Bayesian Hierarchical Modelling for Inferring Genetic Interactions in Yeast

Jonathan Heydari

Thesis submitted for the degree of
Doctor of Philosophy

Institute for Cell & Molecular Biosciences
Newcastle University
Newcastle upon Tyne
United Kingdom

February, 2014

Abstract

Identifying genetic interactions for a given microorganism, such as yeast, is difficult. Quantitative Fitness Analysis (QFA) is a high-throughput experimental and computational methodology for quantifying the fitness of microbial cultures. QFA can be used to compare between fitness observations for different genotypes and thereby infer genetic interaction strengths. Current "naive" frequentist statistical approaches used in QFA do not model between-genotype variation or difference in genotype variation under different conditions. In this thesis, a Bayesian approach is introduced to evaluate hierarchical models that better reflect the structure or design of QFA experiments. First, a two-stage approach is presented: a hierarchical logistic model is fitted to microbial culture growth curves and then a hierarchical interaction model is fitted to fitness summaries inferred for each genotype. Next, a one-stage Bayesian approach is presented: a joint hierarchical model which simultaneously models fitness and genetic interaction, thereby avoiding passing information between models via a univariate fitness summary. The new hierarchical approaches are then compared using a dataset examining the effect of telomere defects on yeast. By better describing the experimental structure, new evidence is found for genes and complexes which interact with the telomere cap. Various extensions of these models, including models for data transformation, batch effects and intrinsically stochastic growth models are also considered.

Contents

List of Figures

List of Tables

Nomenclature

cdc13Δ A null allele (or gene deletion) of *CDC13*

cdc13-1 A point mutation of the wild type gene *CDC13*

CDC13 A wild type gene

H. sapiens Homo sapiens

orf Δ Open reading frame deletion

S. cerevisiae Saccharomyces cerevisiae

ACF   Auto-correlation function

BUGS Bayesian inference Using Gibbs Sampling

Cdc13 A wild type protein

DAVID Database for Annotation, Visualization and Integrated Discovery

DDR   Deoxyribonucleic acid damage response

DNA   Deoxyribonucleic acid

DSB   Double-strand break

dsDNA Double-stranded deoxyribonucleic acid

E-MAP Epistatic Miniarray Profiling

FN   False negative

FP   False positive

GO   Gene ontology

IHM   Interaction hierarchical model

IOD   Integrated optical density

JAGS   Just Another Gibbs Sampler

JHM   Joint hierarchical model

LNA   Linear noise approximation

LNAA   Linear noise approximation of the stochastic logistic growth model with additive intrinsic noise

LNAM   Linear noise approximation of the stochastic logistic growth model with multiplicative intrinsic noise

MCMC   Markov chain Monte Carlo

MDP   Maximum doubling potential

MDR   Maximum doubling rate

ODE   Ordinary differential equation

ORF   Open reading frame

QFA   Quantitative Fitness Analysis

RNA   Ribonucleic acid

RRTR   Román-Román & Torres-Ruiz (2012) logistic growth diffusion process

SDE   Stochastic differential equation

SGA   Synthetic Genetic Array

SGD   Saccharomyces Genome Database

SHM   Separate hierarchical model

SLGM   Stochastic logistic growth model

SLGM+L   Stochastic logistic growth model with Log-normal measurement error

SLGM+N   Stochastic logistic growth model with Normal measurement error

ssDNA   Single-stranded deoxyribonucleic acid

Chapter 1. Introduction

High-throughput screening of microbial culture fitnesses is a powerful tool in biology that can be used to learn about the interaction between genes and proteins in living cells. Fitness, the ability of organisms to survive and reproduce in a specific environment, is of fundamental importance to every living organism. Measuring components of fitness (such as population growth rate) in microbial cultures is a way to directly assess and rank the health of such populations. Genome-wide Quantitative Fitness Analysis (QFA) is a robot-assisted high-throughput laboratory workflow, combining systematic genetic techniques to generate arrays of genetically distinct microbial cultures with quantification and modelling of growth curves to estimate fitnesses (Banks *et al.*, 2012; Addinall *et al.*, 2011). An important reason for carrying out QFA is to compare the fitnesses of cultures with distinct genotypes in order to quantify epistasis (genetic interaction).

In Addinall *et al.* (2011), a frequentist statistical approach is used to model and make inference for significantly interacting genes in a QFA screen comparison. Other large-scale quantitative genetic interaction screening approaches exist, such as Epistatic Miniarray Profiling (E-MAP) (Schuldiner *et al.*, 2006) and Synthetic Genetic Array (SGA) (Tong & Boone, 2006), but we expect QFA to provide higher quality fitness estimates by using a culture inoculation technique which results in a wider range of cell densities during culture growth and by capturing complete growth curves instead of using single time point assays. QFA and alternative genetic interaction screening approaches mentioned above use frequentist statistical methods that cannot account for all sources of experimental variation or estimate evidence of genetic interaction simultaneously and do not partition variation into population, genotype and repeat levels. Further, the frequentist statistical approaches used in the methods above cannot account for relevant prior information.

The first aim of this thesis is to develop new Bayesian models that will better determine genes which significantly interact than the current frequentist approach. Accounting for more sources of variation than the frequentist approach, Bayesian QFA will be able to find genetic interactions within QFA with less error and increased confidence. The new Bayesian QFA will be used to help locate genes that are related to telomere activity in suppressor/enhancer analysis as well as other high throughput experiments such as drug screening.

Analysis of high throughput genetic screen data involves modelling both the experi-

mental structure and its sources of variation. Many underlying sources of variation within the data can be identified in the experimental design. Without fully modelling variation within the experiment, a model may not be able to identify the more subtle interactions. With a Bayesian approach (Bernardo & Smith, 2007) there is more flexibility of model choice, allowing model structure to reflect experimental structure or design. Currently there is no standard frequentist approach which can deal with inference for a hierarchical model that simultaneously models logistic growth parameters and probability of genetic interaction. Using Bayesian hierarchical modelling (Gelman & Hill, 2006), this study looks to extract as much information as possible from valuable QFA data sets. The Bayesian hierarchical approach also allows the borrowing of strength across subjects, helping identify significantly interacting open reading frame deletions ($orf\Delta$s) which otherwise may have been given low significance and overlooked.

Prior distributions are used to incorporate the existing information known about the possible values for parameters. Bayesian analysis can allow the use of Boolean indicators to describe the evidence that each $orf\Delta$ interacts with the query mutation in terms of probability. During the model fitting procedure, we find that $orf\Delta$ fitnesses have a long-tailed distribution around their population mean due to unusually fit, dead or missing $orf\Delta$s. In these instances, the scaled t distribution is used to describe these features.

Following the approach for determining epistasis from the comparison of two QFA screens presented by Addinall *et al.* (2011), the present study develops a two-stage approach to this problem: $i)$ the separate hierarchical model (SHM) is fitted to cell density measurements to estimate fitness, then $ii)$ fitness estimates are input to the interaction hierarchical model (IHM). Next, a unified approach, referred to as the joint hierarchical model (JHM), is developed. The JHM models mutant strain fitnesses and genetic interactions simultaneously, without having to pass information between two different models. The JHM can also allow two important, distinct, microbial fitness phenotypes (population growth rate and carrying capacity) to provide evidence for genetic interaction simultaneously.

Applying the new Bayesian approaches to QFA screen data, the present study is able to identify new genes and complexes that interact with genetic mutation *cdc13-1* in yeast. *cdc13-1* is a genetic mutation which results in dysfunctional telomere maintenance. Telomeres are repetitive regions of deoxyribonucleic acid (DNA) at the end of linear chromosomes. They have been of great interest in recent years as they have been shown to have a role in ageing and cancer (Shay & Wright, 2005).

Current approaches (Addinall *et al.*, 2011) fit a deterministic logistic growth model to yeast QFA data. For logistic growth data sets where stochastic fluctuations are observed, the deterministic model fails to account for the intrinsic noise. To better describe observed yeast QFA data, a stochastic model can be used. Stochastic models simultaneously describe dynamics and noise or heterogeneity in real systems (Chen *et al.*, 2010). For example, stochastic models are increasingly recognised as necessary tools for understanding the behaviour of complex biological systems (Wilkinson, 2011, 2009) and are also used to capture uncertainty in financial market behaviour (Kijima, 2013; Koller, 2012). Many such models are written as continuous stochastic differential equations (SDEs) which often do not have analytical solutions and are slow to evaluate numerically compared to their deterministic counterparts. Simulation speed is often a particularly critical issue when inferring model parameter values by comparing simulated output with observed data (Hurn *et al.*, 2007).

For SDE models where no explicit expression for the transition density is available, it is possible to infer parameter values by simulating a latent process using a data augmentation approach (Golightly & Wilkinson, 2005). However, this method is computationally intensive and not practical for all applications. When fast inference for SDEs is important, for example real-time analysis as part of decision support systems or big data inference problems where simultaneous model fits are made to many thousands of datasets (e.g. Heydari *et al.* (2012)), an alternative approach is needed (Heydari *et al.*, 2013).

The second aim of this thesis is to present a fast approach for stochastic modelling of processes with intractable transition densities and apply this approach to a SDE describing logistic population growth for the first time. One such approach is demonstrated: developing an analytically tractable approximation to the original SDE, by making linear noise approximations (LNAs) (Kurtz, 1970, 1971; Van Kampen, 2011). The present study introduces two new first order LNAs of a stochastic logistic growth model (SLGM) (Capocelli & Ricciardi, 1974), one with multiplicative and one with additive intrinsic noise, which are labelled LNAM and LNAA respectively. The LNA reduces a SDE to a linear SDE with additive noise, which can be solved to give an explicit expression for the transition density.

The Bayesian approach can be applied in a natural way to carry out parameter inference for state space models with tractable transition densities (West & Harrison, 1997). A state space model describes the probabilistic dependence between an observation process variable X_t and state process S_t. The transition density is used to describe the state process S_t and a measurement error structure is chosen to describe the relationship be-

tween X_t and S_t. Transition densities are derived for the LNA approximate models and measurement noise is chosen to be either multiplicative or additive in order to construct a linear Gaussian structure and allow fast inference through the use of a Kalman filter. The Kalman filter (Kalman, 1960) is typically used to infer the hidden state process of interest S_t and is an optimal estimator, minimising the mean square error of estimated parameters. The main assumptions of the Kalman filter are that the underlying system is a linear dynamical system and that all noise is Gaussian (or that the mean and standard deviation of the noise is known). Here the Kalman filter is used to reduce computational time in a parameter inference algorithm by recursively computing the marginal likelihood (West & Harrison, 1997).

It is shown that both of the new diffusion equation models have more realistic growth characteristics at the saturation stage when compared to a related model by Román-Román & Torres-Ruiz (2012) (an approximate model approach which is labeled RRTR) and it is shown that a zero-order LNA of the logistic growth SDE with multiplicative intrinsic noise is equivalent to the RRTR.

This study compares the utility of each of the approximate models during parameter inference by comparing simulations with both synthetic and real datasets. After inference it is shown that the fast approximate methods give similar posterior distributions to the slow arbitrarily exact models. Of the approximate models considered, the RRTR model is shown to be the worst at recovering true parameters of logistic growth data.

The LNA models are an improvement over the RRTR and so should be used for better parameter inference of logistic growth data, as they are just as fast but more accurate. The stochastic modelling approach presented in this study, a LNA followed by a Kalman filter recursion for marginal likelihood computation, is applicable to a range of population growth models or stochastic processes, where fast inference is of importance. The approach presented in this study enables stochastic modelling for a big data genome-wide analysis, where previously a deterministic model, unable to capture the information within the stochasticity of a process, is assumed due to the constraints in computational time associated with large volumes of data. The problems of big data (Boyd & Crawford, 2011) are relatively new and part of an expanding field of research that involves large and complex collections of data sets, typically with large components of noise.

1.1. Quantitative Fitness Analysis

Genome-wide Quantitative Fitness Analysis (QFA) is a robot-assisted high-throughput laboratory workflow, combining systematic genetic techniques to generate arrays of ge-

netically distinct microbial cultures with quantification and modelling of growth curves to estimate fitnesses (Banks *et al.*, 2012; Addinall *et al.*, 2011). A QFA screen can be used to compare the fitnesses of cultures with distinct genotypes in order to quantify genetic interaction.

Genetic interaction strengths are typically estimated by comparing fitnesses in two QFA screens: a control screen and a query screen. QFA output includes fitness estimates for all microbial cultures in an arrayed library including replicate cultures. For example, such a library could be a systematic collection of all non-essential, single gene deletion strains in the model eukaryote *Saccharomyces cerevisiae* (*S. cerevisiae*, brewer's yeast). All strains within a query screen differ from their control screen counterparts by a common condition such as a background gene mutation, drug treatment, temperature or other treatment. To identify strains that show interaction with the query condition, corresponding fitness responses for each strain in the library under the query and control conditions can be compared.

An example of the procedure to create mutant strains to test for genetic interaction using QFA screens is as follows. First a suitable query mutation is chosen, which is relevant to an area of biology of particular interest (e.g. *cdc13-1* for its relevance to telomere capping processes). Next, a library of strains is chosen, within which to search for strains interacting with the query mutation (e.g. a genome-wide library of independent strains with individual, non-essential genes deleted: *orf*Δs). Finally, an appropriate, neutral control background mutation is chosen (e.g. *ura3*Δ) to allow the separation of the effect of background condition from that of the library strains. In most cases, control and query mutations are crossed with the chosen library using Synthetic Genetic Array (SGA) technology (Tong & Boone, 2006). Independent replicate cultures are inoculated and grown across several plates for each strain under each condition to capture biological and technical heterogeneity. Cultures are grown simultaneously and time course images captured by photography. Robotic assistance is required for both culture inoculation and image capture during genome-wide screens which can include approximately 5,000 independent genotypes.

Raw QFA data (photographs) are converted into cell density estimates using the image analysis software Colonyzer (Lawless *et al.*, 2010). Observed changes in cell density over time are converted to fitness estimates for both the control and query strain by fitting logistic growth curves to data. Genetic interactions are identified by finding mutants in the query screen whose fitnesses deviate significantly from predictions given by a theoretical model of genetic independence.

Addinall *et al.* (2011) describe using QFA to infer genetic interactions with telomere-specific query mutations. They use least squares methods to fit logistic growth curves to culture time courses, then generate a univariate fitness estimate for each time course. They use a linear model predicting query strain fitness given control strain fitness, consistent with Fisher's multiplicative model of genetic independence, to test for genetic interaction between the query mutation and each $orf\Delta$. Deviation from the predicted linear relationship between the query and control fitnesses is evidence for genetic interaction between $orf\Delta$ and the query mutation. The significance of observed interactions is assigned using a simple frequentist linear modelling approach. One of the major limitations of the statistical model used in Addinall *et al.* (2011) is that it assumes each $orf\Delta$ fitness has the same variance. It is expected that explicit modelling of heterogeneity will allow more robust identification of interactions, particularly where variability for a particular strain is unusually high (e.g. due to experimental or technical difficulties).

1.1.1. Quantifying fitness

Observing changes in cell number in a microbial culture is the most direct way to estimate culture growth rate, an important component of microbial culture fitness. Direct counting of cell number on a high-throughput scale is not practical and so cell density estimates are made instead from culture photographs taken during QFA. Estimates of the integrated optical density (IOD) generated by the image analysis tool Colonyzer (Lawless *et al.*, 2010) are used to capture cell density dynamics in independent cultures during QFA. Density estimates, scaled to normalise for camera resolution, are gathered for each culture and a dynamic model of population growth, the logistic model $\dot{x} = rx(1-x/K)$ (Verhulst, 1845) (see Section 1.1.2), is fit to the data. Example photographic images of two yeast colonies inoculated by QFA, growing over time, along with corresponding quantitative measures of growth can be seen in Figure 1.3.

For a QFA screen, cultures are typically grown on 384-spot plates over time, where a process called *spotting* is used to inoculate microbial cultures on the plates. The spotting process involves a stage where microbial cultures are first diluted and then the diluted culture is spotted to the plate. Section 2.1.5 describes the spotting process and alternatives in further detail. An example 384-spot plate of yeast cultures is given in Figure 1.1. Yeast cultures in Figure 1.1 are all alive and have similar culture size. A cropped image of 15 yeast cultures from a 384-spot plate is given in Figure 1.2. Yeast cultures in Figure 1.2 have different culture sizes, the smaller cultures have had slow growth relative to the larger cultures. An example of the raw time series data is given in the Appendix, Figure A.1.
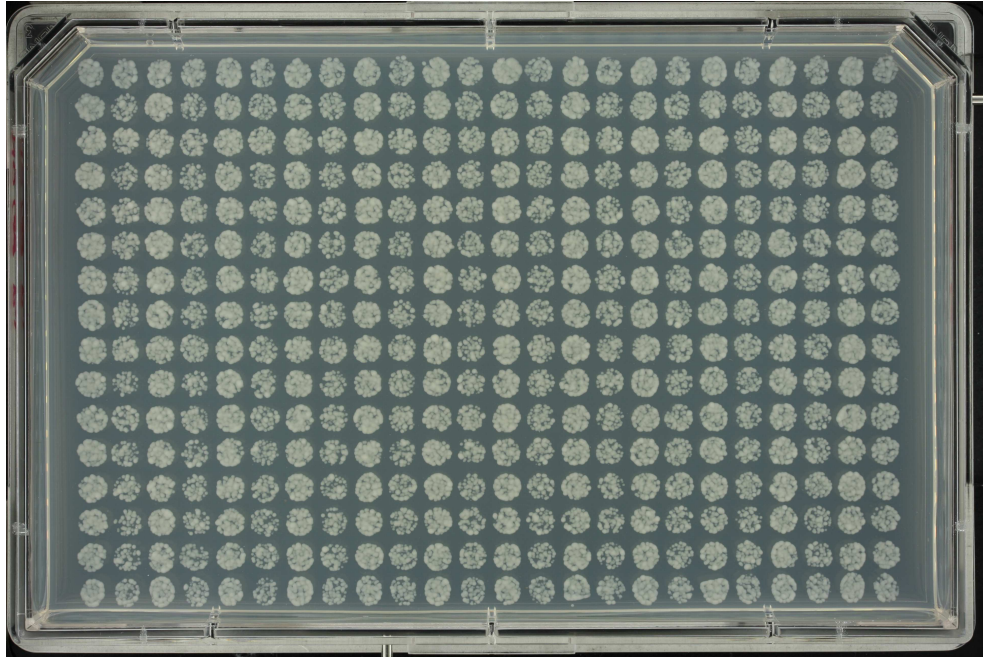
Figure 1.1: Example 384-spot plate image from a yeast quantitative fitness analysis screen, taken approximately 3 days after inoculation. Yeast cultures are spotted and grown in regular arrays on solid agar plates.

Figure 1.2: Cropped image of 15 out of 384 spotted yeast cultures from a 384-spot plate, taken from a quantitative fitness analysis screen. Image taken approximately 3 days after inoculation. Yeast cultures are spotted and grown in regular arrays on solid agar plates.

Further detail on the QFA workflow and alternative 384-spot plate images can be found at (Banks *et al.*, 2012) and `http://research.ncl.ac.uk/qfa/`.

After logistic growth model fitting, estimated logistic growth parameters sets can then be used to determine the fitness of a culture. If required, a univariate fitness definition can be chosen to summarise a set of logistic growth parameters (see Section 1.1.3).
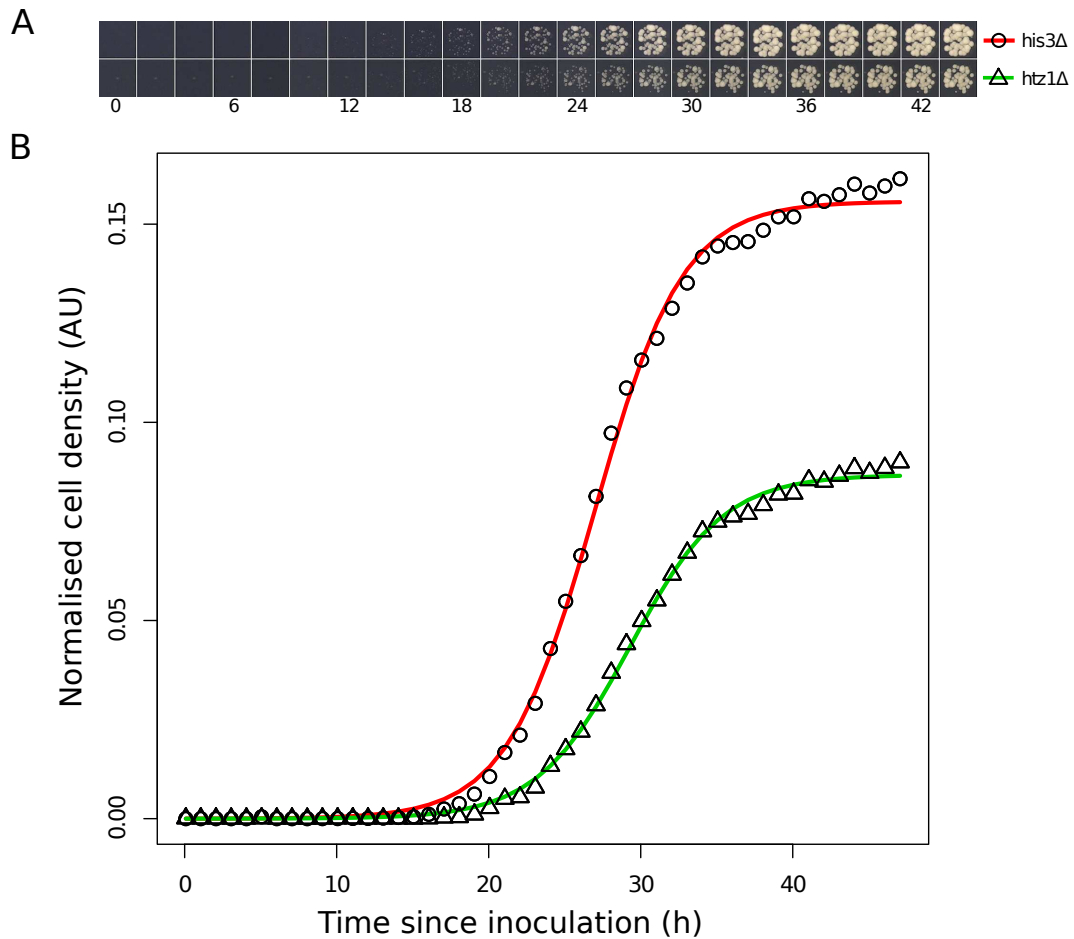
Figure 1.3: A) Timelapse images for two genetically modified *S. cerevisiae* cultures with different genotypes (indicated) corresponding to the time series measurements plotted in panel B. B) Time course cell density estimates derived from analysis of the timelapse images in panel A together with (least squares) fitted logistic growth curves.

1.1.2. The logistic growth model

The logistic model of population growth, an ordinary differential equation (ODE) describing the self-limiting growth of a population of size $x(t)$ at time t, was developed by

Verhulst (1845),

$$\frac{dx(t)}{dt} = rx(t)\left(1 - \frac{x(t)}{K}\right). \tag{1.1}$$

The ODE has the following analytic solution:

$$x(t; \theta) = \frac{KPe^{rt}}{K + P\left(e^{rt} - 1\right)}, \tag{1.2}$$

where $P = x(0)$ and $\theta = (K, r, P)$. The model describes a population growing from an initial size P (culture inoculum density) with an intrinsic growth rate r, undergoing approximately exponential growth which slows as the availability of some critical resource (e.g. nutrients or space) becomes limiting (Jr. *et al.*, 1976). Ultimately, population density saturates at the carrying capacity (maximum achievable population density) K, once the critical resource is exhausted. Appendix A.2 shows how to derive the solution of (1.1), given in (1.2). An example of two different logistic growth trajectories are given by the solid lines in Figure 1.3B. Where further flexibility is required, generalized forms of the logistic growth process (Tsoularis & Wallace, 2002; Peleg *et al.*, 2007) may be used instead (see Section 2.5.2).

1.1.3. Fitness definitions

Culture fitness is an important phenotype, indicating the health of a culture. Several distinct quantitative fitness measures based on fitted logistic model parameters (1.2) can be constructed. Addinall *et al.* (2011) present three univariate measures suitable for QFA: Maximum Doubling Rate (MDR) and Maximum Doubling Potential (MDP) detailed in (1.3), and their product $MDR \times MDP$, where

$$MDR = \frac{r}{log\left(2\frac{K-P}{K-2P}\right)} \ \ \text{and} \ \ MDP = \frac{log\left(\frac{K}{P}\right)}{log(2)}. \tag{1.3}$$

MDR is reciprocal of minimum doubling time T which a cell population takes to reach $2x(0)$, assuming the exponential phase begins at $t = 0$:

$$\frac{x(t)}{x(0)} = 2.$$

We now rearrange to give the following expression for MDR:

$$MDR = \frac{1}{T} = \frac{r}{\log(\frac{2(K-P)}{K-2P})}.$$

MDP is the number of times population size doubles before reaching saturation, assuming geometric progression:

$$x(0) \times 2^{MDP} = K.$$

Rearrange to give the following:

$$MDP = \frac{\log(\frac{K}{P})}{\log 2}.$$

MDR captures the rate at which microbes divide when experiencing minimal intercellular competition or nutrient stress. A strain's growth rate largely dictates its ability to outcompete any neighbouring strains. MDP captures the number of divisions the culture is observed to undergo before saturation. A strain which can divide a few more times than its neighbours in a specific environment also has a competitive advantage.

The choice of a single overall fitness score depends on the aspects of microbial physiology most relevant to the biological question at hand. Typically the fitness definition $MDR \times MDP$ is used in QFA to account for both attributes simultaneously. Other fitness definitions available include cell count, expected generation number and their approximations (Cole *et al.*, 2007).

1.2. Epistasis

Epistasis is the phenomenon where the effects of one gene are modified by those of one or several other genes (Phillips, 1998). Besides the multiplicative model, there are other definitions for epistasis such as additive, minimum and log (Mani *et al.*, 2008). Minimum is a suboptimal approach which may allow "masking" of interactions (Mani *et al.*, 2008). For a typical yeast QFA screen comparison, Addinall *et al.* (2011) assumes a multiplicative interaction model (1.4), but when dealing with measurements on a log scale, it is effectively assuming an additive interaction model (Aylor & Zeng, 2008). This highlights the point that multiplicative and additive models are equivalent if fitness data are scaled appropriately (Cordell, 2002).

1.2.1. Defining epistasis

As presented in Addinall *et al.* (2011), this study assumes Fisher's multiplicative model of genetic independence (1.4) (Cordell, 2002; Phenix *et al.*, 2011), to represent the expected relationship between control strain fitness phenotypes and those of equivalent query strains in the absence of genetic interaction. In this study, we interpret genotypes for which the query strain fitness deviates significantly from this model of genetic independence as interacting significantly with the query mutation. Square bracket notation is used to represent a quantitative fitness measure. For example $[wt]$ and $[query]$ represent wild-type and query mutation fitnesses respectively. "Wild-type" strictly refers to the genotype that is prevalent among individuals in a natural (or wild) population. However, during laboratory cultivation of microbes it is more usual to introduce extra gene mutations to an ancestral lineage that is well established within the scientific community. Working with established lineages allows direct comparison with results from the literature without the confounding effect of sampling genotypes from natural populations, which are considerably more heterogeneous. Thus in context of this thesis, "wild-type" will refer to the reference strain, before additional mutations are introduced. $orf\Delta$ represents an arbitrary single gene deletion strain (i.e. a mutant from the control strain library). $query : orf\Delta$ represents an arbitrary single gene deletion from the query strain library (e.g. crossed with the query mutation). Fisher's multiplicative model of genetic independence is as follows:

$$[query : orf\Delta] \times [wt] = [query] \times [orf\Delta] \tag{1.4}$$

$$\Rightarrow [query : orf\Delta] = \frac{[query]}{[wt]} \times [orf\Delta]. \tag{1.5}$$

In (1.5), $\frac{[query]}{[wt]}$ is a constant for a given pair of QFA screens, meaning that if this model holds, there should be a linear dependence between $[query : orf\Delta]$ and $[orf\Delta]$ for all deletions $orf\Delta$. During genome-wide screens of thousands of independent $orf\Delta$s, it can be assumed that the majority of gene mutations in the library do not interact with the chosen query mutations. Therefore, even if the query or wild-type fitnesses are not available to us, the slope of this linear model can still be estimated by fitting it to all available fitness observations, before testing for strains which deviate significantly from the linear model. Any extra background condition, such as a gene mutation common to both the control and query strains (e.g. triple instead of double deletion strains for the query and control data sets), may change the interpretation or definition of the type of genetic interaction but the same linear relationship is applicable.

1.2.2. Addinall *et al.* (2011) Quantitative Fitness Analysis screen comparison

Addinall *et al.* (2011) present QFA where the logistic growth model (1.2) is fit to experimental data by least squares to give parameter estimates $(\hat{K}, \hat{r})$ for each culture time course (each *orfΔ* replicate). Inoculum density P is assumed known and the same across all *orfΔ*s and their repeats. After inoculating approximately 100 cells per culture, during the first several cell divisions there are so few cells that culture cell densities remain well below the detection threshold of cameras used for image capture and so, without sharing information across all *orfΔ* repeats, P cannot be estimated directly. It is therefore necessary to fix P to the same value for both screens, using an average estimate of P from preliminary least squares logistic growth model fits. Fitting the model to each *orfΔ* repeat separately means there is no sharing of information within an *orfΔ* or between *orfΔ*s when determining $\hat{K}$ and $\hat{r}$. By developing a hierarchical model to share information across *orfΔ* repeats for each *orfΔ* and between *orfΔ*s, estimates for every set of logistic growth curve parameters (K, r) can be improved and therefore for every strain fitness.

Quantitative fitness scores (F_{cm}) for each culture were defined (1.6) (see (1.3) for definitions of MDR and MDP), where

$$F_{cm} = MDR_{cm} \times MDP_{cm}. \tag{1.6}$$

The index c identifies the condition for a given *orfΔ*: $c = 0$ for the control strain and $c = 1$ for the query strain. m identifies an *orfΔ* replicate. Scaled fitness measures $\tilde{F}_{cm}$ are calculated for both the control and query screen such that the mean across all *orfΔ*s for a given screen is equal to 1. After scaling, any evidence that $\tilde{F}_{0m}$ and $\tilde{F}_{1m}$ are significantly different will be evidence of genetic interaction.

The following linear model was fit to the control and query strain scaled fitness measure pairs $\tilde{F}_{cm}$ for each unique *orfΔ* in the gene deletion library:

$$\begin{aligned}
\tilde{F}_{cm} &= \mu + \gamma_c + \varepsilon_{cm}, \ \text{where} \ \gamma_0 = 0 \\
\varepsilon_{cm} &\sim \mathrm{N}(0, \sigma^2), \ \text{where} \ \varepsilon_{cm} \ \text{is i.i.d.}
\end{aligned} \tag{1.7}$$

In (1.7), γ_1 represents the estimated strength of genetic interaction between the control and query strain. If the scaled fitnesses for the control and query strain are equivalent for a particular *orfΔ* such that they are both estimated by some μ, i.e. no evidence of genetic interaction, we would expect $\gamma_c = 0$. The model was fit by maximum likelihood, using the R function "lmList" (Pinheiro & Bates, 2000) with variation assumed to be the

same for all strains in a given screen and the same for both control and query screens. So, for every gene deletion from the library an estimate of γ_1 was generated together with a p-value for whether it was significantly different from zero. False discovery rate (FDR) corrected q-values were then calculated to determine levels of significance for each *orfΔ*. Addinall *et al.* (2011) use the Benjamini-Hochberg test (Benjamini & Hochberg, 1995) for FDR correction. This test is commonly used in genomic analyses as although it assumes independence of test statistics, even if positive correlation exists between tests, the result is that FDR estimates are slightly conservative. Finally a list of *orfΔ* names, ranked by γ magnitudes, was output and *orfΔ*s with q-values below a significance cut-off of 0.05 classed as showing significant levels of genetic interaction with the query mutation.

1.2.3. Fitness plots

Fitness plots are used to show which *orfΔ*s show evidence of genetic interaction from a QFA screen comparison. Figure 1.4 shows an example fitness plot taken from (Addinall *et al.*, 2011). Fitness plots are typically mean *orfΔ* fitnesses for control strains against the corresponding query strains. *orfΔ*s with significant evidence of interaction are highlighted in the plot as red and green for suppressors and enhancers respectively. *orfΔ*s without significant evidence of interaction are in grey. Solid and dashed grey lines are for a simple linear model fit (corresponding to a model of genetic independence) and the line of equal fitness respectively.
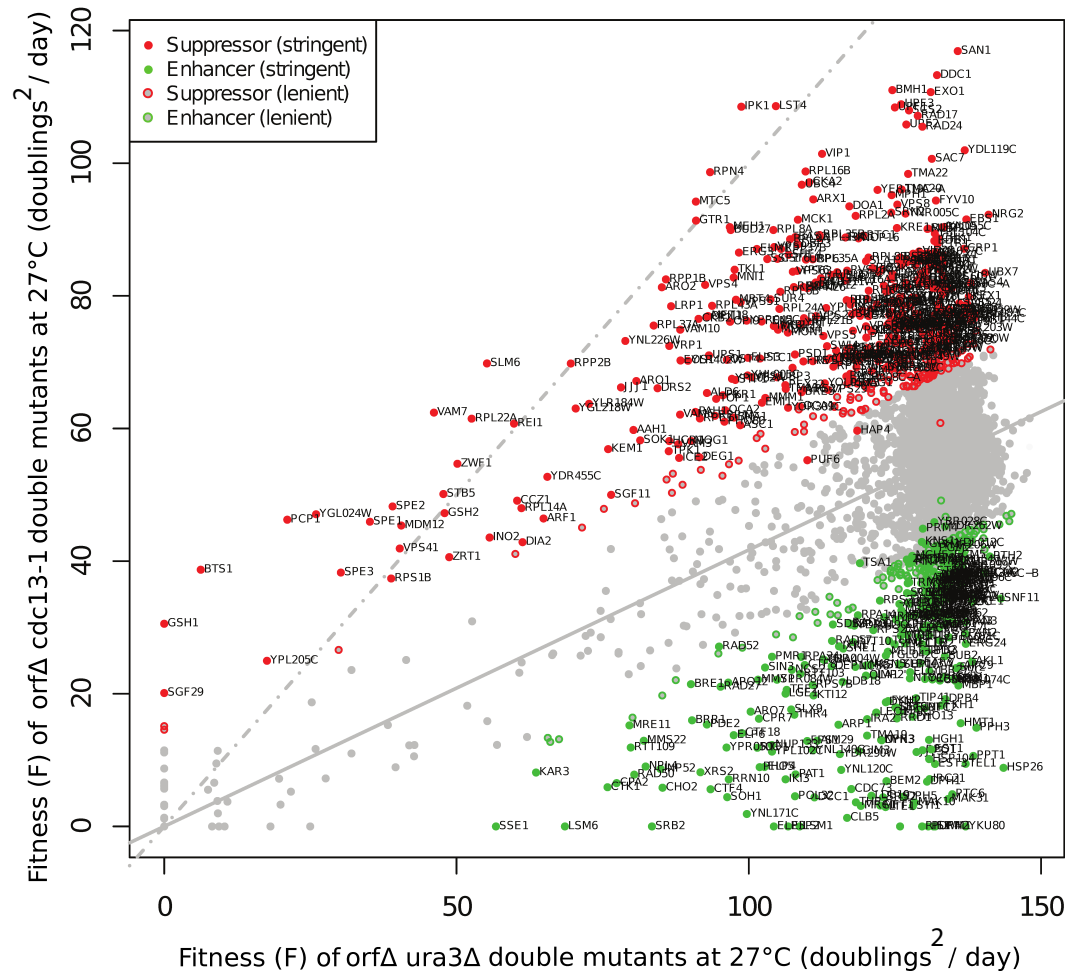
Figure 1.4: Fitness plot taken from Addinall *et al.* (2011). A yeast genome knock out collection was crossed to the *cdc13-1* mutation, or as a control to the *ura3Δ* mutation. 8 replicate crosses were performed for the query and control strains. *orfΔ*s with significant evidence of interaction are highlighted in red and green for suppressors and enhancers respectively. *orfΔ*s without significant evidence of interaction are in grey and have no *orf* name label. Lenient and stringent classification of significant interaction is based on p-values < 0.05 and FDR corrected p-values (q-values) < 0.05 respectively. For a further description on fitness plots, see Section 1.2.3.

1.3. The stochastic logistic growth model

To account for uncertainty about processes affecting population growth which are not explicitly described by the deterministic logistic model, we can include a term describing intrinsic noise and consider an SDE version of the model. Here we extend the ODE in (1.1) by adding a term representing multiplicative intrinsic noise (1.8) to give a model which we refer to as the stochastic logistic growth model (SLGM), which was first introduced by Capocelli & Ricciardi (1974),

$$dX_t = rX_t \left(1 - \frac{X_t}{K} \right) dt + \sigma X_t dW_t, \tag{1.8}$$

where $X_{t_0} = P$ and is independent of Wiener process W_t, $t \geq t_0$. The Wiener process (or standard Brownian motion) is a continuous-time stochastic process, see Section 2.6.1. The Kolmogorov forward equation has not been solved for (1.8) (or for any similar formulation of a logistic SDE) and so no explicit expression for the transition density is available. Román-Román & Torres-Ruiz (2012) introduce a diffusion process approximating the SLGM with a transition density that can be derived explicitly (see Section 5.2).

Alternative stochastic logistic growth models to (1.8) are available. Allen (2010) derives the stochastic logistic growth models given in (1.9) and (1.10) from Markov jump processes (Allen, 2010; Wilkinson, 2011). Firstly,

$$dX_t = rX_t \left(1 - \frac{X_t}{K} \right) dt + \sqrt{rX_t} dW_t, \tag{1.9}$$

where $X_{t_0} = P$ and is independent of W_t, $t \geq t_0$. Secondly,

$$dX_t = rX_t \left(1 - \frac{X_t}{K} \right) dt + \sqrt{rX_t \left(1 + \frac{X_t}{K} \right)} dW_t, \tag{1.10}$$

where $X_{t_0} = P$ and is independent of W_t, $t \geq t_0$.

Note that (1.8) (1.9) and (1.10) are not equivalent to each other. (1.9) and (1.10) are able to describe the discreteness of the Markov jump processes that they approximate (or demographic noise). Demographic noise becomes less significant for large population sizes, therefore (1.9) and (1.10) describe more deterministic growth curves when population size is large (i.e. large carrying capacity K). Equation 1.8 introduces an additional parameter σ, unlike (1.9) and (1.10). The additional parameter in (1.8) allows us to tune the amount of noise in the system that is not directly associated with the noise due to

the discreteness of the process (demographic noise). The additional parameter also gives (1.8) further flexibility for modelling intrinsic noise than (1.9) and (1.10). As the diffusion terms of (1.9) and (1.10) are functions of the logistic growth parameters, for large populations (1.9) and (1.10) can confound intrinsic noise with estimates of logistic growth parameters r and K. For the above reasons, the SLGM in (1.8) is the most appropriate model for estimating logistic growth parameters of large populations, as intrinsic noise does not tend to zero with larger population sizes, unlike (1.9) and (1.10).

1.4. Outline of thesis

A brief outline of thesis is as follows. Chapter 2 gives background to the biological and statistical methods used throughout the thesis. Yeast biology related to the QFA data sets analysed in this study is given as well as an introduction to Bayesian inference.

In Chapter 3 the SHM and IHM models for the new two-stage Bayesian QFA approach are presented. Next, the JHM for the new one-stage Bayesian QFA approach is presented. The chapter is concluded by introducing a two-stage frequentist QFA approach using a random effects model.

In Chapter 4 the new Bayesian approaches are applied to a previously analysed QFA data set for identifying genes interacting with a telomere defect in yeast. The chapter is concluded with an analysis of further QFA data sets with the JHM and two extensions of the JHM; included for further investigation and research.

Chapter 5 begins by introducing an existing logistic growth diffusion equation by Román-Román & Torres-Ruiz (2012). Two new diffusion equations for carrying out fast, Bayesian parameter estimation for stochastic logistic growth data are then presented. The chapter is concluded by comparing inference between the approximate models considered and with arbitrarily exact approaches.

Finally, Chapter 6 presents conclusions on the relative merits of the newly developed Bayesian approaches and stochastic logistic growth models. The chapter is concluded by discussing the broader implications of the results of the studies presented and scope for further research.

Chapter 2. Background

2.1. Yeast biology

Saccharomyces cerevisiae is a species of budding yeast widely used to study genetics. *S. cerevisiae* was the first eukaryotic genome that was completely sequenced (Goffeau *et al.*, 1996). Yeast is ideal for high throughout experimentations as it is easy to use and arrayed libraries of genetically modified yeast strains are readily available or obtainable for experiments (Zeyl, 2000). There are many different observable traits available with *S. Cerevisiae*, such as size, opacity and density. There are about 6000 genes in the *S. Cerevisiae* genome of which 5,800 of these are believed to be true functional genes (Cherry *et al.*, 2012).

Yeasts are ideal for genome-wide analysis of gene function as genetic modification of yeast cells is relatively straightforward and yeast cultures grow quickly. Epistasis identified within a species of yeast may exist in the analogous genes within the human genome (Botstein *et al.*, 1997). Therefore, finding genes involved in epistasis within yeast is of great interest outside the particular experimental species in question.

2.1.1. Telomeres

Telomeres are the ends of linear chromosomes and found in most eukaryotic organisms (Olovnikov, 1996). Telomeres permit cell division and some researchers claim that telomere-induced replicative senescence is an important component of human ageing (Lydall, 2003). They cap (or seal) the chromosome end to ensure genetic stability and are believed to prevent cancer (Shay & Wright, 2005).

In Figure 2.1, a *S. cerevisiae* chromosome is shown with the telomere single-stranded DNA (ssDNA) at the end, where DNA binding proteins such as Cdc13 are bound. Figure 2.1 also shows how telomere maintenance compares between a Homo sapiens (*H. sapiens*) and *S. cerevisiae* chromosome.
Telomere length decreases with each division of a cell until telomere length is very short and the cell enters senescence (Hayflick & Moorhead, 1961), losing the ability to divide. Some cancerous cells up-regulate the enzyme called telomerase which can prevent shortening of telomeres or elongate them, potentially allowing cancerous cells to live indefinitely (Wright & Shay, 1992).
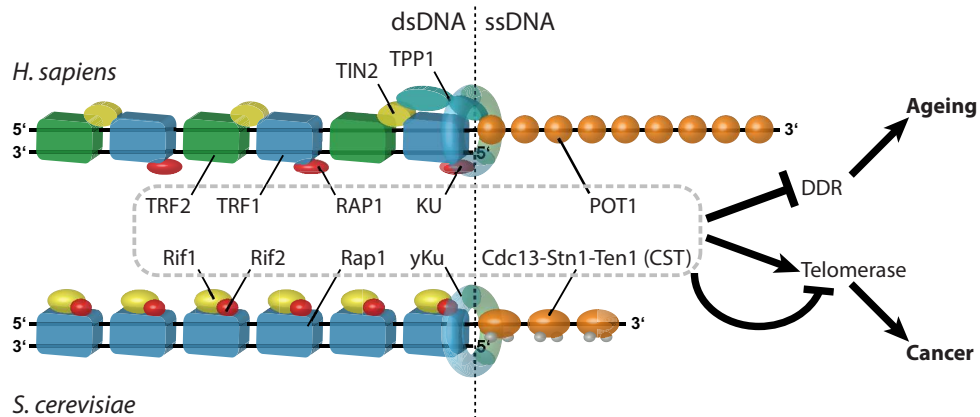
Figure 2.1: Telomere at a chromosome end (diagram and legend taken from Dewar & Lydall (2012)). The telomere cap is evolutionarily conserved. Telomeres are nucleoprotein caps present at the ends of most eukaryotic chromosomes, consisting of double-stranded DNA (dsDNA) with a single-stranded DNA (ssDNA) overhang, bound by dsDNA- and ssDNA-binding proteins. Collectively, the telomere binding proteins "cap" the telomere and serve to regulate telomerase activity and inhibit the DNA damage response (DDR). In budding yeast, the telomeric dsDNA is bound by Rap1, which recruits the accessory factors Rif1 and Rif2. In humans, the telomeric dsDNA is bound by TRF1 and TRF2 (held together by TIN2) and TRF2 recruits RAP1 to telomeres. In budding yeast, Cdc13 binds the telomeric ssDNA and recruits Stn1 and Ten1 to form the CST (Cdc13-Stn1-Ten1) complex, while in humans, the telomeric ssDNA is bound by POT1. In human beings, POT1 and TRF1-TRF2-TIN2 are linked together by TPP1, which may permit the adoption of higher-order structures. In both budding yeast and humans, the Ku complex, a DDR component that binds to both telomeres and Double-strand breaks (DSBs), also binds and plays a protective role.

It is believed that telomeres are partly responsible for ageing; without the enzyme telomerase, a fixed limit to the number of times the cell can divide is set by the telomere shortening mechanism because of the end replication problem (Levy *et al.*, 1992).

2.1.2. The end replication problem

In eukaryote cell replication, shown in Figure 2.2, new strands of DNA are in the $5'$ to $3'$ direction (red arrows), the leading strand is therefore completed in one section whereas the lagging strand must be formed via backstitching with smaller sections known as Okazaki fragments (Lydall, 2003). Figure 2.2 shows how the lagging strand is left with a $3'$ overhang, with the removal of the terminal primer at the end and how the leading strand is left with a blunt end (David Wynford-Thomas, 1997). Telomerase fixes this problem by extending the $3'$ end to maintain telomere length (Levy *et al.*, 1992). Without telomerase, the leading strand is shortened (Olovnikov, 1973) and telomere capping proteins such as Cdc-13 in yeast binds to the ssDNA that remains. Most eukaryotic cells have telomerase

activated and may maintain DNA replication indefinitely. Not all mammalian cells have telomerase activated and it is believed this problem then leads to the shortening of their telomeres and ultimately senescence.
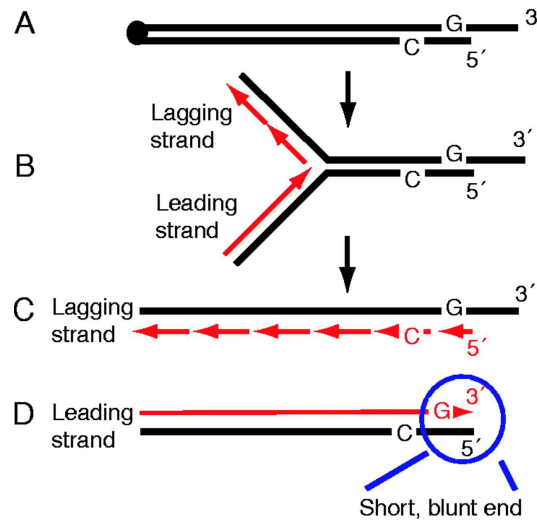
Figure 2.2: The end replication problem (diagram and legend taken from Lydall (2003)). (A) Telomeres in all organisms contain a short $3'$ overhang on the G rich strand. (B) A replication fork moving towards the end of the chromosome. (C) The newly replicated, lagging C strand, will generate a natural $3'$ overhang when the ribonucleic acid (RNA) primer is removed from the final Okazaki fragment, or if the lagging strand replication machinery cannot reach the end of the chromosome. In the absence of nuclease activity the unreplicated $3'$ strand will be the same length as it was prior to replication. (D) The newly replicated leading G strand will be the same length as the parental $5'$ C strand, and blunt ended if the replication fork reaches the end of the chromosome. Therefore the newly replicated $3'$ G strand will be shorter than the parental $3'$ strand and unable to act as a substrate for telomerase because it does not contain a $3'$ overhang. If the leading strand replication fork does not reach the end of the chromosome a $5'$ rather than $3'$ overhang would be generated, but this would not be a suitable substrate for telomerase.

2.1.3. *CDC13* and *cdc13-1*

CDC13 is an essential telomere-capping gene in *S. cerevisiae* (Zubko & Lydall, 2006). The protein Cdc13, encoded by *CDC13*, binds to telomeric DNA (see Figure 2.1), forming a nucleoprotein structure (Lustig, 2001). Cdc13 regulates telomere capping and is part of the CST complex with Stn1 and Ten1 (Wellinger, 2009). This provides protection from degradation by exonucleases such as Exo1. *cdc13-1* is a temperature-sensitive allele of the *CDC13* gene that has temperature sensitivity above $26\,°C$, where the capping ability of the protein is reduced (Nugent *et al.*, 1996). By inducing the temperature sensitivity of *Cdc13-1*, telomere maintenance is disrupted. A lot of research activity for telomere

integrity focuses on the CST complex and often *cdc13* mutations are considered, like *cdc13-1* and *cdc13-5* (see, for example, Anbalagan *et al.*, 2011; Foster *et al.*, 2006).

2.1.4. *URA3*

URA3 is a gene that encodes orotidine 5-phosphate decarboxylase (Cong *et al.*, 2002). *URA3* is used as a genetic marker for DNA transformations, allowing both positive and negative selection depending on the choice of media (Kaneko *et al.*, 2009).

In Addinall *et al.* (2011) *ura3Δ* is used as a control mutation because it is neutral under the experimental conditions. For a QFA comparison, constructing a query mutation such as *cdc13-1* typically involves adding selection markers to the genome. To ensure that the same selection markers are found in both the query and control strains, and that the control and query screens can be carried out in comparable environments, a neutral mutation such as *ura3Δ* can be introduced to the control strain. *URA3* encodes an enzyme called ODCase. Deleting *URA3* causes a loss of ODCase, which leads to a reduction in cell growth unless uracil is added to the media (Jones, 1992). Addinall *et al.* (2011) include uracil in their media so that *ura3Δ* is effectively a neutral deletion, approximating wild-type fitness. As a control deletion, *URA3* is not expected to interact with the query mutation, the library of *orf Δ*s in the control and query screen or any experimental condition of interest such as temperature.

2.1.5. High-throughput methodology for Quantitative Fitness Analysis

To collect enough data to perform QFA (Addinall *et al.*, 2011), a methodology such as high-throughput screening is required (Soon *et al.*, 2013; An & Tolliday, 2009). High-throughput screening is most notably used in the field of biology for genome wide suppressor/enhancer screening and drug discovery. The automation of experimental procedures through robotics, software, sensors and controls allows a researcher to carry out large scale experimentation quickly and more consistently.

