IDENTIFYING INTERACTION EFFECTS IN HIGH-THROUGHPUT STUDIES OF GROWTH

by

Kevin Ushey

B.Sc., The University of British Columbia, 2009

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in

THE FACULTY OF GRADUATE STUDIES

(Statistics)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

August 2011

Abstract

In genomics, a newly emerging way to learn about gene function is through growth curve experiments. In such experiments, different strains of yeast (*Saccharomyces cerevisiae*) – single mutants having one gene knocked out, double mutants having two knocked out – are grown in microtitre plates, with an automated system capturing the size of cell populations over time. These growth curves can provide information on the function(s) of the associated genes. Of particular interest are interaction effects, where the growth of a double mutant is surprising in light of the growth of normal yeast and its two corresponding single mutants.

There is currently a lack of consensus on the best way to analyze growth curve data. For a growth curve experiment, strain fitness must be defined in some way in order to separate and rank strains according to their ability to grow, and it is uncertain which possible definitions of strain fitness have better ability to identify real interaction effects than others. After defining strain fitness, this quantity must be estimated for each strain through either parametric or non-parametric model based approaches, and the approach used can also affect the ability to identify interaction effects. Furthermore, different problems related to the experimental protocol present themselves when attempting to model growth curves, and these need to be accounted for as well.

In this thesis, I will explore and compare some commonly used models and definitions of strain fitness when analyzing growth curves, and relate them concretely to the exponential and logistic models upon which they are built. I will compare and contrast multiple methods used when attempting to analyze growth curve experiments, and seek to propose Area Under the Curve as a definition of strain fitness which prompts a derived variables modeling strategy that performs well in ease of implementation, retains flexibility in assessment of a heterogeneous mix of sigmoidal and non-sigmoidal growth curves, and remains able to identify interaction effects.

Table of Contents

List of Tables

List of Figures

Acknowledgements

Nobody succeeds alone. Throughout my life I've had an immeasurable amount of help in both my personal and academic development, and the years of blood, sweat and tears have finally culminated in this thesis.

First and foremost I thank my supervisor, Dr. Bryan, for her active role in helping me prepare and complete this thesis, as well as her flexibility in allowing me to both pursue other projects and deal with 'real life' throughout my term under her tutelage.

I am indebted to the instructors and professors of the UBC Statistics department, many of whom have helped and given me opportunities even from my time as an undergraduate student within the UBC Department of Statistics.

I would also like to thank Jan and Jessica of the Hieter lab. Our collaborations throughout have helped me to become a more well-rounded statistician, and have taught me much more about how statistics in the 'real world' is done.

Finally, I thank my parents, who despite ill health throughout their lives, have worked slavishly to ensure I was raised happy and well. This thesis would not exist if not for the unending support they've given me throughout my life.

To my loving parents

Chapter 1: Introduction

The genome has now been sequenced for nearly two hundred organisms, and substantial effort is now being dedicated to identifying functional elements within an organism's sequence. The function of a gene, or more specifically, of the protein that it encodes, is of primary interest in genomics. We can gain insight into the function(s) of a certain gene by analyzing how functions in the cell are affected by the deletion of this gene. New technologies are helping to facilitate new experiments in which the effect of the gene deletion on different phenotypes of interest can be assessed. One primary phenotype of interest is the ability of a cell to divide and grow. If a set of genes is involved in a functional pathway that is necessary for normal cellular growth, then one might observe those cells with these genes knocked out might grow worse relative to wild type cells.

Although the genetic function of genes in human cells is of primary interest, because many of the genes involved in cellular maintenance and growth are shared between human and non-human cells, it is useful to study a simpler organism as a means of gaining information about genetic function for human cells. Yeast *(Saccharomyces cerevisiae)* is a particularly useful and well-behaved single celled organism that shares a large amount of genetic information with human cells, and is hence used as a model system for learning about human gene function. By analyzing the growth of *mutant strains* of yeast (yeast having one or more genes knocked out of its genome, relative to the wild type genome), we can learn about how these genes function in the yeast cell, and hence infer how these genes might function in a human cell.

Genetic function is often learned from *growth curve experiments*. In these experiments, different strains of yeast are grown over time, with some proximate measure of the count of cells for a given strain observed over time. The growth curves obtained are typically sigmoidal in shape: growth will be approximately exponential in the earlier stages, and after a period of time environmental constraints will restrict the rate of growth, with growth leveling off at an upper asymptote called the *carrying capacity*. However, different genetic mutations coupled with different properties of the experiment itself might cause some growth curves obtained to be non-sigmoidal in shape.

To analyze a set of growth curves, first a growth phenotype measurable from the growth curves must be chosen and defined. The primary means of learning about genetic function from these growth curve experiments, then, is through estimating and comparing these growth phenotypes over a set of mutant strains. This measured growth phenotype for a particular strain is typically given the name *strain fitness*, and different researchers will choose different growth phenotypes to represent strain fitness. When a parametric growth model is assumed, strain fitness is usually chosen to be a parameter, or function of the parameters, available in that model. However, when faced with a set of growth curves that do not seem to conform to any particular parametric growth model, strain fitness will have to be defined in a model-free way. In general, strain fitness is defined such that strains of yeast which grow faster and/or to higher carrying capacities are assigned higher fitness scores.

A newly emerging topic of interest in genomics is that of *interaction effects*. A gene-gene interaction, or genetic interaction, is said to occur when the strain fitness for a *double mutant* differs strongly from some notion of the expected fitness for that particular double mutant. For example, one might seek to identify genes that operate in compensatory pathways related to cell division – the singular deletions of genes A and B might only have a small effect on strain fitness; however, the deletion of genes A and B together in the organism might lead to heavily dampened growth and even lethality. This would be called synergistic interaction, as the observed strain fitness for this double mutant is much smaller than what one might expect from a combination of the singular deletion effects of genes A and B.

Suppose we are faced with a data from a growth curve experiment in which we have single and double mutants obtained through pairings of a set of query genes q (genes that have typically already been implicated with some function of interest, eg. cancerous tumor growth) with another set of genes g. We want to identify any real interaction effect associated with the different double mutants. We will let θ denote strain fitness, and τ be the additive effect of gene deletion on strain fitness. We can model strain fitness for the double mutant as

$$\theta_{q,g} = \theta_{WT} + \tau_q + \tau_g + \tau_{q,g}$$

Under this model, the effects of gene deletion on strain fitness are assumed to be additive, and the interaction effect between genes q and g is represented by the term $\tau_{q,g}$. Typically, the two genes are assumed to operate independently of each other, and hence prior to analysis we form the assumption that $\tau_{q,g} = 0$. Under this assumption of no interaction, we would expect to observe the strain fitness for a double mutant as

$$\theta_{q,g}^{neut} = \theta_{WT} + \tau_q + \tau_g \qquad \left(\tau_{q,g} = 0\right)$$

Hence, we are interested in forming a statistical test that can assess whether our growth curve data set gives enough evidence for us to conclude whether the different $\tau_{q,g}$ are significantly different from 0. This assumption of 'no interaction' in an additive model is typically given the name *additive neutrality*. Different models for the expected fitness of double mutants, or *neutrality models*, are studied and used throughout the literature, and the choice of neutrality model can have an impact on the results of such a study. (Mani, St Onge, Hartman, Giaever, & Roth, 2008) The framework for analysis of these types of experiments is fairly general, and may also be applied to, for example, *gene-treatment* experiments, in which the strength of a drug is different from gene to gene. (Doostzadeh, Davis, Giaever, Nislow, & Langston, 2007)

Medical outcomes provide a concrete motivation for growth curve experiments. Many experiments are designed with the hope of shedding light towards development of new treatments for cancer. For example, in renal and prostate cancers, it is observed that there is a reduced expression of the gene *ctf8*. To learn more about how this gene functions in conjunction with other genes, a growth curve experiment can be designed that examines and compares how the growth of strains of yeast with the gene *ctf8* removed compare to both wild type, and also to double mutants in which *ctf8* and a different, non-essential gene is deleted. If one were to identify a non-essential gene whose singular deletion does not impact cellular growth strongly, but does have a strong negative effect on the growth on its corresponding double mutant with *ctf8*, one could imagine building a treatment that suppresses the expression of this non-essential gene, which could effectively help to kill the cancerous cells while leaving non-cancerous cells relatively unharmed.

Depending on the assumptions made and growth curves observed, different statistical modeling strategies will be available for different growth curve data sets. I will begin with an introduction to the widely used exponential and logistic growth models, the commonly used definitions of strain fitness associated with these models, and the different methods of estimating these strain fitnesses available. However, it is common for the observed growth curves to depart from the commonly used growth models in different ways. I will describe the different departures from exponential and logistic growth observed in different growth curves, and the consequences these departures can have on their fitting. It is quite common that these departures are so severe that we are unable to use parametric models or their associated definitions of strain fitness. Furthermore, often we find that the growth curves obtained for double mutants of most interest (that is, double mutants that seem to show synthetic interaction) are non-sigmoidal in nature. When faced with severely non-sigmoidal growth curves, we are forced to use a non-parametric definition of strain fitness – a choice which has not been well explored in the literature. I will evaluate a few of the different statistical methodologies that appear commonly in the yeast growth curve literature on their ability to detect significant interaction effects, and propose a derived variables analysis on area under the curve (AUC) and area under the logged curve (AULC) as non-parametric modeling strategies that can effectively bridge the gap between the canonical sigmoidal growth curve shapes expected and the non-sigmoidal growth curves often seen in different growth curve experiments. We will see that many of the methodologies based on exponential or sigmoidal growth break down when faced with a large amount of growth curves, and hence the ease of implementation and scalability of this derived variables analysis, combined with the AUC / AULC's conceptual soundness as a definition of strain fitness, will prove to be very useful when attempting to identify interaction effects from a growth curve experiment.

There are two sets of growth curves that I will use throughout this thesis.

One set of growth curves from the 'Stoepel' data set is almost entirely sigmoidal or exponential in form. In this set of data, the query gene *ctf8* was knocked out in tandem with 45 other genes, with hopes of identifying strong interaction effects. A sample set of growth curves from a single plate in the Stoepel dataset is plotted in Figure 1.1. It is seen that there is

typically a very small amount of noise between growth curves sharing the same deletion status; however, differences from one strain to another can be quite evident. Many of the definitions of strain fitness used in the literature come from the exponential and logistic models prompted by these well-behaved sigmoidal growth curves.

The second set of growth curves come from the dataset hereby called McLellan. Three query genes associated with the progression of cancerous tumour growth, *scc1-73*, *smc1-259*, and *scc2-4*, are knocked out in combination with 31 other genes. Once again, it is hoped that through this experiment different therapeutic targets could be discovered. Figure 1.2 shows a set of growth curves from one plate of the McLellan dataset, on which *scc1* and *smc1* are knocked out in conjunction with a set of non-essential genes. We see a heterogeneous mix of sigmoidal and non-sigmoidal growth curves here, and the presence of these non-sigmoidal growth curves prompts the investigation of model-free definitions of strain fitness.

Recalling that the primary goal is to characterize and compare the interaction effects, it is useful to plot point-wise averaged growth curves for the wild type, single mutants A and B, and double mutant together to obtain a visual representation of whether or not we are seeing genetic interaction. Figure 1.2 presents, from the McLellan data set, these averaged growth curves when inspecting the double mutant *scc1, irc15* for interaction effects. The estimated strain fitness scores, as calculated using AUC as a definition of strain fitness, are represented below the figure in a dot plot to give a visual guide of the differences. Visual inspection shows that the singular deletion of *scc1* or *irc15* has a slightly positive effect on growth relative to wild type, while the growth of the double mutant of *scc1* and *irc15* is markedly lower than the other growth curves. This presents an example of genetic interaction. This is what we wish to determine, for dozens or hundreds of gene pairs at once, in the presence of highly heterogeneous growth curves, and in a fairly automated fashion. Of particular note is the non-sigmoidal nature of the *scc1, irc15* double mutant. Certainly, this double mutant shows a relatively low level of growth representative of genetic interaction; however, the curve's non-sigmoidality would cause difficulties when considering the use of an exponential or logistic growth model. It is the accommodation of these kinds of non-sigmoidal growth curves that is of primary importance in this thesis.

