BIOINFORMATICS

Pages 1–7

Modeling Heterogeneity in Microbial Population Dynamics

Helena Herrmann[1]

[1]School of Computing Science, Newcastle University, UK

MSc Computational Systems Biology
Project Proposal for 14/04/2016
Word Count: 3620

Supervision by Dr Conor Lawless, Institute for Cell and Molecular Biosciences, Newcastle University, UK

1 INTRODUCTION

Cell growth rate is an important component of evolutionary fitness and is thus subjected to great selective forces; a reduced growth rate is generally strongly indicative of a struggling strain. Hence growth rate, a measure of how quickly cells are progressing through the cell cycle, is considered a key cell phenotype. One way of measuring growth rate is through modeling. When modeling cell population dynamics the growth rate parameter is typically measured at the population scale; a scale chosen for technical convenience.

Population scale measurements generally assume that observations are directly transferable to the single cell level. However, there is increasing evidence that, even among isogenic populations, there is considerable heterogeneity in growth rates (e.g. Pin and Baranyi, 2006; Schmidt *et al*., 2012; Levy *et al*., 2012). The idea of phenotypic heterogeneity arising through non-genetic differences (e.g. epigenetics (Bird, 2007), cell age (Ginovart *et al*., 2011), or selection pressure (Navin *et al*., 2011)) is beginning to receive much needed attention as it finds applications in modeling the dynamics of microbial infections, food security assessments, and tumorigenesis dynamics, to name a few. For example, understanding the dynamics underlying isogenic, heterogeneous cell lineages has been marked as a key requirement for developing more effective cures in cancer research (Tabassum and Polyak, 2015). The proposed work in yeast may very well provide a tractable model for such investigations.

This project aims to address the extent to which models are able to accurately capture observable levels of growth rate heterogeneity from single isogenic microbial growth curves and to explore how this inter-cellular variability affects interpretations of population growth rate. While cell growth is generally thought to go through four well-known phases (Figure 1), we would like to assess how these population level observations differ from those at the single lineage level. This project will address the lag phase and the exponential phase, where observed growth rates are most variable and have the greatest impact on apparent lineage fitness.

Models capturing observable levels of growth rate heterogeneity in isogenic microbial cultures will be developed. Novelty will arise from the fact that individual cell lineages will be analyzed separately in order to explore inherent heterogeneities. Finalized models will then be used to explore how intra-cellular variability can affect the interpretation of population growth rate observations.

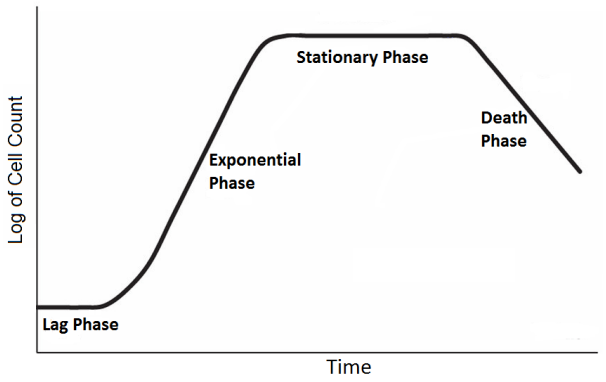

Fig. 1: Diagram of microbial growth phases: (i) lag phase, where inoculated cells adapt to their new environment, (ii) exponential growth phase, where cells divide at a constant growth rate, (iii) growth arrest phase, where system carrying capacity is reached (iv) and death phase, where viable cell counts are declining. The described phases of growth are typically observed at the population level and are often also assumed to apply at the individual cell level (Baranyi, 2002).

It is hypothesized that, when accounting for intrinsic variation in cell division time and growth rate, new mechanistic insights will be gained, since the apparent heterogeneity between populations may be drastically reduced when considering single lineage heterogeneities. In particular, the lag phase will be addressed, as this is physiologically and mathematically the least explored growth phase as of today (Rolfe *et al*., 2012). Ideally, this project will demonstrate that, due to inter-lineage variability in growth rate arising from selection, single lineage and population level observations differ.

