We compare the outputs of a model with susceptibility factors distributed among *Wolbachia*-carriers according to such an all-or-nothing mode to those of the homogeneous model with the same mean susceptibility factor, where the coefficient of variation is zero by construction. A fraction $1-\overline{\sigma}$ is born into the $W_0$ subpopulation and has zero susceptibility, while a proportion $\overline{\sigma}$ is born into $W_1$ and has susceptibility equal to the non-carriers. We still refer to $W$ as the total *Wolbachia*-carrying population, so that $W = W_0 + W_1$. Susceptibility is assumed non-heritable, so upon reproduction the compound $W$ population contributes to both susceptibility subgroups. The system is written as:

$$\begin{aligned}
\frac{dU}{dt} &= aU\left[\frac{U+W(1-s_h)}{N}\right] - (b+kN)U - \lambda U, \\
\frac{dW_1}{dt} &= \overline{\sigma}aW(1-s_f) - \left(\frac{b}{1-s_l}+kN\right)W_1 - \lambda W_1, \qquad (2.7)\\
\frac{dW_0}{dt} &= (1-\overline{\sigma})aW(1-s_f) - \left(\frac{b}{1-s_l}+kN\right)W_0.
\end{aligned}$$

The equation for the non-carrier $U$ population is unchanged, and so are the results in the absence of *Wolbachia*. The limit $\overline{\sigma} = 0$ restores the total protection model just as in the homogeneous system (2.4). Figure 2.3.1(a) also shows, by numerical simulation of system (2.7), that heterogeneity (dashed line) lowers the threshold of invasion in comparison with the homogeneous implementation given the same mean effect on susceptibility, $\overline{\sigma}$ of 0.4, to an degree that increases with the force of infection $\lambda$.

The justification for this facilitated invasion is selection for increased protection, even if susceptibility is not inherited from the parents - this phenomenon has been termed "survivor cohort effect", or "depletion of susceptibles" (O'Hagan *et al.* 2012; Vaupel and Yashin 1985). Susceptible individuals die younger than resistant ones, increasing protection at the population level and consequently increasing their population size relative to the homogeneous case (see Figure 2.3.4) – i.e. in a cohort of same age, resistant individuals accumulate and are over-represented in individuals of the same age (Figure 2.3.4(a)). This build-up of protection also results in larger post-invasion populations (Figure 2.3.1(b), dashed line), and may preclude the scenario of population collapse found under homogeneous protection, by allowing *Wolbachia*-carriers to persist under higher force of infection. Similar analyses could be performed for a heterogeneous distribution of longevity costs (not shown, but it is straightforward to rewrite (2.7) to accommodate a distribution of $s_l$ values, as for susceptibility). A heterogeneous population loses its weaker individuals earlier so that, as a generation ages, death rates decrease in relation to those expected under homogeneity. Heterogeneous fecundity costs, on the other hand, do not influence the threshold as in this case there is no selection for individuals of higher fecundity – traits are not inherited, and more fecund individuals are not overrepresented in later generations, despite variation being present – and unlike the previous *Wolbachia*-associated effects, all groups die at the same rate. Therefore, mean fecundity will not deviate from the initial average.

In sum, only a distribution of trait values that is subject to selection can act to increase fitness at the population level; traits that are not selected keep their initial averages values. Kendall *et al.* (2011) obtain similar results when considering the impact of survival and fe-

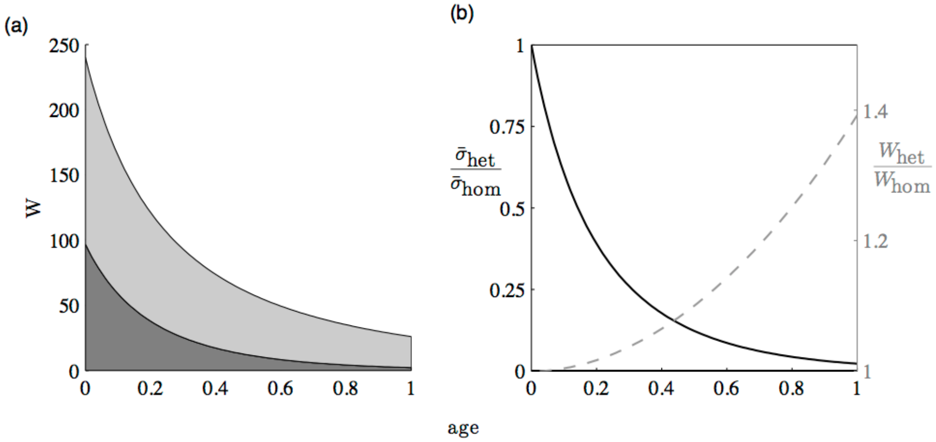


**Figure 2.3.4:** Numbers of individuals in a cohort (model 2.7 with no births, or $a = 0$) with both the resistant population ($W_0$, light grey shading) and the faster-decreasing susceptible population ($W_1$, dark grey) (a). Susceptibility ratio of the heterogeneous *Wolbachia* compared to a homogeneous carrier population ($\bar{\sigma}_{het}/\bar{\sigma}_{hom}$) (solid curve referring to left axis), and total cohort size of heterogeneous *Wolbachia* carriers relative to the homogeneous implementation ($W_{het}/W_{hom}$) (dashed curve referring to right axis) (b). The homogeneous implementation assumes $\sigma = 0.40$, with the heterogeneous all-or-none distribution having the same mean. $\lambda = 2$ in this simulation; other parameters are as in Figure 1.

cundity in population growth. One way for selection to act on fecundity could be through trade-offs with protection, although this is beyond the scope of this paper.

### Continuous distribution of protection

The homogeneous and all-or-nothing distributions of pathogen protection are two idealized extremes. For more realism, a continuous beta distribution of protective effects conferred by *Wolbachia* against *Drosophila* C virus has been estimated for an isogenic *D. melanogaster* host population (Pessoa *et al.* 2014). Following (Gomes *et al.* 2014), a continuous distribution of *Wolbachia* effects is introduced by rewriting the model as:

$$\begin{aligned}\frac{dU}{dt} &= aU\left[\frac{U+W(1-s_h)}{N}\right]-(b+kN)U-\lambda U,\\ \frac{dw(x)}{dt} &= q(x)aW(1-s_f)-\left(\frac{b}{1-s_l}+kN\right)w(x)-x\lambda w(x).\end{aligned} \tag{2.8}$$

where $q(x)$ is a probability density function describing *Wolbachia*-induced protection, and $W = \int_0^1 w(x)dx$ continues to represent the total *Wolbachia*-carrying population. The equation for the non-carrier $U$ population is unchanged, and so are the results in the absence of the symbiont. The limit $\overline{\sigma} = \int_0^1 xq(x)dx = 0$ restores the total protection model just as in the homogeneous system (2.4). Following Pessoa *et al.* (2014), we implement $q(x)$ as a beta distribution with shape parameters $\alpha = 0.30$ and $\beta = 1.10$. As model (2.8) is less amenable to mathematical treatment, the invasion threshold and the population after invasion were obtained numerically. Implementation of the model was by discretizing the beta distribution into 50 equally spaced susceptibility groups between 0 and 1, with subgroup density given by the integral of the beta probability density function in each of the intervals. The cohort selection effect on the continuous distribution is shown in Figure 2.3.5 confirming the theoretical expectations from Figure 2.3.4.

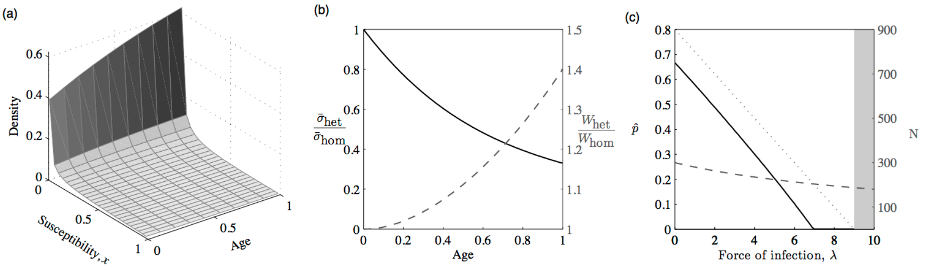


**Figure 2.3.5:** Beta distribution describing the susceptibility of a cohort of *Wolbachia* carriers exposed to natural and infection-induced mortality over age (model 2.8, with $a = 0$) (a). Mean susceptibility of heterogeneous *Wolbachia* carrier population relative to homogeneous implementation ($\bar{\sigma}_{het}/\bar{\sigma}_{hom}$) (solid curve referring to left axis), and total population size of heterogeneous *Wolbachia* carriers relative to homogeneous implementation ($W_{het}/W_{hom}$) (dashed curve referring to right axis) (b). Invasion threshold under beta-distributed protection (solid curve referring to left axis), total insect population after *Wolbachia* invasion ($N_{pos}$) (dashed curve referring to right axis) and in the absence of *Wolbachia* ($N_{pre}$) (dotted line referring to right axis) (c). Shaded area mark a regime where the force of infection precludes persistence of the non-carrier population. The homogeneous implementation assumes $\sigma = 0.21$, and the heterogeneous is initiated with beta-distributed susceptibility ($\bar{\sigma} = 0.21$ and coefficient of variation $1.24$). Other parameters are as in Figure 1.

## 2.4 Discussion

*Wolbachia* has been previously shown to provide their insect hosts with pathogen protection, promoting efforts to provoke invasions with *Wolbachia*-carrying *Ae. aegypti*. Trials in semi-field cages were carried out with sizes of the resident, non-*Wolbachia*-carrying and introduced carrier populations known to fairly good precision, and estimates of the fitness costs of *Wolbachia* available. Using the more costly *w*MelPop-CLA strain, invasions were observed despite the introduced proportion being below the theoretical threshold (Walker *et al.* 2011). Although consistent with numerical simulations of age-structured models (Rasgon *et al.* 2003), to our knowledge no reasons were given for the discrepancy from population-genetic predictions.

Previous studies have shown that parasitic vertically-transmitted symbionts could persist in the population if they protected against more virulent horizontally-transmitted pathogens (Jones *et al.* 2007; Lively *et al.* 2005). Jones *et al.* (2007) explored how systems that support coexistence of different parasites have their stable equilibria changing as a function of ecological and transmission parameters. Fenton *et al.* (2011) expanded the analyses by including CI and therefore showing where, in the parameter space of *Wolbachia*-associated fitness costs, could the symbiont invade. The invasion threshold was calculated for a system without any pathogens, and considering fecundity costs as the only *Wolbachia*-associated effect, which ultimately agreed with population genetics. To our knowledge, a more general analytical treatment of the threshold has not been pursued.

Our results suggest that while population-genetic models accurately predict the effects of fecundity costs on the invasion threshold, predictions for life-shortening and other effects benefit from considering host ecology. More generally, distinct *Wolbachia*-associated effects have independent contributions to the threshold, and are scaled by the ratio of the process which they affect and the birth rate, on which CI acts. Specific consideration of these processes is needed to determine the invasion threshold. Some but not all of these parameters may be straightforward to estimate; death rates could be estimated from published data, for *Ae. aegypti*, for instance, see Maciel-de-Freitas *et al.* (2011) and Southwood *et al.* (1972), or could be directly measured from studies of insect survival. Total burden of disease, on the other hand, could be difficult to accurately estimate.

It is expected that partial protection from pathogens lowers the threshold for *Wolbachia* invasion; heterogeneity further decreases this threshold. Even in the absence of heritability, high variance allows cohort selection effects (O'Hagan *et al.* 2012; Vaupel and Yashin 1985); we have shown that this effect can impact bistable systems by lowering the unstable equilibrium that determines the invasion threshold. This was initially explored with an idealised all-or-nothing distribution of protection, and subsequently by implementing variation in *Wolbachia*-induced protection as estimated for an isogenic population of *D. melanogaster*, the natural host of the *Wolbachia* strains that are being considered for dengue control interventions. The effects described here apply also to heterogeneity in life-shortening effect, due to the same cohort survival bias. In contrast, heterogeneous fecundity does not affect invasion, because there is no selection for higher fecundity, unless it covaries with a selected trait.

Consideration of inheritance and possible trade-offs should be of interest for both invasion analyses, population dynamics (Luo and Koelle 2013) and long term evolution of hosts (Jones *et al.* 2011), pathogens and *Wolbachia*, and has obvious importance to the long term success of intervention strategies based on the symbiont. The framework presented here allows the population dynamics of the symbiont to be readily incorporated into eco-evolutionary and epidemiological models of *Wolbachia*, which has been recognized as the next step towards understanding coevolution of host, symbiont and other parasites in a variety of contexts (Vavre and Charlat 2012).

While the transmission of dengue has had much attention because of its medical importance, mosquito-only flaviviruses have been identified (Calzolari *et al.* 2012), but relatively little is registered about their transmission and general epizootology; nevertheless, those could influence feasibility of provoked invasions ultimately aimed at controlling human disease. We here use a constant force of infection to model overall disease burden due to an ensemble of pathogens (Schmid-Hempel 1998). Transmission may occur horizontally, either directly (see supplementary material) or through an environmental stage, for instance, as proposed by Jones *et al.* (2007); moreover, there is transmission with an intermediate host, as for dengue virus in *Ae. aegypti*, or environmental reservoirs of generalist pathogens, which would not depend much on the prevalence in any one species, and could provide a more or less constant or seasonally fluctuating source of pathogens. Modeling a specific mode of transmission could prove fruitful; however, quantifying the importance of pathogen infection in such systems remains a great challenge. Nevertheless, our approach allows us to clearly show how the interaction of infection-induced mortality and CI affects invasion.

Analytical treatment of host ecology also gives insight beyond just frequency of *Wolbachia* carriers, uncovering regimes where population size post-invasion could be larger than the original, or where the population could be suppressed altogether – both of which are of interest in case of disease control strategies. Population suppression strategies based on cytoplasmic incompatibility have usually been thought in terms of male-only releases (Zabalou *et al.* 2004), or bidirectional incompatibility between different *Wolbachia* strains, possibly with populations artificially sustained through continuous releases (Calvitti *et al.* 2012; Dobson *et al.* 2002). As described here, suppression relies on a population that is able to outcompete residents while being unable to persist; this is possible under an external, environmental effect – here force of infection – and may lead to divergent results between predictions based on caged insects and subsequent field observations.

Extreme scenarios like population suppression by any of the mechanisms described above may be sensitive to unknown (or dynamic, as discussed above) parameter values such as force of infection, differences between lab and field effects, genetic background, and density-dependent regulation; nevertheless, the perspective is an interesting one and warrants further studies. Robustness of these predictions must be carefully considered, as heterogeneity in the traits, again, affects all these possibilities and previously discussed strategies (Calvitti *et al.* 2012; Dobson *et al.* 2002; Zabalou *et al.* 2004), and should be taken into account. To our knowledge, this is the first description of the ecological nature of the invasion threshold, and it provides clear analytical insight into this quantity.

## Acknowledgements

This research was supported by the Portuguese Foundation for Science and Technology (FCT), the European Commission (EC-ICT-231807) and CAPES. CSM is the beneficiary of a doctoral grant from the AXA Research Fund. CJS was partially funded by CNPq and FAPERJ.

# References

Barton NH, Turelli M (2011) Spatial waves of advance with bistable dynamics: cytoplasmic and genetic analogues of allee effects. Am Nat 178:E48–75. doi: 10.1086/661246

Bellows TS (1981) The Descriptive properties of some models for density dependence. J Anim Ecol 50:139–156.

Bian G, Joshi D, Dong Y, *et al.*(2013) Wolbachia invades Anopheles stephensi populations and induces refractoriness to Plasmodium infection. Science 340:748–751. doi: 10.1126/science.1236192

Blagrove M, Arias-Goeta C, AB F, Sinkins S (2012) Wolbachia strain wMel induces cytoplasmic incompatibility and blocks dengue transmission in Aedes albopictus. Proc Natl Acad Sci 109:255–260. doi: 10.1073/pnas.1112021108

Calvitti M, Moretti R, Skidmore AR, Dobson SL (2012) Wolbachia strain wPip yields a pattern of cytoplasmic incompatibility enhancing a Wolbachia-based suppression strategy against the disease vector Aedes albopictus. Parasit Vectors 5:254. doi: 10.1186/1756-3305-5-254

Calzolari M, Zé-Zé L, Růžek D, *et al.*(2012) Detection of mosquito-only flaviviruses in Europe. J Gen Virol 93:1215–1225. doi: 10.1099/vir.0.040485-0

Carrington LB, Lipkowitz JR, Hoffmann AA, Turelli M (2011) A re-examination of Wolbachia-induced cytoplasmic incompatibility in California Drosophila simulans. PLoS One 6:e22565. doi: 10.1371/journal.pone.0022565

Caspari E, Watson GS (1959) On the evolutionary Importance of cytoplasmic sterility in mosquitoes. Evolution (N Y) 13:568–570.

Chrostek E, Marialva MS, Esteves SS, Weinert LA, Martinez J, Jiggins FM, Teixeira L (2014) Wolbachia variants induce differential protection to viruses in Drosophila melanogaster: a phenotypic and phylogenomic analysis. PLoS Genet 9(12): e1003896.

Crain PR, Mains JW, Suh E, *et al.*(2011) Wolbachia infections that reduce immature insect survival: predicted impacts on population replacement. BMC Evol Biol 11:290. doi: 10.1186/1471-2148-11-290

Dobson SL, Fox CW, Jiggins FM (2002) The effect of Wolbachia-induced cytoplasmic incompatibility on host population size in natural and manipulated systems. Proc Biol Sci 269:437–445. doi: 10.1098/rspb.2001.1876

Fenton A, Johnson KN, Brownlie JC, Hurst GDD (2011) Solving the Wolbachia paradox: modeling the tripartite interaction between host, Wolbachia, and a natural enemy. Am Nat 178:333–342. doi: 10.1086/661247

Fisher RA (1937) The wave of advance of advantageous genes. Ann Eugen 7:355–369.

Gomes MGM, Lipsitch M, Wargo AR, et al. (2014) A missing dimension in measures of vaccination impacts. PLoS Pathog 10:e1003849. doi: 10.1371/journal.ppat.1003849

Hancock PA, Godfray HCJ (2012) Modelling the spread of Wolbachia in spatially heterogeneous environments. J R Soc Interface 9:3045–3054. doi: 10.1098/rsif.2012.0253

Hancock PA, Sinkins SP, Godfray HCJ (2011) Population dynamic models of the spread of Wolbachia. Am Nat 177:323–333. doi: 10.1086/658121

Hassell M, Lawton J, May RM (1976) Patterns of dynamical behaviour in single-species populations. J Anim Ecol 45:471–486.

Hedges LM, Brownlie JC, O'Neill SL, Johnson KN (2008) Wolbachia and virus protection in insects. Science (80- ) 322:702. doi: 10.1126/science.1162418

Hilgenboecker K, Hammerstein P, Schlattmann P, *et al.*(2008) How many species are infected with Wolbachia?–A statistical analysis of current data. FEMS Microbiol Lett 281:215–220. doi: 10.1111/j.1574-6968.2008.01110.x

Himler AG, Adachi-Hagimori T, Bergen JE, *et al.*(2011) Rapid spread of a bacterial symbiont in an invasive whitefly is driven by fitness benefits and female bias. Science (80- ) 332:254–256. doi: 10.1126/science.1199410

Hoffmann AA, Montgomery BL, Popovici J, *et al.*(2011) Successful establishment of Wolbachia in Aedes populations to suppress dengue transmission. Nature 476:454–457. doi: 10.1038/nature10356

Hoffmann AA, Turelli M, Harshman LG (1990) Factors affecting the distribution of cytoplasmic incompatibility in Drosophila simulans. Genetics 126:933–948.

Hoffmann AA, Turelli M, Simmons GM (1986) Unidirectional incompatibility between populations of Drosophila simulans. Evolution (N Y) 40:692–701.

Jansen VAA, Turelli M, Godfray HCJ (2008) Stochastic spread of Wolbachia. Proc R Soc B Biol Sci 275:2769–2776. doi: 10.1098/rspb.2008.0914

Jones EO, White A, Boots M (2007) Interference and the persistence of vertically transmitted parasites. J Theor Biol 246:10–17. doi: 10.1016/j.jtbi.2006.12.007

Jones EO, White A, Boots M (2011) The evolution of host protection by vertically transmitted parasites. Proc R Soc B Biol Sci 278:863–870. doi: 10.1098/rspb.2010.1397

Kendall B, Fox G, Fujiwara M, Nogeire T (2011) Demographic heterogeneity, cohort selection, and population growth. Ecology 92:1985–1993.

Lively CM, Clay K, Wade MJ, Fuqua C (2005) Competitive co-existence of vertically and horizontally transmitted parasites. Evol Ecol Res 7:1183–1190.

Luo S, Koelle K (2013) Navigating the devious course of evolution : the importance of mechanistic models for identifying eco-evolutionary dynamics in nature. Am Nat 181:S58–S75. doi: 10.1086/669952

Maciel-de-Freitas R, Aguiar R, Bruno R V., *et al.*(2012) Why do we need alternative tools to control mosquito-borne diseases in Latin America? Mem Inst Oswaldo Cruz 107:828–829.

Maciel-de-Freitas R, Koella JC, Lourenço-de-Oliveira R (2011) Lower survival rate, longevity and fecundity of Aedes aegypti (Diptera: Culicidae) females orally challenged with dengue virus serotype 2. Trans R Soc Trop Med Hyg 105:452–458. doi: 10.1016/j.trstmh.2011.05.006

Maynard-Smith J, Slatkin M (1973) The stability of predator-prey systems. Ecology 54:384–391. doi: http://dx.doi.org/10.2307/1934346

McGraw EA, O'Neill SL (2013) Beyond insecticides: new thinking on an ancient problem. Nat Rev Microbiol 11:181–193. doi: 10.1038/nrmicro2968

Moreira LA, Iturbe-Ormaetxe I, Jeffery JA, *et al.*(2009) A Wolbachia symbiont in Aedes aegypti limits infection with Dengue, Chikungunya, and Plasmodium. Cell 139:1268–1278. doi: 10.1016/j.cell.2009.11.042

O'Hagan JJ, Hernán MA, Walensky RP, Lipsitch M (2012) Apparent declining efficacy in randomized trials: examples of the RV144 HIV vaccine and CAPRISA 004 microbicide trials. AIDS 26:123–126. doi: 10.1097/QAD.0b013e32834e1ce7.Apparent

Osborne SE, Leong YS, O'Neill SL, Johnson KN (2009) Variation in antiviral protection mediated by different Wolbachia strains in Drosophila simulans. PLoS Pathog 5:e1000656. doi: 10.1371/journal.ppat.1000656

Pessoa D, Souto-Maior C, Gjini E, *et al.*(2014) Unveiling time in dose-response models to infer host susceptibility to pathogens. PLOS Comput Biol (in press).

Rasgon JL, Styer LM, Scott TW (2003) Wolbachia-induced mortality as a mechanism to modulate pathogen transmission by vector arthropods. J Med Entomol 40:125–132.

Schmid-Hempel P (1998) Parasites in Social Insects. Princeton University Press.

Smith PG, Rodrigues LC, Fine PE (1984) Assessment of the protective efficacy of vaccines against common diseases using case-control and cohort studies. Int J Epidemiol 13:87–93.

Southwood T, Murdie G, Yasuno M, *et al.*(1972) Studies on the life budget of Aedes aegypti in Wat Samphaya, Bangkok, Thailand. Bull World Health Organ 46:211–226.

Taylor CM, Hastings A (2005) Allee effects in biological invasions. Ecol Lett 8:895–908. doi: 10.1111/j.1461-0248.2005.00787.x

Teixeira L, Ferreira A, Ashburner M (2008) The bacterial symbiont Wolbachia induces resistance to RNA viral infections in Drosophila melanogaster. PLoS Biol 6:e1000002. doi: 10.1371/journal.pbio.1000002

Turelli M (2010) Cytoplasmic incompatibility in populations with overlapping generations. Evolution (N Y) 64:232–241. doi: 10.1111/j.1558-5646.2009.00822.x

Turelli M, Hoffmann AA (1995) Cytoplasmic incompatibility in Drosophila simulans: dynamics and parameter estimates from natural populations. Genetics 140:1319–1338.

Turelli M, Hoffmann AA (1991) Rapid spread of an inherited incompatibility factor in California Drosophila. Nature 353:440–442. doi: 10.1038/353440a0

Vaupel JW, Yashin AI (1985) Heterogeneity's ruses: some surprising effects of selection on population dynamics. Am Stat 39:176–185.

Vavre F, Charlat S (2012) Making (good) use of Wolbachia: what the models say. Curr Opin Microbiol 15:263–268. doi: 10.1016/j.mib.2012.03.005

Walker T, Johnson PH, Moreira LA, *et al.*(2011) The wMel Wolbachia strain blocks dengue and invades caged Aedes aegypti populations. Nature 476:450–453. doi: 10.1038/nature10355

Werren JH, Baldo L, Clark ME (2008) Wolbachia: master manipulators of invertebrate biology. Nat Rev Microbiol 6:741–751. doi: 10.1038/nrmicro1969

Yeap HL, Mee P, Walker T, *et al.*(2011) Dynamics of the "popcorn" Wolbachia infection in outbred Aedes aegypti informs prospects for mosquito vector control. Genetics 187:583–595. doi: 10.1534/genetics.110.122390

Zabalou S, Riegler M, Theodorakopoulou M,(2004) Wolbachia-induced cytoplasmic incompatibility as a means for insect pest population control. Proc Natl Acad Sci U S A 101:15042–15045. doi: 10.1073/pnas.0403853101

*In my whole career, I have never seen a Tajima's D statistic equal to zero.**

Nick Barton

# A

# Stability analysis of the invasion models

Here we elaborate on the local stability analysis of the homogeneous system. The procedures are extensible to the other cases by analogy. Writing the system of differential equation in condensed form as:

$$\frac{d\mathbf{X}}{dt} = \mathbf{F}(\mathbf{X}), \tag{A.1}$$

where $\mathbf{X} = \begin{bmatrix} U & W \end{bmatrix}^T$ and $\mathbf{F}$ is the vector field defined in (2.4), the associated lineralized system can be written as:

$$\frac{d\mathbf{X}}{dt} = \mathbf{J}\mathbf{X}, \tag{A.2}$$

where **J** is the Jacobian matrix of **F** evaluated at some equilibrium. To calculate the Jacobian matrices we have that:

$$\begin{aligned}
\frac{\partial F_1}{\partial U} &= \frac{a(N^2 - W^2 s_h)}{N^2} - b - f(N) - f'(N)U - \lambda \\
\frac{\partial F_1}{\partial W} &= -\frac{aU^2 s_h}{N^2} - f'(N)U \\
\frac{\partial F_2}{\partial U} &= -f'(N)W \\
\frac{\partial F_2}{\partial W} &= a(1 - s_f) - \frac{b}{1 - s_l} - f(N) - f'(N)W - \sigma\lambda.
\end{aligned} \tag{A.3}$$

In the following we specify the density-dependence function as $f(N) = kN$, and proceed to evaluate the Jacobian at each of the three non-trivial equilibrium points and calculate the associated eigenvalues.

## A.1 Trivial equilibrium

The trivial equilibrium, $\mathbf{X}_{pre} = [\begin{array}{cc} 0 & 0 \end{array}]^T$, leads to a Jacobian matrix whose eigenvalues take the form:

$$\begin{aligned}
e1_{pre} &= a\left(1 - \frac{W^2 s_h}{N^2}\right) - b - \lambda \\
e2_{pre} &= a(1 - s_f) - \frac{b}{1 - s_l} - \sigma\lambda.
\end{aligned} \tag{A.4}$$

The first lies between $a(1 - s_h) - b - \lambda$ and $a - b - \lambda$ which, with this level of determination, may be negative or positive. Its upper bound, however, coincides with $f^{-1}(U_{pre})$, whose positivity is ensured by the monotonicity of $f$ when $\mathbf{X}_{pre}$ exists. The second eigenvalue coincides with $f^{-1}(W_{pos})$ and, by a similar argument, is always positive when $\mathbf{X}_{pos}$ exists, establishing that the trivial equilibrium is always unstable when the conditions for *Wolbachia* invasion are met.

## A.2 Pre-invasion equilibrium

The pre-invasion equilibrium, $\mathbf{X}_{pre} = \begin{bmatrix} U_{pre} & W_{pre} \end{bmatrix}^T$, is the same under all the models considered in the main text and given by:

$$U_{pre} = \frac{a-b-\lambda}{k} \qquad \text{and} \qquad W_{pre} = 0, \tag{A.5}$$

which can only represent an existing population when $\lambda < a - b$. Replacing in (A.3) we obtain a Jacobian matrix whose eigenvalues are:

$$\begin{aligned} e1_{pre} &= -a+b+\lambda \\ e2_{pre} &= -as_f - \frac{bs_l}{1-s_l} + (1-\sigma)\lambda. \end{aligned} \tag{A.6}$$

The first is negative when $\mathbf{X}_{pre}$ exists, vanishing at $\lambda = a - b$ when the population goes extinct due to pathogen pressure. The second can be rewritten as $-as_h\hat{p}$, where $\hat{p}$ is the threshold frequency of *Wolbachia*-carriers (2.5) required for invasion. This eigenvalue is therefore negative whenever there is a positive invasion threshold, confirming the stability of the *Wolbachia*-free population when this condition is verified.

## A.3 Post-invasion equilibrium

The post-invasion equilibrium, $\mathbf{X}_{pos} = \begin{bmatrix} U_{pre} & W_{pre} \end{bmatrix}^T$, under model (2.4) is:

$$U_{pos} = 0 \qquad \text{and} \qquad W_{pos} = \frac{a(1-s_f)}{k} - \frac{b}{k(1-s_l)} - \frac{\sigma\lambda}{k}, \tag{A.7}$$

which can represent an existing population when $\lambda < \sigma^{-1}[a(1-s_f) - b(1-s_l)^{-1}]$. Replacing in (A.3) we obtain a Jacobian matrix whose eigenvalues are:

$$\begin{aligned} e1_{pre} &= a(s_f - s_h) + \frac{bs_l}{1-s_l} - (1-\sigma)\lambda \\ e2_{pre} &= -a(1-s_f) + \frac{b}{1-s_l} + \sigma\lambda. \end{aligned} \tag{A.8}$$

The second is negative when $\mathbf{X}_{pos}$ exists, vanishing at $\lambda = \sigma^{-1}[a(1 - s_f) - b(1 - s_l)^{-1}]$ when the population goes extinct due to pathogen pressure. The first can be rewritten as $as_h(\hat{p} - 1)$, which is negative for $\hat{p} < 1$, confirming the stability of the *Wolbachia*-carrier population.

## A.4 Invasion threshold

The unstable equilibrium, $\mathbf{X}_{uns} = \begin{bmatrix} U_{uns} & W_{uns} \end{bmatrix}^T$, under model (2.4) is:

$$\begin{aligned} U_{uns} &= \frac{1-\hat{p}}{k}\left[a(1-s_f) - \frac{b}{1-s_l} - \sigma\lambda\right] \\ W_{uns} &= \frac{\hat{p}}{k}\left[a(1-s_f) - \frac{b}{1-s_l} - \sigma\lambda\right], \end{aligned} \tag{A.9}$$

Stability analysis is less tractable in this case but phase planes obtained numerically in Figure 1 are consistent with a saddle point, therefore always unstable.

*I don't believe in analytical results; I just set up a model and simulate the shit out of it.*.*

Mark Thomas

# 3

# Within-host dynamics in a DENV-Aedes-Wolbachia system

## Author contributions

The author of the thesis, in collaboration with Rafael Maciel-de-Freitas and Gabriel Sylvestre Ribeiro designed the experiments. The author and Gabriel Sylvestre Ribeiro performed the infection experiments and maintenance of the cohort. Gabriel Sylvestre Ribeiro performed the virus quantification. The author performed all statistics, simulations, inference, and analyses in general.

ABSTRACT

Infection is a complex and dynamic process involving a population of invading microbes and the host with its responses aimed at controlling the invasion. Depending on the purpose, infection at the organism level can be described by a process as simple as a coin toss, up to a multi-factorial dynamic model – the former may be adequate as a component of a population model for instance, while the latter is necessary for a comprehensive description of the process from challenge with an infectious inoculum up to establishment or elimination of the pathogen. Often, readouts of laboratory experiments are static, snapshots of the process, assayed under some convenient experimental condition, and therefore cannot thoroughly represent the system. As opposed to the discrete treatment of infection in population models, or the summarized description of typical lab experiments, in this chapter infection is treated as a dynamic process dependent on the initial conditions of the infectious challenge, viral growth, and the host response as functions of time. Here, experimental data is generated for multiple doses of type 1 dengue virus, and viral levels are recorded at different points in time for two populations of mosquitoes: either carrying endosymbiont *Wolbachia* or not. A dynamic microbe-response model is used to describe pathogen growth in the presence of the host immune system, and to infer model parameters for the two populations of insects, revealing a slight – but potentially important – reduction in viral levels along time conferred by the symbiont.

## 3.1 Background

The potential of a pathogen to sustain its transmission in a population is often summarized into general quantities like vectorial capacity, which includes host variables such as population size and biting rate, or more restricted ones like vector competence, which is the ability of a single vector, e.g. one mosquito, to become infected and further transmit the disease.

Because there is no real readout of "ability to be infected and then infect", it is not possible to measure vectorial competence directly, instead it must rely on proxies such as susceptibility to infection or, more concretely, probability of having detectable levels of virus in the saliva.

As shown before (Gomes et al. 2014; Pessoa et al. 2014), measuring dose-independent parameters of susceptibility to infection can be accomplished by constructing a dose-response profile. It has also been shown that, if a proxy for infection response is time dependent, that dependence should be treated appropriately; otherwise, there may be distortions in the estimated parameters (Pessoa et al. 2014). In the case of survival as a readout, a combination of a survival profile for the infected individuals plus another profile for the uninfected can be used to determine the infected proportion of a group challenged by any one dose (Pessoa et al. 2014). This was used to assess the protection conferred by the bacterial endosymbiont *Wolbachia* to its invertebrate host *Drosophila melanogaster*, by comparing these dose-survival profiles for hosts with and without the symbiont.

In the case presented in the next sections, a comprehensive assessment of the response to infectious challenges is conditioned on having an appropriate description of the dynamics of viral particles inside a host. The probability of developing a systemic infection, therefore, is not determined by simple presence of virus at an arbitrary time point. Instead, infection is determined by a dynamic profile that conforms to the establishment and persistence of the pathogen – i.e. full blown, systemic infection – as opposed to pathogens being transiently detected, but being quickly eliminated without greater consequences to the host.

The natural history of a virus population ingested by an insect involves a passage through the gut of the host, where some of the viruses are likely to be killed by physical, chemical, enzymatic or immune processes. The viral particles that survive this first rounds would then need to cross the midgut infection barrier (and escape barrier)(Franz 2015) to proliferate in the body cavity and need to spread to different tissues in order to be transmitted to other hosts, before the host either recovers or dies.

From this cartoon description, it can then be expected that an initially small population of viruses will multiply inside a host and become manyfold larger. Indeed, trying to quantify virus levels very early after inoculation may result in very low or undetectable levels, while later on it could be very easy to find large numbers of viral nucleic acid copies in the host. Therefore, without much of a stretch, the first simple expectation is that if we follow viral levels along time for a host, we should be able to see this growth and sketch a curve describing it.