Hundreds of microbial strains with various gene deletions need to be systematically created, cultured and then have measurable traits quantified. The repeatability of microbial culture growth is ideal to give sufficient sample sizes for identifying both variation and significance in high throughput experimentation (Xu, 2010).

The quality of the quantitative data is critical for identifying significantly interacting genes. To measure the phenotypes of different mutant strains of a micro-organism such as yeast (Zeyl, 2000), a process called *spotting* is used. This process is different to a typical SGA experiment where *pinning* would be used (see, for example, Tong & Boone (2006)).

Pinning is a quicker but less quantitative process where the microbial strains are typically directly pinned to a 1536 plate and allowed to grow until image analysis starts. Spotting on the other hand has a stage where the cultures are diluted and then the dilute culture is spotted in 384 format to give a more accurate reading in image analysis. This in turn gives rise to much more accurate time series data for modelling.

Figure 2.3 illustrates the spotting process. An image opacity measure is typically used as a proxy for the density of microbial colonies. Time lapse photographs are taken of the 384-spot plates after incubation, using high resolution digital cameras, to measure growth. A software package such as Colonyzer (Lawless *et al.*, 2010) can then be used to determine a quantitative measure of fitness from the photographs taken of the cultures grown on the plates. To ensure a consistent method to capture images of microbial colonies, all cameras should be of the same make and model.

2.2. Comparing lists of genes

Upon completing a QFA screen comparison, a list of genes ordered by genetic interaction strength can be obtained. Lists of ordered genes can be used to compare two different statistical approaches for a QFA screen comparison.

A comparison of two lists can be carried out through standard statistical similarity measures such as the Jaccard Index or Spearman's rank correlation coefficient. Observing only the subset of genes showing significant evidence of genetic interaction, two lists of genes can be compared using the Jaccard Index (Cheetham & Hazel, 1969), see Section 2.2.1. The Jaccard index does not account for the ordering of genes and is dependent on the number of interactions identified when the cut-off of genes showing significant evidence of interaction is chosen or influenced by the experimenter. Due to these undesirable properties of the Jaccard index, this method is not appropriate for an unbiased comparison of statistical methods. The Spearman's rank correlation coefficient (Kowalczyk *et al.*, 2004) is able to account for the ordering of genes and is able to account for the whole list of genes available, see Section 2.2.2.

Gene ontology (GO) term enrichment can be used to suggest which list of genetic interactions has the most biological relevance (Consortium, 2004). There are many other alternative approaches available for the comparison of two gene lists (Yang *et al.*, 2006; Lottaz *et al.*, 2006).

Using both Spearman's correlation coefficient and GO term enrichment analysis of gene lists allows for both an unbiased statistical and biological comparison of two lists of ordered genes.
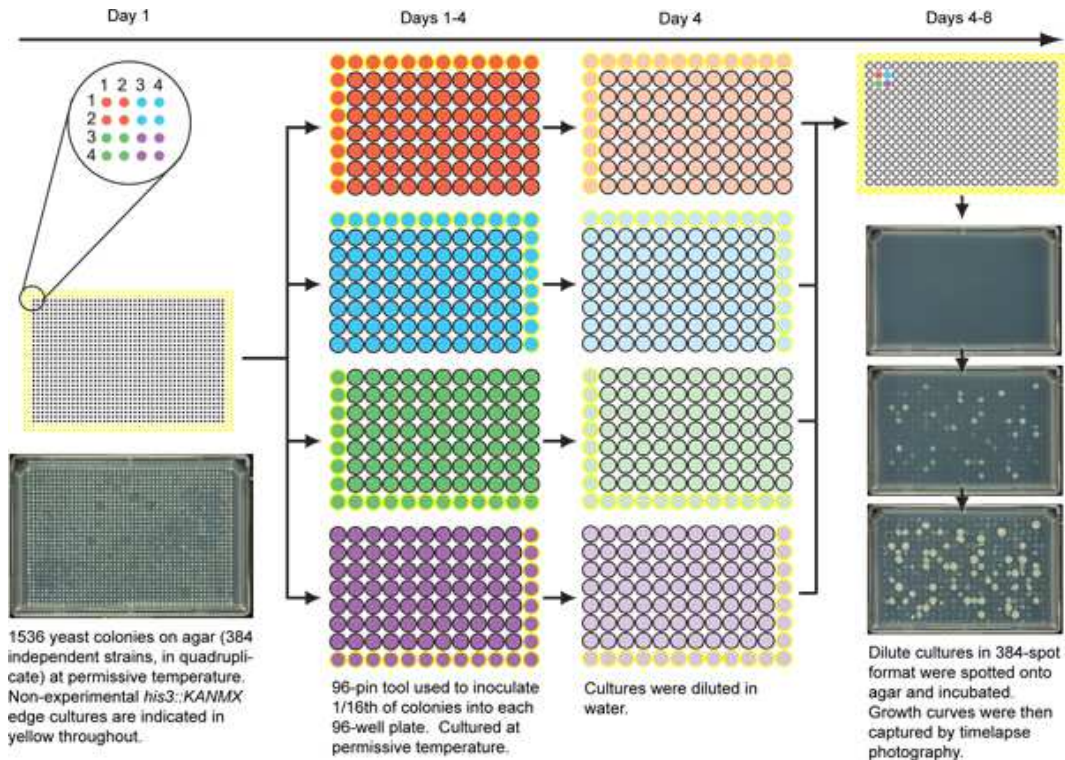
Figure 2.3: The spotting procedure for robotic inoculation of yeast strains in 384-spot format (diagram and legend taken from Banks *et al.* (2012)). This procedure begins with 1536 independent cultures per plate (left). In this typical example, colonies at positions 1,1; 1,2; 2,1 and 2,2 (colored red) are four replicates of the same genotype. *his3::KANMX* cultures in yellow, growing on the edge of the plate, have a growth advantage due to lack of competition and are therefore not examined by Quantitative Fitness Analysis. One of these replicates (e.g. 1,1) is inoculated into liquid growth media in 96-well plates using a 96-pin tool which inoculates 96 out of 1536 colonies each time. In order to inoculate one replicate for each of 384 gene deletions, four different "quadrants" (indicated as red, blue, green and purple) are inoculated into four different 96-well plates containing growth media. After growth to saturation (e.g. 3 days at 20 C), cultures are diluted in water, then the four quadrants from one repeat are spotted in 384-format onto a solid agar plate (right) in the same pattern as the original Synthetic Genetic Array plate (as indicated by color). The process can be repeated to test other replicates: 1,2; 2,1 and 2,2. Example time-lapse images on the right were captured 0.5, 2 and 3.5 days after inoculation.

2.2.1. Jaccard index

For two sample sets, the Jaccard index (Jaccard, 1912; Cheetham & Hazel, 1969) gives a measure of similarity. Where A and B are two sample sets of interest, the Jaccard Index is as follows:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}.$$

The value of J(A,B) can range from 0 to 1, with a larger number for more similarity.

2.2.2. Spearman's rank correlation coefficient

The Spearman's rank correlation coefficient (Spearman, 1987; Kowalczyk *et al.*, 2004) allows comparison of two variables X_i and Y_i, both of sample size n. First, X_i and Y_i are both converted into ranks x_i and y_i. Where there are rank ties or duplicate values, the rank equal to the average of their positions is assigned. The Spearman's rank correlation coefficient is as follows:

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}.$$

The value of ρ can range from -1 to 1. As the relationship between two variables becomes closer to being described by a monotonic function, the larger in magnitude ρ will be.

2.2.3. Gene ontology term enrichment analysis

Gene ontology (GO) term enrichment analysis can give insight to the biological functions of a list of genes (Consortium, 2004). A list of GO terms can be acquired from a list of genes. For yeast the Saccharomyces Genome Database (SGD) (Cherry *et al.*, 2012) can be used to find GO term associations for each gene in the genome. A statistical analysis is carried out to determine which GO terms are most prevalent in a list of genes. The experimenter can then look at GO terms of interest, find out which genes they correspond to and how many are identified in the list.

An unbiased Gene Ontology (GO) term enrichment analyses on a list of genes can be carried out using the software R (R Core Team, 2013) and the bioconductoR package GOstats (Falcon & Gentleman, 2007). There are many other software packages and online services available to carry out a GO term enrichment such as the Database for Annotation, Visualization and Integrated Discovery (DAVID) (Huang *et al.*, 2008, 2009) or the Gene Ontology Enrichment Analysis and Visualization tool (GOrilla) (Eden *et al.*, 2009, 2007).

A GO term clustering analysis is a statistical approach that can be used to follow up a GO term analysis. Information on the relation of GO terms is used in a clustering analysis to find functionally related groups of GO terms. The bioinformatics tool DAVID (Huang *et al.*, 2008, 2009) can be used to carry out GO term clustering (`david.abcc.ncifcrf.gov/`).

2.3. Bayesian inference

A classical (or frequentist) statistical approach typically assumes model unknown parameters are constants and uses the likelihood function to make inference. An alternative methodology is a Bayesian approach (Bernardo & Smith, 2007; Gelman *et al.*, 2003), named after Thomas Bayes (Bayes & Price, 1763). In a Bayesian setting, a parametric model similar to the frequentist approach can be assumed but model parameters are treated as random variables. This feature allows any *prior* knowledge for a given parameter to be incorporated into inference by building a *prior* distribution to describe the information available. We are interested in the *posterior* distribution, that is the probability of the parameters given the evidence. Moreover, where D is the observed data, θ is the set of parameters of interest, we are interested in calculating the *posterior density* $\pi(\theta|D)$. *A priori* knowledge of θ is described by $\pi(\theta)$ and the likelihood of data by $L(D|\theta)$. Using Bayes theorem we obtain the following:

$$\pi(\theta|D) \propto \pi(\theta)L(D|\theta)$$

or
$$Posterior \propto Prior \times likelihood.$$

2.3.1. Markov chain Monte Carlo

In Bayesian inference we are typically interested in sampling from the posterior distribution or one of its marginals, but often this is difficult. Markov Chain Monte Carlo (MCMC) methods are used for sampling from probability distributions (Gamerman, 1997; Gilks *et al.*, 1995). The Monte Carlo name describes the repeated random sampling used to compute results. A Markov chain can be constructed with an equilibrium distribution that is the *posterior* distribution of interest.

A Markov chain $\{X_n, n \in \mathbb{N}^0\}$ is a stochastic process which satisfies the Markov property (or "memoryless" property): for $A \subseteq S$, where S is the continuous state space s.t. $X_n \in S$,

$$P(X_{n+1} \in A | X_n = x, X_{n-1} = x_{n-1}, ..., X_0 = x_0) = P(X_{n+1} \in A | X_n = x),$$

$\forall x, x_{n-1}, ..., x_0 \in S$. The equilibrium distribution $\pi(x)$ is a limiting distribution of a Markov chain with the following two properties. First, there must exist a distribution $\pi(x)$ which is stationary. This condition is guaranteed when the Markov chain satisfies

detailed balance:

$$\pi(x)p(x,y) = \pi(y)p(y,x), \qquad \forall x, y,$$

where $p(x,y)$ is the transition density kernel of the chain. Secondly, the stationary distribution $\pi(x)$ must be unique. This is guaranteed by the ergodicity of the Markov process; see Gamerman (1997) for a definition and sufficient conditions.

2.3.2. Metropolis-Hastings algorithm

The Metropolis-Hastings algorithm (Metropolis *et al.*, 1953; Hastings, 1970) is a MCMC method for obtaining a random sample from a probability distribution of interest (or stationary distribution) (Chib & Greenberg, 1995). With the following procedure a sample from the stationary distribution of the Markov chain can be obtained:

1) Initialise counter $i = 0$ and initialize $X_0 = x_0$

2) From the current position $X_i = x$, generate a candidate value y^* from a proposal density $q(x, y)$.

3) Calculate a probability of acceptance $\alpha(x, y^*)$, where

$$\alpha(x,y) = \begin{cases} \min\left\{1, \frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}\right\} & \text{if } \pi(x)q(x,y) > 0 \\ 1 & \text{otherwise.} \end{cases}$$

4) Accept the candidate value with probability $\alpha(x, y^*)$ and set $X_{i+1} = y^*$, otherwise reject and set $X_{i+1} = x$.

5) Store X_{i+1} and iterate $i = i + 1$.

6) Repeat steps 2-5 until the sample size required is obtained.

The choice of proposal density is important in determining how many iterations are needed to converge to a stationary distribution. There are many choices of proposal distribution (Gamerman, 1997), the simplest case is the symmetric chain. The symmetric chain involves choosing a proposal where $q(x, y) = q(y, x)$, such that step two simplifies

to give the following:

$$\alpha(x, y) = \begin{cases} \min\left\{1, \frac{\pi(y)}{\pi(x)}\right\} & \text{if } \pi(x) > 0 \\ 1 & \text{otherwise.} \end{cases}$$

More general cases are random walk chains and independence chains.

For a random walk chain, the proposed value at stage i is given by $y^* = x_i + w_i$, where w_i are i.i.d. random variables. The distribution for w_i must therefore be chosen, and is typically Normal or Student's t distribution centred at zero. If the distribution for w_i is symmetric, the random walk is a special case of symmetric chains.

For an independence chain, the proposed transition is formed independently of the previous position of the chain, thus $q(x, y) = f(y)$ for some density $f(.)$:

$$\alpha(x, y) = \begin{cases} \min\left\{1, \frac{\pi(y)f(x)}{\pi(x)f(y)}\right\} & \text{if } \pi(x)f(y) > 0 \\ 1 & \text{otherwise.} \end{cases}$$

Parameters within our proposal distribution are known as tuning parameters. They are typically used to adjust the probability of acceptance or improve mixing and must be chosen through some automatic procedure or manually, see Section 2.3.4.

2.3.3. Gibbs sampling

The Gibbs sampler (Gelfand & Smith, 1990; Casella & George, 1992) is a MCMC algorithm for obtaining a random sample from a multivariate probability distribution of interest $\pi(\theta)$, where $\theta = (\theta^1, \theta^2, ..., \theta^d)$. Consider that the full conditional distributions $\pi(\theta_i | \theta_1, ..., \theta_{i-1}, \theta_{i+1}, ..., \theta_d)$, $i = 1, ..., d$ are available. Where it is simpler to sample from conditional distribution than to marginalize by integrating over a joint distribution, the Gibbs sampler is applicable. The following procedure sequentially samples from the full conditional distribution for each parameter, resulting in the probability distribution of interest. The algorithm is as follows:

1) Initialise counter $i = 1$ and parameters $\theta_{(0)} = (\theta^1_{(0)}, \theta^2_{(0)}, ..., \theta^d_{(0)})$.

2) Simulate $\theta^1_{(i)}$ from $\theta^1_{(i)} \sim \pi(\theta^1 | \theta^2_{(i-1)}, ..., \theta^d_{(i-1)})$.

3) Simulate $\theta^2_{(i)}$ from $\theta^2_{(i)} \sim \pi(\theta^2 | \theta^1_{(i)}, \theta^3_{(i-1)}, ..., \theta^d_{(i-1)})$.

4) ...

5) Simulate $\theta_{(i)}^d$ from $\theta_{(i)}^d \sim \pi(\theta^d | \theta_{(i)}^1, ..., \theta_{(i)}^{d-1})$.

6) Store $\theta_{(i)} = (\theta_{(i)}^1, \theta_{(i)}^2, ..., \theta_{(i)}^d)$ and iterate $i = i + 1$.

7) Repeat steps 2-6 until the sample size required is obtained.

To ensure the full conditional distributions for each parameter in a Bayesian model are known and easy to handle, conjugacy can be used. Conjugacy is where the prior is of the same family as the posterior. Conjugacy can be induced by the choice of prior, for example if it is known that a likelihood is Normal with known variance, a Normal prior over the mean will ensure that the posterior is also a Normal distribution.

2.3.4. Convergence issues

To accept output from MCMC algorithms, all chains are required to have reached convergence (Gamerman, 1997; Cowles & Carlin, 1996). Convergence is a requirement to gain unbiased samples of a posterior distribution. Visual and statistical tests can be used to determine if chains have converged, see Section 2.3.5.

Other issues that we must consider for MCMC sampling algorithms are choice of tuning parameters, burn-in period, sample size and thinning, if required. Tuning parameters require a good choice of proposal distribution, preferably with high acceptance rates and good mixing. There are many schemes available for the choice of tuning parameters (Andrieu & Thoms, 2008). Typically tuning parameters are determined during a burn-in period. The burn-in period is a number of iterations which an algorithm must be run for in order to converge to equilibrium. Sample size depends on how many iterations from the posterior are required for both inference and testing convergence. Thinning involves discarding output for iterations of a MCMC algorithm, in order to give less dependent realizations from the posterior distribution.

Extending the length of the burn-in period, sample size and thinning leads to increased computational time. With large data sets and models with a large number of parameters, computation time can become a problem. With a Bayesian modelling approach, computational time associated with MCMC can be much longer than a much simpler least squares approach. This problem is exacerbated when coupled with poor mixing and is likely to

lead the experimenter to simplify their modelling procedure, consequently sacrificing the quality of inference, in order to complete their analysis within a shorter time frame.

2.3.5. Convergence diagnostics

To determine whether chains are true samples from their target distributions, tests for lack of convergence or mixing problems (Gamerman, 1997; Cowles & Carlin, 1996) must be carried out. Typically multiple tests are used to give confidence that the output has convergence. There are many convergence diagnostics for testing chains for convergence, for example the Heidelberg-Welch (Heidelberger & Welch, 1981) and Raftery-Lewis (Raftery & Lewis, 1995) tests. For many convergence diagnostics, summary statistics such as p-values can be used to decide whether convergence has been reached. Visual inspection of diagnostic plots can also be used to determine if convergence has been reached. Trace plots are used to check if samples from the posterior distribution are within a fixed region of plausible values and not exploring the whole range. ACF (auto-correlation function) plots are used to determine serial correlation between sample values of the posterior distribution in order to check for the independence of observations. Density plots are used to check whether a sample posterior distribution is restricted by the choice of prior distribution and determine whether choice of prior is appropriate. Running multiple instances of our MCMC algorithm and comparing chains can also help us decide whether our chains have converged.

2.3.6. Computer programming

To ensure results and inference are reproducible, it is useful to create a computer package so that an analysis can be made in the future without all the code required being re-written. Using freely available software such as the statistical program R (R Core Team, 2013), scripts and commands can be built and shared for easy implementation of code.

Where fast inference is of importance, the choice of programming language is an important consideration. The software package R can also be used as an interface for running code in the C programming language. Statistical code written in the C programming language is typically much faster than using standard R functions or code written in many other programming languages (Fourment & Gillings, 2008).

2.4. Hierarchical modelling

Hierarchical modelling is used to is used to describe the structure of a problem where we believe some population level distribution exists, describing a set of unobserved parameters (Gelman *et al.*, 2003). Examples include pupils nested within classes, children nested within families and patients nested within hospitals. With the pupil-class relationship (2 level-hierarchy), for a given class there may be a number of pupils. We may believe that by being in the same class, pupils will perform similarly in an exam as they are taught by the same teacher. Further, we may have a pupil-class-school relationship (3 level-hierarchy). For a given school, multiple classes exist and in each class there is a number of pupils. We may believe that being within the same school, classes would perform similarly in an exam as they share the same head teacher or school principal.

Hierarchical modelling is used to describe a parent/child relationship (Gelman & Hill, 2006). Repeating the parent/child relationship allows multiple levels to be described. Where a hierarchical structure is known to exist, describing this experimental structure avoids confounding of effects with other sources of variation.

There are many different hierarchical models available, depending on what the experimenter is most interested in (Zuur *et al.*, 2009; Goldstein, 2011). Sharing of information can be built into hierarchical models by the sharing parameters. Allowing parameters to vary at more than one level allows an individual child (subject) effect to be examined. A typical frequentist hierarchical model is built with random effects and has limited distributional assumptions available, whereas a Bayesian hierarchical model is flexible to describe various distributions (Gelman, 2006), see Section 2.4.1.

Plate diagrams allow hierarchical models to be represented graphically (Lunn *et al.*, 2000*b*; Thulasiraman, 1992). Nodes (circles) are used to describe parameters and plates (rectangles) to describe repeating nodes. The use of multiple plates allows nesting to be described.

2.4.1. Distributional assumptions

The flexibility of the Bayesian paradigm allows for models to be built that are otherwise not practical in the frequentist paradigm. More appropriate assumptions can therefore be made to better describe experimental structure and variation in a Bayesian setting (Gelman *et al.*, 2003). For example, inference for a hierarchical *t*-distribution or hierarchical variable section model in a frequentist context is difficult in practise without using MCMC methods that are a more natural fit with Bayesian approaches.

The use of prior distributions allows information from the experimenter and experimental constraints to be incorporated, for instance if a parameter is known to be strictly positive then a positive distribution can be used to enforce this. Truncation can be used to reduce searching posterior areas with extremely low probability.

2.4.2. Indicator variables

Indicator variables are used in variable selection models to describe binary variables (O'Hara & Sillanpaa, 2009). A Bernoulli distributed indicator variable can take the value 0 or 1 to indicate the absence or presence of an effect and can be used to describe binary outcomes such as gender.

2.4.3. The three parameter *t*-distribution

The Student's *t*-distribution has one parameter, namely the degrees of freedom parameter ν which controls the kurtosis of the distribution (Johnson *et al.*, 1995). The Student's *t*-distribution is as follows:

$$t_1(x; \nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}, x \in \mathbb{R}, \nu \in \mathbb{R}^+. \tag{2.1}$$

The ν scale parameter has the effect of increasing the heaviness of the distribution's tails. Adding an additional location parameter μ and scale parameter σ allows further flexibility with the shape of the distribution (Jackman, 2009). The σ scale parameter does not correspond to a standard deviation but does control the overall scale of the distribution. The three parameter *t*-distribution (or scaled *t*-distribution) is then as follows:

$$t_3(x; \mu, \nu, \sigma) = \frac{1}{\sigma} t_1\left(\frac{(x-\mu)}{\sigma}; \nu\right), x \in \mathbb{R}, \nu \in \mathbb{R}^+,$$

where t_1 is given in (2.1).

2.5. Generalisations of the logistic growth model

Where more flexibility than the logistic growth model is required, the logistic growth model (1.1) can be extended by adding parameters (Tsoularis & Wallace, 2002; Jr. *et al.*, 1976). A common extension of the logistic growth model is Richards' growth model (Richards, 1959; Peleg *et al.*, 2007), which adds a single parameter for changing the shape of growth. A more general case to both the logistic and Richards' growth model is

the generalised logistic growth model. Similarly to the logistic growth model (1.1) and its stochastic counterpart (1.8), these more general equations can be extended to diffusion equations if required.

2.5.1. Richards' growth model

Richards' Growth model (Richards, 1959) adds an extra parameter β to the logistic growth equation (1.1). The parameter β affects where maximum growth occurs and consequently the relative growth rate (Tsoularis & Wallace, 2002). Richards' Growth model is as follows:

$$\frac{dx_t}{dt} = rx_t \left[1 - \left(\frac{x_t}{K} \right)^{\beta} \right]. \tag{2.2}$$

The ODE has the following analytic solution:

$$x_t = \frac{K}{(1 + Qe^{-r\beta t})^{\frac{1}{\beta}}},$$
$$\text{where } Q = \left[\left(\frac{K}{P} \right)^{\beta} - 1 \right] e^{\beta t_o},$$

(α, β) are positive real numbers and $t \geq t_0$. When $\beta = 1$, Richards' growth model is equivalent to the logistic growth equation.

2.5.2. Generalised logistic growth model

The generalised logistic growth model adds extra parameters (α, β, γ) to the logistic growth equation (1.1). The extra parameters (α, β, γ) affect where maximum growth occurs, the relative growth rate (Tsoularis & Wallace, 2002) and give a greater selection of curve shapes than the Richards' growth model (2.2). The generalised logistic growth model is as follows:

$$\frac{dx_t}{dt} = rx_t^{\alpha} \left[1 - \left(\frac{x_t}{K} \right)^{\beta} \right]^{\gamma}, \tag{2.3}$$

where (α, β, γ) are positive real numbers and $t \geq t_0$. The generalised logistic growth model cannot in general be integrated to give an analytical solution for x_t. When $\alpha = 1$, $\beta = 1$ and $\gamma = 1$, the generalised logistic growth model is equivalent to the logistic growth equation.

2.6. State space models

A state space model describes the probabilistic dependence between a measurement process Y_t and a state process X_t (West & Harrison, 1997; Durbin *et al.*, 2004). The most basic case of a state space model is as follows:

$$
\begin{aligned}
(X_t|X_{t-1} = x_{t-1}) &\sim f(t, x_{t-1}), \\
(Y_t|X_t = x_t) &\sim g(t, x_t),
\end{aligned}
\tag{2.4}
$$

where f and g are known. A state space model with a linear Gaussian structure has the advantage of allowing us to carry out more efficient MCMC by integrating out latent states with a Kalman filter, instead of imputing all states. The probabilistic representation and the ability to incorporate prior information makes Bayesian inference an appropriate choice for parameter estimation of a state space model.

State space representation provides a general framework for analysing stochastic dynamical systems observed through a stochastic process. A state space model allows us to include both an internal state variable and an output variable in our model. The state-space representation of a stochastic process with measurement error can be given by (2.4) where f is the transition density of the process and g is the assumed measurement error. Inference methods are also readily available to carry out estimation of state space models.

2.6.1. Stochastic differential equations

An ordinary differential equation (ODE) can be used to model a system of interest. For systems with inherent stochastic nature we require a stochastic model. A stochastic differential equation (SDE) is a differential equation where one or more terms include a stochastic process (Wilkinson, 2011; Øksendal, 2010). An SDE differs from an ODE by the addition of a diffusion term, typically a Weiner process, used to describe the intrinsic noise of a given process. A Wiener process (or standard Brownian motion) is a continuous-time stochastic process. A Wiener process $W(t)$, $t \geq 0$, has the following three properties Durrett (1996):

1) $W(0) = 0$.

2) The function $t \to W(t)$ is almost surely everywhere continuous.

3) $W(t)$ has independent increments with $W(t) - W(s) \sim \text{N}(0, t - s)$, for $0 \leq s < t$.

Intrinsic noise from a Weiner process perpetuates the system dynamics of a differential equation. The intrinsic noise is able to propagate though the process, unlike measurement

noise. Instead of inappropriately modelling intrinsic noise by measurement noise, an SDE allows our process to model both system and measurement noise separately.

The simplest case of a stochastic differential equation is of the form:

$$dX(t) = \mu dt + \sigma dW(t),$$

where W denotes a Wiener process. Parameters μ and σ may depend on time and correspond to the drift and diffusion coefficients respectively. The transition density of a stochastic process describes the movement from one state to the next and can be found from the solution of the process.

2.6.2. The Euler-Maruyama method

The Euler-Maruyama method provides an approximate numerical solution of a SDE (Carletti, 2006). For a stochastic process of the form:

$$dX_t = f(X_t)dt + g(X_t)dW_t,$$

where functions f and g are given and W_t is a Wiener process. Given an initial condition $X_0 = x_0$ we can build an Euler-Maruyama approximation of X over an interval $[0, T]$. The Markov chain Y defined below is an Euler-Maruyama approximation to the true solution of X. First we set the initial condition $Y_0 = x_0$. Next, the interval $[0, T]$ is partitioned into N equal subintervals of width $\Delta t > 0$. The Euler-Maruyama approximation is then recursively defined for $1 \leq i \leq N$ as follows:

$$Y_{i+1} = Y_i + f(Y_i)\Delta t + g(Y_i)\Delta W_i,$$

where $\Delta W_i = W_{t_{i+1}} - W_{t_i} \sim \text{N}(0, \Delta t)$. The Euler-Maruyama approximation Y will become a better approximation to the true process X as we increase the size of N.

2.6.3. Kalman filter

The Kalman filter (Kalman, 1960; Welch & Bishop, 1995) is a recursive algorithm that can be used to estimate the state of a dynamic system from a series of incomplete and noisy measurements. The main assumptions of the Kalman filter are that the underlying system is a linear dynamical system and that the noise has known first and second moments. Gaussian noise satisfies the second assumption, for example.

Inference for a state space model (2.4) (see Section 2.6), where both f and g are Gaussian, can be carried out using a Kalman filter. If all noise is zero-mean, uncorrelated and white, then the Kalman filter represents an optimal linear filter (Simon, 2006), even if the noise is not Gaussian. An application of the Kalman filter is given in Section C.5 of the Appendix.

The Kalman filter algorithm is derived as follows: X_{t_i} and Y_{t_i} are the state and measurement processes respectively. w_t and u_t are the state and measurement error respectively, where w_t and u_t are IID, $E[w_t] = 0$, $E[u_t] = 0$, $E[w_t w_t^T] = W_t$ and $E[u_t u_t^T] = U_t$. The Kalman filter can be extended where w_t and u_t are not zero mean. The unobserved latent process is driven by:

$$X_{t_i} | X_{t_{i-1}} \sim \mathrm{N}(G_{t_i} X_{t_{i-1}}, W_{t_i})$$

and the measurement error distribution, relating the latent variable to the observed is given by

$$Y_{t_i} | X_{t_i} \sim \mathrm{N}(F_{t_i}^T X_{t_i}, U_{t_i}),$$

where matrices F_{t_i}, G_{t_i}, U_{t_i} and W_{t_i} are all given. Now, suppose that:

$$X_{t_{i-1}} | Y_{1:t_{i-1}} \sim \mathrm{N}(m_{t_{i-1}}, C_{t_{i-1}}).$$

Incrementing time with $X_{t_i} = G_{t_i} X_{t_{i-1}} + w_{t_{i-1}}$ and condition on $Y_{1:t_{i-1}}$ to give:

$$
\begin{aligned}
X_{t_i} | Y_{1:t_{i-1}} &= G_{t_i} X_{t_{i-1}} | Y_{1:t_{i-1}} + w_{t_i} | Y_{1:t_{i-1}} \\
&= G_{t_i} X_{t_{i-1}} | Y_{1:t_{i-1}} + w_{t_{i-1}},
\end{aligned}
$$

as w_{t_i} is independent of $Y_{1:t_{i-1}}$. We can then show the following using standard multivariate theory:

$$X_{t_i} | Y_{1:t_{i-1}} \sim \mathrm{N}(a_{t_i}, R_{t_i}).$$

where $a_{t_i} = G_{t_i} m_{t_{i-1}}$ and $R_{t_i} = G_{t_i} C_{t_{i-1}} G_{t_i}^T + W_{t_i}$. As $Y_{t_i} = F_{t_i}^T X_{t_i} + u_{t_i}$, and condition on $Y_{1:t_{i-1}}$ to give:

$$
\begin{aligned}
Y_{t_i} | Y_{1:t_{i-1}} &= F_{t_i}^T X_{t_i} | Y_{1:t_{i-1}} + u_{t_i} | Y_{1:t_{i-1}} \\
&= F_{t_i}^T X_{t_i} | Y_{1:t_{i-1}} + u_{t_i},
\end{aligned}
$$

as u_{t_i} is independent of $Y_{1:t_{i-1}}$. We can then show the following using standard multivari-

ate theory:

$$Y_{1:t_i}|Y_{1:t_{i-1}} \sim N(F_{t_i}^T a_{t_i}, F_t^T R_{t_i} F_t + U_{t_i})$$

$Y_{1:t_i}|Y_{1:t_{i-1}}$ and $X_{t_i}|Y_{1:t_{i-1}}$ are therefore jointly Gaussian with the following mean and covariance:

$$\begin{pmatrix} X_{t_i} \\ Y_{1:t_i} \end{pmatrix} \sim MVN \left(\begin{pmatrix} a_{t_i} \\ Y_{t_i} \end{pmatrix}, \begin{pmatrix} R_{t_i} & R_{t_i}F_t \\ F_t^T R_{t_i} & F_t^T R_{t_i}F_t + U_{t_i} \end{pmatrix} \right),$$

Finally, the following multivariate theorem is used:

$$\text{if } \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \sim MVN \left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right),$$

$$\text{then } Y_1|Y_2 = y_2 \sim MVN \left(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(y_2 - \mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \right),$$

to obtain the following:

$$X_{t_i}|Y_{1:t_i} \sim N(m_{t_i}, C_{t_i}),$$
$$\text{where } m_{t_i} = a_{t_i} + R_{t_i}F(F^T R_{t_i}F + U)^{-1}[Y_{t_i} - F^T a_{t_i}] \tag{2.5}$$
$$\text{and } C_{t_i} = R_{t_i} - R_{t_i}F(F^T R_{t_i}F + U)^{-1}F^T R_{t_i}.$$

Parameters m_0 and C_0 must be initialised first, then using the equations in (2.5), m_{t_i} and C_{t_i} can be recursively estimated.

Typically, the Kalman filter is used to make inference for a hidden state process, but it can be used to reduce computational time in algorithms for inferring process hyperparameters by recursively computing the marginal likelihood $\pi(y_{t_{1:N}})$ (West & Harrison, 1997), where

$$\pi(y_{t_{1:N}}) = \prod_{i=1}^{N} \pi(y_{t_i}|y_{t_{1:(i-1)}})$$

and $\pi(y_{t_i}|y_{t_{1:(i-1)}}) = \int_X \pi(y_{t_i}, x_{t_i}|y_{t_{1:(i-1)}})dx_{t_i} = \int_X \pi(y_{t_i}|x_{t_i})\pi(x_{t_i}|y_{t_{1:(i-1)}})dx_{t_i}$ gives a tractable Gaussian integral. The procedure for computing the marginal likelihood $\pi(y_{t_{1:N}})$ using the Kalman filter algorithm is as follows:

1) Initialise with prior knowledge for X_0 and set $i = 1$.

2) Prediction step from $X_{t_{i-1}}|Y_{1:t_i-1}$ to $X_{t_i}|Y_{1:t_i-1}$ (giving $\pi(x_{t_i}|y_{1:t_i-1})$).

3) Calculate and store $\pi(y_{t_i}|y_{1:t_{i-1}})$.

4) Update step to give $X_{t_i}|Y_{1:t_i}$, then iterate $i = i + 1$.

5) Repeat steps 2-4 (and compute $\pi(y_{t_{1:N}})$).

2.6.4. Linear noise approximation

The linear noise approximation (LNA) (Kurtz, 1970, 1971; Van Kampen, 2011) reduces a non-linear SDE to a linear SDE with additive noise, which can be solved (Wallace, 2010; Komorowski *et al.*, 2009). The LNA assumes the solution of a diffusion process Y_t can be written as $Y_t = v_t + Z_t$ (a deterministic part v_t and stochastic part Z_t), where Z_t remains small for all $t \in \mathbb{R}_{\geq 0}$. The LNA is useful when a tractable solution to a SDE cannot be found. Typically the LNA is used to reduce an SDE to a Ornstein-Uhlenbeck process which can be solved explicitly. Ornstein-Uhlenbeck processes are Gaussian, time discretising the resulting LNA will therefore give us a linear Gaussian state space model with an analytically tractable transition density available. The LNA can be viewed as a first order Taylor expansion of an approximating SDE about a deterministic solution (higher order approximations are possible (Gardiner, 2010)). We can also view the LNA as an approximation of the chemical Langevin equation (Wallace *et al.*, 2012). Applications of the LNA to non-linear SDEs are given in Section 5.3 and 5.4.

Chapter 3. Modelling genetic interaction

3.1. Introduction

In this chapter, alternative modelling approaches are developed to better model a QFA screen comparison than the current frequentist Addinall *et al.* (2011) approach. Section 3.2 presents the modelling assumptions for the development of a Bayesian approach. Two Bayesian approaches are then presented in Sections 3.3 and 3.4, incorporating some model assumptions that are not convenient in a frequentist setting. So that our Bayesian models can be compared with a frequentist hierarchical modelling approach, a random effects model is then presented in Section 3.5.

The models in this chapter are compared using previously analysed *S. cerevisiae* QFA screen data in the next chapter. Historic *S. cerevisiae* QFA screen datasets are used to shape the model assumptions adopted in the following sections.

3.2. Bayesian hierarchical model inference

As an alternative to the maximum likelihood approach presented by Addinall *et al.* (2011), we present a Bayesian, hierarchical methodology where *a priori* uncertainty about each parameter value is described by probability distributions (Bernardo & Smith, 2007) and information about parameter distributions is shared across $orf\Delta$s and conditions. Plausible frequentist estimates from across 10 different historic QFA data sets, including a wide range of different background mutations and treatments were used to quantify *a priori* uncertainty in model parameters.

Prior distributions describe our beliefs about parameter values. These should be diffuse enough to capture all plausible values (to capture the full range of observations in the datasets) while being restrictive enough to rule out implausible values (to ensure efficient inference). Inappropriate choice of priors can result in chains drifting during mixing and becoming stuck in implausible regions. Although using conjugate priors would allow faster inference, we find that the conjugate priors available for variance parameters (Gelman, 2006) are either too restrictive at low variance (Inverse-gamma), not restrictive enough at low variance (half-t family of prior distributions) or are non-informative or largely discard the prior information available (Uniform). Our choice for the priors of precision parameters is the non-conjugate Log-normal as we find the distribution is only

restrictive at extremely high and low variances.

We use three types of distribution to model parameter uncertainty: Log-normal, Normal and scaled t-distribution with three degrees of freedom. We use the Log-normal distribution to describe parameters which are required to be non-negative (e.g. parameters describing precisions, or repeat-level fitnesses) or parameter distributions which are found by visual inspection to be asymmetric. We use the Normal distribution to describe parameters which are symmetrically distributed (e.g. some prior distributions and the measurement error model) and we use the t-distribution to describe parameters whose uncertainty distribution is long-tailed (i.e. where using the Normal distribution would result in excessive shrinkage towards the mean). A Normal distribution was considered for describing the variation in *orf∆*s but was found to be inappropriate, failing to assign density at the extreme high and low fitnesses. For example, after visual inspection of frequentist *orf∆* level means about their population mean, we found there to be many unusually fit, dead or missing *orf∆* and concluded that *orf∆* fitnesses would be well modelled by the t-distribution.

Instead of manually fixing the inoculum density parameter P as in Addinall *et al.* (2011) our Bayesian hierarchical models deal with the scarcity of information about the early part of culture growth curves by estimating a single P across all *orf∆*s (and conditions in some of our models). Our new approach learns about P from the data and gives us a posterior distribution to describe our uncertainty about its value.

The new, hierarchical structure implemented in our models (Goldstein, 2011) reflects the structure of QFA experiments. Information is shared efficiently among groups of parameters such as between repeat level parameters for a single mutant strain. An example of the type of Bayesian hierarchical modelling which we use to model genetic interaction can be seen in Yi (2010), where hierarchical models are used to account for group effects.

In Phenix *et al.* (2011) the signal of genetic interaction is chosen to be "strictly ON or OFF" when modelling gene activity. We include this concept in our interaction models by using a Bernoulli distributed indicator variable (O'Hara & Sillanpaa, 2009) to describe whether there is evidence of an *orf∆* interacting with the query mutation; the more evidence of interaction, the closer posterior expectations will be to one.

Failing to account for all sources of variation within the experimental structure, such as the difference in variation between the control and query fitnesses, may lead to inaccurate conclusions. By incorporating more information into the model with prior distributions and a more flexible modelling approach, we will increase statistical power. With an improved analysis it may then be possible for a similar number of genetic interactions

to be identified with a smaller sample size, saving on the significant experimental costs associated with QFA.

Inference is carried out using Markov Chain Monte Carlo (MCMC) methods. The algorithm used is a Metropolis-within-Gibbs sampler where each full-conditional is sampled in turn either directly or using a simple Normal random walk Metropolis step. Due to the large number of model parameters and large quantity of data from high-throughput QFA experiments, the algorithms used for carrying out inference often have poor mixing and give highly auto-correlated samples, requiring thinning. Posterior means are used to obtain point estimates where required.

For the new Bayesian approaches (described in Section 3.3 and 3.4), model fitting is carried out using the techniques discussed above, implemented in C for computational speed, and is freely available in the R package "qfaBayes" at `https://r-forge.r-project.org/projects/qfa`.

3.3. Two-stage Bayesian hierarchical approach

In the following sections, a two-stage Bayesian, hierarchical modelling approach (see Section 3.3.1 and 3.3.2) is presented. The following two-stage Bayesian approach generates *orf△* fitness distributions and infers genetic interaction probabilities separately. For a QFA screen comparison, first the separate hierarchical model (SHM) given in Section 3.3.1, is fit to each screen separately and a set of logistic growth parameter estimates obtained for each time-course. Secondly, each set of logistic growth parameter estimates is converted into a univariate fitness summary and input to the interaction hierarchical model (IHM) given in Section 3.3.2, to determine which genes show evidence of genetic interaction.

3.3.1. Separate hierarchical model

The separate hierarchical model (SHM), presented in Table 3.1, models the growth of multiple yeast cultures using the logistic function described in (1.2). In this first hierarchical model, the logistic model is fit to the query and control strains separately.

In order to measure the variation between *orf△*s, parameters (K^p, σ_o^K) and (r^p, σ_o^r) are included at the population level of the hierarchy. Within-*orf△* variation is modelled by each set of *orf△* level parameters (K_l^o, τ_l^K) and (r_l^o, τ_l^r). Learning about these higher level parameters allows information to be shared across parameters lower in the hierarchy. A three-level hierarchical model is applied to (K, K_l^o, K_{lm}) and (r, r_l^o, r_{lm}), sharing infor-

mation on the repeat level and the *orfΔ* level. Note that *orfΔ* level parameters K_l^o and r_l^o are on the log scale ($e^{K_l^o}$ and $e^{r_l^o}$ are on the scale of the observed data).

Assuming a Normal error structure, random measurement error is modelled by the ν_l parameters (one for each *orfΔ*). Information on random error is shared across all *orfΔ*s by drawing $\log \nu_l$ from a normal distribution parameterised by (ν_p, σ^ν). A two-level hierarchical structure is also used for both the τ_l^K and τ_l^r parameters.

Modelling logistic model parameter distributions on the log scale ensures that parameter values remain strictly positive (a realistic biological constraint). Truncating distributions allows us to implement further, realistic constraints on the data. Truncating $\log r_{lm}$ values greater than 3.5 corresponds to disallowing biologically unrealistic culture doubling times faster than about 30 minutes and truncating of repeat level parameters $\log K_{lm}$ above 0 ensures that no carrying capacity estimate is greater than the maximum observable cell density, which is 1 after scaling.

orfΔ level parameters $e^{K_o^l}$ and $e^{r_o^l}$ are on the same scale as the observed data. Realistic biological constraints (positive logistic model parameters) are enforced at the repeat level, however both $e^{K_o^l}$ and $e^{r_o^l}$, which are assumed to have scaled t-distributions, are truncated below zero to keep exponentiated parameters strictly positive. Most *orfΔ* level logistic growth parameters are distributed in a bell shape around some mean value, it is the unusually fit, dead or missing *orfΔ*s within a typical QFA screen that require the use of a long tailed distribution such as the scaled t-distribution with 3 degrees of freedom. The non-standard choice of a truncated scaled t-distribution with 3 degrees of freedom ensures that the extreme high and low values have probability assigned to them regardless of the population level location and scale parameters for a given QFA screen.

For example, after visual inspection of frequentist *orfΔ* level means about their population mean, we found there to be many unusually fit, dead or missing *orfΔ* and concluded that *orfΔ* fitnesses would be well modelled by the t-distribution.

Identifiability problems can arise for parameters K_{lm} and r_{lm} when observed cell densities are low and unchanging (consistent with growth curves for cultures which are very sick, dead or missing). In these cases, either K_{lm} or r_{lm} can take values near zero, allowing the other parameter to take any value without significantly affecting the model fit. In the Addinall *et al.* (2011) approach identification problems are handled in an automated post-processing stage: for cultures with low K estimates (classified as dead), r is automatically set to zero. Without correcting for identification problems in our Bayesian models, misleading information from implausible values will be shared across our models. Computing time wasted on such identifiability problems is reduced by truncating repeat level

Table 3.1: Description of the separate hierarchical model (SHM). Dependent variable y_{lmn} (scaled cell density measurements) and independent variable t_{lmn} (time since inoculation) are data input to the SHM. $x(t)$ is the solution to the logistic model ODE given in (1.2). l indicates a particular *orfΔ* from the gene deletion library, m indicates a repeat for a given *orfΔ* and n indicates the time point for a given *orfΔ* repeat.

$$l = 1, 2, ..., L \qquad\qquad\qquad \textit{orf}\Delta \text{ level}$$
$$m = 1, ..., M_l \qquad\qquad\qquad \text{Repeat level}$$
$$n = 1, 2, ..., N_{lm} \qquad\qquad\qquad \text{Time point level}$$

Time point level

$$y_{lmn} \sim \text{N}(\hat{y}_{lmn}, (\nu_l)^{-1}) \qquad\qquad \hat{y}_{lmn} = x(t_{lmn}; K_{lm}, r_{lm}, P)$$

Repeat level

$$\log K_{lm} \sim \text{N}(K_l^o, (\tau_l^K)^{-1}) I_{(-\infty,0]} \qquad \log \tau_l^K \sim \text{N}(\tau^{K,p}, (\sigma^{\tau,K})^{-1}) I_{[0,\infty)}$$
$$\log r_{lm} \sim \text{N}(r_l^o, (\tau_l^r)^{-1}) I_{(-\infty,3.5]} \qquad \log \tau_l^r \sim \text{N}(\tau^{r,p}, (\sigma^{\tau,r})^{-1})$$

orfΔ level

$$e^{K_l^o} \sim t(K^p, (\sigma^{K,o})^{-1}, 3) I_{[0,\infty)} \qquad \log \sigma^{K,o} \sim \text{N}(\eta^{K,o}, (\psi^{K,o})^{-1})$$
$$e^{r_l^o} \sim t(r^p, (\sigma^{r,o})^{-1}, 3) I_{[0,\infty)} \qquad \log \sigma^{r,o} \sim \text{N}(\eta^{r,o}, (\psi^{r,o})^{-1})$$
$$\log \nu_l \sim \text{N}(\nu^p, (\sigma^\nu)^{-1}) \qquad\qquad \log \sigma^\nu \sim \text{N}(\eta^\nu, (\psi^\nu)^{-1})$$

Population level

$$\log K^p \sim \text{N}(K^\mu, (\eta^{K,p})^{-1}) \qquad\qquad \log r^p \sim \text{N}(r^\mu, (\eta^{r,p})^{-1})$$
$$\log P \sim \text{N}(P^\mu, (\eta^P)^{-1}) \qquad\qquad \nu^p \sim \text{N}(\nu^\mu, (\eta^{\nu,p})^{-1})$$
$$\tau^{K,p} \sim \text{N}(\tau^{K,\mu}, (\eta^{\tau,K,p})^{-1}) \qquad\qquad \log \sigma^{\tau,K} \sim \text{N}(\eta^{\tau,K}, (\psi^{\tau,K})^{-1})$$
$$\tau^{r,p} \sim \text{N}(\tau^{r,\mu}, (\eta^{\tau,r,p})^{-1}) \qquad\qquad \log \sigma^{\tau,r} \sim \text{N}(\eta^{\tau,r}, (\psi^{\tau,r})^{-1})$$

parameters r_{lm}, preventing the MCMC algorithms from becoming stuck in extremely low probability regions when K_{lm} takes near zero values. Similarly, $\log \tau_l^K$ parameters are truncated below 0 to overcome identifiability problems between parameters K_{lm} and r_{lm} when r_{lm} takes near zero values.

The SHM in Table 3.1 is fit to both the query and control strains separately. Means are taken to summarise logistic growth parameter posterior distributions for each *orf∆* repeat. Summaries $(\hat{K}_{lm}, \hat{r}_{lm}, \hat{P})$ for each *orf∆* repeat are converted to univariate fitnesses F_{clm}, where c identifies the condition (query or control), with any given fitness measure e.g. $MDR \times MDP$ (see (1.3) and Addinall *et al.* (2011)). A problem of the two-stage approach is that we must choose a fitness definition most relevant to the experiment. We choose the same definition used in Addinall *et al.* (2011), $MDR \times MDP$, for the comparison of our methods. An alternative choice of fitness definition could be used given sufficient biological justification. Section 1.1.3 gives the derivations of MDR and MDP. The product of $MDR \times MDP$ is used as it accounts for the attributes of two definitions simultaneously.

The flow of information within the model and how each parameter is related to the data can be seen from the plate diagram in Figure 3.1 (Lunn *et al.*, 2000*b*).
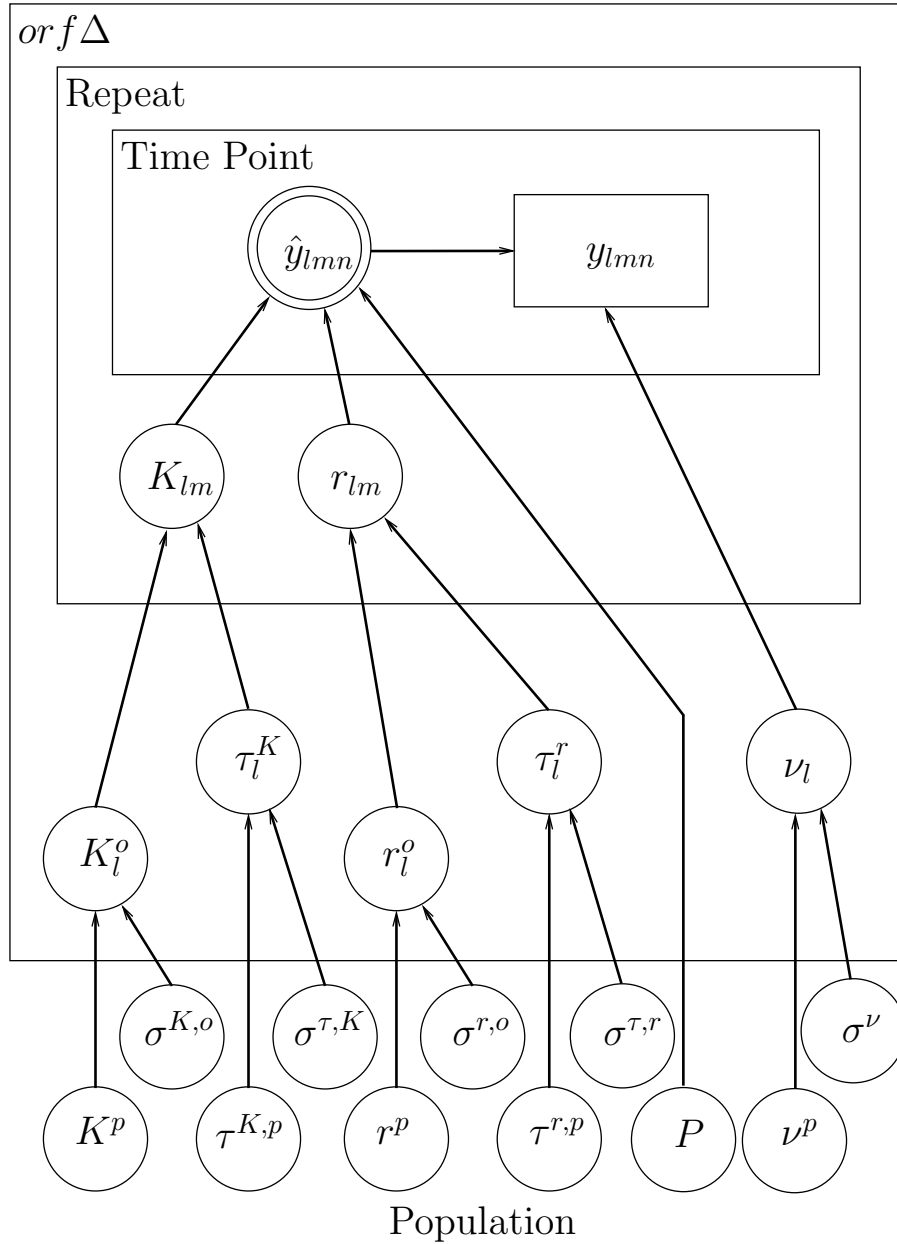
Figure 3.1: Plate diagram for the separate hierarchical model, described in Section 3.3.1. This figure shows the four levels of hierarchy in the SHM model, population, $orf\Delta$ (l), repeat (m) and time point (n). Prior hyperparameters for the population parameters are omitted. A circular node represents a parameter in the model. An arrow from a source node to a target node indicates that the source node parameter is a prior hyperparameter for the target node parameter. Each rectangular box corresponds to a level of the hierarchy. Nodes within multiple boxes are nested and their parameters are indexed by corresponding levels of the hierarchy. The node consisting of two concentric circles corresponds to the models fitted values. The rectangular node represents the observed data.