2 AIMS

1. Find a model which best captures heterogeneity in microbial growth curves through single lineage observations.

Isogenic cell growth exhibits intrinsic stochasticities which may drown in the noise of extrinsic heterogeneities when considering population dynamics. Addressing the effect of isogenic variation on cell lineage dynamics may yield further mechanistic insight for analyzing population growth curves. Can a stochastic model which reduces the apparent heterogeneity in growth rates in the data be found?

2. Explore the implications which single lineage modeling may have on the interpretation of various growth phases.

μQFA video data obtained by Lawless provides little evidence for a lag phase at first sight (Lawless, 2013). It is suspected that when taking cell growth and division time into account, the lag phase may actually be a mere artifact of long right tail in the growth rate distribution as shown for *htz1*Δ (HTZ1_3) in Figure 2, as well as the fact that fitter strains will dominate over time as a result of selection. Can explicit modeling of cell lineages give rise to an apparent lag phase at the population level even though this may not be visible at the single cell level?

Additionally, microbiologists often sample cell populations during the exponential phase in order to improve reproducibility as fitter lineages are likely to dominate the population at this point. It is thus worth analyzing how apparent population heterogeneity affects sampling from different phases.

3. Depending on resource availability and findings, repeat the micro Quantitative Fitness Analysis (μQFA) experiments with the aim of validating model predictions.

Experimentation will involve undertaking microscopic observations of clonal *Saccharomyces cerevisiae* cells to capture the lag and exponential phase of growth in great detail. Existing in-house image analysis software will be used for generating growth curves from the data, where the measured area of clonal cell cultures translates to the total number of cells at a given time. Following an iterative systems biology protocol, this step would allow for cells to be grown at a higher or lower density than the current data, in order to test the predictions and generality of the proposed models.

3 PROPOSED RESEARCH

3.1 Obtained Data

To capture growth rate heterogeneity, a range of models will be developed and their validity tested against the against available data. In order to ensure that the analyses hold under a range of experimental conditions and genetic backgrounds, three data sets will be explored: (i) μQFA data produced by Lawless (unpublished), (ii) high-throughput microscopy assay data produced by Levy *et al*., 2012 and (iii) Ziv *et al*., 2013. All data sets consist of *S. cerevisiae* micro-colony growth curves.

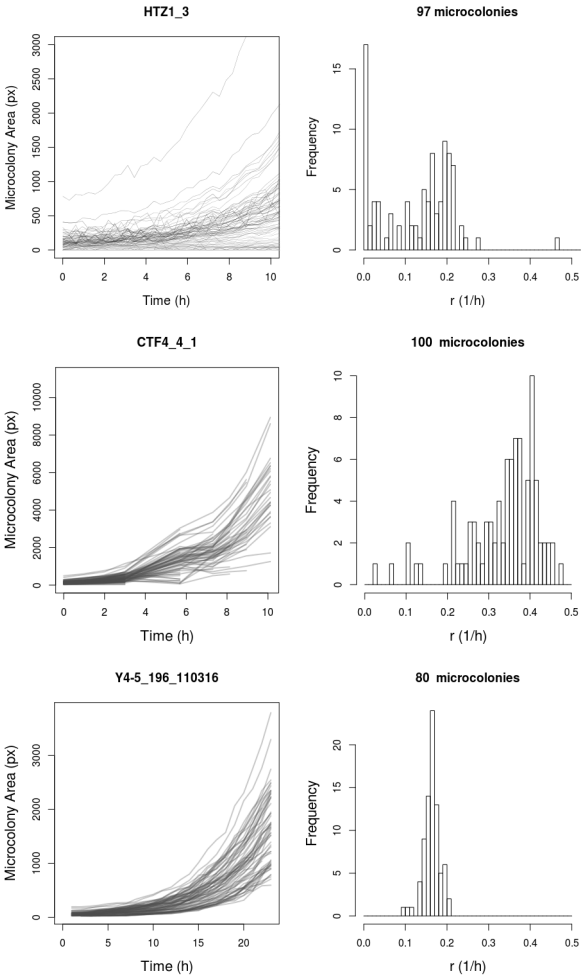

Fig. 2: Example outputs of clonal colony growth curves and the associated growth rate frequencies generated from each of the three data sets: Lawless (top), Levy *et al*., 2012 (middle) and Ziv *et al*., 2013 (bottom). Growth rates were estimated using the `lm` linear regression function in R, whereby the growth rate is equal to the slope of log(area) $\sim$ time.