The second part of this illustration concerns the initial inoculum. The number of viral particles must be large enough to accommodate initial losses, then replicate, and eventually reach high numbers. If the initial virus level is high, it will reach high numbers faster, or more easily; conversely, if the number is small it may take longer, or it may not be able to achieve high numbers at all. Presumably, there may also be some intermediate level, at which either establishment or elimination could occur; therefore, otherwise identical hosts (e.g. isogenic lines) can either become infected or not just because of small fluctuations in the initial conditions, or due to stochasticity along the process of viral replication and elimination by the host.

Considering this illustration of both the initial conditions and time course of infection, we can hypothesize that to fully describe infection we need to describe a range of infective doses ranging from very low probability of infection up to extremely infective, as well as measurements at early time points – when the virus is struggling to succeed in its first replication rounds – time points in-between, and up to later times, when infection is established (or eliminated) and any significant dynamics are no longer observed. In terms of data alone, that would be a complete data representation of infection, as opposed to a typical laboratory condition.

Although the interest in the evolutionary conflicts between host and parasites has a long history – see Frank (1996) and references therein – the development of compartmental models of microbes and associated immune responses within a host lags population transmission models by at least a few decades (Alizon & van Baalen 2008). An early example of theoretical models of the immune system can be seen in Perelson & Oster (1979); models of parasite-immunity interactions took about another decade to become more prevalent (Alizon & van Baalen 2008; Antia & Lipsitch 1996; Sasaki & Iwasa 1991).

The mathematical formulation of within-host dynamics typically includes simple, constant-rate pathogen growth, with immune response-dependent pathogen death; immunity is described as either induced by the pathogen (Antia et al. 1997), constitutive (Hamilton et al.

2008), or both. Less common formulations favor other features as a limiting factors, like host biomass (White et al. 2012), but these disregard immune response altogether.

Importantly, a pathogen-free state is not stable in these models and exists only in the absence of pathogens – deterministically, infection is eventually established once pathogens enter the host. Therefore, they do not contemplate (except through stochasticity alone) a scenario where an initial inoculum can either cause systemic infection or fail to invade. Such a bistable model is potentially trickier to formulate, and may require a few extra parameters; an example is that of Pujol et al. (2009), which has one more parameter than similar ones (Antia et al. 1997; Hamilton et al. 2008), but one less initial unknown, having effectively the same number of degrees of freedom.

In this chapter I will describe an experiment designed to obtain a data set with a structure compatible with a dynamic model, where the time dependence of dengue virus 1 (*DENV-1*) in *Aedes aegypti* hosts is recorded for varying initial infectious doses. Dynamic models are proposed to describe the entire data set, and therefore allow associating a profile of viral growth for a certain initial condition to a probability of full-blown infection. This model-based approach can then be used to compare two populations of *Ae. aegypti* that differ in whether the individuals are carriers of endosymbiont *Wolbachia* or not.

In the chapter appendix the results of a mechanistic within-host model of pathogen replication coupled to immune response is compared to a simpler logistic growth/exponential decay model. Overall the framework proposed here is designed to fully describe infection in terms of initial conditions, pathogen dynamics, and system state. I also discuss the limitations of the current data set and analyses, and how both could be improved in further experiments of the kind.

## 3.2 Methods

### 3.2.1 Preliminary data and general framework

As described in previous chapters, the (binary) outcome of an infectious challenge is unlikely to be captured by a single experimental condition, in that case a single challenge dose (Pessoa et al. 2014), so an additional variable accounting for challenge dose is needed. In the same *Drosophila*-DCV system, not only the proportion infected but also the levels of *Drosophila* C virus on a given day can be seen to be dose dependent, and for a single dose a clear temporal trend can be observed (Zwerschke et al. unpublished) – these features are shown in figure 3.2.1.

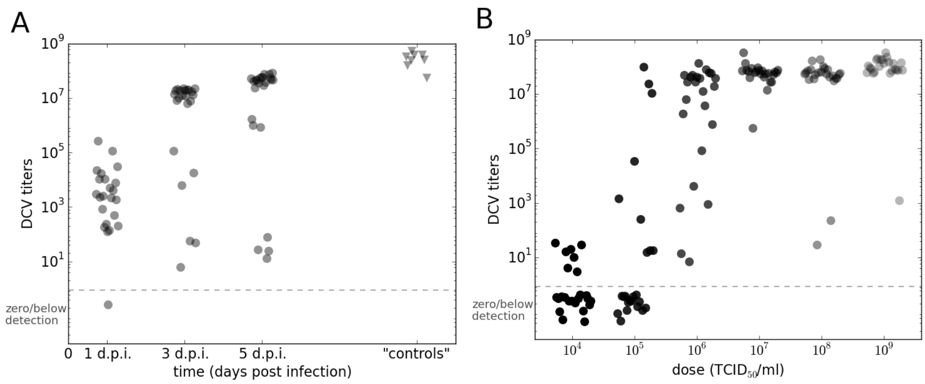


**Figure 3.2.1:** Time course data (at 1, 3, and 5 days post infection) of Drosophila C Virus, on a pricking challenge with $10^8$ TCID$_{50}$ solution (panel A) - “control” referes to an additional group of flies assayed later times whenever they become moribund. Quantification of viral level as a function of pricking challenge dose, assayed at day 5 post infection (panel B). (Reproduced with permission from Zwerschke et al. (unpublished))

Therefore, to describe and compare viral levels it is necessary to account for both the inoculated pathogen dose (henceforth simply *dose*) and the time when viral levels are assayed (i.e. the number of days post infection or simply the *time*). A complete design should include data for all time points and doses. Figure 3.2.2 shows a typical setup where two groups are compared for an arbitrary condition using simple statistical tests, and a multi-factorial design that require more sophisticated methods.

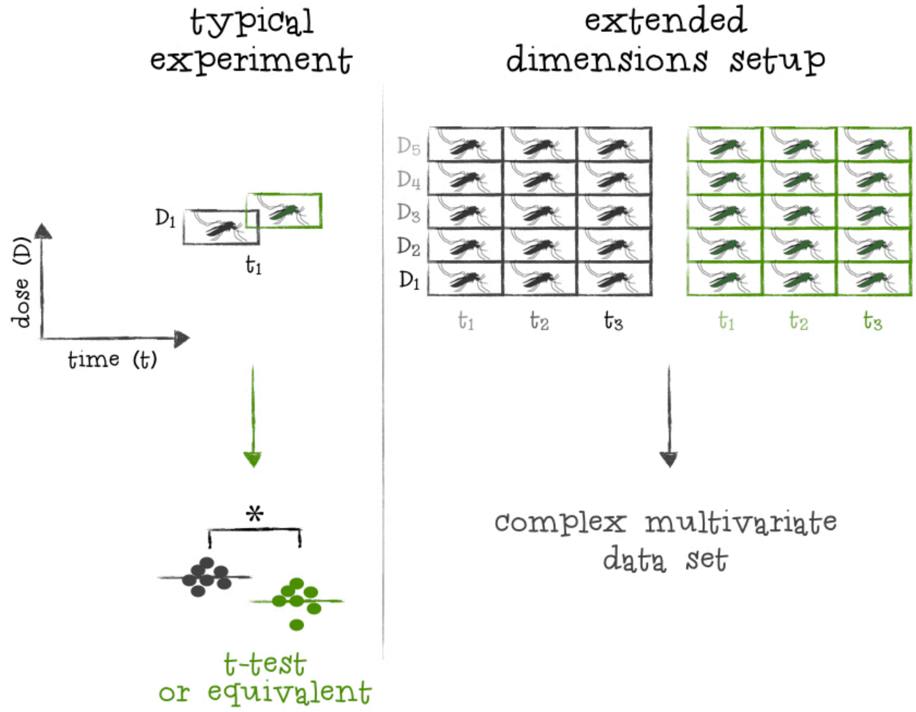


**Figure 3.2.2:** A typical experiment comparing groups under a single condition (left), and a multi-factorial design taking time after challenge and injected dose into account (right).

The analysis framework must make a relevant assessment of the experimental readout considering the controlled experimental variables; these dependent variables are the viral and symbiont levels inside each host. Unlike the typical experiment, where a simple comparison of means between two groups could be applied, here a model that describes titers as a function of time for varying initial doses is necessary. Figure 3.2.3 shows a dynamic system whose components interact to produce an observable outcome, that may be dependent on the initial conditions.

In the next section I describe the formalization of this conceptual model, the precise

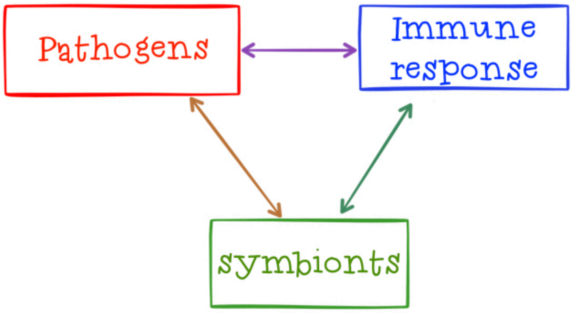


**Figure 3.2.3:** Schematic representation of the system and its interactions. Dynamic feedback between the components determines their levels as functions of time.

structure of the data set, and the detailed experimental methodology carried out.

## 3.2.2 Experiment design and execution

### Rearing of mosquitoes in the laboratory

The *Aedes aegypti* mosquitoes were reared following the protocol described by Consoli & Lourenço-de-Oliveira (1994). Egg laying is done on strips of filter paper, the surface of which the eggs adhere to; afterwards they are put in dechlorinated water 1.5-liter containers for 24 hours for egg eclosion to occur. Larvae were fed with brewer's yeast every two days. Pupae were transfered to cages were adults emerged. Ideal temperature is 25ºC, so 25 ± 3ºC and relative humidity 80 ± 5% was used for 2-3 days, with 10% sugar meals *ad libitum*, and allowing mating up to 24 hours before any infection.

Two distinct mosquito populations were used: *wMelBr* and *wMelTET* (henceforth also simply *w*Mel and *TET*). The *wMelBr* population is the result of the crossing of brazilian mosquitoes originally not carrying the *Wolbachia* with mosquitoes that did carry the symbiont; therefore these are mosquitoes with the genetic background of brazilian mosquitoes, but carrying the *Wolbachia* endosymbiont. The *wMelTET* mosquito population was obtained by treating of the *wMelBr* population with the antibiotic tetracyclin, hence eliminating the endosymbiont. The genetic background is then equivalent between the two lines, with the only difference being the presence or absence of *Wolbachia* symbiont.

### Dengue virus

The dengue virus ($DENV$) samples were collected from serotypes and strains circulating in the city of Rio de Janeiro, Brazil, which were isolated from viremic patients and deposited in the Laboratory for Flaviviruses in Instituto Oswaldo Cruz (IOC-Fiocruz). The first serotype was used ($DENV$-*1*), starting from a sample amplified to hundred million ($10^8$) times the 50% infective dose for tissue culture $\text{TCID}_{50}ml^{-1}$ and stored at $-80^oC$. The original sample was then serially diluted tenfold down to ten thousand times the $\text{TCID}_{50}$ (the dilution volume, milliliter, is implied hereafter unless explicitly stated), therefore resulting in the concentrations of $10^4, 10^5, 10^6, 10^7, 10^8$ $\text{TCID}_{50}$, as well as controls mock-infected with virus-free medium.

### Intrathoracic injections with dengue virus

Anticipating that the precise initial conditions would be very important for the model-based analysis, we decided that by design our experiment would try to bypass any processes other than the intrinsic viral replication ability and the host's immune response against the pathogen; therefore, we chose to inject *DENV-1* directly into the body cavity (haemocoel) of the mosquitoes, in an attempt to reduce variation in the observed viral levels. Here it is worth noting that the mosquitoes midgut is the natural bottleneck to virus proliferation, with physical and chemical barriers such as digestive enzymes, pH changes, as well as epithelial receptors; depositing viral particles directly in the haemocoel bypasses them, and therefore skips some of the normal steps in infection by ingestion of contaminated blood. To achieve high precision on this end we used a nanoliter-precision injector (Nanoject II, Drummond ScientificCompany) that allows for automatic calibration of the injected volume (and as a consequence of viral titer) in order to reduce variation in this initial condition to a minimum. Three injections of precisely 69 *nl* then were used for every single mosquito.

Experimentally executing an oral infection protocol is feasible, nevertheless, and in some ways easier than using direct injection, so I discuss the potential implications of using this unnatural route of infection and the importance of using a more natural oral infection methodology in the last section of this chapter.

It is assumed that *Ae. aegypti* can transmit *DENV* after 10-14 days of being exposed to the virus, when the extrinsic incubation period is completed and it has disseminated to different tissues and organs, finally reaching the insect's saliva. Despite it being common to assay infection once around the time the incubation period is expected to be complete, and viral levels are expected to be higher (and more easily detectable), we decided to quantify *DENV* at three time points: as early as 3 days post-infection (d.p.i.), and then at 7, and 14 d.p.i.

Besides being performed in mosquitoes from the same generation, the whole procedure of infection from dilution of the viral samples to infection and storage of the mosquitoes was performed in a single day, over a period of 12 hours, as to minimize any variation related to the mosquitoes, the viral sample or its dilution, or any other unforeseen time dependent variables. A total of 616 mosquitoes were injected and kept in Falcon tubes, divided such that each condition would have at least 15 mosquitoes at the beginning of the experiment. The mosquitoes were kept at laboratory conditions and fed every day until their time point was due, when each mosquito was individually stored whole and cryopreserved at −80°C until all mosquitoes were taken out at once for quantification of viral RNA, and *Wolbachia* and mosquito DNA.

### Nucleic acid detection

Mosquitoes were macerated and viral RNA was extracted using a commercial kit (High pure viral nucleic acid kit – Roche). RNA detection for each individual was performed by by quantitative real-time polymerase chain reaction (qRT-PCR, or more simply, qPCR), as in Maciel-de-Freitas *et al.* (2011). The qPCR assay was based on a one-step assay (TaqMan Fast Virus 1-step Master Mix – Thermo Fisher Scientific).

*Wolbachia* DNA concentration was individually assayed by detection of the TM513 gene of the symbiont, while *Ae. aegypti* DNA was quantified using the RPS17 housekeeping gene for a stable reference, also through qPCR.

The serially diluted samples used for infection were also quantified to allow calculation of the absolute virus concentration. At least one technical replicate was performed for each quantification of every sample.

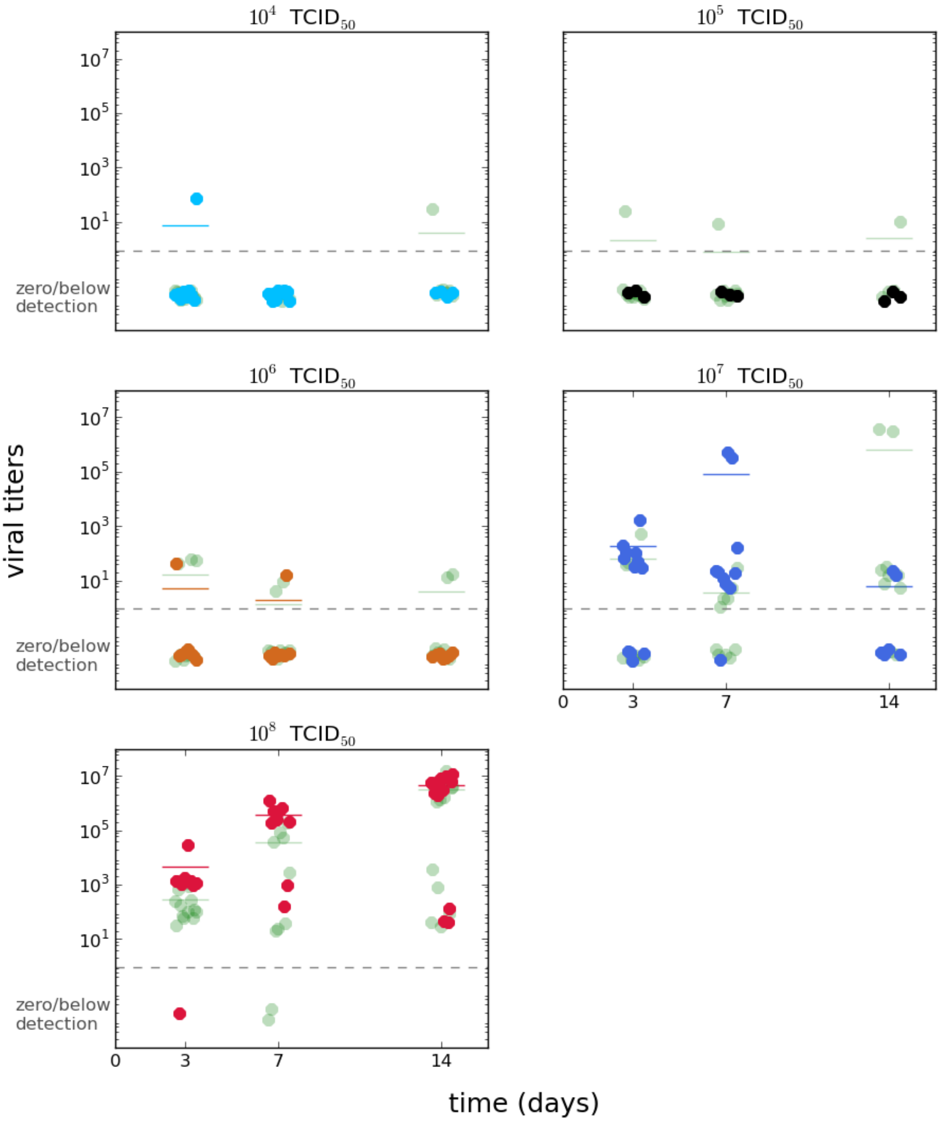


**Figure 3.2.4:** *DENV-1* viral titer data for *TET* group (colors) overlaid to that of the *w*Mel group (light green).

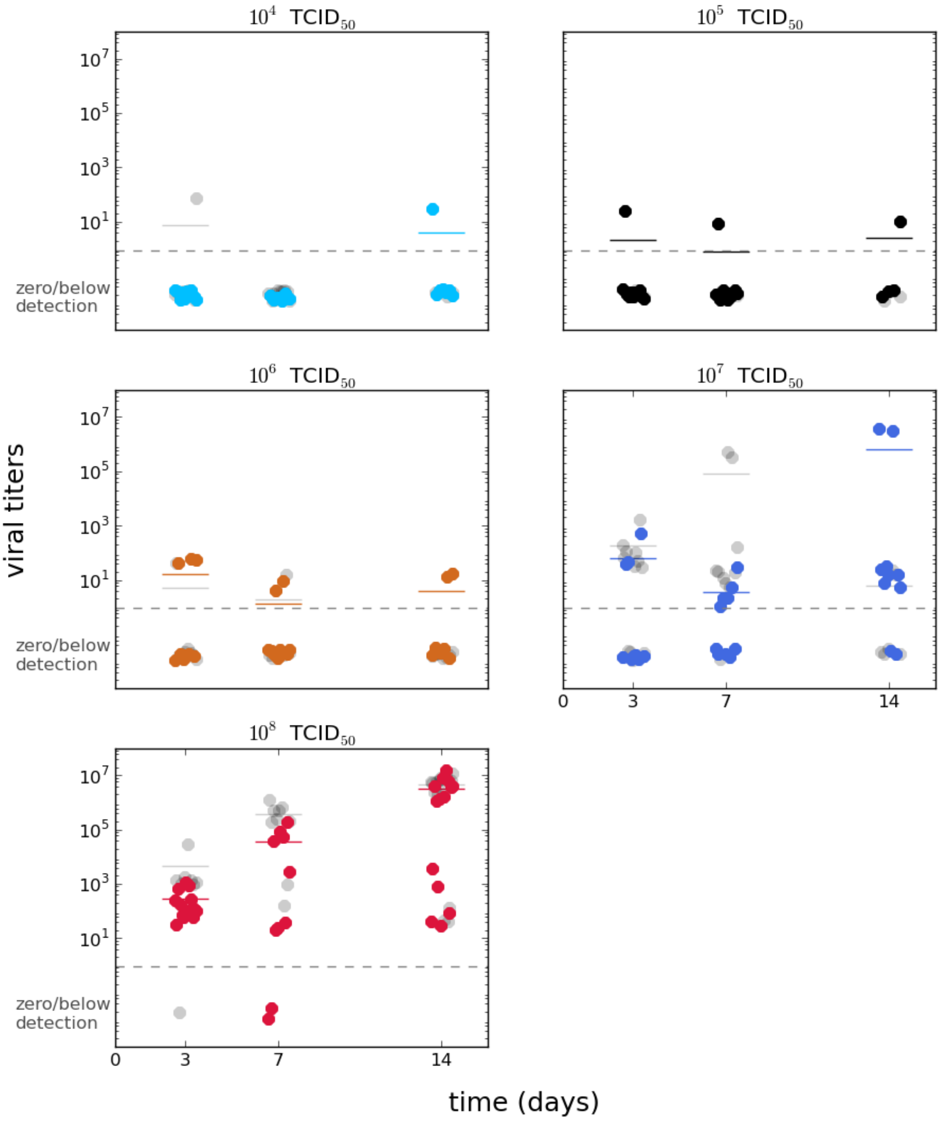


**Figure 3.2.5:** *DENV-1* viral titer data for *w*Mel group (colors) overlaid to that of the *TET* group (light gray).

### Structure of the final data set

The combination of the multiple virus dilutions (plus mock-infected controls) with the different time points resulted then in eighteen different infection conditions, which we expected to be representative of the full course of infection for different initial conditions – in contrast to the traditional single high-dose/late-time assays often used as a proxy for vector competence. Further combined with the two different population conditions, the entire procedure resulted in a quantification of *DENV-1* RNA, *Wolbachia* DNA, and *Ae. aegypti* DNA for each of the 301 samples. Table B.1.1 (appendix B) shows the total numbers of mosquitoes in the experiment identified by groups and subgroups.

It is plausible that mosquitoes infected with higher doses would die in higher proportion, especially in later time points; these effects could result in a survivor bias, or cohort selection, and the distribution of viral titers in some conditions being disproportionally affected. At day zero mosquitoes were uniquely identified by population and challenge dose, but not time point, and mortality was not recorded at each time point of PCR quantification; therefore, it is not possible to formally analyze survival of the different cohorts with this data set. As a function of dose alone losses do not show a clear trend; however, it is not possible to completely eliminate or confirm the hypothesis of cohort bias. This limitation could be overcome in future experiments of this kind.

To produce the final data set used for all analyses hereafter, both the levels of *DENV-1* and *Wolbachia* were normalized by the mosquito's housekeeping gene, expected to remain unchanged for all conditions. The data set for viral levels in the *TET* group is shown in figure 3.2.4. For the *w*Mel group the viral level data is shown in figure 3.2.5.

As for the *Wolbachia* levels, the *TET* group was used as a negative control, and every sample had undetectable qPCR levels of the symbiont in at least one of the technical replicates, as expected. For the *w*Mel group, the levels were computed relative to the same house keeping gene used as a standard for the viral titer data. These are shown in figure 3.2.6.

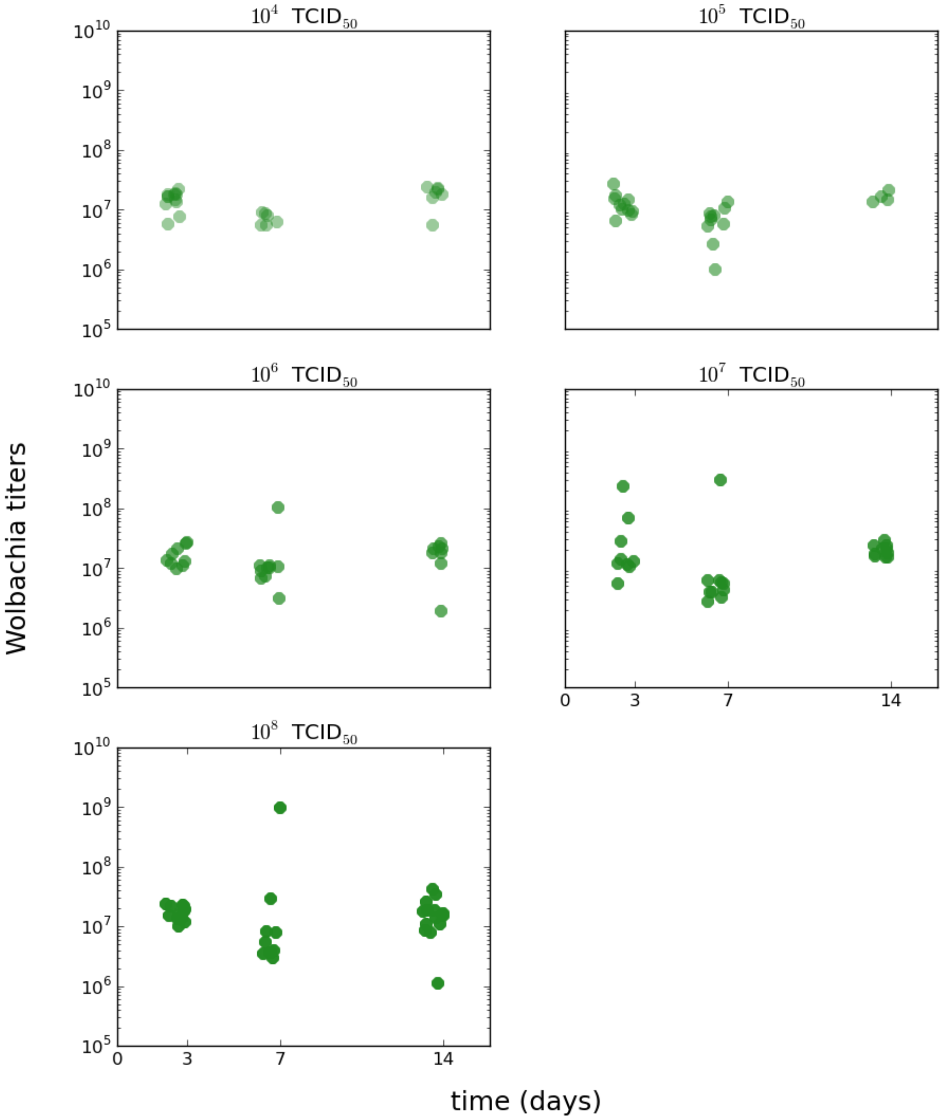


**Figure 3.2.6:** Wolbachia levels in symbiont carrying population.

### 3.2.3 Models

#### Pathogen growth plus immune response bistable model

Here I use a variant of a model previously available in the literature, the only modification being a logistic-like term in of the equations, inducing a carrying capacity (Pujol et al. 2009). In the published model there was no need for such a term because only a binary outcome was of interest – either infection elimination or persistence – and pathogen level data was absent. As such, there was no need to include saturation in growth; once simulations were found to grow towards large numbers the simulation was terminated, and therefore anything after that time point was disregarded.

This model includes one equation for the pathogen population $(P)$, and one equation for a host response $(R)$, shown in 3.1. The explanation of how the two interact is given below.

$$\begin{aligned} \frac{dP}{dt} &= rP - \delta_P PR - kP^2 \\ \frac{dR}{dt} &= a + \lambda P - \gamma R - \delta_R PR \end{aligned} \tag{3.1}$$

In this model $r$ is the growth rate of pathogens and $k$ the other parameter modulating the carrying capacity level at which pathogen population size saturates. Reduction in growth is dependent on a non-linear $\delta_P PR$ term accounting for the rate of destruction of pathogen units which increases with the response $R$.

The host response can be described by pathogen-independent, or constitutive, components $a$, a constant rate of recruitment of the response, minus a linear death rate $\gamma R$, as well as pathogen-dependent, or induced, components $\lambda P$, describing the increased rate of recruitment of the response minus a non-linear term $\delta_R PR$ representing the pathogen-induced destruction, use, or wear of the response.

The model has 7 free parameters, that describe actual biological processes. From these rules bistable dynamics emerge. Parameters are dose-independent; therefore, whether pathogens increase or decrease in numbers depends exclusively on the initial conditions.

Besides using a mechanistic description of the pathogen processes involved, this model also makes predictions about the host response that may cause the decline of the pathogen population, or allow their increase – this response can be likened to a real immune response, despite not representing one explicit immune process, but more of a composite bundle.

### 3.2.4 Statistics and Inference

Means for each time/dose combination were computed and displayed only as a visual aid (shown in figures 3.2.4 and 3.2.5). Significance tests were not performed, and pairwise comparisons between conditions are not reported. Pearson correlations were computed between the viral and *Wolbachia* levels in the *w*Mel group, both for the dose/time points separately as well as for all data regardless of challenge dose and time (i.e. a single correlation for the complete data set). $R^2$ and p–values are reported for those cases.

Bayesian inference of the parameters in the non-linear models was performed using a Markov Chain Monte Carlo implementation in the Python programming language (Patil et al. 2010). Since the data was normalized so that the lowest nonzero value was one and all values were discrete, a poisson distribution was used to compute the likelihood of the parameters given the data.

Uniform priors were used on a wide range of positive values unless otherwise stated, ensuring that posterior values did not concentrate on the higher limit after the procedure; in case that was observed estimation was repeated with wider prior intervals. Convergence was assessed by stability of the chains, and replicate chains were run to make sure the same approximate values were obtained regardless of starting point of the Markov chain (Gelman et al. 2013, chap. 11). *Burn-in* was performed by discarding a number of initial samples according to the total length of the chain

The estimated mean or median, and their confidence intervals were used instead of the frequentist pairwise approach to ascertain differences between groups. The dynamic pathogen-response model was simultaneously fitted to both *TET* and *w*Mel group data sets.

Given the relatively high number of degrees of freedom, there are choices ranging from fitting a completely separate set of parameters to each of the two data sets, up to fitting a single set of parameters to both data sets. As an intuitive compromise between freedom and inferability, we choose to fix the pathogen equation parameters between groups, and allow the host response parameters to differ between groups. In the appendix we use infer alternative sets of parameters for robustness testing and in an attempt to improve inference. We further discuss our choices for parameters that are shared or independent between the two groups in the results section as well.

## 3.3 Results

### 3.3.1 Descriptive data summaries and correlations

From the panels in figure 3.2.4, for the *TET* group, a broad trend can be observed, where the lowest doses have mainly zero/undetectable levels, while the highest doses are several fold greater and appear to increase with time. Looking at the mean values computed from the data, indicated by the horizontal lines, it can be generally observed that lower doses tend to remain low or even become undetectable, while high doses increase on average. There is the exception, however, of the second highest dose ($10^7$ $\text{TCID}_{50}$) at the last time point (day 14 post-infection); it is noteworthy that while the values themselves do not look like outliers, the number of points for that dose and time are comparatively low (due to loss of insects that died throughout the experiment), and may be insufficient to represent the full distribution of viral titers.

As mentioned before, although we cannot confirm or discard that there was an increased mortality for both that dose and time, there are only six data points for that condition, which less likely to be representative of the whole viral titer distribution, especially in the light of the previous time point, the other doses at this time, as well as other preliminary data (not shown). I discuss further the treatment of data points that do not conform to this broad decrease/increase pattern in the following sections.

Beyond the summaries of the mean, it can be observed that the spread in the viral titers increase with time in the case of the highest doses, with the highest values at each point increasing and the lowest decreasing. For the lowest doses, with viral titers being low from the beginning, what is observed is the tendency to go to zero with time. Without formal tests of the quantitative aspects, the broad picture is compatible with that of a viral challenge that can be eliminated by the host, but that otherwise increases with time until some saturating higher level.

As a preliminary analysis of the dynamic trends, we fit a straight line ($viraltiters = a \cdot time + b$) through the first two time points of the highest dose in log scale, as a means of obtaining a crude estimate of the exponential growth rate (which would be straight line in log scale) as the angular parameter, and the initial concentration of the virus, as the logarithm of the linear parameter estimated by this simple, analytic method. The estimates are $a = 0.94$, $b = 3.1$, with a p–value $< 0.001$. Because these estimates are based on a very restricted subset of the data, the values are taken as a rough quantification only. Instead of a hard restriction, the estimates are used as prior information in the bayesian estimation method to constrain the parameter space.

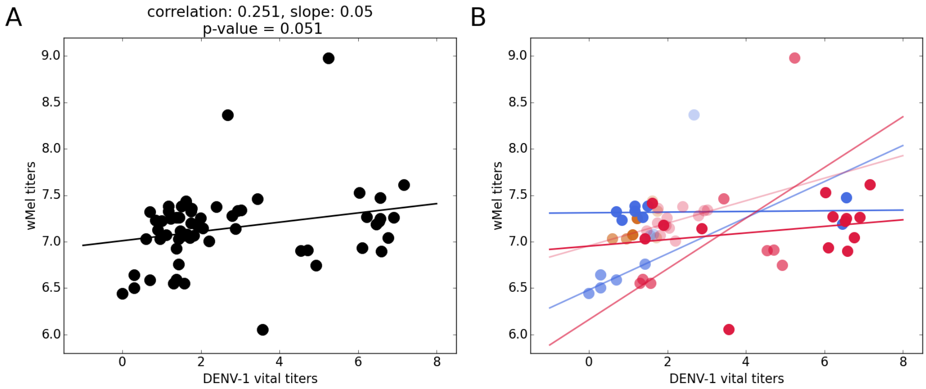


**Figure 3.3.1:** Correlations between *DENV-1*and Wolbachia titers, displayed in log scale for the entire data set (A), and per-dose subsets (B) – color code follows that of the raw data.

In the case of the *w*Mel group viral titer data set (figure 3.2.5) the trends are similar, although the lowest doses seem to have a greater number of data points with detectable titers (at low levels, nevertheless), and the highest seem to increase somewhat more slowly (which supported by a smaller growth, as well as lower mean at the last time point). As for the *Wolbachia* levels, shown in figure 3.2.6, they seem fairly stable and independent of challenge dose or time point; I do not compute any summaries for this data alone.

Because both *DENV-1*and *Wolbachia* titers are available for each individual insect, the correlation between viral and symbiont levels can be computed (figure 3.3.1). We find a significant (p–value $=$ 0.05) but low ($R^2 \approx 0.25$) positive correlation when computing it for the whole data set. Looking at correlations for each dose and time point, I find varying levels of correlation (table 3.3.1) with unclear trend in strength or significance of the correlations. It may be noted that all correlations computed are positive, but the lack of consistency in the significance makes them inconclusive for the most part.

**Table 3.3.1:** Virus/symbiont correlation

| Global | Dose ($TCID_{50}$) | **3** | **7** | **14** | (days post infection) |
|---|---|---|---|---|---|
| 0.25 (0.051) | $\mathbf{10^4}$ | - | - | - | $R^2$ (p–value) |
| | $\mathbf{10^5}$ | - | - | - | |
| | $\mathbf{10^6}$ | - | - | - | |
| | $\mathbf{10^7}$ | - | 0.88 (0.049) | 0.1 (0.81) | |
| | $\mathbf{10^8}$ | 0.50 (0.093) | 0.57 (0.14) | 0.21 (0.474) | |

The lack of clear temporal dynamics in the *Wolbachia* levels and their low correlations with viral levels, gives additional support to the treatment of the symbiont as a factor that can be either present or absent, instead of a dynamic variable of its own.

### 3.3.2 Model predictions

#### Deterministic analysis

In the absence of pathogens, besides the trivial result of all processes in the pathogen equation being zero, i.e. $dP/dt = 0$, the response equation reaches a stable equilibrium as a result of pathogen-independent recruitment $a$ being balanced by pathogen-independent decay (but linearly dependent on the response) $\gamma R$. As a result the baseline response in this model is given by $R^* = a/\gamma$. This stable equilibrium (as denoted by a star henceforth) is therefore given by $(P^*, R^*) = (P^*_{\text{free}}, R^*_{\text{free}}) = (0, a/\gamma)$ (Pujol et al. 2009).