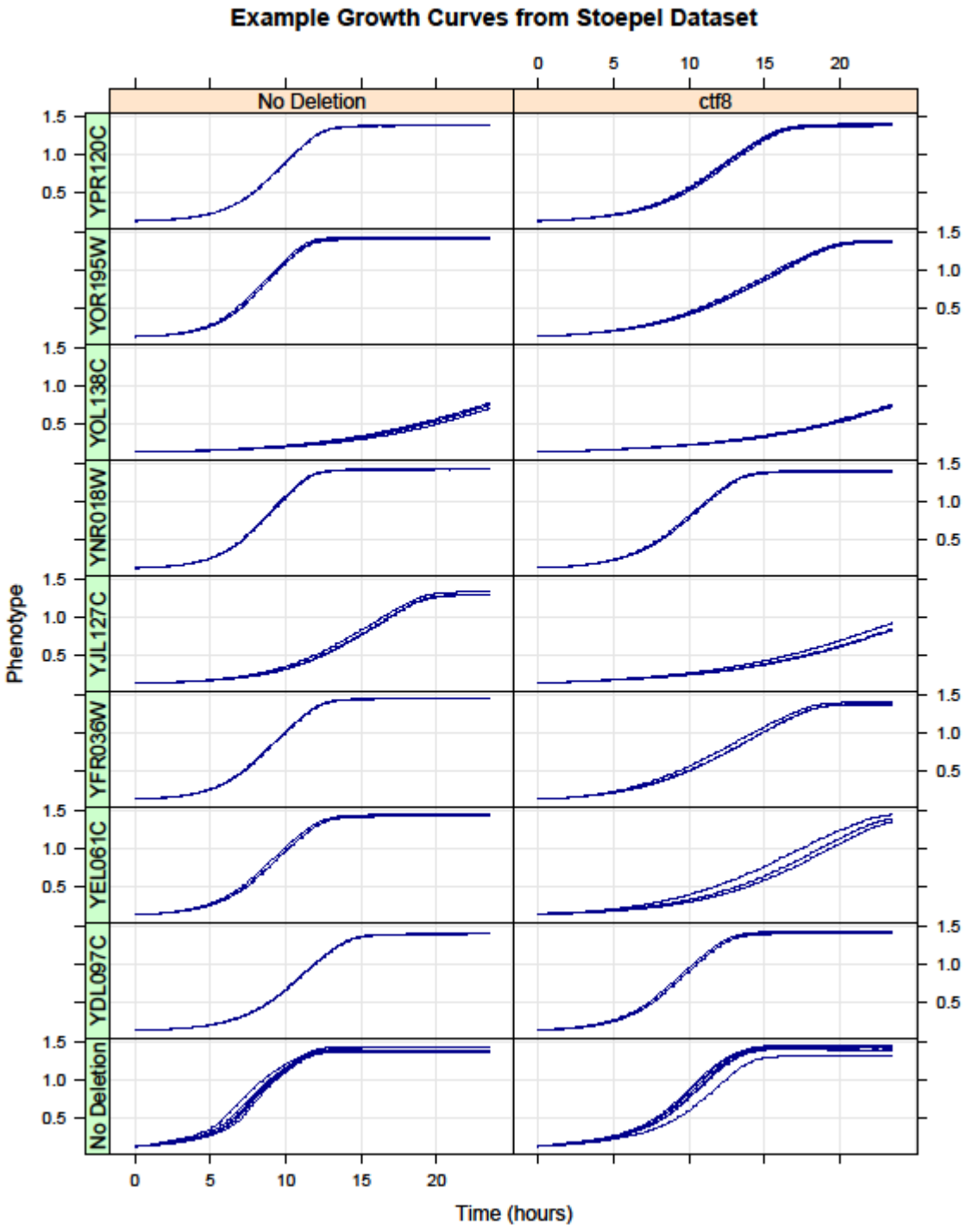

Figure 1.1 Sample set of growth curves from Stoepel data set

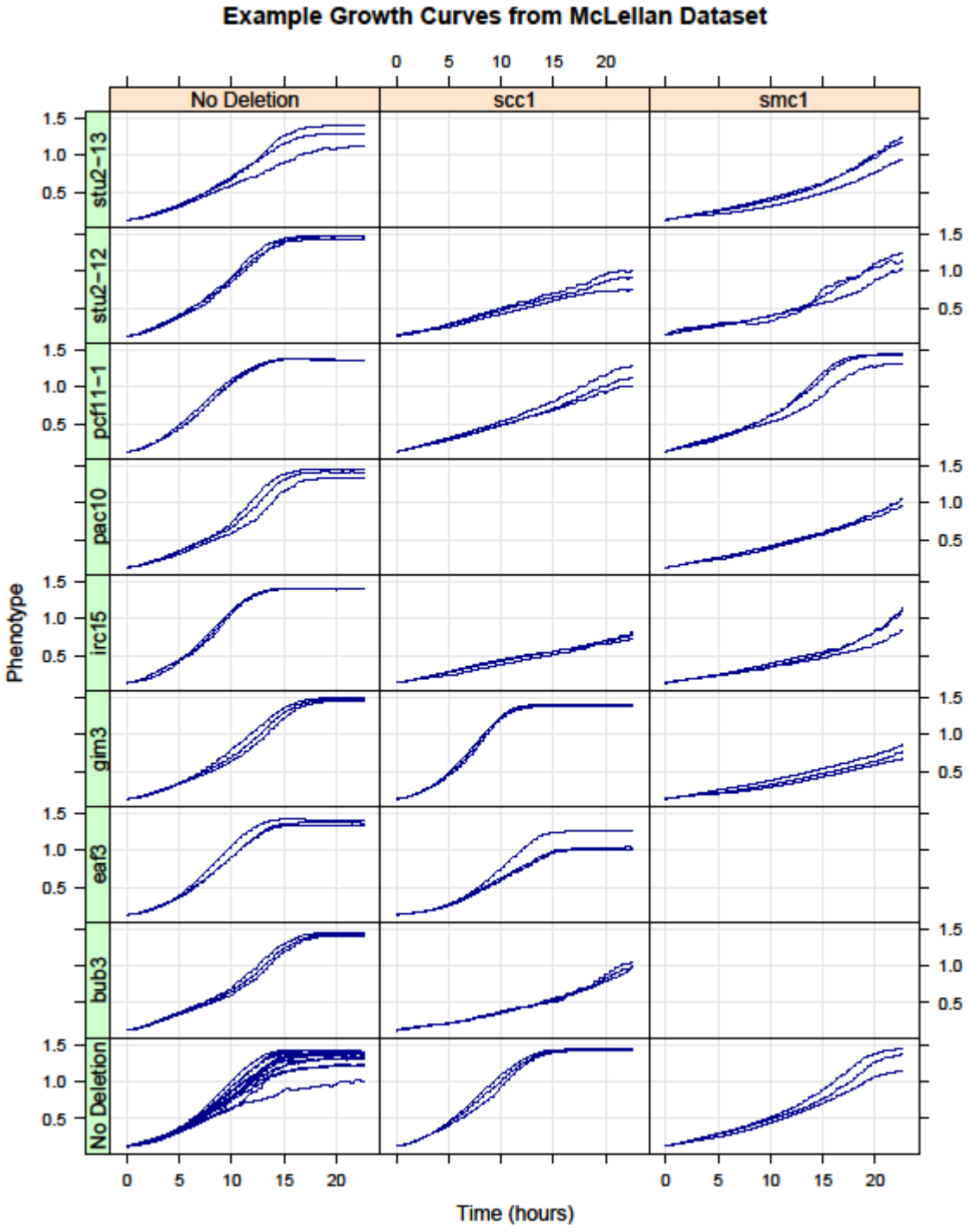

Figure 1.2 Sample set of growth curves from McLellan data set

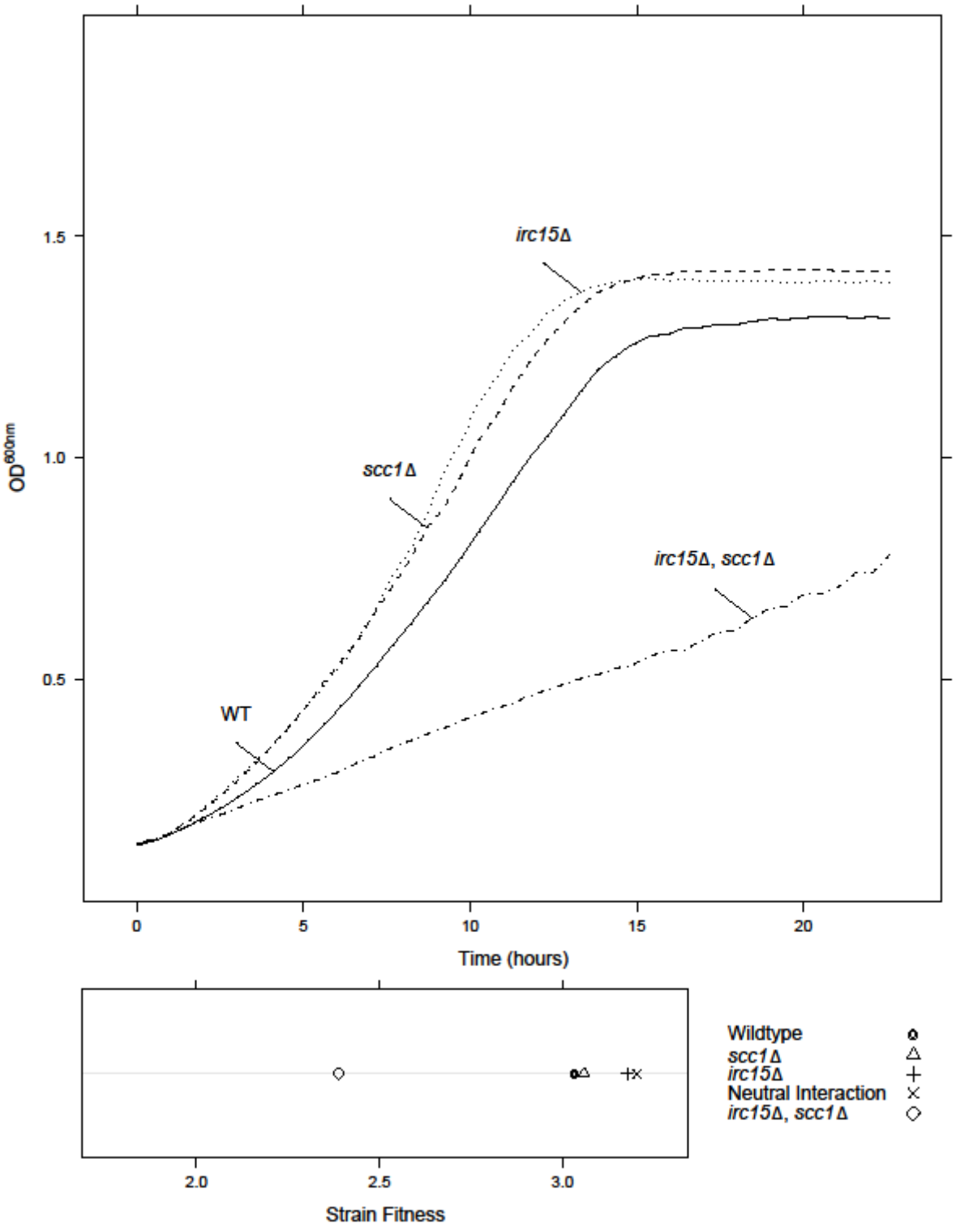

Figure 1.3 Assessing interaction between *irc15, scc1*

Chapter 2: Growth Models and Strain Fitness

This chapter is dedicated to the introduction of the commonly used growth models seen throughout the literature. We begin with a gentle introduction to exponential growth and the commonly used parameterizations of the exponential growth function. Logistic growth will then be introduced as an extension of exponential growth, with the introduction of a growth constraint. Exploration of these models is important as many definitions of strain fitness in growth curve experiments are derived from these growth models.

2.1 Exponential Growth

Exponential growth occurs when the derivative of a function is proportional to the current value of the function. Let $y(t)$ be the value of the function at time t, generally thought of as population size. An exponential growth function will satisfy the following equation

$$\frac{d}{dt}y(t) = \text{constant} \times y(t)$$

Solving for $y(t)$ yields an expression of the following functional form

$$y(t) = y_0 B^t$$

where $y_0 = y(0)$ is the initial quantity. The quantity B can be written in two different ways, corresponding to different natural parameters. The first way is to express B as a base b raised to a power $\tilde{r}(b)$.

$$y(t) = y_0 b^{\tilde{r}(b)t} \qquad \left(B = b^{\tilde{r}(b)}\right)$$

We call $\tilde{r}(b)$ the exponential growth constant and it is the natural parameter used in defining exponential growth this way. It is written as a function of b because its actual value depends on the base chosen. Most commonly, the base b is taken to be e, and I will write $\tilde{r}(e) = r$ for this important special case. This gives the expression

$$y(t) = y_0 e^{rt}$$

For an arbitrary base *b*, the above function may be rewritten as

$$y(t) = y_0 b^{\frac{r}{\log(b)}t} \qquad \tilde{r}(b) = \left(\log b\right)^{-1} r$$

So, $\tilde{r}(b)$ and r will differ only by a multiplicative constant. More generally, two exponential growth constants with arbitrary bases b and b' can be related by the expression

$$\tilde{r}(b) = \frac{\log b'}{\log b}\tilde{r}(b')$$

The second way to parameterize exponential growth is through expressing the quantity *B* as

$$B = b^{1/\tilde{d}(b)} \Rightarrow y(t) = y_0 b^{t/\tilde{d}(b)}$$

Here, $\tilde{d}(b)$ is the time required for *y* to increase by a multiplicative factor of *b*, and we note that by definition, $\tilde{r} = 1/\tilde{d}$. Henceforth I call $\tilde{d}(b)$ the multiplicative time, and it is the time it takes for the function to increase from *y* to *by*. Doubling time – the time required for a population to double in size – is a popular concept and implies a choice of base *b* = 2. This important special case will be specified as $\tilde{d}(2) = d$, and gives rise to the expression

$$y(t) = y_0 2^{t/d}$$

By similar arguments used for the natural parameter $\tilde{r}$, $\tilde{d}(b)$ and $\tilde{d}(b')$ will differ only by a multiplicative constant.

Next, I relate the two of the most common choices in parameterization, *r* and *d*. We can determine the relationship between the two as follows:

$$y_0 e^{rt} = y_0 2^{t/d} \Rightarrow e^r = 2^{1/d}$$

Doing this, we can write

$$r = d^{-1}\log 2, \qquad d = r^{-1}\log 2$$

$$r = \frac{1}{d\log_2 e} \qquad d = \frac{1}{r\log_2 e}$$

Given y_0, exponential growth can be completely parameterized by one parameter, and different choices of parameterization lead to different quantities of interest. We note that the assumption of exponential growth leads to two of the canonical definitions of strain fitness,

$$\theta = r\ , \qquad \text{or} \qquad \theta = d\ ,$$

as only one parameter (assuming y_0 is known or fixed) is required to fully define an exponential growth function.

2.2 Learning *r* and *d* from Exponential Growth

Suppose that one or more points from an exponential growth function $y(t)$ are observed, but the natural parameter of interest r or d is unknown. How might we use these points to calculate our parameter of choice? I first consider low-tech methods in the absence of error. These are methods commonly used in different published papers, and although in such papers these calculations are often formed without specific reference to the exponential growth model, it will be illuminating to identify how they are developed.

2.2.1 Single Time Point

Suppose that it is known that $y(t)$ is growing exponentially, but only one time point $t = t^\star$ is observed. We want to try and determine r or d based on this single point. A calculation for each of the natural parameters can be written as

$$r = \frac{1}{t^\star}\log\frac{y(t^\star)}{y_0}, \qquad d = t^*\left(\log_2\frac{y(t^\star)}{y_0}\right)^{-1}$$

Note that in each case, to calculate r, we require knowledge of y_0, the initial quantity.

2.2.2 Double Time Point

Suppose we observe two points from an exponential growth function $\left(t_k, y(t_k)\right),\ k=\{1,2\},\ t_1<t_2$. Let $\Delta = t_2 - t_1$ be the time difference between the two points. We can calculate each of the natural parameters as

$$r = \frac{1}{\Delta}\log\frac{y(t_2)}{y(t_1)}, \qquad d = \Delta\left(\log_2\frac{y(t_2)}{y(t_1)}\right)^{-1}$$

Note the equivalence of this expression with that of the single time point expression, as knowledge of the point y_0 is required in what we call the single time-point calculation. That is, this form reduces to the single time point form if we take $t_1 = 0$, and $t_2 = t^\star$. The introduction of a second observed point eliminates the need for knowledge of the initial quantity y_0. It is noted that the log-ratio allows us to calculate *r* from two data points from an exponential growth function – that is, when faced with exponential growth the log-ratio is a primary means of learning about *r* or *d*.

