Modelling Competition for Nutrients between Microbial Populations Growing on Solid Agar Surfaces

Author: Daniel Boocock; Supervisor: Dr Conor Lawless

August 19, 2016

ABSTRACT

Motivation: Growth rate is a major component of the evolutionary fitness of microbial organisms. When nutrients are plentiful, fast-growing strains come to dominate populations whereas slower-growing strains are wiped out. This makes growth rate an excellent (a useful) surrogate for the health of cells. Measuring the health of cells grown in different genetic backgrounds or environments can inform about genetic interaction and drug sensitivity. In high-throughput procedures such as QFA and SGA, arrays of microbial cultures are grown on solid agar plates and quantitative fitness estimates are determined from growth measurements. Diffusion of nutrients along gradients in nutrient density arising between fast- and slow-growing neighbours is likely to affect growth rate and fitness estimates. However, current analyses assume that cultures grow independently. We study data from QFA experiments growing *Saccharomyces cerevisiae* to test a mass action kinetic model of nutrient dependent growth and diffusion. We try to correct for competition to provide more accurate and precise fitness estimates.

Results: Don't know what to say yet.

Availability and Implementation: CANS, a Python package developed for the analysis in this paper, is freely available at https://github.com/lwlss/CANS.

1 INTRODUCTION

The bacteria *Eschericia Coli* and yeast *Saccharomyces cerevisae* are unicellular organisms studied as a model prokaryote and eukaryote respectively. Bacteria and yeast grow in colonies, where cells may (be clones originating from a single cell or) belong to different genetic strains originating from different individual cells. In favourable conditions, growth is exponential and this makes growth rate a major component of fitness; faster growing strains quickly come to dominate the population. At a certain point growth becomes limited and a stationary phase is reached. For unicellular organisms, growth rate is equal to cell cycle progression rate and all of the genetic information must be copied before each division. As a result, evolutionary pressure has led to rapidly dividing organisms with compact genomes of essential genes. These genes have been conserved in other species over billions of years of evolution, which is, in part, what makes *E. Coli* and *S. cerevisae* useful as model species. The eukaryote *S. cerevisae*, is particularly useful for the study of other eukaryotes such as humans.

The growth rate of microbial organisms is measurable and is often used to determine fitness. In experiments, cell cultures are commonly grown in two types of medium: on the surface of a nutrient rich solid agar and in a liquid mixture containing nutrients. (REMOVE: In spot tests (phenotypic array), cultures are pinned or inoculated on the surface of a solid agar containing nutrients. In liquid culture assays, cultures are mixed in a liquid medium containing nutrients.) In both cases cultures are incubated and growth is observed. Identical strains can grow differently between the two mediums and disagreement in fitness estimates is currently an issue Baryshnikova *et al.* (2010a) (I couldn't find a paper specifically talking about this issue but they have a correlation plot Fig2a where correlations are worse with a liquid culture study by Jasnos and Korona; in fact the Baryshnikova paper Fig3c seems to say that they had strong correlation in their "high-resolution liquid growth profiling study"). I do not focus on this issue and exclusively study fitness screens using solid agar.

Fitness estimates can be used to infer genetic interaction or drug response and high-throughput methods allow this to be conducted on a genome-wide scale (see e.g. Costanzo *et al.* (2010); Andrew *et al.* (2013)). In a typical genetic interaction screen a strain is made with a mutation in a query gene. Double mutants are created by introducing a second deletion in this strain. By comparing the growth of double mutants with a control containing a neutral deletion, genetic interactions can be inferred. If a strain is fitter than the control then the deletion is said to suppress the defect of the query gene. If a strain is less fit than the control then the deletion is said to enhance the defect of the query gene. Either scenario suggests that the two genes interact and have a related function. Due to redundancy, single deletions are often non-lethal. (Remove: Knock downs and conditional mutations can also be used.) This has allowed Costanzo *et al.* (2010) to explore genetic interactions for $\sim$75% of the *S. cerevisae* genome.

Synthetic Genetic Array (SGA) and Quantitative Fitness Analysis (QFA) are high-throughput methods for obtaining quantitative fitness estimates of microbial cultures grown on solid agar (Baryshnikova *et al.*, 2010b; Banks *et al.*, 2012). Typically one query gene and replicates of several deletions are pinned or inoculated in a rectangular array on a solid agar plate. Many plates with different query genes and deletions are grown in high-throughput to explore whole genomes. I study data from QFA which refers to quantitative estimation of fitness by measurement and fitting of growth curves. In a typical QFA procedure liquid cultures are incolulated onto solid agar (containing nutrients (already mentioned above)) in a 16x24 rectangular array of 384 spots. Inoculum density can be varied to capture more or less of the growth curve and the most dilute cultures are inoculated with $\sim$100 starting cells (Addinall *et al.*, 2011). Plates are grown in incubation and removed to be photographed at timepoints throughout growth.

Photographs are of whole plates and growth typically covers several days to capture both the exponential and stationary growth phases. Colonyzer (Lawless *et al.*, 2010) processes optical density measurements in photographs to produce a timecourse of cell density estimates for each culture. In pasts analysis, the logistic growth model was independently fit to the timecourse of each culture and fitness estimates were defined in terms of parameters of this model: the growth constant r and carrying capacity K. In contrast, SGA typically uses a larger array of 1536 pinned cultures and a single endpoint assay of culture area to quantify growth. The the differential form and solution of the logistic model (Verhulst, 1845) (probably don't need this reference) are given in Equations 1, where C represents cell density and C_{t_0} is cell density at time zero.

$$\dot{C} = rC\left(1 - \frac{C}{K}\right) \tag{1a}$$

$$C(t) = \frac{KC_{t_0}e^{rt}}{K + C_{t_0}(e^{rt} - 1)} \tag{1b}$$

The logistic model is a simple mechanistic model describing self-limiting growth and has a sigmoidal solution. Growth begins exponentially with rate rC and curtails as the population size increases and cells begin to compete for space and nutrients (remove: or interact in some other way). Cell density reaches a final carrying capacity K at the stationary phase. In QFA, nutrients must diffuse through agar to reach cells growing on the surface. It is plausible that the carrying capacity K represents the point at which nutrients either run out or growth becomes limited by the diffusion of nutrients and is approximately stationary. Fitting the logistic model to QFA data requires plate level or culture level parameters for C_{t_0} and culture level parameters for r and K making 769 or 1152 parameters per 384 culture plate.

//Could remove and just discus MDR when I get to the results// The growth constant r could be used as a fitness measure. However, Addinall *et al.* (2011) define a more complicated fitness measure as the product of Maximum Doubling Rate (MDR) and Maximum Doubling Potential (MDP) which they calculate from logistic model parameters. MDR measures the doubling rate at the beginning of the exponential growth phase, when growth is fastest, and MDP is the number of divisions which a culture undergoes from inoculation to the stationary phase.

$$MDR = \frac{r}{log\left(\frac{2(K-C_0)}{K-2C_0}\right)} \tag{2a}$$

$$MDP = \frac{log\left(\frac{K}{C_0}\right)}{log(2)} \tag{2b}$$

To improve the quality of fits, QFA now uses the generalised logistic model which requires an extra shape parameter for each culture. Standard and generalised logistic model r are not equivalent so comparison relies on MDR and MDP as fitness measures. The analysis of QFA data using both models is available through the QFA R package (Lawless *et al.*, 2016). //Could remove and just discus MDR when I get to the results//

//Could remove//Addinall *et al.* (2011) used QFA and *S. cerevisae* to screen for genes involved in telomere stability which is related to ageing and cancer and has implications for human health and disease. Hits from this study have been successfully followed to discover new biology (Holstein *et al.*, 2014). (To be honest I have no idea about the significance of what they found in this paper. We had a more general focus. If I have room I should probably try and sell the potential benefits and past successes of QFA a bit more to expand the motivation. Obviously I will mention the Addinall paper when I describe p15 in the methods.) //Could remove//

Since QFA aims to determine differences in the fitness of microbial strains from measurements of differences in growth, fast and slow growing cultures are often grown side-by-side. Figure 1 shows a section of a QFA plate from a study by Addinall *et al.* where this is the case. (Cultures were inoculated with approximately equal cell density but have grown at different rates to visibly different sizes after $\sim$2.5 days.) Despite starting with the same amount of nutrients and growing at different rates there is a characteristic timescale for the cessation of growth. This suggests a global growth limiting effect which I believe to be caused by an interaction between cultures. I test the hypothesise that the interaction is a competition effect due to the diffusion of nutrients along gradients formed between fast and slow growing neighbours. This has implications for growth estimates; competition will cause growth to appear faster or slower for each neighbour than would be observed if they were grew independently. The experiment shown in Figure 2 provides further support for a nutrient competition effect. The same cultures were grown in alternate columns on two separate plates but with cultures added or removed from the neighbouring columns inbetween. Cultures in Figure 2a), where neighbours were removed, grew faster and larger (how much? I can look at the data myself) than the same cultures in Figure 2b), where neighbours were added. This suggests that an interaction between neighbours is present and may be affecting fitness estimates. Current QFA analysis using the logistic model assumes that cultures grow independently and ignores possible competition effects between neighbours. The sigmoidal curve of the logistic model poorly fits QFA data in many cases and this may be due to competition effects. I aim to fit a network model of nutrient dependent growth and diffusion to QFA data to try to correct for competition and increase the accuracy and precision of fitness estimates.

Could explain the difference in dilute and more concentrated cultures. In the image captions or elsewhere?

Could also talk about quorum sensing and amonia when I get to competition.