In the original formulation by Pujol et al. (2009) the lack of the $-kP^2$ term causes the pathogen-free equilibrium to be the only stable steady state; the other equilibrium being a saddle point for the range of parameter values explored by the authors. The two solutions to that system are the following:

$$\left(\left(P^*_{\text{free}} = 0; R^*_{\text{free}} = \frac{a}{\gamma}\right), \left(P^*_{\text{saddle}} = \frac{a\delta_P - \gamma r}{r\delta_R - \lambda\delta_P}; R^*_{\text{saddle}} = \frac{r}{\delta_P}\right)\right)$$

The unstable equilibrium defines a threshold between the stable pathogen-free state, and endless growth of the pathogens.

The addition of the logistic-like, square term $(-kP^2)$ limits this unbounded growth to a a maximum carrying capacity, with otherwise little qualitative impact in the other equilibria of the system. The pathogen-free solution is unaltered, but the mathematical form of the two other solutions becomes more complicated, as shown below:

$$\begin{aligned}
&\left(\left(P^*_{\text{free}} = 0; R^*_{\text{free}} = \frac{a}{\gamma}\right),\right.\\
&\left(P^*_{\text{systemic}} = \frac{r\omega - \delta\lambda - \gamma k + \sqrt{(\delta\lambda + \gamma k + r\omega)^2 - 4\delta\omega(ak + \lambda r)}}{2k\omega};\right.\\
&\left. R^*_{\text{systemic}} = \frac{r\omega + \delta\lambda + \gamma k - \sqrt{(\delta\lambda + \gamma k + r\omega)^2 - 4\delta\omega(ak + \lambda r)}}{2\delta\omega}\right),\\
&\left(\hat{P} = \frac{r\omega - \delta\lambda - \gamma k - \sqrt{(\delta\lambda + \gamma k + r\omega)^2 - 4\delta\omega(ak + \lambda r)}}{2k\omega};\right.\\
&\left.\left.\hat{R} = \frac{r\omega + \delta\lambda + \gamma k + \sqrt{(\delta\lambda + \gamma k + r\omega)^2 - 4\delta\omega(ak + \lambda r)}}{2\delta\omega}\right)\right)
\end{aligned}$$

For a biologically meaningful, real-numbered solution, it can be verified that the equilibrium value for the pathogens in the second solution is greater than the last, defining the systemic infection equilibrium $(P^*_{\text{systemic}}, R^*_{\text{systemic}})$, and the intermediate equilibrium $(\hat{P}, \hat{R})$ creating a bistability threshold (where the hat henceforth denotes equilibrium at this unstable values specifically).

Therefore, in the presence of pathogens, and in the range of positive values of interest, the three equilibrium states are possible: pathogen elimination, which is stable and reachable by elimination of all pathogens; the other stable equilibrium being one where pathogens are established. The saddle point, which defines a ratio $\hat{P}/\hat{R}$, at which deterministically speaking, a perturbation in favor of the pathogens tilts the system towards pathogen establishment, and conversely, towards pathogen elimination when the ratio favors the host response. Except at these particular value set $(\hat{P}, \hat{R})$, the separatrix that would deterministically define whether the system goes one way or the other is not defined by this ratio – (see Pujol et al. 2009).

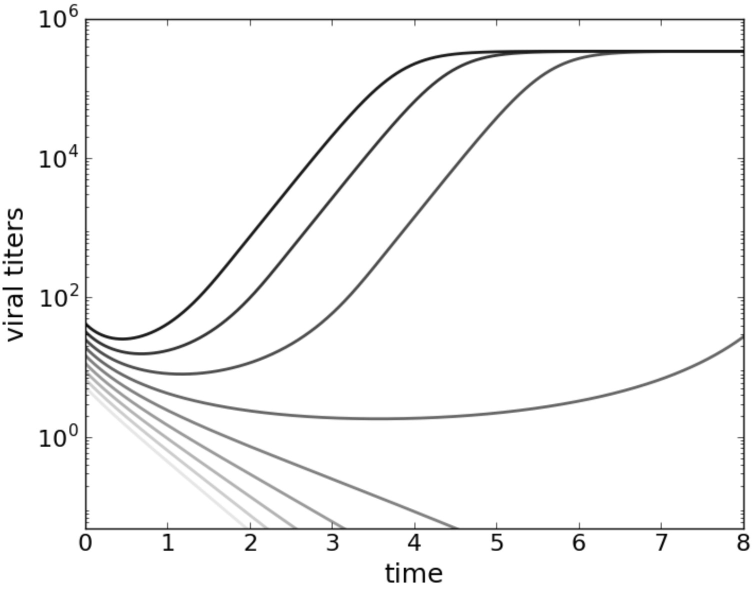


**Figure 3.3.2:** Numerical solutions of the system of differential equations with initial conditions of the response given by their $R^*_{\text{free}}$ value, and and increasing value of P for each darker shade. Parameter values are: $r = 3.5$, $\delta_P = 0.042$, $a = 4.8$, $\gamma = 0.032$, $\delta_R = 0.045$, $\lambda = 0.12$, $k = 10^{-5}$.

These solutions do not determine what is the state of the system at any time; considerations about initial conditions must be made, which can constrain the possible state of the system. In a host that has never seen infection – or that has eliminated any infection long enough before – one can assume the system will be in the pathogen-free equilibrium; introduction of a $P_0$ pathogen inoculum results in the state $(P_0, R^*_{\text{free}})$ as initial condition. Under these constraints, initial conditions are determined by the initial pathogen inoculum alone.

These qualitative features of these solutions can be thought to match those of the data set, so it is a reasonable candidate to explain the data. A graphical example of the behavior predicted by the model under these assumptions is shown in figure 3.3.2, with an increasing initial number of pathogens moving the stable equilibrium from elimination to establishment of the pathogen population.

### 3.3.3 Inference

The experimental data is used to infer parameters for the differential equation model described above. In appendix B a simpler logistic-growth/exponential-decay model is fit to the data; additionally, both models are fit to single doses at a time for methodological checking purposes. Although both models perform somewhat similarly, the pathogen-response model (eqs. 3.1) does not require additional assumptions about initial conditions to produce the entire dynamical spectrum seen in the data.

#### Dynamic pathogen-response model

For the pathogen-response model 3.1, the parameters were estimated from the experimental mosquito challenge data taking into consideration the possible equivalence in some parameters between the two groups. Because *Wolbachia* is generally accepted to have an effect on virus proliferation, a different set of parameters is fitted to the *TET* and *w*Mel data sets; as mentioned above, the symbiont is assumed to affect the response rate parameters.

We take the parameters specific to the pathogen, i.e. the growth rate $r$ and $\delta_P$, to be the same for both groups. We also assume the carrying capacity $k$ is not affected by the presence of *Wolbachia*; that is supported by the observation and comparison of the averages for the lates time point at the highest dose (where carrying capacity is presumed to have been reached), which does not reveal striking differences between the two groups.

Conversely, the response processes are assumed to have different values between the two populations; therefore, the baseline rates recruitment $a$, and natural degradation of the response $\gamma$, as well as the pathogen-dependent recruitment $\lambda$ and degradation rates $\delta_R$ are different between *TET* and *w*Mel population groups.

Initial conditions are also assumed to be equal between groups, since the exact same aliquot and dilution volumes are used for both groups. Therefore, how much each dose is diluted from the highest concentration (which I call the *dosefold* parameter) is fixed at between groups, and so is the highest dose – gamma distributed priors are used for these parameters. Together with the highest concentration $P_0^{\text{high}}$ parameter (equal to $P_0^5$, the fifth and highest dose in this data set), all other initial concentrations ($P_0^1, P_0^2, P_0^3, P_0^4$) can be computed from the two parameters, obtained by simply dividing it by the dilution volume.

Figure 3.3.3A shows a lower infection profile for the $P_0^5$ dose in the *w*Mel group, but higher pathogen levels for the $P_0^4$ trajectory – the lower three doses are essentially zero.

Under the assumptions of the mathematical model, the parameter values have mechanistic biological interpretations. From the posterior distribution of the parameters (figure 3.3.3B), it can be seen that the constitutive recruiting rate ($a$) is smaller in the population with the symbiont, although pathogen-independent decay ($\gamma$) is also lower. Conversely, induced recruiting ($\lambda$) is larger for that population, and in turn pathogen-dependent decay ($\delta_R$) is larger as well.

Generally speaking, the posterior distributions for all parameters are quite narrow, this can either mean that the data is very informative and there is little uncertainty about the "correct" value of the parameters, or that something else is artificially causing the *MCMC* chain to underestimate the uncertainty, possibly affecting the accuracy of the method as well. While it is normally not straightforward to distinguish between the two, even in simple cases or simulation studies with known parameters it is rarely the former. To rule out some gross methodological error, in appendix B the model is fitted to a single dose at a time, which is unlikely to yield very precise estimates, and indeed there is greater uncertainty in the posteriors. Although that confirms that the *MCMC* algorithm is working as

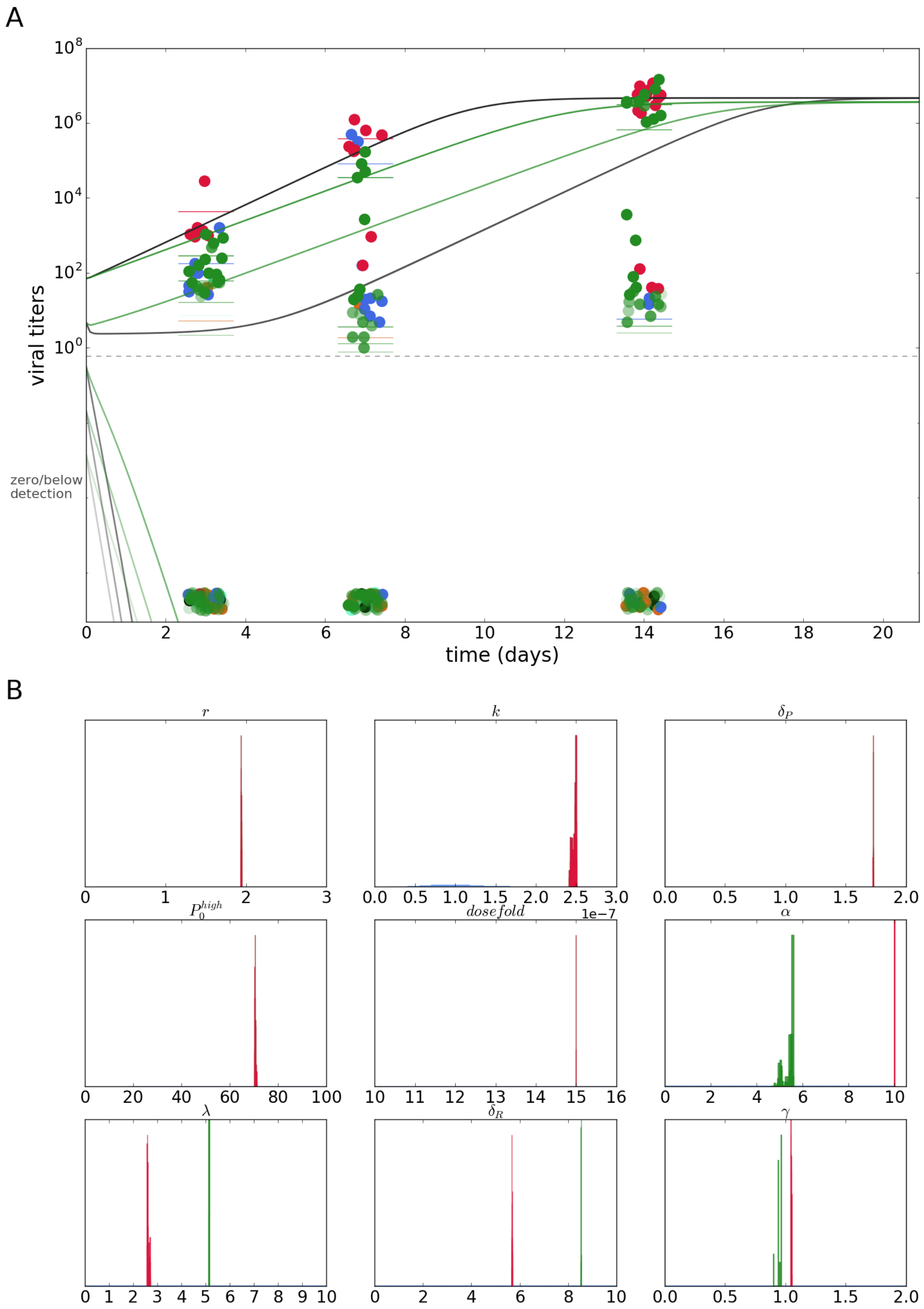


**Figure 3.3.3:** (A) Fit of dynamic host-pathogen model. Grayscale lines represent infection profiles from the *TET* group; green lines are the profiles for *w*Mel group. Green data points are the experimental data for the *w*Mel data, and the remaining the colors the *TET group*. (B) Posterior distribution of estimated parameters. Where parameters are different between groups green color represents the posteriors for the *w*Mel-associated parameters.

intended, it does not exclude that for the entire dataset the inferred values are being artificially constrained – this possibility is further considered in the next subsubsection and discussion section.

In any case, from panel A of the same figure, still, it is clear that the model does not fit the data averages for the second highest dose. Considering the issue of small numbers for the last time point of that dose, and the somewhat unexpectedly low average level, we suspect that may be causing the lack of fit and distorting the trajectory and, ultimately, the model parameters. Therefore, we also fit the model to a subset of the data set that does not include that dose/time point (dose $10^7$ $\mathrm{TCID}_{50}$, 14 d.p.i.).

Figure 3.3.4 shows the results for this subset of the data. Panel A shows a generally lower profile of infection progression along time for the two top doses for the *w*Mel group (for the two lowest prediction of both populations is just zero, or vanishingly small). The middle dose estimate is low for *TET* and nearly steady until around day 10 post infection, decreasing slowly afterwards. It is also near the threshold, where stochasticity is most important in determining whether fixation or elimination occurs for this mosquito population. For the *w*Mel group the middle dose is essentially zero from very early.

The posterior estimates of the underlying parameters of the model for both groups are shown in figure 3.3.4B, and again narrow distributions are obtained. The qualitative relationships between the parameters are maintained, although the actual values change considerably.

Despite the qualitative relationships being maintained, the choice to remove a condition where average titers are lower in the *TET* group is *ad hoc*, and may be seen to bias results towards *Wolbachia* protection; therefore, in appendix B the same analysis is performed on a further reduced subset of the data where a condition in the opposite direction is removed. The results are consistent with the first subset results (figure 3.3.4), not the entire data set (figure 3.3.3). These subset analyses, as well as approaches that eliminate the need for any *post hoc* analysis are further considered next section and in the chapter discussion.

Besides the subsets, inference with alternative sets of free parameters are also performed

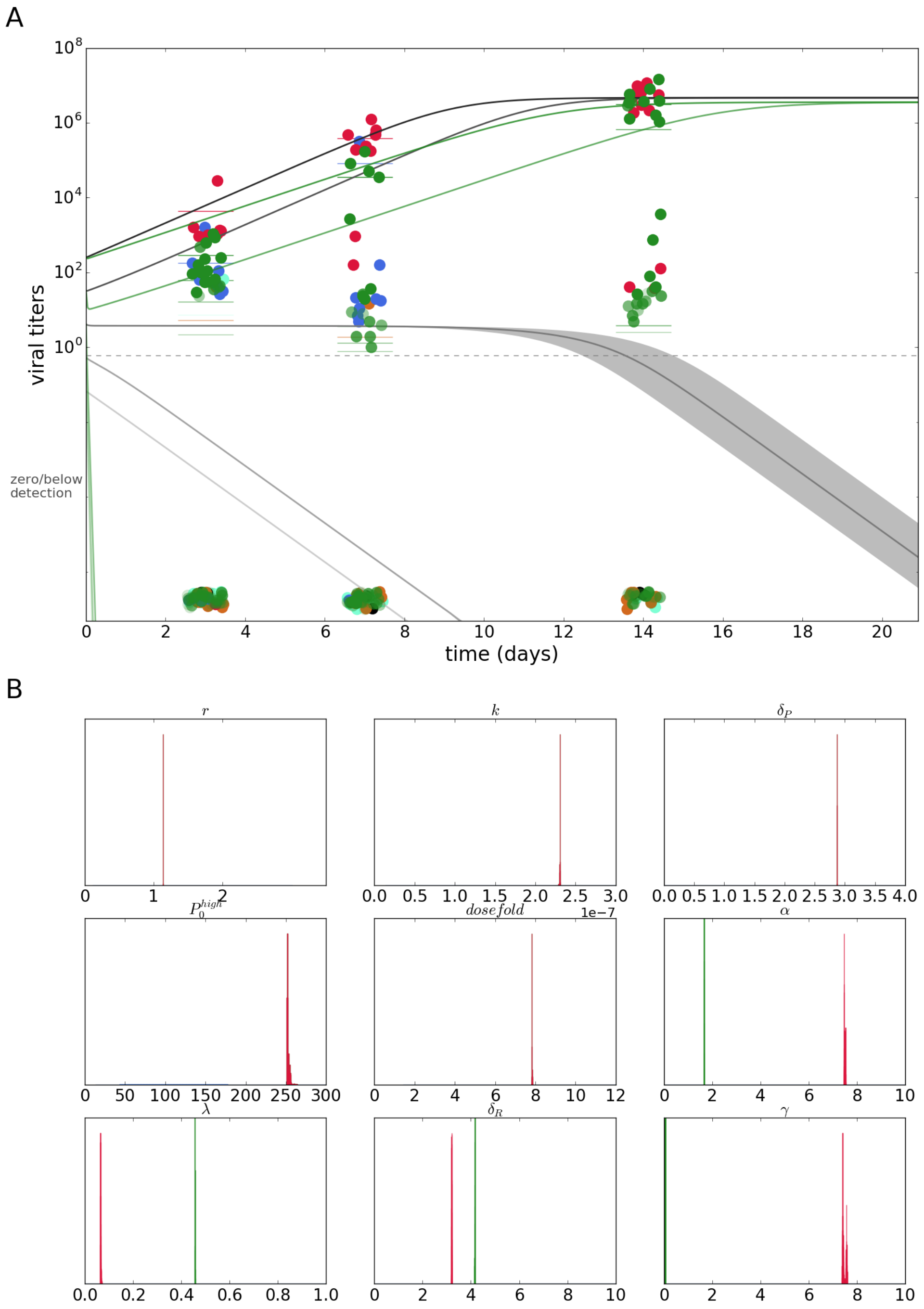


**Figure 3.3.4:** (A) Fit of dynamic host-pathogen model to data subset. (B) Posterior distribution of estimated parameters. Colors are as before.

in appendix B. In sum, initial conditions are allowed to vary and/or pathogen-independent processes are fixed between groups. Overall, results are consistent, with the fit being mostly dependent on the data subset being considered, and the parameter relationships being maintained for any choice of parameter set. Because the main objective of the analysis is the broad comparison between insect populations, which is consistent for the different models, information-theoretic or likelihood-based comparisons of the models are not performed.

Taken together, inference suggests that in the *w*Mel group there are fewer pathogens that go against a stronger induced immune response, thus supporting the hypothesis that *Wolbachia* confers protection to *Aedes aegypti* hosts against *DENV-1*. The protection is observed mostly for higher doses, so the degree of protection is not of a fixed-proportion proportion kind, but is rather described as a property emerging from the dynamic model, possibly with negligible or even incrased risk for intermediate doses.

### Stochasticity in the system state

Stochasticity is expected, and will, cause the system to drift from the deterministic prediction, so the equilibria described are not guaranteed to be reached as a function of initial conditions alone. Stochastically, but influenced by the deterministic vector field, the system will then fluctuate along time either to pathogen elimination or establishment.

An illustration of this stochastic trajectories is shown in figure 3.3.5A. Panel B shows data points sampled from the stochastic simulation at specific time points together with the deterministic output for the same doses – from that panel it looks like the deterministic model could capture the mean behavior of a an ensemble of stochastic runs. A clearer visual comparison, however, would be taking a poisson samples from the deterministic prediction – that is shown in panel C. It is clear that the variation expected by the latter model is a lot smaller than that of the former.

To be clear, in the discrete stochastic model, the noise propagated along time results in a greater variance in the observed levels than that assuming a deterministic trajectory, and

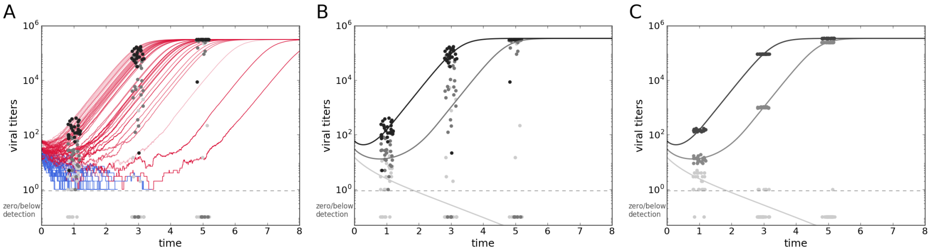


**Figure 3.3.5:** Full stochastic trajectory of the pathogen-response model with data points shown at chosen time points (A). Deterministic trajectory overlaid to data points from stochastic model (B). Deterministic trajectory with data points sampled from a poisson distribution that uses deterministic prediction as distribution parameter (C). Parameter values are: $r = 3.5$, $\delta_P = 0.042$, $a = 4.8$, $\gamma = 0.032$, $\delta_R = 0.045$, $\lambda = 0.12$, $k = 10^{-5}$, $P_0 = 15$, $dosefold = 2$. Time discretization is $5 \cdot 10^{-4}$ days.

at any time point assuming an error model around it. This illustrates the fact that stochastic implementations of dynamic models are parametric descriptions of more complex, time-dependent distributions, for which closed forms may not be available. This process can be analytically shown to represent the likelihood, making it a formal (although not always practical) tool for inference.

Under this model, bistability will tend to increase variance along time, as trajectories above the bistability threshold tend to go further up, and the ones that go below are attracted to zero. One important prediction in such a system is that both high levels of pathogens and zero/undetectable levels are expected to be seen for replicates with the same initial conditions – especially if they are near the bistability threshold.

Increasing the initial pathogen inoculum, or dose, should increase the proportion of individuals with established infections, and reduce the variance in the pathogen levels, just as reducing doses will result in observed levels that are increasingly close or completely at the value of zero. Those are features that can be broadly observed in the *Ae. aegypti* data set (as well as in the *D. melanogaster* preliminary data).

It is worth noting that – for these specific parameters – the simulation shown can take minutes on a desktop computer, as opposed to a fraction of second for the continuous model. If finer time discretization is needed the computational demand will be higher, and it may be infeasible with methods like the Gillespie algorithm (Gillespie 1977).

In any case, the effects of stochasticity in this system are quite striking, and can account for data points that would otherwise be considered outliers, or very incompatible with a deterministic model.

## 3.4 Discussion

The difference in viral titers between two groups cannot be captured by a single arbitrary experimental condition; furthermore, variation between experiments may cause equivalent conditions to have seemingly inconsistent results, but which can in fact be explained by a neglected variable, such as time or dose (Gomes et al. 2014).

Here, for any one group, the viral levels can also clearly be seen to be dose dependent. Taking the dose dimension into consideration in infection experiments has been proposed to be important for obtaining generally applicable parameters (Gomes et al. 2014; Pessoa et al. 2014), and that idea has later been used to get more reliable estimates in *Aedes-dengue-Wolbachia* systems (Ferguson et al. 2015). A temporal trend is also clear, which has long been recognized and is the essence of dynamic models (Antia et al. 1997) – that has recently been also applied to dengue with convenience samples from human patients and more traditional dynamic models (Clapham et al. 2014). By considering a range of doses from harmless to almost-certain infection, as well as a time period from inoculation until establishment, or clearance, of infection, local inconsistencies can be interpreted as minor shifts between the extremes, which do not change the overall picture. The combination of these features provides a more comprehensive understanding of the process of infection.

This kind of data can be interpreted under a dynamic mathematical model with varying initial conditions, as shown by the results in this chapter. Most importantly, the use of a model also allows a full dynamic profile to be constructed through inference from a data set of a reasonable size, as opposed to such a data intensive experiment unfeasible in practice.

The conclusions presented here generally agree with the literature (Hedges et al. 2008; Moreira et al. 2009; Teixeira et al. 2008); however, individual conditions also show the fragility of a typical experimental design. In that context, others have assayed different time points in the hope that both conditions would agree, and the effect would not seem condition-specific; however, if results are contradictory, no conclusions can be drawn from that approach. Under the model proposed here, temporal dynamics have a clear meaning. The profile of infection in the population with *Wolbachia* is always below that of its *TET* counterpart for the same dose. The biological interpretation is that *w*Mel protects *Aedes aegypti* mosquitoes against *DENV-1* through a resistance mechanism that reduces the level of infection, and as a result reduces the risk of a full-blown infection.

Besides the model-predicted viral dynamics, the biological mechanism of the virus protection can be interpreted through the underlying parameters of the model, which can be compared between the two groups. In the *Wolbachia*-carrying population the rate of recruitment of the response ($\lambda$) is estimated to be higher, while the rate of pathogen-induced decay of the response ($\delta_R$) is larger – the former is robust to alternative model parameterizations (table B.6.1, appendix B); the latter is not. In the case of constitutive responses recruitment ($a$) as well as decay ($\gamma$) are lower for the *w*Mel group. Despite some estimates favoring one or other group, the general infection profile is mostly unchanged under alternative model analyses (appendix B).

Accepting a model implies accepting its assumptions, caveats, constructs, and artifacts. By reducing the biology of a highly structured and complex process to a two-equation model much is guaranteed to be lost; nevertheless, there is often a practical limit to how much detail can be included Zwietering (2009). One of the main arguments of this chapter is that traditional analyses can be simplistic and overly reductive; that is not to say that there are no caveats in more sophisticated approaches, and that all results are immediately interpretable as the actual biology of the system. For instance, in previous studies (in a different insect system) it was observed that viral titers were undetectable immediately after inoculation with any dose (Zwerschke et al. unpublished); nevertheless, the model can predict that initial levels are higher than some of the lowest detectable levels. This implies that the $P_0^i$ parameters are a construct from the models (whether it is the pathogen-response, logistic, or the simple regression, all the same); the estimates may be useful, but cannot be interpreted directly as virus concentration at that very early time point.

As always, validation of the claims made here require replication of the experiment, repetition under more natural conditions (like oral infection of mosquitoes), as well as independent assays (when possible) of the parameters estimated by the model. Inference with mathematical models provide estimates of an entire parameter set, subject to the constraints and limitations of the model, and the parameters may or may not have directly measurable quantities that can be independently validated.

While it is clear that the time profile for any single dose can be well described by the pathogen-response model (as well as the logistic model in the appendix B), the entire data may have patterns that are more difficult to describe – like the unusually narrow posteriors – and may point to failures of this particular models, or other aspects of the inference framework. Due to the lack of alternative mechanistic models that describe the broad features of the data, and the difficulty of developing new models from scratch, model selection criteria were not applied here. Simple logistic growth/exponential decrease does not seem to describe the data better than the pathogen-response model; nevertheless, some aspects of the data set seem to be out of the reach of the latter.

One possible direction for improvement is to account for stochasticity in the dynamics (Andrieu & Doucet 2010). The ability of the stochastic version of the model to generate distributions of viral titers with greater variance than its deterministic counterpart can be easily illustrated by a forward simulation. The actual implementation of a stochastic inference method is not as simple as plugging a stochastic simulation model into the numerical solution portion of the inference procedure, and requires a more sophisticated method to track the dynamic system and integrate stochasticity out (Ionides et al. 2006). In the context of *MCMC*-based inference, where simulation parameters are unknown and many iterations are needed, computational costs may be prohibitive, and steep improvements and optimization are probably required. Nevertheless, it is promising for tackling issues with over-dispersed data, which is often the case.

Another avenue for future research is the inclusion of inter-individual variation in key response parameters, such as $\lambda$ and $\delta_R$, which can be mistaken for intra-individual stochasticity (Kendall & Fox 2002). Given the results in chapter 1, asserting the origin of a more heterogeneous distribution of susceptibility is a logical priority. The alternative hypotheses for the mode emergence of this heterogeneity are further speculated on in chapter 5.

Improvements on the experimental side are also possible, the first being the preventable loss of mosquito lives as a consequences of a design flaw inherited from traditional time point experiments: in a typical time point experiment, specific times are predefined, data is recorded only then, so the readouts can be compared for the same time points. Here, since the data points are fitted to a model that has an observable value at any time point, the recording time of the data is essentially arbitrary. Because mosquitoes are monitored and fed every day (and could be monitored every few hours during the day if necessary), moribund mosquitoes could be identified on time, and instead of letting mosquitoes die and discarding them they could be frozen and have their viral levels recorded at that time. Because we did not anticipate a great (or small) loss of mosquitoes, the experiment was not designed to maximize the number of data points, as this would require additional monitoring and storing effort (e.g. storing every mosquito individually and checking them several times a day).

The pathogen-response model produces an inferred dynamic profile of both pathogens and immune response. An independent validation of the model, could therefore be made by correlating the inferred response to a real measure of immune response. However, it must also be observed that the model implementation is an abstraction that aggregates all host immunity into a single differential equation, so a straightforward correlation between that and any proxy for immune response is not trivial. It is also possible that the response does not correlate with a measure of immune response but rather on a combination of factors involved in resistance and tolerance mechanisms, either with and without interference of *Wolbachia*.

Although it is best to eliminate any *ad hoc* treatment, clear lack of fit and low number of data points forced an analysis of a subset of the data. The work present here attempted multiple improvements with respect to data generation, model implementation and inference, but by no means exhausts all possibilities in any one of those components. Some of the weaknesses of the work in this chapter can be addressed by repetition of the experiment, or implementation of more sophisticated methods, none of which are – unfortunately – feasible without considerable more time and effort that can no longer be put into this thesis.

In sum, there are recognized possible improvements, as well as caveats; nevertheless, the results presented here provide the basic framework for a dose- and time-dependent explanation of viral establishment and elimination dynamics. The biological interpretations based on the model agree with simpler assays, and expand the scope of analysis of this tripartite system.

# References

1. Alizon, S. & van Baalen, M. 2008 Acute or Chronic? Within-Host Models with Immune Dynamics, Infection Outcome, and Parasite Evolution. The American Naturalist 172, E244–E256. (doi:10.1086/592404)

2. Andrieu, C., Doucet, A. & Holenstein, R. 2010 Particle Markov chain Monte Carlo methods. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 72, 269–342. (doi:10.1111/j.1467-9868.2009.00736.x)

3. Antia, R. & Lipsitch, M. 1997 Mathematical models of parasite responses to host immune defences. Parasitology 115, 155–167. (doi:10.1017/S003118209700200X)

4. Antia, R., Koella, J. C. & Perrot, V. 1996 Models of the Within-Host Dynamics of Persistent Mycobacterial Infections. Proceedings of the Royal Society B: Biological Sciences 263, 257–263. (doi:10.1098/rspb.1996.0040)

5. Baas, J., Jager, T. & Kooijman, B. 2010 Understanding toxicity as processes in time. Science of The Total Environment 408, 3735–3739. (doi:10.1016/j.scitotenv.2009.10.066)

6. Clapham, H. E., Tricou, V., Van Vinh Chau, N., Simmons, C. P. & Ferguson, N. M. 2014 Within-host viral dynamics of dengue serotype 1 infection. Journal of the Royal Society, Interface / the Royal Society 11, 20140094–20140094. (doi:10.1098/rsif.2014.0094)

7. Consoli, R. A. G. B. & Lourenço-de-Oliveira, R. 1994, Editora

Fiocruz. Principais Mosquitos de Importância Sanitária no Brasil. (doi: http://dx.doi.org/10.7476/9788575412909)

8. Ferguson, N. M. et al. 2015 Modeling the impact on virus transmission of Wolbachia-mediated blocking of dengue virus infection of Aedes aegypti. Sci Transl Med 7, 279ra37–279ra37. (doi:10.1126/scitranslmed.3010370)

9. Frank, S. A. 1996 Host-Symbiont Conflict over the Mixing of Symbiotic Lineages. Proceedings of the Royal Society B: Biological Sciences 263, 339–344. (doi:10.1098/rspb.1996.0052)

10. Franz, A., Kantor, A., Passarelli, A. & Clem, R. 2015 Tissue Barriers to Arbovirus Infection in Mosquitoes. Viruses 7, 3741–3767. (doi:10.3390/v7072795)

11. Frentiu, F. D., Zakir, T., Walker, T., Popovici, J., Pyke, A. T., van den Hurk, A., McGraw, E. A. & O'Neill, S. L. 2014 Limited Dengue Virus Replication in Field-Collected Aedes aegypti Mosquitoes Infected with Wolbachia. PLoS neglected tropical diseases 8, e2688. (doi:10.1371/journal.pntd.0002688)

12. Gelman, A., Carlin, J. B., Stern, H. S. & Rubin, D. B. 2013 Bayesian data analysis, Third Edition. CRC Press. (doi:10.1080/01621459.2014.963405)

13. Gillespie, D. T. 1977 Exact stochastic simulation of coupled chemical reactions. J. Phys. Chem. 81, 2340–2361. (doi:10.1021/j100540a008)

14. Gomes, M. G. M., Lipsitch, M., Wargo, A. R., Kurath, G., Rebelo, C., Medley, G. F. & Coutinho, A. 2014 A Missing Dimension in Measures of Vaccination Impacts. PLoS Pathog. 10, e1003849. (doi:10.1371/journal.ppat.1003849)

15. Hamilton, R., Siva-Jothy, M. & Boots, M. 2008 Two arms are better than one: parasite variation leads to combined inducible and constitutive innate immune responses. Proceedings of the Royal Society B: Biological Sciences 275, 937–945. (doi:10.1098/rspb.2007.1574)

16. Hedges, L. M., Brownlie, J. C., O'Neill, S. L. & Johnson, K. N. 2008 Wolbachia and Virus Protection in Insects. Science 322, 702. (doi:10.1126/science.1162418)

17. Ionides, E. L., Bretó, C. & King, A. A. 2006 Inference for nonlinear dynamical systems. Proc Natl Acad Sci USA 103, 18438–18443. (doi:10.1073/pnas.0603181103)

18. Kendall, B. E. & Fox, G. A. 2002 Variation among Individuals and Reduced Demographic Stochasticity. Conservation Biology 16, 109–116. (doi:10.1046/j.1523-1739.2002.00036.x)

19. Moreira, L. A. et al. 2009 A Wolbachia Symbiont in Aedes aegypti Limits Infection with Dengue, Chikungunya, and Plasmodium. Cell 139, 1268–1278. (doi:10.1016/j.cell.2009.11.042)

20. Patil, A., Huard, D. & Fonnesbeck, C. J. 2010 PyMC: Bayesian Stochastic Modelling in Python. Journal of statistical software 35, 1–81.

21. Perelson, A. S. & Oster, G. F. 1979 Theoretical studies of clonal selection: minimal antibody repertoire size and reliability of self-non-self discrimination. Journal of Theoretical Biology 81, 645–670.