Clonal cultures are lineages derived from a single cell. Population pooling examines purified cell populations in order to learn about the behavior of single cells, thereby assuming that all members of the population behave in the same way. However, as shown in Figure 2, which displays growth curves obtained from clonal cultures grown in the same microplate well, this assumption is difficult to justify. Figure 2 shows that all three of the above data sets provide evidence for existing heterogeneities within clonal cultures, none of which has been

detailed in a publication at the single lineage level. Starting populations are proposed to be split up into individual cells and treated as separate experiments. This will allow for the analysis of inherent stochasticities, which would otherwise drown in population noise.

3.2 Types of Models

Proposed model types include a logistic, deterministic model (Verhult, 1845), a standard Gillespie, birth-only stochastic model (Bailey, 1964; Gillespie, 1977), and a hybrid model which will combine the previous two models. Model implementations will be done in R and Python and should be relatively straightforward.

A traditional, deterministic, logistic population growth model will be taken into consideration for its computational speed, which will be of a great advantage considering the vast amount of available data.

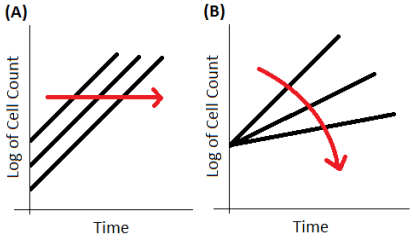

Fig. 3: Types of heterogeneity in the exponential phase which may arise between growth curves as indicated by the red arrows: (A) heterogeneity arising from a difference in starting population at the beginning of the exponential phase, and (B) heterogeneity arising from a difference in growth rate.

However, since cell lineage intrinsic growth-rate heterogeneities should also be taken into consideration, these will have to be captured using a stochastic model. A deterministic model will only be able to capture exponential growth heterogeneities arising from differences in the onset of the exponential phase as illustrated in Figure 3 (A), whereas a stochastic model will additionally be able to capture inter-lineage growth rate heterogeneities, as sketched in Figures 3 (B). Thus, being able to compare stochastic and deterministic modeling may lead to further conclusions about where the greatest heterogeneities lie within in the data and how these affect population level observations. Much of the theoretical ground work for stochastic growth models has been laid by Baranyi (1997, 1998, 2002); however, its application to high-throughput microbial sets of increased precision, which have become available since, remains limited.

3.3 Generating Biologically Meaningful Models

Since the experimentally observed growth of microbial colonies (and, incidentally, that of human fibroblast populations) occurs in a monotonically increasing fashion during the lag and exponential phase, birth-only models will be used, deviating from more traditional birth and death analyses. This will be the first time that birth-only models will be considered in the context of stochastic cell growth.

Furthermore, it may prove sensible to set a lower bound for the time sampling step in the stochastic algorithm (deviating from the original Gillespie algorithm; Gillespie, 1977). Doing so will allow us to incorporate the expected time required for cell growth and division, as an infinitesimal growth and division time holds little biological meaning in the context of cell growth.

Although models are generally proposed to follow a birth-only process (Bailey, 1964), it may also be of interest to see whether there is a low rate of death which can still give rise to monotonically increasing growth curves at the population level and whether this should be incorporated into the model. Alternatively it may be worth incorporating a transition to a non-dividing state, whereby cells remain in the population but cell division is suspended. For example, inocula suspected to be non-dividing cells which remain part of the population as observed in the μQFA data have been circled red in Figure 4.

Lastly, in order to outweigh the computational cost associated with a stochastic model, a hybrid model combining the two approaches is also proposed to be validated against the data. For example, it may be found that most of the inter-lineage variability arises early on in cell growth and thus a stochastic model will initially be considered, but that after a few initial divisions, growth curves closely follow a deterministic model, at which point it may seem sensible to induce a model switch.

3.4 Parameter Inference

Likelihood-free parameter inference techniques will be explored in order to obtain the required parameters from each of the data sets. Various options such as `pyMC` for deterministic modeling, or the particle MCMC described by Wilkinson, 2006, for stochastic modeling, are available foundations to work off. Workflows for parameter inference will be documented for ease of model development. Evidently, only a subset of the available data can be used for parameter inference, as the rest will be required for model testing. Due to the huge amount of data at disposition, up to half of the data (e.g. 2 out of 4 replicates) from each of the sets can safely be used to train the models, as this will leave plenty of growth curves to validate the model against.