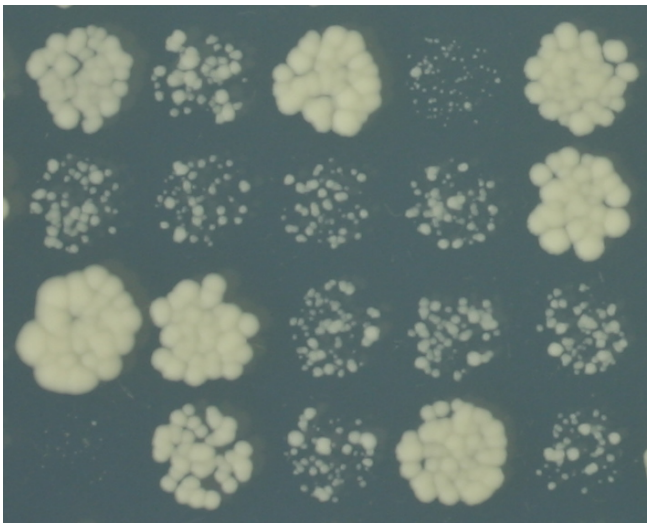

Figure 1: 4x5 section of a QFA plate. *Taken from a 16x24 format solid agar plate inoculated with dilute S. cerevisae cultures. Image captured at ∼2.5d after inoculation and incubation at 27°C.*

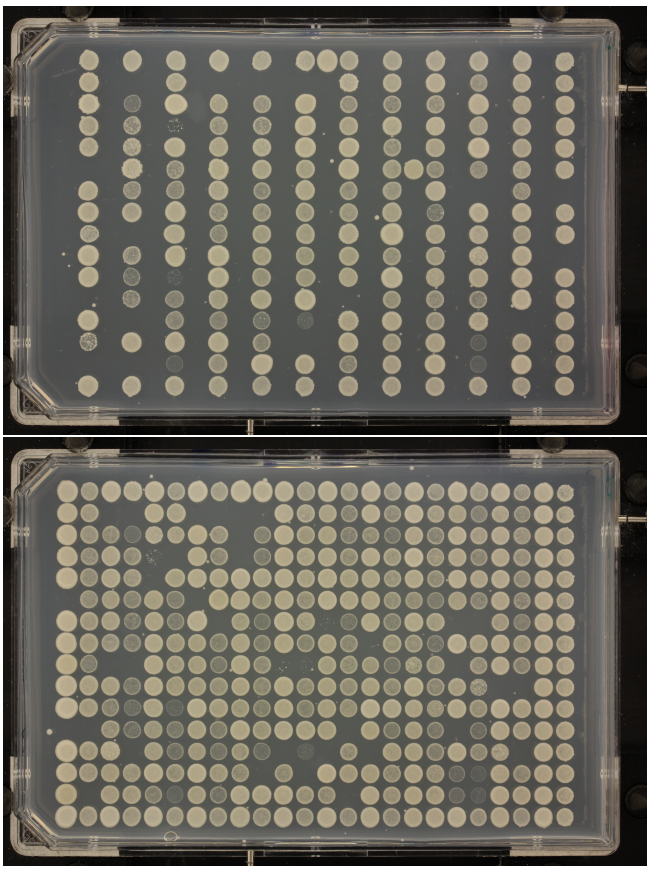

Figure 2: QFA experiment designed to examine competition. *A) QFA plate inocluated with a more concentrated S. Cerevisae inoculum (no cells inocluated on alternate columns). B) Same as in A, but with strains of similar growth rate inoculated in the positions missing in A.*

Competition effects could be dealt with experimentally by randomising the location of cultures on repeated plates. This does not require explicit knowledge or modelling of the source of competition but reduces throughput, so, if possible, a modelling approach is desirable. Poisoning of cultures by a signal molecule such as ethanol, which *S. cerevisae* produces in the metabolism of sugars by fermentation, is another possible source of competition. QFA does not measure nutrients or signal, so if more than one source of competition exists, it becomes very difficult to fit a model and randomisation may be the best approach. QFA data for edge cultures is noisy due to reflections from plate edges. This is only partially corrected for by Colonyzer (Lawless *et al.*, 2010) and as a result data for edge cultures is usually discarded. Addinall *et al.* (2011) grow repeats of a neutral deletion in edge locations, rather than leaving them empty, because of concerns about competition. In an SGA study, Baryshnikova *et al.* (2010a) use statistical techniques to correct for competition between fast and slow growing neighbours in end-point assays of culture area. I hope that modelling competition for nutrients explicitly will better correct for competition using fewer repeats. QFA uses more information than SGA by fitting whole growth curves, rather than a single endpoint assay, so a modelling approach promises to be more powerful. Furthermore, modelling may identify and explain the source of competition. Simulation of an accurate model will allow comparison of experimental designs and exploration of ways to reduce competition effects.

//Diffusion Equation: I am probably going to have to repeat this when I get to the discussion so I could just leave until then.// Reo and Korolev (2014) use a diffusion equation model to simulate nutrient dependent growth of a single bacterial culture on a pertri dish in two-dimensions. They create a sink for nutrients from culture growth and equate the flux of nutrients through culture area with the rate of increase in culture size. They model culture area as varying and keep culture density constant. (This model could be adapted for QFA by keeping culture area constant and allowing culture density to vary.) However, it would be too computationally intensive to fit a similar model to a full QFA plate in three-dimensions, especially if the model is to be used to process many plates from high-throughput experiments. Therefore, a simpler model of nutrient diffusion is required. //Diffusion Equation//

Lawless (link blog) proposed a model of nutrient dependent growth and competition (3,4), hereinafter the competition model, using mass action kinetics and network diffusion. A schematic of the model is drawn in Figure 3. He represents the nutrient dependent division of cells with the reaction equation,

$$C + N \xrightarrow{b} 2C, \tag{3}$$

where C is a cell, N is the amount of nutrient required for one cell division, and b is a rate constant for the reaction. (The identity of the limiting nutrient N is unknown but possible candidates are sugar and nitrogen.) He defines separate reactions (3) with growth constant b_i for each culture, indexed i, on a plate and uses mass action kinetics to derive rate equations for the amount of cells and nutrients associated with each culture, C_i and N_i. This gives the rate equation for C_i

(4a) and the first term in the rate equation for N_i (4b).

$$\dot{C}_i = b_i N_i C_i, \tag{4a}$$

$$\dot{N}_i = -b_i N_i C_i - k_n \sum_{j \epsilon \delta_i} (N_i - N_j). \tag{4b}$$

To arrive at the full competition model, he models the diffusion of nutrients along gradients between a culture i and its closest neighbours δ_i by the second term in (4b), where k_n is a nutrient diffusion constant. This can also be expressed as a series of reactions of the form

$$N_i \xrightarrow{k_n} N_j \tag{5a}$$

$$N_j \xrightarrow{k_n} N_i, \tag{5b}$$

and modelled with mass action kinetics. Unlike the logistic model (1), the competition model has no analytical solution, and must instead be solved numerically. If k_n is set to zero, the competition model reduces to the mass action equivalent of the logistic model, hereinafter the mass-action logistic model, (and has the same sigmoidal solution). (In this limit, parameters of the competition model can be converted in terms of parameters the logistic model (see methods section)). When the competition model is fit to QFA data, C_i is observed and N_i is hidden. Inoculum density, C_{t_0}, is often below detectable levels. By assuming that incolulum density is the same for all cultures and that nutrients are distributed evenly throughout the agar at time zero, plate level initial values of cells and nutrients, C_{t_0} and N_{t_0}, can be used. k_n is assumed to be constant across the plate but must be inferred. There is a growth constant, b_i, for each of 384 cultures on a typical QFA plate making 387 parameters in total. The competition model shares more information between cultures and has less than half the number of parameters of either the standard or generalised logistic model (Banks *et al.*, 2012; Lawless *et al.*, 2016). (If I remove the section on MDR and the generalised logistic model above I will need to add a line of explanation here.)

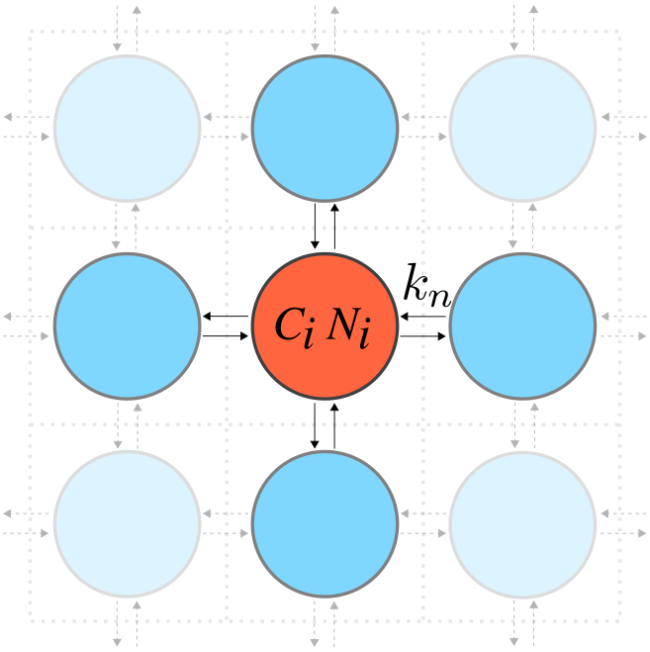

Figure 3: Schematic of the competition model. *Each circle represents a culture, indexed i, growing in a rectangular array on the surface of a nutrient containing solid agar. Arrows represent a network of nutrient diffusion along gradients between cultures. C_i - amount of cells; N_i - amount of nutrients; k_n - plate level nutrient diffusion constant; darker blue circles δ_i - closest neighbours of culture i.*

In QFA, populations begin with ∼100 cells and quickly grow to reach thousands of cells so a deterministic approximation appears valid. Mass action kinetics applies to reactions in a well stirred mixture and is perhaps less valid for a culture growing on solid agar. However, a mass action approximation has been successful in other situations where this assumption is questionable: in the Lotka-Volterra model of predator-prey dynamics (Berryman, 1992) and in signalling and reaction models inside cells (Aldridge *et al.*, 2006; Chen *et al.*, 2010). The order of a reaction also affects the rate equation and the identity and quantity of the nutrient molecule in the (3) is unknown. Reaction (3) also assumes that all nutrients are converted to cells and includes no model of metabolism. I justify the use of the competition model because in the independent limit it has the same solution as the logistic model which has long been used to model microbial growth. Studying the competition model may help us to understand the nature of QFA experiments and, if some assumptions do not hold, it could be used as a first step in developing a more accurate model. Furthermore, collectively fitting the competition model involves a the large number of parameters and data points and will require many simulations to be run. This necessitates the use of an approximate model for computational feasibility. This is especially true if the model is to be used in the analysis of high-throughput data. It is hoped that even an approximate model will be able to measure more reliable growth parameters and better estimate fitness. This will increase the power to infer genetic interaction and drug response which could lead to further discoveries. (For an example of a sucessful QFA study and follow up using the logistic model see Addinall

et al. (2011) and Holstein *et al.* (2014)).

2 METHODS

CANS

To analyse QFA data using the competition model I developed the Python package CANS which can be used for model composition, model simulation, parameter inference, and visualisation of results. CANS accepts cell density timecourses for any size rectangular array. CANS can produce SBML models to document results of parameter inference or for independent validation. It is relatively simple to create and simulate new models involving reactions between species within cultures or between neighbouring cultures, and to fit these provided an initial guess. The CANS package is available at https://github.com/lwlss/CANS.