3.3.2. Interaction hierarchical model

After the SHM fit, the IHM, presented in Table 3.2, can then be used to model estimated fitness scores F_{clm} and determine, for each $orf\Delta$, whether there is evidence for interaction.

Fitnesses are passed to the IHM where query screen fitnesses are compared with control screen fitnesses, assuming genetic independence. Deviations from predicted fitnesses are evidence for genetic interaction. The flow of information within the IHM and how each parameter is related to the data can be seen from the plate diagram in Figure 3.2.

The interaction model accounts for between $orf\Delta$ variation with the set of parameters (Z^p, σ_Z) and within $orf\Delta$ variation by the set of parameters (Z_l, ν_l). A linear relationship between the control and query $orf\Delta$ level parameters is specified with a scale parameter α_1. Any deviation from this relationship (genetic interaction) is accounted for by the term $\delta_l \gamma_{1,l}$. δ_l is a binary indicator of genetic interaction for $orf\Delta\ l$. A scaling parameter α_1 allows any effects due to differences in the control and query data sets to be scaled out, such as differences in genetic background, incubator temperature or inoculum density.

The linear relationship between the control and query fitness scores, consistent with the multiplicative model of genetic independence, described in (1.5), is implemented in the IHM as: $\hat{F} = e^{\alpha_c + Z_l + \delta_l \gamma_{cl}} = e^{\alpha_c} e^{Z_l + \delta_l \gamma_{cl}}$. Strains whose fitnesses lie along the linear relationship defined by the scalar α_1 show no evidence for interaction with the query condition. On the other hand, deviation from the linear relationship, represented by the posterior mean of $\delta_l \gamma_{1,l}$ is evidence for genetic interaction. The larger the posterior mean for δ_l is the higher the probability or evidence there is for interaction, while $\gamma_{1,l}$ is a measure of the strength of interaction. Where the query condition has a negative effect (i.e. decreases fitness on average, compared to the control condition), query fitnesses which are above and below the linear relationship are suppressors and enhancers of the fitness defect associated with the query condition respectively. A list of gene names are ordered by $\delta_l \gamma_{cl}$ posterior means and those $orf\Delta$s with $\hat{\delta}_l > 0.5$ will be classified and labelled as showing "significant" evidence of interaction.

The Bernoulli probability parameter p is our prior estimate for the probability of a given $orf\Delta$ showing evidence of genetic interaction. For a typical yeast QFA screen, p is set to 0.05 as the experimenter's belief before the experiment is carried out is that 5% of our $orf\Delta$s exhibit genetic interactions. Observational noise is quantified by ν_{cl}. The ν_{cl} parameter accounts for difference in variation between condition i.e. the query and control data sets and for difference in variation between $orf\Delta$s.

Table 3.2: Description of the interaction hierarchical model (IHM). F_{clm} are the observed fitness scores, where c identifies the condition for a given *orfΔ*, l identifies a particular *orfΔ* from the gene deletion library and m identifies a repeat for a given *orfΔ*.

$$c = 0, 1 \qquad\qquad \text{Condition level}$$
$$l = 1, ..., L_c \qquad\qquad \text{\textit{orf}Δ level}$$
$$m = 1, ..., M_{cl} \qquad\qquad \text{Repeat level}$$

Repeat level

$$F_{clm} \sim \text{N}(\hat{F}_{cl}, (\nu_{cl})^{-1}) \qquad\qquad \hat{F}_{cl} = e^{\alpha_c + Z_l + \delta_l \gamma_{cl}}$$

orfΔ level

$$e^{Z_l} \sim t(Z^p, (\sigma^Z)^{-1}, 3) I_{[0,\infty)} \qquad\qquad \log \sigma^Z \sim \text{N}(\eta^Z, \psi^Z)$$
$$\log \nu_{cl} \sim \text{N}(\nu^p, (\sigma^\nu)^{-1}) \qquad\qquad \log \sigma^\nu \sim \text{N}(\eta^\nu, \psi^\nu)$$
$$\delta_l \sim Bern(p)$$
$$e^{\gamma_{cl}} = \begin{cases} 1 & \text{if } c = 0; \\ t(1, (\sigma^\gamma)^{-1}, 3) I_{[0,\infty)} & \text{if } c = 1. \end{cases} \qquad \log \sigma^\gamma \sim \text{N}(\eta^\gamma, (\psi^\gamma)^{-1})$$

Condition level

$$\alpha_c = \begin{cases} 0 & \text{if } c = 0; \\ \text{N}(\alpha^\mu, \eta^\alpha) & \text{if } c = 1. \end{cases}$$

Population level

$$\log Z^p \sim N(Z^\mu, (\eta^{Z,p})^{-1}) \qquad\qquad \nu^p \sim \text{N}(\nu^\mu, (\eta^{\nu,p})^{-1})$$

Figure 3.2: Plate diagram for the interaction hierarchical model, described in Section 3.3.2. This figure shows the four levels of hierarchy in the IHM model: population, $orf\Delta$ (l), condition (c) and repeat (m). Prior hyperparameters for population parameters are omitted. Plate diagram notation as in Figure 3.1.

3.4. One-stage Bayesian hierarchical approach

Following from Section 3.3, a one-stage approach for inferring fitness and genetic interaction probabilities separately is presented. All of the SHM and IHM modelling assumptions described in Section 3.3, such as distributional choices and hierarchical structure are inherited by the one stage approach known as the joint hierarchical model (JHM).

3.4.1. Joint hierarchical model

The JHM given in Table 3.3 is an alternative, fully Bayesian version of the two-stage approach described in Section 3.3.1 and 3.3.2. The JHM incorporates the key modelling ideas from both the SHM and the IHM with the considerable advantage that we can learn about logistic growth model, fitness and genetic interaction parameters simultaneously, thereby avoiding having to choose a fitness measure or point estimates for passing information between models. The JHM is an extension of the SHM with the presence or absence of genetic interaction being described by a Bernoulli indicator and an additional level of error to account for variation due to the query condition. Genetic interaction is modelled in terms of the two logistic growth parameters K and r simultaneously. Similar to the interaction model in Section 3.3.2 in Chapter 3.3, linear relationships between control and query carrying capacity and growth rate (instead of fitness score) are assumed: $\left(e^{\alpha_c + K_l^o + \delta_l \gamma_{cl}}, e^{\beta_c + r_l^o + \delta_l \omega_{cl}} \right)$.

By fitting a single JHM, we need only calculate posterior means, check model diagnostics and thin posteriors once. However, the CPU time taken to reach convergence for any given data set is roughly twice that of the two-stage approach for a genome-wide QFA.

The flow of information within the model and how each parameter is related to the data can be seen from the plate diagram in Figure 3.3.

**Table 3.3:** Description of the joint hierarchical model (JHM). The dependent variable y_{clmn} (scaled cell density measurements) and independent variable t_{clmn} (time since inoculation) are input to the JHM. c identifies the condition for a given $orf\Delta$, l identifies a particular $orf\Delta$ from the gene deletion library, m identifies a repeat for a given $orf\Delta$ and n identifies the time point for a given condition and $orf\Delta$ repeat.

$$c = 0, 1 \qquad \text{Condition level}$$
$$l = 1, ..., L_c \qquad orf\Delta \text{ level}$$
$$m = 1, ..., M_{cl} \qquad \text{Repeat level}$$
$$n = 1, ..., N_{clm} \qquad \text{Time point level}$$

Time point level

$$y_{clmn} \sim \mathrm{N}(\hat{y}_{clmn}, (\nu_{cl})^{-1}) \qquad\qquad \hat{y}_{clmn} = x(t_{clmn}; K_{clm}, r_{clm}, P)$$

Repeat level

$$\log K_{clm} \sim \mathrm{N}(\alpha_c + K_l^o + \delta_l\gamma_{cl}, (\tau_{cl}^K)^{-1})I_{(-\infty,0]} \qquad \log \tau_{cl}^K \sim \mathrm{N}(\tau_c^{K,p}, (\sigma_c^{\tau,K})^{-1})I_{[0,\infty)}$$
$$\log r_{clm} \sim \mathrm{N}(\beta_c + r_l^o + \delta_l\omega_{cl}, (\tau_{cl}^r)^{-1})I_{(-\infty,3.5]} \qquad \log \tau_{cl}^r \sim \mathrm{N}(\tau_c^{r,p}, (\sigma_c^{\tau,r})^{-1})$$

orfΔ level

$$e^{K_l^o} \sim t(K^p, (\sigma^{K,o})^{-1}, 3)I_{[0,\infty)} \qquad\qquad \log \sigma^{K,o} \sim \mathrm{N}(\eta^{K,o}, (\psi^{K,o})^{-1})$$
$$e^{r_l^o} \sim t(r^p, (\sigma^{r,o})^{-1}, 3)I_{[0,\infty)} \qquad\qquad \log \sigma^{r,o} \sim \mathrm{N}(\eta^{r,o}, (\psi^{r,o})^{-1})$$
$$\log \nu_{cl} \sim \mathrm{N}(\nu^p, (\sigma^\nu)^{-1}) \qquad\qquad \log \sigma^\nu \sim \mathrm{N}(\eta^\nu, (\psi^\nu)^{-1})$$
$$\delta_l \sim Bern(p)$$

$$e^{\gamma_{cl}} = \begin{cases} 1 & \text{if } c = 0; \\ t(1, (\sigma^\gamma)^{-1}, 3)I_{[0,\infty)} & \text{if } c = 1. \end{cases} \qquad \log \sigma^\gamma \sim \mathrm{N}(\eta^\gamma, \psi^\gamma)$$

$$e^{\omega_{cl}} = \begin{cases} 1 & \text{if } c = 0; \\ t(1, (\sigma^\omega)^{-1}, 3)I_{[0,\infty)} & \text{if } c = 1. \end{cases} \qquad \log \sigma^\omega \sim \mathrm{N}(\eta^\omega, \psi^\omega)$$

Condition level

$$\alpha_c = \begin{cases} 0 & \text{if } c = 0; \\ \mathrm{N}(\alpha^\mu, \eta^\alpha) & \text{if } c = 1. \end{cases} \qquad\qquad \beta_c = \begin{cases} 0 & \text{if } c = 0; \\ \mathrm{N}(\beta^\mu, \eta^\beta) & \text{if } c = 1. \end{cases}$$
$$\tau_c^{K,p} \sim \mathrm{N}(\tau^{K,\mu}, (\eta^{\tau,K,p})^{-1}) \qquad\qquad \log \sigma_c^{\tau,K} \sim \mathrm{N}(\eta^{\tau,K}, (\psi^{\tau,K})^{-1})$$
$$\tau_c^{r,p} \sim \mathrm{N}(\tau^{r,\mu}, (\eta^{\tau,r,p})^{-1}) \qquad\qquad \log \sigma_c^{\tau,r} \sim \mathrm{N}(\eta^{\tau,r}, (\psi^{\tau,r})^{-1})$$

Population level

$$\log K^p \sim \mathrm{N}(K^\mu, (\eta^{K,p})^{-1}) \qquad\qquad \log r^p \sim \mathrm{N}(r^\mu, (\eta^{r,p})^{-1})$$
$$\nu^p \sim \mathrm{N}(\nu^\mu, (\eta^{\nu,p})^{-1}) \qquad\qquad \log P \sim \mathrm{N}(P^\mu, (\eta^P)^{-1})$$
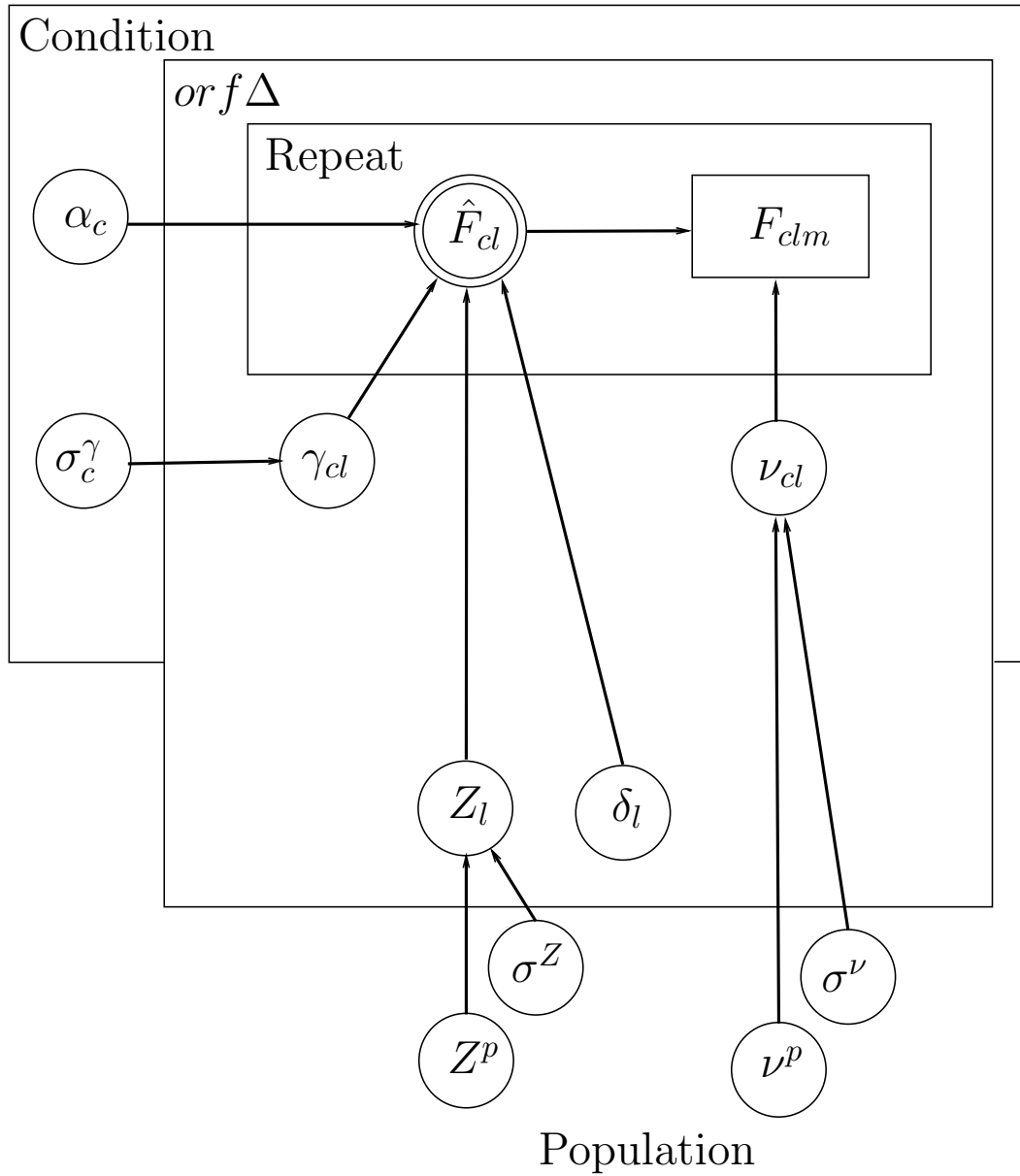
Figure 3.3: Plate diagram for the joint hierarchical model, described in Section 3.4.1. This figure shows the five levels of hierarchy in the JHM model, population, $orf\Delta$ (l), condition (c), repeat (m) and time point (n). Prior hyperparameters for the population parameters are omitted. Plate diagram notation is given in Figure 3.1.

49

3.5. Random effects model

To improve on the Addinall *et al.* (2011) modelling approach whilst remaining within the frequentist paradigm, by accounting for the hierarchical structure of the data, a random effects model (Zuur *et al.*, 2009; Pinheiro & Bates, 2000) can be used. The random effects model (REM) given in Table 3.4 is used to model estimated fitness scores F_{clm} from (1.6) and estimate evidence of interaction for each *orfΔ* simultaneously with a single model fit. Introducing a random effect Z_l allows us to account for between subject variation by estimating a single $\sigma_Z{}^2$. Unlike the Addinall *et al.* (2011) approach, observed values F_{clm} are not scaled and instead a parameter to model a condition effect μ_c is introduced.

γ_{cl} represents the estimated strength of genetic interaction between an *orfΔ* and its query mutation counterpart. For a multiplicative model of epistasis, an additive model is used to describe the log transformed data $f_{clm} = \log(F_{clm} + 1)$, where F_{clm} are the observed fitnesses. We use the Benjamini-Hochberg test to correct for multiple testing in order to make a fair comparison with the (Addinall *et al.*, 2011) approach.

Inference for a frequentist random effects model can be carried out most simply with the R package "lme4" (Bates *et al.*, 2013). For the R code to fit the REM see Section A.3 of the Appendix. In the frequentist paradigm some parameters cannot be modelled as random effects since computational difficulties associated with large matrix computations arise with multiple random effects and very large data sets. Similarly, a more appropriate model with a log-link function in order to model repeat level variation with a normal distribution cannot be fit, due to computational difficulties that arise with non-linear model maximum likelihood algorithms and large data sets. Such computational difficulties cause algorithms for parameter estimation to fail to converge.

Table 3.4: Description of the random effects model (REM). c identifies the condition for a given *orfΔ*, l identifies a particular *orfΔ* from the gene deletion library and m identifies a repeat for a given *orfΔ*.

$$f_{clm} = \mu_c + Z_l + \gamma_{cl} + \varepsilon_{clm}$$

$$\mu_c = \begin{cases} \mu + \alpha & \text{if } c = 0; \\ \mu & \text{if } c = 1. \end{cases} \qquad \gamma_{cl} = \begin{cases} 0 & \text{if } c = 0; \\ \gamma_l & \text{if } c = 1. \end{cases}$$

$$Z_l \sim \mathcal{N}(0, \sigma_Z{}^2) \qquad\qquad\qquad \varepsilon_{clm} \sim \mathcal{N}(0, \sigma^2)$$

Chapter 4. Case Studies

4.1. Introduction

In this chapter, the new Bayesian models developed in Chapter 3 are applied to previously analysed QFA screen data. The one-stage and two-stage Bayesian approaches are compared with the two-stage Addinall *et al.* (2011) and random effects model (REM) approaches for a QFA screen comparison designed to inform the experimenter about telomere biology in *S. cerevisiae*.

After comparing the approaches developed, the one-stage Bayesian joint hierarchical model (JHM) is found to best model a QFA screen comparison. The JHM is then applied to further examples of *S. cerevisiae* QFA screen data to demonstrate the JHM's ability to model different experiments. Two extensions of the JHM are then considered, to account for a batch effect and a transformation effect within a QFA screen comparison. Fitness plots for the further case studies and extensions of the JHM are included for further investigation and research.

The new one-stage Bayesian QFA will be used at first to help identify genes that are related to telomere activity, but the analysis is general enough to be applicable to any high-throughput study of arrayed microbial cultures (including experiments such as drug screening).

4.2. *cdc13-1* 27°C vs *ura3△* 27°C suppressor/enhancer data set

The following analysis is for a QFA experiment comparing query *cdc13-1 s*trains with control *ura3△* strains at 27°C, previously analysed by Addinall *et al.* (2011), to identify genes that show evidence of genetic interaction with the query mutation *cdc13-1*. The ability of the Cdc13 protein produced by *cdc13-1* strains to cap telomeres is reduced at temperatures above 26 °C (Nugent *et al.*, 1996), inducing a fitness defect.

The experimental data used are freely available at `http://research.ncl.ac.uk/colonyzer/AddinallQFA/`. Addinall *et al.* (2011) present a list of interaction strengths and p-values for significance of interaction, together with a fitness plot for this experiment. We will compare lists of genes classified as interacting with *cdc13-1* by the non-hierarchical frequentist approach presented by Addinall *et al.* (2011) and the hierarchical REM with those classified as interacting by our hierarchical Bayesian approaches.

51

4,294 non-essential *orf∆*s were selected from the yeast deletion collection and used to build the corresponding double deletion query and control strains. Independent replicate culture growth curves (time course observations of cell density) were captured for each query and control strain. The median and range for the number of replicates per *orf∆* is 8 and $[8, 144]$ respectively. There are 66 *orf∆* strains that have greater than 8 replicates (for both the control and query screen). More replicates have been tested for this subset of *orf∆*s as a quality control measure to check if 8 replicates are sufficient to generate a stable fitness summary for each *orf∆*. *orf∆*s with high replicate number include a small number of mutations whose phenotypes are well understood in a telomere-defective background, together with some controls and a range of mutations randomly selected from the deletion library. Including genotypes with well characterised phenotypes allows us to leverage expert, domain-specific knowledge to assess the quality of experimental results. The modelling approaches considered can accommodate different numbers of replicates for each *orf∆*, therefore we don't expect systematic bias from the number of repeats. The range for the number of time points for growth curves captured in the control experiment is $[7, 22]$ and $[9, 15]$ in the query experiment. Raw *cdc13-1* 27°C time series data is given in Figure A.1, for example.

As in the Addinall *et al.* (2011) analysis, a list of 159 genes are stripped from our final list of genes for biological and experimental reasons. Prior hyper-parameters for the models used throughout this chapter are provided in Table B.1. Although our priors are informed by frequentist estimates of historical QFA data sets, we ensure our priors are sufficiently diffuse that all plausible parameter values are well represented and that any given QFA data set can be fit appropriately.

The Heidelberg-Welch (Heidelberger & Welch, 1981) and Raftery-Lewis (Raftery & Lewis, 1995) convergence diagnostics are used to determine whether convergence has been reached for all parameters. Posterior and prior densities are compared by eye to ensure that sample posterior distributions are not restricted by the choice of prior distribution. ACF (auto-correlation) plot diagnostics are checked visually to ensure that serial correlation between sample values of the posterior distribution is low, ensuring that the effective sample size is similar to the actual sample size.

To assess how well the logistic growth model describes cell density observations we generate plots of raw data with fitted curves overlaid. Figures 4.1A, 4.1B and 4.1C show time series data for three different mutant strain repeats at 27°C, together with fitted logistic curves. We can see that each *orf∆* curve fit well represents the repeat level estimates as each *orf∆* level (red) curve lies in the region where most repeat level (black)

curves are found. Sharing information between *orf*Δs will also affect each *orf*Δ curve fit, increasing the probability of the *orf*Δ level parameters being closer to the population parameters. Comparing Figures 4.1A, 4.1B and 4.1C shows that the separate hierarchical model (SHM) captures heterogeneity at both the repeat and *orf*Δ levels.

Figure 4.1D demonstrates the hierarchy of information about the logistic model parameter K generated by the SHM for the *rad50*Δ control mutant strain (variation decreases going from population level down to repeat level). Figure 4.1D also shows that the posterior distribution for K is much more peaked than the prior, demonstrating that we have learned about the distribution of both the population and *orf*Δ parameters. Learning more about the repeat level parameters reduces the variance of our *orf*Δ level estimates. The posterior for the first time-course repeat K_{clm} parameter shows exactly how much uncertainty there is for this particular repeat in terms of carrying capacity K.
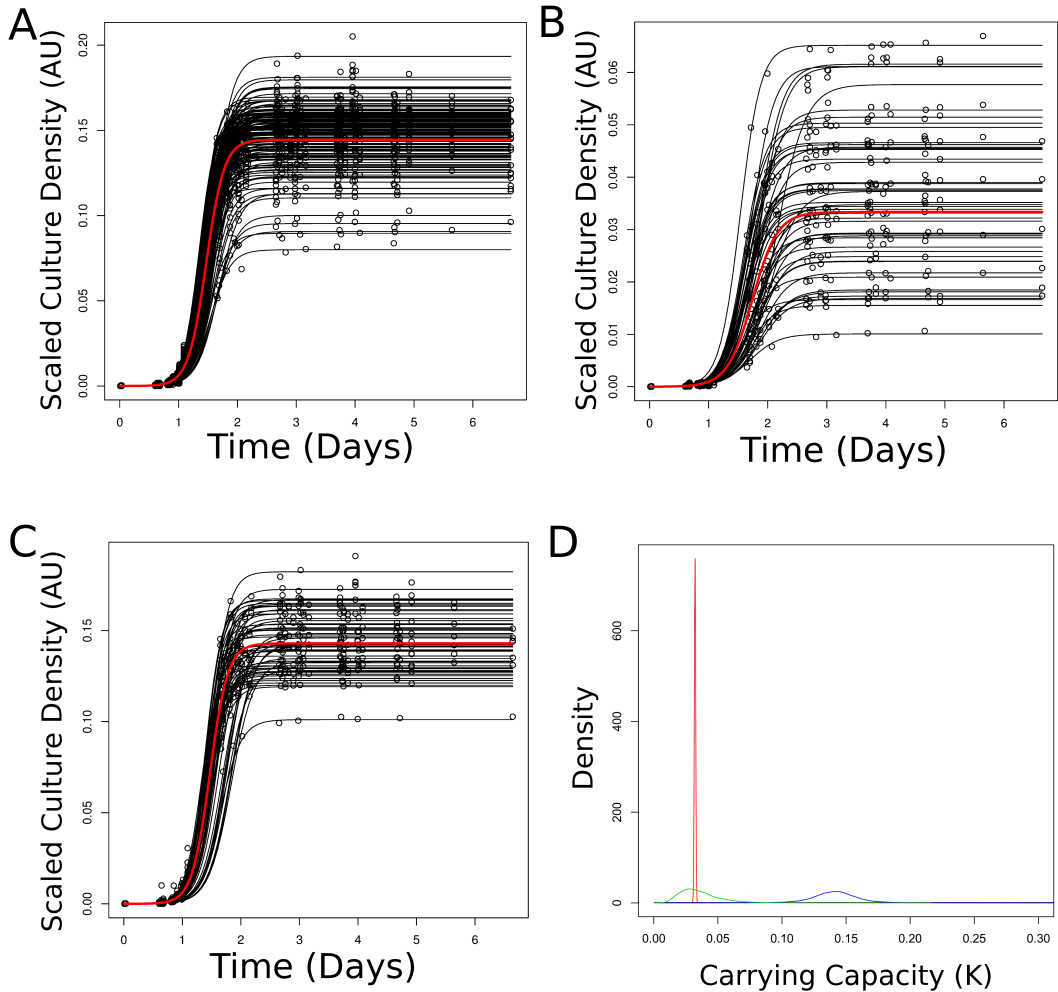
Figure 4.1: Separate hierarchical model (SHM) logistic growth curve fits. Data for *orfΔ* repeats have been plotted in A, B and C, with SHM fitted curves overlaid in black for repeat level parameters and red for the *orfΔ* level parameter fit. A) SHM scatter plot for 144 *his3Δ ura3Δ* repeats at 27°C. B) SHM scatter plot for 48 *rad50Δ ura3Δ* repeats at 27°C. C) SHM scatter plot for 56 *exo1Δ ura3Δ* repeats at 27°C. D) SHM density plot of posterior predictive distributions for *rad50Δ ura3Δ* carrying capacity K hierarchy. The prior distribution for K^p is in black. The posterior predictive for $e^{K_l^o}$ is in blue and for K_{clm} in green. The posterior distribution of the first time-course repeat K_{clm} parameter is in red. Parameters K^p, $e^{K_l^o}$ and K_{clm} are on the same scale as the observed data.

4.2.1. Frequentist approach

Figure 4.2A is a $MDR \times MDP$ fitness plot from Addinall *et al.* (2011) where growth curves and evidence for genetic interaction are modelled using the non-hierarchical frequentist methodology discussed in Section 1.2.2. Figure 4.2B is a $MDR \times MDP$ fitness

plot for the frequentist hierarchical approach REM, described in Table 3.4, applied to the logistic growth parameter estimates used in Addinall *et al.* (2011). The number of genes identified as interacting with *cdc13-1* by Addinall *et al.* (2011) and by the REM are 715 and 315 respectively (Table 4.1). The REM has highlighted many strains which have low fitness. In order to fit a linear model to the fitness data and interpret results in terms of the multiplicative model we apply a log transformation to the fitnesses, thereby affecting the distribution of *orfΔ* level variation.

The REM accounts for between subject variation and allows for the estimation of a query mutation and *orfΔ* effect to be made simultaneously, unlike the model presented by Addinall *et al.* (2011). Due to the limitations of the frequentist hierarchical modelling framework, the REM model assumes equal variances for all *orfΔ*s and incorrectly describes *orfΔ* level variation as Log-normal, assumptions that are not necessary in our new Bayesian approaches.
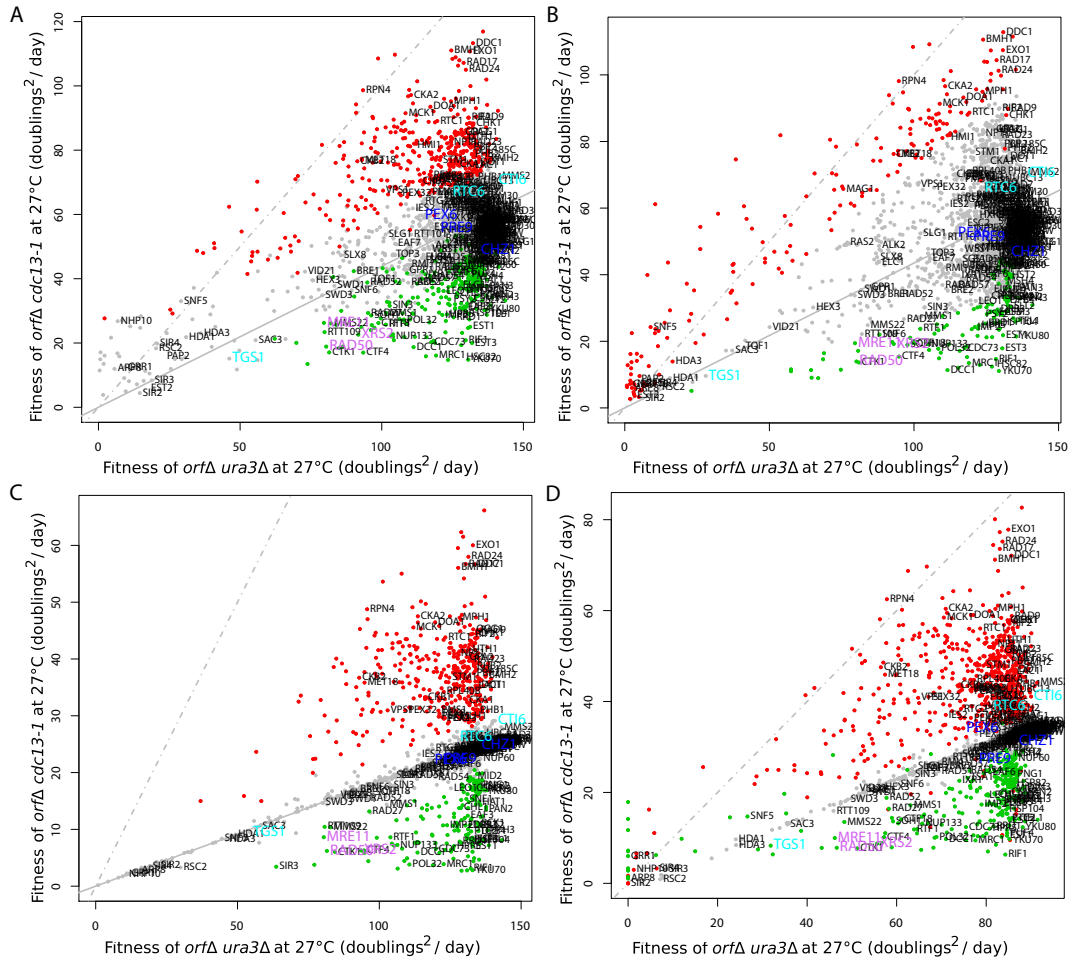
Figure 4.2: Fitness plots with $orf\Delta$ posterior mean fitnesses. Mean $orf\Delta$ level fitness are plotted for the control strains against the corresponding query strains. $orf\Delta$s with significant evidence of interaction are highlighted in red and green for suppressors and enhancers respectively. A) Non-Bayesian, non-hierarchical fitness plot, based on Table S6 from Addinall *et al.* (2011) ($F = MDR \times MDP$). B) Non-Bayesian, hierarchical fitness plot, from fitting the REM to data in Table S6 from Addinall *et al.* (2011) ($F = MDR \times MDP$). C) IHM fitness plot with $orf\Delta$ posterior mean fitness. $orf\Delta$s with significant evidence of interaction are highlighted on the plot as red and green for suppressors and enhancers respectively ($F = MDR \times MDP$). D) JHM fitness plot with $orf\Delta$ posterior mean fitnesses. $orf\Delta$ strains for the JHM plot are classified as being a suppressor or enhancer based on analysis of growth parameter r, meaning occasionally strains can be more fit in the query experiment in terms of $MDR \times MDP$ but be classified as enhancers (green). For panels A and B significant interactors are classified as those with FDR corrected p-values < 0.05. For panels C and D significant interactors have posterior probability $\Delta > 0.5$. To compare fitness plots, labelled genes are those belonging to the following GO terms in Table 4.1: "telomere maintenance", "ageing", "response to DNA damage stimulus" or "peroxisomal organization", as well as the genes identified as interactions only in K with the JHM (see Figure 4.3) (blue), genes interacting only in r with the JHM (cyan) and the MRX complex genes (pink). Solid and dashed grey fitted lines are for the 1-1 line and linear model fits respectively. Alternative fitness plots with each of the GO terms highlighted are given in Section B.2 of the Appendix.

Table 4.1: Number of genes interacting with *cdc13-1* at 27°C identified using each of four approaches: Add (Addinall *et al.*, 2011), REM, IHM and JHM. Number of genes annotated with four example GO terms (telomere maintenance, ageing, response to DNA damage stimulus and peroxisome organisation) are also listed. For the Addinall *et al.* (2011) and REM approach, significant interactors are classified as those with FDR corrected p-values (q-values) < 0.05. The label "half data" denotes analyses where only half of the available experimental observations are used. The JHM uses a $MDR \times MDP$ summary after model fitting to classify suppressors and enhancers, comparable with the other three approaches. The full lists of GO terms for each approach considered are given in a spreadsheet document, freely available online at `http://research.ncl.ac.uk/qfa/HeydariQFABayes/`.

Method	Suppressors	Enhancers	Hits	Suppressors (half data)	Enhancers (half data)	Hits (half data)	Telomere maintenance (N=33)	p value	q value	Aging (N=58)	p value	q value	Response to DNA damage stimulus (N=58)	p value	q value	Peroxisome organization (N=180)	p value	q value
Add	419	296	715	263	192	455	18	1.52E-06	0.0376	16	4.32E-05	0.1863	69	9.28E-12	8.14E-10	13	0.225	0.468
REM	184	131	315	103	86	189	11	2.37E-05	0.0136	10	0.0004	0.0824	49	7.40E-16	1.73E-13	3	0.855	0.914
IHM	404	172	576	252	113	365	14	6.57E-05	0.0051	16	0.0015	0.0445	55	4.60E-09	3.41E-07	10	0.318	0.524
JHM	665	274	939	475	177	601	18	8.22E-05	0.0155	21	0.0015	0.0986	76	3.52E-09	1.99E-07	24	0.002	0.019

4.2.2. Two stage Bayesian approach

Figure 4.2C is an interaction hierarchical model (IHM) fitness plot with $orf\Delta$ level fitness measures generated using the new Bayesian two-stage methodology with fitness in terms of $MDR \times MDP$. 576 genes are identified by the IHM as genetic interactions (Table 4.1). Logistic parameter posterior means are used to generate fitness measures. For a gene (l) from the gene deletion library, (e^{Z_l}) is the fitness for the control and $(e^{\alpha_1 + Z_l + \delta_l \gamma_{c,l}})$ for the query in the IHM. For a gene (l) in the query screen, with no evidence of genetic interaction i.e. $\delta_l = 0$, fitness will be a linear transformation from the control counterpart $(e^{\alpha_1 + Z_l})$. Similar to Figures 4.2A and 4.2B, Figure 4.2C shows how the majority of control strains are more fit than their query strain counterparts, with a mean fitted line lying below the line of equal fitness. Comparing the fitted lines in Figures 4.2A and 4.2B with Figure 4.2C, the IHM shows the largest deviation between the fitted line and the line of equal fitness, is largely due to the difference in P estimated with the SHM for the control and query data sets being scaled out by the parameter α_1. If we fix P in our Bayesian models, similar to the frequentist approach, genetic interactions identified are largely the same, but we then have the problem of choosing P. We recommend estimating P simultaneously with the other model parameters because if the choice of P is not close to the true value, growth rate r estimates must compensate and don't give accurate estimates for

time courses with low carrying capacity K.

It can be seen that many of the interacting *orfΔ*s have large deviations from the genetic independence line. This is because of the indicator variable in the model, used to describe genetic interaction. When there is enough evidence for interaction the Bernoulli variable is set to 1, otherwise it is set to 0. It is interesting to note that non-significant *orfΔ*s, marked by grey points, lie amongst some of the significant strains. Many such points have high variance and therefore we are less confident that these interact with the query mutation. This feature of our new approach is an improvement over that presented in Addinall *et al.* (2011), which always shows evidence for an epistatic effect when mean distance from the genetic independence line is large, regardless of strain fitness variability. An extract from the list of top interactions identified by the IHM is included in Table B.2.

4.2.3. One stage Bayesian approach

Figure 4.2D is a JHM $MDR \times MDP$ fitness plot using the new, unified Bayesian methodology. The $MDR \times MDP$ fitness plot given in Figure 4.2D is for visualisation and comparison with the $MDR \times MDP$ fitness plots of the other approaches considered: the JHM does not make use of a fitness measure. 939 genes are identified by the JHM as genetic interactions (Table 4.1). Posterior means of model parameters are used to obtain the following fitness measures. With the JHM we can obtain an *orf Δ* level estimate of the carrying capacity and growth rate (K, r) for a gene (l). For a gene (l) from the gene deletion library, carrying capacity and growth rate $(e^{K_l^o}, e^{r_l^o})$ are used to evaluate the fitness for the control and $(e^{\alpha_1 + K_l^o + \delta_l \gamma_{c,l}}, e^{\beta_1 + r_l^o + \delta_l \omega_{c,l}})$ for the query. For a gene (l) in the query screen, with no evidence of genetic interaction i.e. $\delta_l = 0$, carrying capacity and growth rate will be linear transformations from the control counterpart $(e^{\alpha_1 + K_l^o}, e^{\beta_1 + r_l^o})$.

Instead of producing a fitness plot in terms of $MDR \times MDP$, it can also be useful to analyse carrying capacity K and growth rate r fitness plots as, in the JHM, evidence for genetic interaction comes from both of these parameters simultaneously, see Figures B.5 and B.6. Fitness plots in terms of logistic growth parameters are useful for identifying some unusual characteristics of *orfΔ*s. For example, an *orfΔ* may be defined as a suppressor in terms of K but an enhancer in terms of r. To enable direct comparison with the Addinall *et al.* (2011) analyses we generated a $MDR \times MDP$ fitness plot, Figure 4.2D. An extract from the list of top interactions identified by the JHM is included in Table B.3.

Table 4.2: Genes interacting with *cdc13-1* at **27°C** and GO terms over-represented in the list of interactions according to each approach A) Number of genes identified for each approach (Add Addinall *et al.* (2011), REM, IHM and JHM) and the overlap between the approaches. 4135 genes from the *S. cerevisiae* single deletion library tested overall. B) Number of GO terms identified for each approach (Add Addinall *et al.* (2011), REM, IHM and JHM) and the overlap between the approaches. 6107 *S. cerevisiae* GO Terms available.

A.		*REM:0*		*REM:1*		**B.**		*REM:0*		*REM:1*	
		Add:0	*Add:1*	*Add:0*	*Add:1*			*Add:0*	*Add:1*	*Add:0*	*Add:1*
IHM:0	*JHM:0*	3097	54	31	10	*IHM:0*	*JHM:0*	5813	21	58	7
	JHM:1	231	78	29	29		*JHM:1*	46	8	6	10
IHM:1	*JHM:0*	1	2	1	0	*IHM:1*	*JHM:0*	20	15	3	12
	JHM:1	30	327	0	215		*JHM:1*	13	54	2	147

4.3. Comparison with previous analysis

4.3.1. Significant genetic interactions

Of the genes identified as interacting with *cdc13-1* (1038, see Table 4.2A) some are identified consistently across all four approaches (215 out of 1038, see Table 4.2A). Of the hits identified by the JHM (939), the majority (639) are common with those in the previously published Addinall *et al.* (2011) approach. However, 231 of 939 are uniquely identified by the JHM and could be subtle interactions which are the result of previously unknown biological processes.

To examine the evidence for some interactions uniquely identified by the JHM in more detail we compared the growth curves for three examples from the group of interactions identified only by the JHM. These examples (*chz1Δ*, *pre9Δ* and *pex6Δ*) are genetic interactions which can be identified in terms of carrying capacity K, but not in terms of growth rate r (see Figure 4.3). By observing the difference between the fitted growth curve (red) and the expected growth curve, given no interaction (green) in Figure 4.3A, 4.3B and 4.3C we test for genetic interaction. Since the expected growth curves in the absence of genetic interaction are not representative of either the data or the fitted curves on the repeat and *orf Δ* level, there is evidence for genetic interaction.

We chose a prior for the probability p of a gene interacting with the background mutation as 0.05. We therefore expected to find 215 genes interacting. The Bayesian models, for which a prior is applicable (IHM and JHM), find more genes than expected (576 and 939 interactions respectively, Table 4.1), demonstrating that information in this dataset can overcome prior expectations. The JHM identifies the highest proportion of genes as hits out of all methods considered, particularly identifying suppressors of *cdc13-1* (Table 4.1). In fact, the JHM identifies more hits than the Addinall *et al.* (2011) approach,
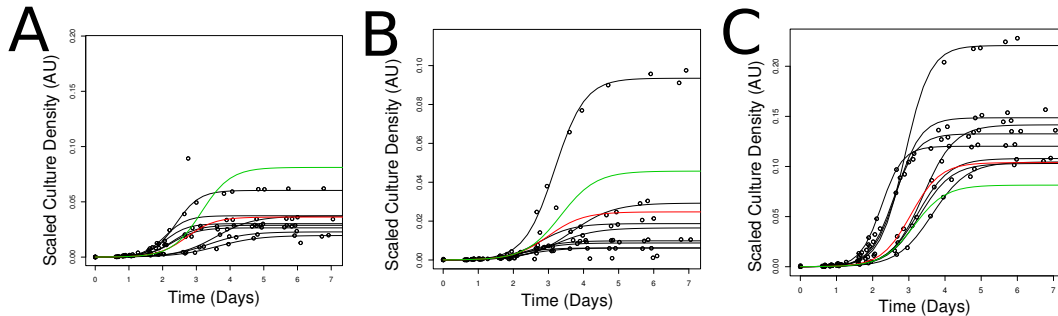
Figure 4.3: Joint hierarchical model (JHM) logistic growth curve fitting. JHM data for *orf*Δ repeats have been plotted in A, B and C, with fitted curves overlaid in black for repeat level parameters, red for the *orf*Δ level query parameter fit and green for the expected *orf*Δ level query parameter fit with no genetic interaction. A) JHM scatter plot for 8 *chz1Δ cdc13-1* repeats. B) JHM scatter plot for 8 *pre9Δ cdc13-1* repeats. C) JHM scatter plot for 8 *pex6Δ cdc13-1* repeats.

even when constrained to using only half of the available data. An important advantage to our new Bayesian approach is that we no longer have the difficulty of choosing a q-value threshold. For the Addinall *et al.* (2011) approach to have similar numbers of interactions to the JHM, a less stringent q-value threshold would have to be justified *a posteriori* by the experimenter.

4.3.2. Previously known genetic interactions

In order to compare the quality of our new, Bayesian hierarchical models with existing, frequentist alternatives, we examined the lists of genetic interactions identified by all the methods discussed and presented here. Comparing results with expected or previously known lists of interactions from the relevant literature, we find that genes coding for the MRX complex (*MRE11*, *XRS2* & *RAD50*), which are known to interact with *cdc13-1* (Foster *et al.*, 2006), are identified by all four approaches considered and can be seen in a similar position in all four fitness plots (Figure 4.2A, 4.2B, 4.2C and 4.2D).

By observing the genes labelled in Figure 4.2A and 4.2B we can see that the frequentist approaches are unable to identify many of the interesting genes identified by the JHM as these methods are unable to detect interactions for genes close to the genetic independence line. The JHM has extracted more information from deletion strain fitnesses observed with high variability than the Addinall *et al.* (2011) approach by sharing more information between levels, consequently improving our ability to identify interactions for genes close to the line of genetic independence (subtle interactions). *CTI6*, *RTC6* and *TGS1* are three examples of subtle interactors identified only by the JHM (interaction in terms of r but not K) which all have previously known telomere-related functions (Franke

et al., 2008; Keogh *et al.*, 2005; Addinall *et al.*, 2008).

We tested the biological relevance of results from the various approaches by carrying out unbiased Gene Ontology (GO) term enrichment analyses on the hits (lists of genes classified as having a significant interaction with *cdc13-1*) using the bioconductoR package GOstats (Falcon & Gentleman, 2007). For the GO term enrichment analysis R code used, see Section B.5 of the Appendix.

All methods identify a large proportion of the genes in the yeast genome annotated with the GO terms "telomere maintenance" and "response to DNA damage stimulus" (see Table 4.1), which were the targets of the original screen, demonstrating that they all correctly identify previously known hits of biological relevance. Interestingly, the JHM identifies many more genes annotated with the "ageing" GO term, which we also expect to be related to telomere biology (though the role of telomeres in ageing remains controversial) suggesting that the JHM is identifying novel, relevant interactions not previously identified by the Addinall *et al.* (2011) screen (see Table 4.1). Similarly, the JHM identifies a much larger proportion of the PEX "peroxisomal" complex (included in GO term: "peroxisome organisation") as interacting with *cdc13-1* (see Table 4.1) including all of those identified in Addinall *et al.* (2011). Many of the PEX genes show large variation in both K and r, an example can be seen in Figure 4.3C for *pex6∆*. Members of the PEX complex cluster tightly, above the fitted line in the fitness plot Figure 4.2D (fitness plots with highlighted genes for GO terms in Table 4.1 are given in Section B.2 of the Appendix), demonstrating that although these functionally related genes are not strong interactors, they do behave consistently with each other, suggesting that the interactions are real. The results of tests for significant over-representation of all GO terms are given in a spreadsheet document, freely available online at `http://research.ncl.ac.uk/qfa/HeydariQFABayes/`.

Overall, within the genes interacting with *cdc13-1* identified by the Addinall *et al.* (2011), REM, IHM and JHM approaches, 274, 245, 266 and 286 GO terms were significantly over-represented respectively (out of 6235 possible GO terms, see Table 4.2B). 147 were common to all approaches and examples from the group of GO terms over-represented in the JHM analysis and not in the Addinall *et al.* (2011) analysis seem internally consistent (e.g. "peroxisome organisation" GO term) and consistent with the biological target of the screen, telomere biology (significant GO terms for genes identified only by the JHM are also included in the spreadsheet document).

Extracts from the list of top interactions identified by both the IHM and JHM are provided in Section B.3. Files including the full lists of genetic interactions for the

IHM and JHM are freely available online at `http://research.ncl.ac.uk/qfa/`
`HeydariQFABayes/`. Alternative fitness plots to Figure 4.2A, B, C & D with gene
labels for those showing significant evidence of genetic interaction are provided in Figure 1.4 and Section B.7. As suppressors and enhancers in the JHM may be in terms of both
K and r, fitness plots in terms of K and r with gene labels for those showing significant
evidence of genetic interaction are given in Figure B.10 and Figure B.11 respectively.

To further compare the similarity of the Bayesian hierarchical models and frequentist
analysis, a table of Spearman's rank correlation coefficients (Spearman, 1987) between
genetic strengths and a $MDR \times MDP$ correlation plot of the JHM versus the Addinall
et al. (2011) are given in Section B.8 of the Appendix.

4.3.3. Hierarchy and model parameters

The hierarchical structure and model choices included in the Bayesian JHM and IHM
are derived from the known experimental structure of QFA. Different levels of variation
for different *orf∆*s are expected and can be observed by comparing distributions of frequentist estimates or by visual inspection of yeast culture images. The direct relationship
between experimental and model structure, together with the richness of detail and number of replicates included in QFA experimental design, reassures us that overfitting is not
an issue in this analysis. For the *ura3∆* 27°C and *cdc13-1* 27°C experiment with 4294
*orf∆*s there are 1.25 times the number of parameters in the JHM (∼200,000) compared to
the two stage REM approach (∼160,000) but when compared to the large number of pairs
of data points (∼830,000) there are sufficient degrees of freedom to justify our proposed
Bayesian models.