22. Pessoa, D., Souto-Maior, C., Gjini, E., Lopes, J. S., Ceña, B., Codeço, C. T. & Gomes, M. G. M. 2014 Unveiling Time in Dose-Response Models to Infer Host Susceptibility to Pathogens. PLoS Comput Biol 10, e1003773–9. (doi:10.1371/journal.pcbi.1003773)

23. Pujol, J. M., Eisenberg, J. E., Haas, C. N. & Koopman, J. S. 2009 The Effect of Ongoing Exposure Dynamics in Dose Response Relationships. PLoS Comput Biol 5, e1000399–12. (doi:10.1371/journal.pcbi.1000399)

24. Sasaki, A. & Iwasa, Y. 1991 Optimal growth schedule of pathogens within a host: Switching between lytic and latent cycles. Theoretical population biology 39, 201–239. (doi:10.1016/0040-5809(91)90036-F)

25. Teixeira, L., Ferreira, Á. & Ashburner, M. 2008 The Bacterial Symbiont Wolbachia Induces Resistance to RNA Viral Infections in Drosophila melanogaster. PLOS Biology 6, e1000002. (doi:10.1371/journal.pbio.1000002)

26. Zwerschke, D., Teixeira, L. et al. Quantitative analysis of Wolbachia protection against viruses (unpublished manuscript).

27. White, S. M., Burden, J. P., Maini, P. K. & Hails, R. S. 2012 Modelling the within-host growth of viral infections in insects. Journal of Theoretical Biology 312, 34–43. (doi:10.1016/j.jtbi.2012.07.022)

28. Zwietering, M. H. 2009 Quantitative risk assessment: Is more complex always better? Simple is not stupid and complex is not always more correct. International Journal of Food Microbiology 134, 57–62. (doi:10.1016/j.ijfoodmicro.2008.12.025)

*Markov Chain Monter Carlo is a great method for losing your hair.**

Simon Tavaré

# B

# Additional inference and validation of within-host models

## B.1 Mosquito numbers and subgrouping data tables

As mentioned in the main text, a total of 616 mosquitoes were divided by symbiont status (absent, *TET* colony, and present *w*Mel) into two groups of approximately half that size, and challenged with five different *DENV-1* doses plus mock-infected controls injected with culture medium. These numbers are shown in table B.1.1.

The mosquitoes actually assayed through qPCR are also shown classified by days post infection (d.p.i.), as well as in aggregate. The difference between the latter and the total number of challenged mosquitoes includes all causes: natural mortality, virus-induced mortality, loss due to freezing, unfreezing, and any other aspects of experimental handling

**Table B.1.1:** Number of mosquitoes challenged and assayed for viral titers

| | **dose / time** | **$10^4$** | **$10^5$** | **$10^6$** | **$10^7$** | **$10^8$** | **ctrl** | **(TCID$_{50}$)** | **total (column sum)** |
|---|---|---|---|---|---|---|---|---|---|
| TET challenged | **all** | **54** | **51** | **55** | **49** | **57** | **38** | | **304** |
| TET qPCR assayed | 3 d.p.i. | 9 | 3 | 8 | 12 | 8 | 8 | | 48 |
| | 7 d.p.i. | 12 | 3 | 8 | 10 | 9 | 1 | | 43 |
| | 14 d.p.i. | 4 | 3 | 7 | 6 | 14 | 3 | | 37 |
| | **all (row sum)** | **25** | **9** | **23** | **28** | **31** | **12** | | **128** |
| *w*Mel challenged | **all** | **51** | **57** | **55** | **54** | **55** | **40** | | **312** |
| *w*Mel qPCR assayed | 3 d.p.i. | 12 | 11 | 9 | 9 | 13 | 11 | | 65 |
| | 7 d.p.i. | 6 | 10 | 10 | 10 | 10 | 6 | | 52 |
| | 14 d.p.i. | 7 | 4 | 8 | 10 | 14 | 6 | | 49 |
| | **all (row sum)** | **25** | **25** | **27** | **29** | **37** | **23** | | **166** |

of the mosquitoes.

Despite the apparently heavy losses in mosquitoes, given the multiplicity of factors causing them, I do not speculate further on specific effects responsible for the losses, and assume they are random for al practical purposes. Limiting losses can be accomplished by taking advantage of the analysis framework and through a more individualized treatment of mosquitoes, which nevertheless implies more intensive work – these are detailed in the discussion section of the main text.

## B.2 Logistic growth/exponential decline model

A logistic growth model is considered to describe simple exponential viral titer growth with saturation at some carrying capacity in the doses where levels are seen to increase.

Therefore there are two parameters to describe one curve. If pathogen levels decrease, growth rate is negative, and the carrying capacity is irrelevant, resulting in an model of exponential decrease. The model can be described as a simple differential equation (and an explicit solution for it also exists):

$$\frac{dP}{dt} = rP - kP^2$$

Despite the logistic-like model not having any mechanistic biological interpretation, it is a simpler, more predictable model that can be used to test the ability of simple dynamic to fit the data.

Because there is no single set of parameters for all five curves of each of the data sets, there are at least a couple of options to fit all curves: either one of the two models (two-parameter logistic or one-parameter exponential, depending on growth or decline characteristics) is used for each curve, which for the total of 5 doses results in between 5 and 10 parameters, or some assumption is used to reduce the total number of parameters. Here, it is assumed that the positive growth parameter as well as the carrying capacity are the same for all doses. The growth rate is reduced by a dose dependent factor $m_i$ (equation B.1) where *i* represents the *ith* dilution.

Intuitively, carrying capacity can be assumed not to change as a function of challenge dose, as it is a property of the host when viral titers become high enough. On the other hand, the growth and the dose-dependent reduction can be treated in a range of ways, from the reduction factor being a free parameter for each dose, to it being a single parameter representing a constant multiplicative reduction as the dose gets lower, i.e. the for the highest, undiluted dose $m_0 = 0$ and growth is given by $r$, for the first dilution $m_1 = m$, for the second $m_2 = 2m$, and so on. I assume this formulation as the null model because of the low number of free parameters: 3, not considering the initial conditions, which may need to be estimated as well:

$$\frac{dP}{dt} = (r - m_i)P - kP^2 \tag{B.1}$$

This model is considered as a simple baseline description of the main features of the

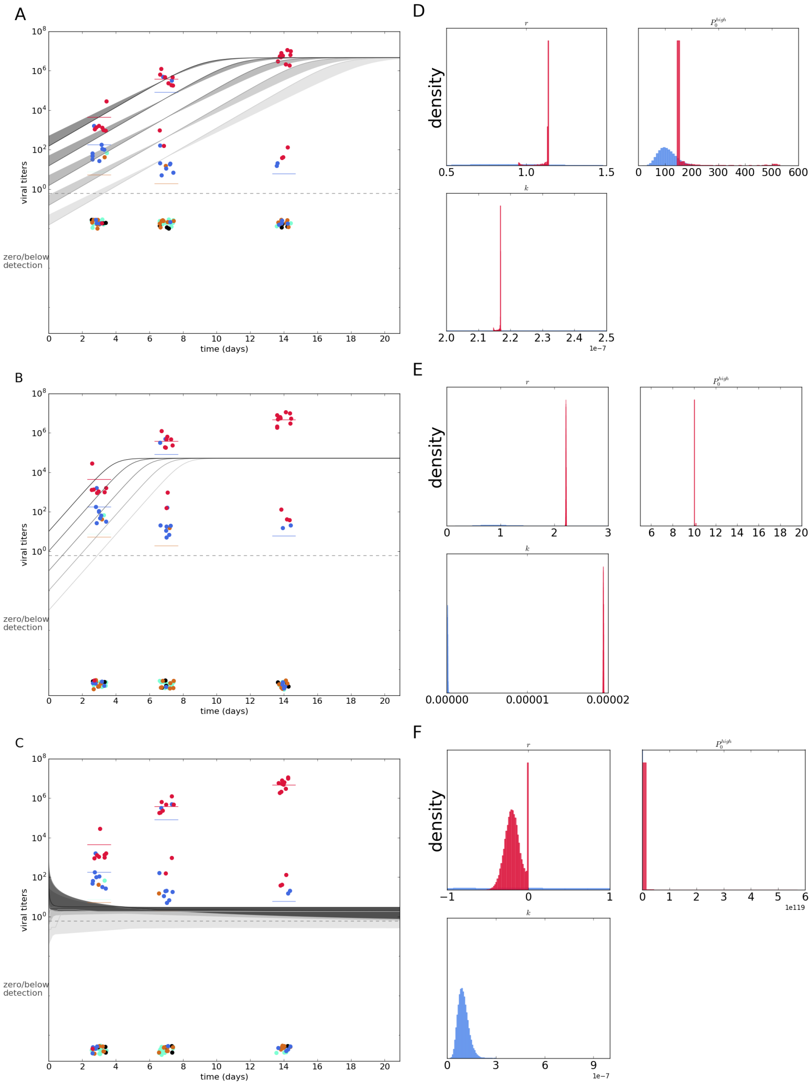


**Figure B.2.1:** Fit of the logistic model to the three highest doses, (A,B,C, starting from the highest), and the posterior distributions of the parameters (D,E,F).

data with the fewest number of parameters I can think of.

## B.3 Single-dose, time course fit

As a first test, single doses can be used to test the ability of the models to fit the data. In the main text a linear regression is performed for the logarithmic transformation of the first two time points of the highest dose; here, logistic growth is fit to the entire time range of the three highest doses.

The fit and confidence intervals of these preliminary fits are shown in figure B.2.1 for the logistic model and figure B.3.1 for the pathogen-response model.

It is necessary to point out that while five curves are shown, only one of them is supposed to fit the data, so in figure B.2.1A the top curve fits the crimson data points, in figure B.2.1B the second highest curve should fit the blue data points, and so on. All other curves are just extrapolations from the single-dose parameter estimates, that nevertheless illustrate the need for multiple-dose time course data to describe the possible infection profiles.

It is also worth noting that what is expected of a good fit is not necessarily that the model passes right on the horizontal lines indicating the means. While this is desirable for the higher doses, where the large numbers would cause the poisson distribution to become normal-like, for lower doses the distribution can be skewed towards zero. Therefore, this alone does not suggest lack of fit. On the other hand, the data is only displayed in base 10 logarithmic scale, so the spread, or more formally the observed variance of the data cannot generally be grasped by a look at the whole data set, but instead deserves a more systematic treatment.

That said, both models can produce the broad patterns of the data, fitting the highest dose quite well (in this case going right through the mean values). For the second highest dose the increase profile is also generally described by either model, although neither model is able to reproduce the decrease in viral levels at the last time point. As I have pointed out in the description of the stochastic pathogen-response model, these low titers can be expected even for high doses, but they are not something that can be predicted from the deterministic trajectory. Also, because I have few data points and low titers on day 14, the carrying capacity can not be well estimated from this dose alone. Finally, for

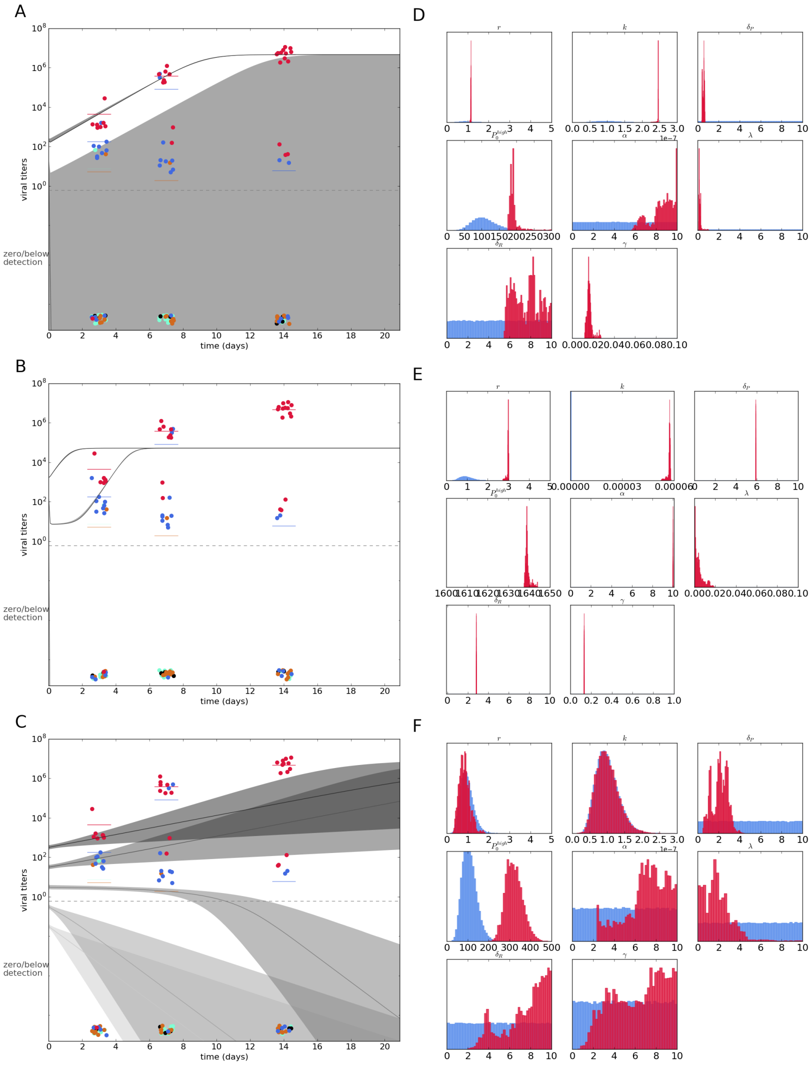


**Figure B.3.1:** Fit of the pathogen-response model to the three highest doses, (A,B,C, starting from the highest), and the posterior distributions of the parameters (D,E,F).

the middle dose, the two models produce somewhat different decrease profiles. Because most data points are zero it is not straightforward to evaluate the best fit. Overall, these inferences show that either model could reproduce general aspects of any single dose.

## B.4 Logistic/exponential model inference

The most important question at this point would be what is the performance of the models to the entire data set. How either model performs under this additional constraint is probably not obvious from the single-dose fits. Here I make the decision to exclude the last time point for the second highest dose in the *TET* group; although it is not something that can be done arbitrarily and without regard to its impact on the analyses, the arguments to do so are mutlifold. Neither model can reproduce the increase-then-decrease shape of the curve, so that neither should *a priori* favored by this, and I expect that there is more information in the other time points and the rest of the data set to make up for this in an unbiased way.

As for the comparison between *wMel* and *TET* groups, it is likely that removing a dose-time point where the mean value for the *TET* group is lower could bias the estimates to those showing *Wolbachia* do protect mosquitoes. Nevertheless, this is the only dose-time point where the means are lower for this group in the two highest doses (in the lower doses the levels being very low anyway). Furthermore the number of replicates happened to be low for this dose-time point, and seemed to contradict a number of previous separate and independent experimental challenges with similar doses for the same two mosquito poulations.

I therefore feel that the confusing effect of an outlier would be worst than any bias removing it could induce. Anyhow, as an analysis control I also make the same analyses removing the second time point (7 d.p.i.) for the second highest dose in the *wMel* group, where it is lower for this group. I also note, that this should not be an issue if a stochastic model is used to fit the data, for instance using a particle filtering algorithm; nevertheless, due to the additional complexity and sophistication of combining that with an MCMC bayesian estimation for such a complex and non-standard model, I do not pursue this here.

That said, the fit of the logistic growth model (Eq. B.1) is shown in figure B.4.1.

The highest dose seems well fit by the simple and familiar shape of logistic-like growth,

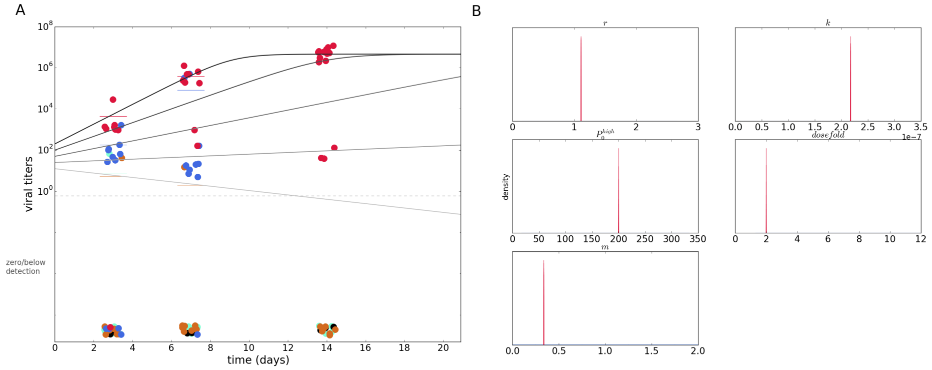


**Figure B.4.1:** Fit of varying-growth logistic growth/exponential decrease model (A, left), and posterior distribution of estimated parameters (B, right).

as expected; the other doses at least visually do not seem to fit as nicely, the second highest dose shows a near linear growth (on log scale). The middle dose shows increase, instead of the apparent decrease shown by the data, and only the lowest dose shows a clear trend of exponetial decrease (seen as a straight line in the semilog plot). Looking at the distribution of posterior estimates, the histograms are quite narrow. It is noteworthy that the initial conditions seem not to be well estimated, with the $P_0^{\text{high}}$ value apparently grossly overestimated, and the fold difference between doses seemingly underestimated. Once again we alert to the distortions of using a visual guide only as criterion to determine goodness of fit; nevertheless, some striking indications of lack of fit are present.

## B.5 Subset data analysis

The pathogen plus response model is given by the two differential equations, respectively:

$$\begin{aligned}\frac{dP}{dt} &= rP - \delta_P PR - kP^2 \\ \frac{dR}{dt} &= a + \lambda P - \gamma R - \delta_R PR\end{aligned} \tag{B.2}$$

Parameters are $r$ pathogen-growth rate, $k$ parameter governing saturation of growth, $\delta_P PR$, response-dependent elimination. Also $a$, constitutive response recruitment rate,

$\gamma R$, natural response decay, and their pathogen-dependent counterparts $\lambda P$ and $\delta_R PR$.

As exemplified by the stochastic implementation of the pathogen-response model, the compound stochasticity along time cannot be fully accounted by the simplification provided by a continuous model with a standard parametric error model such as a normal or poisson distribution around a deterministic prediction. Therefore, even if the correct error model is known, the distribution of data points is bound to be quite different from these parametric distributions by the simple virtue of being augmented in time.

Given the huge computer power that could be necessary to implement inference using stochastic models, the error incurred here by using a continuous model, and the consequence of trying to assess outliers on an *ad hoc* basis is evaluated here by performing the same analysis on the full data set (shown in the main text), and on a different subset of it.

Another subset of the data is tested here by removing an experimental condition showing the opposing trend to that of the main text. The second highest dose ($10^7$ TCID$_{50}$) at the second time point (7 d.p.i.) is considerably lower on average for the *w*Mel group, so removing it should have the opposite effect on the inference as removing the previous one; nevertheless, the general trend in the data for this dose is not as odd as the latter in that the trend could be explained by the model. The results for inference with this subset of the data are shown in figure B.5.1.

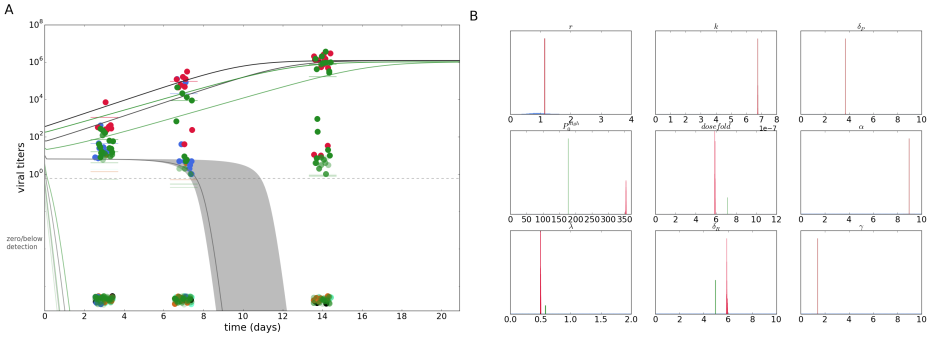


**Figure B.5.1:** (A) Fit of dynamic host-pathogen model for a subset of the data not containing dose $10^7$, 14 d.p.i for the *TET* group, and $10^7$, 7 d.p.i for the *w*Mel group. (B) Posterior distribution of estimated parameters (green is specific to wMel).

In general, these *ad hoc* subdivisions of the data set show that the results are robust to missing data. It therefore seems that the general *Wolbachia* protection is a robust feature of the data set; nevertheless, ideal treatment should include representative distributions for all conditions and adequate treatment of survivor biases, besides the methodological improvements discussed in the main text.

## B.6 Alternative models and submodels

**Table B.6.1:** Shared and distinct parameters for alternative models

| | $\lambda$ $\delta_R$ | $a$ | $\gamma$ | $P_0$ | *dosefold* | |
|---|---|---|---|---|---|---|
| **model 1** | | $\neq$ | $\neq$ | $=$ | $=$ | **(main text)** |
| model 2 | $\neq$ | $=$ | $=$ | $\neq$ | $\neq$ | |
| model 3 | | $=$ | $=$ | $=$ | $=$ | |

In the main text initial conditions and pathogen-associated variables are fixed between the two experimental groups, while response-related parameters are set to be different between groups, in an attempt to minimize assumptions and choose a standard model. Unfortunately, there is no such thing, so it is often wise to consider alternative models or submodels to assess robustness and consistency in the results. Table B.6.1 shows the parameters shared between groups or different ($=$ and $\neq$ signs, respectively). The host-associated $\lambda$ and $\delta_R$ parameters are different for all models, and all pathogen associated parameters are always equal for the two groups, and are omitted for clarity.

In alternative model 2, initial conditions are allowed to vary between the two groups, while $a$ and $\gamma$ are fixed between them. The results for that model are shown in figure B.6.1. Model 3 fixes all parameters between the two groups with the exception of $\lambda$ and $\delta_R$ (figure B.6.2).

The relationship between parameters is maintained between *w*Mel and *TET* for all model variations, that is $\lambda_{wMel} > \lambda_{TET}$, but the same consistency is not observed not for $\delta_R$. Fur-

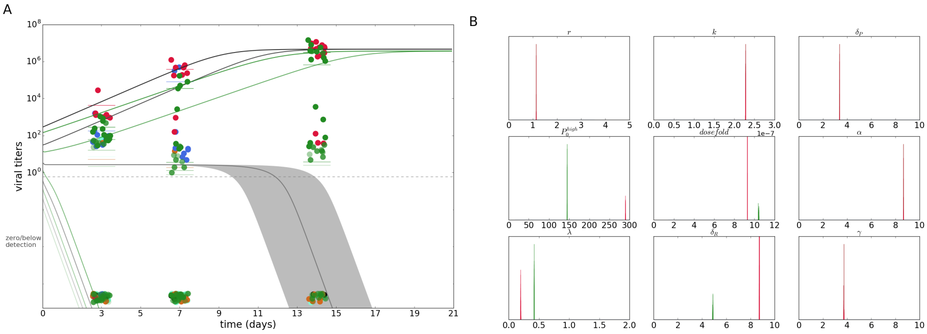


**Figure B.6.1:** (A) Fit of dynamic host-pathogen alternative model 2 ($P_0$ and *dosefold* are different between groups). (B) Posterior distribution of estimated parameters (green is specific to wMel).

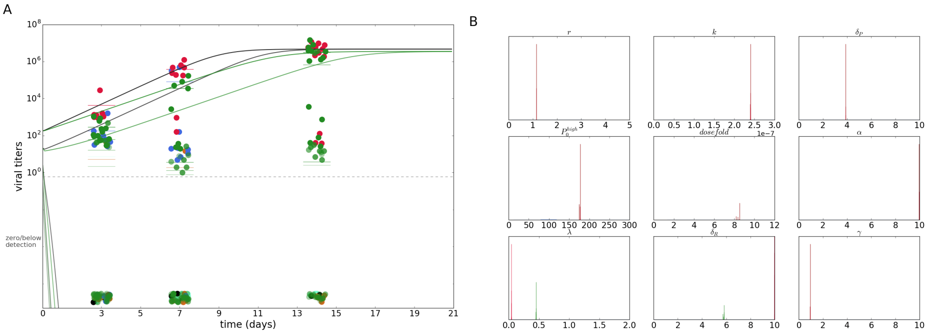


**Figure B.6.2:** (A) Fit of dynamic host-pathogen alternative model 3 ($\lambda$ and $\delta_R$ are the only two parameters that differs between groups). (B) Posterior distribution of estimated parameters (green is specific to wMel).

ther refining of the inference methods, like assessment of intra-individual vs inter-individual heterogeneity, could explain model discrepancies and give more robust results; however, these are beyond the scope of this thesis, and are discussed only as perspectives in the last section of the chapter's main text.

*Taking into account the relevant evolutionary time scale, penguins and viruses are not that different.**

Alexei Drummond

# 4

# Multiple serotype models of dengue virus transmission

## Author contributions

The author of the thesis developed and implemented both the main two-strain mathematical model – adapted from the multiple variants in the published literature – as well as the classic SIR-vector (and multiple compartment variants) in the continuous, discrete stochastic, and individual-based versions. The author did all analytical solutions, performed all simulations, and generated all pseudodata sets, as well as the data sampling algorithms The author implemented all inference algorithms as described in the methods section, and performed all analyses. Flavio Codeço Coelho contributed a finer grained version of a publicly available dengue incidence time series. David Alan Rasmussen contributed code of two Beast 2 Java classes from his published work, as described in the methods.

## ABSTRACT

With around 3 billion people at risk, dengue virus is endemic to many parts of the world. In the Brazilian city of Rio de Janeiro, surveillance measures require notification of new dengue virus cases, and are supplemented by serum collection from patients and sequencing of viral RNA. Phylogenetic analyses have been performed for all serotypes circulating in the country to identify viral genotypes, potentially identify new introductions, and compare viruses presently circulating in the country with those in the past, and of other countries. As a separate type of analysis, a number of mathematical models have been developed to describe dengue virus transmission – particularly qualitative incidence or prevalence patterns – although few have been tested. In this chapter, I show how different mathematical formulations could represent transmission of dengue virus by mosquitoes to humans, how the different model structures entail assumptions about the process, and how these affect outputs qualitatively. Inference from simulated data is used as proof of principle that the kind of data available could be used to accurately estimate all model parameters; however, it is shown that stochasticity may severely hamper efforts to test the models quantitatively. I further implement inference from sequence data for the different models, and compare the performance to that of time series. The methods are applied to the data available for the city of Rio de Janeiro.

## 4.1 Background

The persistence of dengue fever (as well as more severe syndromes caused by dengue virus) constitutes the most extensive viral epidemic transmitted by arthropods, with around 3 billion people at risk worldwide, and 300 million annual cases estimated (Bhatt et al. 2013). The recently recorded expansion in the range of the main transmission vectors, *Aedes aegypti* and *Aedes albopictus* (Kraemer et al. 2015) – presumably due to higher temperatures at temperate regions resulting from climate change – in combination with the emergence of other *Aedes*-transmitted diseases further increased attention to vector control.

Although among the vector transmitted viruses dengue itself has arguably lost some of the attention to Chikungunya and especially to Zika virus (due especially to the previously unknown relationship between the latter and birth defects) at the population level the study of either of these diseases is to a very large extent the study its human and mosquito hosts. The fact that licensed vaccines for these disease was essentially absent – dengue virus had a vaccine in phase 3 clinical trials (Eisen & Moore 2013; Hadinegoro et al. 2015; Villar et al. 2014) that was just recently licensed (WHO 2016) – further highlights the importance of vector control, and of the knowledge about dengue virus transmission in the control of all of the diseases transmitted by the *Aedes* mosquitoes.

One half of the of cycle of dengue virus (*DENV*) – that is, the mosquito-to-human transmission – happens through the bite of an infected *Aedes aegypti* or *Aedes albopictus* mosquito (i.e. the vector in *vector-transmitted disease*); the transmission cycle is completed when an infected human is bitten by a mosquito that in turn becomes infected (Halstead 2007). Although importation from other geographical areas (Salje et al. 2012; Stoddard et al. 2013; Vazquez-Prokopec et al. 2010) as well as sylvatic cycles between non-human primates and other *Aedes* species may play a role in sustaining transmission (Vasilakis et al. 2011), these basic steps of the human-*Aedes* cycle should be enough to create chains of transmission that allow endemicity, and this is considered the primary cycle in explaining dengue virus persistence.

A few details are noteworthy in a general model of dengue virus transmission, which would otherwise conform nicely to that of a generic vector-transmitted mode of propagation. Dengue virus has four antigenically distinct variants – types, or strains – commonly referred to as serotypes (DENV-1 through 4). Anything from a single one to all four of them can be circulating in any one place. If only one serotype is present, a simple description of transmission where a susceptible host gets infected, recovers, and becomes immune to further infection is generally adequate, since infection with a serotype is accepted to confer human hosts lifelong immunity against that same type. If more than one serotype is circulating, infection can happen at least twice (but not more than four times, because unlike influenza, for instance, evolution of the virus does not allow it to escape immunity built against it), one for each previously unseen serotype. In this case multiple infections may need to be accounted for. Also, it could be important to differentiate between strains, as a secondary infection can only be caused by a serotype different from the previous.

Dengue virus transmission has been extensively explored through mathematical models (Johansson et al. 2011). As usual, the disease states of the human hosts have been described by simple extensions of the susceptible-infected-recovered framework, often (but not always) coupled with susceptible-infected description of the mosquito hosts. A mix-and-match of other known or suspected features specific (although possibly not exclusive) to dengue virus have been further added: secondary infections, temporary strain-transcending immunity (cross-protection), enhanced (or reduced) susceptibility to secondary infections, increased lethality in case of severe presentations (often associated to secondary infections, as well as other risk factors such as age or blood-related dysfunctions) (Johansson et al. 2011).

Therefore, dengue virus transmission can be described mathematically by multiple explicit serotypes, which we denote by multiple-serotype models (e.g. a two-serotype model means two different strains of dengue are explicitly described), or by a single explicit virus type, hereafter denoted by SIRX models (which include the classic SIR and SIRS models, as well as intermediate formulations, as described in the methods section). Both formulations purport to describe settings where one or more serotypes may be circulating; in the former description each explicit serotype causes infection once, while in the latter the ensemble of unspecified serotypes causes infection twice or more.

Although seemingly subtle, the conceptual difference between the two modeling approaches is profound: while in the SIRX vector models secondary infections depended exclusively on waiting for recovered individuals to become susceptible again, in the multi-type models strains compete for multiply susceptible individuals. This feature can causes serotype alternation and induce oscillation even in the absence of seasonal forcing. More importantly, this model is less of a caricature of the process of disease transmission, since it is widely accepted that an individual infected with a serotype cannot be infected again by the same strain, rendering the explicit description of the SIRX models technically impossible. Whether one approach or the other is more suitable to describe real dengue epidemics, however, cannot in principle be decided without confronting both models to epidemiological data.

On the epidemiological records side of dengue, some unique patterns are often highlighted in dengue virus epidemics, particularly the oscillations with multianual periods and serotype replacement in successive epidemics (Adams et al. 2006). These can be verified, respectively, from incidence records that show greater number of cases usually around the rainy seasons, and through serological surveys or, more recently, sequencing of circulating viruses. Mathematical models of dengue transmission therefore are built to reproduce these broad patterns; nevertheless, different combinations of anyone's favorite model components may reproduce them, in a way that is indistinguishable from someone else's choice of building blocks. A non-exhaustive list of processes that could produce realistic outputs in a computer simulation include: stochasticity(Otero & Solari 2010), spatial structure (Favier et al. 2005), enhanced secondary infectivity (Nagao & Koelle 2008), "unnatural" transmission routes (Chikaki & Ishikawa 2009).

One of the most hyped effects among the many incorporated one way or another into the mathematical models is that of antibody dependent enhancement, by which a secondary infection would be more severe than the first in the presence of titers of heterologous antibodies against *DENV* (Kliks et al. 1989). The inclusion of the effect has been shown to drive chaotic dynamics even in deterministic mathematical models (Bianco et al. 2009), so as a result it has been suggested that it could be the most important effect modulating the observed somewhat erratic epidemic patterns. The plausibility of the effect is asserted through the observation that in the presence of subneutralizing antibodies invasion of the cell by viruses is facilitated (Guzman & Vazquez 2010); however, in terms of a mathematical model it is not clear if that would translate into increased susceptibility, increased infectivity, or simply a transmission-unrelated increase in virus lethality. If the magnitude of the enhancement could ever be as large as claimed in modeling studies is not established either. Furthermore, it is not clear whether, if present, the effect would be the dominant factor, or if it would be important to the transmission dynamics at all.

Many other effects and combinations would still not exhaust the list of tens of models that purport to explain dengue transmission (Johansson et al. 2011); nevertheless, a quantitative evaluation of the conformity of these models to real data was not done systematically [but see Rasmussen et al. (2014b) for a rare exception]. Here, I put together a model of the human and vector population with secondary infections, either one or two serotypes, and temporary immunity after any infection, but otherwise minimal in what regards any other asymmetries, enhancements, or alternative routes of transmission. I show that minimalistic multi-serotype models can sustain oscillations in the absence of any of the latter effects or stochasticity; I also implement an individual-based model that can simulate both epidemiological and viral evolution. I further develop inference methods to fit these models to time series and multi-serotype sequence data, compare the inference results for the different kinds of simulated data, and apply the estimation method to real data.

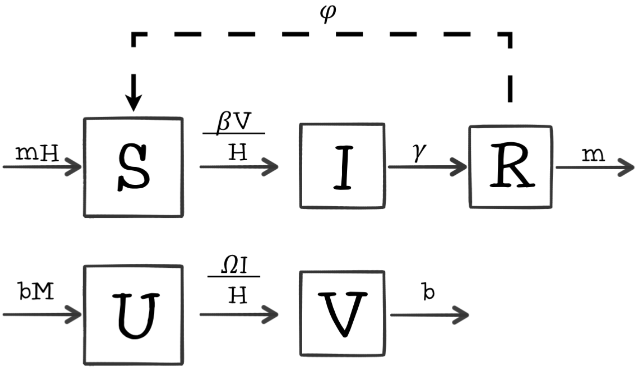


**Figure 4.1.1:** Structure of an SIRS plus vector model, with possible loss of immunity indicated by the dashed arrow. All compartments are subject to natural mortality $m$, but the arrows corresponding to those processes are ommitted in all but the last compartments to avoid repetition and clutter.

## 4.2 Methods

### 4.2.1 SIR model extensions for dengue virus transmission

#### SIRX plus vector models

The simplest model to describe dengue transmission is arguably the vector SIR model, not unlike the first basic models of malaria transmission with a human and mosquito population, although malaria may have an indefinite number of reinfections, making it a SIS model (Ross 1916; Smith et al. 2012). The SIR model assumes human hosts are only susceptible once ($S$), and after infection ($I$) enter the recovered compartment ($R$) permanently, being immune to any further infection by dengue afterwards.

The vector compartment is modeled as a susceptible mosquito compartment ($U$), and an infected one ($V$), from which a mosquito host never exits once it enters – noting that the female mosquitoes are the only ones transmitting disease, so the male population is absent, or implicit.

Alternatively, human hosts may be allowed to lose immunity acquired from past infections and become susceptible again; under that assumption a single host can potentially be reinfected an unlimited number of times before it dies.

The schematic drawing of both model formulations is shown in figure 4.1.1, the only difference between the two being the arrow representing hosts that exit the recovered compartment and reenter the initial susceptible compartment. In this latter case, the structure of the human compartments is that of what is dubbed the SIRS model – the first and last susceptible compartments being the same.