2.2 Solving and fitting

Solving

CANS numerically solves models using one of two methods. The first is slower and uses SciPy's integrate.odeint to solve models written in Python at user supplied timepoints. I vectorised code using NumPy to optimise solving of the competition model by this method. For solving a plate of 384 cultures with cell density observations at 10 unevenly spaced time points, I found that using the Python bindings for libRoadRunner was about 10 times faster. libRoadRunner requires models to be written in SBML so I wrote code using the libSBML Python API to automatically generate SBML versions of the competition model for any size plate. Unlike SciPy's odeint, libRoadRunner only simulates at uniformly spaced timepoints. To fit QFA cell observations, which are not made at fixed time intervals, requires simulated cell amounts at the observed timepoints. For the analysis in (P15 section), where each timecourse has only 10 timepoints, I simulated sequentially between (pairs of) timepoints. This method was slower for the analysis in (Stripes section) where each timecourse had around 50 timepoints. To increase speed I used SciPy's interpolate.splrep to make a 5th order B-spline of cell density timecourses with smoothing condition $s = 1.0$. I evaluated the spline for cell density using SciPy's interpolate.splev at 15 evenly spaced intervals from time zero to the time of the last QFA observation. I then solved these timecourses with one call to RoadRunner.simulate.

2.2.2 Fitting the competition model

I use QFA data after processing with Colonyzer (Lawless *et al.*, 2010). Colonyzer uses integrated optical density measurements in whole plate images as a proxy for cell density. I used [timecourses of] these cell density estimates, which have arbitrary units, throughout my analysis. I fit the competition model using a gradient method and made maximum likelihood estimates of parameters using a normal model of measurement error. For constrained minimisation I used the L-BFGS-B algorithm from SciPy's integrate package.

I determined stopping criteria so that parameters of full-plate simulated data sets, with a small amount of simulated noise, were recovered with high precision. To help the minimizer, I scaled C_{t_0} values by a factor of 10^5 to make parameter values closer in order of magnitude. I ran repeated fits using different parameter guesses for each plate (see Section (P15 and Stripes details)). I set bounds according to Table 1 and checked that best fits had no parameters at a boundary.

Table 1: Parameter bounds. *Used for fitting the competition model to P15 and the Stripes and Filled plates. Bounds on N_{t_0} were applied to both $N_{t_0}^I$ and $N_{t_0}^E$ for internal and edge cultures. "guess" refers to the initial guess (see Section 2.4).*

Parameter	Lower Bound	Upper Bound
C_{t_0}	guess x 10^{-3}	guess x 10^3
N_{t_0}	guess / 2	guess x 2
k_n	0.0	10.0
b	0.0	None

Cultures at the edge of a plate have an advantage because they have access to a greater area of nutrients. I corrected for this using a separate parameter $N_{t_0}^E$ representing a higher initial amount of nutrients in edge cultures. In rate equations involving edge cultures, I scaled edge culture nutrient amount N_i by the ratio $N_{t_0}^I/N_{t_0}^E$, where $N_{t_0}^I$ is the amount of nutrients in internal cultures. The physical interpretation of this correction is that edge cultures have an extra supply of nutrients that can diffuse instantly into the reaction volume. This treatment reduced the error in cell density estimates for cultures one row or column inside the edge and resulted in better fits to internal cultures overall (see Table 2 or Section). Cell density measurements from edge cultures contain more noise due to reflections from plate walls (Lawless *et al.*, 2010). I collectively fit to all cultures and selected best fits based on only the fit to internal cultures.

(Can go to results section or Stripes method section:) QFA data for the Stripes plate contained observations for cultures that were known to be empty. When fitting the competition model, I set growth constant b to zero for these cultures and removed them from fitting.

2.2.3 Fitting the logistic model

Fitting the mass action logistic model requires using culture level N_{t_0} and creating 383 extra parameters. The QFA R package (Lawless *et al.*, 2016) can fit the standard logistic model and has heuristic checks to correct a confounding of parameters that occurs when slow-growing cultures are dominated by noise. I did not have time to implement these checks for the mass action logistic model, so I instead fit the standard logistic model using the QFA R package. This is not equivalent because QFA R does not fit data collectively and instead uses a culture level C_0. However, this is a useful comparison with a method of analysis currently used in QFA (see e.g. Addinall *et al.* (2011)). I do not expect much disagreement of fitness estimates with the mass action logistic model once heuristic checks are implemented. In contrast to the competition model, noisy data from edge cultures was discarded before fitting. I conduct model comparison between the competition and logistic models in sections (Results sections).

2.2.4 Data visualisation

(Do I really need this?) I created plotting functions to visualise fits and simulations of QFA timecourses and to compare the ranking of fitness estimates using the Python package matplotlib.

2.3 Parameter conversion

(Will move to the discussion: The identity of the nutrient molecule is unknown and it is not clear whether metabolism of the nutrient molecule will have a significant effect. If necessary a metabolism reaction could also be modelled.)

When k_n is set to zero, the competition model (4) reduces to the mass action logistic model which has the same sigmoidal solution as the standard logistic model. In this limit, it is possible to equate cells of both models and convert parameters using (6) (see Conor's blog for a derivation).

$$r_i = b_i(C_{t_0} + N_{t_0}) \tag{6a}$$
$$K = (C_{t_0} + N_{t_0}) \tag{6b}$$

The reaction equation of the competition model (3) assumes that all nutrients are converted to cells. This implies that all cultures starting with the same amount of nutrients reach the same final amount of cells. Therefore, to fit the mass action logistic model to QFA data, it is necessary to allow N_{t_0} to vary for each culture which is not physical and, in which case, the mass action logistic model has the same number of parameters (769) as the standard logistic model. (Probably repetition: When I fit the competition model I collectively fit the timecourses of all cultures on a plate using a plate level N_{t_0} and 387 parameters.) Figure 4 shows fits of a single culture on a larger 16x24 format plate using both models. This culture grew faster than its neighbours (not shown) and, according to the competition model, competed for more nutrients. Figure 4a shows the mass-action logistic model fit where N_{t_0} is estimated as being approximately equal to the final cell amount, or carrying capacity K. Figure 4b shows the competition model fit with a plate level N_{t_0} and $kn > 0$. Resimulating with k_n set to zero gives the dashed mass action logistic model curves which are corrected for competition. We can therefore obtain the corrected logistic model r_i and K_i of these curves by converting from competition model estimates of b_i, C_{t_0}, and N_{t_0}. N.B. b is the same for both the solid and dashed curves in Figure 4b.

Competition model C_{t_0} and N_{t_0} are the same for all cultures on a plate. Therefore, by the conversion equations (6), all cultures on a plate have the same carrying capacity K and all $b_i \propto r_i$ by the same factor. Similarly, MDP is the same for all cultures and all $b_i \propto MDR_i$ by the same factor (see Equation 2). Therefore, b is equivalent to all common QFA fitness measures, r, MDR, and $MDR * MDP$ (see e.g. Addinall *et al.* (2011) and Lawless *et al.* (2016)). This makes b a very convenient fitness measure for the competition model; we need not convert to logistic model parameters to compare the fitness rankings of cultures on the same plate. To compare competition model fitness rankings between different plates we can of course use b. However, this is not equivalent to comparing r or MDR as different plates may have different C_{t_0} and N_{t_0}.

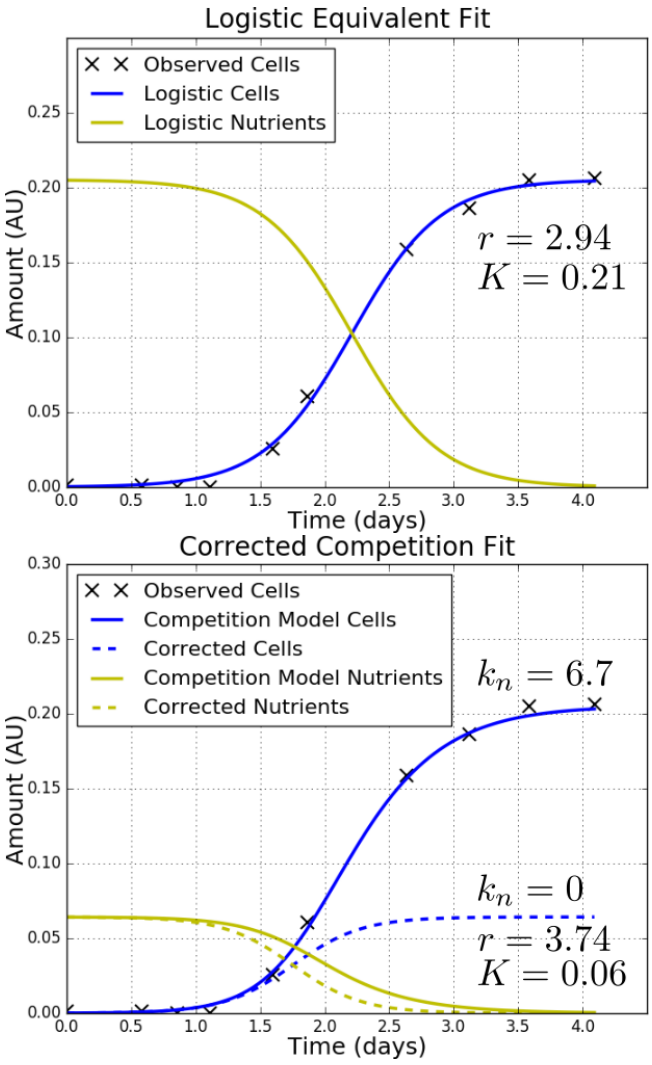

Figure 4: Using the competition model to correct for competition. *Fits are to culture (R10, C3) of P15 which grew faster and reached a higher final cell density than its neighbours (not shown). According to the competition model, this is because this culture competed for more nutrients. To reach the same final cell density, the logistic equivalent model requires a higher amount of starting nutrients for this culture and a different amount for each neighbour. The correction to the competition model simulates how growth would have appeared without competition and allows us to return parameters r and K of the logistic model.*

2.4 Making an initial guess

Achieving good fits of the competition model requires making a good initial guess. To fit the competition model to small simulated zones I could simply use many random parameter guesses. However, for fitting a full plate with 387 parameters the chance of any random guess being close to the "true" values is much smaller and more sophisticated guessing methods are required. I did not understand the disagreement between mass action logistic and competition model estimates of b which is only reduced when parameters are converted to logistic model r and K (see Section **??**). Without this conversion fitness rankings, using b, are inverted between the two models. I instead assumed that there was a more fundamental disagreement between models and developed the "Imaginary Neighbour Model" for guessing competition model b. This allowed good fits to be made. I did not have time to compare imaginary neighbour guessing with logistic model guessing so it is unclear which method is better.