The above expressions for *r* and *d* can be simplified if we choose the two time points with a bit more care. Suppose that two time points t_1 and t_2 are selected from an exponential function, with t_2 chosen to be *g* generations ahead of t_1. That is, $y(t_2) = b^g \times y(t_1)$. I will write $t_2 - t_1 = \Delta(b)$ to emphasize that the length of time chosen will be a function of the base chosen as well. Suppose our interest was focused on multiplicative time $\tilde{d}(b)$. First, we note that:

$$\tilde{d}(b) = \Delta(b)\left(\log_b\frac{y(t_2)}{y(t_1)}\right)^{-1} \Rightarrow \log_b\left(\frac{y(t_2)}{y(t_1)}\right) = \frac{\Delta(b)}{\tilde{d}(b)}$$

Next, we note that we can also re-write $y(t_2) = b^g \times y(t_1)$ as

$$\log_b\left(\frac{y(t_2)}{y(t_1)}\right) = g\,.$$

Combining the above gives the expression $\frac{\Delta(b)}{\tilde{d}(b)} = g$. Rearranging to isolate $\tilde{d}(b)$, we get

$$\tilde{d}(b) = \frac{\Delta(b)}{g}$$

We note that taking the time points to be explicitly g expressions apart greatly simplifies the final expression of d as a function of the time points chosen, and is often done when attempting to determine d through the use of two time points. (Lee et al., 2005; St Onge et al., 2007)

Similarly, one might want to determine $\tilde{r}(b)$ by choosing two time points to be once again g generations apart. Doing this will give a similar expression,

$$\tilde{r}(b) = \frac{g}{\Delta(b)}$$

Determining r and d can be done by taking bases $b = e$ and $b = 2$ respectively.

2.2.3 Error and Estimation of *r*

If the exponential growth model holds exactly true, then only two points are required for determination of r (one if we consider the case where y_0 is known). However, observed data never conforms to the model exactly: there will be some error in our observations of $y(t)$, and these errors will propagate into errors in the estimates of r. It is this error that prompts us to compare different methods of estimating these parameters. There is some notion that, with more data available, or more care taken in choosing the two time points, the precision of an estimator for r should improve. Hence, I will consider a simple model for this error. Suppose we model the error multiplicatively as a log-normal random variable:

$$y(t_k) = y_0 e^{rt_k} \times \exp(\epsilon_k), \qquad \exp(\epsilon_k) \sim LogN(0, \sigma^2) \;\Rightarrow\; \epsilon_k \sim^{iid} N(0, \sigma^2)$$

That is, the error at any particular time point t_k is independent of the error at other time points, and $Var(\epsilon_k) = \sigma^2$ is constant over time. I will consider how this model motivates or

demotivates different potential estimates of *r* specifically; however, similar logic will motivate and demotivate the other available parameterizations as well.

2.2.3.1 Error and the Distance between Two Time Points

First, I show that choosing two time points from an exponential function far apart can improve the precision of an estimator of *r*. Recall the double time point method of calculating *r*, and let $\hat{r}$ be the estimator of *r*. The error will propagate through this model as:

$$\hat{r} = \frac{1}{\Delta}\log\frac{y(t_2)\times\exp(\epsilon_2)}{y(t_1)\times\exp(\epsilon_1)} = \frac{1}{\Delta}\log\frac{y(t_2)}{y(t_1)} + \frac{1}{\Delta}\left(\epsilon_2 - \epsilon_1\right) = r + \frac{1}{\Delta}\left(\epsilon_2 - \epsilon_1\right)$$

We note that the first term in the expression is *r*, while the second term encapsulates the error in estimation of *r*. The variance of this term is

$$Var\left(\frac{1}{\Delta}\left(\epsilon_2 - \epsilon_1\right)\right) = \frac{2}{\Delta^2}\sigma^2$$

Hence, the variance of this estimator is smaller for larger Δ, and this provides a primary motivation for picking two points on an exponential growth function to be as far away as possible.

2.2.3.2 Averaging Multiple Double Time Point Estimates of *r*

Suppose we are once again considering the exponential growth model with multiplicative error, and we are interested in estimating *r*. For an exponential function observed at many points, this quantity could be estimated based on any two different time points. What if we imagined computing a number of estimates of *r* based on multiple double time point measures, and combined these estimates in an attempt to improve the precision in our estimate of *r*? I consider one such scenario.

Suppose that a set of exponential growth constants are calculated sequentially over a set of equally spaced time points t_k, $k = \{1, 2, ..., K\}$. That is, $\Delta_k = t_k - t_{k-1} = \Delta$. We can estimate *r* from each duo of time points t_k and $t_{k\text{-}1}$ as follows:

$$\hat{r}_k = \frac{1}{\Delta}\log\frac{y(t_k)\exp(\epsilon_k)}{y(t_{k-1})\exp(\epsilon_{k-1})}$$

In the absence of error, we should have each $\hat{r}_k$ equal to r. However, in the context of error, each of these estimators will deviate about r, and hence we might expect that an averaging of these terms would lead to an improved estimate of r.

$$\hat{r} = \frac{1}{K-1}\sum_{k=2}^{K}\hat{r}_k = \frac{1}{K-1}\sum_{k=2}^{K}\frac{1}{\Delta}\log\frac{y(t_k)\exp(\epsilon_k)}{y(t_{k-1})\exp(\epsilon_{k-1})}$$

We rearrange the above expression a bit:

$$\hat{r} = \frac{1}{\Delta(K-1)}\sum_{k=2}^{K}\left(\left(\log y(t_k) - \log y(t_{k-1})\right) + \left(\epsilon_k - \epsilon_{k-1}\right)\right)$$

We note that the summation term is a telescoping series, and hence the above expression reduces to

$$\hat{r} = \frac{1}{\Delta(K-1)}\left(\left(\log y(t_K) - \log y(t_1)\right) + \left(\epsilon_K - \epsilon_1\right)\right)$$

Rearranging once again, we obtain the expression

$$\hat{r} = \frac{1}{\Delta(K-1)}\log\frac{y(t_K)\exp(\epsilon_K)}{y(t_1)\exp(\epsilon_1)}$$

We note that $\Delta(K-1)$ is simply the length of time between t_K and t_1, and hence this averaging has reduced the expression to an estimation of r based on only the first and last time points observed. Hence, an averaging of these potential double time point estimators $\hat{r}_k$ is equivalent to the double time point estimator based on just the first and last points observed, and cannot be more precise relative to that double time point estimator.

In the context of error, it seems it would be useful to consider using more than two data points at a time to estimate r, as there is some expectation that the inclusion of more data points in different estimation methods could still lead to improved estimation of r.

2.2.3.3 Fitting a Line

Suppose that exponential growth is observed for a single population, with the aforementioned multiplicative error model. We might want to use all the time points at once to obtain an estimate of r. We observe that, if we take the logarithm of $y(t_k)$, we obtain the expression:

$$\log y(t_k) = \log y_0 + rt_k + \epsilon_k$$

This is the equation for a simple linear model, with intercept equal to log y_0, slope equal to r, and random error term ϵ_k. We could imagine using this model to fit the logged data, and using the slope of the fitted line as an estimate of the exponential growth constant r, and the intercept as a means of estimating y_0 if necessary. By including more and more data in the model fit, we should be able to increase the precision in our estimate of r. The value of least-squares estimators is well understood and hence I do not further explore the mathematical properties of the least-squares estimate of r.

2.2.4 Non-Parametric Definitions of Strain Fitness

I will now begin introduce non-parametric definitions of strain fitness which are not necessarily developed directly from the exponential growth model. However, I will relate these definitions back to the exponential growth model and r where possible. These definitions are introduced now as we will need to consider such non-parametric definitions when considering departures from the exponential model.

2.2.4.1 Specific Rate

The **specific rate (SR)** of a function $y(t)$ is calculated as the ratio of the slope the function at a time t divided by the value of that function at time t:

$$\frac{y'(t)}{y(t)} = SR$$

Recall that exponential growth can be defined through the expression

$$y'(t) = \frac{d}{dt} y(t) = ry(t)$$

Because of this, one potential means of calculating r is through the specific rate. For the exponential growth function, we see that

$$\frac{y'(t)}{y(t)} = r = SR$$

This quantity can be calculated at any time point from an exponentially growing function. Typically in growth curve experiments, exponential growth is only observed in a small region of the growth curve, with sub-exponential growth seen in other areas of the curve. Hence, r is often estimated through the specific rate as (Shah, Laws, Wardman, Zhao, & Hartman, 2007)

$$\hat{r} = \max \frac{y'(t)}{y(t)}$$

It is hoped that taking the maximum will leave the estimate associated with maximal, and hence exponential, growth. In order to estimate these quantities then, one must consider how the slope of the function could be estimated, and different smoothing techniques are often employed before estimation of this derivative. Specific rate has seen some use in the literature as a means of calculating r and hence strain fitness (Shah et al., 2007), and will be explored later in empirical studies.

2.2.4.2 Area Under the Curve, Area Under the Logged Curve

Starting now, I begin to explore potential definitions of strain fitness that begin to deviate from the canonical parameters r and d. Exploration of these definitions will prove more fruitful when considering departures from the exponential growth model; however, it is useful to relate these quantities back to the exponential growth constant r for when the exponential growth model does hold.

Suppose that an exponential growth function is measured continuously, rather than over a set of time points. We might hope that the area under this curve could prove a good characterization of strain fitness. The area under the exponential growth function between two time points t_1 and t_2 can be calculated as

$$AUC = \int_{t_1}^{t_2} y(t)dt = \frac{y_0 e^{r\Delta}}{r}$$

It is noted that the AUC, although certainly not equal to *r*, is still a function of *r*. The AUC has seen some use in the literature as a definition of strain fitness, but I seek to develop it more thoroughly in this thesis. (Addinall et al., 2011; Hartman & Tippery, 2004)

A second quantity I will investigate, the area under the logged curve (AULC), can be calculated as

$$AULC = \int_{t_1}^{t_2} \log y(t)dt = \Delta \log y_0 + r\frac{\Delta^2}{2}$$

As a matter of interest, we note that this expression can be rearranged to be in terms of *r* as

$$r = \frac{2}{\Delta^2}\left(AULC - \Delta \log y_0\right)$$

The utility of these definitions of strain fitness will be explored later in empirical studies when we consider different departures from exponential growth. In actual experiments, because only a finite number of time points will be measured, the integrals will have to be estimated based on the observed data – this too will be discussed later.

2.3 Comparing Two Populations

Suppose we have two populations, or different exponential growth functions, $y_i(t)$ and $y_j(t)$. We might seek to calculate the **difference** in strain fitness $\theta_i - \theta_j$. How might this quantity be calculated, or estimated, based on the previously described methods?

First, I consider the canonical case in which we use *r* as our definition of strain fitness, so that $\theta_i - \theta_j = r_i - r_j$. The previously discussed methods can provide different insights towards these two quantities. We will consider the case where the two populations share the same starting quantity, such that $y_i(0) = y_j(0) = y_0$, as is common in growth curve experiments. This section is introduced first without considering error.

2.3.1 Single Time Point

Suppose the value of two exponential functions is observed at some time $t^\star$. Based on a single time point measure for two populations $y_i(t)$ and $y_j(t)$, it is possible to calculate $r_i - r_j$:

$$r_i - r_j = \frac{1}{t^\star} \log \frac{y_i(t^\star)}{y_j(t^\star)}$$

That is, if the difference between two exponential growth constants r_i and r_j is of primary interest, this quantity can be determined based on a single time point measure from each population. Notice that, assuming that y_0 is shared between the two populations, knowledge of y_0 is not required in estimating this difference. Once again, we see the log-ratio arise naturally as our means of calculating r.

As a side-note, it is worth noting we can re-write the above expression as

$$r_i - r_j \propto \left(\log y_i(t^\star) - \log y_j(t^\star) \right)$$

This gives a primary motivation for using a 2-way ANOVA model on the logged data measured as a means of performing an interaction analysis, a common technique employed in the literature. (McLellan et al., 2009) This will be further explored later in the thesis.

2.3.2 Double Time Point

Since it is possible to determine r as long as two points are observed for a given curve $y_i(t)$, these r may then be compared in whichever way desired. We can consider calculating r for each population and taking the difference,

$$r_i - r_j = \frac{1}{\Delta} \log \frac{y_i(t_2)}{y_i(t_1)} - \frac{1}{\Delta} \log \frac{y_j(t_2)}{y_j(t_1)}$$

or attempting to calculate r based on an average of two single time point measures of $r_i - r_j$ calculated at t_1 and t_2.

$$r_i - r_j = \frac{1}{2} \left(\log \frac{y_i(t_1)}{y_j(t_1)} + \log \frac{y_i(t_2)}{y_j(t_2)} \right)$$

It is easy to show that these expressions are equivalent.

2.3.3 Fitting a Line

In terms of fitting a linear model to the logarithm of $y(t)$, one might consider introducing a categorical covariate so that $r_j - r_i$ could be interpreted directly as a parameter in the model. Define the model:

$$y(t_k) = \log y_0 + r_i t_k + C(\beta t_k) + \epsilon_k$$

where C is a dummy variable equal to 0 if the observation belongs to population i, and 1 if the observation belongs to population j, and ϵ_k is the random error term. β can hence be interpreted as $r_j - r_i$.

2.3.4 Area Under the Curve

By defining either $\theta = AUC$ or $\theta = AULC$, we can easily compute the difference in strain fitness as:

$$\theta = AUC \Rightarrow \theta_i - \theta_j = AUC_i - AUC_j$$
$$\theta = AULC \Rightarrow \theta_i - \theta_j = AULC_i - AULC_j$$

Once again, it is interesting to assess the relationship between AUC, AULC and r under the exponential growth model.

2.3.4.1 AUC for Two Populations

The log ratio of two AUCs can provide insight to the difference between two exponential growth constants r_i and r_j:

$$\frac{1}{\Delta}\log\left(\frac{AUC_i}{AUC_j}\right) = (r_i - r_j) - \frac{\log r_i - \log r_j}{\Delta}$$

If Δ is suitably large relative to $\log r_i - \log r_j$, the second term in the above expression will be small, and the log ratio of AUCs will approximately isolate the difference between the growth constants r_i and r_j. This provides some motivation for calculating the total area under

the curve, in which we choose the time points $t_1 = 0$, and t_2 to be the highest observed time point available.

2.3.4.2 AULC for Two Populations

The difference in two AULCs between different populations can be written as

$$AULC_i - AULC_j = \left(\Delta \log y_0 + r_i \frac{\Delta^2}{2} \right) - \left(\Delta \log y_0 + r_j \frac{\Delta^2}{2} \right) = \frac{\Delta^2}{2} \left(r_i - r_j \right)$$

From this, we note that one could calculate $r_i - r_j$ as

$$r_i - r_j = \frac{2}{\Delta^2} \left(AULC_i - AULC_j \right)$$

We note that knowledge of y_0 is no longer required for calculation of this quantity.