4.3.4. Computing requirements

Our Bayesian hierarchical models require significant computational time. As expected,
the mixing of chains in our models is weakest at population level parameters such as
K_p and α_c. For the *ura3∆* 27°C and *cdc13-1* 27°C dataset, the JHM takes ∼2 weeks
to converge and produce a sufficiently large sample. The two stage Bayesian approach
takes one week (with the IHM part taking ∼1 day), whereas the REM takes ∼3 days and
the Addinall *et al.* (2011) approach takes ∼3 hours. A QFA experiment can take over a
month from start to finish and so analysis time is acceptable in comparison to the time
taken for the creation of the data set but still a notable inconvenience. We expect that
with further research effort, computational time can be decreased by using an improved
inference scheme and that inference for the JHM could be completed in less than a week

without parallelisation. MCMC algorithms are inherently sequential so, parallelisation is not completely trivial and may be considered for future development. Parallelisation may reduce computational time by partitioning the state space into segments that can be updated in parallel (Rosenthal, 2000). For the JHM it may be possible to partition by QFA screens to reduce computational time. Further, parallelisation may be possible across $orf\Delta$s for even further reduction to computational time.

4.3.5. Convergence diagnostics

Evidence of convergence for our Bayesian models in Section 4.2.2 and 4.2.3 can be shown by observing posterior samples from the MCMC samplers used. Figures 4.4, 4.5 and 4.6 show evidence of convergence for a subset of population level parameters from the SHM, IHM and JHM respectively. Posterior samples of 1000 particles are obtained after a burn-in period of 800k and a thinning of every 100 observations for the SHM, IHM and JHM.

Population level parameters are found to have the worst mixing in our models due to the large number of lower level parameters that population level parameter sampling distributions are conditioned upon. We demonstrate how our population parameters have converged with Trace plots, ACF and density plots in Figures 4.4, 4.5 and 4.6. Trace plots show that the posterior samples are bound between a fixed range of values, indicating convergence. Auto-correlation functions do not have any large peaks above the dashed blue line for significant evidence of dependence, showing that each sequential sample value from the posterior distributions are largely uncorrelated with previous values and ensuring that the effective sample size is similar to the actual sample size. ACF plots in Figures 4.5 and 4.6 do show some dependence within our posterior samples but as the ACF decays rapidly before a lag of 5, there is only a small amount that will not be a problem for inference. Density plots show that that there is enough information within the models to give sufficiently peaked single modes, converging around a fixed region of plausible values.

Table 4.3 gives diagnostic statistics for the population parameters considered in Figures 4.4, 4.5 and 4.6. We can see in Table 4.3 that the lowest effective sample size of our model parameters is 324, for the JHM P parameter, followed by 378 for the SHM P parameter. Of all our model parameters, P was found to have the lowest effective sample size, but we are still able to find a large enough sample for our inference. Heidelberg and Welch P-values do not show evidence against the stationary of our chains, using a cut-off of 0.10. The above statistics are calculated for all model parameters and are used to identify where mixing is poor and if our model has reached convergence. All chains are

Table 4.3: Bayesian model convergence statistics for the two-stage approach in Section 4.2.2 and one-stage approach in Section 4.2.3. Heidelberg and Welch P-values and the effective sample size have been calculated for a subset of population level parameters.

Model	Parameter	Effective sample size	Heidelberg and Welch P-value
SHM	K_p	521	0.49
	r_p	441	0.11
	P	378	0.56
	ν_p	1000	0.17
IHM	Z_p	677	0.35
	σ_z	430	0.14
	ν_p	1000	0.46
	α_c	914	0.59
JHM	K_p	473	0.72
	r_p	566	0.12
	P	324	0.12
	ν_p	1000	0.13
	α	407	0.36
	β	808	0.67

accepted for parameter posterior samples in Section 4.2.2 and 4.2.3 as effective sample sizes are found to be greater than 300 and Heidelberg and Welch P-values greater than 0.10 for every chain.
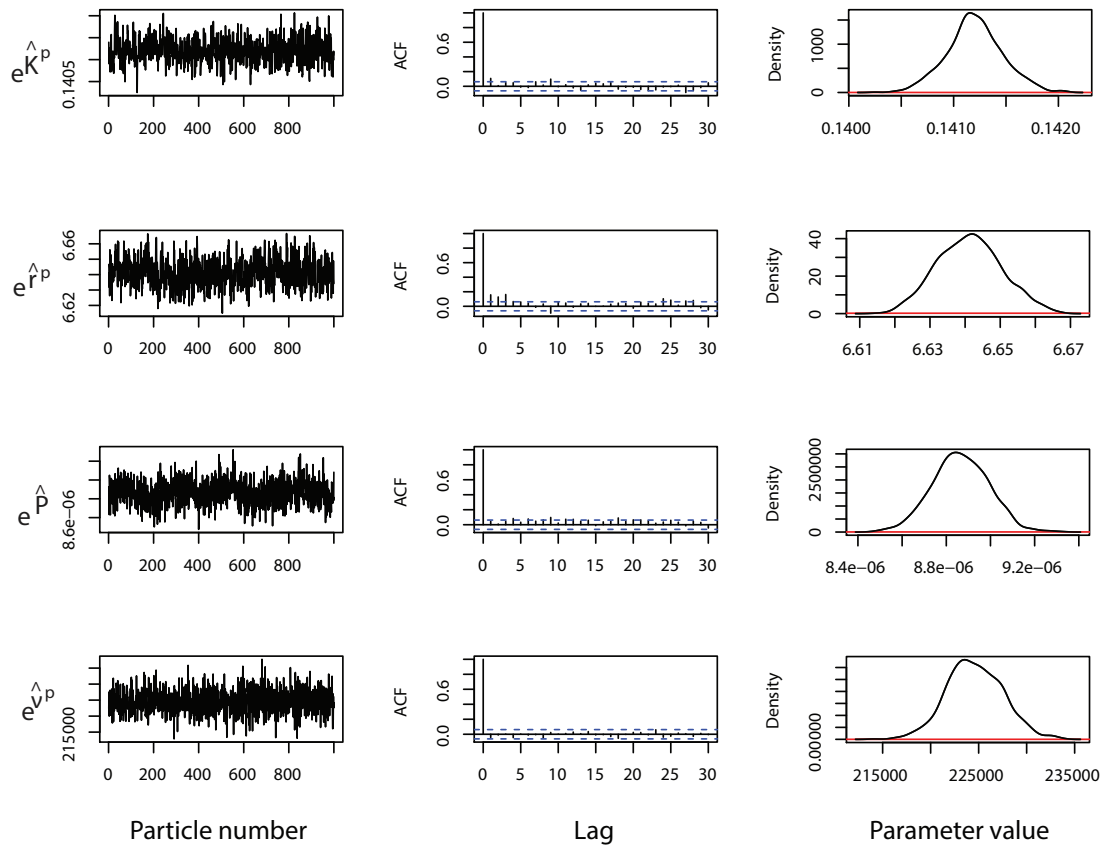
Figure 4.4: Convergence diagnostics for the separate hierarchical model (SHM). Trace, auto-correlation and density plots for the SHM parameter posteriors (sample size = 1000, thinning interval = 100 and burn-in = 800000), see Section 4.2.2. Posterior (black) and prior (red) densities are shown in the right hand column.
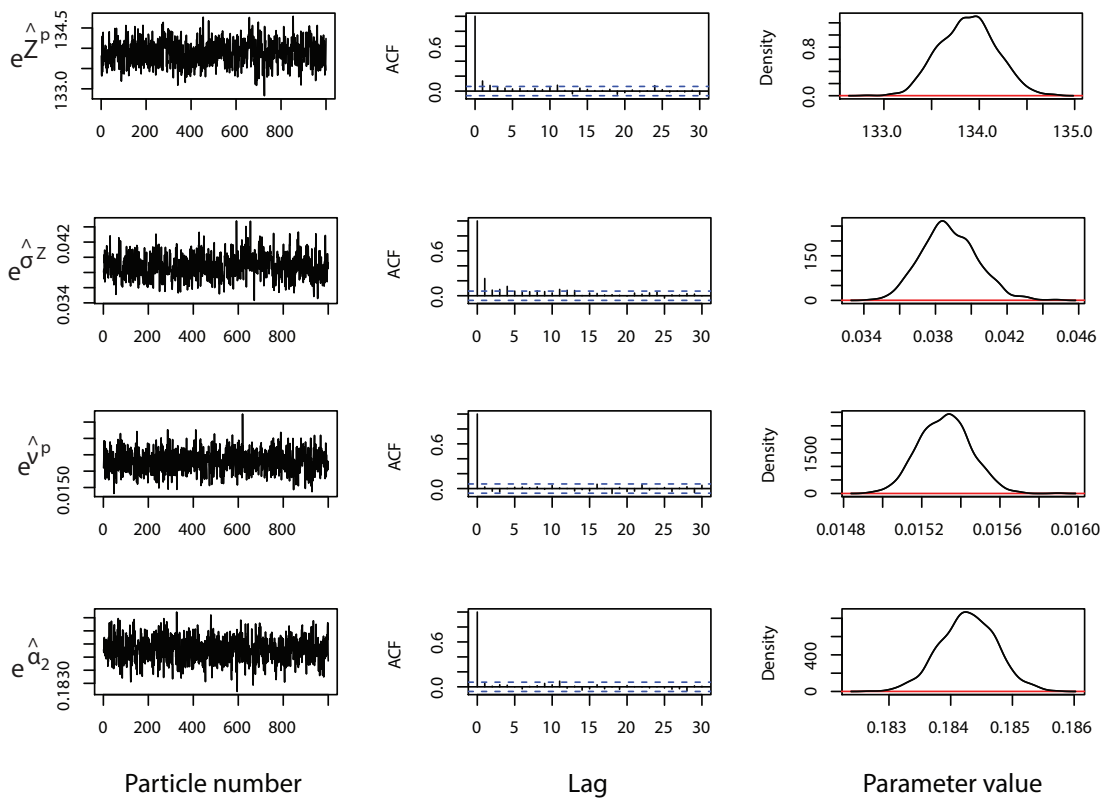
Figure 4.5: Convergence diagnostics for the interaction hierarchical model (IHM). Trace, autocorrelation and density plots for the IHM parameter posteriors (sample size = 1000, thinning interval = 100 and burn-in = 800000), see Section 4.2.2. Posterior (black) and prior (red) densities are shown in the right hand column.
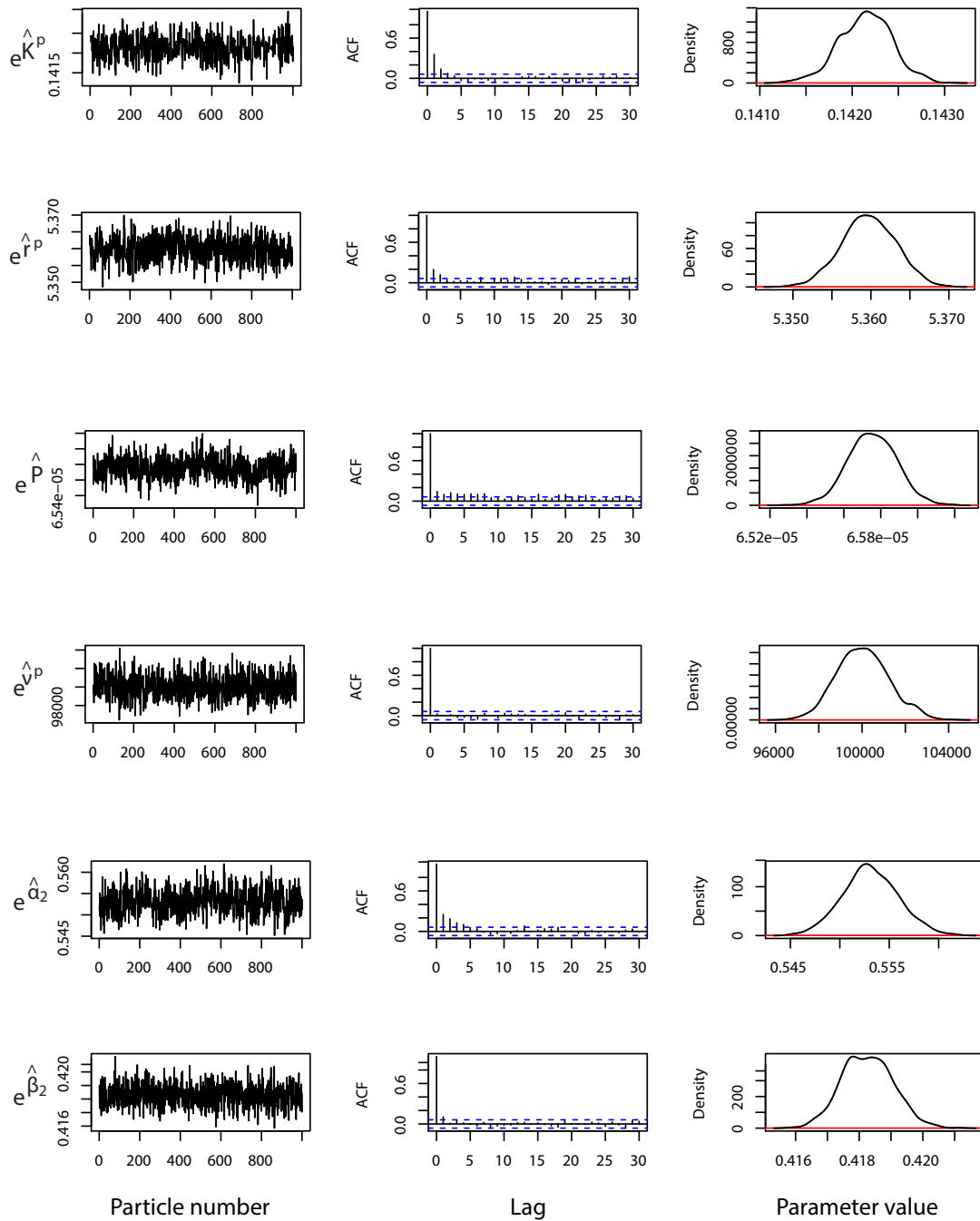
Figure 4.6: Convergence diagnostics for the joint hierarchical model (JHM). Trace, autocorrelation and density plots for the JHM parameter posteriors (sample size = 1000, thinning interval = 100 and burn-in = 800000), see Section 4.2.3. Posterior (black) and prior (red) densities are shown in the right hand column.

67

4.3.6. Simulation study

A simulation study was carried out to compare the performance of the different approaches considered for a simulated QFA screen comparison from the JHM. We believe that the JHM closely models a QFA screen comparison and so by simulating a QFA screen comparison data set from the JHM we will obtain a data set for which we know the full set of true genetic interactions. Simulated JHM data will include important features of QFA screen comparison data, such as a hierarchical structure and genetic interaction in terms of both K and r.

Two simulated QFA screens where generated, a control and query screen with some condition effect in the query. Each screen consists of 4300 *orfΔ*s and 8 logistic growth time-course repeats for each *orfΔ*. Each time-course consists of 10 measurements, evenly distributed across 6 days. 430 genes were set as genetic interactors in the query screen. The true Population level parameters are chosen from frequentist estimates of 10 historic data sets, *orfΔ* and repeat level parameters are then generated from the JHM structure in Table 3.3 and growth time-course data simulated.

Table 4.4 shows the number of true genetic interactions identified, suppressors and enhancers, as well as false positives (FPs) and false negatives (FN) for each of the approaches considered. As expected, the JHM identifies the largest number of true genetic interactions. The number of suppressors identified by the JHM is higher than the Addinall *et al.* (2011), REM and IHM but for enhancers, all methods perform very similarly. Performance of the different methods can be observed through the FP and FN rates. From Table 4.4 we can calculate FP and FN rates, where FP rate$= 1-$"sensitivity" and FN rate$= 1-$"specificity". FP rates for the Addinall *et al.* (2011), REM, IHM and JHM are 0.078, 0.042, 0.006 and 0.002 respectively. The JHM has the lowest FP rate when compared to the other approaches available. Frequentist approaches Addinall *et al.* (2011) and REM have large FP rates when compared to the two Bayesian approaches. The Addinall *et al.* (2011) approach has more false positives than true genetic interactions. FN rates for the Addinall *et al.* (2011), REM, IHM and JHM are 0.488, 0.570, 0.593 and 0.270 respectively. Two-stage approaches Addinall *et al.* (2011), REM and IHM have large FP rates when compared to the JHM. The Addinall *et al.* (2011), REM and IHM have $\sim$200 false negatives, approximately double the number identified by the JHM ($\sim$100). Observing the genes that have been missed by the two-stage approaches, we find that they often fail to identify genetic interactions when evidence is weak in only K or r, even if there is sufficient evidence in the other parameter such that the JHM can identify the genetic interaction.

From our simulation study we have been able to show that the two-stage frequentist approaches have high false positives and false negatives. From the number of false positives identified for each method, we can see that the non-hierarchical Addinall *et al.* (2011) approach has the worst performance, followed by the hierarchical two-stage approaches. As expected, the JHM is the best approach when we consider a simulated hierarchical data set with genetic interaction in terms of K and r, as the two-stage approaches fail to capture more subtle genetic interactions.

Table 4.4: Simulation study with a joint hierarchical model (JHM) simulated dataset. A QFA screen comparison was generated from the JHM and 430 genes are set as genetic interactors, see Section 4.3.6. Applications of the (Addinall *et al.*, 2011), REM, two-stage Bayesian (IHM) and one-stage Bayesian (JHM) approaches are made to the JHM simulated dataset and performance compared. Suppressors and enhancers are defined in terms of $MDR \times MDP$.

Model	True interactions identified (N=430)	True Suppressors (N=274)	True Enhancers (N=156)	False Positives	False Negatives	Sensitivity	Specificity
Addinall *et al.* (2011)	220	158	62	303	210	0.922	0.512
REM	185	100	85	163	245	0.958	0.430
IHM	175	130	45	23	255	0.994	0.407
JHM	314	256	58	8	116	0.998	0.730

4.4. Bayesian inference code comparison

Inference for the Bayesian hierarchical models in this thesis is carried out using code written in the C programming language. To see how our code compares to commonly used software available for carrying out inference for Bayesian models, we have tested posterior samples for our C code and equivalent code using Just Another Gibbs Sampler (JAGS) software (written in C++) (Plummer, 2003) . We carry out our JAGS analysis within the R package "rjags" (Plummer, 2010) which provides a more familiar framework for an R user implementing the JAGS software. The BUGS (Bayesian inference Using Gibbs Sampling) language (Lunn *et al.*, 2000*a*) is used to describe models in JAGS. The SHM, IHM and JHM have each been described with the BUGS language in Section B.6 of the Appendix.

For the following comparison we use a subset from the *cdc13-1* 27°C vs *ura3Δ* 27°C suppressor/enhancer data set described in Section 4.2. A subset of 50 *orf Δ*s (for both the control and query) are chosen, each with 8 time-course repeats. With a smaller data set we are able to collect large posterior sample sizes, sufficient to carry out a comparison between posterior samples. Density plots are used to visually compare the similarity of the posterior samples from the C and JAGS code. The KolmogorovSmirnov test (Huber-Carol, 2002) and unpaired two-sample Student's t-test (Witte & Witte, 2009) are used to

Table 4.5: Unpaired t-test and Kolmagorov-Smirnov p-values comparing posterior samples from the joint hierarchical model (JHM) using both C and Just Another Gibbs Sampler (JAGS) software. An extract of JHM parameters are given for both the C programming language and JAGS software. Posterior means are also included for both approaches. t-tests are carried out on the log posterior samples i.e. $\hat{K}_p$ in place of $e^{\hat{K}_p}$ to assume normality.

Parameter	C Code posterior mean	JAGS posterior mean	t-test (with log posterior samples)	Kolmagorov-Smirnov test
$e^{\hat{K}_p}$	0.143	0.143	0.452	0.401
$e^{\hat{r}_p}$	4.639	4.641	0.424	0.482
$e^{\hat{P}}$	$2.537 \cdot 10^{04}$	$2.517 \cdot 10^{04}$	0.137	0.116
$e^{\hat{\nu}_p}$	$7.402 \cdot 10^{04}$	$7.416 \cdot 10^{04}$	0.250	0.190
$e^{\hat{\alpha}_c}$	0.304	0.304	0.203	0.140
$e^{\hat{\beta}_c}$	0.384	0.384	0.156	0.146

test for significant difference between posterior samples from our C and JAGS code.

A comparison of posterior samples for our most sophisticated model, the JHM, is given below. Posterior samples of 100k particles are obtained after a burn-in period of 1000k and a thinning of every 100 observations for both the C and JAGS code. Computational time for the C and JAGS code is ~ 30 hours and ~ 400 hours respectively. The minimum effective sample size per second (ESS$_{\text{min}}$/sec) for the C and JAGS code is ~ 1 and ~ 0.1 respectively, demonstrating that the C code is $\sim 10 \times$ faster.

Figure 4.7 gives density plots for an extract of JHM parameters for the C and JAGS software. Visually there is no significant difference between the posterior sample density plots in Figure 4.7. Of the parameters shown, the weakest effective sample size (~ 80000ESS) is for the initial inoculum parameter P, but this is sufficiently large enough ESS to test if posterior samples show a significant difference. Table 4.5 demonstrates further that there is no significant difference found between the parameters shown. The unpaired t-test for log posterior samples (for normality assumption) and Kolmogorov-Smirnov test p-values are all greater than 0.10 for the parameters given, including the inoculum density parameter P. Overall we find no significant evidence against the C code and JAGS code sampling from the same posterior distributions.

As carrying out inference using C is ~ 10 times faster than the JAGS equivalent code we prefer the C code for our Bayesian hierarchical models. Obtaining sufficiently sized independent posterior samples of our posterior distributions for a larger data set of ~ 4000 *orf* Δs, we estimate our C code to be at least more than $\sim 50 \times$ faster than the equivalent JAGS as we find the JAGS code to have exponential computational costs as we introduce larger data sets. JAGS is very useful for model exploration as it is fast and simple to describe complex models. The JAGS software is so prohibitively slow for the JHM, that an experimenter is likely to not carry out such inference and use a more simple or faster
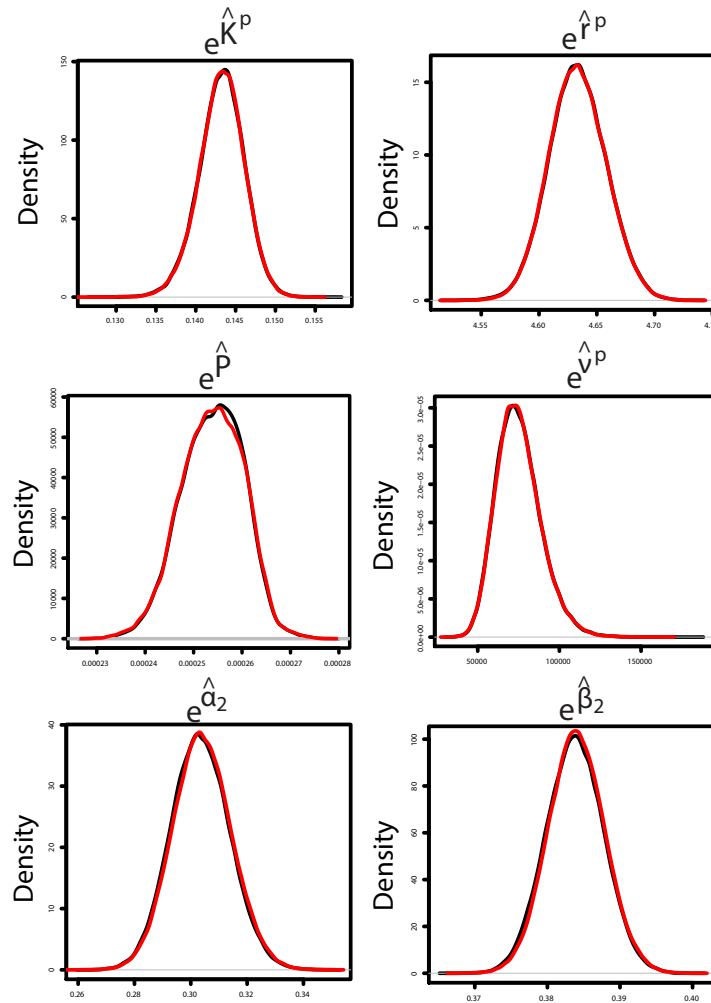
Figure 4.7: Density plots for posterior samples from the joint hierarchical model (JHM) using the C programming language (red) and Just Another Gibbs Sampler (black) software. Density plots for the JHM parameter posteriors (sample size = 100000, thinning interval = 100 and burn-in = 1000000).

method, justifying the use of the C programming language to carry out inference. Further improvements such as the introduction of parallelisation may lead to more favourable computational times in the future.

4.5. Further case studies

In this section we briefly introduce different data sets that may be considered for further investigation and research. We can also see how the JHM performs for different experimental conditions by applying the JHM to different QFA screen comparisons, see $MDR \times MDP$ fitness plots in Figures 4.8-4.11. The data sets used in Figures 4.8-4.11 are currently unpublished from the Lydall lab. For each of the data sets, the JHM in Table 3.3 is applied with the prior hyper-parameters in Table B.1. Posterior samples of 1000 particles are obtained after a burn-in period of 800k, and a thinning of every 100 observations. Similarly to Section 4.3.5, chains from our MCMC sampler are accepted where the effective sample sizes are greater than 300 and Heidelberg and Welch P-values are greater than 0.10 for every chain. As in the Addinall *et al.* (2011) analysis, each experiment has a list of 159 genes stripped from our final list of genes for biological and experimental reasons. Results for the *cdc13-1exo1△* **27°**C vs *cdc13-1* **27°**C and *cdc13-1rad9△* **27°**C vs *cdc13-1* **27°**C experiments have further genes removed for biological and experimental reasons, 23 and 13 genes respectively (a total of 182 and 172 genes respectively).

Figure 4.8 is a *cdc13-1exo1△* **27°**C vs *cdc13-1* **27°**C suppressor/enhancer analysis for finding genes that interact with *exo1* in a telomere maintenance defective background (*cdc13-1* at **27°**C). Similarly, Figure 4.9 is a *cdc13-1rad9△* **27°**C vs *cdc13-1* **27°**C suppressor/enhancer analysis for finding genes that interact with *rad9* in a telomere maintenance defective background. Figure 4.10 is a *yku70△* **37°**C vs *ura3△* **37°**C suppressor/enhancer analysis for finding genes that interact with *yku70* at high temperature. Figure 4.11 is an example of a temperature sensitivity experiment, for finding genes that interact with the high temperature of **37°**C. Figures 4.8-4.11 demonstrate that the JHM can capture different linear relationships that are above or below the 1-1 line. Curvature of the data in Figures 4.8-4.11 suggests that the linear relationships modelled by the JHM may be improved through linearising transformations of the data. Extending the JHM to account for the curvature in the data may improve our model fit and allow to better determine genes which significantly interact.

Table 4.6 compares the number of suppressors and enhancers estimated for each of the experiments considered. The experiments in Table 4.6 have similar numbers of genetic interactions, ranging from 358 to 511, but much lower than the *cdc13-1***27°**C vs *ura3△* **27°**C experiment which has 939. The experiments introduced in this section also differ from the *cdc13-1* **27°**C vs *ura3△* **27°**C experiment as they have more enhancers than suppressors, further demonstrating the JHM's ability to model different experimental

situations and the non-restrictive choice of priors (Table B.1).

Table 4.6: Number of joint hierarchical model (JHM) interactions for QFA datasets given in Section 4.5. Interactions for each dataset is split into suppressors and enhancers. The number of interactions found with the extensions to the joint hierarchical model (see Section 4.6) are also given. Each QFA screen comparison consists of 4294 *orf*△s. Results for all experiments have a list of 159 genes removed from the final list of interactions for biological and experimental reasons. Results for the *cdc13-1exo1*△ **27**°C vs *cdc13-1* **27**°C and *cdc13-1rad9*△ **27**°C vs *cdc13-1* **27**°C experiments have further genes removed for biological and experimental reasons, 23 and 13 genes respectively (a total of 182 and 172 genes respectively).

Query screen	*Control screen*	*Interactions*	*Suppressors*	*Enhancers*
cdc13-1exo1△ **27**°C	*cdc13-1* **27**°C	388	81	307
cdc13-1rad9△ **27**°C	*cdc13-1* **27**°C	358	73	285
yku70△ **37**°C	*ura3*△ **37**°C	511	104	407
ura3△ **37**°C	*ura3*△ **20**°C	460	138	322

Model for *cdc13-1* **27**°C vs *ura3*△ **27**°C experiment	*Interactions*	*Suppressors*	*Enhancers*
JHM	939	665	274
JHM-Batch	553	378	174
JHM-Transformation	901	658	243

Table 4.7A shows the overlap in genes with significant evidence of genetic interactions between the different QFA comparisons considered. The largest number of overlapping genetic interactions are found with the *cdc13-1*△ **27**°C vs *ura*△ **27**°C experiment, overlapping with 301 and 263 genes from the *cdc13-1exo1*△ **27**°C vs *cdc13-1* **27**°C and *cdc13-1rad9*△ **27**°C vs *cdc13-1* **27**°C experiment respectively. The *cdc13-1*△ **27**°C vs *ura*△ **27**°C, *cdc13-1exo1*△ **27**°C vs *cdc13-1* **27**°C and *cdc13-1rad9*△ **27**°C vs *cdc13-1* **27**°C experiments are expected to overlap most as they are designed to find genes interacting in a *cdc13-1* background. The smallest number of overlapping genetic interactions are found with the *ura3*△ **37**°C vs *ura3*△ **20**°C and *yku70*△ **37**°C vs *ura3*△ **37**°C experiment. The *ura3*△ **37**°C vs *ura3*△ **20**°C and *yku70*△ **37**°C vs *ura3*△ **37**°C experiments are expected to have the least overlap as they are not designed to find genes interacting in a *cdc13-1* background. The *yku70*△ **37**°C vs *ura3*△ **37**°C experiment is designed to look at telomeres, but instead of disrupting the telomere capping protein Cdc13 using *cdc13-1*, a *yku70*△ mutation is made such that the protein Yku70 (a telomere binding protein which guides the enzyme telomerase to the telomere (Addinall *et al.*, 2011)) is no longer produced by the cell. Further *ura3*△ **37**°C vs *ura3*△ **20**°C is designed to

investigate temperature sensitivity only.

Table 4.7B shows the overlap in significant GO terms between the different QFA comparisons considered. The largest number of overlapping significant GO terms are found with the *cdc13-1Δ* **27°C** experiment, overlapping with ~150 GO terms for each experiment. The smallest overlap with *cdc13-1Δ* **27°C** vs *uraΔ* **27°C** experiment is 110 GO terms with the *ura3Δ* **37°C** vs *ura3Δ* **20°C** experiment. The smallest number of overlapping genetic interactions are for the *ura3Δ* **37°C** vs *ura3Δ* **20°C** experiment, followed by *yku70Δ* **37°C** vs *ura3Δ* **37°C**, with ~110 and ~120 GO terms overlapping with the other experiments respectively. Similarly to the overlap of genes with significant evidence of genetic interaction, the overlap of significant GO terms shows that our *cdc13-1* background experiments share the most GO terms and that the temperature sensitivity experiment *ura3Δ* **37°C** vs *ura3Δ* **20°C** has the least overlap.

We have shown that the JHM can successfully model different experimental data sets, Figures 4.8-4.11 are included as a reference for further research. Of the different experiments we can see that *cdc13-1* **27°C** vs *ura3Δ* **27°C** is the most dissimilar to the other experiments due to the large number of genetic interactions, 939 in total (see Table 4.6). The next largest number of genetic interactions is 511 with the *yku70Δ* **37°C** vs emphura3Δ **37°C** experiment, which is approximately half the genes found for the *cdc13-1* **27°C** vs *ura3Δ* **27°C** experiment. Tables 4.7A and 4.7B show that the overlap between QFA comparisons is as expected using the JHM, with the closer related experiments sharing the most overlap. To account for the curvature of the data observed in Figures 4.8-4.11 we introduce a JHM with linearising transformations in the next section. Further research may include developing models that can incorporate multiple QFA comparisons to find evidence of genetic interactions between query screens and incorporate more information within our models.

Table 4.7: Overlap between different QFA comparisons for genes interacting and gene ontology terms over-represented in lists of interactions. For a fair comparison, any genes removed from the results of a QFA comparison for biological and experimental reasons are removed for all experiments, therefore results for all experiments have a list of 195 genes (159+23+13, see Table 4.6) removed from the final list of interactions for biological and experimental reasons. A) Number of genes identified for each QFA comparison and the overlap between QFA comparisons. 4099 genes from the *S. cerevisiae* single deletion library are considered. B) Number of GO terms identified for each approach and the overlap between QFA comparisons. 6094 *S. cerevisiae* GO Terms available.

A.	*cdc13-1Δ* **27°C** vs *uraΔ* **27°C**	*cdc13-1exo1Δ* **27°C** vs *cdc13-1* **27°C**	*cdc13-1rad9Δ* **27°C** vs *cdc13-1* **27°C**	*yku70Δ* **37°C** vs *ura3Δ* **37°C**	*ura3Δ* **37°C** vs *ura3Δ* **20°C**
cdc13-1Δ **27°C** vs *uraΔ* **27°C**	926	N/A	N/A	N/A	N/A
cdc13-1exo1Δ **27°C** vs *cdc13-1* **27°C**	301	386	N/A	N/A	N/A
cdc13-1rad9Δ **27°C** vs *cdc13-1* **27°C**	263	245	355	N/A	N/A
yku70Δ **37°C** vs *ura3Δ* **37°C**	252	155	146	506	N/A
ura3Δ **37°C** vs *ura3Δ* **20°C**	223	152	149	164	455

B.	*cdc13-1Δ* **27°C** vs *uraΔ* **27°C**	*cdc13-1exo1Δ* **27°C** vs *cdc13-1* **27°C**	*cdc13-1rad9Δ* **27°C** vs *cdc13-1* **27°C**	*yku70Δ* **37°C** vs *ura3Δ* **37°C**	*ura3Δ* **37°C** vs *ura3Δ* **20°C**
cdc13-1Δ **27°C** vs *uraΔ* **27°C**	282	N/A	N/A	N/A	N/A
cdc13-1exo1Δ **27°C** vs *cdc13-1* **27°C**	142	188	N/A	N/A	N/A
cdc13-1rad9Δ **27°C** vs *cdc13-1* **27°C**	151	130	212	N/A	N/A
yku70Δ **37°C** vs *ura3Δ* **37°C**	150	119	125	245	N/A
ura3Δ **37°C** vs *ura3Δ* **20°C**	110	100	112	119	195

Figure 4.8: *cdc13-1exo1Δ* **27°**C vs *cdc13-1* **27°**C joint hierarchical model (JHM) fitness plot with *orfΔ* posterior mean fitnesses. The JHM does not does not make use of a fitness measure such as $MDR \times MDP$ but the fitness plot is given in terms of $MDR \times MDP$ for comparison with other approaches which do. *orfΔ* strains are classified as being a suppressor or enhancer based on one of the two parameters used to classify genetic interaction, growth parameter r, this means occasionally strains can be more fit in the query experiment in terms of $MDR \times MDP$ but be classified as enhancers (green). Further fitness plot explanation and notation is given in Figure 4.2.

Figure 4.9: *cdc13-1rad9Δ* **27°C** vs *cdc13-1* **27°C** joint hierarchical model (JHM) fitness plot with *orfΔ* posterior mean fitnesses. The JHM does not does not make use of a fitness measure such as $MDR \times MDP$ but the fitness plot is given in terms of $MDR \times MDP$ for comparison with other approaches which do. *orfΔ* strains are classified as being a suppressor or enhancer based on one of the two parameters used to classify genetic interaction, growth parameter r, this means occasionally strains can be more fit in the query experiment in terms of $MDR \times MDP$ but be classified as enhancers (green). Further fitness plot explanation and notation is given in Figure 4.2.
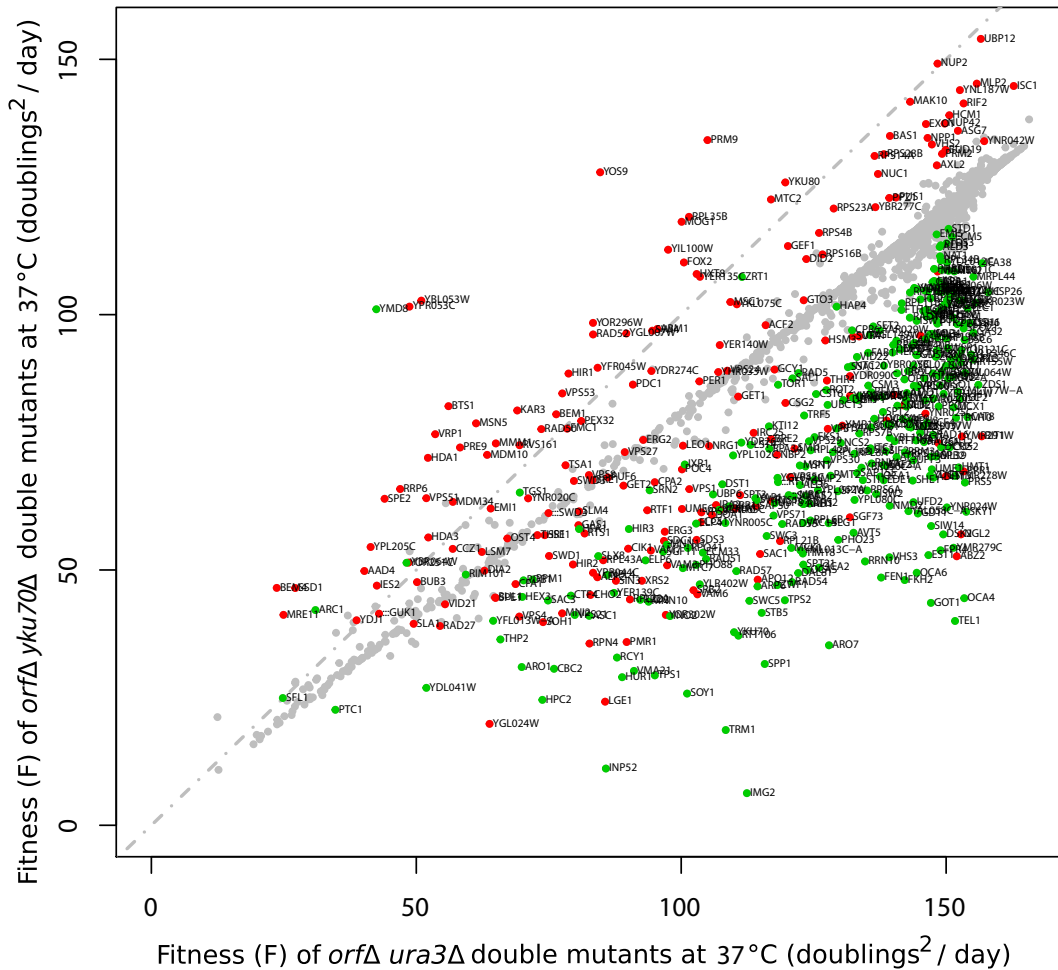
Figure 4.10: *yku70Δ* **37°C** vs *ura3Δ* **37°C** joint hierarchical model (JHM) fitness plot with *orfΔ* posterior mean fitnesses. The JHM does not does not make use of a fitness measure such as $MDR \times MDP$ but the fitness plot is given in terms of $MDR \times MDP$ for comparison with other approaches which do. *orfΔ* strains are classified as being a suppressor or enhancer based on one of the two parameters used to classify genetic interaction, growth parameter r, this means occasionally strains can be more fit in the query experiment in terms of $MDR \times MDP$ but be classified as enhancers (green). Further fitness plot explanation and notation is given in Figure 4.2.
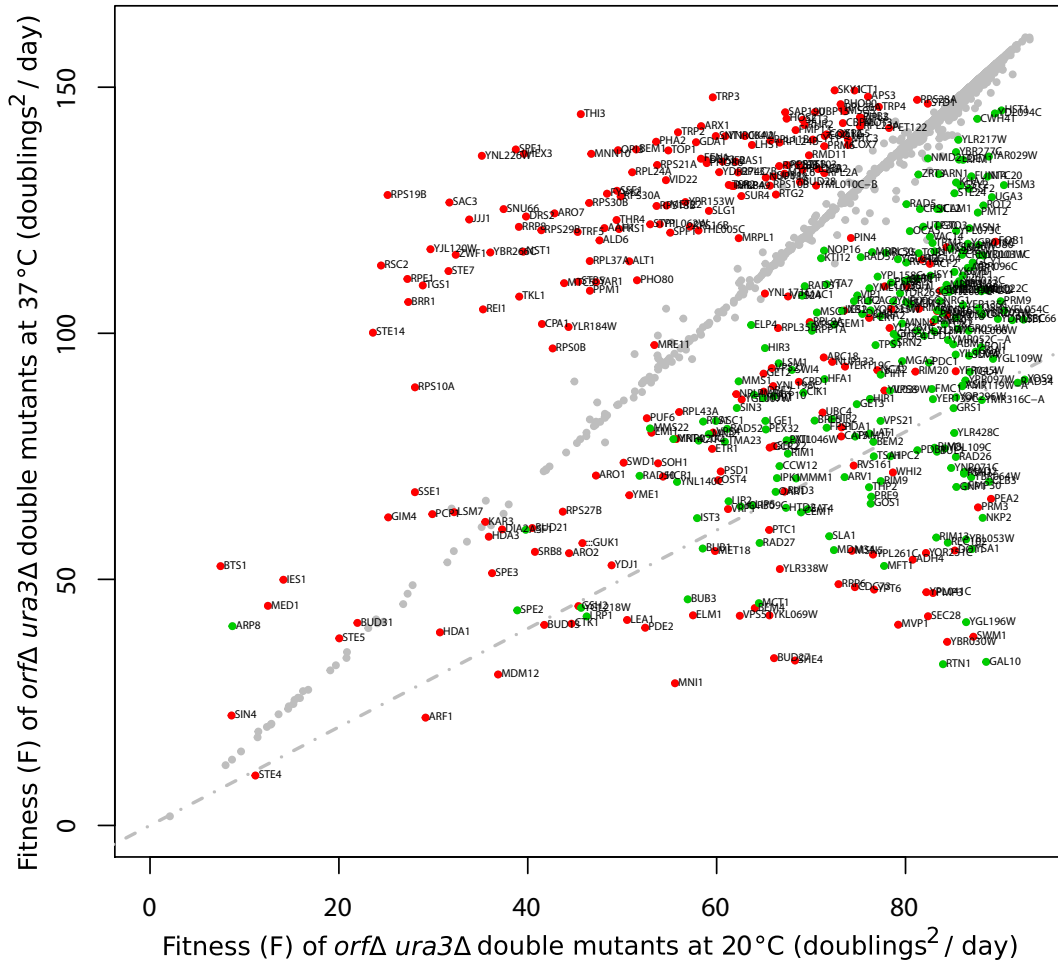
Figure 4.11: *ura3Δ* **37°C** vs *ura3Δ* **20°C** joint hierarchical model (JHM) fitness plot with *orfΔ* posterior mean fitnesses. The JHM does not does not make use of a fitness measure such as $MDR \times MDP$ but the fitness plot is given in terms of $MDR \times MDP$ for comparison with other approaches which do. *orfΔ* strains are classified as being a suppressor or enhancer based on one of the two parameters used to classify genetic interaction, growth parameter r, this means occasionally strains can be more fit in the query experiment in terms of $MDR \times MDP$ but be classified as enhancers (green). Further fitness plot explanation and notation is given in Figure 4.2.

4.6. Extensions of the joint hierarchical model

In this section we briefly introduce two new extensions of the JHM for further investigation and research. An extension to the JHM, given in Table 3.3, is to consider a batch effect. Batch effects are technical sources of variation from the handling of experimental cultures (Leek *et al.*, 2010; Chen *et al.*, 2011). Batch effects can be confounded with the biology of interest, leading to misleading results and conclusions.

A QFA screen comparison is carried out between two QFA screens. Each QFA screen consists of multiple 384 plates grown over time (see Figure 2.3), typically with each *orf Δ* repeat on a different 384 plate. For the *cdc13-1* **27°**C vs *ura3Δ* **27°**C experiment, each QFA screen is built of 120 384 spot plates (240 total unique plates). Each 384 plate is created sequentially and may be created by a different experimenter. The 384 plates may therefore differ due to factors that the experimenters do their best to control such as the amount of nutrition in a plate, temperature, or other environmental effects. Where *orf Δ* repeats are carried out across multiple plates, differences in plates can therefore be captured by introducing a batch effect into the model.

Through careful planning and improved experimental design, batch effects can be reduced or removed. When we are unable to improve our experimental design any further we may be interested in accounting for a batch effect within our model. Introducing parameters to model batch effects in our experiment we can account for any differences between the 240 384 spot plates. A JHM with batch effects (JHM-B), described in Table 4.9, will be able to improve inference by including more of the experimental structure. The model in Table 4.9 introduces a batch effect κ_b and λ_b, for a plate b, to capture any batch effect in carrying capacity K and growth rate r respectively. A batch effect will be estimated within the model and consequently any confounding with *orf Δ* level carrying capacity K and growth rate r parameters will be removed. Using frequentist estimates of the batch effects in the QFA screens, a normal prior was chosen to describe batch effect parameters, allowing either a positive or negative effect to be incorporated for each *orf Δ* repeat in terms of K and r.

Another extension of the JHM is to consider a transformation to linearise the relationship describing genetic independence in the JHM. When carrying out linear regression we may be interested in linearising the data to improve the linear relationship (Kutner *et al.*, 2005). There are many different transformations used for linearising data, the most common are log and power transformations. Power transformations are families of power

functions that are typically used to stabilise variance and make our data more Normal distribution-like. For a variable x, a power function is of the form $f : x \mapsto cx^r$, for $c, r \in \mathbb{R}$, where c and r are constant real numbers. The Box-Cox transformation (Box & Cox, 1964) is a particular case of power transformation that is typically used to transform data and linearise a relationship within a data set.

Without linearising our data, we may not be describing genetic independence within our model correctly, leading to misleading results and conclusions. A JHM with transformations (JHM-T), described in Table 4.10, will be able to improve inference by ensuring a more linear relationship is made between the control and query screen. Genetic independence within the JHM is described as a linear relationship (see Sections 1.2.1 and 3.4.1) for both carrying capacity K and growth rate r. We may not believe there to be a perfectly linear relationship between the control and query for both K and r. Introducing a power transformation for the model of genetic independence in terms of K and r can allow us to linearise the relationship and better model genetic independence. The model in Table 4.10 introduces the transformation parameters ϕ and χ at an $orf\Delta$ level for both the carrying capacity K and growth rate r respectively, where $\phi > 0$ and $\chi > 0$. The "vanilla" JHM assumes an additive model of epistasis with $(\alpha_c + K_l^o + \delta_l\gamma_{cl}, \beta_c + r_l^o + \delta_l\omega_{cl})$, where α_c and β_c are the scale parameters, as we are considering log $orf\Delta$ parameters. The "vanilla" JHM effectively assuming a multiplicative model on the original scale of the data i.e. $(e^{\alpha_c}e^{K_l^o + \delta_l\gamma_{cl}}, e^{\beta_c}e^{r_l^o + \delta_l\omega_{cl}})$. By introducing new parameters ϕ and χ to scale the control and query data $\left(\frac{\alpha_c + K_l^o + \delta_l\gamma_{cl}}{\phi}, \frac{\beta_c + r_l^o + \delta_l\omega_{cl}}{\chi}\right)$ we can expect to have a power transformation with the control and query on the original scale of the data $\left[\left(e^{\alpha_c}e^{K_l^o + \delta_l\gamma_{cl}}\right)^{\frac{1}{\phi}}, \left(e^{\beta_c}e^{r_l^o + \delta_l\omega_{cl}}\right)^{\frac{1}{\chi}}\right]$. The transformation parameters give the same transformation to both the control and query screens. Our model will learn about ϕ and χ, adjusting the relationship of genetic independence and consequently those identified as genetic interaction. Choosing to include a multiplicative transformation parameter where the model describes genetic independence (as an additive model) will give the model the flexibility to adjust the linear relationship between the control and query screens. Prior hyper-parameter choice for the transformation effect must be strictly positive and centred at 1 (no transformation effect) and so a gamma distribution with a mean of 1 is chosen for both χ and ϕ.

Figures 4.12 and 4.13 show JHM-B and JHM-T $MDR \times MDP$ fitness plots respectively, for the *cdc13-1* **27°C** vs *ura3Δ* **27°C** experiment. Prior hyper-parameter choices for the models are given Table B.1. Bayesian inference and MCMC methods for the JHM in Ta-

ble 3.3 is carried out similarly for both the JHM-B and JHM-T. Posterior samples of 1000 particles are obtained after a burn-in period of 800k, and a thinning of every 100 observations. Similarly to Section 4.3.5, chains from our MCMC sampler are accepted where the effective sample sizes are greater than 300 and Heidelberg and Welch P-values are greater than 0.10 for every chain. Similarly to the other previous modelling approaches considered (including the "vanilla" JHM), a list of 159 are stripped from our final list of genes for biological and experimental reasons.

The JHM-B fit in Figure 4.12 has many less interactions on the plot than the "vanilla" JHM fitness plot, this may be evidence of a plate effect existing. The JHM-T fit in Figure 4.13 is largely the same as the "vanilla" JHM fitness plot. It is worth noting that the JHM-T model fit in Figure 4.13 has posterior mean estimates of $\hat{\phi} = 0.96$ and $\hat{\chi} = 0.87$, 2dp, suggesting that a transformation may only exist in terms of r.

Table 4.6 compares the number of suppressors and enhancers estimated for the two extensions of the JHM. The JHM-B reduces the number of genetic interactions from the "vanilla" JHM from 939 to 553, and similarly reduces the number of suppressors and enhancers. Therefore from the "vanilla" JHM to the JHM-B, there is approximately a 41% reduction of genes identified as showing significant evidence of genetic interaction, strong evidence for the presence of a batch effect. The JHM-T is more similar to the JHM with 901 interactions, reducing both suppressors and enhancers by a small amount. Therefore from the "vanilla" JHM to the JHM-T, there is approximately a 4% reduction of genes identified as showing significant evidence of genetic interaction, a much smaller reduction from the JHM than that observed with the JHM-B.