From previous research done at Newcastle University, there exists a full QFA data set (Addinall *et al.*, 2011); thus all of the available microcolony growth curves from the μQFA data can be used to obtain the models. Trained models can then be used to simulate a QFA population growth curve and compare predictions (based on the single lineage μQFA data) with the population QFA observations. This will have the

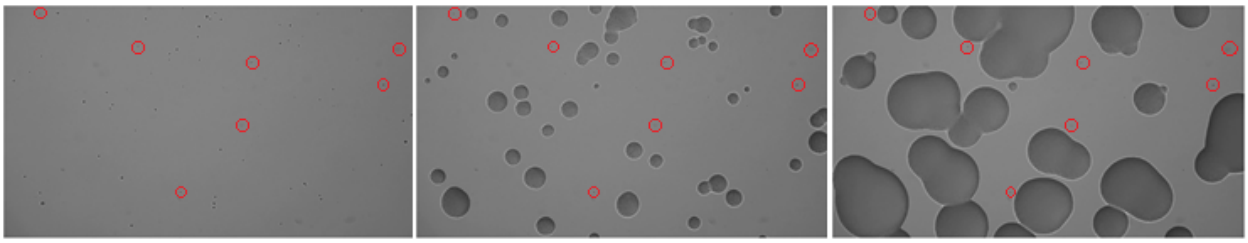

Fig. 4: Colony growth based on the μQFA data after inoculation (left), after $\sim$ 1 hour (center) and after $\sim$ 10 hours (right), where non-growing inocula (suspected non-dividing cells) have been circled in red.

added advantage of showing the distinction between single lineage averages and observed population growth rates driven by selection.

3.5 Model Fit

Finalized models will be validated using either the Bayesian Information Criterion (BIC) (Schwarz, 1978) or Bayes factors (Kass and Raftery, 1995). Christensen *et al*., 2011, provide a useful introduction to both model comparison techniques. Both the BIC and Bayes factors select from a range of models the one which corresponds to the greatest posterior probability, the difference lying in that the BIC does not require explicitly specified priors whereas Bayes factors do (Bollen *et al*., 2012). An appropriate comparison technique will thus have to be chosen based on confidence in establishing prior distributions. Notably, although very similar in implementation, the BIC would always be chosen over the Akaike Information Criterion (AIC) as it penalizes over-fitting more stringently (Burnham and Anderson, 2002), which will prove valuable when considering models with additional biological parameters as described above. As always, following a Bayesian paradigm over a frequentist one not only incorporates estimation uncertainty but also parameter uncertainty (Christensen *et al*., 2011). The model with the best fit will then be used to explore the mechanistic implications of microbial cell growth.

3.6 Model Predictions and Mechanistic Implications

Upon parameter realization and model validation, mechanistic implications can be analyzed. Additionally, it may prove valuable to return to the lab in order test model predictions. Freely available *ibidi* microscopy slides (16 well), which can be used for generating microbial growth curves, can be used. For the μQFA experimental design, please refer to Lawless, 2012.

As for the lag phase, there currently exists no explicit definition in the literature. Divergent definitions either define the lag phase to occur before the first cell division (Baty and Delignette-Muller, 2004) or to last over multiple cell divisions during which exponential growth has not yet begun (Pin and Baranyi, 2006). Rolfe *et al*., 2012 are one of the only research groups that define (yet do not act upon) this distinction, using the terms *lag phase* and *delay phase*. These discrepancies can be even further subdivided in that some texts within the existing literature consider a biological definition of the lag phase (time until maximum growth acceleration; Buchanan and Cygnarowicz, 1990) or a mathematical definition (time until the tangent to the growth curve intersects that of the exponential growth phase; Baranyi, 2002).