2.3.4.3 Estimation of AUC, AULC

To estimate the AUC / AULC, we must consider how to estimate a continuous integral based on a finite set of data points. In this thesis, I perform the approximation using composite Simpson's rule, which is a method of approximating an integral by the area under a set of quadratic functions. Suppose I wish to integrate a continuous function $f(x)$ from points a to b. Values from $f(x)$ are observed at points x_0, x_1, …, x_{K-1}, and x_K, with lower endpoint $a = x_0$ and upper endpoint $b = x_K$. The integral $\int_a^b f(x)dx$ can be approximated as

$$\approx \frac{(x_K - x_0)}{3K} \left[f(x_0) + 4f(x_1) + 2f(x_2) + 4f(x_3) + 2f(x_4) + \ldots + 4f(x_{K-1}) + f(x_K) \right]$$

The amount of error due to this approximation method should be very small, and its impact is considered negligible. Of more interest is the impact of error on estimation of AUC or AULC. Rather than an explicit derivation of how error might propagate when calculating the AUC, I opt to give a heuristic argument for why the AUC should be insensitive to random error. There are two forces that are reducing the effect of random error on the measure of the AUC. First, Simpson's rule itself is a quadratic interpolator, and this interpolation should reduce the variance in the estimator of AUC, but with potential increases in bias. (Hastie & Tibshirani,

1990) Secondly, under the assumption that $\epsilon_k \sim^{iid} N(0, \sigma^2)$, the effect of positive random deviations on the AUC should be cancelled out, 'on average', by the effect of negative random deviations. This argument should hold true for the AULC as well, and provides the primary motivation for calculating the AUC / AULC over the largest range of time points available – typically an initial time point and a common terminal time point.

2.3.5 A Small Comparative Study of Fitness Measures

First, I consider estimation of strain fitness from a single exponential function with multiplicative error. I generate 10000 curves from the model:

$$y(t_k) = e^{0.2t_k} \times \exp(\epsilon_k), \qquad \epsilon_k \sim^{iid} N(0, 0.1^2)$$

Suppose $y_0 = 1$ is known. Time points are equally spaced and taken at 49 points $t_k = \{0, 0.5, ..., 23.5, 24\}$, similar to a 24-hour growth curve experiment. The figure below shows five of the generated exponential curves in colour over the true function in black. It is worth stating that the amount of random error in these curves, while small, is still larger than what we might expect to see in a growth curve experiment. However, it is still worth investigating empirically the performance of different estimators under such a model.

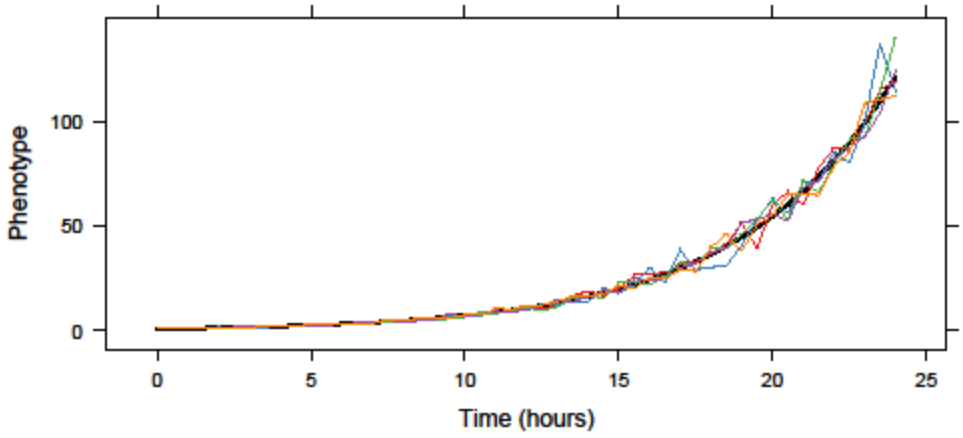

Figure 2.1 Example exponential curves

I explicitly compare the double time point (DTP), simple linear model (LM), AUC, AULC and specific rate (SR) measures to compare their precision in estimation of strain fitness. Note that the DTP, LM, and SR methods are all methods for estimation of r; hence, I compare these estimators relative to the true r, while AUC and AULC are compared to the 'true' values under the model. For $y_0 = 1$, $r = 0.20$ and $\Delta = 24$, we have AUC = 607.5521 and AULC = 57.6. The best estimation methods will be closer to their associated true strain fitness 'on average'; hence, I will evaluate each of these methods in terms of their *coefficient of variation* (CV):

$$CV(\hat{\theta}) = \frac{SE(\hat{\theta})}{\theta}$$

To ensure the methods are comparable, I use the knowledge that $y_0 = 1$ in each estimation method. This implies, for example, that the LM approach is done with forcing the intercept to be log(1) = 0.

For the specific rate, the derivative is estimated in two ways. First, I use the same method as calculated in Shah et al (SR1), without any pre-smoothing. The slope at time point t_k is estimated as:

$$y'(t_k) \approx \frac{y(t_k) - y(t_{k-1})}{t_k - t_{k-1}}$$

Secondly, I use a cubic smoothing spline (SR2) to estimate $y'(t)$ through the function `smooth.spline` in R, with default values for each argument. (Hastie & Tibshirani, 1990) For these two scenarios, I consider estimation of r as $\hat{r} = \max\left(y'(t)/y(t)\right)$.

Finally, I consider one last estimator based on the specific rate (SR3) whereby I estimate r by $\hat{r} = avg\left(y'(t_k)/y(t_k)\right)$, with $y'(t)$ estimated with smoothing splines, and *avg* denoting the mean slope taken over time points t_k.

From the 10000 estimates produced from each method, the coefficients of variation (CV) are computed:

	DTP	LM	AUC	AULC	SR1	SR2	SR3
CV	2.95%	0.515%	2.25%	0.601%	12.3%	52.1%	3.63%

Table 2.1 Single population comparison of methods

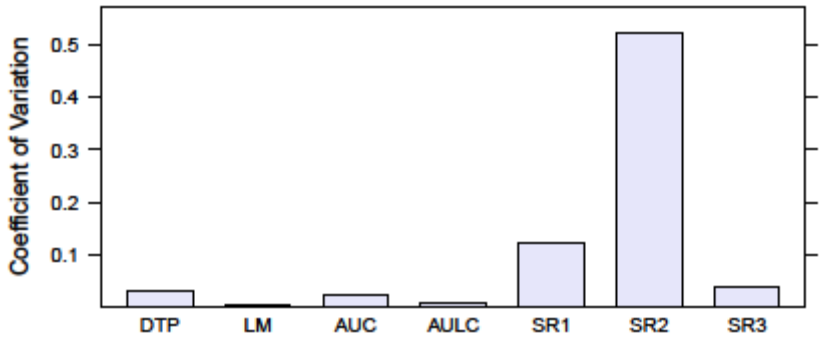

Figure 2.2 Single population comparison of methods

As expected, the linear model approach performs the best in estimation of *r*, producing the estimator with lowest coefficient of variation. However, we note that AULC is a very close contender. The high coefficient of variation as measured for the specific rates SR1 and SR2 suggest they may be poor estimators of *r*. However, SR3 seems a fairly good non-parametric estimator of *r*.

Next, we look at comparison of two populations. Consider the same model with error as before, but now with two populations with growth constants $r_1 = 0.2$ and $r_2 = 0.15$. I generate 10000 curves with 49 time points for each as described before and estimate the difference $r_1 - r_2$ through the previously developed formulae for each method.

Note that, under this model, we have $r_1 - r_2 = 0.05$, $AUC_1 - AUC_2 = 363.5639$, and $AULC_1 - AULC_2 = 14.4$.

	DTP	LM	AUC	AULC	SR1	SR2	SR3
CV	16.8%	2.90%	3.20%	3.38%	374%	78%	19.9%

Table 2.2 Two population comparison of methods

Based on the table, we see that all methods of estimating the specific rate are fairly high in variance – even with smoothing. At this point, I discount SR as a potential non-parametric estimator of *r* in lieu of AUC and AULC, which will be my primary alternative definitions of strain fitness.

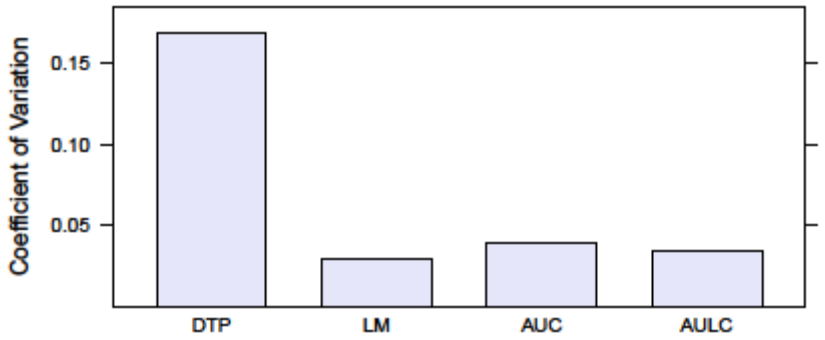

Figure 2.3 Two population comparison of methods

Once again, LM is the winner; however, the race is even tighter than before between LM, AUC, and AULC. The difference in CV between the three methods is very small, lending credence to their further assessment throughout the thesis.

As a reminder, AUC and AULC are being proposed as quantities that remain useful definitions of strain fitness when faced with more serious model departures; however, seeing that their performance is quite good when the exponential growth model holds true is reassuring. It will be interesting to see which of the two performs better when faced with different kinds of departures from exponential growth.

In reality, for a growth curve experiment, exponential growth is observed only in the earlier stages of the experiment; the entirety of a growth curve is typically sigmoidal in shape due to constraints inhibiting growth. Hence, logistic growth models are proposed next as an extension of the exponential growth models, allowing for this constraint on growth.

2.4 Logistic Growth Models

Logistic growth is conceptualized under a similar framework to exponential growth, with the introduction of a growth constraint. In a growth curve experiment, as the population of yeast cells increases, more members must fight over a limited pool of resources, toxic buildup of waste dampens growth, and space constraints prevent growing beyond certain bounds. This provides motivation for modeling this growth constraint as a function of current population size. I write

$$\frac{d}{dt}y(t) = f(y) \times y(t)$$

to indicate the growth rate of this function depends on some non-increasing function of the current population size, as well as the current population size. This function $f(y)$ should be approximately equal to the constant r at early time points, and should decrease as y gets larger. Suppose we choose to model it as a linear function of the population size,

$$f(y) = r - \lambda y \ , \qquad \lambda > 0$$

Substituting this into the above expression gives

$$\frac{d}{dt}y(t) = y(r - \lambda y) = ry - \lambda y^2$$

We note two key features: for small y, $dy/dt \approx ry$, and hence growth is approximately exponential in the early stages. We observe the derivative is zero when $y = 0$ or $y = r/\lambda$, so we anticipate that $y(t)$ will have horizontal asymptotes at 0 and r/λ. Therefore, we introduce a new parameter $B = r/\lambda$, and rewrite the expression for the derivative as

$$\frac{d}{dt}y(t) = \frac{r}{B}y(B - y)$$

Solving this differential equation yields the expression,

$$\log\frac{y}{B - y} = rt + C$$

where C is a constant of integration, whose value will be determined by imposing the initial condition that $y(0) = y_0$. First, we note that at time $t = 0$, we can write,

$$C = \log \frac{y_0}{B - y_0}$$

Substituting back into the previous expression, we find

$$rt + C = -r(t - \frac{1}{r} \log \frac{B - y_0}{y_0})$$

We form the definition

$$t_{mid} = \frac{1}{r} \log \frac{B - y_0}{y_0}$$

and finally determine a final expression after solving for $y(t)$:

$$y(t) = \frac{B}{1 + \exp[r(t_{mid} - t)]}$$

This is called the simple logistic function, and is fully determined by y_0, r, and B (noting that t_{mid} is just a function of the other three parameters). y_0 and r have the same interpretations as before, B is called the carrying capacity and is the upper asymptote of the curve, and t_{mid} is the time at which 50% of total growth is reached. Finally, we also recall that B is a function of r and the growth constraint term λ and hence the most natural triplet of parameters is (y_0, r, λ). However, for interpretabilities' sake we write the simple logistic function in terms of y_0, r, B and t_{mid}, with our triplet of free parameters being (y_0, r, B). Once again, in growth curve experiments it is common to fix y_0 through experimental protocol so the two free parameters are hence r and B.

2.4.1 Logistic Growth Models in the Literature

Rather than the aforementioned methods of estimating *r* or *d*, there are a number of papers which explicitly employ the logistic growth function in an effort to estimate strain fitness. I will explore two of these cases in a bit of detail, in an attempt to deconstruct the authors' definitions of strain fitness. This should also help to demonstrate to the reader the flexibility one has in defining strain fitness when modeling growth with the logistic function. Under the exponential growth model, only one parameter is necessary to fully parameterize the model, and hence parameter choice is a question of interpretation. Under the logistic growth model, the presence of new parameters opens new avenues in defining strain fitness as different functions of these parameters.

2.4.1.1 Addinall et al.

A version of the simple logistic model was fit to ensembles of growth curves, each observed over a set of seven to fourteen time points. Addinall et al. parameterized the model in the following way:

$$y(t) = \frac{By_0e^{rt}}{B + y_0(e^{rt} - 1)}$$

This expression is identical to my parameterization of the simple logistic function, with t_{mid} removed as it has been replaced by its corresponding expression in terms of y_0, r, and B. y_0 was set by experimental protocol to be constant across all the different growth curves assessed. Estimates of y_0, r and B were obtained using a least squares fit.

Addinall et al. define strain fitness as a function of these parameters, with strain fitness being the product of two quantities: the maximal doubling rate *MDR*, and the maximal doubling potential *MDP*. These quantities are calculated as the solutions to the following two equations, using their parameterization of the simple logistic function:

$$\frac{y(MDR^{-1})}{y_0} = 2 \qquad\qquad y_0 \times 2^{MDP} = B$$

$$MDR = \frac{r}{\log\left(\frac{2(B-y_0)}{B-2y_0}\right)}, \qquad\qquad MDP = \frac{\log B - \log y_0}{\log 2}$$

Note that these definitions are using 'doubling time' as a focus; hence, even though the exponential base *e* is used in the logistic model, the actual estimates of strain fitness that are of interest to the researchers are stated in base 2 terms – hence, the transformations used. *MDR* is calculated as the inverse of the minimum doubling time observed over their sigmoidal curve; that is, under the assumption that growth is closest to exponential growth in the early stages. *MDP* is the number of doublings the culture is inferred to have undergone, based on the fitted parameters obtained. It is a measure that is larger for curves that reach a higher upper asymptote *B*. Once again, it is transformed to be interpretable in base 2.