The compartments and parameters are very much standard: $m$ being the human host mortality rate; $\beta$ the mosquito to human transmission coefficient, or rate; $\gamma$ the human recovery rate; $\varphi$ is the immunity loss rate (which is equal to zero in the SIR version of the model); $b$ is the mosquito host mortality rate; $\Omega$ is the human to mosquito transmission rate. Additionally, it is assumed that the birth rates are the same as the mortality rates for each host species; therefore the population sizes stay constant: $H$ is the value set for the human population, while $M$ is the size of the female mosquito population. The mathematical formulation of the SIRS dynamics of transmission is given by the system of equations (4.1).

$$\begin{aligned}\frac{dS}{dt} &= mH + \varphi R - \frac{\beta}{H}VS - mS \\ \frac{dI}{dt} &= \frac{\beta}{H}VS - \gamma I - mI \\ \frac{dR}{dt} &= \gamma I - \varphi R - mR \\ \frac{dU}{dt} &= bM - \frac{\Omega}{H}IU - bU \\ \frac{dV}{dt} &= \frac{\Omega}{H}IU - bV\end{aligned} \tag{4.1}$$

In addition to vector and human basic demographic and epidemiological parameters, it is also common to assume seasonal forcing of the vector population, emulating changing conditions from more to less favorable throughout the year, usually attributable to either hot/cold, or humid/dry seasons. The result of that function is then added either to the birth or death rate of the mosquito population, resulting in a deterministic sinusoidal oscillation.

For clarity the seasonality function is not introduced in this first display of the mathematical system (eqs. 4.1), but is detailed in the following system (4.2) instead.

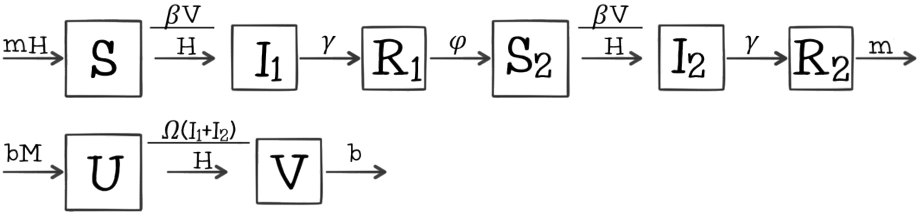


**Figure 4.2.1:** Structure of an SIR plus vector model, with explicit number of possible reinfections (in the particular case illustrated the vector SIRx2).

Given that dengue virus has a finite number of serotypes, it is expected that any one host can only be infected with dengue a few times in a lifetime; therefore, secondary infections can be modeled by explicit compartments for the secondarily or further infected.

In this case, the total number of times a single individual can be reinfected has a hard limit given by the number of infected compartments in the model, which are never revisited. This is straightforwardly modeled by a series of SIR compartments chained together (i.e. an SIR structure followed by another SIR).

The identities of individual serotypes are not explicit in this model, but sequential infections implicitly describe this feature of dengue virus transmission. The case with two consecutive infections is shown in figure 4.2.1.

The compartmental structure of the vector population is unaltered, while the human host population follows a susceptible-infected-recovered path, being immune to reinfection for the time they stay in the first recovered compartment ($R_1$). After that, loss of immunity takes individuals to a second susceptible state ($S_2$, unlike the SIRS model where the first state is revisited), where individuals are again susceptible to infection by infected mosquitoes.

In case of infection they move to the secondarily infected compartment ($I_2$), and after recovery they move to the last compartment ($R$), where they are recovered and can only exit by the ultimate process of death.

The model parameters are the same as the previous model and describe exactly the same processes as before, with the single and only slight exception being that $\varphi$ describes a path of waning immunity through different compartments due to the general model structure. The mathematical formulation of the model with two sequential infected compartments is given by the system of equations (4.2).

As mentioned above, system (4.2) also has a seasonality term acting on mosquito birth rates. This consists of a time dependent cosine function with argument $2\pi$ multiplied by time itself plus a phase variable $\delta$; this assumes these variables are given in years, resulting in an oscillation period of one year, but can trivially be transformed into months, weeks, or days, for instance, by dividing by the appropriate factor. The cosine multiplies the amplitude scaling parameter $a$; this factor is added to the constant birth rate, resulting in a population whose size oscillates between $(1 \pm a)M$.

$$\begin{aligned}
\frac{dS}{dt} &= mH - \frac{\beta}{H}VS - mS \\
\frac{dI_1}{dt} &= \frac{\beta}{H}VS - (\gamma + m)I_1 \\
\frac{dR_1}{dt} &= \gamma I_1 - (\varphi + m)R_1 \\
\frac{dS_2}{dt} &= \varphi R_1 - \frac{\beta}{H}VS_2 - mS_2 \\
\frac{dI_2}{dt} &= \frac{\beta}{H}VS_2 - (\gamma + m)I_2 \\
\frac{dR_2}{dt} &= \gamma I_2 - mR_2 \\
\frac{dU}{dt} &= bM[1 + a \cdot cos(2\pi(t + \delta))] - \frac{\Omega}{H}(I_1 + I_2)U - bU \\
\frac{dV}{dt} &= \frac{\Omega}{H}(I_1 + I_2)U - bV
\end{aligned} \tag{4.2}$$

In contrast to the three-compartment SIRS model I refer to this model as the SIRx2 model (given that the human population states are described by a SIR plus SIR, or two times the SIR structure), where once-recovered individuals would again become susceptible to infection. Because *DENV* has 4 human-infecting serotypes infections may further be tertiary or quaternary; unless explicitly stated, I hereafter refer to all infections after the first simply as secondary, as opposed to primary. SIRx3 and SIRx4 models where individuals infected twice or three times, respectively, become once again susceptible can be built by straightforwardly extending the SIRx2; therefore, I do not show specific schemes or systems of equations for those.

In the infinity limit these models become the SIRS model, except human hosts enter an infinite number of new compartments instead of entering the same compartments an infinite number of times, provided of course hosts stay alive long enough. Depending on rates of infection and death, a smaller number of compartments may be enough for a human host to have enough new compartments for many lifetimes of repeated infection, in which case the model will also approach the SIRS model even with a finite number of re-infections.

I refer to this entire class of models as SIRX, which include the shown SIR, SIRS, as well as the SIRx2 (or any other number between two and infinity) models. In any of those cases, however, the identity of multiple serotypes are only implicit in the fact that human hosts can have secondary infections, because there is only one class of infected mosquitoes that transmit to all susceptible humans regardless.

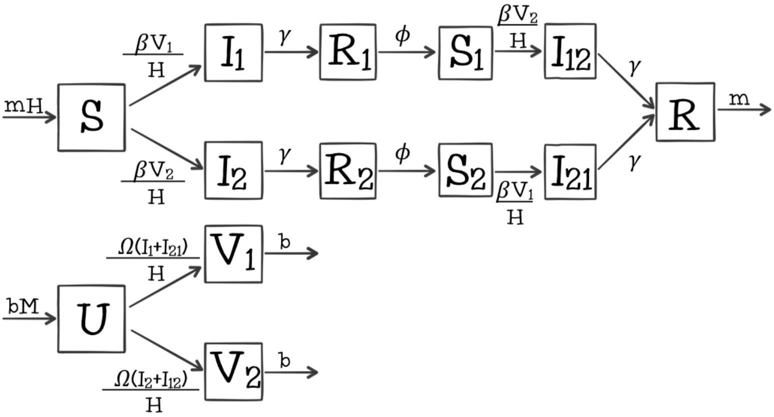


**Figure 4.2.2:** Structure of an explicit two-serotype model and its rate parameters.

### Multiple-serotype infections in the descriptions of dengue virus incidence

A more complete description of multiple-serotype transmission is one that differentiates not only between primary and secondary infections but also disinguishing which serotype causes infection each time. This requires not only a series of compartments, but also parallel paths that describe the order in which the multiple serotypes cause infection. For two serotypes, for instance DENV-1 and DENV-2, it is accepted that a human host could be infected twice, once for each serotype; that is accounted for by the two sequential infected compartments of the SIRx2 model described previously.

Here I wish to account for the order of infection: a human host can either be infected by DENV-1 and then DENV-2, or by DENV-2 and then DENV-1; this creates two alternative paths which are shown in figure 4.2.2 – e.g. $I_{12}$ denotes individuals first infected with serotype 1 and now infected with serotype 2.

Unlike the SIRx2 models the mosquito hosts can either harbour one or the other serotype; therefore, also in contrast to the previous formulations, a secondary human infection depends on transmission from a mosquito with different serotype from that of the primary. The mathematical description of this explicit two-serotype model is given by the system of equations (4.3).

$$
\begin{aligned}
\frac{dS}{dt} &= mH - \frac{\beta}{H}(V_1 + V_2)S - mS \\
\frac{dI_1}{dt} &= \frac{\beta}{H}V_1 S - (\gamma + m)I_1 \\
\frac{dR_1}{dt} &= \gamma I_1 - (\phi + m)R_1 \\
\frac{dS_1}{dt} &= \phi R_1 - \frac{\beta}{H}V_2 S_1 - mS_1 \\
\frac{dI_{12}}{dt} &= \frac{\beta}{H}V_2 S_1 - (\gamma + m)I_{12} \\
\frac{dI_2}{dt} &= \frac{\beta}{H}V_2 S - (\gamma + m)I_2 \\
\frac{dR_2}{dt} &= \gamma I_2 - (\phi + m)R_2 \\
\frac{dS_2}{dt} &= \phi R_2 - \frac{\beta}{H}V_1 S_2 - mS_2 \\
\frac{dI_{21}}{dt} &= \frac{\beta}{H}V_1 S_2 - (\gamma + m)I_{21} \\
\frac{dR}{dt} &= \gamma(I_{12} + I_{21}) - mR \\
\frac{dU}{dt} &= bM[1 + a \cdot cos(2\pi(\delta + t))] - \frac{\Omega}{H}(I_1 + I_{21})U - \frac{\Omega}{H}(I_2 + I_{12})U - bU \\
\frac{dV_1}{dt} &= \frac{\Omega}{H}(I_1 + I_{21})U - bV_1 \\
\frac{dV_2}{dt} &= \frac{\Omega}{H}(I_2 + I_{12})U - bV_2
\end{aligned}
\tag{4.3}
$$

Although the model described by the system of equations (4.3) does contain (more than a couple) SIR-like components, and could possibly be seen as a parallelization of the SIRx2 model, the level of complexity arising from further introducing *DENV*-specific features is considerably higher. It may be more useful to look at it as two interacting epidemics (Rohani et al. 2003), since it is unlikely that either the compound output or the individual serotype dynamics can be predicted to trivially conform to that of its more well known building blocks.

Regarding the basic reproductive number, $R_0$, because it is calculated with regard to a fully susceptible population, the secondary compartments do not cause the models to differ in that matter, and the quantity is given by $R_0 = \frac{\beta\Omega}{b(m+\gamma)}\frac{M}{H}$. If such differences are actually verified in the output of the model, it would be due to the structure of the immunological states believed to represent a population at risk of a multi-serotype dengue epidemic, since all other processes are present in the previous models as well.

On top of the structuring of basic epidemiological and demographic processes (Keeling & Rohani 2011), additional complexity may appear by introducing asymmetries between serotypes (e.g. serotype 1 more infective than serotype 2, or causing infection for a longer period of infection) or order-dependent rates of infection (e.g. secondary infections more or less likely than primary), which are plausible for various biological or medical reasons, and are likely in comparison to the narrow null hypothesis of perfect symmetry.

Other common extensions that are known to be present to some degree, unlike more complex hypothesized immunological or epidemiological effects, include exposed compartments (describing individuals that harbor the pathogen but cannot transmit it yet), spatial structure of transmission, heterogeneity in contact rates, susceptibility to infection, or in infectivity, gamma-shaped (as opposed to exponential) host survival, and many others. Nevertheless, none of these are included in the models used here, I discuss some of these later in the text, although it may become clear by the section on inference which are the difficulties of including too many parameters however simple and concrete the processes may be.

### Implementation

The models were initially implemented as continuous ordinary differential equations (ODEs), solved by numerical methods to approximating the deterministic solutions of the system; commonly available as ODE solver functions in multiple programming languages such as Matlab, Python, and R languages, which were used at different times with no particular preference for either.

## 4.2.2 Individual-based models

Discrete, stochastic, individual-based versions of the models described above were also implemented. Besides the importance of the randomness of the events in the epidemiological model, it was important to be able to simulate not only the epidemiological outputs, but also the evolution of viral sequences. Because there is no direct continuous approximation neither to the appearance of a random mutation, nor to a genetic sequence of nucleotides, the most straightforward way of simulating evolution is to explicitly attribute viral sequences to infected individuals, and allowing them to randomly acquire new mutations as they get transmitted.

Implementations were done in both C++ programming language by using a previous implementation (Gordo et al. 2009; Gordo & Campos 2007), and later by adapting the algorithm to the Python programming language to take advantage of the random number implementations in the latter.

In brief, an "Individual" class was created to have a "sequences" attribute (which was empty if the individual was not infected), and each human or mosquito host was an instance of that class. At each time step ($\Delta t$, which multiplies all probabilities hereafter mentioned), the number of new infections was drawn from a random binomial distribution, since the maximum number of infections is bounded by the total number of susceptible individuals; the probability parameters were equal to the force of infection (e.g. $\lambda = \beta V_1/H$ for infections caused by mosquitos infected with serotype 1) and number of trials equal to the susceptible population (e.g. the susceptible to all $S$, or to type 1 $S_2$, equivalently.) Because the infectivity of all individuals is assumed to be the same, the sequence infecting each new host is randomly drawn from the pool of all existing sequences from the previous time step. After a successful infection of a new individual, new mutations have the opportunity to arise with probability determined by a per-genome mutation rate – mutations are assumed to follow an infinite alleles and sites model, so every new mutation is one that was not previously present in the population. Mutations do not affect any of the model parameters, so evolution is completely neutral. The number of sequences inside a single infected host can be greater than one, in which case this within-host population undergoes a Wright-Fisher sampling step (Wakeley 2009).

The number of new births of the mosquito population, apart from the sinusoidal additional factor, is drawn from a poisson distribution with mean $\mu_{vec} = b$; deaths are drawn from a binomial distribution with probability parameter also $\mu_{vec} = b$, so the total population is expected to fluctuate around the initial value. The human population is assumed to be strictly constant; that is enforced by the number of births being exactly equal to the number of stochastic deaths – this condition can be easily relaxed, however.

Otherwise, as a general rule, the number of events at each time step was drawn from a random binomial (and when applicable, multinomial) distribution where the number of trials was given by the number of individuals in the compartment, and the parameter for probability of success in each trial was given by the rates in the model (e.g. the number of individuals recovered from an infected compartment $I_1$ is given by a random binomial distribution draw with parameters $I_1$ and $\gamma$). If competing processes were present, a multinomial was used instead. If the number of events was not bounded, for instance by the size of the compartment, a poisson distribution was used instead (e.g. the number of mosquito births, or number of new mutations).

The output of this implementation is both a stochastic time series of susceptibles, infected, recovered, and incidences (i.e. the randomly drawn number of new infections recorded at each step in the human and vector populations), and the pool of all extant pathogen sequences for each serotype at all or selected time points (for convenience, split into mosquito and human harbored sequences).

### 4.2.3 Epidemiological surveillance data (and pseudodata)

In countries like Brazil, where notification of dengue is compulsory to doctors, the records commonly consist of periodically reported new cases into the Information System for Incident Notification [Sistema de Informação de Agravos de Notificação] (SINAN). As with many other common endemic diseases, laboratory confirmation is not routine, so diagnostic relies mainly on clinical criteria; neither the serotype causing the infection is normally recorded, nor if the infection is primary or not. Therefore, typical time series do not distinguish between serotypes or sequential infections; what would be available would be a series of equally spaced, discrete values representing the number of cases of "dengue fever" generically defined reported every month or week (SINAN).

The time series data set used here is from the Brazilian SINAN, with the absolute number of new weekly cases of dengue from the year 2009 until 2013 in the city of Rio de Janeiro, when three large incidence peaks are observed. More specific diagnostics data have also been occasionally produced in the form of serological surveys, although these were obtained for specific studies and small cohorts. Because this kind of data is sparse and difficult to access, I do not use any such data, but merely note that it exists for the disease and location I am (mainly) concerned with, even if in a fragmented way.

Relatively recently, routine surveillance of dengue started to include genetic data of the virus. It is now routine activity to isolate samples from patients and obtain nucleotide sequence from the virus in the isolate; an immediate result is the identification of serotypes, and possibly of more specific variants such as genotype (a finer grained distinction within each serotype) as well as the relationship to strains previously found elsewhere (dos Santos et al. 2002). This data set therefore consists of somewhat sparsely sampled viral isolates sequenced along several years – the total number is in the order of tens for the city of Rio de Janeiro – and is to a great extent available as part of published studies (Araújo et al. 2009; de Bruycker-Nogueira et al. 2015; Castro et al. 2013, 2012; Miagostovich et al. 2003, 2006; DeSimone et al. 2004), as well as in public databases for genetic sequences such as GenBank (Benson et al. 2015).

The individual-based models described in the previous subsection were designed in a way that could directly reproduce the form observed in the real data available. Unless there is specific interest in greater details, whenever a model is simulated I try to store the output in a format that mimics the amount, type and level of aggregation, period and interval of collection, and any other feature pertinent to a specific data set. When used in the same way as the real data (e.g. for parameter inference), I call a data set of this sort synthetic, or pseudo-data.

Generally summarizing the two types of pseudo data sets used here, the time series are weekly records of all new cases (the weekly sum of daily-generated incidences), and the second type are genetic sequences. The genetic equivalent of the all-inclusive time series data would be a sequence (or group of sequences) for every newly infected individual time-stamped with the week (or any other time step) when it appeared. That is impractical even for simulated data set, and for a series of reasons it is essentially impossible in real epidemics.

Instead, the pseudo-sequence data is a sample from different times, where the number of sequences from each time point is proportional to the number of cases then, i.e. it is a sample of sequences over one long period, weighted by epidemic size at each of multiple small intervals. A total number of 100 sequences per serotype over the course of a few years was assumed to be a sufficient number, comparable to that of previous data sets used for similar purposes (Rasmussen et al. 2014b), and considered feasible in a city with a population on the order of 10 million, and outbreaks on the order of a few tens of thousands (SINAN). It is also comparable (though larger) in size to data sets collected for other purposes in the city of Rio de Janeiro (DeSimone et al. 2004) When needed, the specific pseudo and real data sets are detailed at the pertinent results sections.

The objective is to obtain pseudo data sets suitable for inference purposes. It is difficult to establish beforehand what the most informative sampling scheme would be (Frost et al. 2015); that is a question on its own right that is not explored here.

### 4.2.4 Bayesian inference from time series

Analytical solutions for the systems adopted in this chapter are not available; therefore, a numerical approximation to the continuous solution was obtained through an ODE solver whenever needed.

Because the number of new cases in any given week is an integer number I chose to use a poisson distribution: the likelihood of the observed value is computed using the sum of all possible human infected states as the poisson parameter (as modeled by compartments, i.e. the total number of dengue cases of any kind in the model output), and the total likelihood is therefore the product of that over all time points in the series – or, more conveniently, the sum of their logarithms.

A binomial distribution could as well be used, in which case its probability parameter could be given by the forces of infection and the number of trials would be given by the susceptible populations at risk. While that would be possible, and possibly mimic more accurately our simulation model and the bounds in the maximum number of infections, it is more cumbersome to add the parameters coming from the different compartments and, more importantly, the poisson distribution expects greater or equal variance when compared to the binomial, and therefore can accommodate any overdispersion in the data.

The bayesian Markov Chain Monte Carlo (*MCMC*) inference algorithm was implemented in Python language using the PyMC module (Patil et al. 2010). Unless detailed otherwise, as a norm gamma-shaped priors were used for parameters that have independently estimated or commonly accepted values, and priors uniform over a wide range were used otherwise. As technical criteria for quality of the inference, the following criteria were used (Gelman et al. 2013, chap. 11): Markov chains were run until the likelihood and posterior traces converged to a maximum and attained stationarity with that regard; for all results shown, replicates of the chains were run to assure mixing (unless otherwise specified, model fit and posterios are computed and shown for individual chains only); initial sampling corresponding to one tenth of the total number of iterations was discarded as warm-up or burn-in period. Correlation between parameters were computed at the end of the chain for at least one replicate when parameters seemed systematically biased.

Most critical was the time limit of around one month that was imposed for practical reasons; that allowed chains as long as a few million iterations on a dedicated high-capacity computer. Most implementations did not seem to have issues with the likelihood converging with a few hundred thousand iterations; nevertheless, more subtle issues were observed in some cases and are discussed in the results and discussion sections.

I applied this estimation method to both simulated data sets, as well as to a time series of dengue incidence from the city of Rio de Janeiro running from 2009 to 2013 (SINAN).

The estimation method used does not therefore account for noise in the system state. Stochastic solutions can be used instead; however, doing so is not as simple as replacing a deterministic solution by a single stochastic one (which may be only slightly slower to obtain), but requires a considerably more sophisticated and a computationally much more intensive method method to estimate the likelihood Andrieu & Doucet (2010); Ionides et al. (2006), with possibly a couple of thousand simulations at each iteration of the Markov Chain. I discuss these so called Sequential Monte Carlo or particle filters as perspectives in the end of this chapter.

### 4.2.5 Population genetics and phylodynamic inference

Although the individual-based model can produce a simulated data set that mimics a set of viral sequences sampled at arbitrary time points along time, unlike with the time series data there is no simple way of calculating the likelihood of the model parameters given that kind of data.

The conceptually most straightforward method is probably the following: make up metrics that are assumed to be representative of the data; simulate the model; compute the same metrics for a large number of model outputs; and try to find the model parameters that best approximate the real data. In spite of the gross oversimplification of this description, this is the basis of Approximate Bayesian Computation (or *ABC*) methods.

An alternative framework to compute the likelihood of a model given sequence data – and arguably a more elegant one, at least in the sense that it is based on a full likelihood expression – relies on the bifurcating properties of the trees that connect related sequences, or conversely (with time flowing backwards) the coalescence of a set of related samples into a common ancestor. The latter gives the name to the coalescent theory, or simply, *the coalescent*, as described by Kingman (1982). Since then the problem of estimating a Wright-Fisher (or Moran) population size from sequence data using the coalescent has been extended to varying environments (Griffiths & Tavaré 1994), to implicitly defined population functions (Frost & Volz 2010), and more generally to structured populations

(Volz 2012).

The Beast 2 software (Bayesian Evolutionary Analysis by Sampling Trees) contains many basic as well as advanced implementations of coalescent-based *MCMC* inference (Bouckaert et al. 2014); it is written in the Java programming language, and uses an XML file to input the actual sequence data and the specifications to the multiple classes involved in the calculations. A phylodynamics package is also available, which among many things includes implementations of the SIR model (Kühnert). Other software for that purpose is also available, notably as packages for the R programming language (Paradis; Volz et al.). Beast 2 was chosen due to the existence of a general community of users, its openness and extensibility, apparent propensity to phylodynamics implementations, and especially the helpfulness of some of its earliest developers.

Nevertheless, there were no tools built in to the core software, nor any extension packages (including the phylodynamics package developed by Kühnert) that were suitable for direct application to the inference using the models pertinent to dengue virus transmission as describe above. For one, a structured implementation is needed – in the simplest case, viral sequences may be either in the human or in the mosquito host, and intuitively coalescence can only happen if they are inside the same hosts (unfortunately, further details about structure in the coalescent for this kind of model are way out of scope, but see Volz (2012) for a complete description of how structure is incorporated in the context of disease transmission models). The implementation was greatly facilitated by code shared from a previous implementation from Rasmussen et al. (2014b), which consisted of an epidemiological model Java class and a structured coalescent likelihood computation class; these extensions could almost directly be applied to the SIR-vector model, and could straightforwardly be used for the SIRS model as well, and more or less easily adapted to all SIRX models.

The implementation of the two-serotype models, however, required additional tinkering. Because *DENV* serotypes are only  60% similar by sequences, they are usually treated separately (de Bruycker-Nogueira et al. 2015; Castro et al. 2013, 2012). Therefore, they are not expected to find a common ancestor in the recent period of a few years of dengue transmission, in which case two separate trees are needed. The two-serotype epidemiological model produces the poulation dynamics function for both serotypes, so a single Java class was implemented with this model; however, the likelihoods have to be calculated for each tree, so two separate classes were implemented that used separate trees and substitution model parameters, but fetched (different, serotype-specific parts) of the same population model output. The combination of the two tree likelihoods was then the global likelihood of the one-epidemiology, two-tree model, and from then on could be used by the *MCMC* algorithm in Beast 2 with no additional tinkering needed to perform inference. The same quality checking criteria as in the time series inference were used; one month of real time amounted to a few tens of millions of iterations in Beast 2 for the most complex models. Additionally, Beast 2 computed the effective number of samples (Gelman et al. 2013, chap. 11) as an additional metric to assess convergence.

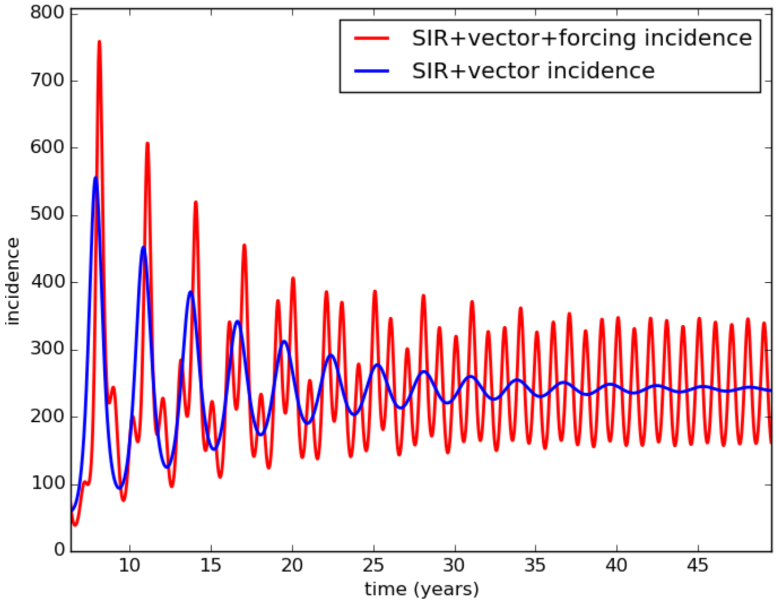


**Figure 4.2.3:** SIR+vector models with (red) and without seasonal forcing (blue) Parameter values in are: $\Omega = \beta = 1.0, b = 0.1, m = 3.65 \cdot 10^{-5}, \gamma = 0.14, a = 0.1, H = 1000000, M = 513789$.

## 4.3 Results

### 4.3.1 Forward model simulation

#### Predictions for disease incidence and prevalence

Figure 4.2.3 shows a deterministic simulation of the SIR-vector model without, and with seasonal forcing in the birth rates. It is worth noting, that although the amplitude of seasonal forcing is of the magnitude of 10%, oscillations can quite easily be greater than that through resonance of the natural (damped) oscillations and the forcing (Dushoff et al. 2004). The average incidence, however, is roughly the same in the presence or absence of forcing.

In the case of waning immunity, secondary infections are also possible, which potentially increases the total incidence at any point in time. Figure 4.3.1 shows the total incidence for the vector transmitted SIR, SIRx2, SIRx4, and SIRS models; for a set of parameter values, changing only the transmission rate within a certain range can result in a very similar incidence profile even in a deterministic setting.

Although for larger transmission rates the number of susceptibles may dominate and cause the models to output distinguishable time series, it must be acknowledged that for some parameter sets, model structure is not easily identifiable from the incidence pattern. Therefore, not only parameter values, but also model structure, are likely to be critical to describe the transmission of dengue virus, and can have a great impact in the parameters estimated.

The structure of the two-serotype model is not directly comparable to the previous models; nevertheless, the output of the model with the same common set of parameters (except for $\beta$) is shown in figure 4.3.2, except for transmission.

Besides the obvious difference that this model has explicit series for both serotypes, the incidence pattern is less regular. A more striking qualitative difference is the fact that the numbers of infected individuals get much closer to zero (at least for individual serotypes); in a deterministic model it means just that, but in a stochastic model that may mean in-

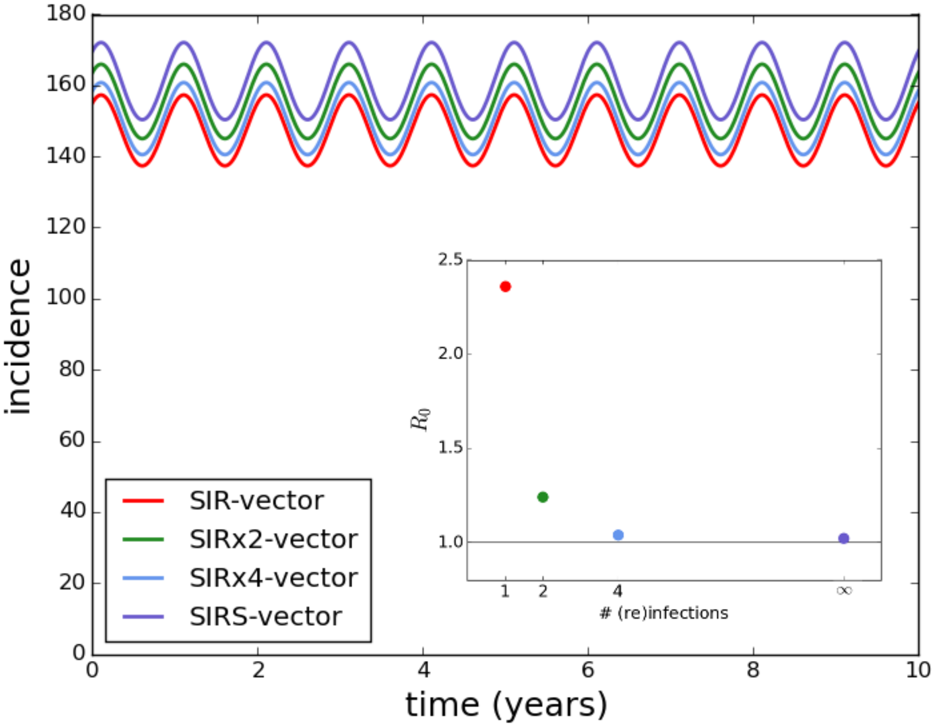


**Figure 4.3.1:** SIR+vector models in some of its variants: standard vector SIR (red), SIRx2 (green), SIRx4 (light blue), and SIRS (purple), showing approximately the same incidence levels at or near near the oscillatory equilibrium. The single changing parameter is transmission intensity for the SIR, SIRx2, SIRx4 and SIRS models, respectively: $\Omega = \beta = 0.2536; 0.1835; 0.1680; 0.1665$. Additional parameter values in are: $b = 0.1, m = 3.65 \cdot 10^{-5}, \gamma = 0.14, \varphi = 0.00165, a = 0.02, \delta = 0, H = 1000000, M = 513789$.

creased probability of extinction.

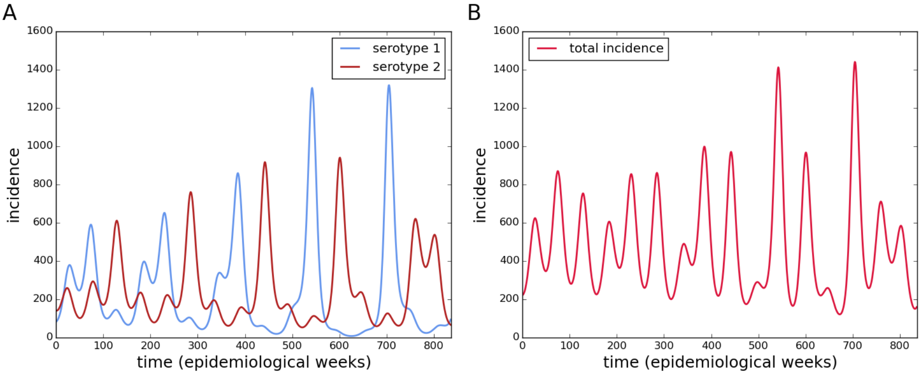


**Figure 4.3.2:** Incidence dynamics for an explicit two-serotype dengue virus model. Shown incidence is that for each serotype separately (a), and the sum of new infections for both serotypes (b). Parameter values are: $b = 0.1, \Omega = 0.7, \beta = 0.7, m = 3.65 \cdot 10^{-5}, \gamma = 0.14, \phi = 0.00165, a = 0.02, \delta = 0, H = 1000000, M = 513789$.

## Discrete time and individual-based implementations of vector transmission of DENV

Discrete time and stochasticity in the system state transitions can significantly affect the oscillations in resonating systems (Dushoff et al. 2004). This is illustrated by the incidence outputs of both the vector SIR, and more extremely the vector SIRS models, in figure 4.3.3.

The probability of extinction of the pathogen is generally low in those cases, although it is clear that numbers are much lower for the case without reinfection (vector SIR), such that if extinction doesn't commonly occur, the trajectories are nevertheless much more prone to the general process noise. Similarly, the intermediate models with two and four infections are prone to effects of discretization and stochasticity. Figure 4.3.4 shows the results for the remaining SIRX models.

For the chosen parameters, extinction is more likely in the two-serotype model Figure 4.3.5 shows the results for this model.

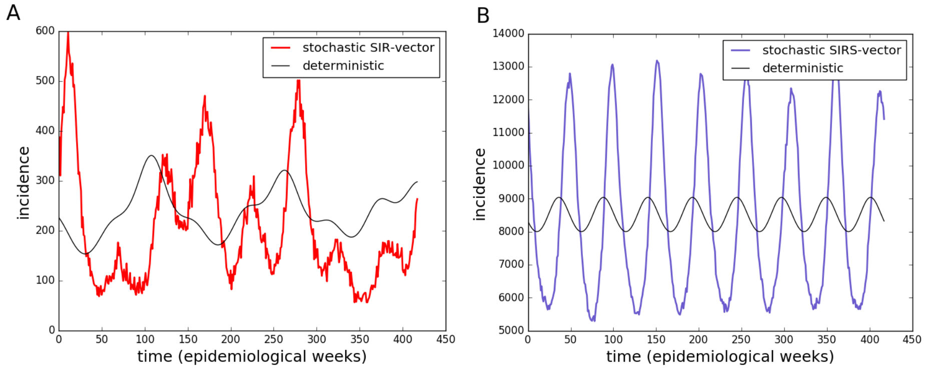


**Figure 4.3.3:** Stochastic output of vector SIR model (a), and of SIRS model (b). Gray lines show solutions to the deterministic system. Parameter values are: $b = 0.1, m = 3.65 \cdot 10^{-5}, \gamma = 0.14, \varphi = 0.00165, \alpha = 0.02, \delta = 0, H = 1000000, M = 513789.$, and $\Omega = \beta = 0.7$ for the SIR, and $\Omega = \beta = 0.47$, for the SIRS.

### Summary of forward modeling predictions

Models that are structured differently but with the same parameters, and therefore same basic reproductive number, will have different outputs and observed incidence and prevalence levels. Conversely, qualitative aspects of the observed disease data can be reproduced by different combinations of model structure and parameters. It is not at all trivial to determine what aspects of the model are known, or best approximate reality, which can be treated as nuisances, and which of them are robust to structure or parametrization changes.

In the case of genetic diversity, it is even harder to distinguish clear relationships between the real data and the model predictions. Summaries are useful to make broad assessments about the data, but are usually not suitable for finer grained comparisons, nor model testing. A quantitative comparison is therefore needed to assess which model best represents reality, and what parameter values explain the processes of disease transmission.