It is hypothesized that the lag phase itself may not be visible in single lineages, but arises as an artifact of intrinsic noise at the population level. It is suspected that when using stochastic modeling to reduce the apparent heterogeneity in the data, the lag phase is merely the result of a long-tailed growth rate distribution. Correlations between a short lag phase and yeast cell age (negative correlation) and a short lag phase and inoculum size (positive correlation) exist, implying that cell growth and cell division can indeed occur almost immediately after inoculation (Ginovart *et al*., 2011). Because the μQFA data by Lawless provides very little evidence for a lag phase *per se* (Lawless, 2013), it is suspected that the lag phase can be apparent at the population level, despite being absent at the single lineage level. In the μQFA data, cells divide almost immediately, with some clones dividing more quickly than others. Thus it is suspected that the lag phase arises at the population level as a result of inter-lineage variability in growth rate which arises as a result of competition between clonal cell lineages. It may very well be that as a result of competition, a wide range of growth rates is observed, which in turn results in very noisy population observations.

A second mechanistic implication worth investigating lies in the standard practice of micro-biologists to sample cell populations from dynamic growth phases rather than stationary ones in order to improve the reproducibility of their results. Given the dynamic aspect of cell growth it would be false to conclude that the maximum likelihood estimator of the initial growth rate distribution applies over time. This is due to the fact that weaker strains (i.e. those with a reduced growth rate) are slower at producing offspring; thus, after a given amount of time the population will be dominated by faster growing strains. This is a likely explanation for why it has been noted that samples later on during the time course (i.e. the exponential phase) are more reproducible. However, sampling from the exponential phase again assumes members of the population to behave in the same way. Looking at single cell lineages as

proposed will allow for assessing the implications of sampling at different growth phases and thus analyzing the validity of experimental design procedures.

4 OBJECTIVES

The above proposed research has been summarized in work packages shown below. The outlined objectives provide a step-by-step guidance for project advancement. All workflows will be documented and all models will be packaged to maximize accessibility, reproducibility and impact in the wider research community.

Model packaging will occur during the development stage so that finalized models are used for the analysis steps. Dissertation write-up will occur throughout the project in the form of blog posts, whereby each stage as outlined in the objectives is to be completed before moving on to the next one. A full time-line for each of the objectives is provided; Figure 5 displays a Gantt chart following these objectives. Time assigned for task completion is mostly generous as this project allows for much refinement and addition of detail should there be time (e.g. testing models with more biological information and obtaining further experimental data). One week for catching-up is scheduled in to account for unexpected difficulties, should these arise.

1. **Preparation and Data Exploration**

   a. Ensure data accessibility by converting all three data sets into a general format which can be accessed using R and Python.

   b. Write scripts to extract growth curves from each of the data sets. Look at pulling out single growth curves and pulling out all growth curves for a single spot or genotype.

   c. Use calibration curves to convert area vs. time curves to cell count vs. time curves.

   Write up: Stage 1 Introduction and Background Reading (reuse project proposal). Generate plots to visualise the raw data for why the impact of heterogeneity is being researched.

2. **Model Development**

   a. Develop parameter inference workflows to learn about microbial growth rates.

   b. Fit a deterministic model to all data using a Bayesian hierarchical model in `pyMC` using a subset of the available data.

   c. Fit a stochastic birth-only model to a subset of the available data.

   d. Fit a hybrid model to to a subset of the available data. Can a sensible cut-off for model switching be determined?

   e. Remaining data can then be used to explore heterogeneity; for example, the *his3*Δ and the *htz11*Δ strains are considered in all three data sets. Consider comparing μQFA and QFA data.

   f. Can the models be improved by adding more biological information? Consider using the time required for cell growth and division as a lower bound for time sampling in the algorithm of the stochastic model. Potentially consider more complicated models which include some kind of death, or alternatively a switch to a non-dividing