By choosing a definition of fitness $\theta = MDR \times MDP$, Addinall et al. seek to give high measures of fitness to strains which grow quickly (*MDR)*, and strains which have undergone more doublings (*MDP*).

One thing worth noting about the author's definition of *MDR* is that, as $B \to \infty$, the expression simplifies to

$$MDR^* = \frac{r}{\log 2}$$

This helps clarify that their definition of *MDR* is very much poking at *r*; more specifically, *r* transformed to be interpretable in base 2, as *MDP* is.

From this short review, we see that there is some flexibility in how strain fitness is defined. Although the canonical way to define strain fitness in some sense remains *r*, because we can imagine that *B* is in fact related to a particular strain of yeast's ability to divide and not entirely determined by experimental protocol, we are given new avenues into definitions of strain fitness that are functions of both *r* and *B*. However, the authors did not construct a

comparative measure of the ability of their definition of strain fitness to identify interaction effects relative to other measures.

2.4.1.2 Shah et al.

Shah et al. compared and evaluated a variety of different growth curve modeling techniques over a set of agar yeast culture arrays. Sigmoidal growth curves were collected from different strains of yeast using automated image analysis, over approximately seven generations. Different definitions of strain fitness were compared over three different models used for the growth curves: 1) raw model-free growth curves, 2) a smoothed spline model, and 3) the simple logistic model. The primary definition of strain fitness assessed in this paper was the maximum specific rate, which is calculated as

$$MSR = \max_{t_k} \frac{y'(t_k)}{y(t_k)}$$

The slope is estimated at each time point, curve by curve, as

$$y'(t_k) \approx \frac{y(t_k) - y(t_{k-1})}{t_k - t_{k-1}}$$

Hence, these *MSR* are computed over 1) the raw growth curves, 2) the spline-smoothed growth curves, and 3) the fitted phenotype obtained after fitting of the logistic model. The authors were interested if the choice of model would affect the variation across replicates in *MSR*. They note that there is a substantial reduction in this variation when the spline model is assumed (relative to the model free fits), and further reduction when the logistic model is assumed.

The authors also included AUC as a definition of strain fitness, in order to see how it performed under the three different models. Interestingly, Shah et al. note that the variation in the measured AUCs did not change much across the different models chosen. The fact that the AUC can be insensitive to the choice of model gives me further justification for its use as a measure of strain fitness, as I seek to propose it when faced with a set of growth curves that do not follow a single parametric model uniformly. A strict comparison of the performance

of each of these measures of strain fitness in their ability to identify interaction effects was, however, not performed.

2.4.2 Four Parameter Logistic Model

A modest extension of the simple logistic function produces what is generally known as the four parameter logistic function. We introduce a new parameter A in order to take control of the lower asymptote of this function, and wish to keep our upper asymptote locked at B. The exponential growth constant remains unchanged as r. We accomplish this by replacing terms in the simple logistic function as follows: we introduce an additive factor A, the parameter B is replaced with $B - A$, and y_0 is replaced with $y_0 - A$. First, I work out the impact on t_{mid}.

$$t_{mid} = \frac{1}{r}\log\frac{B - y_0}{y_0} \qquad \text{(from the simple logistic function)}$$

$$\tilde{t}_{mid} = \frac{1}{r}\log\frac{B - A - (y_0 - A)}{y_0 - A} \qquad \text{(replacing terms as described earlier)}$$

$$\tilde{t}_{mid} = \frac{1}{r}\log\frac{B - y_0}{y_0 - A}$$

Hence, we obtain an equation for $y(t)$ for the four parameter logistic function as

$$y(t) = A + \frac{B - A}{1 + \exp[r(\tilde{t}_{mid} - t)]}$$

We have the restrictions $A < y(t) < B$, and typically restrict ourselves to $r > 0$ and $t_{mid} > 0$. Once again, $\tilde{t}_{mid}$ is simply a function of other parameters.

Note that the restriction $A < y(t) < B$ also explicitly implies $A < y_0$. Although we hope the values of the two are closely related under certain restrictions for parameters in this model, they are certainly do not represent the same quantity.

We verify that A and B are the lower and upper asymptotes respectively. First, we note

$$\lim_{t \to -\infty} \exp[r(\tilde{t}_{mid} - t)] \to \infty, \qquad \lim_{t \to \infty} \exp[r(\tilde{t}_{mid} - t)] \to 0,$$

and use this fact to verify that

$$\lim_{t\to-\infty} y(t) = A, \qquad \lim_{t\to\infty} y(t) = B.$$

Finally, we note that the function is symmetric about t_{mid}; more specifically, the function $y(t+\tilde{t}_{mid})-(B+A)/2$ is odd:

$$y(t+\tilde{t}_{mid})-\frac{B+A}{2} = \frac{B-A}{1+\exp[-rt]}-\frac{B-A}{2}$$

Note that the value of this function at time $t = 0$ is 0, and, calling the above function $h(t)$, we can easily verify that

$$-h(t) = h(-t)$$

Empirical studies later will show that the lower asymptote A is typically significantly greater than 0, hence providing evidence that including A as a parameter and using the four parameter logistic model, rather than the simpler three parameter logistic model, is necessary. Henceforth in the thesis I will be primarily referencing the four parameter logistic model and will opt to call the parameter $\tilde{t}_{mid}$ simply t_{mid}; however, the reader is reminded that while $\tilde{t}_{mid}$ and t_{mid} are conceptually the same, they have slightly different definitions.

2.5 Strain Fitness and the Four Parameter Logistic Function

Strain fitness can be constructed as different functions of the parameters *A, B, r* and t_{mid}. Recall that, in the context of exponential growth, if y_0 is shared between all the growth curves of interest, then the only possible way that these curves could differ is through their exponential growth constants. This provides a primary motivation for the use of *r* or *d* as a definition of strain fitness. Hence, the canonical way of comparing growth curves when a logistic model is assumed is through comparison of the different exponential growth constants *r*. In such cases it is hoped that fitting a logistic model will improve the precision and reduce any potential biases in estimating *r*, relative to other methods that might rely on exponential or non-parametric models. However, one might be motivated to use t_{mid} or *B* depending on their needs. If someone wished to assign higher fitness scores to strains with a

higher carrying capacity, then one might seek to compare the different *B* estimated. Or, as is often done in the literature, a fitness score can be constructed that gives higher scores to strains that both grow quickly and grow to a higher carrying capacity, and vice versa. (Addinall et al., 2011) The AUC / AULC are measures that fit these criteria, and while they might each be viewed as a function of the parameters fit if the four parameter logistic model is assumed, their utility comes in their ability to remain estimable even under growth curves that depart severely from the exponential or logistic growth models.

With the exponential and logistic growth models now introduced, I next discuss growth curve experiments, the types of growth curves one might observe in these experiments, and the severity of different kinds of departures from exponential and logistic growth models observed in these growth curves.

Chapter 3: Growth Curve Experiments

Before proceeding onto a number of empirical studies exploring the benefits and drawbacks of the various means of defining strain fitness, it is prudent to discuss the structure of a growth curve experiment.

Figure 3.1 A 96-well microtitre plate

In a typical growth curve experiment, yeast cells are grown on $8 \times 12 = 96$ well microtitre plates, with each well in a plate containing yeast cells in some growth medium that have had one or two genes from its genome knocked out. The experiment is typically constructed such that the initial quantity of yeast cells y_0 in each well is the same (though not necessarily known explicitly), and the entire plate is run through a plate reader and measured over a set period of time, typically 24 hours. Over this time period, the level of growth (measured by the optical density of a particular wavelength of light, called the OD reading) in each well is measured at a certain number of equally spaced time points. The plates are run through the plate reader one at a time in each experiment, and depending on the pre-known or pre-assessed average fitness of yeast strains on a particular plate, they may be grown at different

temperatures as to help facilitate the growth of sicker strains. Strains are typically grown in triplicate on a particular plate as to measure a particular strain of yeast's growth variability.

The following tree gives a visual summary of the structure of a growth curve experiment.

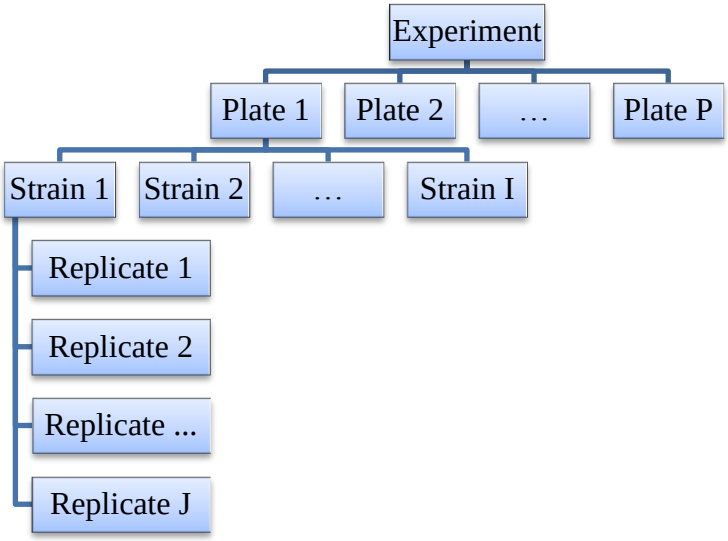

Figure 3.2 Structure of a growth curve experiment

From a growth curve experiment, it is hoped that strain fitness, or some measure of a particular strain of yeast's ability to multiply after the deletion of certain genes or introduction of DNA-damaging media, could be assessed. Through these experiments, we can learn about how certain genes function within a cell to help with cell division, or identify pathways in which multiple genes operate together.

In the absence of growth constraints, we would expect to see exponential growth for each of these strains of yeast – this provides motivation for comparing strains by their exponential growth constants r. (Recall that, under the exponential growth model with the assumption that y_0 is shared by all the growth curves, the only way curves could differ is through r). However, the yeast cells are unable to grow without bound due to space constraints, competition for resources among the yeast cells, and altered kinetics of the cell cycle due to gene deletion. Because of this, the growth curves observed are typically sigmoidal in shape, with earlier stages well approximated by exponential growth. As the population of cells in a particular well increases, the rate of growth will slow until a final carrying capacity is reached, whereby the level of observed growth has stabilized. The carrying capacity is not

entirely determined by external experimental factors, and can vary both from curve to curve and from strain to strain. The sigmoidal shape of growth curves prompts the exploration of sigmoidal models, with primary consideration in this thesis given to the four parameter logistic model.

3.1 Parametric Modeling Choices for Growth Curves

Given the sigmoidal shape of growth curves, it is common to fit a fully parametric model to the observed growth curves (Addinall et al., 2011; Kennedy et al., 2011; McLellan et al., 2009): most commonly the three and four parameter logistic functions are employed, but other functions are available, such the Gompertz or Richards functions. (Kahm, Hasenbrink, Lichtenberg-Frat'e, Ludwig, & Kschischo, 2010) Typically, viable parametric functions have a lower and upper asymptote, and are approximately exponential in the early stages.

The choice of model depends on how one chooses to model the growth constraint. The observed growth curves seen in this thesis will be discussed in terms of their relationship to, and departures from, the four parameter logistic model. However, this thesis is more concerned with estimating strain fitness when faced with vastly non-sigmoidal growth curves, and hence other sigmoidal functions are unexplored.

3.2 Sigmoidal and Exponential Growth Curves

I present some example sigmoidal and exponential growth curves from the Stoepel data set.

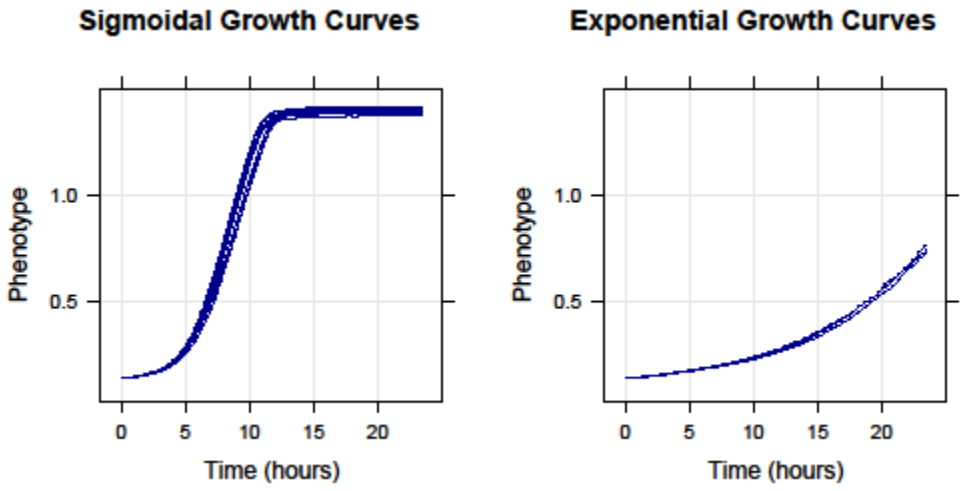

Figure 3.3 Sigmoidal and exponential growth curves

Most parametric models one would use in a growth curve experiment would expect the above forms of growth curves. When faced with a body of growth curves that is entirely composed of exponential and logistic growth curves, then we are very strongly motivated to use *r* or *d* as our definition of strain fitness, and thus the primary concern is how one might best extract *r* from that body of growth curves. When faced entirely with exponential growth curves, one could fit an exponential growth model; when faced entirely with sigmoidal growth curves, one could fit a logistic growth model. More difficult is considering a modeling strategy that can accommodate exponential and sigmoidal growth curves together.

3.2.1 Estimating *r* from a set of Exponential and Sigmoidal Growth Curves

Suppose we have a body of growth curves in which we have a mix of well-behaved exponential and sigmoidal growth curves. Under the exponential or logistic growth models, one might choose $\theta = r$ as their definition of strain fitness. I briefly consider two main strategies for estimating *r* from each growth curve, and will develop them more completely in later chapters.