Table 4.8A shows that the number of genes that overlap with the genes identified by the "vanilla" JHM is 531 and 886 for the JHM-B and JHM-T respectively. Therefore the number of genes identified as interacting by the "vanilla" JHM and now no longer identified is 408 and 53 for the JHM-B and JHM-T respectively. This further demonstrates the large reduction in genetic interactions when using the JHM-B, suggesting that a batch effect is present within the data. The number of genes newly identified as showing significant evidence of genetic interaction by the JHM-B and JHM-T is 22 and 15 respectively. These numbers are small relative to the number of genes that are no longer identified, indicating that the biggest change from the "vanilla" JHM is that the JHM-B and JHM-T are more stringent for determining significant genetic interactions. Table 4.8A shows that the "vanilla" JHM and JHM-T have similar overlap with the Addinall *et al.* (2011), REM and IHM approaches. The JHM-B has much less overlap with the Addinall *et al.* (2011) approach than the "vanilla" JHM does, reducing the overlap from 649 to 498, indicating

Table 4.8: Genes interacting with *cdc13-1* at **27°**C and GO terms over-represented in the list of interactions according to each approach A) Number of genes identified for each approach (Add Addinall *et al.* (2011), REM, IHM, JHM, JHM-B and JHM-T) and the overlap between the approaches. 4135 genes from the *S. cerevisiae* single deletion library are considered. B) Number of GO terms identified for each approach (Add Addinall *et al.* (2011), REM, IHM, JHM, JHM-B and JHM-T) and the overlap between the approaches. 6107 *S. cerevisiae* GO Terms available. See Tables 4.2A and 4.2B for further details on the overlap between the "vanilla" models (Add Addinall *et al.* (2011), REM, IHM, JHM).

A.	*Add*	*REM*	*IHM*	*JHM*	*JHM-B*	*JHM-T*
JHM	649	273	572	939	N/A	N/A
JHM-B	498	239	468	531	553	N/A
JHM-T	628	276	572	886	535	901

B.	*Add*	*REM*	*IHM*	*JHM*	*JHM-B*	*JHM-T*
JHM	219	165	216	286	N/A	N/A
JHM-B	223	170	217	204	265	N/A
JHM-T	215	160	219	267	206	293

that the changes lead to an approach that is even more dissimilar from the Addinall *et al.* (2011) approach.

Table 4.8B shows that the overlap in significant GO terms for the JHM-T and JHM-B with the JHM is 204 and 267 respectively. There are 286 (see Table 4.8B) significant GO terms found with the "vanilla" JHM, meaning there is a reduction of approximately 29% and 7% with the JHM-B and JHM-T respectively, demonstrating the difference of our new approaches from "vanilla" JHM. Table 4.8B also shows that the "vanilla" JHM, JHM-B and JHM-T all have a similar number of overlap in significant GO terms with the Addinall *et al.* (2011), REM and IHM approaches.

We have introduced two potential ways of further extending the JHM to better model a QFA screen comparison, Figures 4.12 and 4.13 are included as a reference for further research. The JHM-B has made large changes to our results by reducing the number of hits, see Table 4.6. Further research may involve investigating the behaviour of an alternative JHM-B with tighter priors for the batch effect parameters so we can see how the additional parameters affect the model fit in more detail. Further research for the JHM-T would involve developing an alternative JHM-T where different transformations are made for the control and query screens. We find that the largest difference with the JHM-B and JHM-T is that they are more stringent for determining genetic interactions than the "vanilla" JHM. Currently we prefer the "vanilla" JHM until further model exploration and analysis such as simulation studies are carried out to further investigate how the JHM-B and JHM-T affect our results.

Table 4.9: Description of the joint hierarchical model with batch effects. b identifies the batch which an *orf*Δ repeat belongs to. Further model notation is defined in Table 3.3

$$c = 0, 1 \qquad \qquad \text{Condition level}$$
$$l = 1, ..., L_c \qquad \qquad orf\Delta \text{ level}$$
$$m = 1, ..., M_{cl} \qquad \qquad \text{Repeat level}$$
$$n = 1, ..., N_{clm} \qquad \qquad \text{Time point level}$$
$$b = 1, ..., B \qquad \qquad \text{Batch}$$

Time point level

$$y_{clmn} \sim \mathrm{N}(\hat{y}_{clmn}, (\nu_{cl})^{-1}) \qquad\qquad \hat{y}_{clmn} = x(t_{clmn}; K_{clm}, r_{clm}, P)$$

Repeat level

$$\log K_{clm} \sim \mathrm{N}(\alpha_c + \kappa_b + K_l^o + \delta_l\gamma_{cl}, (\tau_{cl}^K)^{-1})I_{(-\infty,0]} \qquad \log \tau_{cl}^K \sim \mathrm{N}(\tau_c^{K,p}, (\sigma_c^{\tau,K})^{-1})I_{[0,\infty)}$$
$$\log r_{clm} \sim \mathrm{N}(\beta_c + \lambda_b + r_l^o + \delta_l\omega_{cl}, (\tau_{cl}^r)^{-1})I_{(-\infty,3.5]} \qquad \log \tau_{cl}^r \sim \mathrm{N}(\tau_c^{r,p}, (\sigma_c^{\tau,r})^{-1})$$

orfΔ level

$$e^{K_l^o} \sim t(K^p, (\sigma^{K,o})^{-1}, 3)I_{[0,\infty)} \qquad\qquad \log \sigma^{K,o} \sim \mathrm{N}(\eta^{K,o}, (\psi^{K,o})^{-1})$$
$$e^{r_l^o} \sim t(r^p, (\sigma^{r,o})^{-1}, 3)I_{[0,\infty)} \qquad\qquad \log \sigma^{r,o} \sim \mathrm{N}(\eta^{r,o}, (\psi^{r,o})^{-1})$$
$$\log \nu_{cl} \sim \mathrm{N}(\nu^p, (\sigma^\nu)^{-1}) \qquad\qquad \log \sigma^\nu \sim \mathrm{N}(\eta^\nu, (\psi^\nu)^{-1})$$
$$\delta_l \sim Bern(p)$$

$$e^{\gamma_{cl}} = \begin{cases} 1 & \text{if } c = 0; \\ t(1, (\sigma^\gamma)^{-1}, 3)I_{[0,\infty)} & \text{if } c = 1. \end{cases} \qquad \log \sigma^\gamma \sim \mathrm{N}(\eta^\gamma, \psi^\gamma)$$

$$e^{\omega_{cl}} = \begin{cases} 1 & \text{if } c = 0; \\ t(1, (\sigma^\omega)^{-1}, 3)I_{[0,\infty)} & \text{if } c = 1. \end{cases} \qquad \log \sigma^\omega \sim \mathrm{N}(\eta^\omega, \psi^\omega)$$

Condition level

$$\alpha_c = \begin{cases} 0 & \text{if } c = 0; \\ \mathrm{N}(\alpha^\mu, \eta^\alpha) & \text{if } c = 1. \end{cases} \qquad\qquad \beta_c = \begin{cases} 0 & \text{if } c = 0; \\ \mathrm{N}(\beta^\mu, \eta^\beta) & \text{if } c = 1. \end{cases}$$
$$\tau_c^{K,p} \sim \mathrm{N}(\tau^{K,\mu}, (\eta^{\tau,K,p})^{-1}) \qquad\qquad \log \sigma_c^{\tau,K} \sim \mathrm{N}(\eta^{\tau,K}, (\psi^{\tau,K})^{-1})$$
$$\tau_c^{r,p} \sim \mathrm{N}(\tau^{r,\mu}, (\eta^{\tau,r,p})^{-1}) \qquad\qquad \log \sigma_c^{\tau,r} \sim \mathrm{N}(\eta^{\tau,r}, (\psi^{\tau,r})^{-1})$$

Population level

$$\log K^p \sim \mathrm{N}(K^\mu, (\eta^{K,p})^{-1}) \qquad\qquad \log r^p \sim \mathrm{N}(r^\mu, (\eta^{r,p})^{-1})$$
$$\nu^p \sim \mathrm{N}(\nu^\mu, (\eta^{\nu,p})^{-1}) \qquad\qquad \log P \sim \mathrm{N}(P^\mu, (\eta^P)^{-1})$$

Batch

$$Log \, \kappa_b \sim \mathrm{N}(\kappa^p, (\eta^\kappa)^{-1}) \qquad\qquad Log \, \lambda_b \sim \mathrm{N}(\lambda^p, (\eta^\lambda)^{-1})$$

Table 4.10: Description of the joint hierarchical model with transformations. Model notation is defined in Table 3.3

$$c = 0, 1 \qquad\qquad \text{Condition level}$$
$$l = 1, ..., L_c \qquad\qquad orf\Delta \text{ level}$$
$$m = 1, ..., M_{cl} \qquad\qquad \text{Repeat level}$$
$$n = 1, ..., N_{clm} \qquad\qquad \text{Time point level}$$

Time point level

$$y_{clmn} \sim \text{N}(\hat{y}_{clmn}, (\nu_{cl})^{-1}) \qquad\qquad \hat{y}_{clmn} = x(t_{clmn}; K_{clm}, r_{clm}, P)$$

Repeat level

$$\log K_{clm} \sim \text{N}(\frac{\alpha_c + K_l^o + \delta_l \gamma_{cl}}{\phi}, (\tau_{cl}^K)^{-1})I_{(-\infty,0]} \qquad \log \tau_{cl}^K \sim \text{N}(\tau_c^{K,p}, (\sigma_c^{\tau,K})^{-1})I_{[0,\infty)}$$

$$\log r_{clm} \sim \text{N}(\frac{\beta_c + r_l^o + \delta_l \omega_{cl}}{\chi}, (\tau_{cl}^r)^{-1})I_{(-\infty,3.5]} \qquad \log \tau_{cl}^r \sim \text{N}(\tau_c^{r,p}, (\sigma_c^{\tau,r})^{-1})$$

orfΔ level

$$e^{K_l^o} \sim t(K^p, (\sigma^{K,o})^{-1}, 3)I_{[0,\infty)} \qquad\qquad \log \sigma^{K,o} \sim \text{N}(\eta^{K,o}, (\psi^{K,o})^{-1})$$
$$e^{r_l^o} \sim t(r^p, (\sigma^{r,o})^{-1}, 3)I_{[0,\infty)} \qquad\qquad \log \sigma^{r,o} \sim \text{N}(\eta^{r,o}, (\psi^{r,o})^{-1})$$
$$\log \nu_{cl} \sim \text{N}(\nu^p, (\sigma^\nu)^{-1}) \qquad\qquad \log \sigma^\nu \sim \text{N}(\eta^\nu, (\psi^\nu)^{-1})$$
$$\delta_l \sim Bern(p)$$

$$e^{\gamma_{cl}} = \begin{cases} 1 & \text{if } c = 0; \\ t(1, (\sigma^\gamma)^{-1}, 3)I_{[0,\infty)} & \text{if } c = 1. \end{cases} \qquad \log \sigma^\gamma \sim \text{N}(\eta^\gamma, \psi^\gamma)$$

$$e^{\omega_{cl}} = \begin{cases} 1 & \text{if } c = 0; \\ t(1, (\sigma^\omega)^{-1}, 3)I_{[0,\infty)} & \text{if } c = 1. \end{cases} \qquad \log \sigma^\omega \sim \text{N}(\eta^\omega, \psi^\omega)$$

Condition level

$$\alpha_c = \begin{cases} 0 & \text{if } c = 0; \\ \text{N}(\alpha^\mu, \eta^\alpha) & \text{if } c = 1. \end{cases} \qquad\qquad \beta_c = \begin{cases} 0 & \text{if } c = 0; \\ \text{N}(\beta^\mu, \eta^\beta) & \text{if } c = 1. \end{cases}$$
$$\tau_c^{K,p} \sim \text{N}(\tau^{K,\mu}, (\eta^{\tau,K,p})^{-1}) \qquad\qquad \log \sigma_c^{\tau,K} \sim \text{N}(\eta^{\tau,K}, (\psi^{\tau,K})^{-1})$$
$$\tau_c^{r,p} \sim \text{N}(\tau^{r,\mu}, (\eta^{\tau,r,p})^{-1}) \qquad\qquad \log \sigma_c^{\tau,r} \sim \text{N}(\eta^{\tau,r}, (\psi^{\tau,r})^{-1})$$

Population level

$$\log K^p \sim \text{N}(K^\mu, (\eta^{K,p})^{-1}) \qquad\qquad \log r^p \sim \text{N}(r^\mu, (\eta^{r,p})^{-1})$$
$$\nu^p \sim \text{N}(\nu^\mu, (\eta^{\nu,p})^{-1}) \qquad\qquad \log P \sim \text{N}(P^\mu, (\eta^P)^{-1})$$
$$\phi \sim \Gamma(\phi^{shape}, \phi^{scale}) \qquad\qquad \chi \sim \Gamma(\chi^{shape}, \chi^{scale})$$
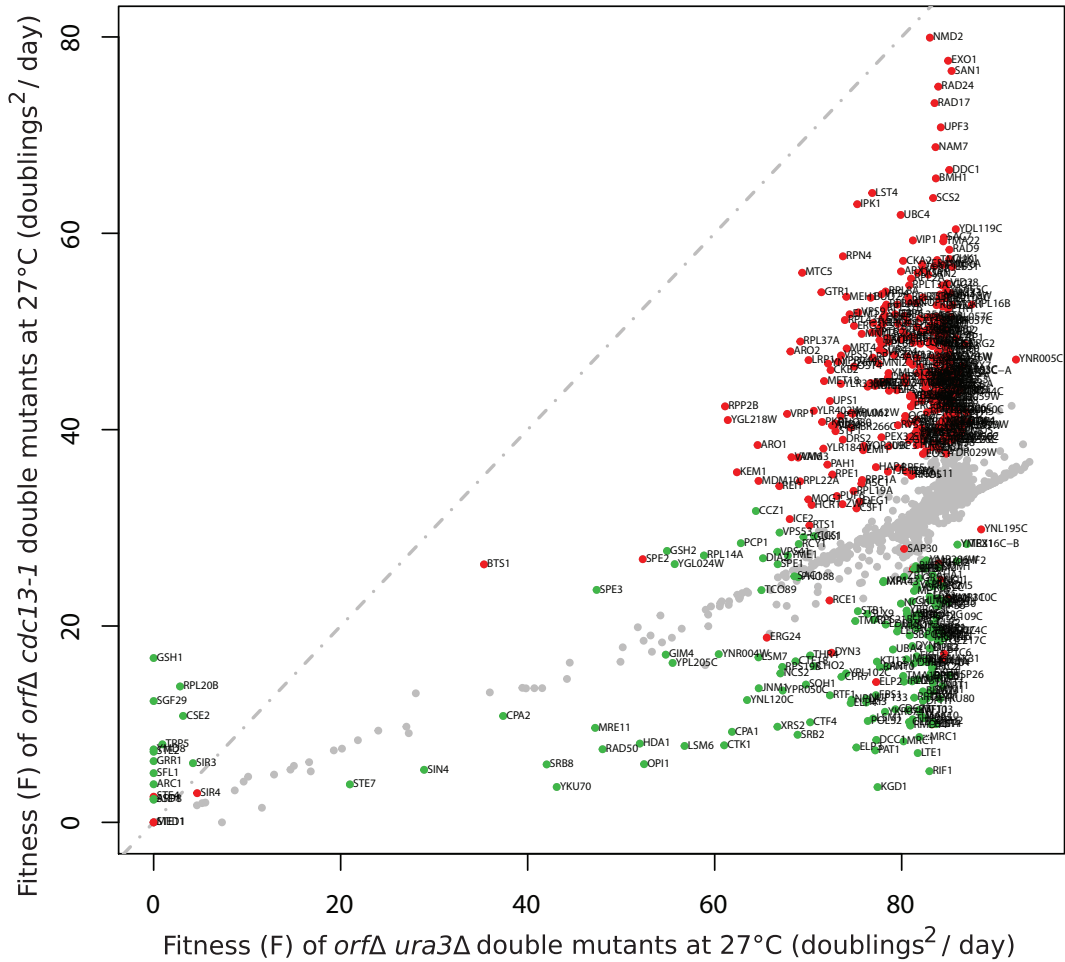
Figure 4.12: *cdc13-1* **27°**C vs *ura3Δ* **27°**C joint hierarchical model with Batch effect (JHM-B) fitness plot with *orfΔ* posterior mean fitnesses. The JHM does not does not make use of a fitness measure such as $MDR \times MDP$ but the fitness plot is given in terms of $MDR \times MDP$ for comparison with other approaches which do. *orfΔ* strains are classified as being a suppressor or enhancer based on one of the two parameters used to classify genetic interaction, growth parameter r, this means occasionally strains can be more fit in the query experiment in terms of $MDR \times MDP$ but be classified as enhancers (green). Further fitness plot explanation and notation is given in Figure 4.2.
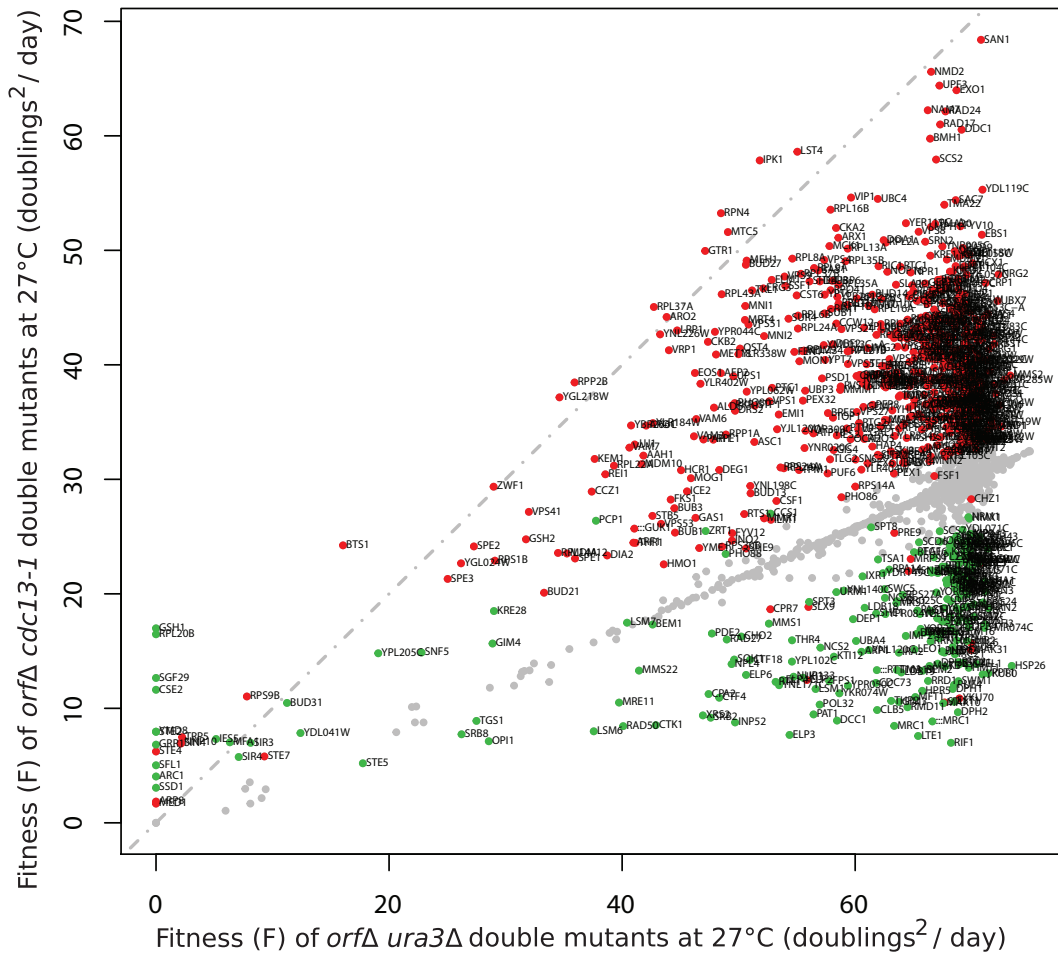
Figure 4.13: *cdc13-1* **27°C** vs *ura3Δ* **27°C** joint hierarchical model with transformations (JHM-T) fitness plot with *orfΔ* posterior mean fitnesses. The JHM does not does not make use of a fitness measure such as $MDR \times MDP$ but the fitness plot is given in terms of $MDR \times MDP$ for comparison with other approaches which do. *orfΔ* strains are classified as being a suppressor or enhancer based on one of the two parameters used to classify genetic interaction, growth parameter r, this means occasionally strains can be more fit in the query experiment in terms of $MDR \times MDP$ but be classified as enhancers (green). Further fitness plot explanation and notation is given in Figure 4.2.

Chapter 5. Fast Bayesian parameter estimation for stochastic logistic growth models

5.1. Introduction

In this Chapter, fast approximations to the stochastic logistic growth model (SLGM) (Capocelli & Ricciardi, 1974) (see Section 1.3) are presented. The SLGM is given by the following diffusion equation:

$$dX_t = rX_t \left(1 - \frac{X_t}{K} \right) dt + \sigma X_t dW_t, \tag{5.1}$$

where $X_{t_0} = P$ and is independent of W_t, $t \geq t_0$.

A deterministic logistic growth model (see Section 1.1.2) is unable to describe intrinsic error within stochastic logistic growth time course data. Consequently a deterministic model may lead to less accurate estimates of logistic growth parameters than a SDE, which can describe intrinsic noise. So that random fluctuations present within observed yeast QFA data (1.1) can accounted for as intrinsic noise instead of being confounded within our measurement error we are interested in using the SLGM in (5.1), instead of its deterministic counterpart (1.1). Alternative stochastic logistic growth equations exist (see Section 1.3) but we find (5.1) to be the most appropriate as intrinsic noise does not tend to zero with larger population sizes.

The SLGM (5.1) is analytically intractable and therefore inference requires relatively slow numerical simulation. Where fast inference is of importance such as real-time analysis or big data problems, we can use model approximations which do have analytically tractable densities, enabling fast inference. For large hierarchical Bayesian models (see Chapter 3), computational time for inference is typically long, ranging from one to two weeks using a deterministic logistic growth model. Inference for large hierarchical Bayesian models using the SLGM would increase computational time considerably (computational time is roughly proportional to the number of time points longer) with relatively slow numerical simulation approaches, therefore we may be interested in using approximate models that will allow us to carry out fast inference.

First an approximate model developed by Román-Román & Torres-Ruiz (2012) is introduced. Two new approximate models are then presented using the linear noise approximation (LNA) (Wallace, 2010; Komorowski *et al.*, 2009) of the SLGM. The model

proposed by Román-Román & Torres-Ruiz (2012) is found to be a zero-order noise approximation.

The approximate models considered are compared against each other for both simulated and observed logistic growth data. Finally, the approximate models are compared to "exact" approaches.

5.2. The Román-Román & Torres-Ruiz (2012) diffusion process

Román-Román & Torres-Ruiz (2012) present a logistic growth diffusion process (RRTR) which has a transition density that can be written explicitly, allowing inference for model parameter values from discrete sampling trajectories.

The RRTR is derived from the following ODE:

$$\frac{dx_t}{dt} = \frac{Qr}{e^{rt} + Q} x_t, \tag{5.2}$$

where $Q = \left(\frac{K}{P} - 1\right) e^{rt_0}$, $P = x_{t_0}$ and $t \geq t_0$. The solution to (5.2) is given in (1.2) (it has the same solution as (1.1)).

Román-Román & Torres-Ruiz (2012) see (5.2) as a generalisation of the Malthusian growth model with a deterministic, time-dependent fertility $h(t) = \frac{Qr}{e^{rt}+Q}$, and replace this with $\frac{Qr}{e^{rt}+Q} + \sigma W_t$ to obtain the following approximation to the SLGM:

$$dX_t = \frac{Qr}{e^{rt} + Q} X_t dt + \sigma X_t dW_t, \tag{5.3}$$

where $Q = \left(\frac{K}{P} - 1\right) e^{rt_0}$, $P = X_{t_0}$ and is independent of W_t, $t \geq t_0$. The process described in (5.3) is a particular case of the Log-normal process with exogenous factors, therefore an exact transition density is available (Gutiérrez *et al.*, 2006). The transition density for Y_t, where $Y_t = \log(X_t)$, can be written:

$$
\begin{aligned}
&(Y_{t_i}|Y_{t_{i-1}} = y_{t_{i-1}}) \sim \mathrm{N}\left(\mu_{t_i}, \Xi_{t_i}\right), \\
&\quad \text{where } a = r, \quad b = \frac{r}{K}, \\
&\mu_{t_i} = \log(y_{t_{i-1}}) + \log\left(\frac{1 + be^{-at_i}}{1 + be^{-at_{i-1}}}\right) - \frac{\sigma^2}{2}(t_i - t_{i-1}) \text{ and} \\
&\Xi_{t_i} = \sigma^2(t_i - t_{i-1}).
\end{aligned}
\tag{5.4}
$$

5.3. Linear noise approximation with multiplicative noise

We now take a different approach to approximating the SLGM (5.1), which will turn out to be closer to the exact solution of the SLGM than the RRTR (5.3). Starting from the original model (5.1), we apply Itô's lemma (Itô, 1944; Øksendal, 2010):

$$df(t, X_t) = \frac{df}{dt}dt + \mu\frac{df}{dx}dt + \frac{1}{2}\sigma^2\frac{d^2 f}{dx^2}dt + \sigma\frac{df}{dx}dW_t, \tag{5.5}$$

with the transformation $f(t, X_t) \equiv Y_t = \log X_t$. After deriving the following partial derivatives:

$$\frac{df}{dt} = 0, \qquad \frac{df}{dx} = \frac{1}{X_t} \quad \text{and} \quad \frac{d^2 f}{dx^2} = -\frac{1}{X_t^2},$$

we can obtain the following Itô drift-diffusion process:

$$dY_t = \left(r - \frac{1}{2}\sigma^2 - \frac{r}{K}e^{Y_t}\right)dt + \sigma dW_t. \tag{5.6}$$

The log transformation from multiplicative to additive noise, gives a constant diffusion term, so that the LNA will give a good approximation to (5.1). The LNA reduces a non-linear SDE to a linear SDE with additive noise. The LNA can be viewed as a first order Taylor expansion of an approximating SDE about a deterministic solution. We now separate the process Y_t into a deterministic part v_t and a stochastic part Z_t so that $Y_t = v_t + Z_t$ and consequently $dY_t = dv_t + dZ_t$. We choose v_t to be the solution of the deterministic part of (5.6):

$$dv_t = \left(r - \frac{1}{2}\sigma^2 - \frac{r}{K}e^{v_t}\right)dt. \tag{5.7}$$

We now redefine our notation as follows: $a = r - \frac{\sigma^2}{2}$ and $b = \frac{r}{K}$. Equation 5.7 is then solved for v_t:

$$v_t = \log\left(\frac{aPe^{aT}}{bP(e^{aT} - 1) + a}\right), \tag{5.8}$$

where $T = t - t_0$. We now write down an expression for dZ_t, where $dZ_t = dY_t - dv_t$:

$$dZ_t = \left(a - be^{Y_t}\right)dt + \sigma dW_t - \left(a - be^{v_t}\right)dt$$

We then substitute in $Y_t = v_t + Z_t$ and simplify the expression to give

$$dZ_t = b\left(e^{v_t} - e^{v_t + Z_t}\right)dt + \sigma dW_t. \tag{5.9}$$

As dZ_t is a non-linear SDE it cannot be solved explicitly, we use the LNA (see Section 2.6.4) to obtain a linear SDE that we can solve explicitly. We apply the LNA by making a first-order approximation of $e^{Z_t} \approx 1 + Z_t$ and then simplify to give

$$dZ_t = -be^{v_t} Z_t dt + \sigma dW_t. \tag{5.10}$$

This process is a particular case of the time-varying Ornstein-Uhlenbeck process, which can be solved explicitly. The transition density for Y_t (derivation in Appendix C.1) is then:

$$(Y_{t_i}|Y_{t_{i-1}} = y_{t_{i-1}}) \sim \mathrm{N}\left(\mu_{t_i}, \Xi_{t_i}\right),$$

$$\text{redefine } y_{t_{i-1}} = v_{t_{i-1}} + z_{t_{i-1}}, Q = \left(\frac{\frac{a}{b}}{P} - 1\right) e^{at_0},$$

$$\mu_{t_i} = y_{t_{i-1}} + \log\left(\frac{1 + Qe^{-at_{i-1}}}{1 + Qe^{-at_i}}\right) + e^{-a(t_i - t_{i-1})}\frac{1 + Qe^{-at_{i-1}}}{1 + Qe^{-at_i}} z_{t_{i-1}} \text{ and}$$

$$\Xi_{t_i} = \sigma^2 \left[\frac{4Q(e^{at_i} - e^{at_{i-1}}) + e^{2at_i} - e^{2at_{i-1}} + 2aQ^2(t_i - t_{i-1})}{2a(Q + e^{at_i})^2}\right]. \tag{5.11}$$

The LNA of the SLGM with multiplicative intrinsic noise (LNAM) can then be written as

$$d \log X_t = [dv_t + be^{v_t} v_t - be^{v_t} \log X_t]\, dt + \sigma dW_t,$$

where $P = X_{t_0}$ and is independent of W_t, $t \geq t_0$.

Note that the RRTR given in (5.3) can be similarly derived using a zero-order noise approximation ($e^{Z_t} \approx 1$) instead of the LNA.

5.4. Linear noise approximation with additive noise

As in Section 5.3, we start from the SLGM, given in (5.1). Without first log transforming the process, the LNA will lead to a worse approximation to the diffusion term of the SLGM, but we will see in the coming sections that there are nevertheless advantages. We separate the process X_t into a deterministic part v_t and a stochastic part Z_t so that $X_t = v_t + Z_t$ and consequently $dX_t = dv_t + dZ_t$. We chose v_t to be the solution of the deterministic part of (5.1):

$$dv_t = \left(rv_t - \frac{r}{K}v_t^2\right) dt. \tag{5.12}$$

We now redefine our previous notation as follows: $a = r$ and $b = \frac{r}{K}$. Equation 5.12 is then solved for v_t:

$$v_t = \frac{aPe^{aT}}{bP(e^{aT} - 1) + a}.$$ (5.13)

We now write down an expression for dZ_t, where $dZ_t = dX_t - dv_t$:

$$dZ_t = \left(aX_t - bX_t^2\right) dt + \sigma X_t dW_t - \left(av_t - bv_t^2\right) dt.$$

We then substitute in $X_t = v_t + Z_t$ and simplify the expression to give

$$dZ_t = (a - 2bv_t)Z_t - bZ_t^2 dt + \left(\sigma v_t + \sigma Z_t\right) dW_t.$$

As dZ_t is a non-linear SDE it cannot be solved explicitly, we use the LNA (see Section 2.6.4) to obtain a linear SDE that we can solve explicitly. We now apply the LNA, by setting second-order term $-bZ_t^2 dt = 0$ and $\sigma Z_t dW_t = 0$ to obtain

$$dZ_t = (a - 2bv_t)Z_t dt + \sigma v_t dW_t.$$ (5.14)

This process is a particular case of the Ornstein-Uhlenbeck process, which can be solved. The transition density for X_t (derivation in Appendix C.3) is then

$$(X_{t_i}|X_{t_{i-1}} = x_{t_{i-1}}) \sim N(\mu_{t_i}, \Xi_{t_i}),$$

where $x_{t_{i-1}} = v_{t_{i-1}} + z_{t_{i-1}}$,

$$\mu_{t_i} = x_{t_{i-1}} + \left(\frac{aPe^{aT_i}}{bP(e^{aT_i} - 1) + a}\right) - \left(\frac{aPe^{aT_{i-1}}}{bP(e^{aT_{i-1}} - 1) + a}\right)$$
$$+ e^{a(t_i - t_{i-1})} \left(\frac{bP(e^{aT_{i-1}} - 1) + a}{bP(e^{aT_i} - 1) + a}\right)^2 Z_{t_{i-1}} \text{ and}$$ (5.15)
$$\Xi_t = \frac{1}{2}\sigma^2 aP^2 e^{2aT_i} \left(\frac{1}{bP(e^{aT_i} - 1) + a}\right)^4$$
$$\times [b^2 P^2 (e^{2aT_i} - e^{2aT_{i-1}}) + 4bP(a - bP)(e^{aT_i} - e^{aT_{i-1}})$$
$$+ 2a(t_i - t_{i-1})(a - bP)^2].$$

The LNA of the SLGM, with additive intrinsic noise (LNAA) can then be written as

$$dX_t = \left[bv_t^2 + (a - 2bv_t) X_t\right] dt + \sigma v_t dW_t,$$

where $P = X_{t_0}$ and is independent of W_t, $t \geq t_0$.

5.5. Simulation and Bayesian inference for the stochastic logistic growth model and approximations

To compare the accuracies of each of the three approximate models in representing the SLGM, we first compare simulated forward trajectories from the RRTR, LNAM and LNAA with simulated forward trajectories from the SLGM (Figure 5.1). We use the Euler-Maruyama method (Carletti, 2006) (see Section 2.6.2) with very fine discretisation to give arbitrarily exact simulated trajectories from each SDE.

The LNAA and LNAM trajectories are visually indistinguishable from the SLGM (Figures 5.1 A, C & D). On the other hand, population sizes simulated with the RRTR display large deviations from the mean as the population approaches its stationary phase (Figures 5.1A & B). Figure 5.1E further highlights the increases in variation as the population approaches stationary phase for simulated trajectories of the RRTR, in contrast to the SLGM and LNA models.
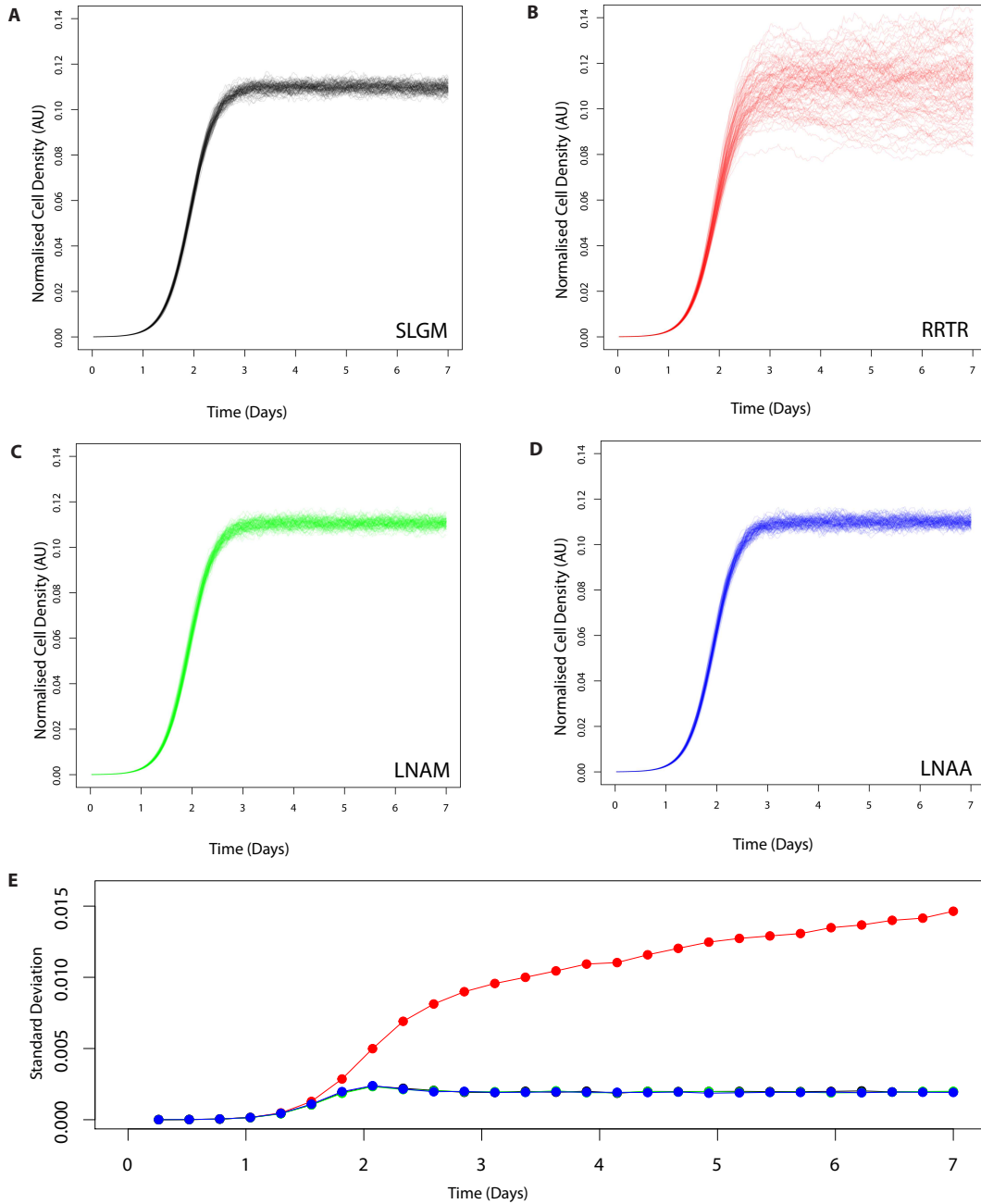
Figure 5.1:   Forward trajectories (No. of simulations=100) for the stochastic logistic growth model and approximations. See Table 5.1 for parameter values. A) The stochastic logistic growth model (SLGM). B) The Román-Román & Torres-Ruiz (2012) (RRTR) approximation. C) The linear noise approximation with multiplicative intrinsic noise (LNAM). D) The linear noise approximation with additive intrinsic noise (LNAA). E) Standard deviations of simulated trajectories over time for the SLGM (black), RRTR (red), LNAM (green) and LNAA (blue).

5.5.1. Bayesian parameter inference with approximate models

To compare the quality of parameter inference using each of these approximations we simulated synthetic time-course data from the SLGM and combined this with either Log-normal or Normal measurement error. Carrying out Bayesian inference with broad priors (see (5.16) and (5.17)) we compared the parameters recovered using each approximation with those used to generate the synthetic dataset. The synthetic time-course datasets consist of 27 time points generated using the Euler-Maruyama method with very fine intervals (Carletti, 2006).

We formulate our inference problem as a dynamic linear state space model (West & Harrison, 1997). The advantage of a state space formulation is that we are then able to build a Kalman filter to carry out fast parameter inference. We can take advantage of a linear Gaussian structure and construct a Kalman filter recursion for marginal likelihood computation (Appendix C.5). By choosing to match the measurement error structure to the intrinsic error of our models we can build a linear Gaussian structure. We therefore assume Log-normal (multiplicative) error for the RRTR and LNAM, and for the LNAA we assume Normal (additive) measurement error. Dependent variable y_{t_i} and independent variable $\{t_i, i = 1, ..., N\}$ are data input to the model (where t_i is the time at point i and N is the number of time points). X_t is the state process, describing the population size.

The state space model for the RRTR and LNAM is as follows:

$$\log(y_{t_i}) \sim \mathrm{N}(X_{t_i}, \nu^2),$$
$$(X_{t_i}|X_{t_{i-1}} = x_{t_{i-1}}) \sim \mathrm{N}\left(\mu_{t_i}, \Xi_{t_i}\right), \text{ where } x_{t_i} = v_{t_i} + z_{t_i}, \tag{5.16}$$

μ_{t_i} and Ξ_{t_i} are given by (5.4) and (5.11) for the RRTR and LNAM respectively. Priors are as follows:

$$\log X_0 \equiv \log P \sim \mathrm{N}(\mu_P, \tau_P{}^{-1}), \qquad \log K \sim \mathrm{N}(\mu_K, \tau_K{}^{-1}), \qquad \log r \sim \mathrm{N}(\mu_r, \tau_r{}^{-1}),$$
$$\log \nu^{-2} \sim \mathrm{N}(\mu_\nu, \tau_\nu{}^{-1}), \qquad \log \sigma^{-2} \sim \mathrm{N}(\mu_\sigma, \tau_\sigma{}^{-1})I_{[1,\infty]}.$$

Bayesian inference is carried out with broad priors such that estimated parameter values are not heavily influenced by our choice. See Table C.1 for prior hyper-parameter values. Log-normal prior distributions are chosen to ensure positive logistic growth parameters and precision parameters are strictly positive. Our prior for $\log \sigma^{-2}$ is truncated below 1 to avoid unnecessary exploration of extremely low probability regions, which could be caused by problems identifying ν, for example when $\log \nu^{-2}$ takes large val-

ues, and to ensure that intrinsic noise does not dominate the process. Our choice of 1 for the truncation threshold is made by observing forward simulations from our processes and choosing a value for $\log \sigma^{-2}$ where intrinsic noise is so large that the deterministic part of the process is masked, consequently making the LNA a bad approximation. We also find that truncating $\log \sigma^{-2}$ is more preferable to truncating $\log \nu^{-2}$ as truncating $\log \nu^{-2}$ does not alleviate the identifiability problem without being very restrictive for the measurement error structures.

The state space model for the LNAA is as follows:

$$y_{t_i} \sim \mathrm{N}(X_{t_i}, \nu^2),$$
$$(X_{t_i}|X_{t_{i-1}} = x_{t_{i-1}}) \sim \mathrm{N}\left(\mu_{t_i}, \Xi_{t_i}\right), \text{ where } x_{t_i} = v_{t_i} + z_{t_i}, \quad (5.17)$$

μ_{t_i} and Ξ_{t_i} are given by (5.15). Priors are as in (5.16). Measurement error for the observed values is Normal so that we have a linear Gaussian structure. The state space models in (5.16) and (5.17) have different measurement error structures. So that a fair comparison can be made between (5.16) and (5.17), we choose our priors so that the marginal moments for the measurement error of our models is not too dissimilar, particularly at the earliest stage where most growth is observed.

To see how the inference from our approximate models compares with slower "exact" models, we consider Euler-Maruyama approximations (Kloeden & Platen, 1992) of (5.1) and of the log transformed process, using fine intervals. We use the approach of (Golightly & Wilkinson, 2005) to carry out inference of our "exact" models. A single site update algorithm is used to update model parameters and the Euler-Maruyama approximation of the latent process in turn. Given these approximations we can construct a state space model for an "exact" SLGM with Log-normal measurement error (SLGM+L) and similarly for the SLGM with Normal measurement error (SLGM+N), priors are as in (5.16).

Our inference makes use of a Kalman filter to integrate out the state process. The Kalman filer allows for fast inference compared to slow numerical simulation approaches that impute all states. The algorithm for our approximate models is the Metropolis-within-Gibbs sampler with a symmetric proposal (Gamerman & Lopes, 2006). Full-conditionals are sampled in turn to give samples from the joint posterior distribution:

$$\pi(K, r, P, \sigma, \nu, X_{t_{1:N}}, y_{t_{1:N}}),$$

where $X_{t_{1:N}}$ is the latent process and $y_{t_{1:N}}$ is the observed data, for N observed data points. The Metropolis-within-Gibbs sampler algorithm is as follows:

1) Initialise counter $i = 1$ and parameters $K_{(0)}, r_{(0)}, \sigma_{(0)}, P_{(0)}, \nu_{(0)}$

2) Simulate $K_{(i)}$ from $K \sim \pi(K|\nu_{(i-1)}, r_{(i-1)}, \sigma_{(i-1)}, P_{(i-1)}, y_{t_{1:N}})$

3) Simulate $r_{(i)}$ from $r \sim \pi(r|\nu_{(i-1)}, K_{(i)}, \sigma_{(i-1)}, P_{(i-1)}, y_{t_{1:N}}$

4) Simulate $\sigma_{(i)}$ from $\sigma \sim \pi(\sigma|\nu_{(i-1)}, K_{(i)}, r_{(i)}, P_{(i-1)}, y_{t_{1:N}})$

5) Simulate $P_{(i)}$ from $\nu \sim \pi(P|\nu_{(i-1)}, K_{(i)}, r_{(i)}, \sigma_{(i)}, y_{t_{1:N}})$

6) Simulate $\nu_{(i)}$ from $\nu \sim \pi(\nu|K_{(i)}, r_{(i)}, \sigma_{(i)}, P_{(i)}, y_{t_{1:N}})$

7) Repeat steps 2-6 until the sample size required is obtained.

We find the mixing for our algorithm is improved when we have intermediate steps between sampling from the $\sigma_{(i)}$ and $\nu_{(i)}$ full conditionals. Each update in our algorithm is accomplished by a Metropolis-Hastings step using a Kalman filter. Acceptance ratios are calculated for each update during a burn-in period. To improve the computational speed of our inference, further research may involve using an algorithm where we jointly update our parameters. Posterior means are used to obtain point estimates and standard deviations for describing variation of inferred parameters. The Heidelberger and Welch convergence diagnostic (Heidelberger & Welch, 1981) is used to determine whether convergence has been achieved for all parameters.

Computational times for convergence of our MCMC schemes (code is available at `https://github.com/jhncl/LNA.git`) can be compared using estimates for the minimum effective sample size per second (ESS$_{min}$/sec) (Plummer *et al.*, 2006). The average ESS$_{min}$/sec of our approximate model (coded in C) is $\sim$100 and "exact" model $\sim$1 (coded in JAGS (Plummer, 2010) with 15 imputed states between time points, chosen to maximise ESS$_{min}$/sec). We find that our C code is typically twice as fast as the simple MCMC scheme used by JAGS, indicating that our inference is $\sim$50$\times$ faster than an "exact" approach. A more efficient "exact" approach could speed up further, say by another factor of 5, but our approximate approach will at least be an order of magnitude faster.

We use a burn-in of 600,000 and a thinning of 4,000 to obtain a final posterior sample size of 1,000 for MCMC convergence of all our models.

To compare the approximate models ability to recover parameters from the SLGM with simulated Log-normal measurement error, we simulate data and carry out Bayesian inference. Figure 5.2 shows that all three approximate models can capture the synthetic time-course well, but that the RRTR model is the least representative with the largest amount of drift occurring at the saturation stage, a property not found in the SLGM or the two new LNA models. Comparing forwards trajectories with measurement error (Figure 5.2), the "exact" model is visually similar to all our approximate models, but least similar to the RRTR. Further, Table 5.1 demonstrates that parameter posterior means are close to the true values and that standard deviations are small for all models and each parameter set. By comparing posterior means and standard deviations to the true values, Table 5.1 shows that all our models are able to recover the three different parameter sets considered.

To compare the approximations to the SLGM with simulated Normal measurement error, we simulate data and carry out Bayesian inference. Figure 5.4 shows that of our approximate models, only the LNAA model can appropriately represent the simulated time-course as both our models with Log-normal measurement error, the RRTR and LNAM do not closely bound the data. Comparing forwards trajectories with measurement error (Figure 5.4), the "exact" model is most visually similar to the LNAA, which shares the same measurement error structure. Further, Table 5.1 demonstrates that only our models with Normal measurement error have posterior means close to the true values and that standard deviations are larger in the models with Log-normal measurement error. Observing the posterior means for K for each parameter set (Table 5.1), we can see that the RRTR has the largest standard deviations and that, of the approximate models, its posterior means are furthest from both the true values and the "exact" model posterior means. Comparing LNA models to the "exact" models with matching measurement error, we can see in Table 5.1 that they share similar posterior means and only slightly larger standard deviations. Example posterior diagnostics given in Figure 5.3, demonstrate that posteriors are distributed tightly around true values for our LNAA and data from the SLGM with Normal measurement error.

5.5.2. Application to observed yeast data

We now consider which diffusion equation model can best represent observed microbial population growth curves taken from a Quantitative Fitness Analysis (QFA) experiment
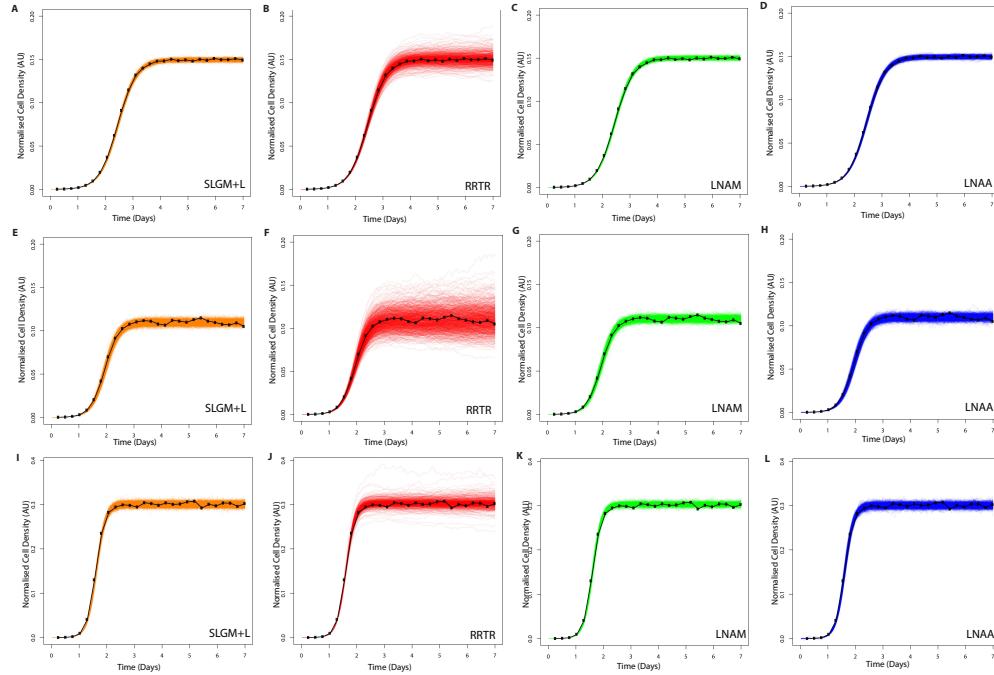
Figure 5.2: Forward trajectories with measurement error for the stochastic logistic growth model and approximations, simulated from parameter posterior samples (sample size=1000). Model fitting is carried out on SLGM forward trajectories with Log-normal measurement error (black), for three different sets of parameters (see Table 5.1). See (5.16) or (5.17) for model and Table C.1 for prior hyper-parameter values. Each row of figures corresponds to a different time course data set, simulated from a different set of parameter values, see Table 5.1. Each column of figures corresponds to a different model fit: A), E) & I) SLGM+L (orange). B), F) & J) RRTR model with lognormal error (red). C), G) & K) LNAM model with lognormal error (green). D), H) & L) LNAA model with normal error (blue). See Table 5.1 for parameter posterior means and true values.