2.4.1 Guessing initial amounts

Recall from the competition model reaction equations (3 and 5) that nutrients can only diffuse or be converted to cells. Thus, assuming that reactions are nearly complete at the end of cell observations and that $C_{t_0} \ll N_{t_0}$, the total initial amount of nutrients, N_{Tot}, can be estimated using,

$$N_{Tot} = n_I N^I_{t_0} + n_E N^E_{t_0} \approx C_F, \tag{7}$$

where C_F is the total of final cell measurements, n_I and n_E are the numbers of internal and edge cultures, and $N^I_{t_0}$ and $N^E_{t_0}$ are initial nutrient amounts for internal and edge cultures (see Section 2.2.2). Using (7) and an estimate for the ratio of area associated with edge cultures to area associated with internal cultures, $A_r = A^E/A^I = N^E_{t_0}/N^I_{t_0}$, I made guesses of $N^I_{t_0}$ and $N^E_{t_0}$ using,

$$\begin{aligned} N^I_{t_0} &= N_{Tot}/(n_I + n_E A_r) \\ N^E_{t_0} &= N_{Tot}/(n_I/A_r + n_E). \end{aligned} \tag{8}$$

When $A_r = 1$, (8) reduces to the initial nutrient guess for the one initial nutrient parameter model. I used $A_r = 1.4$.
In QFA using dilute cultures, C_{t_0} falls below the level of detection. I did not estimate initial guesses of C_{t_0} and instead ran multiple fits over a range of C_{t_0} values in logspace. What was the range?

2.4.2 Guessing b

To guess competition model b_i I used the imaginary neighbour model to quickly fit individual cultures. The model is based on the reaction and rate equations of the competition model (3–5) but tries to replicate the diffusion of nutrients into and out of a culture using imaginary fast and slow growing neighbours with different nutrient diffusion constants $k_{n,f}$ and $k_{n,s}$. The growth constants of the fast and slow growing cultures are b_f and b_s. A schematic of the model is drawn in Figure (ref). To fit the model to QFA data, I fixed C_{t_0} and $N^I_{t_0}$ for all cultures by the initial guesses (see Section 2.4.1); I fixed b_f at a range of different guesses, and fixed $b_s = 0$; I allowed b, $k_{n,f}$, and $k_{n,s}$ to vary. I determined the number, n, of each neighbour from the guess of $N^I_{t_0}$ and the range of final cell amounts, such that the culture with the highest observed final cell density had enough slow growing neighbours to provide all of the nutrients necessary to reach this final cell density. I solved the imaginary neighbour model using SciPy's odeint. I fit using a gradient method as in Section 2.2.2.

2.4.3 Guessing k_n

Simulations of the competition model using sets of b parameters drawn from different normal distributions have linear relationships between variance in final cell amount and nutrient diffusion constant k_n. I simulated guessed parameters C_0, N_0, and b_i with a range of different k_n values and used linear regression to parameterise the straight line. I then took the variance in final cell amount for real data and guessed k_n from the straight line.

2.5 Development of a genetic algorithm

2.6 Model comparison using a single QFA plate

2.7 Cross-plate calibration and validation

3 RESULTS

3.1 Guessing

N_0 estimated from average final cell amounts. See formula in code for two N_0 estimation.

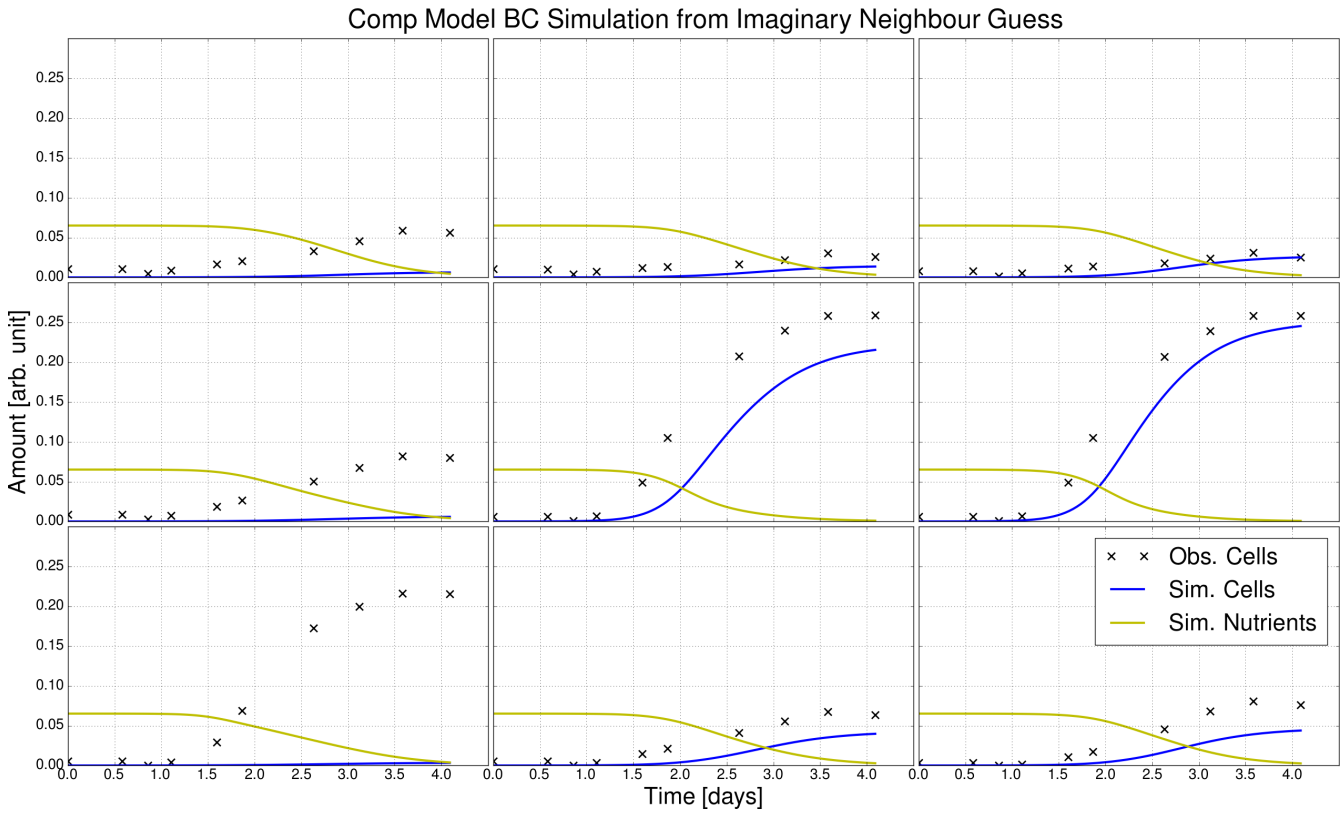

Figure 5: Competition model simulation using parameters from imaginary neighbour guessing. *Shows a 3x3 zone with top-left coordinate (5, 18) from P15 with background cdc13-1 at 27°C.*

3.2 Competition Model Fitting to P15

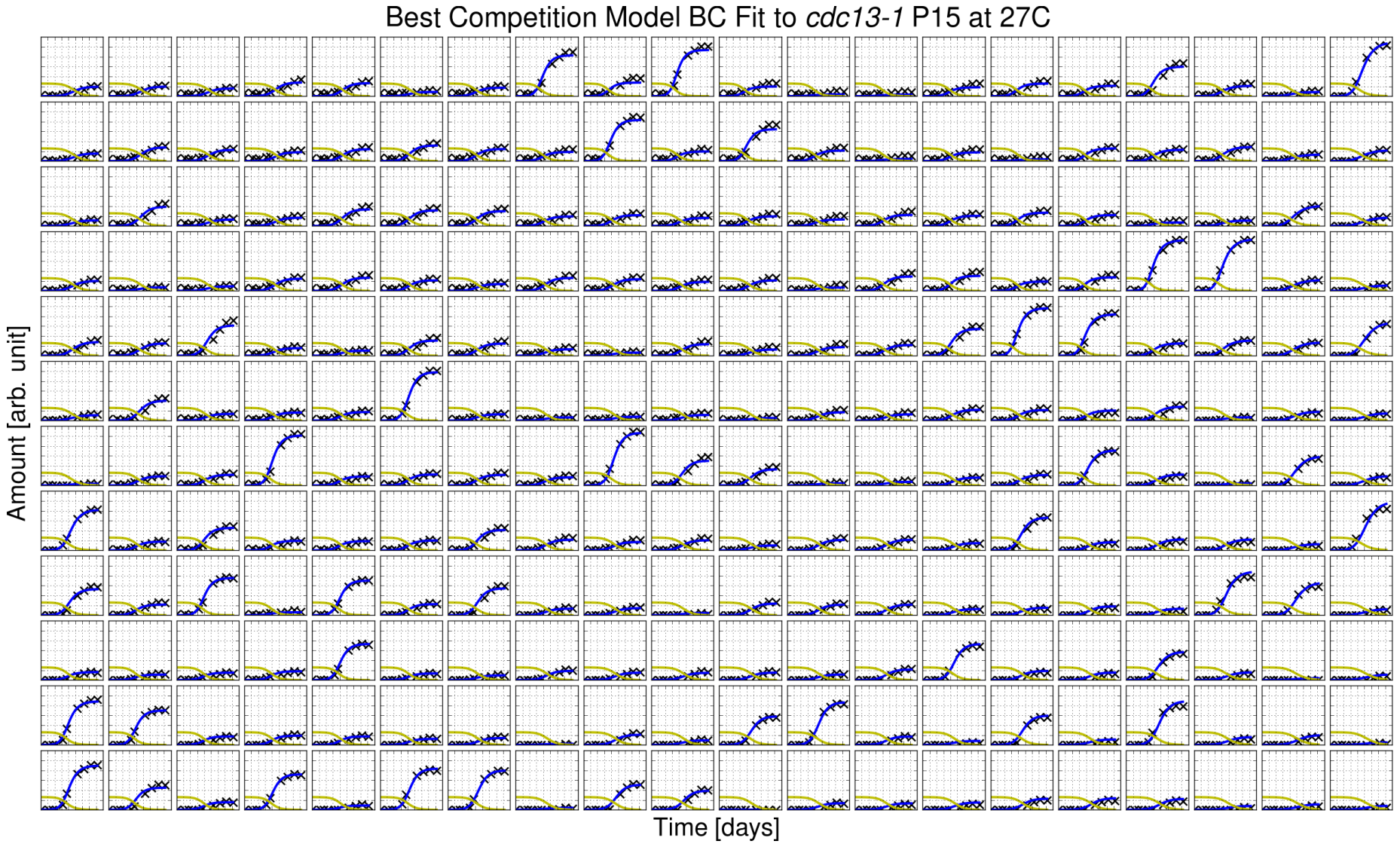

Figure 6: Fit of the competition model to a QFA plate. *Data is for a 16x24 format plate (P15) with a background mutation cdc13-1 incubated at 27°C. The plate contains 6 repeats of 50 genetic strains randomly arranged across the internal cultures. Repeats of a single strain are used for all edge cultures (removed in the plot). Model output for state variable, cell population size (blue curve), is fit to observed data (black crosses). Model predictions for unobserved variable (nutrient amount) are also plotted (yellow).*