3.2.1.1 Focusing on Exponential Growth

To estimate *r* using a method predicated on the assumption of exponential growth, one must first isolate the portion of a sigmoidal growth curve that is indistinguishable, in some sense, from exponential growth. This is difficult: if one were interested in isolating the 'exponential' portion of a growth curve and using one of the previously discussed methods to calculate *r*, they may have to first discard the initial censored / sub-exponential observations, but also pick a range of observations small enough such that the growth constraint has not pulled the observed growth too far from exponential growth. In other words, one must consider performing both **left-truncation** and **right-truncation** on the growth curves observed. This presents a catch-22: the region selected should be large enough as to provide a good estimate of *r*; however, the larger the region, the further the potential departures from exponential growth. Typically in the literature such regions have been chosen through ad-hoc methods which can be adequate on a project to project basis. (St Onge et al., 2007)

3.2.1.2 Focusing on Logistic Growth

When we are faced with a mix of exponential and sigmoidal growth curves, the primary problem in fitting the logistic model is that it is now overparametrized for the exponential growth curves. Because the carrying capacity for these exponential growth curves has not been adequately observed in such data, the logistic model will be unable to obtain fitted parameters. The primary solution in such cases is to tether the upper asymptotes of these exponential growth curves to some set of the sigmoidal growth curves in the data set. For example, we might force any exponential growth curves to share a fitted upper asymptote B with the wild type curves, so that the other parameters can become estimable. Once again, this process is not easily automatable, but still workable on a project to project basis.

3.2.2 Departures from the Logistic Growth Model

There are a number of departures from the logistic growth model in these growth curve experiments that we should be aware of when forming an analysis.

3.2.2.1 OD Reader Minimum Read Level Leads to Censored Observations

The true phenotype that might be measured in the earliest stages of growth is censored due to the read level induced by the growth media used. When the yeast cell population in a particular well is very small, the contribution of growth media to the OD read obtained will overshadow the effect of the yeast cells. Hence, over the curve, we observe the maximum of this minimum read level, and the true cell population that might be observed. Furthermore, this implies that y_0 is censored in the data as well. The plot ahead illustrates this, with the measured level of growth plotted in red, and the true level of growth (below the minimum read level) plotted in blue. This censoring of initial observations forces us to be very careful in interpretation of the lower asymptote A, and also notes that any estimate of y_0 obtained from the logistic model will be difficult to interpret.

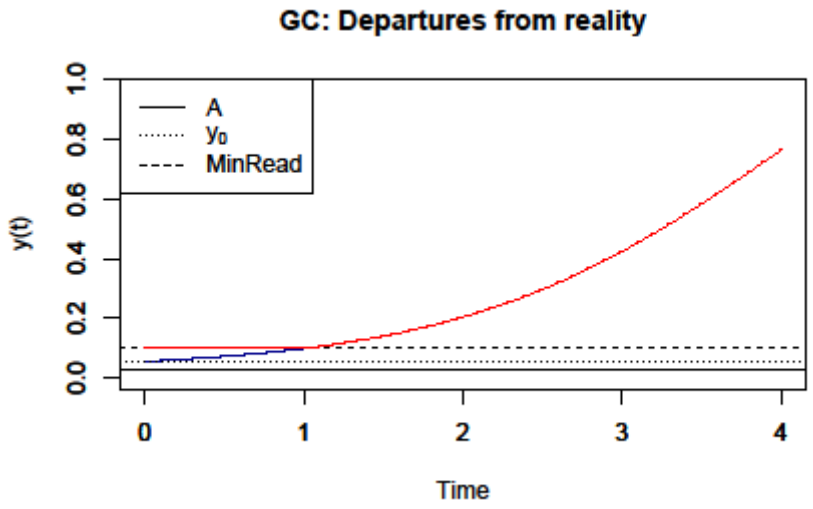

Figure 3.4 Early observations in a growth curve experiment are censored.

3.2.2.2 Action of the Growth Constraint

It is worth stating early that the logistic growth function does not provide a perfect model of the growth constraint. The effect of the growth constraint in observed growth curves is often quite muted in the earlier observed parts of the growth curve, but comes into effect strongly and sharply after a certain point. This is in contrast to the symmetric shape of the logistic growth function, which typically over-estimates the effect of the growth constraint in the earlier parts of an observed growth curve. This will be more fully explored later; however, this systematic departure is still quite small.

An example of this departure is presented with a single wild type curve from the Stoepel data set. I fit the four parameter logistic model by non-linear least squares to this single curve and overlay the fitted values, using the R function `nls`.

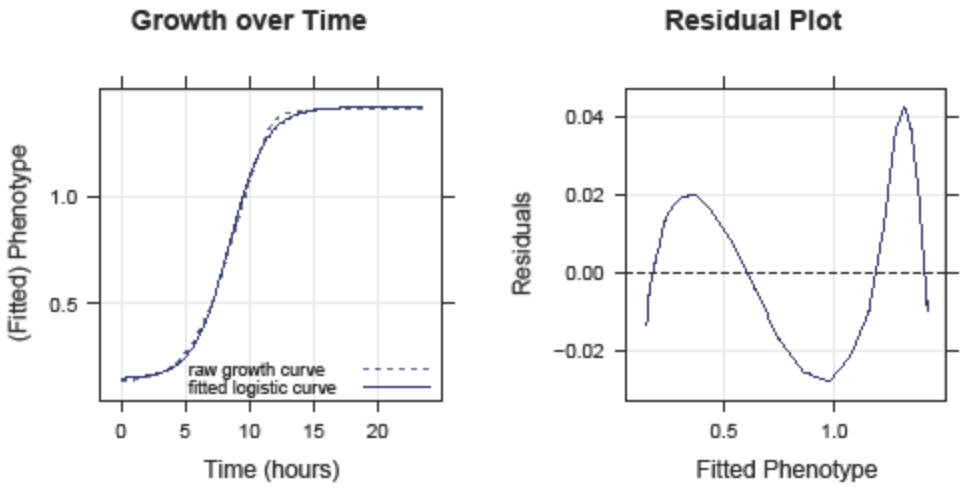

Figure 3.5 Illustrating systematic lack of fit of logistic model

Although the amount of lack of fit is small, we do see a systematic lack of fit in the accompanying residual plot. In particular, we notice that the growth curve levels off much more quickly than the fitted logistic curve, which makes a more gradual transition towards its upper asymptote. This is emphasized by the large peak in the residual plot observed for the larger fitted values.

3.3 Non-Sigmoidal Growth Curves

Depending on the effect of either the gene mutation or the solution in which the yeast is grown, the observed growth curves may not be sigmoidal at all.

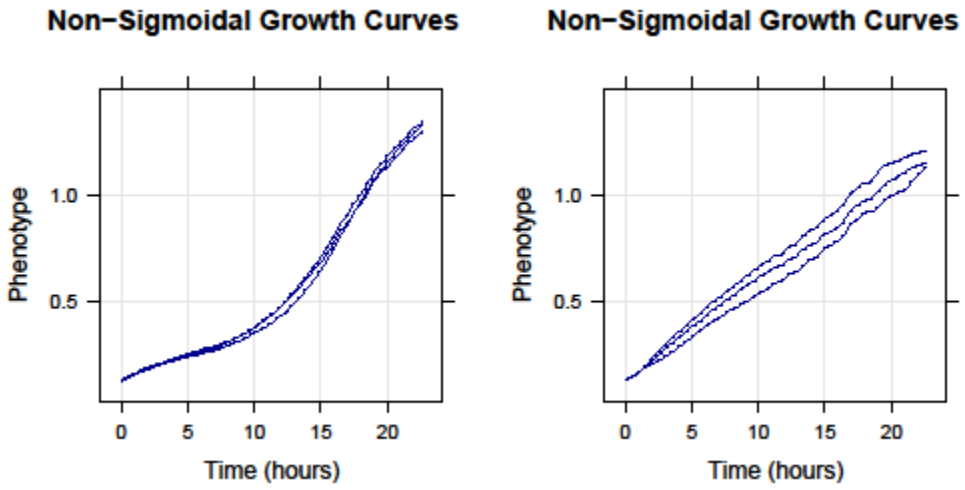

Figure 3.6 Non-sigmoidal growth curves

The two sets of non-sigmoidal growth curves are drawn from the McLellan data set. The curves in the left panel are *smc1, rps31* double mutants, while the curves in the right panel are *scc2, pac10* double mutants.

It is quite apparent that, for growth curves of this shape, neither the exponential nor logistic growth models will be appropriate. Given a large enough data set, we will inevitably encounter these non-exponential, non-sigmoidal growth curves, perhaps in abundance. It is the presence of these kinds of growth curves in a data set that prompt investigation of model-free definitions of strain fitness. That is, for growth curves that deviate systematically from the exponential / logistic growth models, the canonical parameters used in defining strain fitness eg. r are no longer relevant descriptors, and so we might seek a definition of strain fitness that can accommodate both sigmoidal and non-sigmoidal curves together, while remaining a conceptually sensible definition of strain fitness – hence, AUC and AULC.

With the different kinds of growth curves introduced, I now proceed to a set of interaction studies to illustrate the strengths and pitfalls in employment of the logistic model, relative to methodology using AUC or AULC.

Chapter 4: Interaction Analysis Study

This chapter is dedicated to the development of a complete implementation of a growth curve analysis. I will explore different modeling solutions that can lead to estimation of the interaction effects desired through the use of a microcosm of growth curves.

4.1 Outline of Methodology

In a growth curve analysis we seek to use measured growth curve data to identify interaction effects. First, I outline the road map for the analysis of a set of growth curves.

Figure 4.1 Outline of growth curve analysis

For example,

1) The growth model (eg. logistic) would allow us to model growth over time,
2) The normalization model helps to correct for the plate (nuisance) effects, and
3) The interaction model allows us to infer directly on interaction effects; that is, the effect on strain fitness of the deletion of two genes in tandem.

There are two main classes of approaches that I will attempt to develop.

First, I attempt to develop a single model that can encompass all of the three above models entirely within itself. Hence, estimated interaction effects could be collected, alongside standard errors, directly from a single model with normalization built in. This will henceforth be called the single model solution.

Secondly, as we will see in this chapter, I will have to deviate from the single model solution as different computational road blocks make fitting of a single model solution impractical

and unfeasible. The second approach accomplishes each of the above stages in a sequential, step-wise process called a derived variables analysis. (Diggle, Heagerty, Liang, & Zeger, 2002) So, one might consider fitting a growth model, extracting parameter estimates, discarding the standard error associated with these estimates, and then pushing these values themselves through the normalization and interaction models.

Another notion hiding in this road map is the definition of strain fitness. Under a single model approach, the embedment of the normalization and interaction models forces us to restrict our definition of strain fitness to be a single parameter used in that model, eg *r*. However, under a DVA approach, we have considerably more freedom. We can assume a parametric model and define strain fitness to be a (function of) the parameters used: for example, $MDR \times MDP$ as in Addinall et al. Alternatively, we can opt to use a non-parametric definition of strain fitness as well: for example, AUC / AULC.

4.1.1 Single Model Solution

As discussed before, we would ideally want to write down a model which can model growth over time, plate effects, and allow for extraction of interaction effects. I will first write down the 'ideal' model that, in a perfect world, would be fit in such a scenario. Let *i* denote strain, *j* denote culture or replicate, *k* denote time and *p* denote plate. I formulate an 'ideal' model as:

$$y_{ijp}(t_k) = A_{ijp} + \frac{B_{ijp} - A_{ijp}}{1 + \exp(r_{ijp}[t_{mid_{ijp}} - t_k])} + \epsilon_{ijkp}, \qquad \epsilon_{ijkp} \sim^{iid} N(0, \sigma_\epsilon^2)$$

For each parameter $\alpha \in \{A, B, r, t_{mid}\}$ in the model, we might include fixed and random effect terms:

$$\alpha_{ijp} = \mu_i + \beta_{ij} + \gamma_p$$

μ_i is a fixed effect denoting the mean parameter value associated with a strain *i*, β_{ij} is a random effect term accounting for within-strain replication; that is, it allows for variation between replicates to be modeled for, and γ_p is a random effect associated with plate.

We will assume that the random effect terms β_{ij} and γ_p are independent – that is, there is no link between plate-to-plate variability and the variability between replicates.

Next, we could apply contrasts such that the interaction model we use is imbedded in our definition of strain fitness. Recall that, for a query gene *q* and a non-essential gene *g*, the strain fitness can be decomposed as

$$\theta_{q,g} = \theta_{WT} + \tau_q + \tau_g + \tau_{q,g}$$

Hence, we could imbed this interaction model in the model used for each parameter α_{ijp} as (decomposing index *i* into two terms, *q* and *g*) as:

$$\begin{aligned} \alpha_{(q,g)jp} &= \mu_{(q,g)} + \beta_{(q,g)j} + \gamma_p \\ \alpha_{(q,g)jp} &= \left(\theta_{WT} + \tau_q + \tau_g + \tau_{q,g}\right) + \beta_{(q,g)j} + \gamma_p \end{aligned}$$

Once again, inference on the interaction effects $\tau_{q,g}$ is the primary goal.

4.1.2 Derived Variables Analysis

The alternate route towards this final interaction model is through a derived variables analysis (DVA). Rather than attempting to incorporate the growth, normalization and interaction models in one tidy package, we might opt to perform each step sequentially. Hence, I outline my implementations of the three different models that will be used throughout this chapter.

4.1.2.1 The Growth Model

As before, different growth models are available in a DVA, with different models giving different possible definitions of strain fitness. If we specify a growth model explicitly (e.g. the four parameter logistic model), then strain fitness can be defined as some function of *A, B, r* and t_{mid}. Furthermore, we can consider different allocations of fixed and random effects to each of the four parameters. In this regard, I will consider the four parameter logistic function with different allocations of fixed and random effects throughout the chapter. We may also neglect to specify a growth model and simply define strain fitness as some

calculable quantity for each growth curve – eg, AUC or AULC. However, by foregoing a parametric model, we are unable to include fixed and random effects explicitly.

After defining a growth model (or choosing not to), we can define strain fitness. Once again I call strain fitness θ, and consider next how to normalize for plate effects.

4.1.2.2 The Normalization Model

We can exploit the fact that wild type growth curves are grown on each plate in order to enact normalization. We assume that, in the absence of plate effects, the average wild type fitness should not vary much across plates. Hence, we use this variation in average wild type fitness to normalize for plate effects in the following way. Let γ_p be the effect for plate *p*, and let *ref* be the reference plate selected. All the estimates of strain fitness will be normalized relative to this reference plate as

$$\hat{\theta}_{ijp}^{norm} = \hat{\theta}_{ijp} - \hat{\gamma}_p, \qquad \hat{\gamma}_p = avg(\hat{\theta}_{WT}^{ref}) - avg(\hat{\theta}_{WT}^{p})$$

where the average wild type strain fitness measured on plate *p* is denoted as $avg(\hat{\theta}_{WT}^{p})$. After obtaining these normalized fitness measures, we next push them through the interaction model.

4.1.2.3 The Interaction Model

After obtaining the normalized fitness measures, we can perform a 2-way ANOVA analysis through use of the interaction model defined before:

$$\theta_{q,g} = \theta_{WT} + \tau_q + \tau_g + \tau_{q,g} + \epsilon_{q,g}, \qquad \epsilon_{q,g} \sim N(0, \sigma^2)$$

From this model, we can obtain t-statistics for the interaction terms of the form

$$T_{q,g} = \frac{\hat{\theta}}{SE(\hat{\theta})}$$

and with these we can then begin assessing the interaction effects for statistical significance, rank them, and so forth.