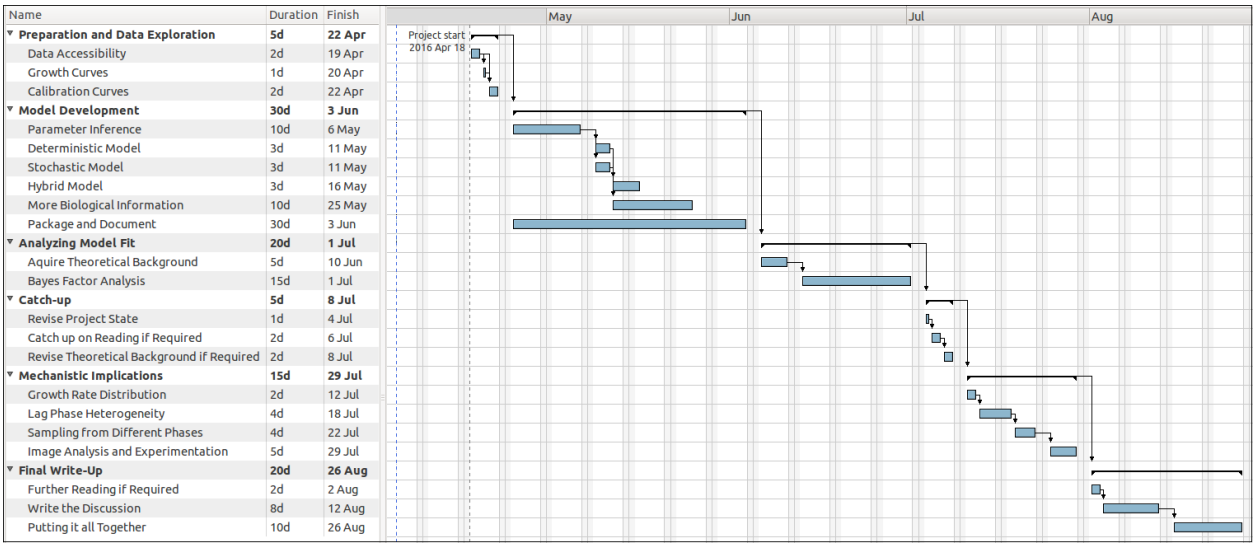

Fig. 5: Gantt chart for the above outline objectives for completing the MSc Dissertation by August 26th, 2016.

state, while maintaining monotonic increasing growth curves. Is there a low rate of death which can still give increasing curves?

g. Package and document the models.

Write up: Stage 2 - Model development. State model assumptions, implications, and validity. Analyze how these model fits differ and why. Generate plots to demonstrate the accuracy of each of the developed models. State each of the required parameters and the biological significance of the parameters in the context of the model.

3. **Analyzing Model Fit**

a. Which model seems to exhibit the closest fit to the data?

b. Compare model fit using Bayesian Information Criterion (BIC) or Bayes Factors. Both naturally penalize for over-fitting.

Write up: Stage 3 Model Exploration and Validity

4. **Exploring the Mechanistic Implications**

a. Explore heterogeneity in the data; analyze the growth rate distribution.

b. How much of the apparent heterogeneity in growth rate is reduced when considering stochasticity?

c. Does a lag phase *per se* even exist within the data or can this be explained by heterogeneity?

d. See how sampling from different phases is affected by heterogeneity.

e. Image analysis and experimentation using *ibidi* sample microscopy slides; try and obtain μQFA data with higher resolution in order to further prove model assumptions (depending on resource availability).

Write up: Stage 4 Mechanistic implications on lag phase. Include figures that emphasize each of the concluded implications.

5. **Finalizing**

a. Submission to *arXiv* and possible journal submission.

Write up: Stage 5 Discussion & putting it all together; final Dissertation for submission.

5 RESEARCH SIGNIFICANCE

With the availability of high-throughput technology, microbiology is progressing to become a data-rich science. This leads to the limiting factors in scientific advances no longer resting in the amount of available data but in the quantitative analyses performed on them. This project makes efficient use of a vast range of existing, expensive, experimental data sets by approaching them in a new way. As outlined, novelty lies in that single cell lineages will be considered in order to explore population intrinsic heterogeneities. This will be the first time that single-lineage stochastic, deterministic and hybrid models describing microbial growth will be considered alongside each other. Furthermore, if the apparent heterogeneity in population parameters is reduced by considering single lineage variations in clonal cell cultures as proposed, the explored mechanistic implications and the developed models will have vast applications ranging from experimental design procedures to quantitative risk assessments in food security and tumorigenesis treatments.

REFERENCES

Addinall,S.G., Holstein,E.M., Lawless,C., Yu,M., Champan,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**, e1001362.

Baranyi,J. (1997) Simple is good as long as it is enough. *Food Microbiol.*, **14**, 189-192.

Baranyi,J. (1998) Comparison of stochastic and deterministic concepts of bacterial lag. *J. of Theor. Biol.*, **192**, 403-408.