The forward simulation approach is quite convenient when good estimates are available for all or most parameters, and there is good confidence in the model structure, or the outputs are robust or easy to evaluate for unknown parameters. Even for a reasonably large

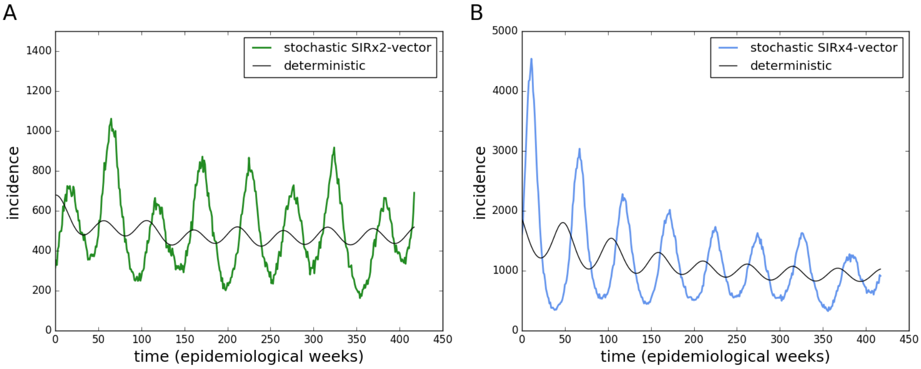


**Figure 4.3.4:** Stochastic output of vector SIRx2 model (a), and of SIRx4 model (b). Gray lines show solutions to the deterministic system. Parameter values are: $b = 0.1, \Omega = \beta = 0.7, m = 3.65 \cdot 10^{-}5, \gamma = 0.14, \varphi = 0.00165, a = 0.02, \delta = 0, H = 1000000, M = 513789$.

amount of synthetic (or pseudo-) data it is feasible to perform computation of genetic summaries in a reasonably short time without much algorithm optimization effort.

For complex, non-linear models it may, however, be difficult to thoroughly perform sensitivity analyses for more than a couple parameters and identify disease dynamics compatible with real data. For genetic diversity computations that rely on individual-based simulations it is even more costly and time-consuming to adopt the forward approach.

### 4.3.2 Inference: quantitative hypothesis testing

#### Time series-based inference

The *MCMC* algorithm produces samples of the joint posterior, which consist of a series of parameter sets accepted by the method, the fit can be empirically computed by simulating the model for a representative subset of these samples, and computing the mean or median values of the model output. Credibility intervals can be similarly computed by taking the score at some percentile (e.g. a 95% CI results if 2.5% and 97.5% are chosen). We present the mean computed as described above and call it the "fit" hereafter, together with the 95% credibility intervals, unless otherwise stated.

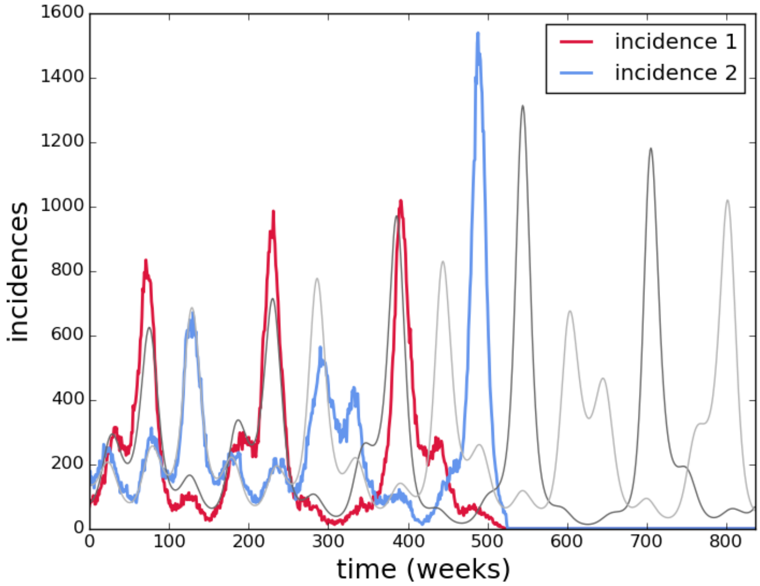


**Figure 4.3.5:** Stochastic output of two-sreotype model. Gray lines show solutions to the deterministic system. Parameter values are the same as before.

For inference purposes, the ratio of mosquito population $M$ to human hosts $H$ is the estimated parameter, denoted hereafter as $M_{ratio}$. The mean time of immunity before becoming susceptible, that is $1/\phi$ or $1/\varphi$ are the parameters actually estimated, and are denoted just as such, the inverse of the original parameter.

The easiest data set the inference method can fit is a deterministic simulation with some noise added afterwards to the continuous solution. This best case but unlikely scenario serves as proof of concept that there would be enough information in a data set with this format. Figure 4.3.6, panel A shows the fit of the two-serotype model to pseudodata produced by a deterministic simulation of 209 weeks (approximately four years) with poisson noise added to each of the 209 time points, i.e. the model used for estimation is the same as that used to produce the simulated data.

The fit shown in is extremely good, considering the pseudodata is the output of a highly nonlinear model, and that only an aggregated and partial observation of the system is used (i.e. the sum of the changes in the infected compartment ). Therefore, while this is the best case scenario, it is not a given that it would be possible to adequately fit the model, and furthermore estimate the parameters accurately – the parameters could be structurally unidentifiable, or the amount or type of data could not allow accurate estimation of the parameters. The estimates of the model parameters are shown in figure 4.3.6, panel B, which displays the prior and posterior distributions. Most estimates are quite precise even in the absence of informative priors.

The case of the exact same estimation method applied to a data set produced by the stochastic individual-based model simulation instead is shown in figure 4.3.7. While the fit to the data is still quite good, the effect of stochasticity on the posterior estimates can be observed as pronounced biases in some posterior distributions.

Unlike the previous case, gamma priors were used for the initial population states, and they are nevertheless not as well estimated as before, but contain the true values (not shown). Particularly important is the estimate for $R_0$, which is significantly lower than the true value and the estimate of the temporary cross protection time (i.e. strain-transcending immunity – the inverse of the rate of immunity waning $\phi$ – only present in models with secondary infections), which is precisely estimated.

The observed biases are similar to those observed with a simpler vector SIR model (appendix C, figure C.1.1), and the fact that these are not observed with the continuous simulation suggests that model complexity or data structure are not the main factors.

Indeed, fixing some parameters and/or initial conditions can improve estimation of the remaining parameters. Appendix C shows that fixing all other epidemiological parameters except for $R_0$ improves it estimate and fixing initial conditions improves it further (figure C.1.2). However, it is not necessarily true that the more parameters fixed, the better the estimates, as leaving a few of the epidemiological parameters results in the best estimate of $R_0$ when compared to the above (figure C.1.3).

The *MCMC* algorithm allows empirical computation of the correlation between the parameters, since it relies on repeatedly sampling the posterior distribution of parameters. The biases could therefore be at least in part attributed to correlation between the estimates

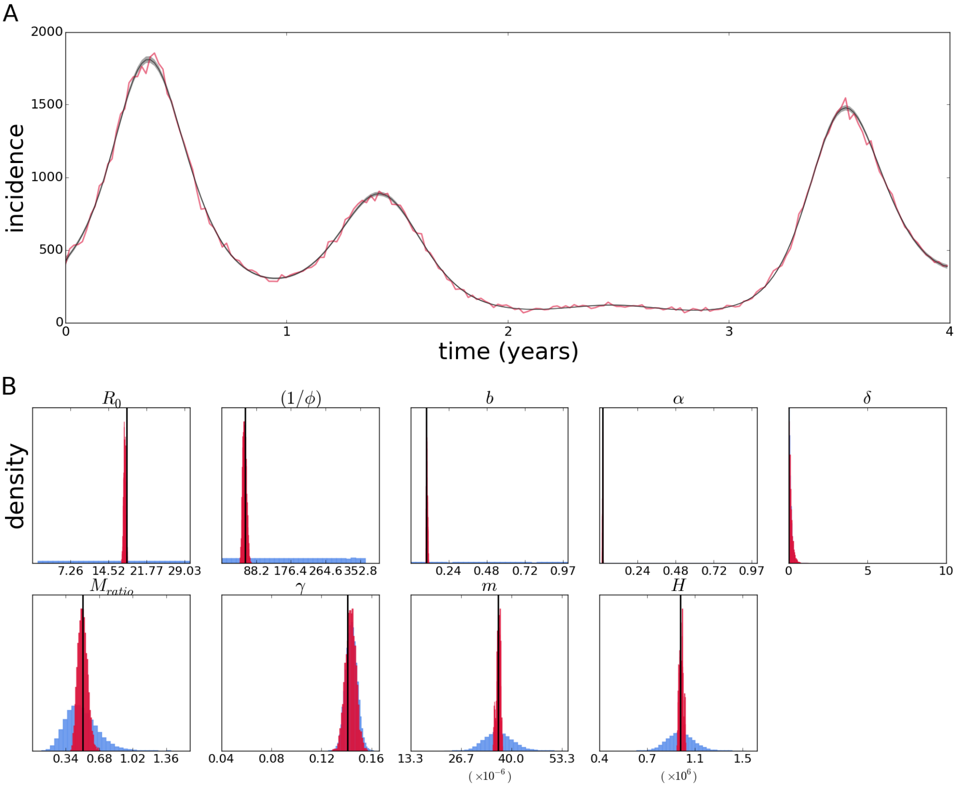


**Figure 4.3.6:** Fit of the continuous two-serotype model with poisson likelihood to data simulated from the continuous model with poisson noise added afterwards (A). Posterior distributions (crimson) of epidemiological parameters (B) – prior distributions are shown in light blue (possibly not visible if the densities are very low compared to the posterior density). True values, i.e. parameter values used to simulate the data set, are shown as vertical black lines.

(not shown, but see discussion section); however, the problem seems to go beyond lack of identifiability of specific parameter combinations, since allowing some parameters to accommodate uncertainty can improve inference results. I further discuss parameter correlation, fixing parameters, uncertainty in estimates, and possible methods to get around these issues in the next section.

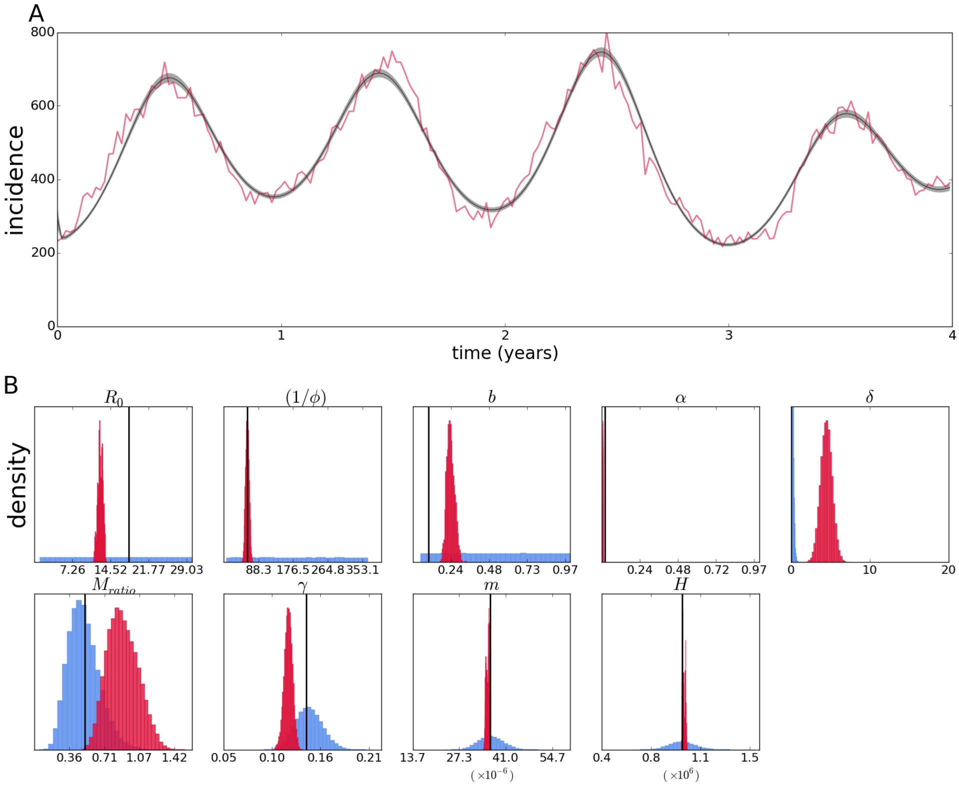


**Figure 4.3.7:** Fit of the two-seroytpe continuous model with poisson likelihood to data simulated from its stochastic version (A). Posterior distributions (crimson color) and priors (light blue) of epidemiological parameters and initial conditions (B).

Because incidence data most likely does not differentiate between infecting serotypes, a time series gives no information about the number of circulating serotypes, and therefore it is not possible to decide on single or multi-serotype models based on that alone. For inference purposes a number of circulating serotypes must be assumed when a transmission model is used.

The previous results assume the correct model structure (although not the correct error model) is known; with real data other factors such as model misspecification can introduce further sources of errors. Appendix C shows that using the vector SIRx2 model to estimate parameters from a time series simulated from the two-serotype model can throw estimates off (particularly that of $R_0$); using the same model on data simulated from the vector SIR model can have equally problematic results. That highlights the importance of testing different alternative models for the same data set, when it is not possible to favor any specific model based on the data alone.

It is in principle possible to distinguish the infecting serotype, and even the order of infection, for time series data; however, that would require elaborated tests and/or record keeping for current and previous infections for every single recorded case. Sequence based inference relies on a sample of cases, not all possible records, and differentiates infecting serotype by default, which can get around some of these issues.

### Genealogy-based inference

The aim of this section is to show results of methods comparable to that of time series-based inference. In Rio de Janeiro, the available dengue virus sequence data seems to be a result of concentrated efforts to obtain data representative of each single epidemic period, not an overall data set with particular features. To my knowledge, there is no specific goal for the entirety of the available data as to the total size, sampling interval, or representativity of both epidemic and inter-epidemic periods, for instance. Therefore, the simulated data set used for inference here is created to mimic more of an ideal yet feasible data set to collect. In practice that was done by deciding on the total number of sequences desired and sampling with a constant probability that would yield that expected value – given that binomial sampling was done every week, periods with greater number of cases would be more represented in the data.

For a constant probability to be specified, the total number of cases in the entire interval should be known before sampling, which cannot be the case in the real world; even if that was known a probability of sampling a sequence cannot be directly decided on by researchers. Nevertheless, the computational sampling scheme mimics the real world in that the greater the number of infected individuals, the more report to hospitals and health centers, and the probability of obtaining consent to get biological samples and sequencing them can be set by an arbitrary rate (e.g. a goal to obtain a sample from one out of 100 patients) that would yield a total number of samples over a usual epidemic.

For the vector SIR model, a pseudodata set was sampled from the last four years of a sixteen year run resulting in 106 simulated sequences with random mutations. The results from estimation based on that data are shown in figure 4.3.8. The posterior contains the true values of the parameters, although some are slightly biased, not centered around the true value.

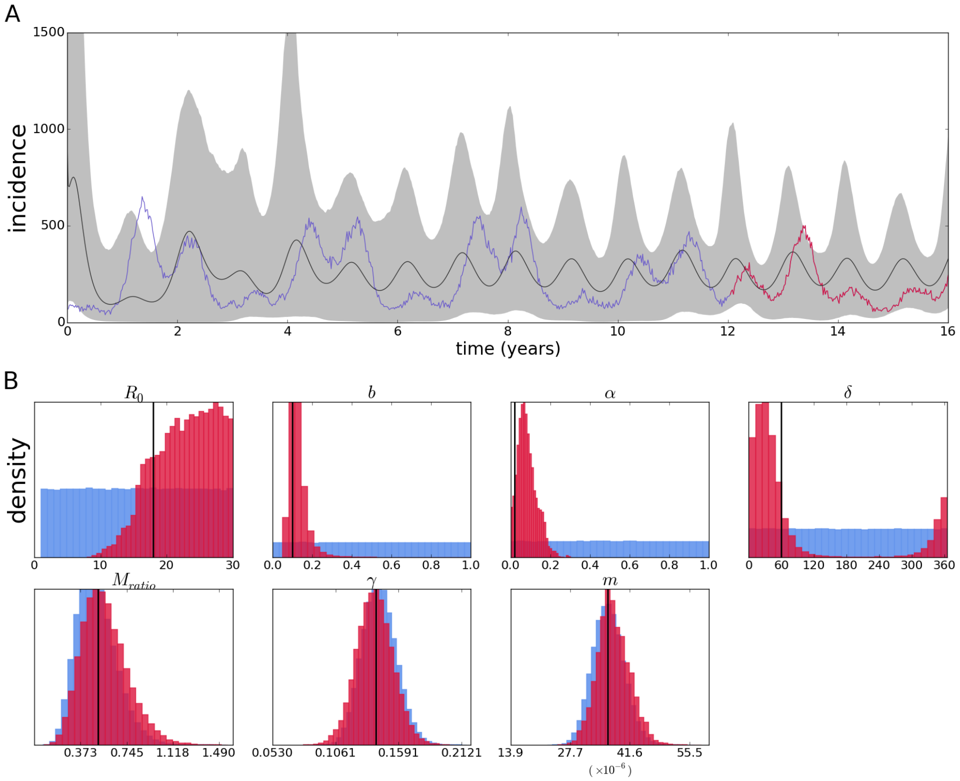


**Figure 4.3.8:** Reconstruction of the epidemic with a vector SIR model from sequences simulated by its corresponding individual-based model(A). Red portion of the time series denotes period from which samples were taken. Posterior distributions (crimson color) and priors (light blue) of epidemiological parameters (B).

Again an expected value and confidence intervals are computed from the sets of parameters in each step of the Markov Chain; however, figure 4.3.8A does not show a fit, since the time series data is not used for estimation, but rather a reconstruction of the incidence patterns. Nevertheless, this reconstruction may accurately depict the mean incidence if the sequence data is informative enough, and here it indeed reproduces oscillations on the same order as those in the data. The phase of the oscillations are not perfectly synchronized, and the amplitude is more regular than the actual incidence (although this is expected due to the stochastic data as opposed to the deterministic nature of the inference method; this is further discussed in the next section).

For the two-serotype model, the pseudo-sequences used here for inference were produced by the same run of the individual-based model as the time series pseudodata (in the previous subsection) – as would be the case with a real epidemic, where all possible new cases are recorded as incidence, and some of them are sampled for sequencing. The pseudo-sequence data is sampled from the same 209 weeks for which pseudo-incidence time series was recorded; the total simulation length was 16 years (therefore, the pseudo-time series in the previous subsection actually consists of the last 4 years of this longer run. Only the sequence data is used to produced the results that follow.

Inference using 119 type 1 sequences and 104 type 2 sequences sampled according to the above scheme from a population of simulated individuals is shown in figure 4.3.9.

The general *MCMC* settings are the same as with the time series wherever applicable, despite them being different implementations as explained in the methods section. Also, there is no way around the fact that some parameters in this estimation are absent and do not apply to the previous case – for instance the origin time of the epidemic (which for a time series is trivially defined as the first time point), the mutation rate and the tree itself (both of which have no impact on the incidence series). Otherwise, the estimates are generally comparable.

Biases are not nearly as pronounced as in the inference with the time series; Besides, the variables with gamma-distributed priors seem not to get much information from the likelihood and stay almost unaltered compared to their priors. The origin parameter, the time before the present when the epidemics starts, is fixed at 5852 days; this choice is

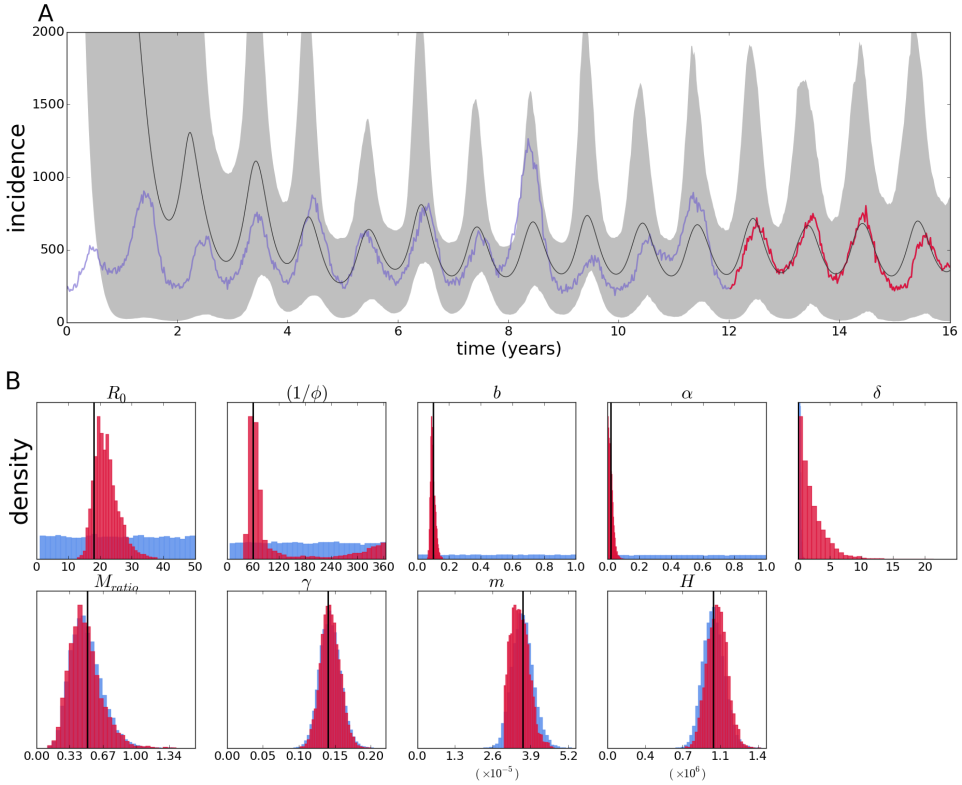


**Figure 4.3.9:** Reconstruction of the epidemic with a two-serotype model from sequences simulated by its corresponding individual-based model (A). Posterior distributions (crimson color) and priors (light blue) of epidemiological parameters (B).

discussed in the following sections. On the other hand, parameters with uninformative priors such as the mosquito mortality rate are quite well estimated; moreover, the basic reproductive number is quite accurately estimated.

### Inference for epidemiological data from the city of Rio de Janeiro, Brazil

We apply the methods tested above to epidemiological data from the city of Rio de Janeiro, Brazil. Weekly incidence data for a period of approximately four years between the end of 2009 and that of 2013 was used to fit both the vector SIRx2 and the two-serotype models.

Given the caveats observed for the simulated data, preliminary estimation was performed using the *MCMC* method for time series described above. In addition to the continuous model described above, a fixed scale (or reporting rate) parameter multiplies the incidence output to account for underreporting of cases in the recorded series – i.e. a parameter value lower than one means the actual epidemic is larger than the observed time series record.

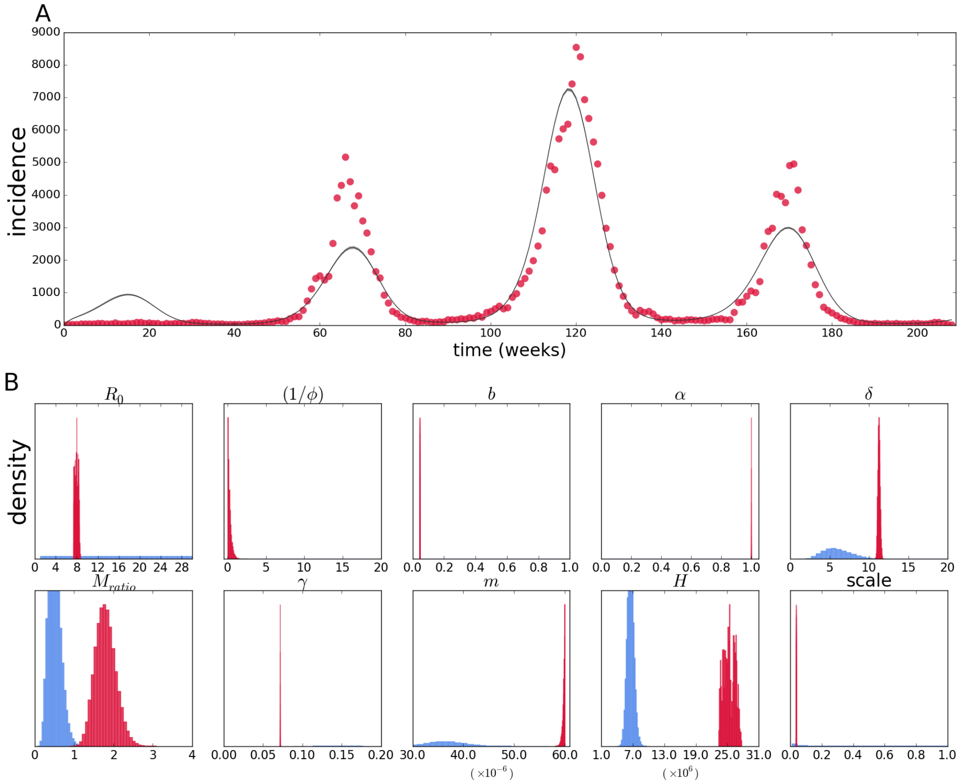


**Figure 4.3.10:** Fit of the vector SIRx2 continuous model with poisson likelihood to epidemiological data from Rio de Janeiro (A). Posterior distributions (crimson color) and priors (light blue) of epidemiological parameters and initial conditions (B).

The parameter estimates for the vector SIRx2 model are shown as posteriors (Figure 4.3.10B); the parameters with independent estimates (included in the inference method as prior probabilities) deviate from that expectation, notably the population size $H$, and human mortality rate $m$, both of which are grossly overestimated. It could be that there is a biological explanation for the departures from the independent estimates, but as seen in the previous subsection, bias and correlation between parameters is likely to affect the estimates. Another well accepted value is recovery happening in around a week $(1/\gamma)$; the posterior confidently places that at around twice that time. The biological parameters of interest without accepted values place the mosquito lifespan $(1/b)$ at around 22 days, and temporary immunity period $(1/\phi)$ at 1 day, while $R_0$ is estimated to be around 8. Parameters of more epidemiological interest are the reporting rate ("scale" parameter) close to only 4% and the vector to human ratio of close to 2 mosquitoes for every human host. The general fit (Figure 4.3.10A) is good, although the model predicts a lower incidence at the first and third peaks, as well as an early outbreak around the 20th epidemiological week, which is not present in the epidemiological data.

Next, the two-serotype model is fitted to the same data set; the results are shown in figure 4.3.11. The general fit is visibly improved. Interestingly, again some of the better known, or accepted, parameters deviate from expectation in the same direction as with the vector SIRx2 model; human lifespan $(1/m)$ is again underestimated, but population size $H$ is inferred to be close to the census-recorded number. Recovery $\gamma$ is again estimated to be twice as slow as the commonly accepted rate of a seven day long infection; mosquito mortality rate $b$ also has a similar value to that in the one-serotype model.

The robustness of the estimates can be interpreted as a strong signal in the data for these parameters, although it is difficult to know how much they are affected by the deviation in the independently-estimated values, as well as the others with no accepted values such as the temporary immunity period $1/\phi$. The robustness of some of the estimates can be further tested by fixing the parameters with well accepted values $(\gamma, m, H)$. These results are in appendix C; figure C.1.6 shows that while preventing parameter values very inconsistent with prior beliefs or independent estimates, most estimates are affected by the choice of free parameters estimated.

Contrary to the vector SIRx2 model, the estimate for $R_0$ is on the high end of the uni-

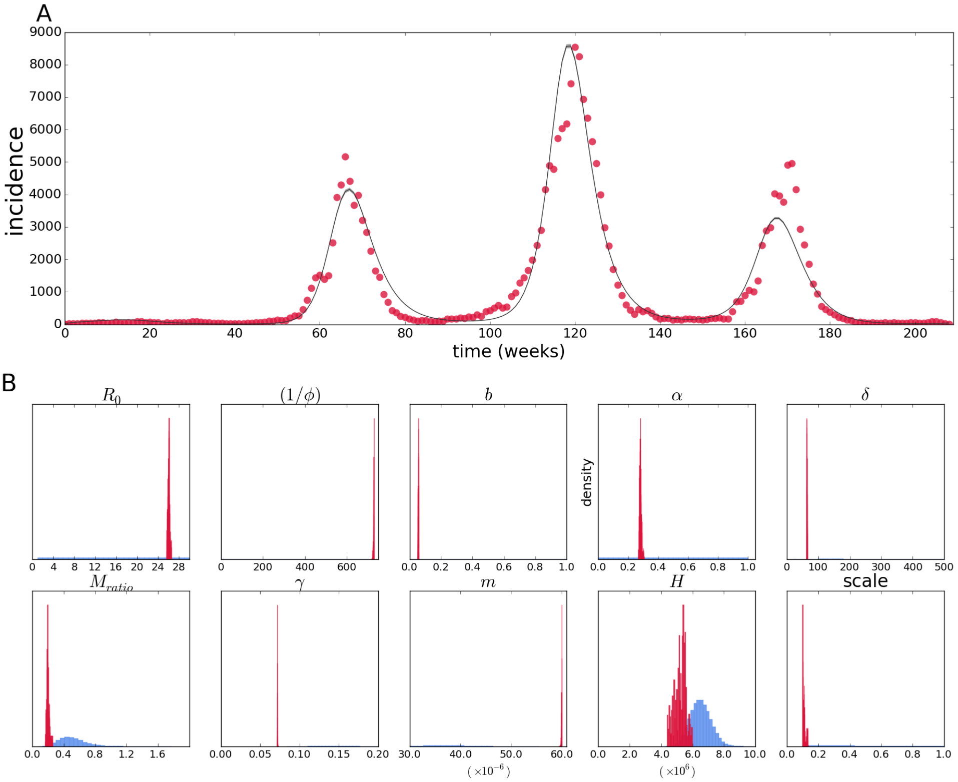


**Figure 4.3.11:** Fit of the two-serotype continuous model with poisson likelihood to epidemiological data from Rio de Janeiro (A). Posterior distributions (crimson color) and priors (light blue) of epidemiological parameters and initial conditions (B).

form prior probability assigned to the variable, around 26; the mosquito to human ratio is also opposite to the estimate of the previous model, being about twice as low as the mean of the prior distribution at 0.2, and the cross protection time is in the complete opposite side of the range at 729 days. Mosquito lifespan is close to the previous estimate, 18 days, and the reporting rate is on the order of 10%

The values inferred, however, must not be taken at face value, considering the limitations observed in the inference using the simulated data. We discuss that further in the next section.

## 4.4 Discussion

The issue of modeling dengue virus transmission is far from trivial, as is probably modeling most infectious diseases. Writing on a paper or programming a SIR-like model into a computer is indeed straightforward; that in its simplicity that alone can capture unintuitive features from disease transmission is quite surprising, but going beyond that is a long and winding road, and extending the basic models in the right direction even more so. It may also be reasonable to ignore variation and selection in some cases (even out of necessity), which greatly simplifies model formulation; still under these simplified models there is plenty to criticize in the existent body of work in modeling dengue transmission (Johansson et al. 2011). Basic simplifications whose consequences are taken for granted may also rear up their ugly heads when the model needs to be confronted to real data.

Inference is a logical next step, although it is an arduous task that may not be as rewarding as straightforward simulation. Despite perfection being far out of sight, the critical exercise can be a constructive one. The difficulties in the case of dengue are illustrated by the issue of multiple serotype interactions, and more subtle effects that are expected – like asymmetry between serotypes or antibody dependent enhancement.

From the forward simulation point of view, it is clear that the structure of the model radically affects the observed incidence – in what can be understood as the crucial factor of the availability of susceptibles – therefore interpreting epidemic series in terms of a single preferred model and its associated parameters is guaranteed to be problematic. The impact of structural differences for a given parameter set – as shown by the forward simulations – is evident, others may not be so clear beforehand, but are indisputable once we become aware of it; for instance, heterogeneity in any rates is more than expected, it must be present, and can have dramatic consequences (Gomes et al. 2012, 2016). These effects are likely to be larger than those of small asymmetries.

A baseline problem of this order is likely not to be solved by adding parameters and processes to the same model structure (Johansson et al. 2011), but instead the basic structures and multiple extensions have to be systematically compared. In theory this problem could be solved by formally comparing the performance of all available models against real data;

in practice, the task is a harder one: a record of a time series that aggregates all kinds of infections may lack the information necessary to distinguish between alternative models.

Identifiability analyses may uncover structural features of the model that may prevent inference of particular parameters, and may point to reparametrization of combinations that prevent structural identifiability issues (Bellu et al. 2007); however, it can be cumbersome to implement for larger models (Eisenberg et al. 2013), and it is uncommon (and possibly not feasible) that researchers in the different communities go about all the methods that could improve their results. The analysis of the data simulated by a continuous model strongly suggests that there are no severe structural issues with the method, although I did not perform any of the above-mentioned analyses. Alternative to structural identifiability analyses are more empirical assessments of parameter inferability (Toni et al. 2009), and model comparison based on information criteria (Gelman et al. 2013, chap. 7).

Formal model selection criteria rely on a reasonable fit by the different models; conversely, there should be enough information in the data to grant support to the most appropriate model – for instance, a more or less regular oscillating pattern produced by a two-serotype model may be easily reproduced by a one-serotype model, and the latter may be favored for having fewer parameters; what is more, a single cosine function (or a couple of them) could fit the pattern with fewer of parameters, but that does not change the fact that the data was produced by, and the correct model is still, a two-serotype model.

An example of the difficulties mentioned above are the estimates obtained here from incidence data of the city of Rio de Janeiro. Similar parameter values for different models suggest robustness in the estimates; nevertheless, large deviations from independent estimates may call into question these supposedly robust estimates. Conversely, disparate values for different models may point to the inadequacy of one of the models, and lack of robustness of estimates under model misspecification, but it does not guarantee that either estimate is correct – it is not possible to check deviation from the correct value if no independent estimate is available. Fixing the known parameters can be an additional constraint to the parameter space; however, it may force the remaining parameters into erroneous values due to the decreased flexibility in the model.

The more fine-grained the data is, the more precise can the comparison between models be, but that assumes that some of the models can explain the data well enough. It may, however, be the case that more detailed data causes the models only to fail more miserably than before. In the end, all models are approximations (at best, reasonable ones, but most likely rather crude ones), so that the multiple aspects of model and inference framework must also improve as data becomes better and more plentiful.

The use of sequence data for model-based inference presents itself as both an exciting perspective as well as a challenge: on the one hand it can make important distinctions such as genotypes and serotypes of pathogens, and by design allow inference about the entire population to be derived from a sample. It also carries, by default, information about more than one time series at the same time – i.e. incidence, prevalence and, if applicable, migration (Volz 2012). On the other hand fitting a model to sequences relies on elaborate constructs that are difficult to visualize and evaluate – some of the improvements that come with new kinds of data are therefore not without new issues.

One thing that seems to be unique to epidemiological models is that the data associated to the pathogens can be acquired simultaneously in different formats, so inference with one kind of data may be independently validated by another kind (e.g. sequence-based reconstruction of the epidemic can be compared to incidence data, as shown for pseudo data). Alternatively, these and other kinds of data (serological, vector population data, etc) can be used in combination to improve estimation, if the problem is scarcity of data.