4.1.2.4 Motivating a Derived Variables Analysis

With the DVA approach I use now introduced, I will now motivate why it should still prove a useful means of performing a growth curve analysis.

In asking whether or not a DVA approach is appropriate, we are essentially asking whether the ANOVA model we use as our interaction model is appropriate. Let's outline the main assumptions. First, recall our interaction model:

$$\theta_{q,g} = \theta_{WT} + \tau_q + \tau_g + \tau_{q,g} + \epsilon_{q,g}, \qquad \epsilon_{q,g} \sim N(0, \sigma^2)$$

Hence, I outline the main assumptions:

1) Normality of residuals. While the normality assumption is perhaps overly optimistic, it is still more defensible than the normality of random effects terms in a single model approach. The distribution of residuals will be briefly explored later for different methods.

2) Independence. We should be fairly comfortable in assuming independence, since the growth of one particular replicate should not give us information about another replicate.

3) Homogeneity of variance. That is, the variation we see in replicates does not depend on strain; this variance is common to each strain.

Finally, Diggle et al. note two major requirements for a derived variables analysis to be valid, both which are related to the homogeneity of variance assumption: there should be no missing values, and there should be an equal number of time point measures for each growth curve. (Diggle et al., 2002) In a growth curve experiment, missing values are typically very rare, as the entire measurement process is automated and well-controlled. However, methods requiring truncation would produce growth curves measured over a varying amount of time points, and in doing so may violate the equal variance assumption made in ANOVA.

4.1.3 Comparing Normalization Models for each Class of Methods

A derived variable analysis gives us some more control over the normalization model, relative to a single model based solution. With DVA, normalization can be performed after the naïve strain fitness estimates are calculated by exploiting the fact that wild type curves are grown on each plate. We use the fact that, in the absence of plate effects, average wild type fitness should not vary much; hence, we use the variation observed between plates in wild type fitness to characterize the plate effects.

For the single model solution, I explicitly consider a random plate effect. By attempting to include a plate effect in our single model solution, we are implicitly assuming that the expected average fitness of curves is equal across plates. That is, the normalization performed on plate effects when done within a single model is done to equalize the average strain fitness estimate across plates, under the assumption that these differences in average fitness across plates is due entirely to the plates. However, if the average fitness on a particular plate was particularly low due directly to the sicker strains grown on that plate, the embedded normalization model would incorrectly adjust strain fitness up for each strain.

4.2 Empirical Study with a Microcosm of Growth Curves

Recall that the growth curves measured in the Stoepel data set have the canonical sigmoidal form that supports the use of exponential or logistic growth models in determining strain fitness. Hence, these growth curves will provide a good starting point for exploration of the different ways to perform an interaction analysis. To begin, I select a small set of growth curves that are quite typical of other single- and double-mutant strains obtained in this data set, and of other growth curves seen in the literature.

Throughout the thesis, R and an accompanying package `nlme` is used for model fitting and parameter estimation. (Pinheiro, Bates, DebRoy, Sarkar, & Team, 2011)

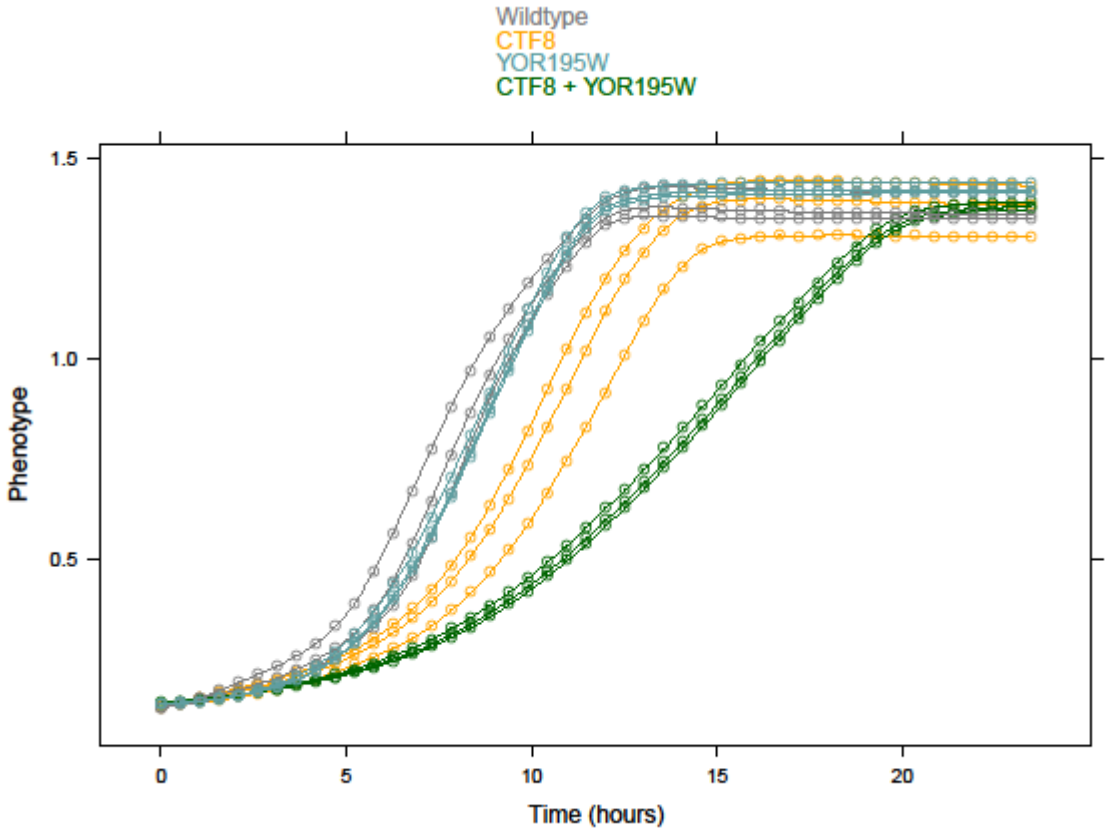

Figure 4.2 A microcosm of growth curves

I have selected three replicates each of wild type, CTF8 single mutant, YOR195W single mutant and CTF8 + YOR195W double mutant from a single plate, for a total of twelve growth curves from this plate. The three growth curves observed for the double mutant strain, CTF8 + YOR195W, appear noticeably 'sicker' than the other growth curves assessed. Because the observed strain fitness of the double mutant seems to be, in some sense, less than the 'combined' strain fitness of the other single mutants, we would likely call this a synthetic interaction. The variation observed between strains seems large relative to the variation between replicates for a given strain. Because these growth curves are collected only over a single plate, normalization is unnecessary and is left unconsidered throughout this section.

First, I introduce the contenders from each class of models, and then evaluate their performance over some different criteria.

4.2.1 Single Model Solutions

I will consider two models in particular which can encompass the entirety of the growth and interaction model. I will use the four parameter logistic model specifically, and seek to develop two different models. I will use *r* as my definition of strain fitness. First, I present the four parameter logistic model once again:

$$y_{ijp}(t_k) = A_{ijp} + \frac{B_{ijp} - A_{ijp}}{1 + \exp(r_{ijp}[t_{mid_{ijp}} - t_k])} + \epsilon_{ijkp}, \qquad \epsilon_{ijkp} \sim^{iid} N(0, \sigma_\epsilon^2)$$

I consider two different models for the parameters as:

1) $\alpha_{ij} = \mu_i$ (fixed strain effects only, **FE**)

2) $\alpha_{ij} = \mu_i + \beta_{ij}$ (fixed strain effects and random well effects, **FE+RE**)

The interaction model will be embedded in the growth model, and hence all forms of error present will be pulled through the interaction analysis.

4.2.2 Derived Variables Analysis Solutions

In the DVA modeling solutions I propose, I explicitly keep the growth and interaction models separate. Hence, I will consider the same growth models as described before, in addition to a model-free solution in which strain fitness is defined as AUC and AULC. I encapsulate the set of methods I will assess under this DVA approach.

Growth Model	Normalization Model	Interaction Model
Logistic function, FE + RE $\theta_{ij} = r_{ij} = \mu_i + \beta_{ij}$	$\hat{\theta}_{ijp}^{norm} = \hat{\theta}_{ijp} - \hat{\gamma}_p,$ $\hat{\gamma}_p = avg(\hat{\theta}_{WT}^{ref}) - avg(\hat{\theta}_{WT}^{p})$	$\theta_{q,g} = \theta_{WT} + \tau_q + \tau_g + \tau_{q,g} + \epsilon_{q,g}$ $\epsilon_{q,g} \sim^{iid} N(0, \sigma^2)$
AUC (No growth model assumed) $\theta_{ij} = AUC = \int_0^{t_{\max}} y_{ij}(t)dt$		
AULC (No growth model assumed) $\theta_{ij} = AULC = \int_0^{t_{\max}} \log y_{ij}(t)dt$		

4.2.3 Results

It should be noted that, although suppressed in this thesis, there is still substantial difficulty in successfully fitting the single model based solutions. The R function `nlme` is very sensitive to the start values chosen when fitting these models, and I found the default starting values chosen inadequate to fitting the models. Regardless, with some hand-holding I am able to coerce the models to successfully fit the data.

First, I present fitted and residual plots from the logistic growth models used, to help give the reader an idea of how well these two models can fit this small set of data. The fitted growth curves from the model are plotted in bold colour, while the raw growth curves are underlain in the same colour, but with dashed lines.

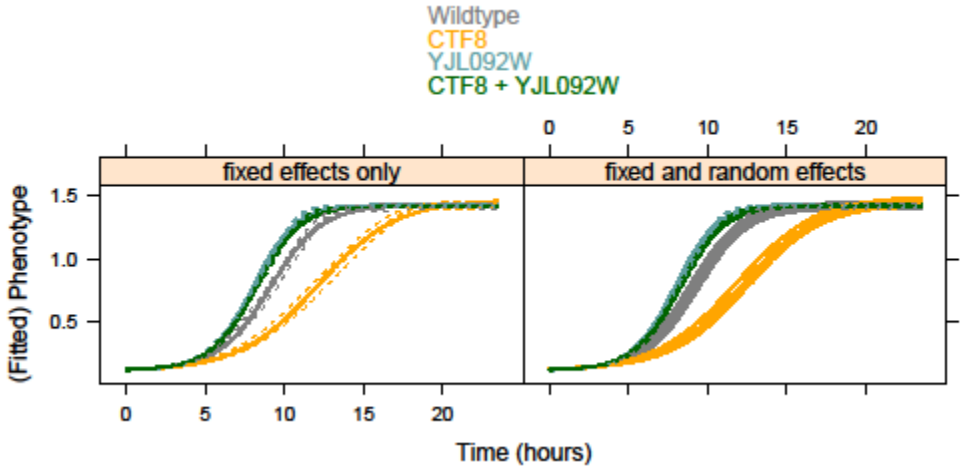

Figure 4.3 Plots of fitted values over raw growth curves

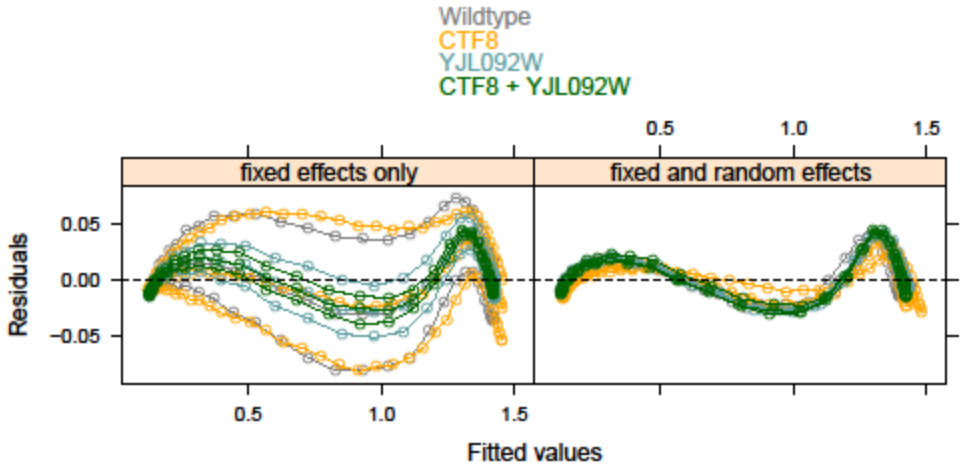

Figure 4.4 Residual plots

There are two main things of interest to note from these plots. First, the inclusion of random effects allows us to model growth curves directly at the replicate level, and we see a substantial improvement in the fit via the side-by-side residual plots. However, we do note that there is still a small amount of systematic deviation from the logistic model, suggesting that one might consider exploration of different growth models if there were concerned about this departure. We do note that the amount of random error is nearly non-existent – almost all of the error left by the fixed + random effects model is systematic.

Note that I choose to leave out DVA FE from this point on, as strain fitness is not calculated at the replicate level under this model. That is, all 3 replicates for a particular strain would be assigned the same strain fitness value. Hence, I present next t-statistics associated with the interaction effect $\tau_{q,g}$ for this microcosm.

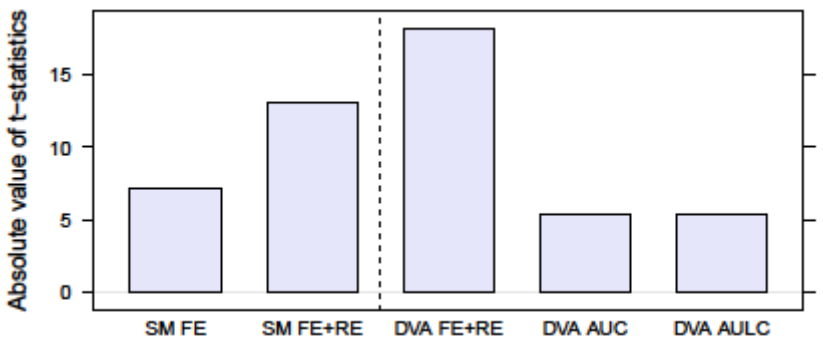

Figure 4.5 Barchart of t-statistics comparing single model, DVA approaches

	Single Model		Derived Variables Analysis		
	FE	**FE+RE**	**FE+RE**	**AUC**	**AULC**
t-statistic	-7.18	-13.14	-18.21	-5.41	-5.34

Table 4.1 Statistical significance by modeling approach for microcosm of curves

It seems that assuming a model when the growth curves actually do present sigmoidal growth greatly improves the statistical significance assessed. Comparing FE+RE approaches, we see that a DVA affords a higher t-statistic and hence more statistical significance than the SM approach. Although we do see t-statistics relatively high in magnitude for the DVA AUC / AULC approaches (that is, high enough to be statistical significant at most sensible cut offs), they are quite a bit smaller than those produced under the logistic growth model in each scenario. This small study suggests that, when faced with sigmoidal growth curves, we may be well served to take the extra effort in fitting a parametric model.