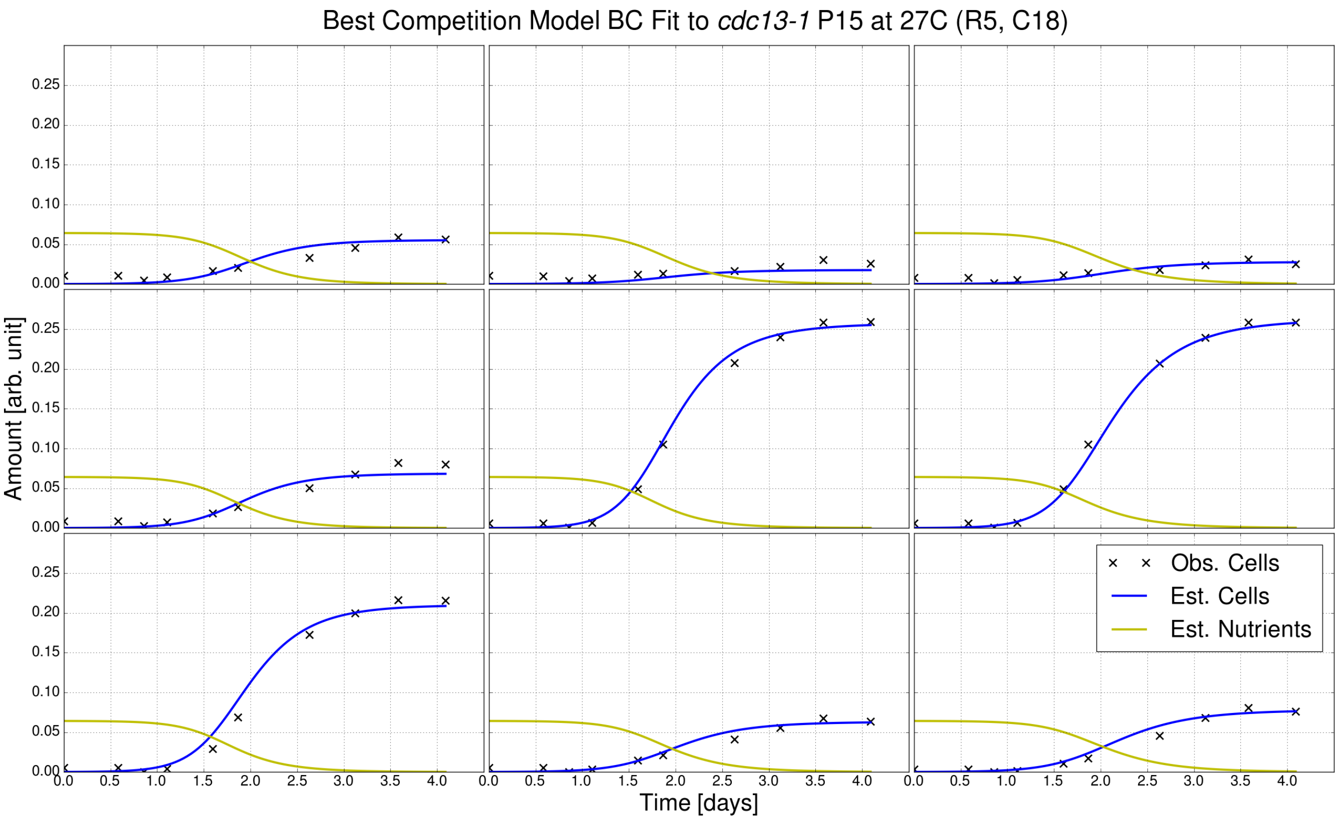

Figure 7: A 3x3 zone from Figure 6 with top-left coordinate (5, 18).

Some text Some text.

3.3 Evaluating the treatment of boundaries

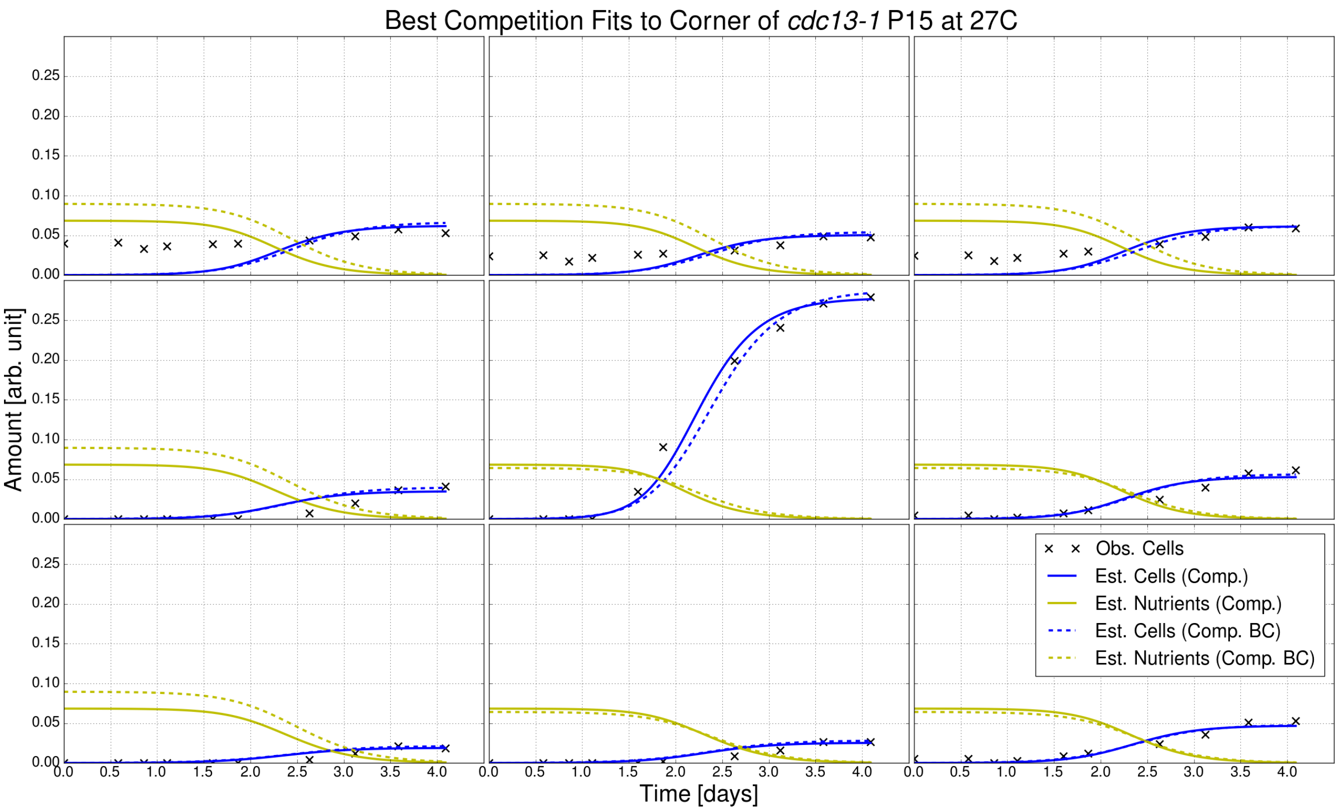

Figure 8: Treatment of boundary conditions in fits of the competition model. *The top left corner of a 16x24 QFA plate fitted with two versions of the competition model, the first has a single initial nutrient amount for all cultures, the second has a separate initial nutrient amount for edge cultures.*

Table 2: Average error in objective function for one a two N_0 parameter competition models. *Values are for the same fits as in Figure 8 and have been scaled by 10^4. Averages are for cultures belonging to the areas indicated by the column "Cultures". "Next to edge" refers to cultures one in from the edge. "Internal" refers to all cultures but the edge.*

Cultures	One N_0	Two N_0
Edge	35.9	36.5
Next to edge	9.54	7.98
Internal	6.67	6.30
All	12.4	12.2

Table 2

Some text Some text

Some text Some text.

3.4 Agreement of b rankings

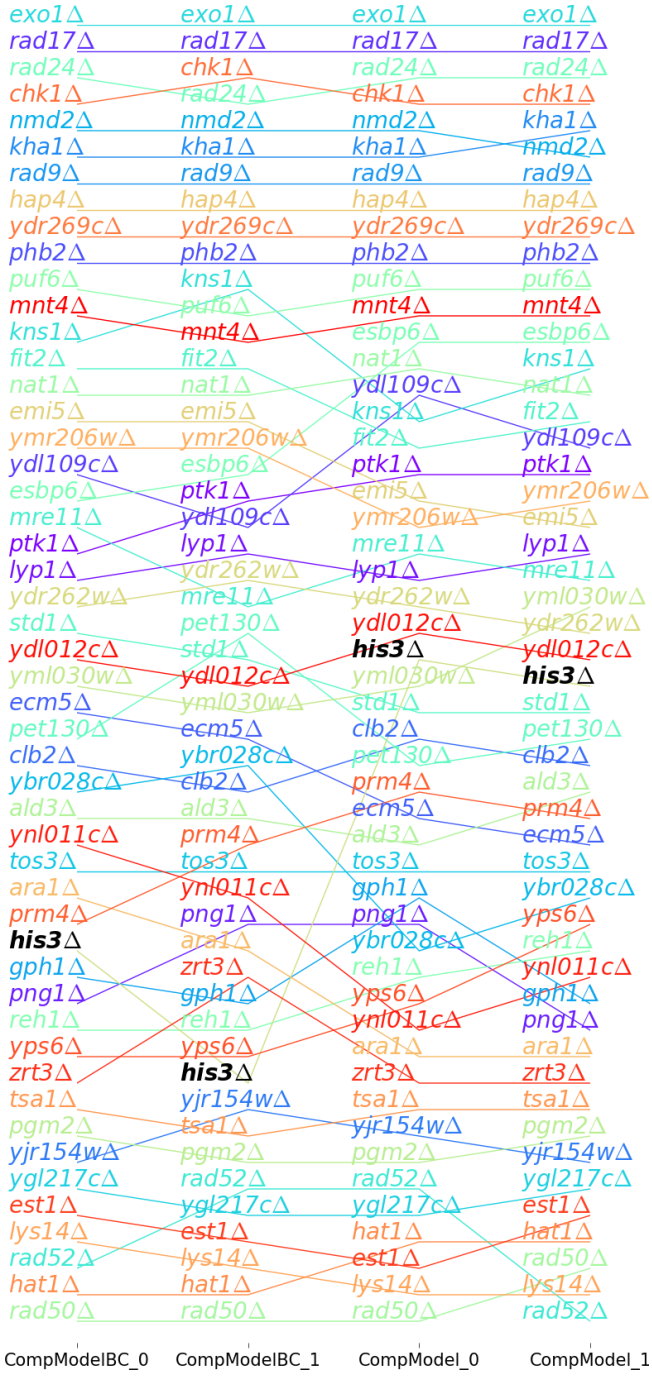

Figure 9: Comparison of b ranking for the best five competition model fits to P15. *Ranking is calculated from the mean b estimate from the six repeats or each strain.*