Baranyi,J. (2002) Stochastic modelling of bacterial lag phase. *Int. J. Food Microbiol.*, **73**, 277-294.

Baty,F., Delignette-Muller,M.-L. (2004) Estimating the bacterial lag time: which model, which precision? *Int. J. Food Microbiol.*, **91**, 261-277.

Bailey,N.T.J. *The Elements Of Stochastic Processes*. New York: Wiley, 1964. Print.

Bird,A. (2007) Perceptions of epigenetics. *Nature*, **447**, 396-398.

Bollen,K.A., Ray,S., Zavisca.J., Harden,J.J. (2012) A comparison of Bayes factor approximation methods including two new methods. *Sociol. Methods Research*, **41**, 294-324.

Buchanan,R.L., Cygnarowicz,M.L. (1990) A mathematical approach toward defining and calculating the duration of the lag phase. *Food Microbiol.*, **7**, 237-240.

Burnham,K.P., Anderson, D.R. *Model selection and multimodel inference*.New York: Springer, 2002. Print.

Christensen,R., Johnson,W., Branscum,A., Hanson.T.E. *Bayesian ideas and data analysis: an introduction for scientists and statisticians.* Boca Raton: CRC Press Taylor & Francis, 2011. Print.

Gillespie,D.T. Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.*, **81**, 2340-2361.

Ginovart,M., Prats,C., Portell,X., Silbert,M. (2011) Exploring the lag phase and growth initiation of a yeast culture by means of an individual-based model. *Food Microbiol.*, **28**, 810-817.

Kass,R.E., Raftery,A.E. (1995) Bayes factors. *J. Americ. Statistic. Assoc.*, **90**, 773-795.

Lawless,C. (2012) μQFA: Like QFA only awesomer. *SBSB Seminar*, Newcastle University, UK; http://lwlss.net/talks/uqfa.

Lawless,C. (2013) A discrete stochastic logistic model of cell lineages. *SBSSB Seminar*, Newcastle University, UK; http://lwlss.net/talks/discstoch.

Levy,S.F., Ziv,N., Siegal,M.L. (2012) Bet hedging in yeast by heterogeneous, age-correlated expression of a stress protectant. *PLoS Biol.*, **10**, e1001325.

Navin,N., Kendall,J., Troge,J., Andrews.P., Rodgers,L., McIndoo,J., Cook,K., Stepansky,A., Levy,D., Esposito,D., Muthuswamy,L., Krasnitz,A., McCombie,W.R., Hicks,J., Wigler,M. (2011) Tumor evolution inferred by single-cell sequencing. *Nature*, **472**, 90-94.

Pin,C., Baranyi,J. (2006) Kinetics of single cells: Observation and modeling of a stochastic process. *Appl. Environ. Microbiol.*, **72**, 2163-2169.

Rolfe,M.D., Rice,C.J., Lucchini,S., Pin,C., Thompson,A., Cameron,A.D.S., Alston,M., Stringer,M.F., Betts,R.P., Baranyi,J., Peck,M.W., Hinton,J.C.H. (2011) Lag phase is a distinct growth phase that prepares bacteria for exponential growth and involves transient metal accumulation. *J. of Bacteriol.*, **194**, 686-701.

Tabassum,D.P., Polyak,K. (2015) Tumorigenesis: it takes a village. *Nature Reviews*, **15**, 473-483.

Schmidt,M., Creutziger,M., Lenz.P. (2012) Influence of molecular noise on the growth of single cells and bacterial populations. *PLoS One*, **7**, e29932.

Schwarz,G.E. (1978) Estimating the dimension of a model. *Annals. of Statistics*, **6**, 461-464.

Verhulst,P.-F. (1845) Recherches mathmatiques sur la loi d'accroissement de la population [Mathematical Researches into the Law of Population Growth Increase]. *Nouveaux Mmoires de l'Acadmie Royale des Sciences et Belles-Lettres de Bruxelles*, **18**, 1-42.

Wilkinson,D.J. *Stochastic Modelling for Systems Biology.* Boca Raton: Taylor & Francis, 2006. Print.

Ziv,N., Siegal,M.L., Gresham,D. (2013) Genetic and nongenetic determinants of cell growth variation assessed by high-throughput microscopy. *Mol. Biol. Evol.*, **30**, 25682578.