This aspect is probably an important contrast to coalescent-based methods in fields like conservation genetics, where great strides have been made to incorporate processes like recombination (McVean & Cardin 2005), structured environments (as opposed to change in effective population size) (Mazet et al. 2016) and even allow inference from a single recombining genome (Li & Durbin 2011), but where very little validation of alternative models exists, especially with alternative types of data, which are not available for hundreds or thousands of years ago. Coalescent methods in epidemiology offer the opportunity of trying the methods, and validating them with more stringent criteria.

Increasingly it seems that scarcity is not the main problem (Pybus et al. 2013), but rather the difficulty of inferring multi-dimensional parameter sets (or constraining their space enough via independent estimation of individual parameters), and formally comparing full models in a way that makes the estimates actually useful, in addition to more subtle aspects of data collection (Frost et al. 2015).

Although its effects are clear in the individual-based simulations, the issue of stochasticity in the system state was not directly tackled here; therefore, a deterministic model can be forced to fit a different trajectory only by changing its parameters, even if the deviation is caused by chance. Methods such as particle filtering (or Sequential Monte Carlo) allow tracking of the stochastic system state along time (Ionides et al. 2006); these have been incorporated into genealogy-based estimation (Rasmussen et al. 2014a, 2011), potentially solving the issue of the effect of stochasticity in the population immunological states. These methods apply straightforwardly to population immunological states along time, which in the case on inference from time series is directly correspondent to the likelihood; in the coalescent-based estimation that is one of the components of the inference model, but it is not as easy to illustrate genetic drift in a similar way. Implementing methods that account for stochasticity is beyond the scope of this thesis, although the results shown strongly suggest that the simple fact that stochasticity is present can hamper progress in an otherwise simple task.

It can be tempting to focus specifically on the fine-tuning mathematical models, development new inference methods, and on extensive efforts to gather comprehensive data sets, but it is important to take into account how the weaknesses of each of the steps compound into a larger impediment. Concentrating particularly into just one of these (or other even more particular) aspects may prevent a realistic use of data-driven, model-based analysis. I have shown how model structure, assumptions about stochasticity, prior information and data requirements all deserve specific treatments lest the lack thereof introduces or amplifies biases and inaccuracies in the results, and therefore hope to have contributed to the integration of model building, inference frameworks, as well as future efforts to gather epidemiological data of different kinds.

# References

1. Adams, B., Holmes, E. C., Zhang, C., Mammen, M. P., Nimmannitya, S., Kalayanarooj, S. & Boots, M. 2006 Cross-protective immunity can account for the alternating epidemic pattern of dengue virus serotypes circulating in Bangkok. Proc Natl Acad Sci U S A 103, 14234–14239. (doi:10.1073/pnas.0602768103)

2. Andrieu, C., Doucet, A. & Holenstein, R. 2010 Particle Markov chain Monte Carlo methods. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 72, 269–342. (doi:10.1111/j.1467-9868.2009.00736.x)

3. Araújo, J. M. G. de, Bello, G., Schatzmayr, H. G., Santos, F. B. D. & Nogueira, R. M. R. 2009 Dengue virus type 3 in Brazil: a phylogenetic perspective. Mem. Inst. Oswaldo Cruz 104, 526–529.

4. Bellu, G., Saccomani, M. P., Audoly, S. & D'Angiò, L. 2007 DAISY: A new software tool to test global identifiability of biological and physiological systems. Computer Methods and Programs in Biomedicine 88, 52–61. (doi:10.1016/j.cmpb.2007.07.002)

5. Benson, D. A., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. & Sayers, E. W. 2015 GenBank. Nucleic Acids Research 43, D30–D35. (doi:10.1093/nar/gku1216)

6. Bhatt, S. et al. 2013 The global distribution and burden of dengue. Nature 496, 504–507. (doi:10.1038/nature12060)

7. Bianco, S., Shaw, L. B. & Schwartz, I. B. 2009 Epidemics with multistrain interactions:

the interplay between cross immunity and antibody-dependent enhancement. Chaos 19, 043123. (doi:10.1063/1.3270261)

8. Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C.-H., Xie, D., Suchard, M. A., Rambaut, A. & Drummond, A. J. 2014 BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. PLoS Comput Biol 10, e1003537. (doi:10.1371/journal.pcbi.1003537)

9. de Bruycker-Nogueira, F., Nogueira, R. M. R., Faria, N. R. D. C., Simões, J. B. S., Nunes, P. C. G., de Filippis, A. M. B. & Santos, dos, F. B. 2015 Insights of the genetic diversity of DENV-1 detected in Brazil in 25years: Analysis of the envelope domain III allows lineages characterization. Infection, Genetics and Evolution 34, 126–136. (doi:10.1016/j.meegid.2015.07.007)

10. Castro, M. G., de Nogueira, F. B., Nogueira, R. M. R., Lourenço-de-Oliveira, R. & Santos, dos, F. B. 2013 Genetic variation in the 3' untranslated region of dengue virus serotype 3 strains isolated from mosquitoes and humans in Brazil. Virol. J. 10, 3. (doi:10.1186/1743-422X-10-3)

11. Castro, M. G. de et al. 2012 Dengue virus type 4 in Niterói, Rio de Janeiro: the role of molecular techniques in laboratory diagnosis and entomological surveillance. Mem. Inst. Oswaldo Cruz 107, 940–945. (doi:10.1590/S0074-02762012000700017)

12. Chikaki, E. & Ishikawa, H. 2009 A dengue transmission model in Thailand considering sequential infections with all four serotypes. J Infect Dev Ctries 3, 711–722.

13. Drummond, A. J. & Bouckaert, R. R. 2015 Bayesian evolutionary analysis with BEAST. Cambridge: Cambridge University Press. (doi:10.1017/cbo9781139095112)

14. Dushoff, J., Plotkin, J. B., Levin, S. A. & Earn, D. J. D. 2004 Dynamical resonance can account for seasonality of influenza epidemics. Proc Natl Acad Sci U S A 101, 16915–16916. (doi:10.1073/pnas.0407293101)

15. Eisen, L. & Moore, C. G. 2013 Aedes( Stegomyia) aegyptiin the Continental United States: A Vector at the Cool Margin of Its Geographic Range. Journal of Medical Entomology 50, 467–478. (doi:10.1603/ME12245)

16. Eisenberg, M. C., Robertson, S. L. & Tien, J. H. 2013 Identifiability and estimation of multiple transmission pathways in cholera and waterborne disease. Journal of Theoretical Biology 324, 84–102. (doi:10.1016/j.jtbi.2012.12.021)

17. Favier, C., Schmit, D., Müller-Graf, C. D. M., Cazelles, B., Degallier, N., Mondet, B. & Dubois, M. A. 2005 Influence of spatial heterogeneity on an emerging infectious disease: the case of dengue epidemics. Proceedings of the Royal Society B: Biological Sciences 272, 1171–1177. (doi:10.1098/rspb.2004.3020)

18. Frost, S. D. W., Pybus, O. G., Gog, J. R., Viboud, C., Bonhoeffer, S. & Bedford, T. 2015 Eight challenges in phylodynamic inference. Epidemics 10, 88–92. (doi:10.1016/j.epidem.2014.09.001)

19. Frost, S. D. W. & Volz, E. M. 2010 Viral phylodynamics and the search for an 'effective number of infections'. Philosophical transactions of the Royal Society of London. Series B, Biological sciences 365, 1879–1890. (doi:10.1098/rstb.2010.0060)

20. Gelman, A., Carlin, J. B., Stern, H. S. & Rubin, D. B. 2013 Bayesian data analysis, Third Edition. CRC Press. (doi:10.1080/01621459.2014.963405)

21. Gomes, M. G. M., Aguas, R., Lopes, J. S., Nunes, M. C., Rebelo, C., Rodrigues, P. & Struchiner, C. J. 2012 How host heterogeneity governs tuberculosis reinfection? Proceedings of the Royal Society B: Biological Sciences 279, 2473–2478. (doi:10.1098/rspb.2011.2712)

22. Gomes, M. G. M., Gjini, E., Lopes, J. S., Souto-Maior, C. & Rebelo, C. 2016 A theoretical framework to identify invariant thresholds in infectious disease epidemiology. Journal of Theoretical Biology 395, 97–102. (doi:10.1016/j.jtbi.2016.01.029)

23. Gordo, I., Gomes, M. G. M., Reis, D. G. & Campos, P. R. A. 2009 Genetic diversity in the SIR model of pathogen evolution. Plos One 4, e4876. (doi:10.1371/journal.pone.0004876)

24. Gordo, I. & Campos, P. R. A. 2007 Patterns of genetic variation in populations of infectious agents. BMC evolutionary biology 7, 116. (doi:10.1186/1471-2148-7-116)

25. Griffiths, R. C. & Tavaré, S. 1994 Sampling Theory for Neutral Alleles in a Varying Environment. Philosophical Transactions of the Royal Society of London B: Biological Sciences 344, 403–410. (doi:10.1098/rstb.1994.0079)

26. Guzman, M. G. & Vazquez, S. 2010 The complexity of antibody-dependent enhancement of dengue virus infection. Viruses 2, 2649–2662. (doi:10.3390/v2122649)

27. Hadinegoro, S. R. et al. 2015 Efficacy and Long-Term Safety of a Dengue Vaccine in Regions of Endemic Disease. N. Engl. J. Med. 373, 1195–1206. (doi:10.1056/NEJMoa1506223)

28. Halstead, S. B. 2007 Dengue. Lancet 370, 1644–1652. (doi:10.1016/S0140-6736(07)61687-0)

29. Ionides, E. L., Bretó, C. & King, A. A. 2006 Inference for nonlinear dynamical systems. Proc Natl Acad Sci U S A 103, 18438–18443. (doi:10.1073/pnas.0603181103)

30. Johansson, M. A., Hombach, J. & Cummings, D. A. T. 2011 Models of the impact of dengue vaccines: A review of current research and potential approaches. Vaccine 29, 5860–5868. (doi:10.1016/j.vaccine.2011.06.042)

31. Keeling, M. J. & Rohani, P. 2008 Modeling infectious diseases in humans and animals. Princeton University Press.

32. Kingman, J. F. C. 1982 The coalescent. Stochastic Processes and their Applications 13, 235–248. (doi:10.1016/0304-4149(82)90011-4)

33. Kliks, S. C., Nisalak, A., Brandt, W. E., Wahl, L. & Burke, D. S. 1989 Antibody-Dependent Enhancement of Dengue Virus Growth in Human Monocytes as a Risk Factor for Dengue Hemorrhagic Fever.

34. Koelle, K. & Rasmussen, D. A. 2012 Rates of coalescence for common epidemiological models at equilibrium. Journal of the Royal Society, Interface / the Royal Society 9, 997–1007. (doi:10.1098/rsif.2011.0495)

35. Kraemer, M. U. G. et al. 2015 The global distribution of the arbovirus vectors Aedes aegypti and Ae. albopictus. eLife 4, e08347. (doi:10.7554/eLife.08347)

36. Kuehnert D. Phylodynamics package. https://github.com/BEAST2-Dev/phylodynamics. Accessed: 2016-06-15

37. Li, H. & Durbin, R. 2011 Inference of human population history from individual whole-genome sequences. Nature 475, 493–496. (doi:10.1038/nature10231)

38. Mazet, O., Rodríguez, W., Grusea, S., Boitard, S. & Chikhi, L. 2016 On the importance of being structured: instantaneous coalescence rates and human evolution–lessons for ancestral population size inference? Heredity (Edinb) 116, 362–371. (doi:10.1038/hdy.2015.104)

39. McVean, G. A. T. & Cardin, N. J. 2005 Approximating the coalescent with recombination. Philosophical Transactions of the Royal Society of London B: Biological Sciences 360, 1387–1393. (doi:10.1098/rstb.2005.1673)

40. Miagostovich, M. P., Sequeira, P. C., Santos, Dos, F. B., Maia, A., Nogueira, R. M. R., Schatzmayr, H. G., Harris, E. & Riley, L. W. 2003 Molecular typing of dengue virus type 2 in Brazil. Rev. Inst. Med. Trop. Sao Paulo 45, 17–21.

41. Miagostovich, M. P., Santos, dos, F. B., Fumian, T. M., Guimarães, F. R., da Costa, E. V., Tavares, F. N., Coelho, J. O. & Nogueira, R. M. R. 2006 Complete genetic characterization of a Brazilian dengue virus type 3 strain isolated from a fatal outcome. Mem. Inst. Oswaldo Cruz 101, 307–313.

42. Nagao, Y. & Koelle, K. 2008 Decreases in dengue transmission may act to increase the incidence of dengue hemorrhagic fever. Proceedings of the National Academy of Sciences of the United States of America 105, 2238–2243. (doi:10.1073/pnas.0709029105)

43. Otero, M. & Solari, H. G. 2010 Stochastic eco-epidemiological model of dengue disease transmission by Aedes aegypti mosquito. Math Biosci 223, 32–46. (doi:10.1016/j.mbs.2009.10.005)

44. Paradis, E. coalescentMCMC R Package. http://colgem.r-forge.r-project.org/ Accessed: 2016-06-15

45. Patil, A., Huard, D. & Fonnesbeck, C. J. 2010 PyMC: Bayesian Stochastic Modelling in Python. Journal of statistical software 35, 1–81.

46. Pybus, O. G., Fraser, C. & Rambaut, A. 2013 Evolutionary epidemiology: preparing for an age of genomic plenty. Philosophical transactions of the Royal Society of London. Series B, Biological sciences 368, 20120193–20120193. (doi:10.1098/rstb.2012.0193)

47. Rasmussen, D. A., Volz, E. M. & Koelle, K. 2014 Phylodynamic Inference for Structured Epidemiological Models. PLoS Comput Biol 10, e1003570. (doi:10.1371/journal.pcbi.1003570)

48. Rasmussen, D. A., Boni, M. F. & Koelle, K. 2014 Reconciling phylodynamics with epidemiology: the case of dengue virus in southern Vietnam. Mol Biol Evol 31, 258–271. (doi:10.1093/molbev/mst203)

49. Rasmussen, D. A., Ratmann, O. & Koelle, K. 2011 Inference for Nonlinear Epidemiological Models Using Genealogies and Time Series. PLoS Comput Biol 7, e1002136. (doi:10.1371/journal.pcbi.1002136)

50. Rico-Hesse, R. 2010 Dengue virus virulence and transmission determinants. Curr. Top. Microbiol. Immunol. 338, 45–55. (doi:10.1007/978-3-642-02215-9_4)

51. Rohani, P., Green, C. J., Mantilla-Beniers, N. B. & Grenfell, B. T. 2003 Ecological interference between fatal diseases. Nature 422, 885–888. (doi:10.1038/nature01542)

52. Ross, R. 1916 An Application of the Theory of Probabilities to the Study of a priori Pathometry. Part I. Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences 92, 204–230. (doi:10.1098/rspa.1916.0007)

53. Salje, H. et al. 2012 Revealing the microscale spatial signature of dengue transmission and immunity in an urban population. Proceedings of the National Academy of Sciences of the United States of America 109, 9535–9538. (doi:10.1073/pnas.1120621109)

54. Santos, dos, F. B., Miagostovich, M. P., Nogueira, R. M. R., Edgil, D., Schatzmayr, H. G., Riley, L. W. & Harris, E. 2002 Complete nucleotide sequence analysis of a Brazilian dengue virus type 2 strain. Mem. Inst. Oswaldo Cruz 97, 991–995.

55. De Simone, T. S., Nogueira, R. M. R., Araújo, E. S. M., Guimarães, F. R., Santos, F. B., Schatzmayr, H. G., Souza, R. V., Teixeira Filho, G. & Miagostovich, M. P. 2004 Dengue virus surveillance: the co-circulation of DENV-1, DENV-2 and DENV-3 in the State of Rio de Janeiro, Brazil. Trans. R. Soc. Trop. Med. Hyg. 98, 553–562. (doi:10.1016/j.trstmh.2003.09.003)

56. Sistema de Informações de Agravos de Notificação : http://ces.ibge.gov.br/base-de-dados/metadados/ministerio-da-saude/sistema-de-informacoes-de-agravos-de-notificacao-sinan.html. Accessed: 2016-06-23

57. Smith, D. L., Battle, K. E., Hay, S. I., Barker, C. M., Scott, T. W. & McKenzie, F. E. 2012 Ross, macdonald, and a theory for the dynamics and control of mosquito-transmitted pathogens. Public Library of Science. (doi:10.1371/journal.ppat.1002588)

58. Stoddard, S. T. et al. 2013 House-to-house human movement drives dengue virus transmission. Proceedings of the National Academy of Sciences of the United States of America 110, 994–999. (doi:10.1073/pnas.1213349110)

59. Toni, T., Welch, D., Strelkowa, N., Ipsen, A. & Stumpf, M. P. H. 2009 Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. Journal of The Royal Society Interface 6, 187–202. (doi:10.1098/rsif.2008.0172)

60. Vasilakis, N., Cardosa, J., Hanley, K. A., Holmes, E. C. & Weaver, S. C. 2011 Fever from the forest: prospects for the continued emergence of sylvatic dengue virus and its impact on public health. Nat. Rev. Microbiol. 9, 532–541. (doi:10.1038/nrmicro2595)

61. Vazquez-Prokopec, G. M., Kitron, U., Montgomery, B., Horne, P. & Ritchie, S. A. 2010 Quantifying the spatial dimension of dengue virus epidemic spread within a tropical urban environment. PLoS neglected tropical diseases 4, e920. (doi:10.1371/journal.pntd.0000920)

62. Villar, L. et al. 2015 Efficacy of a tetravalent dengue vaccine in children in Latin America. N. Engl. J. Med. 372, 113–123. (doi:10.1056/NEJMoa1411037)

63. Volz, E. M., Koelle, K. & Bedford, T. 2013 Viral phylodynamics. PLoS Comput Biol 9, e1002947. (doi:10.1371/journal.pcbi.1002947)

64. Volz, E. M. 2012 Complex Population Dynamics and the Coalescent Under Neutrality. Genetics 190, 187–201. (doi:10.1534/genetics.111.134627)

65. Volz, E. M. Ratmann, O., Severson, E. R. rcolgem R Package http://colgem.r-forge.r-project.org/ Accessed: 2016-06-15

59. Wakeley J. 2009 Coalescent Theory: An Introduction. Greenwood Village: Roberts & Company Publishers.

60. WHO 2016 Dengue vaccine: WHO position paper – July 2016. Wkly. Epidemiol. Rec. 91, 349–364.

61. Wikramaratna, P. S., Simmons, C. P., Gupta, S. & Recker, M. 2010 The effects of tertiary and quaternary infections on the epidemiology of dengue. Plos One 5, e12347. (doi:10.1371/journal.pone.0012347)

*"Inference is not the inverse of a hypothesis test."*

Andrew Gelman

# C

# Additional MCMC chains for dengue transmission models

## C.1 Time series-based inference

### C.1.1 One-serotype model inference

The results for the inference with the one-serotype model from simulated data is shown in figure C.1.1. Biases similar in magnitude are also observed for this simpler model.

### C.1.2 Inference of a reduced set of parameters

The inference from time series generated by a continuous model with noise added independently to each time point serves as proof-of-concept for time series as suitable data to

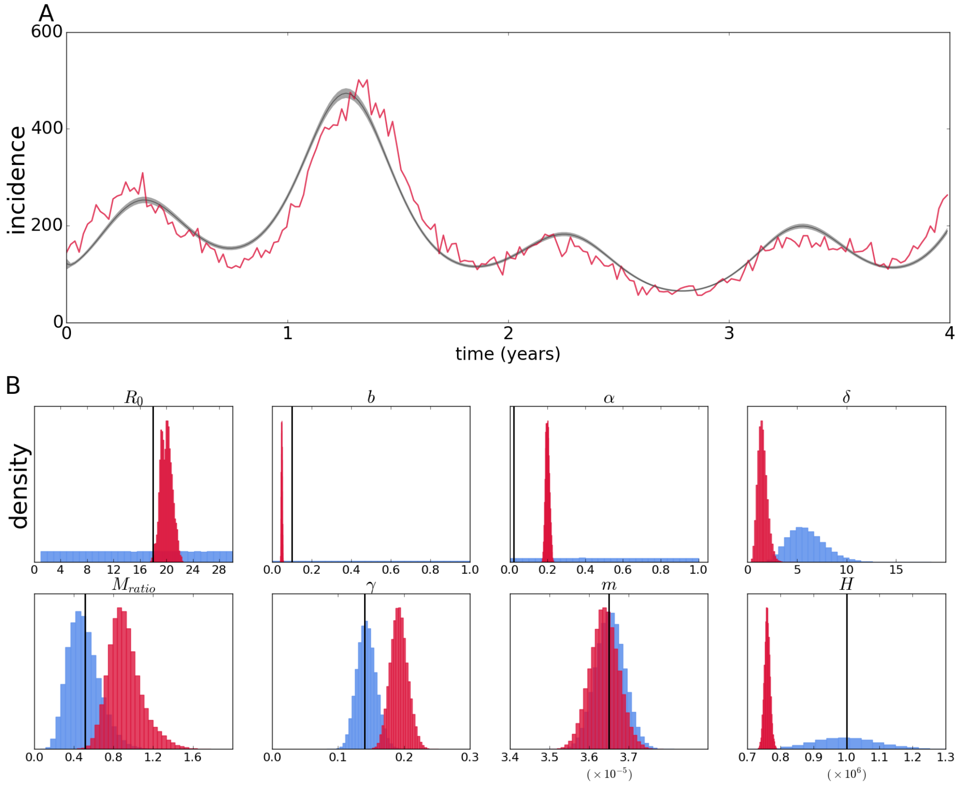


**Figure C.1.1:** Fit of the one-seroytpe model to stochastically simulated data (A). Posterior distributions (crimson color) and priors (light blue) of epidemiological parameters and initial conditions (B)

infer the parameters in the model; nevertheless, the jump to inference from a fully stochastic model is quite a big one, and testing some intermediate conditions for inference could be warranted. Assuming all epidemiological rate parameters except for $R_0$ were known, the estimates for the parameter are shown in figure C.1.2, for the case where the initial conditions are inferred and that when these are also assumed to be known.

The bias in the estimate is of the order of around 10% in the first case, and less than 5% when $R_0$ is the only estimated parameter.

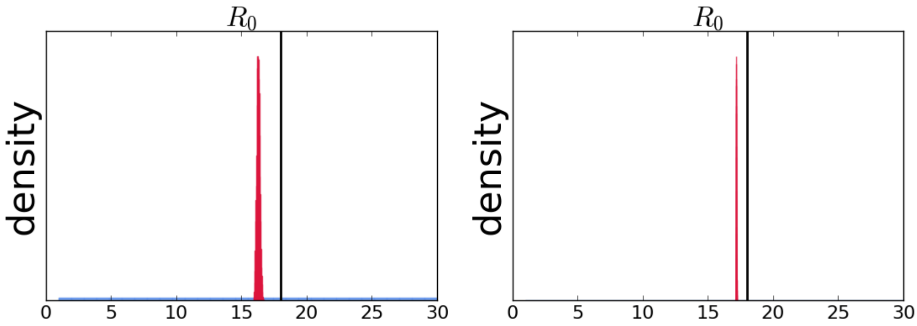


**Figure C.1.2:** Estimates of $R_0$ as single epidemiological parameter for the case where it is estimated together with the initial conditions of the system (left), or as the only unknown parameter (right)

Interestingly, when assuming that initial conditions are known, but estimating other parameters for which independent estimates are difficult to obtain, $R_0$ is underestimated by less than 2%.

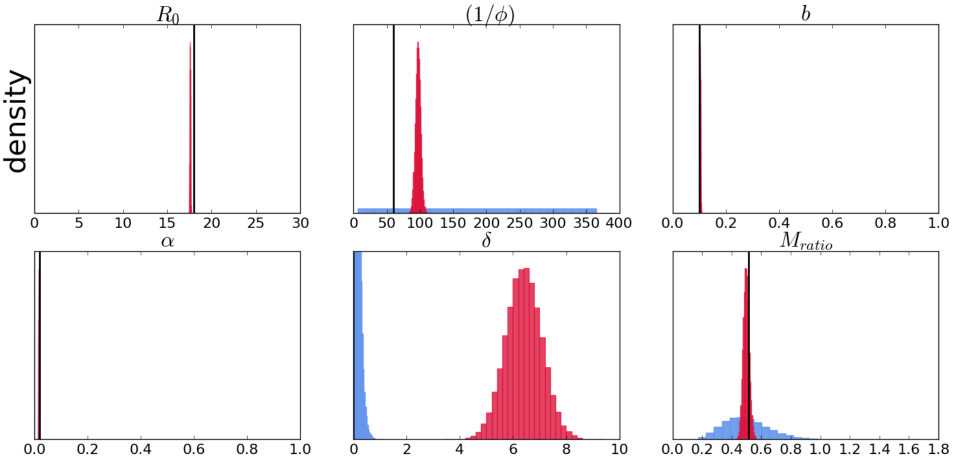


**Figure C.1.3:** Estimates of a subset of epidemiological parameters assuming initial conditions are known.

These results shown that identifiability and estimation of multiple parameters is an issue, but that the estimation of initial conditions alone can bias the estimation. It is worth noting that even when the known initial values are fixed – not estimated – the time series is still subject to stochasticity, which is not fully accounted by this method or the assumption of known parameters, and can therefore bias the estimates.

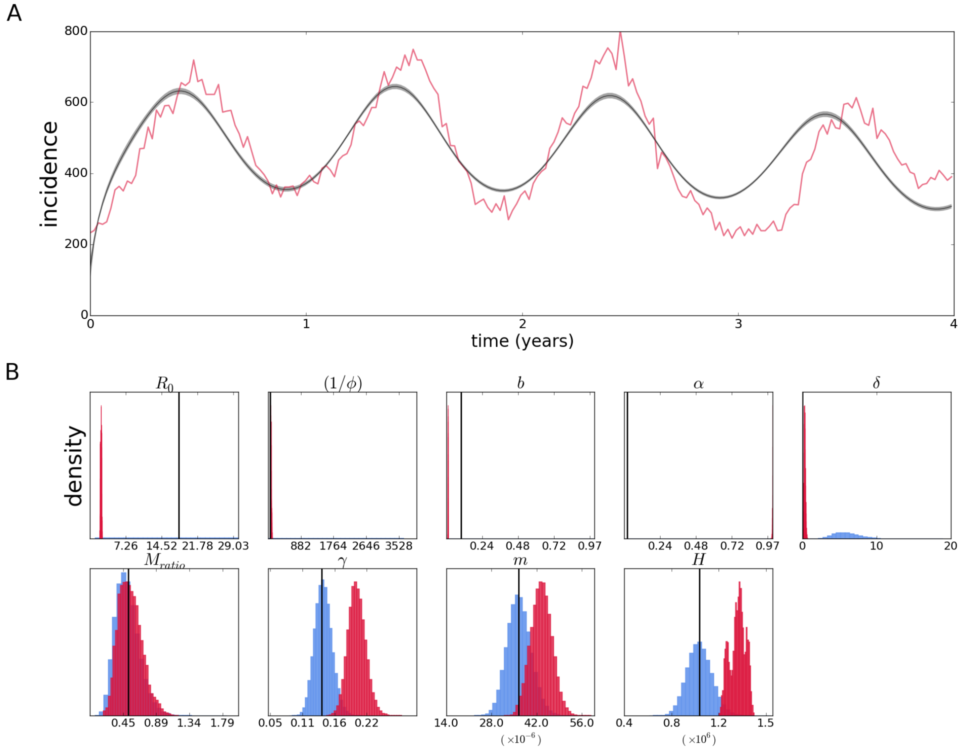


**Figure C.1.4:** Fit of the vector-SIRx2 continuous model with poisson likelihood to data simulated from a stochastic two-serotype model (A). Posterior distributions (crimson color) and priors (light blue) of epidemiological parameters and initial conditions (B).

## C.1.3 Inference under model misspecification

The vector SIR model is a submodel of the vector SIRx2 and the two-serotype model, both of which collapse into it when the immunity waning parameters equals zero (the latter also collapses when only one serotype is circulating). While it is theoretically possible that either model will correctly estimate the appropriate parameters to be zero, in practice the over-specification of the models may affect parameter estimates; conversely, under-specification is guaranteed not to have all parameters, and yet the estimates for the parameters that are still included could be robust to that kind of misspecification.

The result of fitting a one serotype model to a single time series recorded from two-serotypes interacting (but indistinguishable from the data) is shown in figure C.1.4.

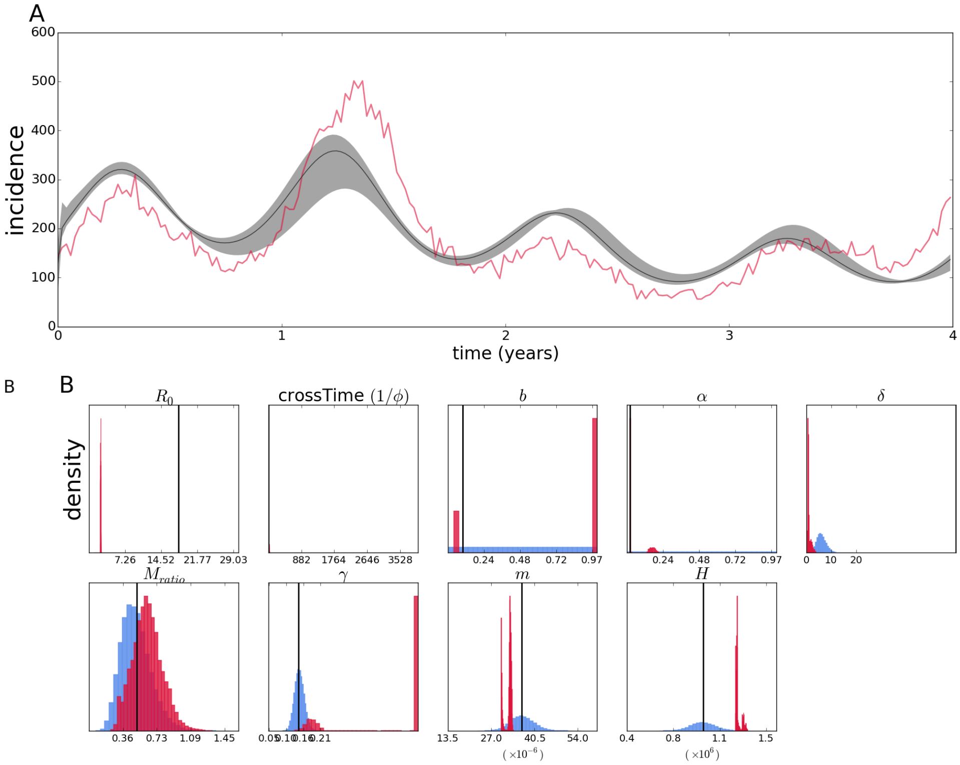


**Figure C.1.5:** Fit of the vector-SIRx2 continuous model with poisson likelihood to data simulated from a stochastic vector-SIR model (A). Posterior distributions (crimson color) and priors (light blue) of epidemiological parameters and initial conditions (B).

Conversely, a model with more than one infected compartment can be fit to data from a single type of infection; that kind over overfitting is shown in figure C.1.5.

While the fit is visually acceptable, the inference is unable to estimate most parameter values correctly; unexpectedly, convergence seems better against the data simulated by a two-serotype model, although the biases are clear.

This highlights the importance of having enough alternative models to test against the data, and shows that even seemingly convergent estimates may hide significant biases.

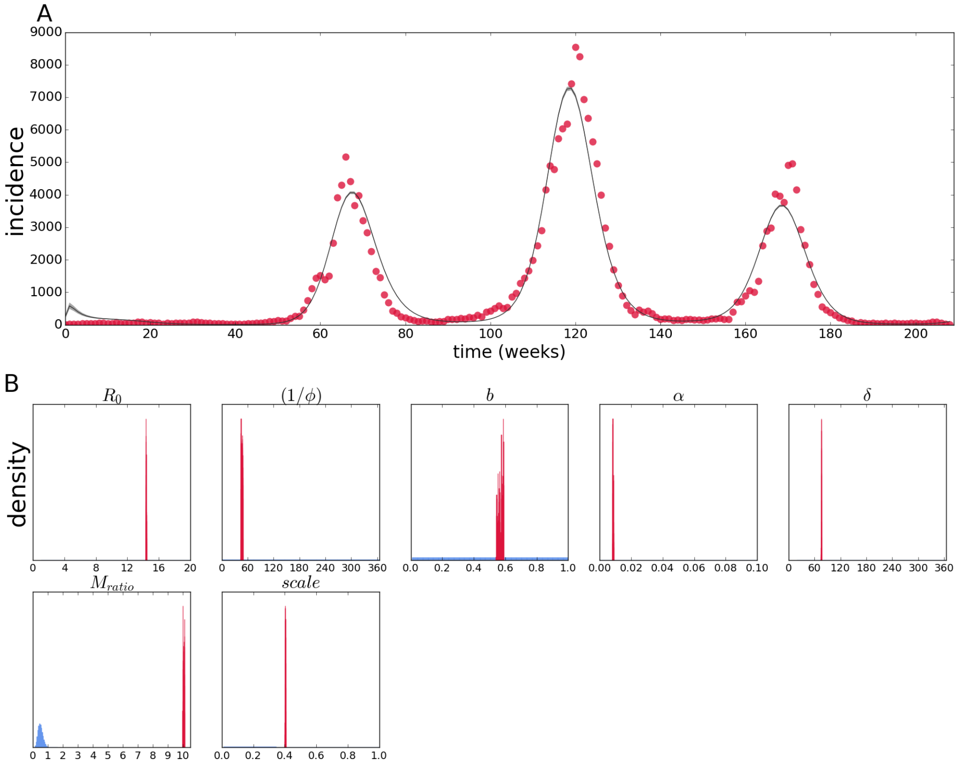


**Figure C.1.6:** Fit of the two-serotype continuous model with poisson likelihood to epidemiological data from Rio de Janeiro (A), with $\gamma$, $m$, $H$ parameters fixed. Posterior distributions (crimson color) and priors (light blue) of epidemiological parameters and initial conditions (B).

### C.1.4 Inference of a reduced set of parameters for robustness testing

The results for inference from epidemiological data with the $\gamma$, $m$, and $H$ parameters fixed are shown in figure C.1.6. The model fit is not visually affected by the fixed parameters. Parameters $\varphi$ and $M_{ratio}$ are radically affected; mosquito mortality rate $b$ also changes considerably, while the rest shift more slightly. Particularly, $R_0$ is shifted down to 14.

## C.2 Genealogy-based inference

### C.2.1 Alternative priors

Considering the effect of stochasticity on the system state, as well as the apparent lack of information about it in the data, it is expected that different priors for these parameters may affect the inference. Arguably the strongest assumption about the initial conditions is that they were completely known, and therefore are kept fixed for estimation of the remaining parameters for the epidemiological rates. If the conditions are observed, or estimated in any manner, they can be treated as less of a certainty and incorporated probabilistically, into the likelihood function in addition to the coalescent process, or time series data.

A weaker imposition is to consider information about the system state as priors in the bayesian method. That too would have different gradations, ranging from strong priors with a mode at a given value, uniform priors over some larger or narrower ranges, or some other sensible choice of priors. In any case, the choice must be justified, but is probably not the strongest assumption embedded into the inference framework. I show the effect of alternative choice of priors for the population initial conditions on the estimates, and detail in which case the priors could be an appropriate choice.