3.5 Comparison of fitness ranking

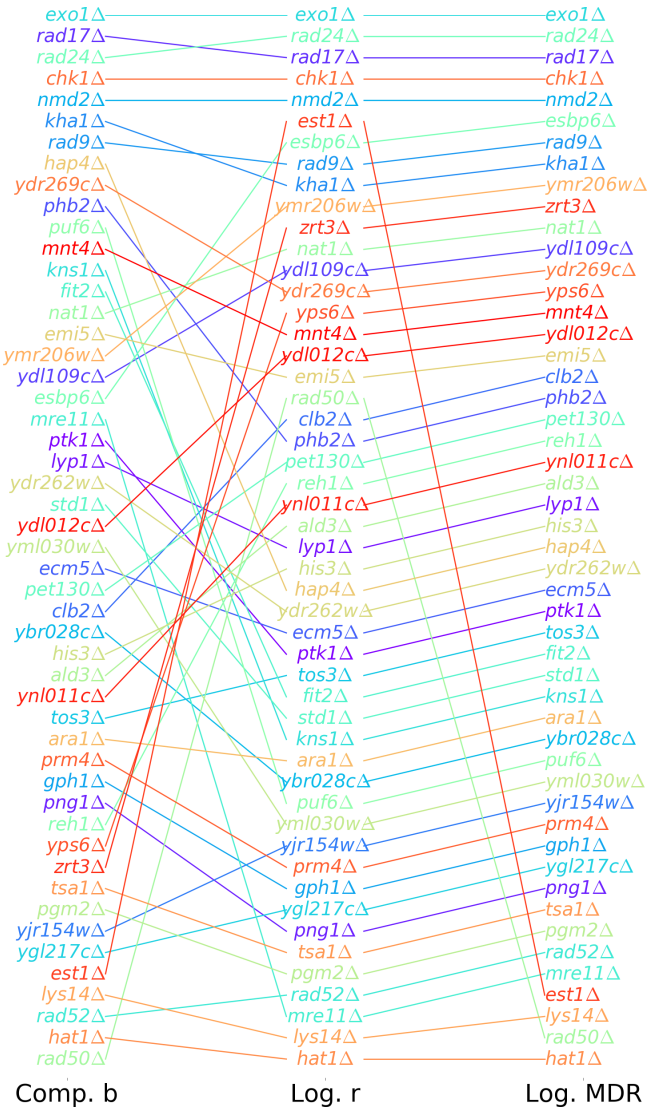

Figure 10: Comparison of r ranking for fits of the competition and logistic model to P15. *Competition model r was converted from b, N_0, and C_0 from the best competition model estimate. Logistic r was taken from fits using the QFA R package which makes heuristic checks for slow growing cultures.*

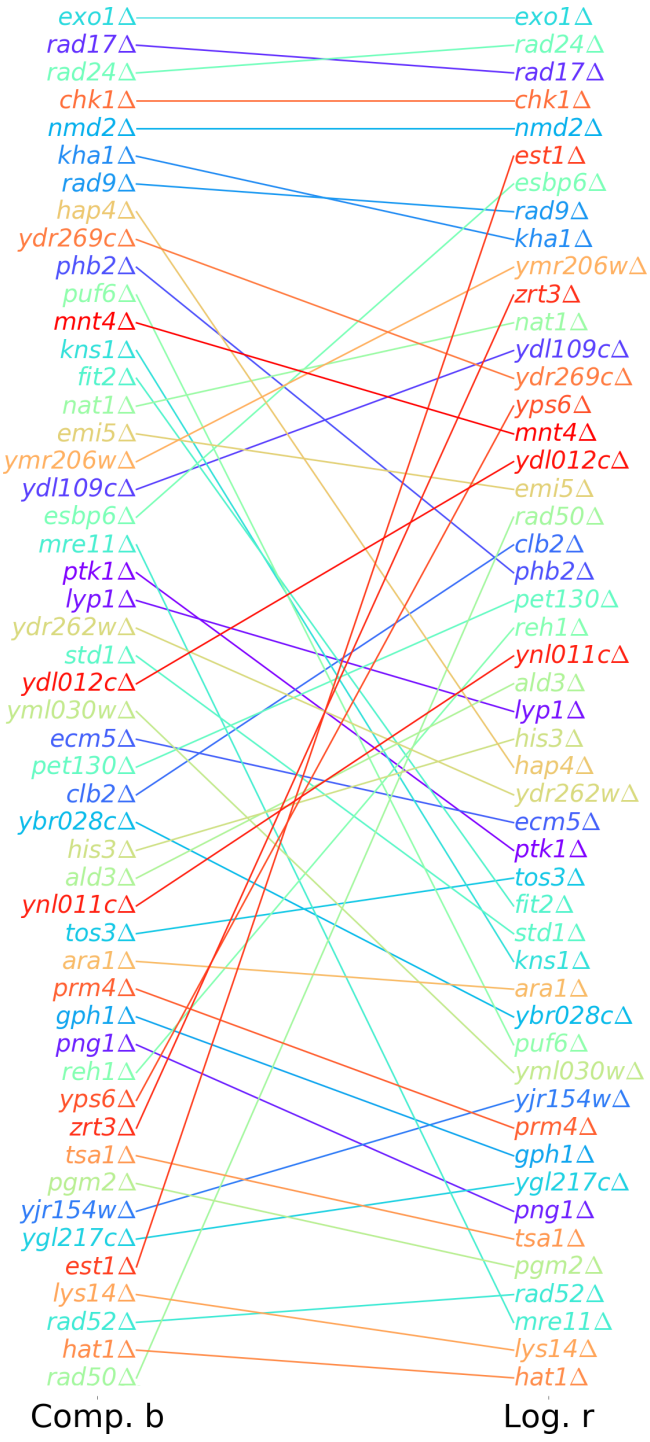

Figure 11: Comparison of r ranking for fits of the competition and logistic model to P15. *Fitnesses of genetic strains are ranked most to least fit from top to bottom. Competition model r was converted from b, N_0, and C_0 from the best competition model estimate. Logistic r and MDR were taken from logistic model fits using the QFA R package which makes heuristic checks for slow growing cultures.*

3.6 Comparison of Variation in Fitness Estimates

Use repeats on plate 15 (6 per deletion) to calculate coefficient of variation (COV) of estimated r or MDR.

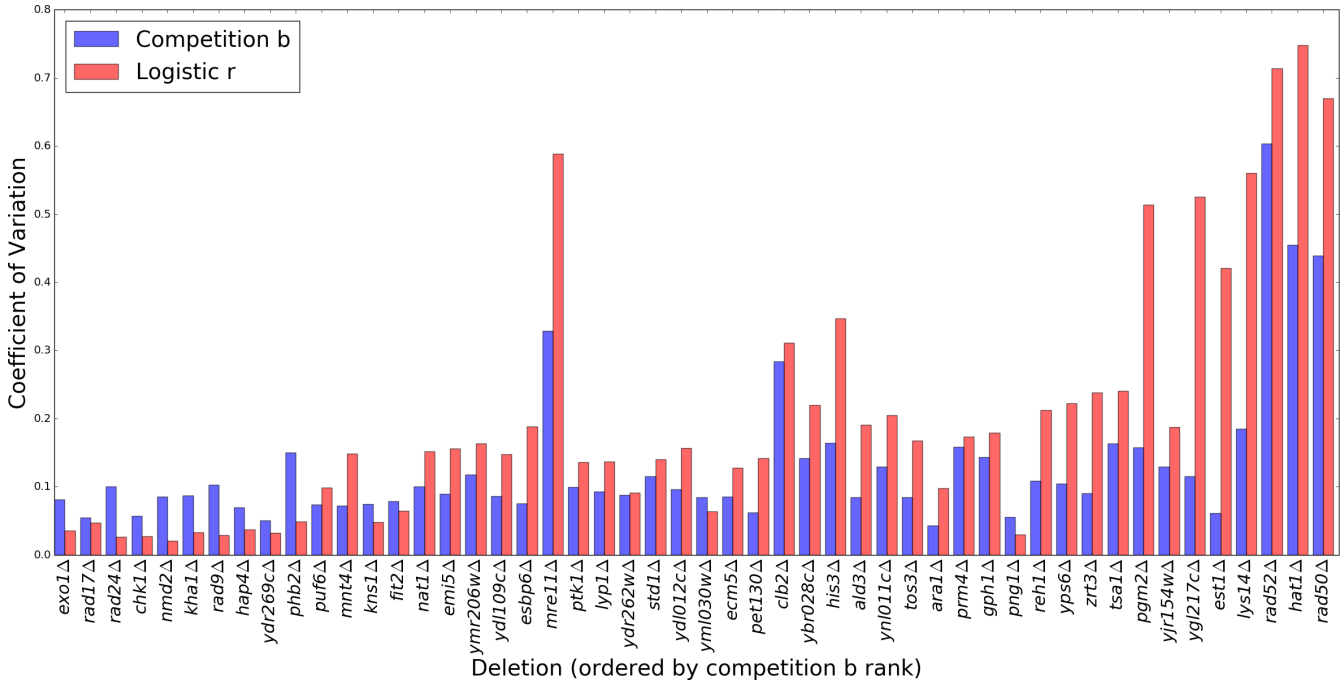

Figure 12: Coefficient of variation of r estimates. *Strains are ordered left to right along the horizontal axis by highest to lowest competition model r ranking. Fits are for the competition model, the QFA R logistic model, and the logistic equivalent model.*