(Section 1.1) (Addinall *et al.*, 2011; Banks *et al.*, 2012), see Figure 5.5. The data consists of scaled cell density estimates over time for budding yeast *Saccharomyces cerevisiae*. Independent replicate cultures are inoculated on plates and photographed over a period of 5 days. The images captured are then converted into estimates of integrated optical density (IOD, which we assume are proportional to cell population size), by the software package Colonyzer (Lawless *et al.*, 2010). The dataset chosen for our model fitting is a representative set of 10 time-courses, each with 27 time points. Once we have chosen the most appropriate stochastic model we can then look to apply our chosen model to logistic growth data from the QFA screens used throughout Chapter 4 in the future.

As in Figure 5.4, we see that the LNAA model is the only approximation that can appropriately represent the time-course and that both the RRTR and LNAM fail to bound the data as tightly as the LNAA (Figure 5.5). Our two "exact" models are visually similar
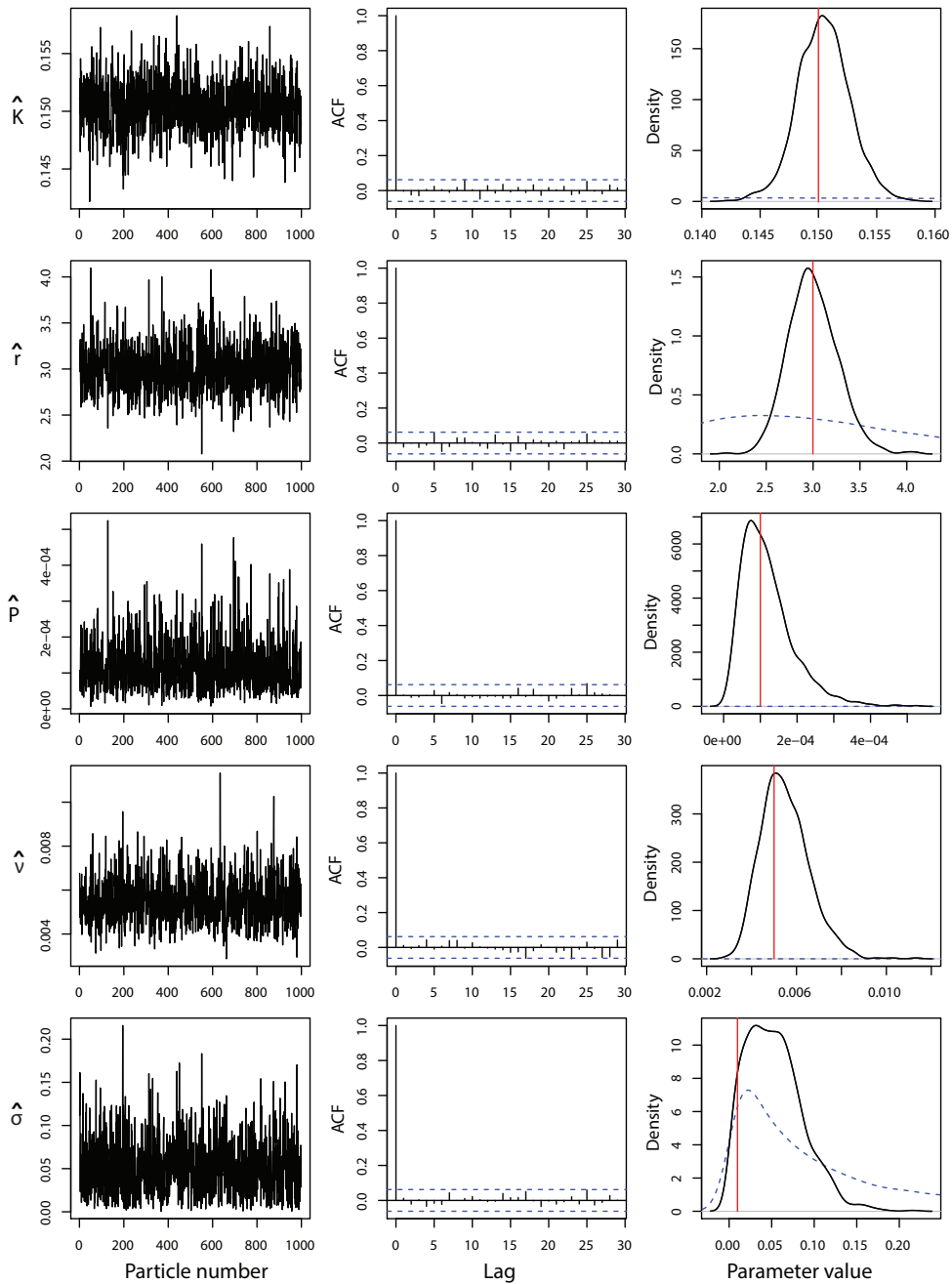
Figure 5.3: Convergence diagnostics for the linear noise approximation of the stochastic logistic growth model with additive intrinsic noise (LNAA) fit to simulated stochastic logistic growth data with Normal measurement error, see Figure 5.4D. Trace, auto-correlation and density plots for the (LNAA) parameter posteriors (sample size = 1000, thinning interval = 4000). Posterior density (black), prior density (dashed blue) and true parameter values (red) are shown in the right hand column.

Figure 5.4: Forward trajectories with measurement error, simulated from inferred parameter posterior samples (sample size=1000). Model fitting is carried out on SLGM forward trajectories with Normal measurement error (black), for three different sets of parameters (see Table 5.1). See (5.16) or (5.17) for model and Table C.1 for prior hyper-parameter values. Each row of figures corresponds to a different time course data set, simulated from a different set of parameter values, see Table 5.1. Each column of figures corresponds to a different model fit: A), E) & I) SLGM+N (pink). B), F) & J) RRTR model with lognormal error (red). C), G) & K) LNAM model with lognormal error (green). D), H) & L) LNAA model with normal error (blue). See Table 5.1 for parameter posterior means and true values.

to our approximate models with the same measurement error, with the SLGM+N most similar to the LNAA and the SLGM+L to the RRTR and LNAM. This is as expected due to matching measurement error structures. Table 5.1 summarises parameter estimates for the observed yeast data using each model. The variation in the LNAA model parameter posteriors is much smaller than the RRTR and LNAM, indicating a more appropriate model fit. Comparing the LNA models and "exact" models with matching measurement error, we can see in Table 5.1 that they share similar posterior means and standard deviations for all parameters and in particular, they are very similar for both K and r, which are important phenotypes for calculating fitness (Addinall *et al.*, 2011).

In Table 5.2, to compare quality of parameter inference for 10 observed yeast timecourses with each approximate model. Mean squared error (MSE) for 1000 posterior sample forward simulations are calculated for each yeast time course and summed to give

Table 5.1: Bayesian state space model parameter posterior means, standard deviations and true values for Figures 5.2, 5.4 and 5.5. True values for the simulated data used for Figures 5.1, 5.2 and 5.4 are also given.

Panel	Model	$\hat{K}$		$\hat{r}$		$\hat{P}$		$\hat{\nu}$		$\hat{\sigma}$	
Figure 5.2, SLGM with lognormal error											
A	SGLM+L	0.150	(0.001)	2.982	(0.014)	$1.002 \cdot 10^{-04}$	$(1.112 \cdot 10^{-06})$	$3.860 \cdot 10^{-03}$	$(2.127 \cdot 10^{-03})$	0.017	(0.005)
B	RRTR	0.150	(0.003)	2.990	(0.011)	$9.931 \cdot 10^{-05}$	$(1.069 \cdot 10^{-06})$	$5.684 \cdot 10^{-03}$	$(2.360 \cdot 10^{-03})$	0.012	(0.006)
C	LNAM	0.150	(0.001)	2.988	(0.013)	$9.980 \cdot 10^{-05}$	$(1.124 \cdot 10^{-06})$	$4.140 \cdot 10^{-03}$	$(2.180 \cdot 10^{-03})$	0.016	(0.005)
D	LNAA	0.150	(0.001)	3.005	(0.020)	$9.647 \cdot 10^{-05}$	$(2.946 \cdot 10^{-06})$	$3.099 \cdot 10^{-05}$	$(2.534 \cdot 10^{-05})$	0.019	(0.003)
E	SGLM+L	0.110	(0.001)	3.975	(0.047)	$5.054 \cdot 10^{-05}$	$(1.568 \cdot 10^{-06})$	$6.159 \cdot 10^{-03}$	$(5.527 \cdot 10^{-03})$	0.051	(0.014)
F	RRTR	0.109	(0.007)	3.984	(0.035)	$5.046 \cdot 10^{-05}$	$(1.137 \cdot 10^{-06})$	$5.928 \cdot 10^{-03}$	$(4.596 \cdot 10^{-03})$	0.037	(0.009)
G	LNAM	0.110	(0.001)	3.985	(0.046)	$5.043 \cdot 10^{-05}$	$(1.580 \cdot 10^{-06})$	$6.188 \cdot 10^{-03}$	$(5.191 \cdot 10^{-03})$	0.052	(0.013)
H	LNAA	0.110	(0.001)	3.959	(0.067)	$5.207 \cdot 10^{-05}$	$(4.310 \cdot 10^{-06})$	$4.540 \cdot 10^{-05}$	$(4.395 \cdot 10^{-05})$	0.059	(0.010)
I	SGLM+L	0.300	(0.001)	5.997	(0.029)	$1.962 \cdot 10^{-05}$	$(4.041 \cdot 10^{-07})$	$9.543 \cdot 10^{-03}$	$(4.035 \cdot 10^{-03})$	0.024	(0.015)
J	RRTR	0.301	(0.004)	6.015	(0.017)	$1.943 \cdot 10^{-05}$	$(2.835 \cdot 10^{-07})$	$1.241 \cdot 10^{-02}$	$(2.307 \cdot 10^{-03})$	0.008	(0.006)
K	LNAM	0.300	(0.001)	6.015	(0.031)	$1.953 \cdot 10^{-05}$	$(4.202 \cdot 10^{-07})$	$8.943 \cdot 10^{-03}$	$(4.252 \cdot 10^{-03})$	0.027	(0.016)
L	LNAA	0.300	(0.001)	6.037	(0.067)	$1.895 \cdot 10^{-05}$	$(1.502 \cdot 10^{-06})$	$8.122 \cdot 10^{-05}$	$(1.596 \cdot 10^{-04})$	0.047	(0.008)
Figure 5.4, SLGM with normal error											
A	SLGM+N	0.150	(0.002)	3.099	(0.085)	$9.299 \cdot 10^{-05}$	$(7.305 \cdot 10^{-06})$	$5.326 \cdot 10^{-03}$	$(1.009 \cdot 10^{-03})$	0.059	(0.030)
B	RRTR	0.213	(0.123)	1.368	(0.263)	$4.552 \cdot 10^{-03}$	$(2.118 \cdot 10^{-03})$	$2.539 \cdot 10^{-01}$	$(1.097 \cdot 10^{-01})$	0.419	(0.129)
C	LNAM	0.171	(0.033)	1.580	(0.271)	$5.241 \cdot 10^{-03}$	$(2.048 \cdot 10^{-03})$	$2.054 \cdot 10^{-01}$	$(7.805 \cdot 10^{-02})$	0.473	(0.051)
D	LNAA	0.150	(0.002)	2.990	(0.262)	$1.189 \cdot 10^{-04}$	$(7.099 \cdot 10^{-05})$	$5.490 \cdot 10^{-03}$	$(1.060 \cdot 10^{-03})$	0.053	(0.033)
E	SLGM+N	0.109	(0.001)	4.183	(0.074)	$4.390 \cdot 10^{-05}$	$(4.129 \cdot 10^{-06})$	$9.679 \cdot 10^{-04}$	$(2.806 \cdot 10^{-04})$	0.057	(0.012)
F	RRTR	0.157	(0.087)	2.631	(0.337)	$4.398 \cdot 10^{-04}$	$(1.678 \cdot 10^{-04})$	$1.040 \cdot 10^{-01}$	$(1.009 \cdot 10^{-01})$	0.374	(0.162)
G	LNAM	0.116	(0.009)	3.019	(0.374)	$4.967 \cdot 10^{-04}$	$(1.397 \cdot 10^{-04})$	$3.346 \cdot 10^{-02}$	$(4.309 \cdot 10^{-02})$	0.475	(0.044)
H	LNAA	0.110	(0.001)	4.010	(0.158)	$5.012 \cdot 10^{-05}$	$(1.443 \cdot 10^{-05})$	$1.093 \cdot 10^{-03}$	$(3.638 \cdot 10^{-04})$	0.053	(0.013)
I	SLGM+N	0.305	(0.003)	5.267	(0.125)	$3.263 \cdot 10^{-04}$	$(3.407 \cdot 10^{-05})$	$1.119 \cdot 10^{-02}$	$(1.974 \cdot 10^{-03})$	0.045	(0.031)
J	RRTR	0.314	(0.057)	3.030	(0.233)	$1.307 \cdot 10^{-03}$	$(2.897 \cdot 10^{-04})$	$2.228 \cdot 10^{-01}$	$(3.708 \cdot 10^{-02})$	0.075	(0.086)
K	LNAM	0.313	(0.020)	3.392	(0.430)	$1.118 \cdot 10^{-03}$	$(3.269 \cdot 10^{-04})$	$1.176 \cdot 10^{-01}$	$(8.435 \cdot 10^{-02})$	0.360	(0.165)
L	LNAA	0.302	(0.002)	5.862	(0.523)	$2.890 \cdot 10^{-05}$	$(2.599 \cdot 10^{-05})$	$8.774 \cdot 10^{-03}$	$(1.466 \cdot 10^{-03})$	0.041	(0.028)
Figure 5.5, observed yeast data											
A	SLGM+L	0.110	(0.007)	4.098	(0.299)	$7.603 \cdot 10^{-06}$	$(3.206 \cdot 10^{-06})$	$3.457 \cdot 10^{-01}$	$(5.319 \cdot 10^{-02})$	0.113	(0.109)
B	SLGM+N	0.110	(0.003)	3.905	(0.173)	$1.044 \cdot 10^{-05}$	$(3.086 \cdot 10^{-06})$	$1.852 \cdot 10^{-04}$	$(7.460 \cdot 10^{-05})$	0.167	(0.028)
C	RRTR	0.114	(0.026)	3.764	(0.201)	$1.079 \cdot 10^{-05}$	$(3.155 \cdot 10^{-06})$	$3.379 \cdot 10^{-01}$	$(4.840 \cdot 10^{-02})$	0.078	(0.077)
D	LNAM	0.110	(0.011)	3.777	(0.216)	$1.077 \cdot 10^{-05}$	$(3.277 \cdot 10^{-06})$	$3.362 \cdot 10^{-01}$	$(5.137 \cdot 10^{-02})$	0.104	(0.108)
E	LNAA	0.109	(0.003)	3.832	(0.198)	$1.069 \cdot 10^{-05}$	$(3.680 \cdot 10^{-06})$	$1.769 \cdot 10^{-04}$	$(6.607 \cdot 10^{-05})$	0.164	(0.033)

True values		K	r	P	ν	σ
Figures 5.1, panels A, B, C and D		0.11	4	0.00005	N/A	0.05
Figures 5.2 and 5.4, panels A, B, C & D		0.15	3	0.0001	0.005	0.01
Figures 5.2 and 5.4, panels E, F, G and H		0.11	4	0.00005	0.001	0.05
Figures 5.2 and 5.4, panels I, J, K and L		0.3	6	0.0002	0.01	0.02

Table 5.2: Total mean squared error (MSE) for 10 observed yeast growth time courses, each with 1000 forward simulated time-courses with measurement error. Parameter values are taken from posterior samples. Standard Deviations give the variation between the sub-total MSEs for each yeast time course fit (n=10).

Model	SLGM+N	SLGM+L	RRTR	LNAM	LNAA
Total MSE	29.847	100.165	600.601	99.397	30.959
Standard Deviation	1.689	8.391	55.720	9.263	2.030

a Total MSE for each model. It is clear that the RRTR is the worst overall representation of the 10 yeast time courses, with the highest total MSE and a much larger total MSE than the "exact" SLGM+L. It is interesting to see there is a very similar total MSE for the SLGM+L and LNAM, and similarly for the SLGM+N and LNAA, demonstrating that our approximations perform well.

Once the most appropriate approximate stochastic model is chosen, we can incorporate the SDE within our Bayesian hierarchical models described in Section 3. Currently
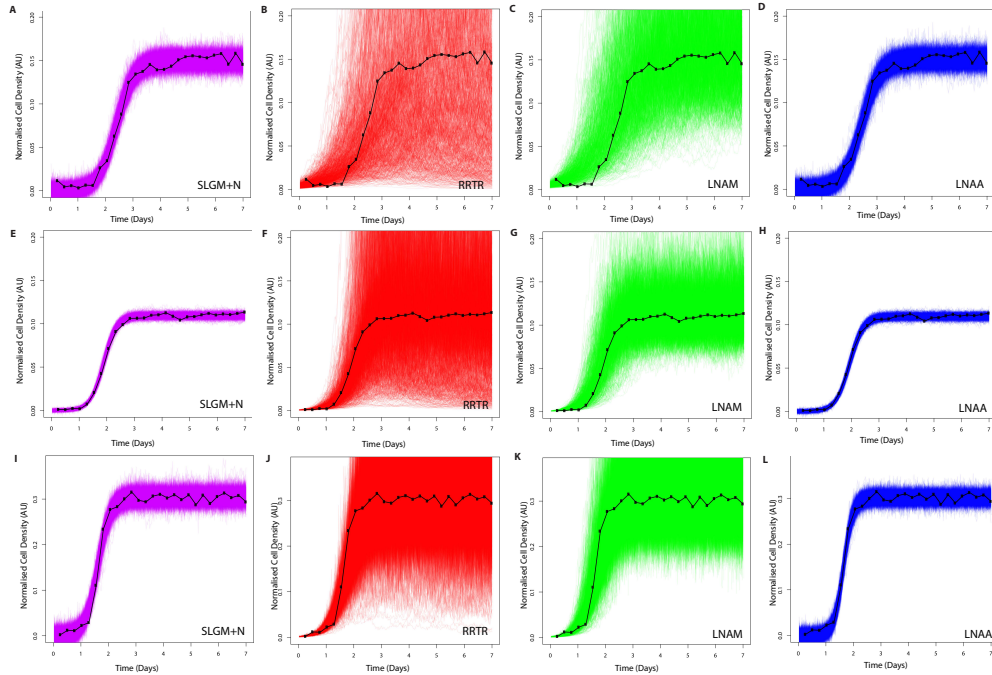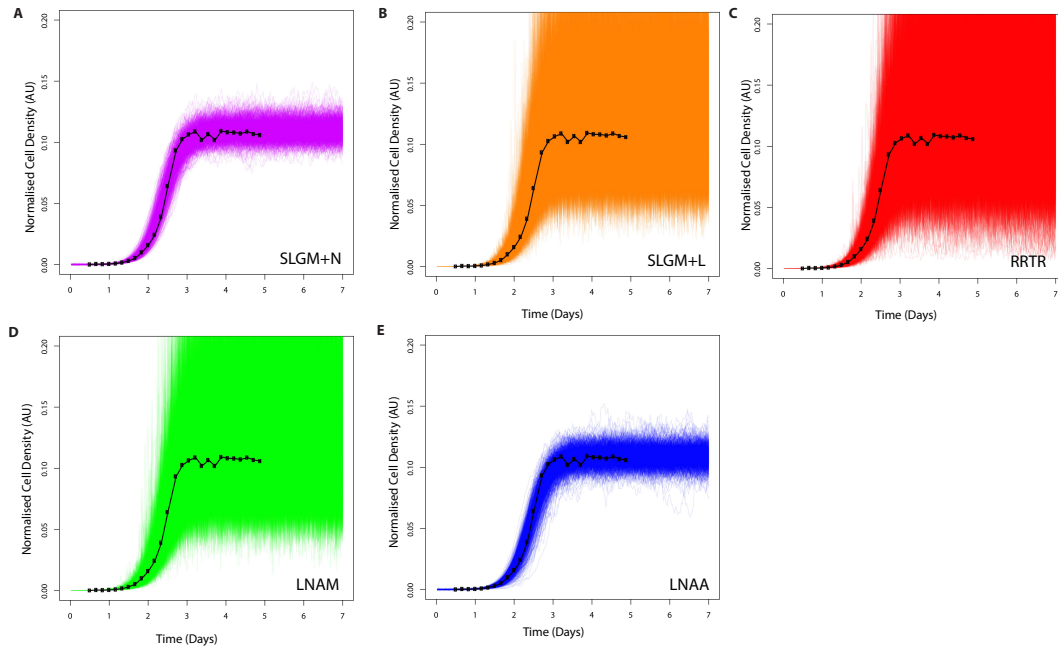
Figure 5.5: Forward trajectories with measurement error, simulated from inferred parameter posterior samples (sample size=1000). Model fitting is carried out on observed yeast time-course data (black). See (5.16) or (5.17) and Table C.1 for prior hyper-parameter values. See Table 5.1 for parameter posterior means. A) SLGM+N (pink). B) SLGM+L (orange). A) RRTR model with Log-normal error (red). B) LNAM model with log-normal error (green). C) LNAA model with Normal error (blue).

the Bayesian hierarchical models described in Section 3 have long computational times, ~2 weeks for the joint hierarchical model (JHM) (~1 week with further optimisations) and so extending these models using slow numerical methods would lead to prohibitively slow computational times that we estimate to take ~3-6 months (with 4294 *orf* Δs, ~8 repeats and ~27 time points). Inference using the Kalman filter will allow the Bayesian hierarchical models to carry out stochastic modelling at a greatly reduced computational time (~10× faster) compared to an arbitrarily exact approach.

Chapter 6. Conclusions and future work

We have joined a hierarchical model of microbial growth with a model for genetic interaction in order to learn about strain fitnesses, evidence for genetic interaction and interaction strengths simultaneously. By introducing Bayesian methodology to QFA we have been able to model the hierarchical nature of the experiment and expand the multiplicative model for genetic interaction to incorporate many sources of variation that previously had to be ignored.

We proposed two new Bayesian hierarchical model approaches to replace the current statistical analysis for identifying genetic interactions within a QFA screen comparison. Both the new two-stage and one-stage approaches give similar results but have different interpretations. The two-stage approach fits the SHM followed by the IHM, with univariate point estimate fitness definitions generated as an intermediate step. The two-stage approach can therefore be regarded as a Bayesian hierarchical version of the Addinall *et al.* (2011) approach. In contrast, the one-stage approach fits the JHM, which does not require a univariate definition of fitness, recognising that fitness is a multi-faceted concept, allowing interaction to be identified by either growth rate (logistic parameter r) or final biomass (logistic parameter K) achievable by a given genotype. Our one-stage approach is a new method of detecting genetic interaction that further develops the interpretation of epistasis within QFA screens.

Hierarchical methods are able to account for the many sources of variation that exist within QFA data by accurately reflecting QFA experimental design, which is known. A hierarchical, frequentist approach using random effects, namely the REM is presented in order to improve on the Addinall *et al.* (2011) approach. Due to the lack of flexibility with modelling assumptions in the standard frequentist modelling paradigm, the REM is unsuitable for modelling the distribution of $orf\Delta$ level variation on a log scale or for simultaneously modelling genetic interaction and logistic growth curves.

The data from which logistic parameter estimates are derived during QFA are the result of a technically challenging, high-throughput experimental procedure with a diverse range of possible technical errors. Our Bayesian, hierarchical models allow us the flexibility to make distributional assumptions that more closely match the data. This allows us to switch between modelling parameter uncertainty with Normal, Log-Normal and Student's t distribution where appropriate.

QFA experimental design is intrinsically multilevel and is therefore more closely modelled in our hierarchical scheme. Consequently the JHM and IHM capture sources of variation not considered by Addinall *et al.* (2011). By sharing information across levels in the hierarchy, our models have allowed us to learn more about *orf*Δs with weaker genetic interaction. Our more flexible model of variance also avoids misclassification of individual genotypes with high variance as having significant interactions. Without fully accounting for the variation described in the Bayesian hierarchical models, the previous Addinall *et al.* (2011) approach may have relatively poor power to detect subtle interactions, obscuring potential novel observations.

Many subtle, interesting genetic interactions may remain to be investigated for the example dataset we present: QFA to understand telomere capping using *cdc13-1*. The JHM is better able to identify subtle interactions (see Figure 4.3). In our two-stage approaches, univariate fitness measures such as $MDR \times MDP$ are used in the intermediate steps, occasionally causing interaction in terms of one parameter to be masked by the other. For example, strains with little evidence for interaction with a background mutation in terms of growth rate but with strong evidence of interaction in terms of carrying capacity are sometimes classified as interactors using the JHM (see Figure 4.3). The JHM has identified genes that have not been identified as showing genetic interaction in the Addinall *et al.* (2011) or two-stage Bayesian analysis, for example *CHZ1*, which is thought to be related to telomere activity (Wan *et al.*, 2011).

As expected, many genes previously unidentified by Addinall *et al.* (2011) have been identified as showing evidence of interaction using both of our Bayesian hierarchical modelling approaches. Some genes which have been identified only by the JHM (see Figure 4.2D), such as those showing interaction only in terms of r, are found to be related to telomere biology in the literature. Currently there is not sufficient information available to identify the proportion of identified interactions that are true hits and so we use unbiased GO term enrichment analyses to confirm that the lists of genetic interactions closely reflect the true underlying biology. GO term annotations relevant to telomere biology are available for well-studied genes in the current literature. Unsurprisingly all of the approaches considered closely reflect the most well-known GO terms (see Table 4.1).

Computational time for the new Bayesian approach ranges from one to two weeks for one of the datasets presented in Addinall *et al.* (2011). This compares favourably with the time taken to design and execute the experimental component of QFA (approximately six weeks). Time and resources used to follow up the results of a QFA screen comparison can be saved with the Bayesian approaches suggested, allowing genes to be

chosen for further investigation with increased confidence. With an improved analysis it may be possible to detect more genetic interactions with the same sample size, allowing us to systematically detect and rank interactions genome-wide. Overall we recommend a JHM or "Bayesian QFA" for analysis of current and future QFA data sets as it accounts for more sources of variation than the Addinall *et al.* (2011) QFA methodology. With the JHM we have outlined new genes with significant evidence of interaction in the *ura3∆* 27°C and *cdc13-1* 27°C experiment. The full lists of genetic interactions for both the two and one stage Bayesian hierarchical approaches as well as lists of significant GO terms are freely available online at `http://research.ncl.ac.uk/qfa/HeydariQFABayes/`. The new Bayesian hierarchical models we present here will also be suitable for identifying new genes showing evidence of genetic interaction in backgrounds other than telomere activity. We hope that further, reductionist lab work by experimental biologists will give additional insight into the mechanisms by which the new genes we have uncovered interact with the telomere.

In this thesis we have also presented two new diffusion processes for modelling logistic growth data where fast inference is required: the linear noise approximation (LNA) of the stochastic logistic growth model (SLGM) with multiplicative noise and the LNA of the SLGM with additive intrinsic noise (labelled as the LNAM and LNAA respectively). Both the LNAM and LNAA are derived from the linear noise approximation of the stochastic logistic growth model (SLGM). The new diffusion processes approximate the SLGM more closely than an alternative approximation (RRTR) proposed by Román-Román & Torres-Ruiz (2012). The RRTR lacks a mean reverting property that is found in the SLGM, LNAM and LNAA, resulting in increasing variance during the stationary phase of population growth (see Figure 5.1).

We compared the ability of each of the three approximate models and the SLGM to recover parameter values from simulated datasets using standard MCMC techniques. When modelling stochastic logistic growth with Log-normal measurement error we find that our approximate models are able to represent data simulated from the original process and that the RRTR is least representative, with large variation over the stationary phase (see Figure 5.2). When modelling stochastic logistic growth with Normal measurement error we find that only our models with Normal measurement error can appropriately bound data simulated from the original process (see Figure 5.4). We also compared parameter posterior distribution summaries with parameter values used to generate simulated data after inference using both approximate and "exact" models (see Table 5.1). We find that,

when using the RRTR model, posterior distributions for the carrying capacity parameter K are less precise than for the LNAM and LNAA approximations. We also note that it is not possible to model additive measurement error while maintaining a linear Gaussian structure (which allows fast inference with the Kalman filter) when carrying out inference with the RRTR. We conclude that when measurement error is additive, the LNAA model is the most appropriate approximate model.

To test model performance during inference with real population data, we fitted our approximate models and the "exact" SLGM to microbial population growth curves generated by quantitative fitness analysis (QFA) (see Figure 5.5). We found that the LNAA model was the most appropriate for modelling experimental data. It seems likely that this is because a Normal error structure best describes this particular dataset, placing the LNAM and RRTR models at a disadvantage. We demonstrate that arbitrarily exact methods and our fast approximations perform similarly during inference for 10 diverse, experimentally observed, microbial population growth curves (see Table 5.2) which shows that, in practise, our fast approximations are as good as "exact" methods. We conclude that our LNA models are preferable to the RRTR for modelling QFA data.

It is interesting to note that, although the LNAA is not a better approximation of the original SGLM process than the LNAM, it is still quite reasonable. Figures 5.1A and 5.1D show that the SLGM and LNAA processes are visually similar. Figure 5.1E demonstrates that forward trajectories of the LNAA also share similar levels of variation over time with the SLGM and LNAM.

Fast inference with the LNAA gives us the potential to develop large hierarchical Bayesian models for genome-wide QFA datasets, using a diffusion equation and realistic computational resources

Here, we have concentrated on a biological model of population growth. However, we expect that the approach we have demonstrated: generating linear noise approximations of stochastic processes to allow fast Bayesian inference with Kalman filtering for marginal likelihood computation, will be useful in a wide range of other applications where simulation is prohibitively slow.

Further work involves extending the Bayesian hierarchical models in Chapter 3 with the approximate stochastic logistic growth models and methods for carrying out inference described in Chapter 5. By accounting for the random fluctuations within the logistic growth data we will be able to improve our logistic growth parameter estimates.

We have demonstrated how to incorporate a batch effect or a transformation effect to

the joint hierarchical model in Section 4.5. Introducing a batch or transformation effect into our models will allow us to capture even further experimental variation. Fitness plots for further case studies given in Section 4.5 and extensions of the joint hierarchical model given in Section 4.6 are included for experimental biologists to investigate further.

A related experiment to the QFA screen comparison analysed within this thesis is the "all-by-all" QFA experiment (in early development at the time of writing). The "all-by-all" QFA experiment begins with a control plate consisting of N $orf\Delta$s. For each of the N $orf\Delta$s a new query plate is created, each query plate consists of the control plate and an additional background mutation related to one of the N $orf\Delta$s. In total there will be $N+1$ unique plates (including the control plate). Where a standard QFA comparison looks for genes that interact with a single query mutation (or condition), the "all-by-all" QFA experiment aims to find genetic interactions for multiple query mutations (N) simultaneously. The "all-by-all" experiment therefore incorporates more information and investigates more potential genetic interactions than a standard QFA comparison. We expect that the Bayesian hierarchical modelling and genetic interaction modelling developed in this thesis will be used to create models for describing the "all-by-all" QFA experiment as well as many other similar experiments in the future.

By improving our software we may be able to reduce computational time for inference. Currently the code for implementing the Bayesian models described in this thesis is written in the C programming language which can be run as standalone software or through an R package "qfaBayes", available at `https://r-forge.r-project.org/projects/qfa`. The computational speed of our C code used for inference could be improved by parallel implementation, taking advantage of a multi-core processor to carry out tasks simultaneously. With faster computational times we expect to reduce the time for a typical QFA comparison with the JHM from $\sim$2 weeks to less than a week.

Currently the information available on true genetic interactions and biological processes in yeast is limited and so we rely on objective analyses such as simulation studies to give unbiased comparisons between the approaches considered. The biological processes of many genes in the yeast genome are yet to be identified so we are unable to use GO term enrichment analysis as a "gold standard" for comparing the results of our approaches. Information used to build a gene ontology is typically well known and taken from well understood experiments, we expect that subtle genetic interactions which we are interested in finding will have little information available. QFA screen comparisons are designed to learn biology which is not already fully understood and so a biological comparison between the different approaches considered is difficult. Simulation studies

(see Section 4.3.6) give us the ability to compare the different approaches and the effects of modelling more experimental structure.

A typical QFA comparison is a large and complex data set corresponding to around 400,000 time series, posing considerable computational, as well as statistical challenges. With a Bayesian approach we are able to evaluate complex hierarchical models to better reflect the structure or design of genome-wide QFA experiments. Bayesian variable selection methods embedded within a large hierarchical model allow us to describe genetic interaction and use prior information to incorporate physical and biological constraints within our models. We have shown that Bayesian hierarchical modelling of large and complex data gives us the advantage of increased modelling flexibility compared to a frequentist approach, allowing us to better describe the experimental structure or design. For the reasons above, a QFA screen comparison or any other highly structured experimental dataset is better modelled using a Bayesian hierarchical modelling approach when compared to an alternative frequentist approach.

Overall this thesis presents improved modelling approaches to the current non-hierarchical frequentist approach for a QFA screen comparison. The research contained in this thesis illustrates how Bayesian inference gives us further modelling flexibility, allowing us to better describe the known experimental structure. Further, our modelling approaches and assumptions are transferable outside QFA screen experiments where we wish to capture as much experimental structure as possible. The results from our temperature sensitive *cdc13-1* QFA experiment results will give further insight to the telomere and consequentially aging and cancer in yeast and potentially the human genome (Botstein *et al.*, 1997).

Appendix A. QFA data set sample, solving the logistic growth model and random effects model R code

A.1. *cdc13-1* Quantitative Fitness Analysis data set sample

```
                  Image.Name Row Col X.Offset Y.Offset Area Trimmed Threshold Intensity Edge.Pixels Colony.Color.R Colony.Color.G Colony.Color.B
1 DLR00012888-2009-10-20_08-14-14   3  22     3162     477  250    1595  94.51286     36215           8             74             93             99
2 DLR00012888-2009-10-20_12-45-39   3  22     3147     519  767    5515  85.42326     34187          10             68             86             93
3 DLR00012888-2009-10-20_16-09-30   3  22     3165     510  507    4868  84.43037     32764           4             69             86             93
4 DLR00012888-2009-10-21_08-42-49   3  22     3144     490 1195   37605  96.04073     58816           1            100            117            115
5 DLR00012888-2009-10-21_15-59-40   3  22     3145     489 2258  112046  99.54796    127655           3            118            131            128
6 DLR00012888-2009-10-22_07-58-08   3  22     3150     481 6291  495068  98.88876    504778           0            142            150            140

  Background.Color.R Background.Color.G Background.Color.B Edge.length Tile.Dimensions.X Tile.Dimensions.Y      Growth      Barcode
1                 71                 90                 96         759               137               138 0.0003308422 DLR00012888
2                 64                 82                 90         554               137               138 0.0011439464 DLR00012888
3                 62                 80                 89         570               137               138 0.0010097427 DLR00012888
4                 72                 91                 95         948               137               138 0.0078002004 DLR00012888
5                 72                 90                 95        1186               137               138 0.0232410916 DLR00012888
6                 69                 84                 89         947               137               138 0.1026892593 DLR00012888

         Date.Time        Inoc.Time Treatments                                                            Medium
1 2009-10-20_08-14-14 2009-10-19_17-30-00        27 SD/MSG -ARG -HIS -LEU -LYS +Canavanine +clonNAT +G418 +Hygromycin +Thialysine
2 2009-10-20_12-45-39 2009-10-19_17-30-00        27 SD/MSG -ARG -HIS -LEU -LYS +Canavanine +clonNAT +G418 +Hygromycin +Thialysine
3 2009-10-20_16-09-30 2009-10-19_17-30-00        27 SD/MSG -ARG -HIS -LEU -LYS +Canavanine +clonNAT +G418 +Hygromycin +Thialysine
4 2009-10-21_08-42-49 2009-10-19_17-30-00        27 SD/MSG -ARG -HIS -LEU -LYS +Canavanine +clonNAT +G418 +Hygromycin +Thialysine
5 2009-10-21_15-59-40 2009-10-19_17-30-00        27 SD/MSG -ARG -HIS -LEU -LYS +Canavanine +clonNAT +G418 +Hygromycin +Thialysine
6 2009-10-22_07-58-08 2009-10-19_17-30-00        27 SD/MSG -ARG -HIS -LEU -LYS +Canavanine +clonNAT +G418 +Hygromycin +Thialysine

         Screen.Name RepQuad MasterPlate.Number Timeseries.order Library.Name   ORF Gene  Background Expt.Time              ID
1 cdc13-1 rad9D SDLv2 Rpt4       4                  1                1 BooneSDLv2 YAL002W VPS8 CDC13-1RAD9 0.6140509 DLR0001288810322
2 cdc13-1 rad9D SDLv2 Rpt4       4                  1                2 BooneSDLv2 YAL002W VPS8 CDC13-1RAD9 0.8025347 DLR0001288810322
3 cdc13-1 rad9D SDLv2 Rpt4       4                  1                3 BooneSDLv2 YAL002W VPS8 CDC13-1RAD9 0.9440972 DLR0001288810322
4 cdc13-1 rad9D SDLv2 Rpt4       4                  1                4 BooneSDLv2 YAL002W VPS8 CDC13-1RAD9 1.6339005 DLR0001288810322
5 cdc13-1 rad9D SDLv2 Rpt4       4                  1                5 BooneSDLv2 YAL002W VPS8 CDC13-1RAD9 1.9372685 DLR0001288810322
6 cdc13-1 rad9D SDLv2 Rpt4       4                  1                6 BooneSDLv2 YAL002W VPS8 CDC13-1RAD9 2.6028704 DLR0001288810322
```

Figure A.1: *cdc13-1* Quantitative Fitness Analysis data set sample. Notable columns include "ORF", "Expt.Time" and "Growth". "ORF" indicates which *orf*Δ strain the row corresponds to. "Expt.Time" indicates the time in days from the *orf*Δ strain being spotted (Addinall *et al.*, 2011). "Growth" gives an adjusted measure of cell culture density from the image analysis for a given *orf*Δ strain and time point. Generated from Colonyzer output files with the qfa R package, freely available at `http://qfa.r-forge.r-project.org/`.

A.2. Solving the logistic growth model

The solution to the logistic growth ODE (1.1) can be obtained as follows. First we factor the right side of (1.1) and rearrange to give:

$$\frac{dx(t)}{x(t)\left(1 - \frac{x(t)}{K}\right)} = rdt.$$

We now rearrange further using a partial fractions expansion and integrate over both sides of the equation:

$$\int \frac{dx(t)}{x(t)} + \frac{\frac{1}{K}dx(t)}{\left(1 - \frac{x(t)}{K}\right)} = \int rdt. \tag{A.1}$$

Integrating the first component on the left side of (A.1) we obtain the following, where c_1 is an unknown constant:

$$\int \frac{dx(t)}{x(t)} = \log(x(t)) + c_1.$$

Integrating the second component on the left side of (A.1) we obtain the following, where c_2 is an unknown constant:

$$\frac{1}{K} \int \frac{dx(t)}{1 - \frac{x(t)}{K}} = -\log(1 - \frac{x(t)}{K}) + c_2.$$

Integrating the right side of (A.1) we obtain the following, where c_3 is an unknown constant:

$$\int rdt = rt + c_3.$$

Solving the integrals in (A.1) we obtain the following, where $c_4 = c_3 - c_1 - c_2$ is an unknown constant:

$$\log\left(\frac{x(t)}{1 - \frac{x(t)}{K}}\right) = rt + c_4.$$

Rearranging our equation, we obtain the following:

$$\frac{x(t)}{1 - \frac{x(t)}{K}} = e^{rt+c_4}.$$

We now apply initial conditions, $P = x_0$ and rearrange to obtain an expression for c_4:

$$c_4 = \log\left(\frac{P}{1 - \frac{P}{K}}\right).$$

We now substitute in our expression for c_4 to give:

$$\log\left(\frac{x(t)}{1 - \frac{x(t)}{K}}\right) = rt + \log\left(\frac{P}{1 - \frac{P}{K}}\right)$$

Finally, we rearrange to give (1.2).

A.3. Random effects model R code

```
library(lme4) #http://cran.r-project.org/web/packages/lme4/index.html
#http://research.ncl.ac.uk/colonyzer/AddinallQFA/Logistic.zip and extract zip file
#alternatively http://research.ncl.ac.uk/colonyzer/AddinallQFA/
#"Table S8 Logistic Output Files - 36MB .zip file"
aa<-read.delim("cSGA_v2_r1_Logistic.txt",header=T,skip=1,sep="\t")
#...
bb<-read.delim("Adam_cdc13-1_SDLV2_REP1_Logistic.txt",header=T,skip=0,sep="\t")
#...
aa<-aa[aa$Treatments==27,]
bb<-bb[bb$Treatments==27,]
aa<-aa[!aa$Row==1,]
aa<-aa[!aa$Row==16,]
aa<-aa[!aa$Col==1,]
aa<-aa[!aa$Col==24,]
bb<-bb[!bb$Row==1,]
bb<-bb[!bb$Row==16,]
bb<-bb[!bb$Col==1,]
bb<-bb[!bb$Col==24,]


ORFuni=ORFuni_a=unique(aa$ORF)
ORFuni_b=unique(bb$ORF)
L=length(ORFuni_a)
NoORF_a=NoORF_b=aaa=bbb=numeric()
for (i in 1:L){
 NoORF_a[i]=nrow(aa[aa$ORF==ORFuni[i],])
 NoORF_b[i]=nrow(bb[bb$ORF==ORFuni[i],])
 aaa<-rbind(aaa,aa[aa$ORF==ORFuni[i],])
 bbb<-rbind(bbb,bb[bb$ORF==ORFuni[i],])
}
a=b=numeric(0)
K_lm=aaa$Trimmed.K
P_a=43
r_lm=aaa$Trimmed.r
for (i in 1:length(r_lm)){
 if(K_lm[i]<=2*P_a){K_lm[i]=2*P_a+0.01;r_lm[i]=0;}
 a[i]=(r_lm[i]/log(2*max(0,K_lm[i]-P_a)/max(0,K_lm[i]-2*P_a)))*(log(K_lm[i]/P_a)/log(2));
}
K_lmb=bbb$Trimmed.K
P_b=43
r_lmb=bbb$Trimmed.r
for (i in 1:length(r_lmb)){
 if(K_lmb[i]<=2*P_b){K_lmb[i]=2*P_b+0.01;r_lmb[i]=0;}
 b[i]=(r_lmb[i]/log(2*max(0,K_lmb[i]-P_b)/max(0,K_lmb[i]-2*P_b)))*(log(K_lmb[i]/P_b)/log(2));
}

condition<-factor(c(rep("a",length(a)),rep("b",length(b))))
subject=numeric()
for (i in 1:L){
 subject=c(subject,rep(i,NoORF_a[i]))
}
for (i in 1:L){
 subject=c(subject,rep(i,NoORF_b[i]))
}
subcon=subject
subcon[1:length(a)]=0
subcon<-factor(subcon)
subject<-factor(subject)
f=c(a,b)
data=data.frame(f,subject,condition,subcon)
data$lf=log(data$f+1)
data$subcon<-C(data$subcon,sum)
bk<-contrasts(data$subcon)
contrasts(data$subcon)=bk[c(nrow(contrasts(data$subcon)),1:(nrow(contrasts(data$subcon))-1)),]
model1<-lmer(lf~subcon+(1|subject),data=(data),REML=F)
```

Appendix B. Bayesian hierarchical modelling

B.1. Hyper-parameter values for Bayesian hierarchical modelling

Table B.1: Hyper-parameter values for Bayesian hierarchical modelling of quantitative fitness analysis data. Hyper-parameter values for the separate hierarchical model (SHM), interaction hierarchical model (IHM) and joint hierarchical model (JHM) are provided.