In the case where there is an estimate for the state of every population described by the model, but as mentioned above it is not treated as true, or incorporated into the likelihood, gamma-distributed priors could be used to make the *MCMC* chain tend towards those values unless there is stronger information in the likelihood that brings them elsewhere. The results for this case, all else being equal to the main text, are shown in figure C.2.1. Since there seems not to be a strong signal to estimate the population state parameters

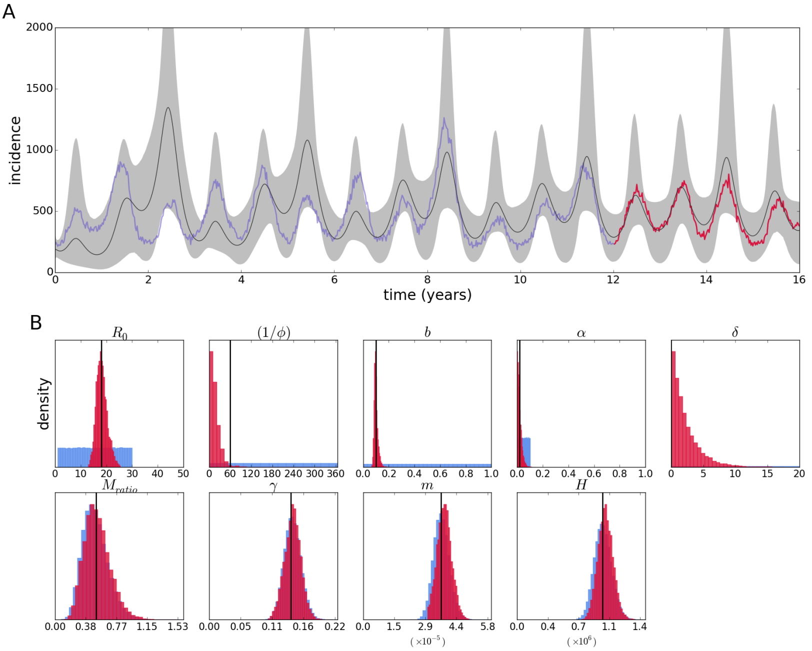


**Figure C.2.1:** Reconstruction of the epidemic (A), and posterior distributions (crimson color) and priors (light blue) of epidemiological parameters (B) for the two-serotype model and gamma-distributed priors for the initial conditions.

when using priors that are uniform or nearly so, using these parameters will essentially recover the prior distribution. The estimate for $R_0$ is more precise, although $1/\varphi$ is biased to lower values, and $\alpha$ is less precise. Otherwise, the difference is not striking.

The opposite of the above is using uniform priors over an extremely range of possible values (since here we know that the total population is a million people, and that that is enforced by the gamma-distributed prior on that parameter, this is roughly translated to uniform over the range from zero to a million, which could also be made even larger for the sake of it). For these wide uniform priors, the results are shown in figure C.2.2.

For the human infected compartments shown in panel C of figure C.2.2, instead of having the posteriors squeezed close to zero, they are distributed over a wider interval,

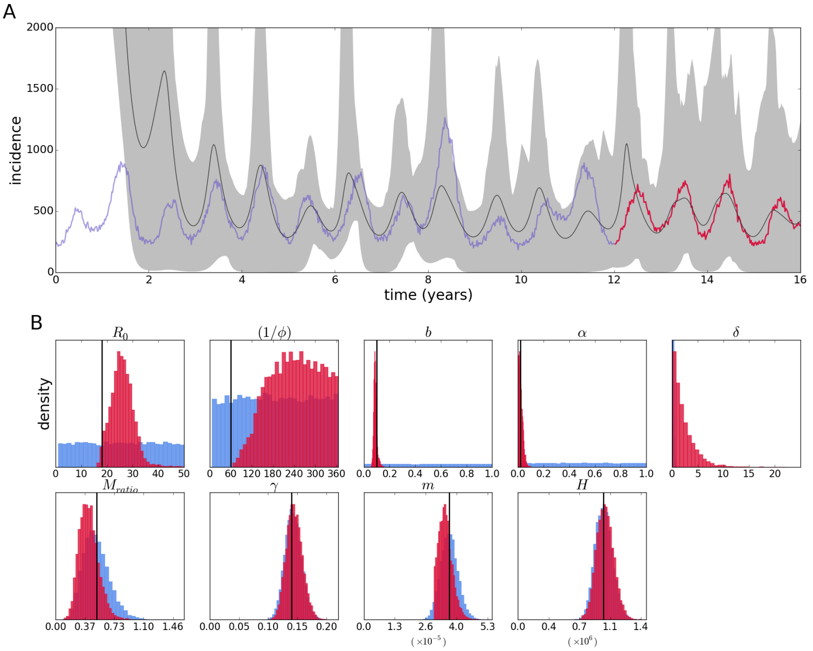


**Figure C.2.2:** Reconstruction of the epidemic (A), and posterior distributions (crimson color) and priors (light blue) of epidemiological parameters (B) for the two-serotype model and wide uniform priors for the initial conditions.

although the susceptible and recovered parameters do not differ notably from what was observed before. The most striking difference is the vector infected compartments are estimated to have values on the higher extreme of the distribution, which in this case is known to be wrong, as both human and vector infections oscillate in around a few thousand of infections at any point in time. As for the epidemiological parameters, shown in panel B, $R_0$ is overestimated in this case, and $1/\varphi$ is overestimated with very little precision, covering most of the uniform prior (with notable exception of the actual value). Nevertheless, reconstruction of the epidemic is not severely affected, but does have wider credibility intervals.

A less conservative use of uniform priors is that with different ranges depending on the populations. Because the human infected compartments are at least partially observed, it may be reasonable to define a narrower uniform distribution for these variables More generally it is easy to do that for the simulated data because it is known what is a reasonable range for each variable; nevertheless, when that is unknown or doubtful there is the risk of limiting the interval too much and excluding the correct range. The results for this specification for the priors is shown in figure C.2.3. The estimates of the epidemiological parameters (figure C.2.3, panel B) are similar to those obtained before, with somewhat less precision in the estimate of $R_0$ compared to the results in the main text, or the gamma-distributed priors, but the posterior for $1/\varphi$ includes the true value better the the latter - although the distribution is quite wide. Regarding the population states, the posteriors are not very precise compared to the priors, but even that can mean a better precision than

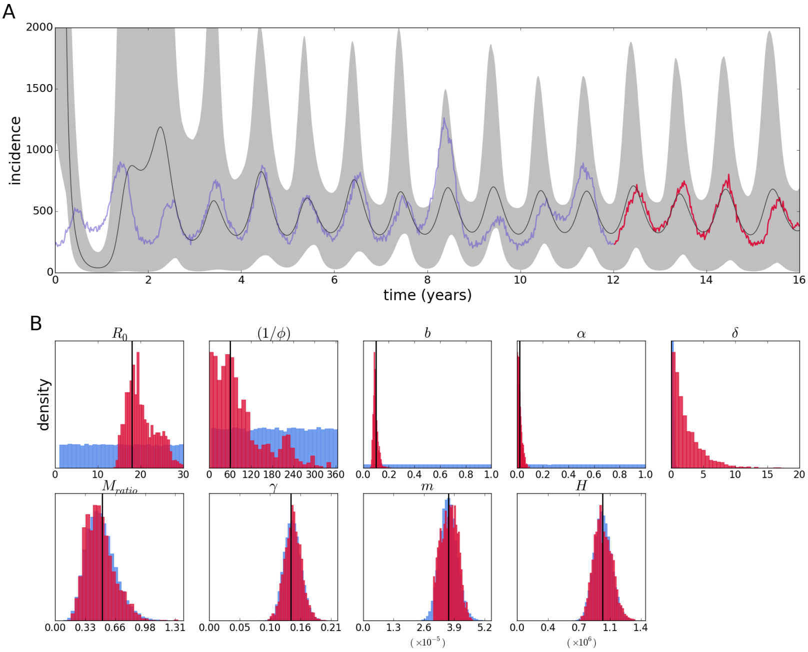


**Figure C.2.3:** Reconstruction of the epidemic (A), and posterior distributions (crimson color) and priors (light blue) of epidemiological parameters (B) for the two-serotype model and wide uniform priors for the initial conditions.

the posteriors obtained with wide uniform priors, since the latter can span 3 or 4 orders of magnitude more. The reconstruction of the epidemic has somewhat narrower credibility intervals, and overall inference is improved compared to the wide priors.

### C.2.2 One serotype subset for two-serotype inference

Furthermore, inference for the two-serotype model can be made by computing the likelihood of the coalescent process for only one serotype, so results are shown for that case in conditions similar to that of the two serotypes. The results for gamma-distributed priores are shown in figure C.2.4. For completely flat priors, using the subset of the data corre-

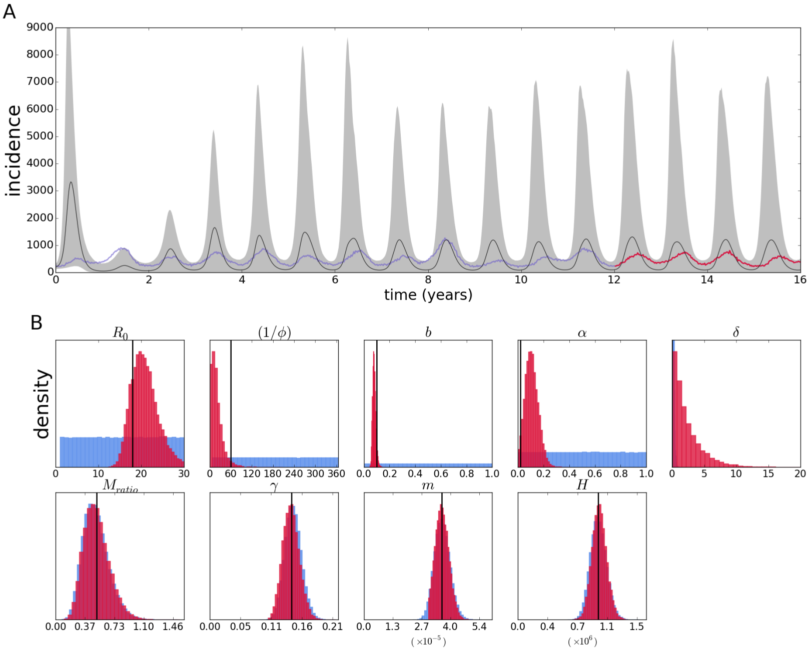


**Figure C.2.4:** Reconstruction of the epidemic (A), and posterior distributions (crimson color) and priors (light blue) of epidemiological parameters (B) for the two-serotype model and gamma-distributed priors for the initial conditions, using data from only one of the serotypes ("serotype 2").

sponding to the second seroytpe, the results are shown in figure C.2.5. For the counterpart of the data set, that is the first serotype, the estimates are shown in figure C.2.6. In this case, the estimates can be seen to be worse than the alternative serotype; because the amount of data is comparable, the difference could in principle be attributed to stochasticity in the mutations that happened to appear and were sampled in each data set.

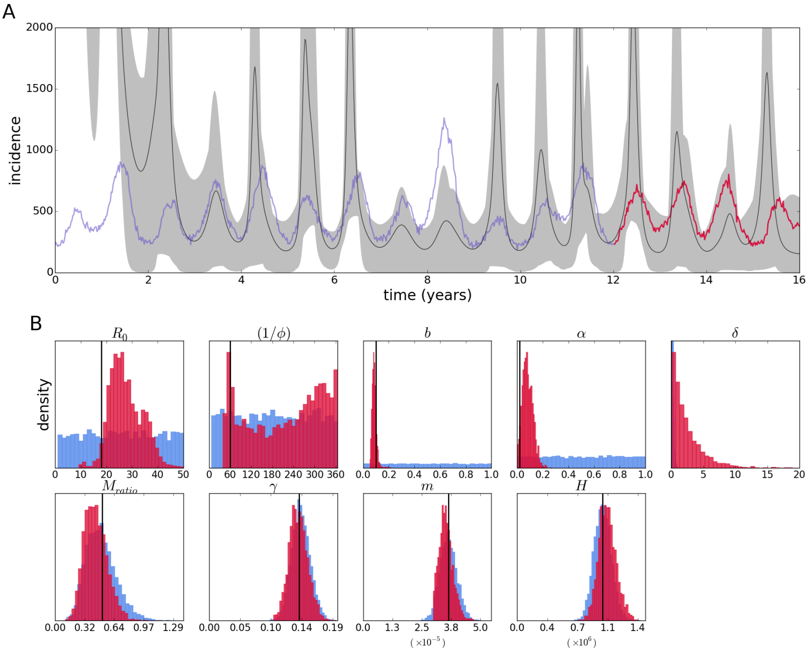


**Figure C.2.5:** Reconstruction of the epidemic (A), and posterior distributions (crimson color) and priors (light blue) of epidemiological parameters (B) for the two-serotype model and wide uniform priors for the initial conditions, using data from only one of the serotypes ("serotype 2").

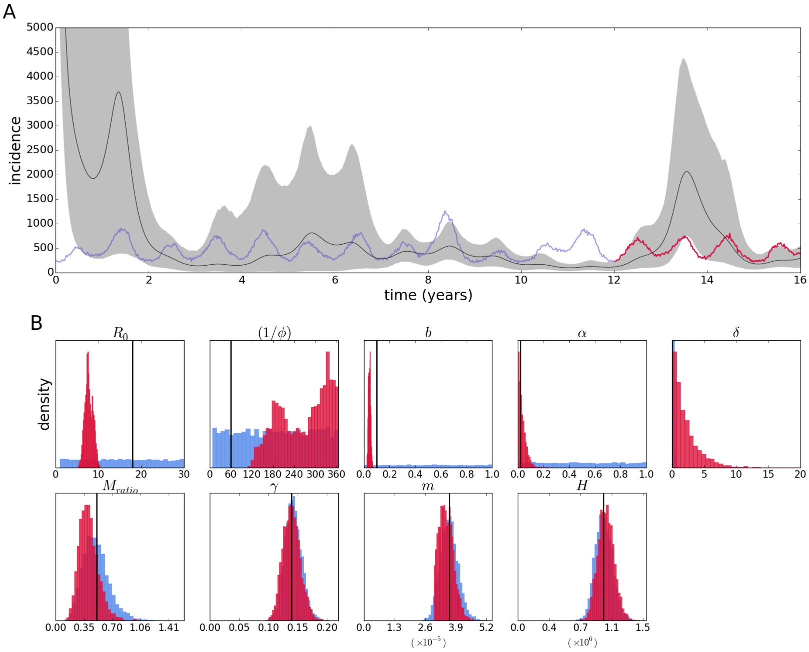


**Figure C.2.6:** Reconstruction of the epidemic (A), and posterior distributions (crimson color) and priors (light blue) of epidemiological parameters (B) for the two-serotype model and wide uniform priors for the initial conditions, using data from only one of the serotypes (“serotype 1”).

*"It is clear, then, that the idea of a fixed method, or of a fixed theory or rationality, rests on too naive a view of man and his social surroundings. To those who look at the rich material provided by history, and who are not intent on impoverishing it in order to please their lower instincts, their craving for intellectual security in the form of clarity, precision, "objectivity", "truth", it will become clear that there is only one principle that can be defended under all circumstances and in all stages of human development. It is the principle: anything goes."*

Paul Feyerabend

# 5

# General discussion, perspectives and conclusions

## 5.1 Modeling, and integrating or confronting multiple organization scales

Infection is a process that develops at many different levels, from molecular and cell processes, to within-host proliferation, transmission to different host, and sustaining a population-wide epidemic. It cannot be said that any one scale of organization is dominant over all others; it likely depends on the magnitude of the effects and how they carry over to the higher levels (or feedback to the lower ones). Nevertheless, it is almost always necessary to abstract from the complete integration between all of them and focus on one or a couple of scales at most.

Evolution will act on the mechanisms at all levels, provided again that the magnitude of the process is relevant at one or more scales, and there is enough variation in the traits to allow change in the population of host, pathogen, symbionts, or any other participants in the interactions; the result will be a continuous competition or some equilibrium between them (Souto-Maior 2015b). More generally, and ignoring the somewhat artificial separation between species, the system is just like evolution of different traits in a single species subject to trade-offs or changing environmental constraints. Alternatively, in the exceptional case that selection is not at all important for the system (or, more likely, that it is not significant at the time scale considered) neutral evolution still sculpts defined patterns that contain information about the population growth, size, and structure (Volz 2012), making genomes into a sort of recording machine.

The systems of interest in this thesis were concentrated on two scales: host and population. Population models (as in chapters 2 and 4) reduce the host to a rate or probability, e.g. infection of a host is treated as a random number draw with a binary outcome – successful or failed transmission – and the model makes sense only insofar as a group of these simplistic hosts mimic the behavior of a real epidemic (Keeling & Rohani 2008). Some of these probabilities or rates can be be measured, or estimated directly at the host level – e.g. the recovery rate can be estimated to be the inverse of the usual recovery period of 7 days for many viral diseases, while death rate can be estimated to be the inverse of the average life span of around 70 years, for instance, or a more refined age-dependent death rate if the population model is age-structured – while others are the emerging host-level outcome of a within-host process, like the susceptibility/transmission rate. Because the transmission rate is derived as a product of contact and susceptibility (Keeling & Rohani 2008), this emergent property can be derived as a contact-outcome, or dose-response model(Gomes et al. 2014; Pessoa et al. 2014) – and that is the purpose of chapter 1.

Alternatively, the within-host process can be described explicitly (as in chapter 3); in this case the host is treated as the environment where virus replication competes with its immune system, and is a dynamic process instead of an instantaneous outcome. At this point, it should be clear that integrating both levels (while conceptually and mathematically possible) would be extremely complicated, and require accurate (reasonably so, at least) estimates of individual parameters. Alternatively, the set of parameters for this multiscale model could be inferred if enough data is available, but the task is probably no easier, and despite a few serious attempts, the approach is has not been and entirely successful one (Day et al. 2011; Mideo et al. 2011). Instead of a "total modeling" approach, I argue that accepting the limitations of descriptions at any one level and trying to make progress from comparing the outcomes at the interfaces of the different levels is a more fruitful strategy.

For instance, the work described in chapter 1 forwards the idea that the susceptibility parameter or distribution can be estimated from a dose-response model (Gomes et al. 2014; Pessoa et al. 2014); the within-host model in chapter 3 describes the dynamics of viral titers ultimately resulting in establishment or clearing of (i.e. probability of) infection for any given dose – therefore both approaches yield a dose-response relationship, and ultimately they should be compatible. Beyond the connection between levels, the results at the interfaces can be contrasted. The dose-response models in chapter 1 infer a homogeneous probability of infection for hosts without *Wolbachia* and distributed, or heterogeneous, probability for symbiont-carrying hosts (Pessoa et al. 2014). In terms of the within-host model this can be seen in at least two perspectives: heterogeneity is an emergent property of a different set of microbe-response parameters, which generate a shallow dose-infectivity profile; or, heterogeneous susceptibility at the host level stems from heterogeneity in the within-host (sub-organismal) parameters.

The latter hypotheses were not directly tested, but it is in principle possible to compare the output of the within-host results (for instance stochastic runs of the model in chapter 3), with that of the host-level dose-response model (chapter 1). This remains as a perspective for future work. This piecewise, more artisanal than automated approach can be seen as defeatist by the more megalomaniac or believers in silver bullets, but in the face of the limited success of the more data-intensive, brute force approaches, it can also be argued that it is a cleverer one after all.

## 5.2 Time-dependent variables and extending partially observed data sets for dynamics models

In any given system, not all variables are recorded simultaneously. Even if a model with multiple variables is used to interpret the system, most of the time only a fraction of them is observed. In the case of disease transmission models, it may be of interest to describe incidence or prevalence (which relate to the new and current infected individuals), but the other compartments are also necessary to fully describe the state of the system. When estimating the parameters of the model, including the estimate of incidence/prevalence along time, the estimate of other compartments – such as recovered individuals – is also produced; therefore, two possibilities arise: this data can also be obtained and used for inference, or it can be independently compared to the estimate obtained from the main data set (e.g. incidence alone).

There is of course a reason why the number of recovered individuals is normally not recorded: it would require assaying every single person (or at least a reasonably large sample of people) at every point in time for antibodies against the pathogen, as opposed to only observing individuals who develop the disease. A third possibility that requires less effort and puts less weight in this additional data, would be trying to record the state of the system (i.e. the value of all compartments) at a single, early point in time, and using the data as prior information for the initial state of the system.

For systems such as a within-host microbe-response model, the easiest variable to be recorded is pathogen levels (as described in chapter 3); other dynamic variables are inferred from that. Unlike epidemiological models, where unobserved variables are nevertheless easily defined, an immune-response variable can be hard to specify, instead being a conceptual construct that aggregates observables, but lacking a defined metric. Despite not being nearly as straightforward as the previous example, trying to observe a correlation of the inferred response with an independently-recorded time series of host responses could nevertheless be a good indication that the model is representing a relevant component of host immunity

Generally speaking, different kinds of data can be combined to increase power of the inference, or at times left out to allow independent validation of the estimates; that can be especially important when sequence data is used, since it is not trivial to visually inspect the model fit – this becomes abundantly clear when attempting to reconstruct an epidemic profile form sequences: if an epidemic reconstructed from sequences bears no similarity to actual incidence/prevalence data, it is probably not a good one.

## 5.3 Data collection in the field, experimental assays in the laboratory (and the likelihood of a quantitative model explaining them)

Given enough information about the workings of any system, a mathematical model can be built to describe these processes and represent the main outputs. The structure of a model is invariably an abstraction or approximation at best, but even in the unlikely event the model structure is correct for all practical purposes, the unknown parameter values still preclude outright simulation of the process to reproduce real data or predict future outcomes.

Parameters need to be estimated, either independently or by joint inference of the entire parameter set. As mentioned above some parameters are directly measurable, e.g. a population size can be estimated by going out and counting every individual, others need some sort of model (whether simple or complex), e.g. an exponential growth rate can be estimated by plotting the numbers (of say, bacteria) on a logarithmic scale – in this case it is assumed that the base of the exponent is known (for instance base 2 or $e$), and that growth is the only active process. Under more complex models, the assumptions are looser and more uncertain, but otherwise it is unlikely that anything can be inferred at all – it is important to accept that fact: there is no such thing as assumption-free approaches, any more than there is free lunch.

While a model can be coherent from a mathematical and/or biological point of view, for inference purposes it is desirable that the specification of the model parameters makes them identifiable, e.g. Watterson's $\theta = 4N_e\mu$ is estimated because it is not possible to separately estimate the effective population size ($N_e$) and mutation rate ($\mu$), but only their product (Charlesworth & Charlesworth 2010). Algebraic methods exist to assess identifiability of parameters in models such as the SIR model, but slightly more complicated models may require more general methods (Eisenberg et al. 2013), and the approaches may become unfeasible as large scale models are formulated (Eisenberg 2013). Often, as is the case in this work, an empirical approach where simulating simple data and making test inferences is used, although might not be as elegant, it can still be very useful to assess inferability (Toni et al. 2009).

Under a likelihood-based inference framework, the available data will determine the likelihood function; however, except for the simplest cases (two- or three-dimensional functions) it is impossible to visualize these surfaces, and its complexity normally makes it impossible to assess beforehand if an inference algorithm will be able to find the optimal values, and whether more data is needed. It is possible to test estimation methods against a known parameter set by simulating a data structure similar to that collected or obtained in a lab, that is, simulating the actual experiment: model, sampling, stochasticity, expected losses, etc; however, the parameters of the experiment are unknown, and one (or many) simulated data sets can be indication, but not proof success with real data.

In the case of SIR-like models, once these basic structure or variations are accepted (or under a simulation study scenario), some unknown parameter values may be restricted by independent estimates or ballpark figures, and in principle the job is facilitated. With exception of the notoriously difficult to estimate transmission (compound of contact and susceptibility) rate (Anderson & May 1981) (and in the case of multiple pathogen strains the cross-protection factor or time) most parameters have either been estimated, or restricted to a narrow range, so the most important issue becomes whether the structure of the data available contains enough information in the likelihood function about the unknown parameters. In the case of a dengue incidence time series, there is basically an oscillating pattern, and it is not difficult to imagine that many different models could reproduce that general pattern – although it is also not possible to rule out that there are unique features that favor a specific model.

In the case of viral sequences, there is the advantage of separating dengue serotypes from the start. The coalescent process incorporates birth rates, population size, as well as structure (Volz 2012), which in terms of vector transmission models (as formulated in chapter 4) represent multiple time series from different hosts. Conceptually this amounts to having specific data about both human (primary and secondary infections) and mosquito incidence and prevalences, for both serotyopes, i.e. having 14 different time series instead of a single incidence one – nevertheless, these are informative only indirectly through polymorphisms in viral genomes. Given that sequence based inference eliminates some of the biases observed in estimates from time series, it would be important to assess whether including multiple explicit time series (per-serotype, mosquito prevalence, etc) for inference corrects the more pronounced biases.

Often the discussion is about producing or gathering more or "better" data, but rarely does it ever come up whether these idealized data sets are informative enough through the lens of the likelihood function. It could be argued that despite the increased quality and abundance of data (Pybus et al. 2013), and interesting new perspectives (Eames et al 2015), basic steps remain challenging (Frost et al. 2015), and seemingly simple tasks like simulating some data and recovering the right parameters can have disappointing results. I would argue that this may stem from placing most efforts on either the side of data gathering or method development, but neglecting the more tedious intermediate steps like the effects of data sampling, of interactions among recombination and structure (and other details). It is understandable that that kind of work can be seen as a minor improvement and less rewarding, but unless there are additional incentives to do so, maybe the community will be at risk of having very fancy methods that never work in practice.

Stochasticity is by definition at the center of inference endeavors, and its presence alone may spell the difference between very accurate or very poor results (as shown for the two-serotype model). Using continuous approximations is guaranteed to fail somehow, although it is important to know how bad and in which way they do fail. Inference methods that adequately accommodate stochasticity exist (Ionides et al. 2006), but they may require additional tinkering to work with processes like the coalescent and a tree parameter (Rasmussen et al. 2014a; Smith et al. 2016). The computational costs can also be prohibitive depending on the specific implementation – which happens to be the case for the within-host model in this thesis – nevertheless incorporating stochasticity is likely to have greater effects than an ensemble of system-specific deterministic processes.

Abstaining from using models altogether is a solution to all these problems only insofar as scientists want to give all hope of making progress; therefore, a concerted effort of all these fronts, including the tedious bits, is essential. The issues are general when trying to estimate anything with any kind of data, and the work presented in this thesis illustrates broader concerns for furthering model-based approaches in biology and other complex systems.

## 5.4 Final remarks

Many of the issues pointed out have existing and well researched solutions, although they may require intense experimental design and performance, extensive data collection, and systematic implementation of sophisticated analyses frameworks (and potentially the computational demand associated to it); presumably each of those has been separately achieved in a safe, limited environment, but going all out on the biggest problems may never happen if there are no real incentives to develop the tedious parts

A very good example are models of dengue transmission themselves, alternative formulations can be grouped by the dozen of publications in reputable journals, as listed by a single review from Johansson et al. (2011) but possibly getting up to the hundreds in total – so there is not shortage of models to be tested as hypotheses. There certainly is enough data, or the potential to collect extensive and detailed data sets (Mondini et al. 2009; Rasmussen et al. 2014), and sophisticated methods exist that, though computationally intensive, could be implemented without requiring supercomputer infrastructure to run. And nevertheless full-model estimation in published papers can be counted on the fingers of one hand, twice probably.

Trying to push the frontier of research further in the most meaningful way can therefore be a thankless task, consuming hours, days, months, years, or the time of an entire PhD. For an individual there is no guarantee that the effort will be deemed worthy of the time invested in that particular endeavor, as opposed to some more "relevant" or "sexy" research – that is especially true if the progress is measured by century-old metrics (Lozano et al.

2012; Saha & Christakis 2003). Partial success is nearly tantamount to failure: no bravery medal; no honorable mention. That is a great way of getting an entire community stuck with mediocre practices.

Therefore, not all of the goals I set at the beginning of the research work towards Ph.D. were accomplished, for a number of reasons, and not always was I capable of identifying the best approach to solve a problem, or guarantee all the tools were used to make sure all bases were covered. However, I have always tried to try learn things I did not know before, and try do create new things that I believed no one knew before. I believe the greatest (and there are many, and big) limitations of the work summarized in this thesis stem not from choosing a difficult, ungrateful, and apparently less accomplishing path, but not being able to push the boundaries even farther back by trying even more unconventional ideas (whether it was a question of time limitation, exhaustion, frustration, or the many obstacles in the way of anyone taking on themselves to do such a task).

If it is to be anything, I hope this thesis expresses that desire to try harder and different, not for oneself but really to burst through the frontiers instead of chipping away at it.

# References

1. Anderson, R. M. & May, R. M. 1981 The Population Dynamics of Microparasites and Their Invertebrate Hosts. Philosophical Transactions of the Royal Society of London B: Biological Sciences 291, 451–524. (doi:10.1098/rstb.1981.0005)

2. Charlesworth, B. 2010. Elements of evolutionary genetics. Roberts Publishers.

3. Day, T., Alizon, S. & Mideo, N. 2011 Bridging scales in the evolution of infectious disease life histories: theory. Evolution 65, 3448–3461. (doi:10.1111/j.1558-5646.2011.01394.x)

4. Eames, K., Bansal, S., Frost, S. & Riley, S. 2015 Six challenges in measuring contact networks for use in modelling. Epidemics 10, 72–77. (doi:10.1016/j.epidem.2014.08.006)

5. Eisenberg, M. 2013 Generalizing the differential algebra approach to input-output equations in structural identifiability. arXiv:1302.5484v1 [q-bio.QM].

6. Eisenberg, M. C., Robertson, S. L. & Tien, J. H. 2013 Identifiability and estimation of multiple transmission pathways in cholera and waterborne disease. Journal of Theoretical Biology 324, 84–102. (doi:10.1016/j.jtbi.2012.12.021)

7. Frost, S. D. W., Pybus, O. G., Gog, J. R., Viboud, C., Bonhoeffer, S. & Bedford, T. 2015 Eight challenges in phylodynamic inference. Epidemics 10, 88–92. (doi:10.1016/j.epidem.2014.09.001)

8. Gomes, M. G. M., Lipsitch, M., Wargo, A. R., Kurath, G., Rebelo, C., Medley, G. F. & Coutinho, A. 2014 A Missing Dimension in Measures of Vaccination Impacts. PLoS Pathog. 10, e1003849. (doi:10.1371/journal.ppat.1003849)

9. Ionides, E. L., Bretó, C. & King, A. A. 2006 Inference for nonlinear dynamical systems. Proc Natl Acad Sci U S A 103, 18438–18443. (doi:10.1073/pnas.0603181103)

10. Johansson, M. A., Hombach, J. & Cummings, D. A. T. 2011 Models of the impact of dengue vaccines: A review of current research and potential approaches. Vaccine 29, 5860–5868. (doi:10.1016/j.vaccine.2011.06.042)

11. Keeling, M. J. & Rohani, P. 2008 Modeling infectious diseases in humans and animals.

12. Lozano, G. A., Larivière, V. & Gingras, Y. 2012 The weakening relationship between the impact factor and papers' citations in the digital age. Journal of the American Society for Information Science and Technology 63, 2140–2145. (doi:10.1002/asi.22731)

13. Mideo, N., Nelson, W. A., Reece, S. E., Bell, A. S., Read, A. F. & Day, T. 2011 Bridging scales in the evolution of infectious disease life histories: application. Evolution 65, 3298–3310. (doi:10.1111/j.1558-5646.2011.01382.x)

14. Mondini, A. et al. 2009 Spatio-temporal tracking and phylodynamics of an urban dengue 3 outbreak in São Paulo, Brazil. PLoS neglected tropical diseases 3, e448. (doi:10.1371/journal.pntd.0000448)

15. Pessoa, D., Souto-Maior, C., Gjini, E., Lopes, J. S., Ceña, B., Codeço, C. T. & Gomes, M. G. M. 2014 Unveiling Time in Dose-Response Models to Infer Host Susceptibility to Pathogens. PLoS Comput Biol 10, e1003773–9. (doi:10.1371/journal.pcbi.1003773)

16. Pybus, O. G., Fraser, C. & Rambaut, A. 2013 Evolutionary epidemiology: preparing for an age of genomic plenty. Philosophical transactions of the Royal Society of London. Series B, Biological sciences 368, 20120193–20120193. (doi:10.1098/rstb.2012.0193)

17. Rasmussen, D. A., Boni, M. F. & Koelle, K. 2014 Reconciling phylodynamics with epidemiology: the case of dengue virus in southern Vietnam. Mol Biol Evol 31, 258–271. (doi:10.1093/molbev/mst203)

18. Rasmussen, D. A., Volz, E. M. & Koelle, K. 2014 Phylodynamic Inference for Structured Epidemiological Models. PLoS Comput Biol 10, e1003570. (doi:10.1371/journal.pcbi.1003570)

19. Saha, S., Saint, S. & Christakis, D. A. 2003 Impact factor: a valid measure of journal quality? J Med Libr Assoc 91, 42–46.

20. Smith, R. A., Ionides, E. L. & King, A. A. 2016 Infectious Disease Dynamics Inferred from Genetic Data via Sequential Monte Carlo. bioRxiv, 096396. (doi:10.1101/096396)

21. Souto-Maior, C. 2015 Host–Symbiont–Pathogen–Host Interactions: Wolbachia, Vector-Transmitted Human Pathogens, and the Importance of Quantitative Models of Multipartite Coevolution. In Reticulate Evolution, pp. 207–230. Cham: Springer International Publishing. (doi:10.1007/978-3-319-16345-1_8)

22. Toni, T., Welch, D., Strelkowa, N., Ipsen, A. & Stumpf, M. P. H. 2009 Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. Journal of The Royal Society Interface 6, 187–202. (doi:10.1098/rsif.2008.0172)

23. Volz, E. M. 2012 Complex Population Dynamics and the Coalescent Under Neutrality. Genetics 190, 187–201. (doi:10.1534/genetics.111.134627)

# D

# Published articles and book chapters

1. Pessoa, D., Souto-Maior, C., Gjini, E., Lopes, J. S., Ceña, B., Codeço, C. T. & Gomes, M. G. M. 2014 Unveiling Time in Dose-Response Models to Infer Host Susceptibility to Pathogens. PLoS Comput Biol 10, e1003773–9. (doi:10.1371/journal.pcbi.1003773)

2. Souto-Maior, C., Lopes, J. S., Gjini, E., Struchiner, C. J., Teixeira, L. & M Gomes, M. G. 2015 Heterogeneity in symbiotic effects facilitates Wolbachia establishment in insect populations. Theoretical Ecology 8, 53–65. (doi:10.1007/s12080-014-0235-7)

3. Souto-Maior, C. 2015 Host–Symbiont–Pathogen–Host Interactions: Wolbachia, Vector-Transmitted Human Pathogens, and the Importance of Quantitative Models of Multipartite Coevolution. In Reticulate Evolution, pp. 207–230. Cham: Springer International Publishing.
(doi:10.1007/978-3-319-16345-1_8)

4. Gomes, M. G. M., Gjini, E., Lopes, J. S., Souto-Maior, C. & Rebelo, C. 2016 A theoretical framework to identify invariant thresholds in infectious disease epidemiology. Journal of Theoretical Biology 395, 97–102.
(doi:10.1016/j.jtbi.2016.01.029)

ITQB-UNL | Av. da República, 2780-157 Oeiras, Portugal
Tel (+351) 214 469 100 | Fax (+351) 214 411 277

www.itqb.unl.pt