3.7 Cross-plate validation

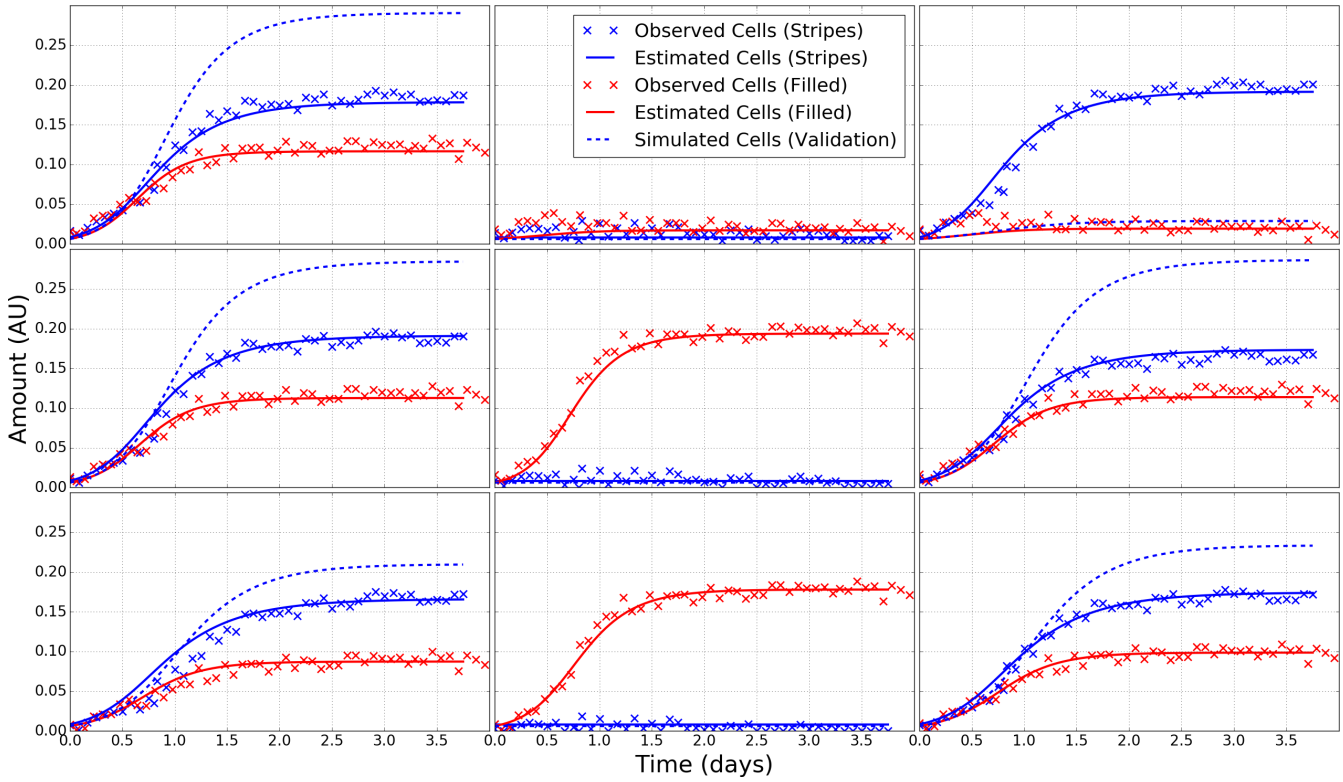

Figure 13: Calibration and validation of the competition model. *I fit the competition model to the 16x24 format "Stripes" and "Filled" plates in Figure 2. The plot shows cell measurements and estimates for both plates for a 3x3 section with top left coordinates (R9, C10). I took the parameters estimates for the "Filled" plate (calibration) and set growth constant, b, to zero for cultures in the empty columns of the "Stripes" plate. I then simulated using these parameters to produce the dashed blue curve (validation). If the model is working correctly, the dashed blue curve should resemble the "Stripes" data (blue crosses).*

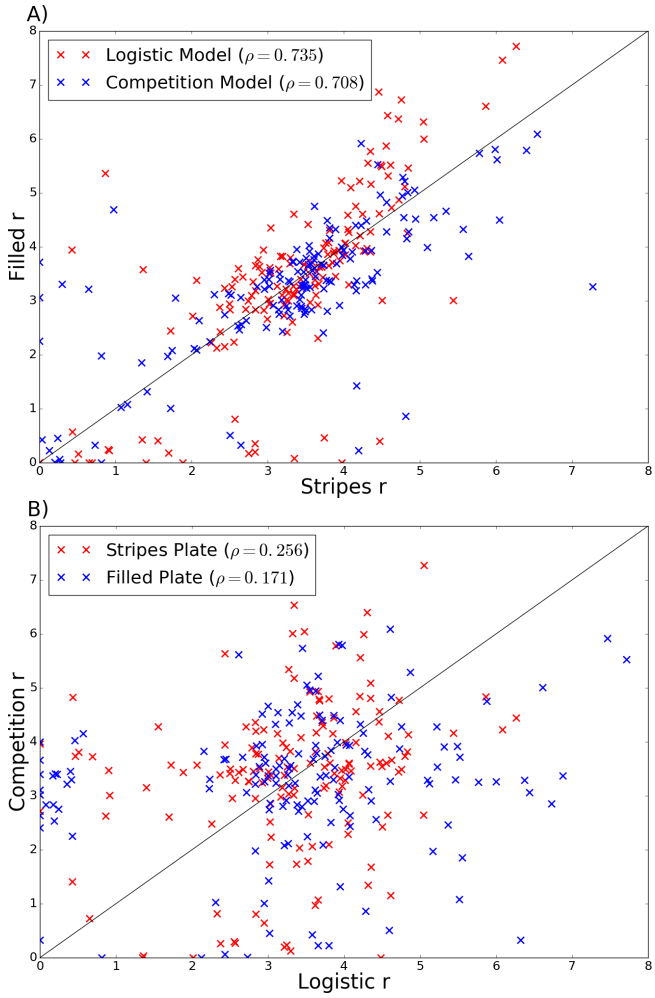

Figure 14: Correlation of r estimates for "Stripes" and "Filled" plates. *A) Correlation of r estimates between plates for logistic and competition models. B) Correlation of r estimates between logistic and competition models for both plates. I fit the competition model and independent model to the "Stripes" and "Filled" plates in Figure 2. I converted competition model b to logistic model r. I only used data for cultures that were common between the two plates common and removed edge cultures. The Pearson correlation coefficient, ρ, is shown in the legends. The line $y = x$ is also plotted.*

3.8 Towards a genetic algorithm

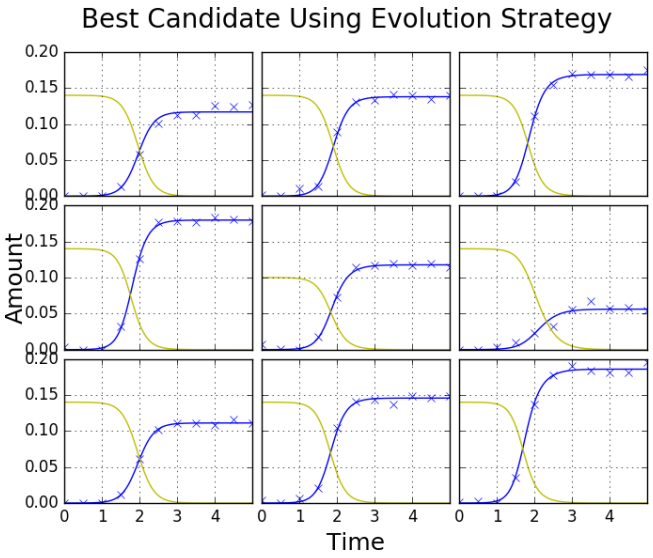

Figure 15: *Genetic algorithm fit to a 3x3 simulation. MIGHT TAKE A LITTLE BIT OF WORK TO REPRODUCE AND COULD USE PARAMETERS FROM THE BEST P15 FIT RATHER THAN JUST PICKING/RANDOMIZING. NEED TO CHECK THAT PLATE LEVEL PARAMETERS WERE ALSO EVOLVED.*

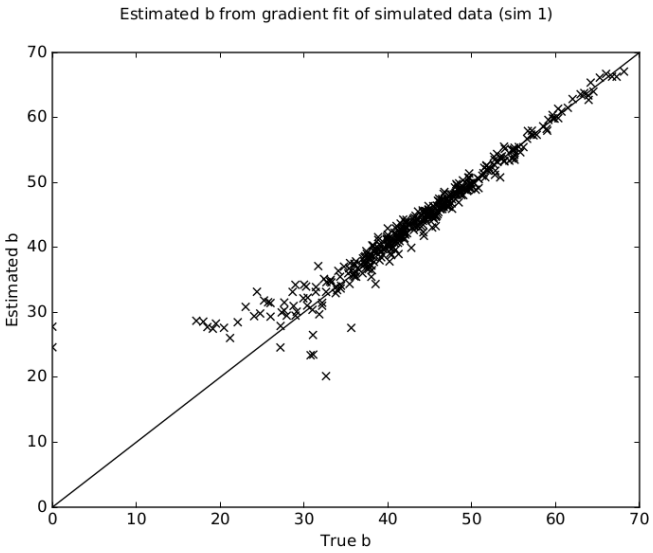

Figure 16: *Recovery of true b values from a gradient method with fixed plate level parameters. I simulated timecourses from the best five (which model? all BC?) fits to p15, fixed the true plate level parameters, and used a gradient method to recover b. This plot shows the worst case from the five sets of values.*

4 DISCUSSION

Fits of the competition model (see Figures 6 and 7) use less parameters and are qualitatively better than fits of either the logistic or generalised logistic model from the QFA R package (Addinall *et al.*, 2011; Lawless *et al.*, 2016). Competition model ranking of growth estimates for repeats on P15 (see Figure 10) agree with the logistic model rankings from Addinall *et al.* (2011) for the fastest and slowest growers (and with rankings from independent spot tests (refs) CHECK THIS).

However, there is much disagreement in the rank of other strains. (Could also do with the correlation plot for P15). The reliability of growth estimates was not improved using the competition model for the fastest growing strains on P15 (see Figure 12). This may be due to the effect of noise dominated cell observations from slow growing cultures on collectively fit parameters. This did not affect the the logistic model which fit to cultures individual. The greater reliability of estimates for slow growing cultures could be entirely due to collective fitting rather than to correcting for competition. (Could do with p-values on figure; also bold HIS3 everywhere) Unfortunately, the change in order for middle rankings is unlikely to identify new genetic interactions because significance is determined by comparison to a neutral deletion also in this range. Although improved, uncertainty in estimates for the slowest growers is still much higher than for the fastest meaning that the power to infer genetic reactions is not dramatically improved. (HIS3 has about half the variance for the competition model and this could be significant.)