SHM & JHM		SHM & JHM		JHM		IHM		JHM-B & JHM-T	
Parameter Name	Value	Parameter Name	Value	Parameter Name	Value	Parameter Name	Value	Parameter Name	Value
$\tau^{K,\mu}$	2.20	$\eta^{r,p}$	0.13	α^{μ}	0.00	Z_{μ}	3.66	κ^{p}	0.00
$\eta^{\tau,K,p}$	0.02	ν^{μ}	19.82	η^{α}	0.25	$\eta^{Z,p}$	0.70	η^{κ}	1.17
$\eta^{K,o}$	-0.79	$\eta^{\nu,p}$	0.02	β^{μ}	0.00	η^{Z}	0.10	λ^{p}	0.00
$\psi^{K,o}$	0.61	P^{μ}	-9.04	η^{β}	0.25	ψ^{Z}	0.42	η^{λ}	1.17
$\tau^{r,\mu}$	3.65	η^{P}	0.47	p	0.05	η^{ν}	0.10	ϕ^{shape}	100.00
$\eta^{\tau,r,p}$	0.02			η^{γ}	-0.79	ψ^{ν}	2.45	ϕ^{scale}	0.01
$\eta^{r,o}$	0.47			ψ^{γ}	0.61	ν^{μ}	2.60	χ^{shape}	100.00
$\psi^{r,o}$	0.10			η^{ω}	0.47	$\eta^{\nu,p}$	0.05	χ^{scale}	0.01
η^{ν}	-0.83			ψ^{ω}	0.10	α^{μ}	0.00		
ψ^{ν}	0.86			$\eta^{\tau,K}$	2.20	η^{α}	0.31		
K^{μ}	-2.01			$\psi^{\tau,K}$	0.02	p	0.05		
$\eta^{K,p}$	0.03			$\eta^{\tau,r}$	3.65	η^{γ}	0.10		
r^{μ}	0.97			$\psi^{\tau,r}$	0.02	ψ^{γ}	0.42		

B.2. *cdc13-1* 27°C vs *ura3Δ* 27°C fitness plots with gene ontology terms highlighted

Figure B.1: Alternative fitness plots with *orf Δ* posterior mean fitnesses. Labels for the "telomere maintenance" gene ontology term are highlighted in blue. A) Non-Bayesian, non-hierarchical fitness plot, based on Table S6 from Addinall et al. (2011) ($F = MDR \times MDP$). B) Non-Bayesian, hierarchical fitness plot, from fitting REM to data in Table S6 from Addinall et al. (2011) ($F = MDR \times MDP$). C) IHM fitness plot with *orf Δ* posterior mean fitness ($F = MDR \times MDP$). D) JHM fitness plot with *orf Δ* posterior mean fitnesses. *orf Δ* strains are classified as being a suppressor or enhancer based on analysis of growth parameter r. Further fitness plot explanation and notation is given in Figure 4.2.
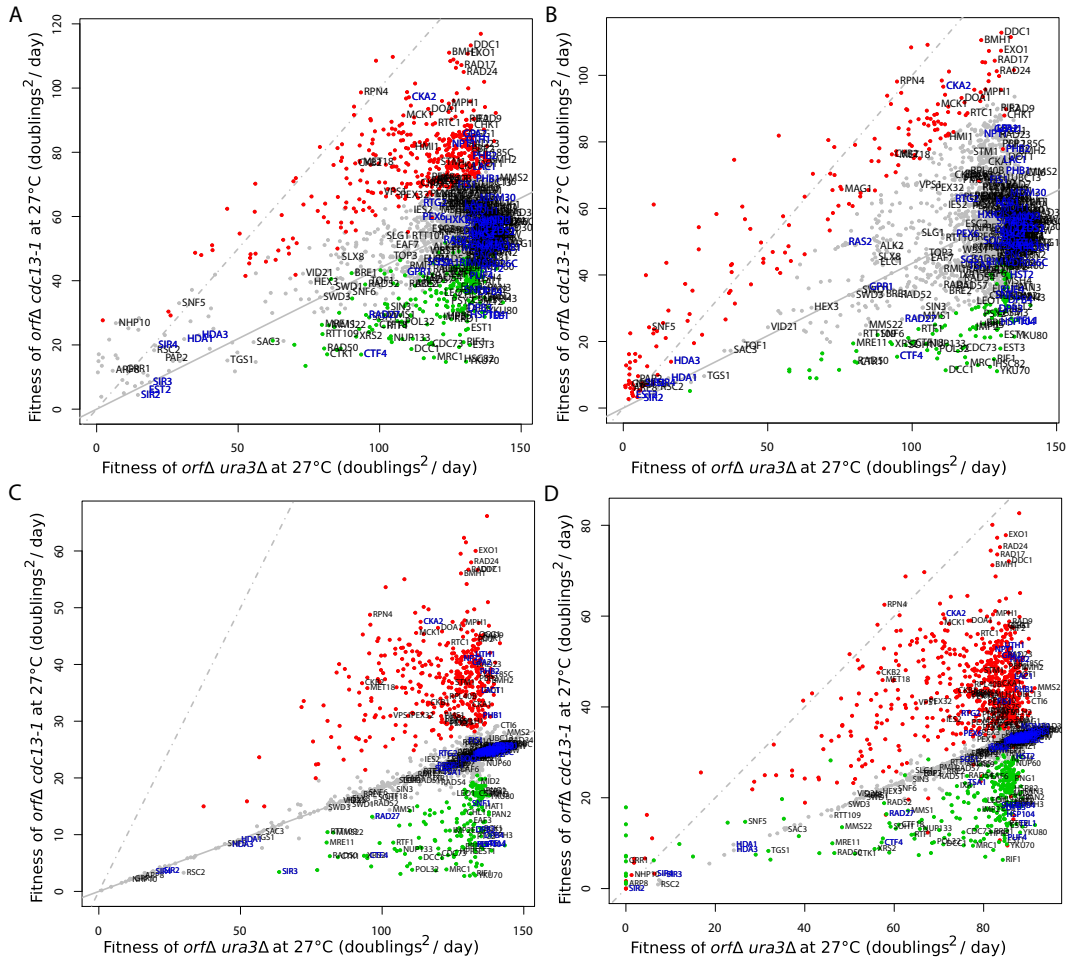
Figure B.2: Alternative fitness plots with *orfΔ* posterior mean fitnesses. Labels for the "ageing" gene ontology term are highlighted in blue. A) Non-Bayesian, non-hierarchical fitness plot, based on Table S6 from Addinall et al. (2011) ($F = MDR \times MDP$). B) Non-Bayesian, hierarchical fitness plot, from fitting REM to data in Table S6 from Addinall et al. (2011) ($F = MDR \times MDP$). C) IHM fitness plot with *orfΔ* posterior mean fitness ($F = MDR \times MDP$). D) JHM fitness plot with *orfΔ* posterior mean fitnesses. *orfΔ* strains are classified as being a suppressor or enhancer based on analysis of growth parameter r. Further fitness plot explanation and notation is given in Figure 4.2.
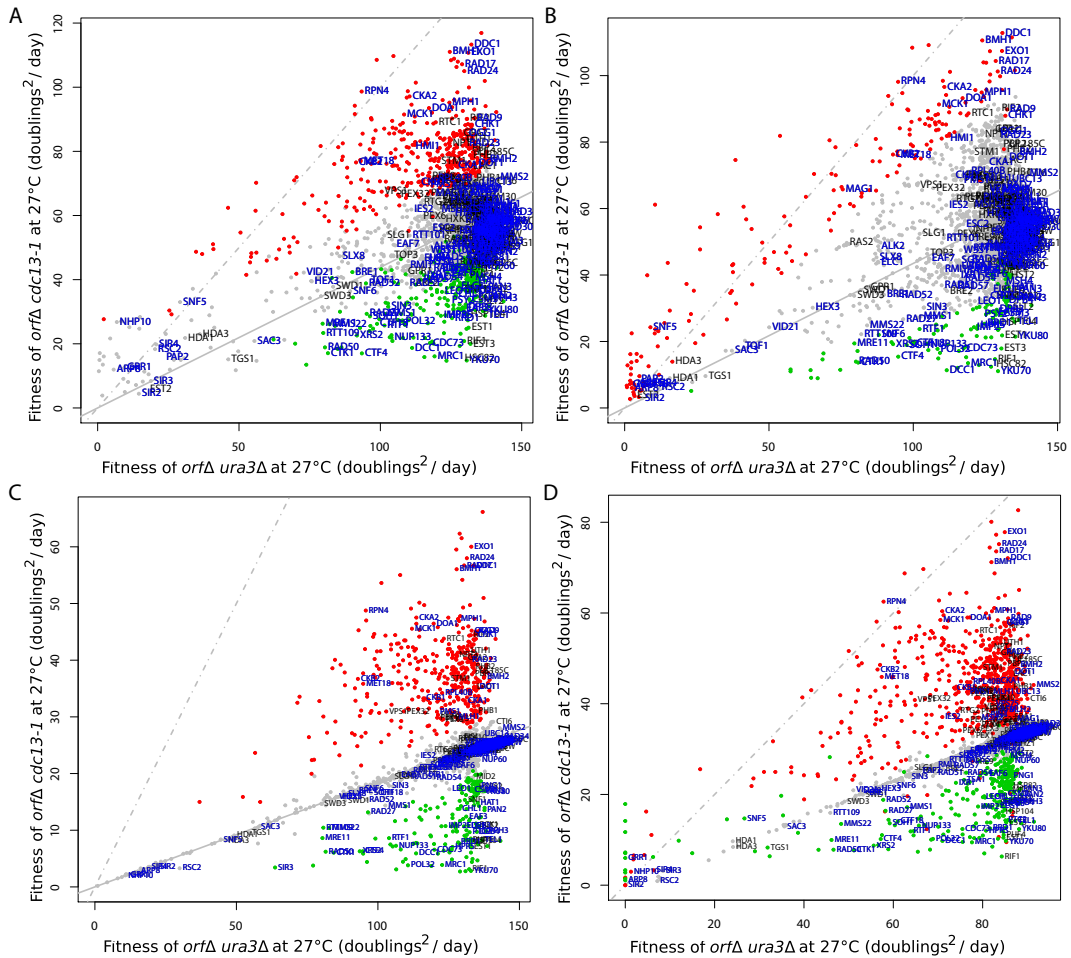
Figure B.3: Alternative fitness plots with *orf*Δ posterior mean fitnesses. Labels for the "response to DNA damage" gene ontology term are highlighted in blue. A) Non-Bayesian, non-hierarchical fitness plot, based on Table S6 from Addinall et al. (2011) ($F = MDR \times MDP$). B) Non-Bayesian, hierarchical fitness plot, from fitting REM to data in Table S6 from Addinall et al. (2011) ($F = MDR \times MDP$). C) IHM fitness plot with *orf*Δ posterior mean fitness ($F = MDR \times MDP$). D) JHM fitness plot with *orf*Δ posterior mean fitnesses. *orf*Δ strains are classified as being a suppressor or enhancer based on analysis of growth parameter r. Further fitness plot explanation and notation is given in Figure 4.2.
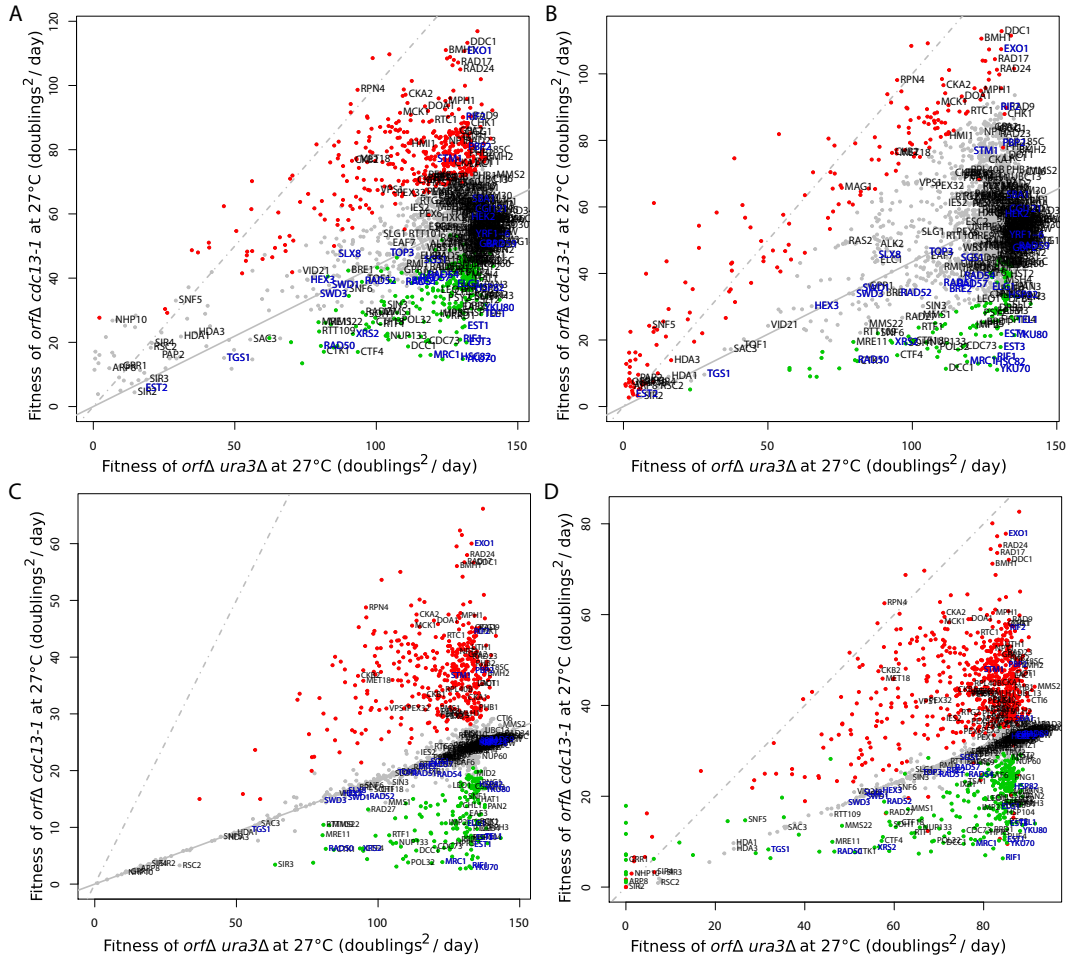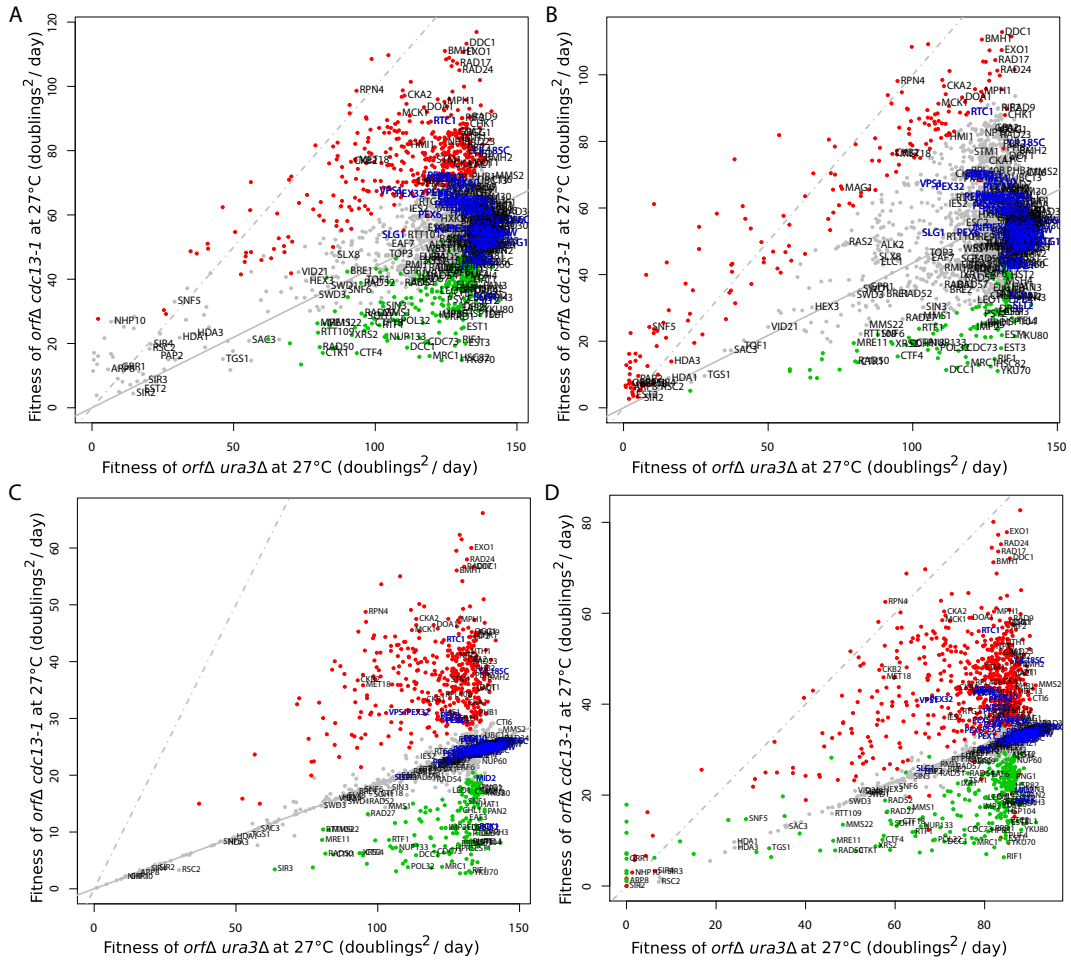
Figure B.4: Alternative fitness plots with *orf∆* posterior mean fitnesses. Labels for the "peroxisomal organisation" gene ontology term are highlighted in blue. A) Non-Bayesian, non-hierarchical fitness plot, based on Table S6 from Addinall et al. (2011) ($F = MDR \times MDP$). B) Non-Bayesian, hierarchical fitness plot, from fitting REM to data in Table S6 from Addinall et al. (2011) ($F = MDR \times MDP$). C) IHM fitness plot with *orf∆* posterior mean fitness ($F = MDR \times MDP$). D) JHM fitness plot with *orf∆* posterior mean fitnesses. *orf∆* strains are classified as being a suppressor or enhancer based on analysis of growth parameter r. Further fitness plot explanation and notation is given in Figure 4.2.

B.3. Lists of top genetic interactions for the two-stage and one-stage Bayesian approaches

Table B.2: Sample of interaction hierarchical model top genetic interactions with *cdc13-1* at **27°C**

Type of Interaction	Gene Name	Probability of Interaction δ_l	Strength of Interaction $e^{(\delta_l \gamma_l)}$	Position in Addinall (2011)
Suppressor	IPK1	1.00	2.87	10
	LST4	1.00	2.77	13
	RPN4	1.00	2.76	17
	MTC5	1.00	2.66	20
	GTR1	1.00	2.64	38
	NMD2	1.00	2.62	3
	SAN1	1.00	2.62	16
	UPF3	1.00	2.58	21
	RPL37A	1.00	2.56	121
	NAM7	1.00	2.53	22
	RPP2B	1.00	2.52	120
	YNL226W	0.99	2.49	126
	YGL218W	1.00	2.46	250
	MEH1	1.00	2.45	45
	ARO2	1.00	2.45	68
	EXO1	1.00	2.45	1
	BUD27	1.00	2.43	46
	RAD24	1.00	2.39	4
	RPL16B	1.00	2.39	33
	RPL43A	1.00	2.39	150
Enhancer	:::MRC1	1.00	0.11	35
	YKU70	1.00	0.11	31
	STI1	1.00	0.11	42
	RIF1	1.00	0.13	36
	ELP3	1.00	0.16	82
	CLB5	1.00	0.17	58
	MRC1	1.00	0.17	63
	DPH2	1.00	0.18	24
	POL32	1.00	0.19	113
	MAK31	1.00	0.19	37
	SWM1	1.00	0.20	25
	LTE1	1.00	0.21	48
	MAK10	1.00	0.22	44
	ELP2	1.00	0.22	77
	PAT1	1.00	0.24	144
	DPH1	1.00	0.25	55
	SRB2	0.99	0.25	174
	THP2	1.00	0.26	67
	MFT1	1.00	0.26	52
	LSM6	0.97	0.26	389

See `http://research.ncl.ac.uk/qfa/HeydariQFABayes/IHM_strip.txt` for the full list.

Table B.3: Sample of joint hierarchical model top genetic interactions with *cdc13-1* at **27°**C

Type of Interaction	Gene Name	Probability of Interaction δ_l	Strength of Interaction $e^{(\delta_l \gamma_l)}$	Strength of Interaction $e^{(\delta_l \omega_l)}$	Strength of Interaction $MDR \times MDP$	Position in Addinall (2011)
Suppressor	CSE2	1.00	490.51	0.48	11.71	838
in K	SGF29	1.00	273.69	0.68	14.16	580
	GSH1	1.00	78.79	0.92	17.89	281
	YMD8	1.00	59.31	0.65	7.05	2022
	YGL024W	1.00	28.13	1.18	13.33	151
	RPS9B	1.00	24.67	1.12	10.24	801
	GRR1	1.00	22.51	0.67	5.99	1992
Suppressor	BTS1	1.00	19.27	2.29	19.65	201
in r	IPK1	1.00	5.56	2.26	44.81	10
	NMD2	1.00	2.96	2.19	48.51	3
	SAN1	1.00	2.37	2.17	48.70	16
	LST4	1.00	5.79	2.14	44.14	13
	RPN4	1.00	8.00	2.12	40.46	17
	UPF3	1.00	3.16	2.07	45.25	21
Suppressor in	SAN1	1.00	2.37	2.17	48.70	16
$MDR \times MDP$	NMD2	1.00	2.96	2.19	48.51	3
	UPF3	1.00	3.16	2.07	45.25	21
	EXO1	1.00	2.89	2.06	45.04	1
	IPK1	1.00	5.56	2.26	44.81	10
	LST4	1.00	5.79	2.14	44.14	13
	NAM7	1.00	3.02	2.04	43.00	22
Enhancer	YKU70	1.00	0.01	1.09	−23.44	31
in K	STI1	1.00	0.01	1.20	−21.60	42
	RIF1	1.00	0.01	0.63	−26.17	36
	:::MRC1	1.00	0.01	0.83	−23.15	35
	MAK31	1.00	0.02	1.18	−18.19	37
	CLB5	1.00	0.02	0.87	−19.54	58
	MRC1	1.00	0.02	0.81	−20.40	63
Enhancer	PAT1	1.00	1.71	0.28	−18.30	144
in r	PUF4	1.00	2.00	0.31	−21.61	34
	YKU80	1.00	2.15	0.33	−21.68	32
	RTT103	1.00	2.54	0.34	−17.87	153
	LSM1	0.99	2.13	0.34	−16.20	101
	GIM3	0.99	0.93	0.35	−19.70	132
	INP52	0.96	0.86	0.36	−14.50	345
Enhancer in	RIF1	1.00	0.01	0.63	−26.17	36
$MDR \times MDP$	LTE1	1.00	0.06	0.40	−23.96	48
	YKU70	1.00	0.01	1.09	−23.44	31
	:::MRC1	1.00	0.01	0.83	−23.15	35
	DPH2	1.00	0.04	0.56	−23.11	24
	EST1	1.00	0.12	0.46	−22.20	5
	MAK10	1.00	0.04	0.59	−21.92	44

See `http://research.ncl.ac.uk/qfa/HeydariQFABayes/JHM_strip.txt` for the full list.

B.4. *cdc13-1* 27°C vs *ura3Δ* 27°C fitness plots for the joint hierarchical model in terms of carrying capacity and growth rate parameters

Figure B.5: Joint hierarchical model (JHM) carrying capacity fitness plot with *orfΔ* posterior mean fitnesses. *orfΔ* strains are classified as being a suppressor or enhancer based on carrying capacity parameter K. Significant interactors have posterior probability $\Delta > 0.5$. To compare fitness plots, labelled genes are those belonging to the following gene ontology terms in Table 4.1: "telomere maintenance", "ageing", "response to DNA damage stimulus" or "peroxisomal organization", as well as the genes identified as interactions only in K with the JHM (see Figure 4.3) (blue), genes interacting only in r with the JHM (cyan) and the MRX complex genes (pink).
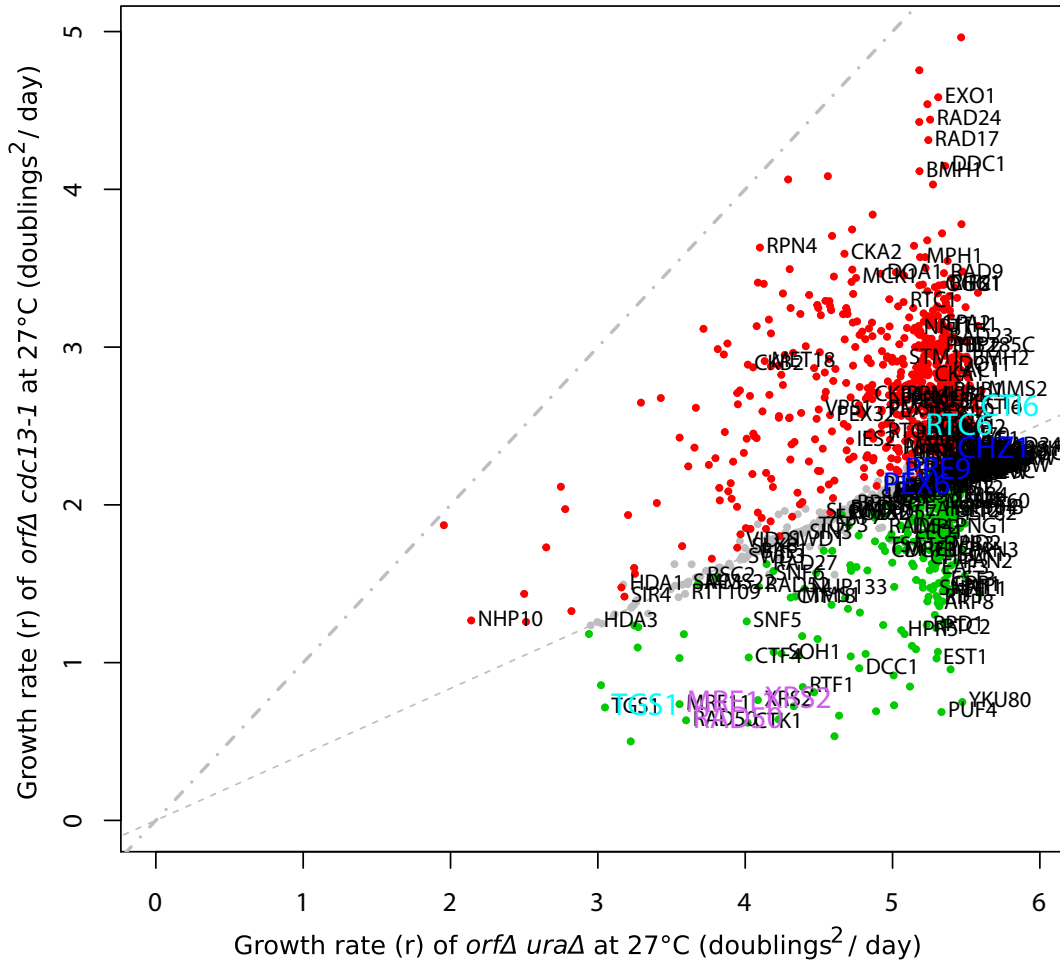
Figure B.6: Joint hierarchical model (JHM) growth rate fitness plot with *orf*Δ posterior mean fitnesses. *orf*Δ strains are classified as being a suppressor or enhancer based on growth parameter r. Significant interactors have posterior probability $\Delta > 0.5$. To compare fitness plots, labelled genes are those belonging to the following gene ontology terms in Table 4.1: "telomere maintenance", "ageing", "response to DNA damage stimulus" or "peroxisomal organization", as well as the genes identified as interactions only in K with the JHM (see Figure 4.3) (blue), genes interacting only in r with the JHM (cyan) and the MRX complex genes (pink).

B.5. Gene ontology term enrichment analysis in R

```
source("http://bioconductor.org/biocLite.R")
biocLite("GOstats")
biocLite("org.Sc.sgd.db")
###################
library(GOstats) # GO testing tool package
library(org.Sc.sgd.db) # yeast gene annotation package
genes=read.table("JHM_strip.txt", header=T)
UNIVSTRIP=genes[,2]
genes<-as.vector(genes[genes[,3]>0.5,2])
genes<-unique(genes)
ensemblIDs=as.list(org.Sc.sgdPMID2ORF)
univ=unlist(ensemblIDs)
univ=univ[!is.na(univ)]
length(univ)
length(unique(univ))
univ=unique(univ)
all=as.vector(univ)
all=all[all%in%UNIVSTRIP]
length(all)
ontology=c("BP")
vec<-genes%in%univ
genes<-genes[vec]
params_temp=new("GOHyperGParams", geneIds=genes,
universeGeneIds=all,
 annotation="org.Sc.sgd.db", categoryName="GO",
 ontology=ontology, pvalueCutoff=1,
 testDirection = "over")
results=hyperGTest(params_temp)
results=summary(results)
results$qvalue<-p.adjust(results$Pvalue,method="BH")
```

B.6. Code for Just Another Gibbs Sampler software

B.6.1. Separate hierarchical model code

```
model {
for (l in 1:N){
 for (m in 1:NoORF[l]){
  for (n in 1:NoTime[(NoSum[l]+m)]){
   y[m,n,l] ~ dnorm(y.hat[m,n,l], exp(nu_l[l]))
   y.hat[m,n,l] <- (K_lm[(NoSum[l]+m)]
     *P*exp(r_lm[(NoSum[l]+m)]*x[m,n,l]))
     /(K_lm[(NoSum[l]+m)]+P*(exp(r_lm[(NoSum[l]+m)]*x[m,n,l])-1))
  }
  K_lm[(NoSum[l]+m)]<- exp(K_lm_L[(NoSum[l]+m)])
  K_lm_L[(NoSum[l]+m)] ~ dnorm(K_o_l_L[l],exp(tau_K_l[l]))T(,0)
  r_lm[(NoSum[l]+m)]<- exp(r_lm_L[(NoSum[l]+m)])
  r_lm_L[(NoSum[l]+m)] ~ dnorm(r_o_l_L[l],exp(tau_r_l[l]))T(,3.5)
 }
 K_o_l_L[l]<- log(K_o_l[l])
 K_o_l[l] ~ dt( exp(K_p), exp(sigma_K_o),3)T(0,)
 r_o_l_L[l]<- log(r_o_l[l])
 r_o_l[l] ~ dt( exp(r_p), exp(sigma_r_o),3)T(0,)
 nu_l[l] ~ dnorm(nu_p,  exp(sigma_nu) )
 tau_K_l[l]~dnorm(tau_K_p,exp(sigma_tau_K))T(0,)
 tau_r_l[l]~dnorm(tau_r_p,exp(sigma_tau_r))
}
K_p ~ dnorm(K_mu,eta_K_p)
r_p ~ dnorm(r_mu,eta_r_p)
nu_p ~ dnorm(nu_mu,eta_nu_p)
P<-exp(P_L)
P_L ~ dnorm(P_mu,eta_P)
tau_K_p ~ dnorm(tau_K_mu,eta_tau_K_p)
sigma_tau_K ~ dnorm(eta_tau_K,psi_tau_K)
tau_r_p ~ dnorm(tau_r_mu,psi_tau_r)
sigma_tau_r ~ dnorm(eta_tau_r,psi_tau_r)
sigma_nu~dnorm(eta_nu,psi_nu)
sigma_K_o ~ dnorm(eta_K_o,psi_K_o)
sigma_r_o ~ dnorm(eta_r_o,psi_r_o)
}
```

B.6.2. Interaction hierarchical model code

```
model {
for (l in 1:N){
 for (c in 1:2){
  for (m in 1:NoORF[l,c]){
   y[m,c,l]~ dnorm(exp(alpha_c[c]
     +delta_l[l,c]*gamma_cl_L[l,c])*Z_l[l],exp(nu_cl[l+(c-1)*N]))
  }
  nu_cl[l+(c-1)*N]~dnorm(nu_p,exp(sigma_nu))
 }
 Z_l[l]~dt(exp(Z_p),exp(sigma_Z),3)T(0,)
 delta_l[l,1]<-0
 delta_l[l,2]~dbern(p)
 gamma_cl_L[l,1]<-0
 gamma_cl_L[l,2]<-log(gamma_l[l])
 gamma_l[l]~dt(1,exp(sigma_gamma),3)T(0,)
}
alpha_c[1]<-0
alpha_c[2]~dnorm(alpha_mu,eta_alpha)
Z_p~dnorm(Z_mu,eta_Z_p)
nu_p~dnorm(nu_mu,eta_nu_p)
sigma_Z~dnorm(eta_Z,psi_Z)
sigma_nu~dnorm(eta_nu,psi_nu_p)
sigma_gamma~dnorm(eta_gamma,psi_gamma)
}
```

B.6.3. Joint hierarchical model code

```
model {
for (l in 1:N){
 for (c in 1:2){
  for (m in 1:NoORF[l,c]){
   for (n in 1:NoTime[NoSum[l,c]+m,c]){
    y[m,n,l,c] ~ dnorm(y.hat[m,n,l,c],exp(nu_cl[l+(c-1)*N]))
    y.hat[m,n,l,c] <- (K_clm[(SHIFT[c]+NoSum[l,c]+m)]
     *P*exp(r_clm[(SHIFT[c]+NoSum[l,c]+m)]*x[m,n,l,c]))
     /(K_clm[(SHIFT[c]+NoSum[l,c]+m)]+P*(exp(r_clm[(SHIFT[c]+NoSum[l,c]+m)]
     *x[m,n,l,c])-1))
    }
   K_clm[(SHIFT[c]+NoSum[l,c]+m)]<-exp(K_clm_L[(SHIFT[c]+NoSum[l,c]+m)])
   K_clm_L[(SHIFT[c]+NoSum[l,c]+m)] ~ dnorm(alpha_c[c]+K_o_l_L[l]
   +(delta_l[l,c]*gamma_cl_L[l,c]),exp(tau_K_cl[l+(c-1)*N]))T(,0)
   r_clm[(SHIFT[c]+NoSum[l,c]+m)]<-exp(r_clm_L[(SHIFT[c]+NoSum[l,c]+m)])
   r_clm_L[(SHIFT[c]+NoSum[l,c]+m)] ~ dnorm(beta_c[c]+r_o_l_L[l]
   +(delta_l[l,c]*omega_cl_L[l,c]),exp(tau_r_cl[l+(c-1)*N]))T(,3.5)
   }
  tau_K_cl[l+(c-1)*N]~dnorm(tau_K_p_c[c],exp(sigma_tau_K_c[c]))T(0,)
  tau_r_cl[l+(c-1)*N]~dnorm(tau_r_p_c[c],exp(sigma_tau_r_c[c]))
nu_cl[l+(c-1)*N]~dnorm(nu_p,exp(sigma_nu))
 }
 K_o_l_L[l]<- log(K_o_l[l])
 K_o_l[l] ~ dt(exp(K_p),exp(sigma_K_o),3)T(0,)
 r_o_l_L[l]<- log(r_o_l[l])
 r_o_l[l] ~ dt(exp(r_p),exp(sigma_r_o),3)T(0,)
 delta_l[l,1]<-0
 delta_l[l,2]~dbern(p)
 gamma_cl_L[l,1]<-0
 gamma_cl_L[l,2]<-log(gamma_l[l])
 gamma_l[l]~dt(1,exp(sigma_gamma),3)T(0,)
 omega_cl_L[l,1]<-0
 omega_cl_L[l,2]<-log(omega_l[l])
 omega_l[l]~dt(1,exp(sigma_omega),3)T(0,)
}
alpha_c[1]<-0
alpha_c[2]~dnorm(alpha_mu,eta_alpha)
beta_c[1]<-0
beta_c[2]~dnorm(beta_mu,eta_beta)
K_p~dnorm(K_mu,eta_K_p)
r_p~dnorm(r_mu,eta_r_p)
nu_p~dnorm(nu_mu,eta_nu_p)
P <- exp(P_L)
P_L ~dnorm(P_mu,eta_P)
sigma_K_o~dnorm(eta_K_o,psi_K_o)
sigma_r_o~dnorm(eta_r_o,psi_r_o)
tau_K_p_c[1]~dnorm(tau_K_mu,eta_tau_K_p)
tau_K_p_c[2]~dnorm(tau_K_mu,eta_tau_K_p)
tau_r_p_c[1]~dnorm(tau_r_mu,eta_tau_r_p)
tau_r_p_c[2]~dnorm(tau_r_mu,eta_tau_r_p)
sigma_tau_K_c[1]~dnorm(eta_tau_K,psi_tau_K)
sigma_tau_K_c[2]~dnorm(eta_tau_K,psi_tau_K)
sigma_tau_r_c[1]~dnorm(eta_tau_r,psi_tau_r)
sigma_tau_r_c[2]~dnorm(eta_tau_r,psi_tau_r)
sigma_nu~dnorm(eta_nu,psi_nu)
sigma_gamma~dnorm(eta_gamma,psi_gamma)
sigma_omega~dnorm(eta_omega,psi_omega)
}
}
```

B.7. Additional *cdc13-1* 27°C vs *ura3*△ 27°C fitness plots

Figure B.7: Alternative non-Bayesian, hierarchical fitness plot, from fitting the random effects model (REM) to data in Table S6 from Addinall *et al.* (2011) ($F = MDR \times MDP$). *orf*△s with significant evidence of interaction are highlighted in red and green for suppressors and enhancers respectively. *orf*△s without significant evidence of interaction are in grey and have no *orf* name label. Significant interactors are classified as those with FDR corrected p-values < 0.05.
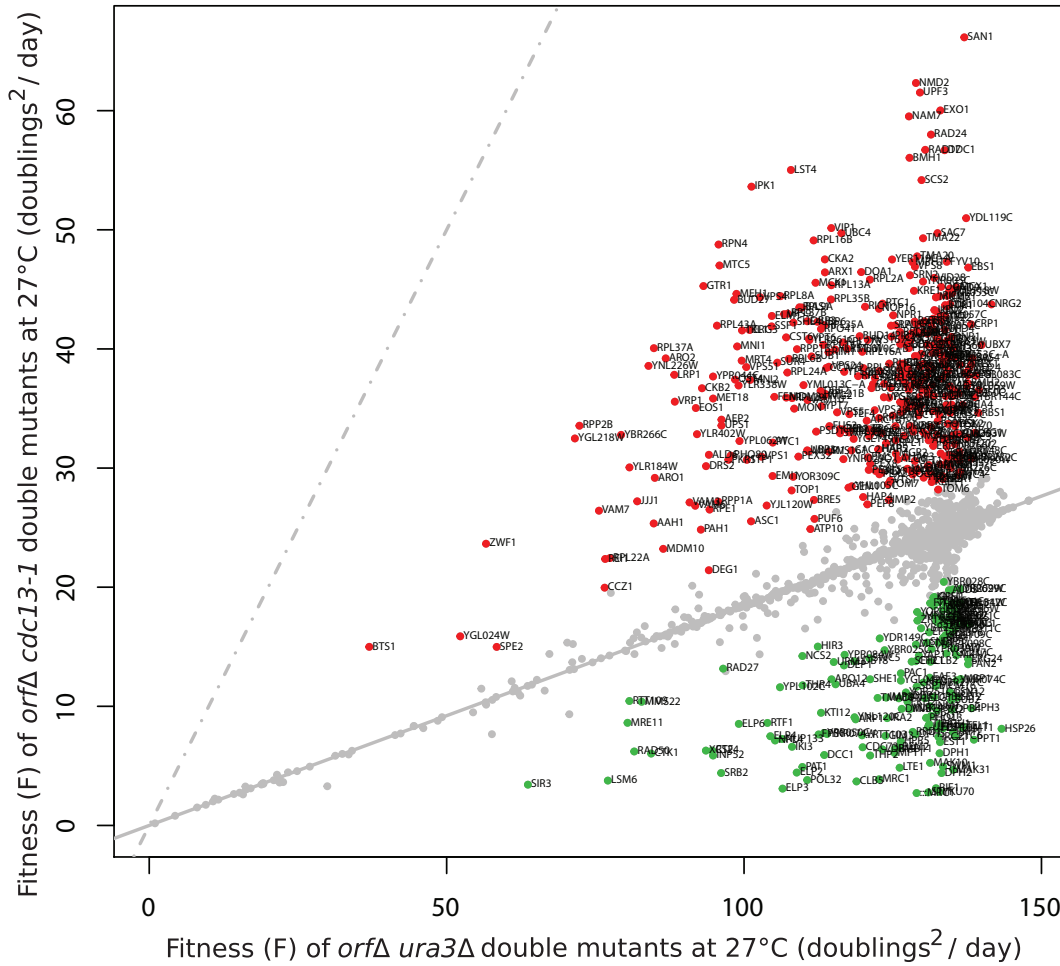
Figure B.8: Alternative interaction hierarchical model (IHM) fitness plot with *orf∆* posterior mean fitness. *orf∆*s with significant evidence of interaction are highlighted on the plot as red and green for suppressors and enhancers respectively ($F = MDR \times MDP$). Solid and dashed grey fitted lines are for the IHM linear model fit. *orf∆*s with significant evidence of interaction are highlighted in red and green for suppressors and enhancers respectively. *orf∆*s without significant evidence of interaction are in grey and have no *orf* name label. Significant interactors have posterior probability $\Delta > 0.5$.

Figure B.9: Alternative joint hierarchical model (JHM) fitness plot with *orf*Δ posterior mean fitnesses. The JHM does not does not make use of a fitness measure such as $MDR \times MDP$ but the fitness plot is given in terms of $MDR \times MDP$ for comparison with other approaches which do. *orf*Δ strains are classified as being a suppressor or enhancer based on one of the two parameters used to classify genetic interaction, growth parameter r, this means occasionally strains can be more fit in the query experiment in terms of $MDR \times MDP$ but be classified as enhancers (green). *orf*Δs with significant evidence of interaction are highlighted in red and green for suppressors and enhancers respectively. *orf*Δs without significant evidence of interaction are in grey and have no *orf* name label. Significant interactors have posterior probability $\Delta > 0.5$.
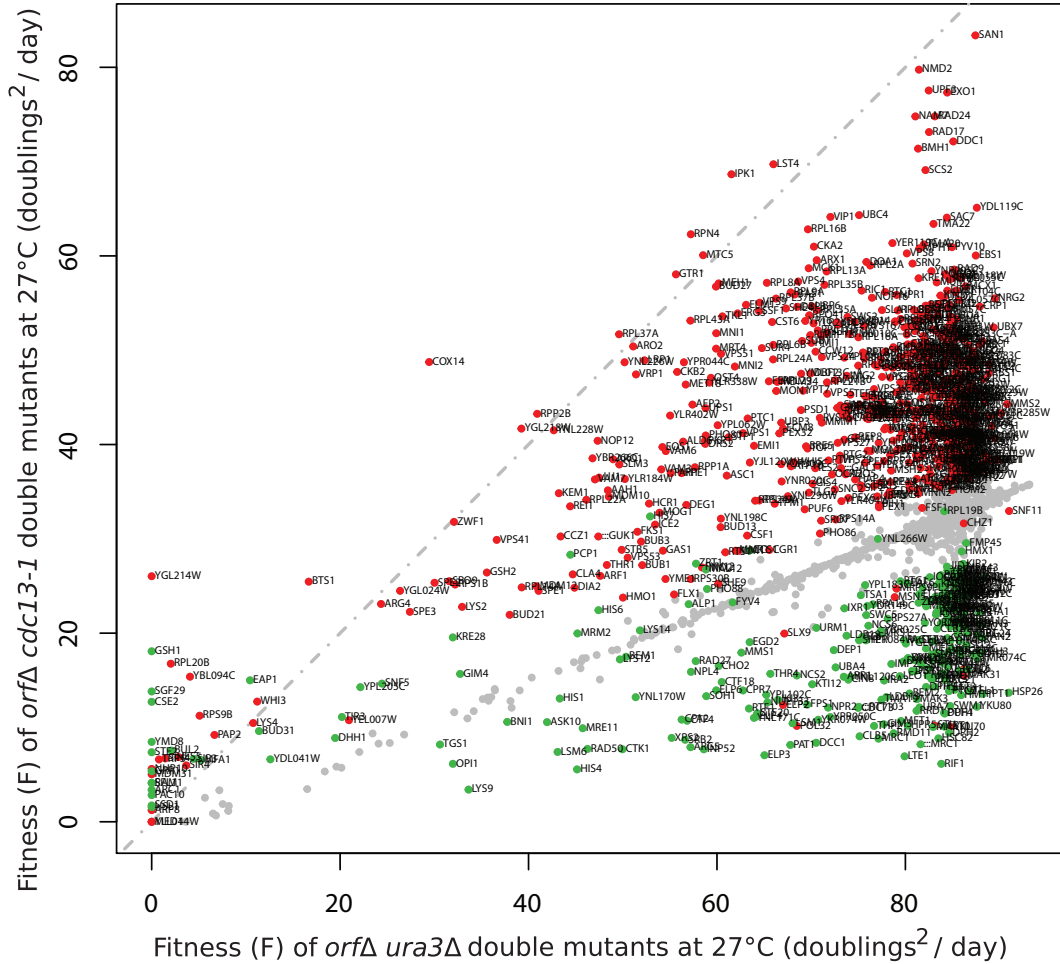
Figure B.10: Joint hierarchical model (JHM) carrying capacity fitness plot with *orf*Δ posterior mean fitnesses. *orf*Δ strains are classified as being a suppressor or enhancer based on carrying capacity parameter K. *orf*Δs with significant evidence of interaction are highlighted in red and green for suppressors and enhancers respectively. *orf*Δs without significant evidence of interaction are in grey and have no *orf* name label. Significant interactors have posterior probability $\Delta > 0.5$.
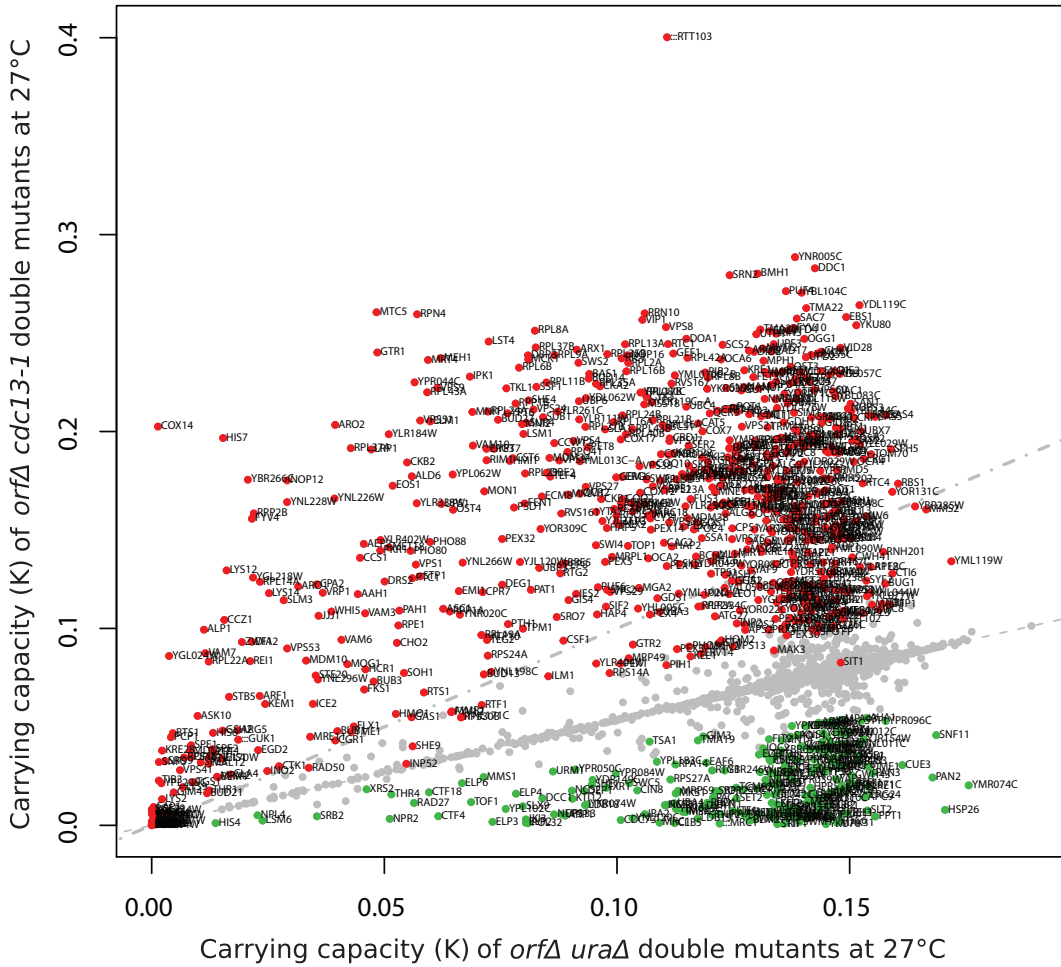
Figure B.11: Joint hierarchical model (JHM) growth rate fitness plot with *orfΔ* posterior mean fitnesses. *orfΔ* strains are classified as being a suppressor or enhancer based on growth parameter r. *orfΔ*s with significant evidence of interaction are highlighted in red and green for suppressors and enhancers respectively. *orfΔ*s without significant evidence of interaction are in grey and have no *orf* name label. Significant interactors have posterior probability $\Delta > 0.5$.
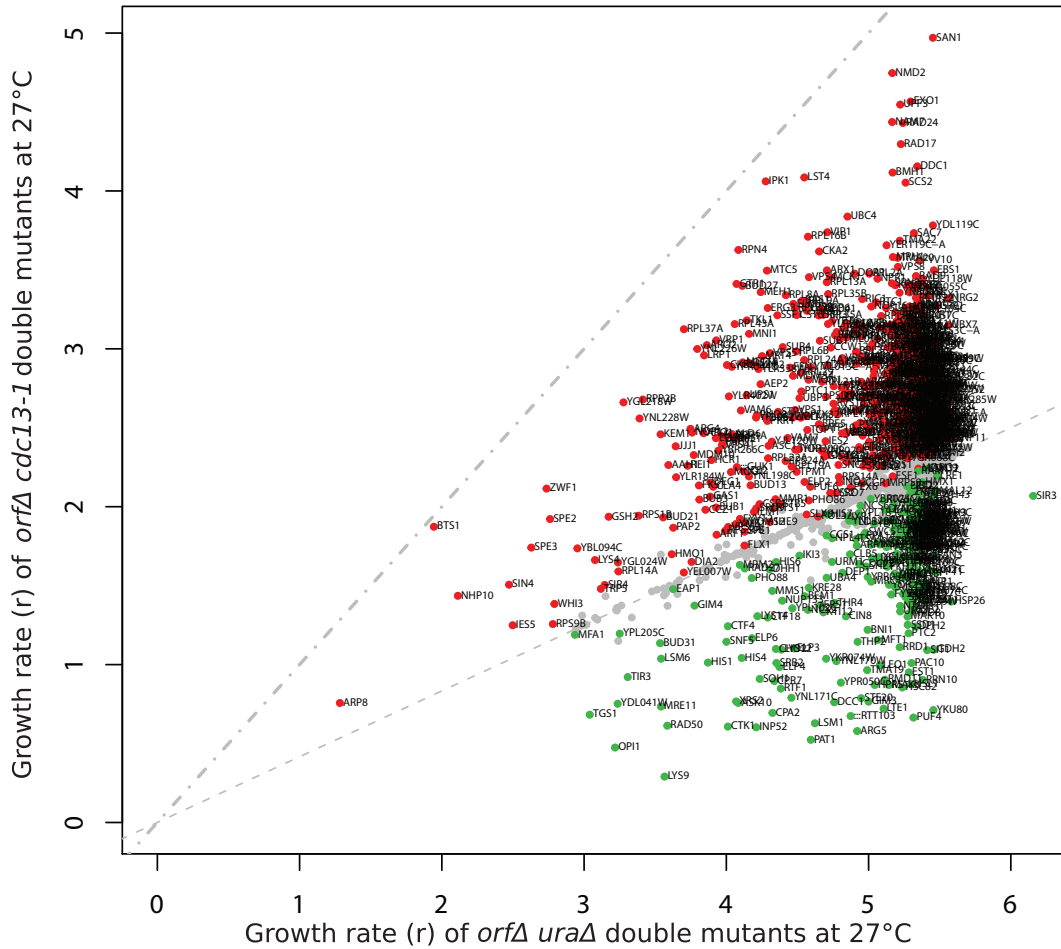
B.8. Correlation between methods

The Addinall et al. (2011) approach has its highest correlation with the IHM, followed by the JHM and then the REM. The REM correlates least well with the JHM while showing the same correlation with both the Addinall et al. (2011) approach and the IHM. The correlation between the IHM and the JHM is the largest observed between any of the methods, demonstrating the similarity of our Bayesian hierarchical methods.

Table B.4: Spearman's rank correlation coefficients for magnitudes from genetic independence, between Addinall et al. (2011), random effects approach (REM), interaction hierarchical model (IHM) and joint hierarchical model (JHM) approaches

Method	*Method*			
	Addinall et al. (2011) QFA	*REM QFA*	*IHM QFA*	*JHM QFA $(MDR \times MDP)$*
Addinall et al. (2011) QFA,	1	0.77	0.89	0.88
REM QFA,		1	0.77	0.75
IHM QFA,			1	0.95
JHM QFA $(MDR \times MDP)$,				1

The $MDR \times MDP$ correlation plot of the JHM versus the Addinall et al. (2011) approach demonstrates the similarity (Pearson correlation=0.90) and differences between the two approaches in terms of $MDR \times MDP$. We can see how the results differ between the JHM and Addinall et al. (2011), with a kink at the origin due to the JHM allowing shrinkage of non-interacting genes towards the fitted line.
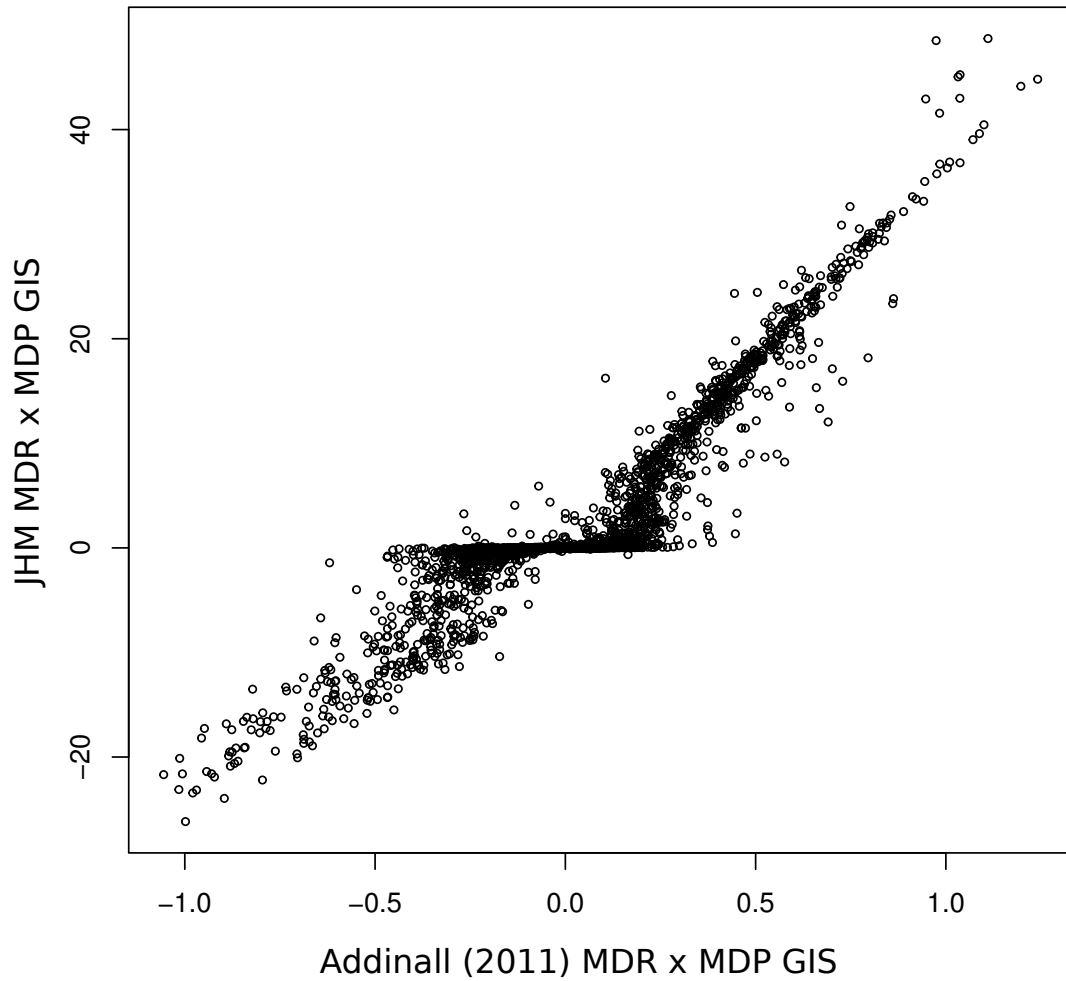
Figure B.12: $MDR \times MDP$ genetic interaction correlation plot of JHM versus Addinall et al. (2011) (Pearson correlation=0.90).