Fitting the logistic model to slower growing cultures requires heuristic checks to correct for confounding between r and K. The QFA R implementation appears to have some issues. The strains *est1*Δ and *rad50*Δ have dramatic changes in ranking between logistic model r and MDR (see Figure 10). In independent spot tests (ref) these are very sick strains and I confirmed this in the raw QFA images by visual inspection. High r and low K have been erroneously fit to both strains. This is corrected for when converting to MDR which agrees with the competition model ranking and independent validation and is more similar to the fitness measure ($MDR \times MDP$) used in the original analysis by Addinall *et al.* (2011). For other cultures it appears that encroachment of fast growing cultures into neighbours is affecting cell density estimates made by Colonyzer (Lawless *et al.*, 2010). In logistic fits some growth curves are still in the exponential phase at the end of observations and this may be another fitting issue. If repeated, the plate from Addinall *et al.* (2011) should be run with a lower concentration of nutrients in the agar so that the stationary phase can be reached before cultures start to merge.

I looked at plate images from QFA and Colonyzer to investigate other discrepancies. *mre11*Δ is a weak growing strain (ref validation) which was misclassified as healthy by the competition model but not the logistic model. One repeat contained unusual heterogeneity, which may be natural or the results of contamination, and may explain the discrepancy. (Should take median value next time). *hap4*Δ appears to be much healthier than *zrt3*Δ which agrees with the competition model but not the logistic model. Although the precision of estimates is similar for both models, the competition model appears to be more accurate. Unfortunately, I lack independent data for validation of the middle strains.

Recent work Herrmann and Lawless suggests that direct measures of C_{t_0} may not be reliable due to heterogeneity between cells in the same inoculum; many cells do not grow and only the fastest growing cells contribute significantly to the final population. A plate level C_{t_0} also seems inappropriate but having extra parameters for the staring cell density of each culture is undesirable. Only a small amount of nutrients is used when cultures are small. Therefore, cultures could be grown for a short time before making direct cell density measurements that may be more accurate. QFA inocula use cells taken from the stationary phase where there might be more heterogeneity (ref). It may be possible to increase the reliability of fitness estimates by taking inocula from the exponential growth phase or using a higher starting density to average out effects.

The Stripes and Filled plates used a higher inoculum density and had very few noise-dominated cultures. Compared to P15, this would have reduced noise in collectively fit competition model estimates and would not have required heuristic checks to be employed for the logistic model. This may therefore be a fairer comparison than P15. Correlation of fitness estimates between plates in Figure **??**a was similar for both models. This is despite not finding global minima with the competition model. However, correlation between models for the same plate in Figure 14b is poor. (Could definitely do with P15 correlations to compare). There are issues with validation for both models (see Figure 13); the logistic model does not account for differences between plates at all and the competition model overcorrects. As I lack independent data for validation it is difficult to decide which to believe. (Unfortunately, these plates lack repeats so I could not study the reliability of estimates on the same plate. (I believe that we have more issues with accuracy than precision anyway). It would be informative to repeat P15 with C_{t_0} at a measurable level. In any case, it is clear that the competition model could be improved.

4.1 Future work

I was unable to find global minima using a gradient method (see Section 2.2.2) to fit the competition model. I began work on a genetic algorithm method of solving but lacked time to complete this. I did however find that, with fixed plate level parameters, it is possible to reliably return b_i with a gradient method (see Figure 16. This offers the potential to use a hierarchical genetic algorithm where candidate plate level parameters are fixed in gradient fits of culture level parameters. Alternatively, a pure hierarchical genetic algorithm may work (i.e. where b_i are also evolved). A hierarchical Bayesian approach to fitting the competition model, similar to that of Heydari *et al.* (2016) for the logistic model, could also return global minima but might be slow. Current best fits, which are different local minima, have well correlated fitness rankings and make similar overcorrections for competition. This suggests that there is a more fundamental issue with the model.

It would be informative to validate the independent limit of the competition model to determine whether a mass action approximation is valid and whether it is correct to ignore the effect of metabolism on nutrient and final cell density. I suggest to validate first in liquid cultures, where the assumption of a well stirred mixture is more valid, and then attempt to validate for single cultures grown on agar, which more closely resembles QFA. Growth on a surface has a lower dimensionality and may be diffusion limited so a fractal kinetics model

may be required (Kopelman, 1988; Savageau, 1995). Nutrients (sugars, nitrogen, etc.) in QFA agars are of a standard composition, designed to reduce the excess of any single nutrient (check QFA paper and cite). It would be helpful to know and control the identity of the limiting nutrient using a different formula of agar. With nitrogen, rather than sugar, as the limiting nutrient, we are less likely to have to model metabolism.

Estimates of the nutrient diffusion constant k_n were fairly high such that nutrients diffused readily between neighbours and were nearly depleted when growth stopped. It may be that growth becomes limited by the diffusion of nutrients through agar before all nutrients are depleted and that nutrients are not well approximated as being evenly distributed withing the spatial scales that we model. Using a finer grid could reduce the overcorrection seen in Figure 13. Reo and Korolev (2014) use the diffusion equation (with Neumann and Dirichlet boundary conditions) to simulate nutrient dependent growth of a single bacterial culture on a pertri dish in two-dimensions. They create a sink for nutrients from culture growth and equate the flux of nutrients through culture area with the rate of increase in culture size. They model culture area as varying and keep culture density constant. This model could be adapted for QFA by keeping culture area constant and allowing culture density to vary. A mass action kinetic model of reaction (3) could be used for culture growth and the nutrient sink. It is computationally unfeasible to use such a detailed model to fit a whole plate but simulation could be very informative.

If we find that competition for nutrients is not responsible for the interaction between neighbours, for instance if growth becomes limited by diffusion of nutrients in the agar before nutrients from neighbours can be accessed, then we could instead model signalling by ethanol poisoning. This may be modelled similarly to how the competition model models nutrient diffusion and much of the code could be reused. (Quorum sensing via the molecule amonia could also be having an effect (ref)). If there is any combination of competition, metabolism, signalling, or arrest contributing significantly to differences in the growth of cultures and the interaction between neighbours then it will be difficult to separate them when fitting a model to data. We may have to develop ways to calibrate effects in isolation and use this information when fitting to high-throughput data. We only have observations for cells.

4.2 Improvements and other recommendations

It is quicker to fit to small zones of a plate but, as these have a larger proportion of edge cultures, boundary conditions become important. Growing smaller arrays in isolation would help to speed up the development process.

Each culture is surrounded by a different group of neighbours. The imaginary neighbour guess could be improved by using a range of b_f values to fit each culture and selecting b from the best fit. It would also be good to compare with guesses from the logistic model. I suspect that a gradient method will still fail to find a global minimum.

Edge cultures must be included when fitting the competition model and this may have contributed significantly to error. When fitting the competition model noise might be better dealt with by leaving edge cultures empty. A different treatment of boundaries could also be used by modelling empty cultures outside edges rather than the approach in Section (2.2.2).

In order to make sure that competition effects were present in data, we made a dramatic change between the stripes and filled plates. We could have first validated the model against a smaller change, by varying between slower and faster growing cultures rather that none and very strong growing cultures. If the model works well between such plates, it may work well for the majority of QFA experiments which typically have smaller differences between cultures than the data we studied. If we did want to test the in an extreme case we could have inoculated fast growing cultures next certain strains and not others to try to induce a change in ranking for which the competition model might compensate better than the logistic model.

It is desirable to have a mechanistic model including nutrient dependent growth so that we can use a plate level N_{t_0} eliminate the dependence on heuristic checks.

REFERENCES

Addinall, S.G. *et al.* (2011) Quantitative fitness analysis shows that nmd proteins and many other protein complexes suppress or enhance distinct telomere cap defects. *PLoS Genet*, **7**, 4, 1–16.

Aldridge, B.B. *et al.* (2006) Physicochemical modelling of cell signalling pathways. *Nature cell biology*, **8**, 11, 1195–1203.

Andrew, E.J. *et al.* (2013) Pentose phosphate pathway function affects tolerance to the g-quadruplex binder tmpyp4. *PLoS ONE*, **8**, 6, 1–10.

Banks, A. *et al.* (2012) A quantitative fitness analysis workflow. `http://www.jove.com/video/4018/a-quantitative-fitness-analysis-workflow`.

Baryshnikova, A. *et al.* (2010a) Quantitative analysis of fitness and genetic interactions in yeast on a genome scale. *Nature Methods*, **7**, 12, 1017–24. Copyright - Copyright Nature Publishing Group Dec 2010; Last updated - 2014-09-19.

Baryshnikova, A. *et al.* (2010b) Synthetic genetic array (sga) analysis in saccharomyces cerevisiae and schizosaccharomyces pombe. *Methods in enzymology*, **470**, 145–179.

Berryman, A.A. (1992) The orgins and evolution of predator-prey theory. *Ecology*, **73**, 5, 1530–1535.

Chen, W.W. *et al.* (2010) Classic and contemporary approaches to modeling biochemical reactions. *Genes & development*, **24**, 17, 1861–1875.

Costanzo, M. *et al.* (2010) The genetic landscape of a cell. *science*, **327**, 5964, 425–431.

Heydari, J. *et al.* (2016) Bayesian hierarchical modelling for inferring genetic interactions in yeast. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **65**, 3, 367–393.

Holstein, E.M. *et al.* (2014) Interplay between nonsense-mediated mrna decay and {DNA} damage response pathways reveals that stn1 and ten1 are the key {CST} telomere-cap components. *Cell Reports*, **7**, 4, 1259 – 1269.

Kopelman, R. (1988) Fractal reaction kinetics. *Science*, **241**, 4873, 1620–1626.

Lawless, C. *et al.* (2010) Colonyzer: automated quantification of micro-organism growth characteristics on solid agar. *BMC Bioinformatics*, **11**, 1, 1–12.

Lawless, C. *et al.* (2016) *qfa: Tools for Quantitative Fitness Analysis (QFA) of Arrayed Microbial Cultures Growing on Solid Agar Surfaces*. R package version 0.0-42/r678.

Reo, Y.J. and Korolev, K. (2014) Modeling of Nutrient Diffusion and Growth Rate in Bacterial Colonies.

Savageau, M.A. (1995) Michaelis-menten mechanism reconsidered: implications of fractal kinetics. *Journal of theoretical biology*, **176**, 1, 115–124.

Verhulst, P. (1845) Recherches mathematiques sur la loi d'accroissement de la population. *Nouveaux memoires de l'Academie Royale des Sciences et Belles-Lettres de Bruxelles*, **18**, 14–54.