Appendix C. Stochastic logistic growth modelling

C.1. Linear noise approximation of the stochastic logistic growth model with multiplicative intrinsic noise solution

First we look to solve dZ_t, given in equation (5.10). We define $f(t) = -be^{v_t} = -\frac{baPe^{aT}}{bP(e^{aT}-1)+a}$ to obtain the following,

$$dZ_t = f(t)Z_t dt + \sigma dW_t.$$

In order to match our initial conditions correctly, $Z_0 = 0$. Define a new process $U_t = e^{-\int_{t_0}^{t} f(s)ds} Z_t$ and solve the integral,

$$\int_{t_0}^{t} f(s)ds = \int_{t_0}^{t} -\frac{baPe^{aS}}{bP(e^{aS}-1)+a}ds = \log\left(\frac{a}{bP(e^{aT}-1)+a}\right),$$

where, $S = s - t_0$ and $T = t - t_0$. Apply the chain rule to U_t,

$$dU_t = e^{-\int_{t_0}^{t} f(s)ds} dZ_t - f(t)e^{-\int_{t_0}^{t} f(s)ds} Z_t dt.$$

Now substitute in $dZ_t = f(t)Z_t dt + \sigma dW_t$ and simplify to give

$$dU_t = e^{-\int_{t_0}^{t} f(s)ds} \sigma dW_t.$$

Apply the following notation $\phi(t) = e^{\int_{t_0}^{t} f(s)ds} = \frac{a}{bP(e^{aT}-1)+a}$ and $\psi(t) = \sigma$ to give

$$dU_t = \phi(t)^{-1}\psi(t)dW_t.$$

U_t, has the following solution,

$$U_t = U_0 + \int_{t_0}^{t} \phi(s)^{-1}\psi(s)dW_s.$$

As $U_t = \phi(t)^{-1}Z_t$, Z_t then has the following solution (Arnold, 2013),

$$Z_t = \phi(t)\left[Z_0 + \int_{t_0}^{t} \phi(s)^{-1}\psi(s)dW_s\right].$$

Finally, the distribution at time t is $Z_t | Z_0 \sim N(M_t, E_t)$ (Arnold, 2013), where $M_t = \phi(t) Z_0$ and $E_t = \phi(t)^2 \int_{t_0}^t \left[\phi(s)^{-1} \psi(s) \right]^2 ds$.

Further, $M_t = \frac{a}{bP(e^{aT}-1)+a} Z_0$ and $E_t = \sigma^2 \left[\frac{a}{bP(e^{aT}-1)+a} \right]^2 \int_{t_0}^t \left[\frac{a}{bP(e^{aS}-1)+a} \right]^{-2} ds$.

As $\int_{t_0}^t \left[\frac{a}{bP(e^{aS}-1)+a} \right]^{-2} ds = \frac{b^2 P^2 (e^{2aT}-1) + 4bP(a-bP)(e^{aT}-1) + 2aT(a-bP)^2}{2a^3}$,

$$
\begin{aligned}
E_t =& \sigma^2 \left[\frac{a}{bP(e^{aT}-1)+a} \right]^2 \left[\frac{b^2 P^2 (e^{2aT}-1) + 4bP(a-bP)(e^{aT}-1) + 2aT(a-bP)^2}{2a^3} \right] \\
=& \sigma^2 \left[\frac{b^2 P^2 (e^{2aT}-1) + 4bP(a-bP)(e^{aT}-1) + 2aT(a-bP)^2}{2a \left(bP(e^{aT}-1)+a \right)^2} \right].
\end{aligned}
$$

Taking our solutions for v_t (5.8) and Z_t, we can now write our solution for the LNA to the log of the logistic growth process (5.6).

As $Y_t = v_t + Z_t$,

$$
Y_t | Y_0 \sim \mathcal{N} \left(\log \left[\frac{aPe^{aT}}{bP(e^{aT}-1)+a} \right] + M_t, E_t \right).
$$

Note: $\frac{aPe^{aT}}{bP(e^{aT}-1)+a}$ has the same functional form as the solution to the deterministic part of the logistic growth process (5.1) and is equivalent when $\sigma = 0$ (such that $a = r - \frac{\sigma^2}{2} = r$).

Further, as Y_t is normally distributed, we know $X_t = e^{Y_t}$ will be log normally distributed and

$$
X_t | X_0 \sim \log \mathcal{N}(\log \left(\frac{aPe^{aT}}{bP(e^{aT}-1)+a} \right) + M_t, E_t).
$$

Alternatively set $Q = \left(\frac{\frac{a}{b}}{P} - 1 \right) e^{at_0}$,

$$
X_t | X_0 \sim \log \mathcal{N}(\log \left(\frac{\frac{a}{b}}{1 + Qe^{-at}} \right) + M_t, E_t).
$$

From our solution to the log process we can obtain the following transition density

$$
(Y_{t_i}|Y_{t_{i-1}} = y_{t_{i-1}}) \sim \mathrm{N}\left(\mu_{t_i}, \Xi_{t_i}\right),
$$

$$
\text{where } y_{t_{i-1}} = v_{t_{i-1}} + z_{t_{i-1}}, Q = \left(\frac{\frac{a}{b}}{P} - 1\right)e^{at_0},
$$

$$
\mu_{t_i} = y_{t_{i-1}} + \log\left(\frac{1 + Qe^{-at_{i-1}}}{1 + Qe^{-at_i}}\right) + e^{-a(t_i - t_{i-1})}\frac{1 + Qe^{-at_{i-1}}}{1 + Qe^{-at_i}}z_{t_{i-1}} \text{ and}
$$

$$
\Xi_{t_i} = \sigma^2\left[\frac{4Q(e^{at_i} - e^{at_{i-1}}) + e^{2at_i} - e^{2at_{i-1}} + 2aQ^2(t_i - t_{i-1})}{2a(Q + e^{at_i})^2}\right].
$$

C.2. Zero-order noise approximation of the stochastic logistic growth model

After obtaining (5.7) in Section 5.3, we can derive the RRTR logistic growth diffusion process as follows. First our expression for dv_t, given in (5.7), is approximated by setting $\sigma^2 = 0$,

$$dv_t = \left(r - \frac{1}{2}\sigma^2 - \frac{r}{K}e^{v_t} \right) dt = \left(r - \frac{r}{K}e^{v_t} \right) dt.$$

We now write down an expression for dZ_t, where dY_t is given in (5.6) and $dZ_t = dY_t - dv_t$,

$$dZ_t = \left(r - \frac{1}{2}\sigma^2 - \frac{r}{K}e^{Y_t} \right) dt + \sigma dW_t - \left(r - \frac{r}{K}e^{v_t} \right) dt.$$

We can then rearrange and simplify to give the following,

$$dZ_t = \left(\frac{r}{K} \left[e^{v_t} - e^{Y_t} \right] - \frac{1}{2}\sigma^2 \right) dt + \sigma dW_t.$$

We now substitute in $Y_t = v_t + Z_t$,

$$dZ_t = \left(\frac{r}{K} \left[e^{v_t} - e^{v_t + Z_t} \right] - \frac{1}{2}\sigma^2 \right) dt + \sigma dW_t.$$

We now apply a zero order LNA by setting $e^{Z_t} = 1$ to obtain,

$$dZ_t = \left(\frac{r}{K} \left[e^{v_t} - e^{v_t} \right] - \frac{1}{2}\sigma^2 \right) dt + \sigma dW_t.$$

We can then simplify to give the following,

$$dZ_t = -\frac{1}{2}\sigma^2 dt + \sigma dW_t. \tag{C.1}$$

Differentiating v_t, given in (5.8), with respect to t we can obtain an alternative expression for dv_t,

$$dv_t = \frac{a(a - bP)}{bP(e^{aT} - 1) + a} dt = \frac{r(K - P)}{K + P(e^{rT} - 1)} dt, \tag{C.2}$$

where $T = t - t_0$. We now write down our new expression for Y_t, where $dY_t = dv_t + dZ_t$, given (C.2) and (C.1),

$$dY_t = \left(\frac{r(K - P)}{K + P(e^{aT} - 1)} - \frac{1}{2}\sigma^2 \right) dt + \sigma dW_t$$

or alternatively by setting $Q = \left(\frac{K}{P} - 1\right) e^{a t_0}$,

$$dY_t = \left(\frac{Qr}{e^{rt} + Q} - \frac{1}{2}\sigma^2\right) dt + \sigma dW_t.$$

We can then apply Itô's lemma (5.5) (Itô, 1944) with the transformation $f(t, Y_t) \equiv X_t = e^{Y_t}$. After deriving the following partial derivatives:

$$\frac{df}{dt} = 0, \qquad \frac{df}{dx} = e^{Y_t} \quad \text{and} \quad \frac{d^2 f}{dx^2} = e^{Y_t},$$

we can obtain the following Itô drift-diffusion process:

$$dX_t = \frac{Qr}{e^{rt} + Q} X_t dt + \sigma dW_t,$$

which is exactly the RRTR logistic diffusion process presented by Román-Román & Torres-Ruiz (2012).

C.3. Linear noise approximation of the stochastic logistic growth model with additive intrinsic noise solution

First we look to solve dZ_t, given in (5.14). We define $f(t) = a - 2bv_t$ to obtain the following,

$$dZ_t = f(t)Z_t dt + \sigma v_t dW_t.$$

In order to match our initial conditions correctly, $Z_0 = 0$. Define a new process $U_t = e^{-\int_{t_0}^t f(s)ds} Z_t$ and solve the integral,

$$\int_{t_0}^t f(s)ds = \int_{t_0}^t (a - 2bV_s)ds = aT - 2\log\left(\frac{bP(e^{aT} - 1) + a}{a}\right),$$

as $\int_{t_0}^t V_s ds = \frac{1}{b}\log\left(\frac{bP(e^{aT}-1)+a}{a}\right)$, where $S = s - t_0$ and $T = t - t_0$. Apply the chain rule to U_t,

$$dU_t = e^{-\int_{t_0}^t f(s)ds} dZ_t - f(t)e^{-\int_{t_0}^t f(s)ds} Z_t dt.$$

Now substitute in $dZ_t = f(t)Z_t dt + \sigma v_t dW_t$ and simplify to give,

$$dU_t = e^{-\int_{t_0}^t f(s)ds} \sigma v_t dW_t.$$

Apply the following notation $\phi(t) = e^{\int_{t_0}^t f(s)ds} = e^{aT}\left(\frac{a}{bP(e^{aT}-1)+a}\right)^2$ and $\psi(t) = \sigma v_t$ to give,

$$dU_t = \phi(t)^{-1}\psi(t)dW_t.$$

U_t has the following solution,

$$U_t = U_0 + \int_{t_0}^t \phi(s)^{-1}\psi(s)dW_s.$$

As $U_t = \phi(t)^{-1}Z_t$, Z_t has the following solution (Arnold, 2013),

$$Z_t = \phi(t)\left[Z_0 + \int_{t_0}^t \phi(s)^{-1}\psi(s)dW_s\right].$$

Finally the distribution at time t is $Z_t|Z_0 \sim N(M_t, E_t)$ (Arnold, 2013), where $M_t = \phi(t)Z_0$ and $E_t = \phi(t)^2 \int_{t_0}^t \left[\phi(s)^{-1}\psi(s)\right]^2 ds$.

$$M_t = e^{aT}\left(\frac{a}{bP(e^{aT} - 1) + a}\right)^2 Z_0$$

and

$$E_t = \left(e^{aT} \left(\frac{a}{bP(e^{aT} - 1) + a} \right)^2 \right)^2 \int_{t_0}^{t} \left[e^{aS} \left(\frac{a}{bP(e^{aS} - 1) + a} \right)^2 \right]^{-2} \sigma^2 V_s^2 ds$$

$$= \sigma^2 \left(e^{aT} \left(\frac{a}{bP(e^{aT} - 1) + a} \right)^2 \right)^2$$

$$\times \int_{t_0}^{t} \left[e^{aS} \left(\frac{a}{bP(e^{aS} - 1) + a} \right)^2 \right]^{-2} \left[\frac{aPe^{aS}}{bP(e^{aS} - 1) + a} \right]^2 ds$$

$$= \sigma^2 \left(e^{aT} \left(\frac{a}{bP(e^{aT} - 1) + a} \right)^2 \right)^2$$

$$\times \int_{t_0}^{t} \left[e^{-2aS} \left(\frac{a}{bP(e^{aS} - 1) + a} \right)^{-4} \right] \left[\frac{aPe^{aS}}{bP(e^{aS} - 1) + a} \right]^2 ds$$

$$= \sigma^2 \left(e^{aT} \left(\frac{1}{bP(e^{aT} - 1) + a} \right)^2 \right)^2 \int_{t_0}^{t} \left[a^2 P^2 \left(\frac{1}{bP(e^{aS} - 1) + a} \right)^{-2} \right] ds,$$

as $\int_{t_0}^{t} \left(\frac{1}{bP(e^{aS} - 1) + a} \right)^{-2} ds = \frac{b^2 P^2 (e^{2aT} - 1) + 4bP(a - bP)(e^{aT} - 1) + 2aT(a - bP)^2}{2a}$,

$$E_t = \frac{1}{2} \sigma^2 a P^2 e^{2aT} \left(\frac{1}{bP(e^{aT} - 1) + a} \right)^4$$

$$\times \left[b^2 P^2 (e^{2aT} - 1) + 4bP(a - bP)(e^{aT} - 1) + 2aT(a - bP)^2 \right].$$

Taking our solutions for v_t (5.13) and Z_t, we can obtain the following transition density

$$(X_{t_i} | X_{t_{i-1}} = x_{t_{i-1}}) \sim N(\mu_{t_i}, \Xi_{t_i}),$$

$$\text{where } x_{t_{i-1}} = v_{t_{i-1}} + z_{t_{i-1}},$$

$$\mu_{t_i} = x_{t_{i-1}} + \left(\frac{aPe^{aT_i}}{bP(e^{aT_i} - 1) + a} \right) - \left(\frac{aPe^{aT_{i-1}}}{bP(e^{aT_{i-1}} - 1) + a} \right)$$

$$+ e^{a(t_i - t_{i-1})} \left(\frac{bP(e^{aT_{i-1}} - 1) + a}{bP(e^{aT_i} - 1) + a} \right)^2 Z_{t_{i-1}} \text{ and}$$

$$\Xi_{t_i} = \frac{1}{2} \sigma^2 a P^2 e^{2aT_i} \left(\frac{1}{bP(e^{aT_i} - 1) + a} \right)^4$$

$$\times [b^2 P^2 (e^{2aT_i} - e^{2aT_{i-1}}) + 4bP(a - bP)(e^{aT_i} - e^{aT_{i-1}})$$

$$+ 2a(t_i - t_{i-1})(a - bP)^2].$$

C.4. Prior hyper-parameters for Bayesian state space models

Table C.1: Prior hyper-parameters for Bayesian sate space models, Log-normal with mean (μ) and precision (τ)

Parameter Name	Value
μ_K	$\log(0.1)$
τ_K	2
μ_r	$\log(3)$
τ_r	5
μ_P	$\log(0.0001)$
τ_P	0.1
μ_σ	$\log(100)$
τ_σ	0.1
μ_ν	$\log(10000)$
τ_ν	0.1

C.5. Kalman filter for the linear noise approximation of the stochastic logistic growth model with additive intrinsic noise and Normal measurement error

To find $\pi(y_{t_{1:N}})$ for the LNAA with Normal measurement error we can use the following Kalman Filter algorithm. First we assume the following:

$$\theta_{t_i}|y_{1:t_i} \sim \mathrm{N}(m_{t_i}, C_{t_i}),$$
$$m_{t_i} = a_{t_i} + R_{t_i} F(F^T R_{t_i} F + U)^{-1}[y_{t_i} - F^T a_{t_i}],$$
$$C_{t_i} = R_{t_i} - R_{t_i} F(F^T R_{t_i} F + U)^{-1} F^T R_{t_i}$$

and initialize with $m_0 = P$ and $C_0 = 0$. Now suppose that,

$$\theta_{t_i}|y_{1:t_{i-1}} \sim \mathrm{N}(a_{t_i}, R_{t_i}),$$
$$a_{t_i} = G_{t_i} m_{t_{i-1}}$$
$$\text{and } R_{t_i} = G_{t_i} C_{t_{i-1}} G_{t_i}^T + W_{t_i}.$$

The transition density distribution, see (5.15) is as follows:

$$\theta_{t_i}|\theta_{t_{i-1}} \sim \mathrm{N}(G_{t_i}\theta_{t_{i-1}}, W_{t_i})$$

or equivalently $(X_{t_i}|X_{t_{i-1}} = x_{t_{i-1}}) \sim \mathrm{N}\left(\mu_{t_i}, \Xi_{t_i}\right)$, where $x_{t_{i-1}} = v_{t_{i-1}} + z_{t_{i-1}}$,

$$\theta_t = \begin{pmatrix} 1 \\ X_{t_i} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ H_{\alpha,t_i} & H_{\beta,t_i} \end{pmatrix} \begin{pmatrix} 1 \\ X_{t_{i-1}} \end{pmatrix}$$
$$= G_{t_i}\theta_{t_{i-1}},$$
$$G_{t_i} = \begin{pmatrix} 1 & 0 \\ H_{\alpha,t_i} & H_{\beta,t_i} \end{pmatrix}, \quad W_{t_i} = \begin{pmatrix} 0 & 0 \\ 0 & \Xi_{t_i} \end{pmatrix}$$

where $H_{\alpha,t_i} = H_\alpha(t_i, t_{i-1}) = v_t - V_{t-1}e^{a(t_i - t_{i-1})}\left(\dfrac{bP(e^{aT_{i-1}} - 1) + a}{bP(e^{aT_i} - 1) + a}\right)^2$

and $H_{\beta,t_i} = H_\beta(t_i, t_{i-1}) = e^{a(t_i - t_{i-1})}\left(\dfrac{bP(e^{aT_{i-1}} - 1) + a}{bP(e^{aT_i} - 1) + a}\right)^2.$

The measurement error distribution is as follows:

$$y_{t_i}|\theta_{t_i} \sim \mathrm{N}(F^T \theta_{t_i}, U)$$
$$\text{or equivalently } y_{t_i}|\theta_{t_i} \sim \mathrm{N}(X_{t_i}, \sigma_\nu^2),$$
$$\text{where } F = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \text{ and } U = \sigma_\nu^2.$$

Matrix Algebra:

$$a_{t_i} = G_{t_i} m_{t_{i-1}}$$
$$= \begin{pmatrix} 1 & 0 \\ H_{\alpha,t_i} & H_{\beta,t_i} \end{pmatrix} \begin{pmatrix} 1 \\ m_{t_{i-1}} \end{pmatrix} = \begin{pmatrix} 1 \\ H_{\alpha,t_i} + H_{\beta,t_i} m_{t_{i-1}} \end{pmatrix}$$

$$R_{t_i} = G_{t_i} C_{t_{i-1}} G_{t_i}^T + W_{t_i}$$
$$= \begin{pmatrix} 0 & 0 \\ 0 & H_{\beta,t_i}{}^2 c_{t_{i-1}}^2 \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 0 & \Xi_{t_i} \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & H_{\beta,t_i}{}^2 c_{t_{i-1}}^2 + \Xi_{t_i} \end{pmatrix}$$

$$C_{t_{i-1}} = \begin{pmatrix} 0 & 0 \\ 0 & c_{t_{i-1}}^2 \end{pmatrix}$$

$$R_{t_i} F (F^T R_{t_i} F + U)^{-1} = \begin{pmatrix} 0 & 0 \\ 0 & H_{\beta,t_i}{}^2 c_{t_{i-1}}^2 + \Xi_{t_i} \end{pmatrix} \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$
$$\times \left[\begin{pmatrix} 0 & 1 \end{pmatrix} \begin{pmatrix} 0 & 0 \\ 0 & H_{\beta,t_i}{}^2 c_{t_{i-1}}^2 + \Xi_{t_i} \end{pmatrix} \begin{pmatrix} 0 \\ 1 \end{pmatrix} + \sigma_\nu^2 \right]^{-1}$$
$$= \left[\left(H_{\beta,t_i}{}^2 c_{t_{i-1}}^2 + \Xi_{t_i} + \sigma_\nu^2 \right) \right]^{-1} \begin{pmatrix} 0 \\ H_{\beta,t_i}{}^2 c_{t_{i-1}}^2 + \Xi_{t_i} \end{pmatrix}$$

$$m_{t_i} = a_{t_i} + R_{t_i} F (F^T R_{t_i} F + U)^{-1} [y_{t_i} - F^T a_{t_i}]$$

$$= \begin{pmatrix} 1 \\ H_{\alpha,t_i} + H_{\beta,t_i} m_{t_{i-1}} \end{pmatrix}$$

$$+ \left[\left(H_{\beta,t_i}{}^2 c_{t_{i-1}}^2 + \Xi_{t_i} + \sigma_\nu^2 \right) \right]^{-1}$$

$$\times \begin{pmatrix} 0 \\ H_{\beta,t_i}{}^2 c_{t_{i-1}}^2 + \Xi_{t_i} \end{pmatrix} \left[y_{t_i} - \begin{pmatrix} 0 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ H_{\alpha,t_i} + H_{\beta,t_i} m_{t_{i-1}} \end{pmatrix} \right]$$

$$= \begin{pmatrix} 0 \\ H_{\alpha,t_i} + H_{\beta,t_i} m_{t_{i-1}} + \frac{H_{\beta,t_i}{}^2 c_{t_{i-1}}^2 + \Xi_{t_i}}{H_{\beta,t_i}{}^2 c_{t_{i-1}}^2 + \Xi_{t_i} + \sigma_\nu^2} \left[y_{t_i} - H_{\alpha,t_i} - H_{\beta,t_i} m_{t_{i-1}} \right] \end{pmatrix}$$

$$C_{t_i} = R_{t_i} - R_{t_i} F (F^T R_{t_i} F + U)^{-1} F^T R_{t_i}$$

$$= \begin{pmatrix} 0 & 0 \\ 0 & H_{\beta,t_i}{}^2 c_{t_{i-1}}^2 + \Xi_{t_i} \end{pmatrix}$$

$$- \left[\left(H_{\beta,t_i}{}^2 c_{t_{i-1}}^2 + \Xi_{t_i} + \sigma_\nu^2 \right) \right]^{-1}$$

$$\times \begin{pmatrix} 0 \\ H_{\beta,t_i}{}^2 c_{t_{i-1}}^2 + \Xi_{t_i} \end{pmatrix} \left[\begin{pmatrix} 0 & 1 \end{pmatrix} \begin{pmatrix} 0 & 0 \\ 0 & H\beta, t_i{}^2 c_{t_{i-1}}^2 + \Xi_{t_i} \end{pmatrix} \right]$$

$$= \begin{pmatrix} 0 & 0 \\ 0 & H_{\beta,t_i}{}^2 c_{t_{i-1}}^2 + \Xi_{t_i} - \frac{\left(H_{\beta,t_i}{}^2 c_{t_{i-1}}^2 + \Xi_{t_i} \right)^2}{H_{\beta,t_i}{}^2 c_{t_{i-1}}^2 + \Xi_{t_i} + \sigma_\nu^2} \end{pmatrix}$$

With m_{t_i} and C_{t_i} for $i = 1 : N$, we can evaluate a_{t_i}, R_{t_i} and $\pi(x_{t_i}|y_{t_{1:(i-1)}})$ for $i = 1 : N$. We are interested in $\pi(y_{t_{1:i}}) = \prod_{i=1}^{N} \pi(y_{t_i}|y_{t_{1:(i-1)}})$, where $\pi(y_{t_i}|y_{t_{1:(i-1)}}) = \int_x \pi(y_{t_i}|x_{t_i}) \pi(x_{t_i}|y_{t_{1:(i-1)}}) dx_{t_i}$ gives a tractable Gaussian integral. Finally,

$$\log \pi(y_{t_{1:N}}) = \sum_{i=1}^{N} \log \pi(y_{t_i}|y_{t_{1:(i-1)}})$$

$$= \sum_{i=1}^{N} \left[-\log \left(\sqrt{2\pi(\sigma_f^2 + \sigma_g^2)} \right) - \frac{(\mu_f - \mu_g)^2}{2(\sigma_f^2 + \sigma_g^2)} \right],$$

$$\text{where } \mu_f - \mu_g = y_{t_i} - a_{t_i} = y_{t_i} - H_{\alpha,t_i} - H_{\beta,t_i} m_{t_{i-1}}$$

$$\text{and } \sigma_f^2 + \sigma_g^2 = \sigma_\nu^2 + R_{t_i} = \sigma_\nu^2 + H_{\beta,t_i}{}^2 c_{t_{i-1}}^2 + \Xi_{t_i}.$$

Procedure

1. Set $i = 1$. Initialize $m_0 = P$ and $C_0 = 0$.

2. Evaluate and store the following log likelihood term:

$$\log \pi(y_{t_i}|y_{t_{1:(i-1)}}) = \left[-\log \left(\sqrt{2\pi(\sigma_f^2 + \sigma_g^2)} \right) - \frac{(\mu_f - \mu_g)^2}{2(\sigma_f^2 + \sigma_g^2)} \right],$$

where $\mu_f - \mu_g = y_{t_i} - H_{\alpha,t_i} - H_{\beta,t_i}m_{t_{i-1}}$ and $\sigma_f^2 + \sigma_g^2 = \sigma_\nu^2 + H_{\beta,t_i}{}^2 c_{t_{i-1}}^2 + \Xi_{t_i}$.

3. Create and store both m_{t_i}, and C_{t_i},

where $m_{t_i} = H_{\alpha,t_i} + H_{\beta,t_i}m_{t_{i-1}} + \frac{H_{\beta,t_i}{}^2 c_{t_{i-1}}^2 + \Xi_{t_i}}{H_{\beta,t_i}{}^2 c_{t_{i-1}}^2 + \Xi_{t_i} + \sigma_\nu^2} \left[y_{t_i} - H_{\alpha,t_i} - H_{\beta,t_i}m_{t_{i-1}} \right]$

and $c_{t_i}^2 = H_{\beta,t_i}{}^2 c_{t_{i-1}}^2 + \Xi_{t_i} - \frac{\left(H_{\beta,t_i}{}^2 c_{t_{i-1}}^2 + \Xi_{t_i} \right)^2}{H_{\beta,t_i}{}^2 c_{t_{i-1}}^2 + \Xi_{t_i} + \sigma_\nu^2}.$

4. Increment i, $i=(i+1)$ and repeat steps 2-3 till $\log \pi(y_{t_N}|y_{t_{1:(N-1)}})$ is evaluated.

5. Calculate the sum:

$$\log \pi(y_{t_{1:N}}) = \sum_{i=1}^{N} \log \pi(y_{t_i}|y_{t_{1:(i-1)}}).$$

Bibliography

ADDINALL, S. G., DOWNEY, M., YU, M., ZUBKO, M. K., DEWAR, J., LEAKE, A., HALLINAN, J., SHAW, O., JAMES, K., WILKINSON, D. J., WIPAT, A., DUROCHER, D. & LYDALL, D. 2008 A genomewide suppressor and enhancer analysis of cdc13-1 reveals varied cellular processes influencing telomere capping in Saccharomyces cerevisiae. *Genetics* **180** (4), 2251–2266.

ADDINALL, S. G., HOLSTEIN, E.-M., LAWLESS, C., YU, M., CHAPMAN, K., BANKS, A. P., NGO, H.-P., MARINGELE, L., TASCHUK, M., YOUNG, A., CIESIOLKA, A., LISTER, A. L., WIPAT, A., WILKINSON, D. J. & LYDALL, D. 2011 Quantitative Fitness Analysis Shows That NMD Proteins and Many Other Protein Complexes Suppress or Enhance Distinct Telomere Cap Defects. *PLoS Genet* **7** (4), e1001362.

ALLEN, L. J. 2010 *An Introduction to Stochastic Processes with Biology Applications, Second Edition*. Chapman and Hall/CRC.

AN, W. F. & TOLLIDAY, N. J. 2009 Introduction: cell-based assays for high-throughput screening. *Methods in molecular biology (Clifton, N.J.)* **486**, 1–12.

ANBALAGAN, S., BONETTI, D., LUCCHINI, G. & LONGHESE, M. P. 2011 Rif1 Supports the Function of the CST Complex in Yeast Telomere Capping. *PLoS Genet* **7** (3), e1002024.

ANDRIEU, C. & THOMS, J. 2008 A tutorial on adaptive MCMC. *Statistics and Computing* **18** (4), 343–373.

ARNOLD, L. 2013 *Stochastic Differential Equations: Theory and Applications*. Dover Publications, Incorporated.

AYLOR, D. L. & ZENG, Z.-B. 2008 From Classical Genetics to Quantitative Genetics to Systems Biology: Modeling Epistasis. *PLoS Genet* **4** (3), e1000029.

BANKS, A., LAWLESS, C. & LYDALL, D. 2012 A Quantitative Fitness Analysis Workflow. *J. Vis. Exp* **66**, e4018.

BATES, D., MAECHLER, M. & BOLKER, B. 2013 *lme4: Linear mixed-effects models using S4 classes*. R package version 0.999999-2.

BAYES & PRICE 1763 An Essay towards Solving a Problem in the Doctrine of Chances. By the Late Rev. Mr. Bayes, F. R. S. Communicated by Mr. Price, in a Letter to John Canton, A. M. F. R. S. *Philosophical Transactions (1683-1775)* **53**, 370–418.

BENJAMINI, Y. & HOCHBERG, Y. 1995 Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)* **57** (1), 289–300.

BERNARDO, J. & SMITH, A. 2007 *Bayesian Theory*. John Wiley & Sons Canada, Limited.

BOTSTEIN, D., CHERVITZ, S. A. & CHERRY, M. 1997 Yeast as a Model Organism. *Science* **277** (5330), 1259–1260.

BOX, G. E. P. & COX, D. R. 1964 An Analysis of Transformations. *Journal of the Royal Statistical Society. Series B (Methodological)* **26** (2), 211–252.

BOYD, D. & CRAWFORD, K. 2011 Six Provocations for Big Data. *Social Science Research Network Working Paper Series* .

CAPOCELLI, R. & RICCIARDI, L. 1974 Growth with regulation in random environment. *Kybernetik* **15** (3), 147–157.

CARLETTI, M. 2006 Numerical solution of stochastic differential problems in the biosciences. *J. Comput. Appl. Math.* **185** (2), 422–440.

CASELLA, G. & GEORGE, E. I. 1992 Explaining the Gibbs Sampler. *The American Statistician* **46** (3), 167–174.

CHEETHAM, A. H. & HAZEL, J. E. 1969 Binary (Presence-Absence) Similarity Coefficients. *Journal of Paleontology* **43** (5), 1130–1136.

CHEN, C., GRENNAN, K., BADNER, J., ZHANG, D., GERSHON, E., JIN, L. & LIU, C. 2011 Removing batch effects in analysis of expression microarray data: An evaluation of six batch adjustment methods. *PLoS ONE* **6** (2), e17238.

CHEN, Y., LAWLESS, C., GILLESPIE, C. S., WU, J., BOYS, R. J. & WILKINSON, D. J. 2010 CaliBayes and BASIS: integrated tools for the calibration, simulation and storage of biological simulation models. *Briefings in Bioinformatics* **11** (3), 278–289.

146

CHERRY, J. M., HONG, E. L., AMUNDSEN, C., BALAKRISHNAN, R., BINKLEY, G., CHAN, E. T., CHRISTIE, K. R., COSTANZO, M. C., DWIGHT, S. S., ENGEL, S. R., FISK, D. G., HIRSCHMAN, J. E., HITZ, B. C., KARRA, K., KRIEGER, C. J., MIYASATO, S. R., NASH, R. S., PARK, J., SKRZYPEK, M. S., SIMISON, M., WENG, S. & WONG, E. D. 2012 Saccharomyces Genome Database: the genomics resource of budding yeast. *Nucleic acids research* **40** (Database issue).

CHIB, S. & GREENBERG, E. 1995 Understanding the Metropolis-Hastings Algorithm. *The American Statistician* **49** (4), 327–335.

COLE, D. J., RIDOUT, M. S., MORGAN, B. J. T., BYRNE, L. J. & TUITE, M. F. 2007 Approximations for expected generation number. *Biometrics* **63** (4), 1023–1030.

CONG, Y.-S., WRIGHT, W. E. & SHAY, J. W. 2002 Human Telomerase and Its Regulation. *Microbiology and Molecular Biology Reviews* **66** (3), 407–425.

CONSORTIUM, G. O. 2004 The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Research* **32** (suppl 1), D258–D261.

CORDELL, H. J. 2002 Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Human molecular genetics* **11** (20), 2463–2468.

COWLES, M. K. & CARLIN, B. P. 1996 Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review. *Journal of the American Statistical Association* **91**, 883–904.

DAVID WYNFORD-THOMAS, D. K. 1997 The end-replication problem.

DEWAR, J. M. & LYDALL, D. A. 2012 Similarities and differences between uncapped telomeres and DNA double-strand breaks. *Chromosoma* **121** (2), 117–130.

DURBIN, J., HARVEY, A., KOOPMAN, S. & SHEPHARD, N. 2004 *State Space and Unobserved Component Models: Theory and Applications*. Cambridge University Press.

DURRETT, R. 1996 *Stochastic Calculus: A Practical Introduction*. Taylor & Francis.

EDEN, E., LIPSON, D., YOGEV, S. & YAKHINI, Z. 2007 Discovering Motifs in Ranked Lists of DNA Sequences. *PLoS Comput Biol* **3** (3), e39.

EDEN, E., NAVON, R., STEINFELD, I., LIPSON, D. & YAKHINI, Z. 2009 Gorilla: a tool for discovery and visualization of enriched go terms in ranked gene lists. *BMC Bioinformatics* **10** (1), 48.

FALCON, S. & GENTLEMAN, R. 2007 Using GOstats to test gene lists for GO term association. *Bioinformatics* **23** (2), 257–8.

FOSTER, S. S., ZUBKO, M. K., GUILLARD, S. & LYDALL, D. 2006 MRX protects telomeric DNA at uncapped telomeres of budding yeast *cdc13-1* mutants. *DNA Repair* **5** (7), 840 – 851.

FOURMENT, M. & GILLINGS, M. 2008 A comparison of common programming languages used in bioinformatics. *BMC Bioinformatics* **9** (1), 82.

FRANKE, J., GEHLEN, J. & EHRENHOFER-MURRAY, A. E. 2008 Hypermethylation of yeast telomerase RNA by the snRNA and snoRNA methyltransferase Tgs1. *J. Cell. Sci.* **121** (Pt 21), 3553–3560.

GAMERMAN, D. 1997 *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*, 1st edn. Chapman & Hall.

GAMERMAN, D. & LOPES, H. 2006 *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. Taylor & Francis.

GARDINER, C. 2010 *Stochastic Methods: A Handbook for the Natural and Social Sciences*. Springer.

GELFAND, A. E. & SMITH, A. F. M. 1990 Sampling-Based Approaches to Calculating Marginal Densities. *Journal of the American Statistical Association* **85** (410), 398–409.

GELMAN, A. 2006 Prior distributions for variance parameters in hierarchical models. *Bayesian analysis* **1** (3), 515–533.

GELMAN, A., CARLIN, J. B., STERN, H. S. & RUBIN, D. B. 2003 *Bayesian Data Analysis, Second Edition (Chapman & Hall/CRC Texts in Statistical Science)*, 2nd edn. Chapman and Hall/CRC.

GELMAN, A. & HILL, J. 2006 *Data Analysis Using Regression and Multilevel/Hierarchical Models*, 1st edn. Cambridge University Press.

GILKS, W., RICHARDSON, S. & SPIEGELHALTER, D. 1995 *Markov Chain Monte Carlo in Practice*. Taylor & Francis.

GOFFEAU, A., BARRELL, B. G., BUSSEY, H., DAVIS, R. W., DUJON, B., FELD-MANN, H., GALIBERT, F., HOHEISEL, J. D., JACQ, C., JOHNSTON, M., LOUIS, E. J., MEWES, H. W., MURAKAMI, Y., PHILIPPSEN, P., TETTELIN, H. & OLIVER, S. G. 1996 Life with 6000 genes. *Science* **274** (5287), 546–567.

GOLDSTEIN, H. 2011 *Multilevel Statistical Models*. Wiley.

GOLIGHTLY, A. & WILKINSON, D. J. 2005 Bayesian Inference for Stochastic Kinetic Models Using a Diffusion Approximation. *Biometrics* **61** (3), 781–788.

GUTIÉRREZ, R., RICO, N., ROMÁN-ROMÁN, P. & TORRES-RUIZ, F. 2006 Approximate and generalized confidence bands for some parametric functions of the lognormal diffusion process with exogenous factors. *Scientiae Mathematicae Japonicae* **64** (2), 313–330.

HASTINGS, W. K. 1970 Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57** (1), 97–109.

HAYFLICK, L. & MOORHEAD, P. 1961 The serial cultivation of human diploid cell strains. *Experimental Cell Research* **25** (3), 585 – 621.

HEIDELBERGER, P. & WELCH, P. D. 1981 A spectral method for confidence interval generation and run length control in simulations. *Commun. ACM* **24** (4), 233–245.

HEYDARI, J. J., LAWLESS, C., LYDALL, D. A. & WILKINSON, D. J. 2012 Bayesian hierarchical modelling for inferring genetic interactions in yeast. *In submission* .

HEYDARI, J. J., LAWLESS, C., LYDALL, D. A. & WILKINSON, D. J. 2013 Fast Bayesian parameter estimation for stochastic logistic growth models, arXiv:1310.5524.

HUANG, D. W., SHERMAN, B. T. & LEMPICKI, R. A. 2008 Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols* **4** (1), 44–57.

HUANG, D. W., SHERMAN, B. T. & LEMPICKI, R. A. 2009 Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research* **37** (1), 1–13.

HUBER-CAROL, C. 2002 *Goodness-Of-Fit Tests and Model Validity*. Birkhäuser Boston.

HURN, A. S., JEISMAN, J. I. & LINDSAY, K. A. 2007 Seeing the Wood for the Trees: A Critical Evaluation of Methods to Estimate the Parameters of Stochastic Differential Equations. *Journal of Financial Econometrics* **5** (3), 390–455.

ITÔ, K. 1944 Stochastic integral. *Proceedings of the Imperial Academy* **20** (8), 519–524.

JACCARD, P. 1912 The distribution of the flora in the alpine zone.1. *New Phytologist* **11** (2), 37–50.

JACKMAN, S. 2009 *Bayesian Analysis for the Social Sciences*. John Wiley & Sons.

JOHNSON, N., KOTZ, S. & BALAKRISHNAN, N. 1995 *Continuous univariate distributions. Wiley series in probability and mathematical statistics: Applied probability and statistics* v. 2. Wiley & Sons.

JONES, M. E. 1992 Orotidylate decarboxylase of yeast and man. *Curr. Top. Cell. Regul.* **33**, 331–342.

JR., M. E. T., JR., E. L. B., KIRK, K. A. & PRUITT, K. M. 1976 A theory of growth. *Mathematical Biosciences* **29** (34), 367 – 373.

KALMAN, R. E. 1960 A New Approach to Linear Filtering and Prediction Problems. *Transactions of the ASME–Journal of Basic Engineering* **82** (Series D), 35–45.

KANEKO, S., TANAKA, T., NODA, H., FUKUDA, H., AKADA, R. & KONDO, A. 2009 Marker-disruptive gene integration and URA3 recycling for multiple gene manipulation in Saccharomyces cerevisiae. *Applied Microbiology and Biotechnology* **83** (4), 783–789.

KEOGH, M. C., KURDISTANI, S. K., MORRIS, S. A., AHN, S. H., PODOLNY, V., COLLINS, S. R., SCHULDINER, M., CHIN, K., PUNNA, T., THOMPSON, N. J., BOONE, C., EMILI, A., WEISSMAN, J. S., HUGHES, T. R., STRAHL, B. D., GRUNSTEIN, M., GREENBLATT, J. F., BURATOWSKI, S. & KROGAN, N. J. 2005 Cotranscriptional set2 methylation of histone H3 lysine 36 recruits a repressive Rpd3 complex. *Cell* **123** (4), 593–605.

KIJIMA, M. 2013 *Stochastic Processes with Applications to Finance, Second Edition*. CRC Press.

KLOEDEN, P. & PLATEN, E. 1992 *Numerical Solution of Stochastic Differential Equations*. Springer.

KOLLER, M. 2012 *Stochastic Models in Life Insurance*. Springer.

KOMOROWSKI, M., FINKENSTADT, B., HARPER, C. V. & RAND, D. A. 2009 Bayesian inference of biochemical kinetic parameters using the linear noise approximation. *BMC Bioinformatics* **10**, 343.

KOWALCZYK, T., PLESZCZYNSKA, E. & RULAND, F. 2004 *Grade Models and Methods for Data Analysis: With Applications for the Analysis of Data Populations*. Springer.

KURTZ, T. G. 1970 Solutions of ordinary differential equations as limits of pure jump Markov processes. *Journal of Applied Probability* **7** (1), 49–58.

KURTZ, T. G. 1971 Limit theorems for sequences of jump Markov processes approximating ordinary differential processes. *Journal of Applied Probability* **8** (2), 344–356.

KUTNER, M. H., NACHTSHEIM, C. J., NETER, J. & LI, W. 2005 *Applied linear statistical models.*, 5th edn. McGraw-Hill Irwin.

LAWLESS, C., WILKINSON, D. J., YOUNG, A., ADDINALL, S. G. & LYDALL, D. A. 2010 Colonyzer: automated quantification of micro-organism growth characteristics on solid agar. *BMC Bioinformatics* **11**, 287.

LEEK, J. T., SCHARPF, R. B., BRAVO, H. C. C., SIMCHA, D., LANGMEAD, B., JOHNSON, W. E., GEMAN, D., BAGGERLY, K. & IRIZARRY, R. A. 2010 Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature reviews. Genetics* **11** (10), 733–739.

LEVY, M. Z., ALLSOPP, R. C., FUTCHER, A., GREIDER, C. W. & HARLEY, C. B. 1992 Telomere end-replication problem and cell aging. *Journal of Molecular Biology* **225** (4), 951 – 960.

LOTTAZ, C., YANG, X., SCHEID, S. & SPANG, R. 2006 OrderedLista bioconductor package for detecting similarity in ordered gene lists. *Bioinformatics* **22** (18), 2315–2316.

LUNN, D., THOMAS, A., BEST, N. & SPIEGELHALTER, D. 2000*a* WinBUGS - A Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing* **10** (4), 325–337.

LUNN, D. J., THOMAS, A., BEST, N. & SPIEGELHALTER, D. 2000*b* WinBUGS - a Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing* **10**, 325–337.

LUSTIG, A. J. 2001 Cdc13 subcomplexes regulate multiple telomere functions.

LYDALL, D. 2003 Hiding at the ends of yeast chromosomes: telomeres, nucleases and checkpoint pathways. *Journal of Cell Science* **116**, 4057–4065.

MANI, R., ST.ONGE, R. P., HARTMAN, J. L., GIAEVER, G. & ROTH, F. P. 2008 Defining genetic interaction. *Proceedings of the National Academy of Sciences* **105** (9), 3461–3466.

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** (6), 1087–1092.

NUGENT, C. I., HUGHES, T. R., LUE, N. F. & LUNDBLAD, V. 1996 Cdc13p: A Single-Strand Telomeric DNA-Binding Protein with a Dual Role in Yeast Telomere Maintenance. *Science* **274** (5285), 249–252.

O'HARA, R. B. & SILLANPAA, M. J. 2009 A Review of Bayesian Variable Selection Methods: What, How and Which. *Bayesian Analysis* **4**, 85.

ØKSENDAL, B. 2010 *Stochastic Differential Equations: An Introduction with Applications*. Springer.

OLOVNIKOV, A. 1973 A theory of marginotomy: The incomplete copying of template margin in enzymic synthesis of polynucleotides and biological significance of the phenomenon. *Journal of Theoretical Biology* **41** (1), 181 – 190.

OLOVNIKOV, A. M. 1996 Telomeres, telomerase, and aging: Origin of the theory. *Experimental Gerontology* **31** (4), 443 – 448.

PELEG, M., CORRADINI, M. G. & NORMAND, M. D. 2007 The logistic (Verhulst) model for sigmoid microbial growth curves revisited. *Food Research International* **40** (7), 808 – 818.

PHENIX, H., MORIN, K., BATENCHUK, C., PARKER, J., ABEDI, V., YANG, L., TEPLIAKOVA, L., PERKINS, T. J. & KRN, M. 2011 Quantitative Epistasis Analysis and Pathway Inference from Genetic Interaction Data. *PLoS Comput Biol* **7** (5), e1002048.

PHILLIPS, P. C. 1998 The Language of Gene Interaction. *Genetics* **149** (3), 1167–1171.

PINHEIRO, J. C. & BATES, D. M. 2000 *Mixed Effects Models in S and S-Plus*. Springer.

PLUMMER, M. 2003 JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*.

PLUMMER, M. 2010 rjags: Bayesian graphical models using MCMC R package version 2.1.0-10, `http://CRAN.R-project.org/package=rjags`.

PLUMMER, M., BEST, N., COWLES, K. & VINES, K. 2006 CODA: Convergence Diagnosis and Output Analysis for MCMC. *R News* **6** (1), 7–11.

R CORE TEAM 2013 *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

RAFTERY, A. E. & LEWIS, S. M. 1995 The Number of Iterations, Convergence Diagnostics and Generic Metropolis Algorithms. In *Practical Markov Chain Monte Carlo (W.R. Gilks, D.J. Spiegelhalter and S. Richardson, eds.)*, pp. 115–130. Chapman and Hall.

RICHARDS, F. J. 1959 A Flexible Growth Function for Empirical Use. *J. Exp. Bot.* **10** (2), 290–301.

ROMÁN-ROMÁN, P. & TORRES-RUIZ, F. 2012 Modelling logistic growth by a new diffusion process: Application to biological systems. *Biosystems* **110** (1), 9–21.

ROSENTHAL, J. S. 2000 Parallel computing and monte carlo algorithms. *Far east journal of theoretical statistics* **4** (2), 207–236.

SCHULDINER, M., COLLINS, S., WEISSMAN, J. & KROGAN, N. 2006 Quantitative genetic analysis in saccharomyces cerevisiae using epistatic miniarray profiles (e-maps) and its application to chromatin functions. *Methods* **40** (4), 344 – 352, chromatin and Transcriptional Regulation.

SHAY, J. W. & WRIGHT, W. E. 2005 Senescence and immortalization: role of telomeres and telomerase. *Carcinogenesis* **26** (5), 867–874.

SIMON, D. 2006 *Optimal State Estimation: Kalman, H Infinity, and Nonlinear Approaches*. Wiley.

SOON, W. W., HARIHARAN, M. & SNYDER, M. P. 2013 High-throughput sequencing for biology and medicine. *Molecular Systems Biology* **9** (1).

SPEARMAN, C. 1987 The proof and measurement of association between two things. By C. Spearman, 1904. *The American journal of psychology* **100** (3-4), 441–471.

THULASIRAMAN, K. & SWAMY, M. N. S. 1992 *Directed Graphs*, pp. 97–125. John Wiley and Sons, Inc.

TONG, A. H. & BOONE, C. 2006 Synthetic genetic array analysis in Saccharomyces cerevisiae. *Methods Mol Biol* **313**, 171–192.

TSOULARIS, A. & WALLACE, J. 2002 Analysis of logistic growth models. *Mathematical Biosciences* **179** (1), 21 – 55.

VAN KAMPEN, N. 2011 *Stochastic Processes in Physics and Chemistry*. Elsevier Science.

VERHULST, P. F. 1845 Recherches mathématiques sur la loi d'accroissement de la population. *Nouveaux mémoires de l'Academie Royale des Science et Belles-Lettres de Bruxelles* **18**, 1–41.

WALLACE, E. W., GILLESPIE, D. T., SANFT, K. R. & PETZOLD, L. R. 2012 Linear noise approximation is valid over limited times for any chemical system that is sufficiently large. *IET Syst Biol* **6** (4), 102–115.

WALLACE, E. W. J. 2010 A simplified derivation of the Linear Noise Approximation, arXiv:1004.4280.

WAN, Y., CHEN, W., XING, J., TAN, J., LI, B., CHEN, H., LIN, Z., CHIANG, J.-H. & RAMSEY, S. 2011 Transcriptome profiling reveals a novel role for trichostatin A in antagonizing histone chaperone Chz1 mediated telomere anti-silencing. *FEBS Letters* **585** (15), 2519 – 2525.

WELCH, G. & BISHOP, G. 1995 An Introduction to the Kalman Filter. *Tech. Rep.*. Chapel Hill, NC, USA.

WELLINGER, R. J. 2009 The CST complex and telomere maintenance: the exception becomes the rule. *Mol. Cell* **36** (2), 168–169.

WEST, M. & HARRISON, J. 1997 *Bayesian Forecasting and Dynamic Models*, 2nd edn. New York: Springer-Verlag.

WILKINSON, D. 2011 *Stochastic Modelling for Systems Biology, Second Edition*. Taylor & Francis.

WILKINSON, D. J. 2009 Stochastic modelling for quantitative description of heterogeneous biological systems. *Nature Reviews Genetics* **10** (2), 122–133.

WITTE, R. & WITTE, J. 2009 *Statistics*. John Wiley & Sons.

WRIGHT, W. E. & SHAY, J. W. 1992 The two-stage mechanism controlling cellular senescence and immortalization. *Experimental Gerontology* **27** (4), 383 – 389, special Issue Human Diploid Fibroblast-like Cells as a Model System for the Study of Senescence.

XU, J. 2010 *Microbial Population Genetics*. Caister Academic Press.

YANG, X., BENTINK, S., SCHEID, S. & SPANG, R. 2006 Similarities of ordered gene lists. *J Bioinform Comput Biol* **4** (3), 693–708.

YI, N. 2010 Statistical analysis of genetic interactions. *Genetics research* **92** (5-6), 443–459.

ZEYL, C. 2000 Budding yeast as a model organism for population genetics. *Yeast* **16** (8), 773–784.

ZUBKO, M. K. & LYDALL, D. 2006 Linear chromosome maintenance in the absence of essential telomere-capping proteins. *Nat. Cell Biol.* **8** (7), 734–740.

ZUUR, A., IENO, E., WALKER, N., SAVELIEV, A. & SMITH, G. 2009 *Mixed Effects Models and Extensions in Ecology with R*. Springer.