4.3 Extending the Study – Other Microcosms of Curves

I have picked a fairly 'ideal' microcosm of growth curves from the Stoepel data set, which is essentially made up of the kinds of growth curves that the logistic model would expect. However, it will be interesting to see how much model fitting success I have in each scenario, when I consider all possible microcosms of the form produced earlier. To this end, I cycle through all the double mutants on all plates in the Stoepel data set to produce sets of data similar to seen before – 3 wild type curves, 3 curves from each of the 2 single mutants,

and 3 of the double mutant curves. In total, I collect 59 of these possible data sets, and attempt model fitting under each of the previously described approaches.

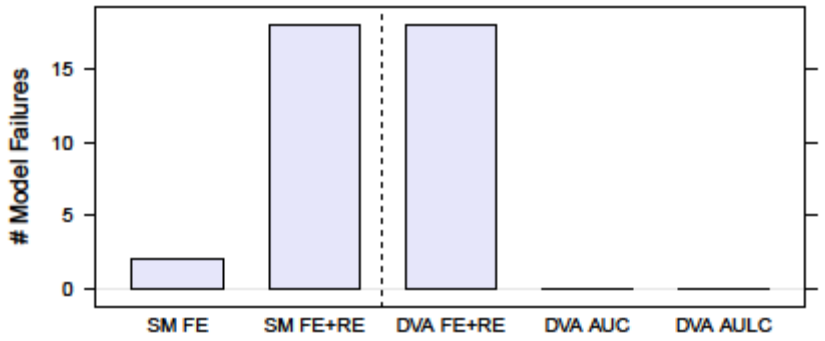

Figure 4.6 Barchart – # model fit failures by method

This represents the beginnings of computational difficulties one faces when using the four parameter logistic growth model. Even for these small sets of growth curves that follow the exponential / logistic models fairly closely, we begin to see a large number of model fit failures. This is due to a combination of `nlme`'s sensitivity to initial starting values, plus the presence of exponential growth curves (ie, those growth curves for which the carrying capacity is unobserved) causing our chosen model to become overparametrized. Of course, the computational difficulties in a DVA approach are essentially nil, so we observe no failure under the AUC, AULC method.

I next plot the t-statistics calculated under the AUC method, with different plotting symbols used depending on whether the SM FE+RE model was successfully fit or not.

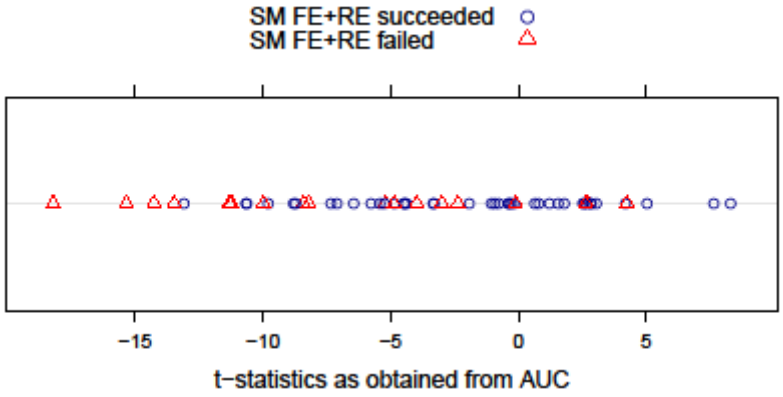

Figure 4.7 Success rate of single model solution

What we see here is fairly alarming. The gene pairs where we see the largest interaction effects are those where the logistic model is failing – that is, the interaction effects that are of greatest interest to us are the ones causing the logistic model the most trouble. Furthermore, this is with a small set of fairly well-behaved growth curves. We could imagine that scaling up this model to span across plates would be virtually impossible, given that we are seeing failure at such early stages. This reflects my experience in attempting to fit such models to entire sets of growth data in R.

While I have had success in fitting the logistic growth model to an entire plate's worth of growth curves, it was not without great pains in both manually selecting suitable starting values, as well as tethering of the upper asymptote of sicker strains to the wild types observed on that plate – a task that requires manual intervention and is impractical to automate. Hence, we are motivated by practicality to employ a DVA approach even when faced with exponential and logistic growth curves.

In the next chapter, I leave this microcosm of growth curves, and enact different DVA analyses on entire sets of growth curve data. There, I will further evaluate the different methodologies available to us in analysis of a growth curve data set.

Chapter 5: Large-Scale Empirical Studies

This chapter is dedicated to comparing each the previously described methods of estimating and testing interaction effects over entire sets of growth curve data. There are two sets of growth curve data that I will be assessing in this chapter: the well-behaved Stoepel set of growth curves, in which all curves are either logistic or exponential in form, and the more difficult McLellan data set, where we observe a heterogeneous mix of exponential, sigmoidal, and non-sigmoidal growth curves. It should be noted that, for each of these data sets, the gene pairs assessed have been selected based on prior expectation that we should observe significant interaction effects. Hence, the other genes *g* are not randomly sampled from the genome, and we are not attempting to 'find needles in a haystack' in terms of interacting gene pairs.

There are two main criteria on which I will assess these methods: the number of statistically significant interaction effects found at some pre-defined cutoff, and the overall concordance between each method. Recalling the discussion in Chapter 4, all of the methodology seen here will be done in the context of a derived variables analysis.

5.1 Introduction to Data

First, I outline the structure of both the Stoepel and McLellan data sets. Although I have outlined different qualities of each of these data sets briefly throughout the thesis, I will state explicitly the structure of each data set now.

5.1.1 Stoepel Growth Curves

The goal in this experiment was to identify synthetic sick interactions between the query gene *ctf8* and 45 other non-essential genes. Strains of yeast were grown on 96-well plates, one plate per day, over 10 different days. Each plate was grown at 26^{o}C over a time period of 24 hours, and a total of 46 observations are made over each well. Three replicates were included for each strain analyzed, with higher levels of replication for wild type and *ctf8* single mutants. The growth curves in this experiment are all sigmoidal or exponential in shape, with those exponential curves being the sicker strains that did not have enough time to reach carrying capacity.

5.1.2 McLellan Growth Curves

The goal in this experiment was to identify synthetic sick interactions between three cohesion query genes (*scc1-73, smc1-259, scc2-4*), and 28 non-essential genes. Strains of yeast were grown for 24 hours on 11 different plates, each run on a separate day. Of these 11 plates, five were grown at a temperature of 26^{o}C, and six at 30^{o}C. On each plate there were fifteen replicate wells for the wild type strain and three replicate wells for each of the other strains analyzed. A total of 45 observations were taken over equally-spaced time points. In this experiment, we observe a heterogeneous mix of sigmoidal and non-sigmoidal growth curves, which prompts my exploration of AUC as a definition of strain fitness.

The effect of temperature on growth is very difficult to model as its effect seems to depend on the sickness of the curve assessed. Because of this, two parallel analyses will be run over each temperature, and interpretation of results is restricted to that particular temperature.

5.2 Introduction to Methodology

First, I introduce the methods I plan on using for performing the interaction analyses, as well as my implementations. Recall that any method of estimating *r* based on exponential growth will require identifying the 'exponential' part of a growth curve. To this end, I use different heuristic strategies accompanying each method to perform this identification. Concordance of methods provides some evidence that they are performing equally well; however, we may see different estimation methods identify more significant interaction effects than others.

5.2.1 Single Time Point

Recall that the single time point measure of *r* required knowledge of y_0 to compute. In these experiments, this quantity is unknown and difficult to estimate, so I opt to implement a similar single time point approach by defining fitness for a particular growth curve as

$$\theta_i = \frac{1}{t^\star} \log y(t^\star)$$

Leaving y_0 out of the expression now makes θ_i a bad estimator of r_i; however, we note that for two populations, the difference can be written as

$$r_i - r_j = \frac{1}{t^\star}\left(\log y_i(t^\star) - \log y_j(t^\star)\right) + \frac{1}{t^\star}\left(\log y_0 - \log y_0\right) = \frac{1}{t^\star}\left(\log y_i(t^\star) - \log y_j(t^\star)\right)$$

That is, because y_0 is a quantity shared between the two growth curves, it falls out of the above expression – and by the structure of a growth experiment, this should hold for all pairs of growth curves. So, even if our definition of strain fitness does not capture *r*, it will still capture interaction effects of interest on the difference in *r*. The next question, then, is how $t^\star$ might be chosen.

5.2.1.1 Choosing a Suitable Time Point

As a means of side-stepping truncation algorithms, I choose the time point for which the variance in growth measured at that point, across all curves and plates in the experiment, is maximized:

$$t^\star = \arg\max_{t_k} Var(y(t_k))$$

This ensures that I choose a time point such that the growth curves are, at least according to this criterion, maximally distinguishable from one another. The faster growing wild type growth curves have had time to separate from the slower growing sick strains. Although I am not guaranteed to identify a point lying in the exponential-looking part of a growth curve, in practice this algorithm works well.

5.2.2 Double Time Point

The double time point method of calculating *r* is very dependent on the assumption of exponential growth. I opt to choose the time points based on a heuristic method. I will choose the first time point to be the one for which 20% of growth has been reached on average, and the second time point to be the one for which 40% of growth has been reached on average for all growth curves in the experiment. I call these lower and upper time points t_L and t_U respectively. These time points are chosen to reflect an example truncation one would wish to perform in discarding initial sub-exponential observations, as well as later non-exponential observations. Similar truncation rules are enacted in the literature, typically according to how many generations of exponential growth one sees in the experiment. (St Onge et al., 2007)

After truncation is performed, *r* can be computed according to the expression

$$r = \frac{1}{\Delta} \log \frac{y(t_U)}{y(t_L)}$$

where $\Delta = t_U - t_L$. This chosen time is common to each growth curve, and hence differences are discovered based on the differences in phenotype measured at these two time points.

5.2.3 Fitting a Line

In attempting to determine *r* using the linear model described previously, we are required to consider both left- and right-truncation of growth curves again. I opt to use the same truncation strategy as defined before, with all time points within that region used in the model fit. Hence, a linear model of the following form is fit:

$$\log y(t_k) = \log y_0 + r t_k + \epsilon_k, \qquad \epsilon_k \sim^{iid} N(0, \sigma^2),$$

where $t_k \in \{t_L, t_{L+1}, ..., t_{U-1}, t_U\}$.

5.2.4 Four Parameter Logistic Model

The four parameter logistic model is used only for fitting of growth curves from the Stoepel data set, as fitting the logistic growth model to the McLellan growth curves is both unfeasible and undesirable due to the presence of non-sigmoidal growth curves. Hence, the following discussion is given relative to the Stoepel data set.

Due to computational difficulties, I was not able to fit one full model for the entire set of growth curves. Rather, the four parameter logistic model is fit on a plate-by-plate basis, with a derived variables analysis planned for the estimates of *r* obtained. Hence, the following model is fit plate-by-plate, once again letting *i* denote strain, *j* denote replicate and *k* denote time:

$$y_{ij}(t_k) = A + \frac{B_{ij} - A}{1 + \exp[r_{ij}(t_{mid_{ij}} - t_k)]}$$

Fixed effects are used as follows:

1) *A:* A universal parameter estimate of A is assumed; hence, all growth curves on a particular plate contribute to one global estimate of A. This is done in the spirit of the fact that y_0 is shared between all growth curves, even if it is censored.

2) B: Strain-specific fixed effects are used for the carrying capacity B, such that each strain can obtain its own fixed-effect estimate of B. Four of the non-essential genes assessed (YOL138C, YJL127C, YDR161W, YEL061C) required rescuing. That is, because the upper asymptote B was not sufficiently observed for the growth curves belonging to these strains, I force these curves to share their upper asymptote with the wild type curves on that particular plate so that the model remains identifiable.

3) t_{mid}: Strain-specific fixed effects on t_{mid} are included in the model.

4) r: Strain-specific fixed effects on r are included in the model.

Finally, random effects are included for both of r and t_{mid}; although there was some desire to include random effects for B as well, I was unable to successfully fit the model when augmenting it in that way.

The model is fit in R via the `nlme` function for each plate, with manually selected starting values chosen so that the model can be successfully fit.

5.2.5 AUC and AULC

The AUC and AULC are computed curve-by-curve using composite Simpson's rule, as previously described. No data manipulation or truncation is required for their computation. It is worth noting that there are no ad-hoc decisions to be made in the computation of AUC or AULC.

5.3 Normalization and Interaction Models

Having already introduced the normalization and interaction models, I will be brief here. The normalization model exploits the fact that wild type growth curves are grown over each of the plates in a data set. We normalize the strain fitness θ_{ijp} on plate p according to reference plate *ref* through the model:

$$\hat{\theta}_{ijp}^{norm} = \hat{\theta}_{ijp} - \hat{\gamma}_p, \qquad \hat{\gamma}_p = avg(\hat{\theta}_{WT}^{ref}) - avg(\hat{\theta}_{WT}^{p})$$

Once again, $avg(\hat{\theta}_{WT}^{p})$ represents the mean estimated strain fitness for wild type curves on plate p. Post-normalization estimates of strain fitness are then brought forward in an interaction analysis by fitting the ANOVA model:

$$\theta_{q,g} = \theta_{WT} + \tau_q + \tau_g + \tau_{q,g} + \epsilon_{q,g}, \qquad \epsilon_{q,g} \sim^{iid} N(0, \sigma^2)$$

Once again, inference on the interaction effects $\tau_{q,g}$ is of primary interest. In particular, I seek to identify interaction effects that are significantly less than zero.

5.4 Evaluation of Different Methods of Assessing Interaction Effects

I now apply each of the previously described methods of identifying interaction effects to each data set, as done through a derived variables analysis. Throughout figures in this section, I will use the short-hand "STP" for single time point, "DTP" for double time point, "Line" for the linear model fit to truncated and logged curves, "FPL" for the four parameter logistic model, "AUC" for area under the curve, and "AULC" for area under the logged curve. STP, DTP, Line and FPL will all use *r* as a definition of strain fitness, while AUC and AULC methods use AUC and AULC as definitions of strain fitness respectively.

Because we have different definitions of strain fitness (namely, *r*, AUC and AULC), it makes most sense to consider comparison of standardized interaction effects, rather than raw estimates. Hence, throughout the thesis I will consider comparison of the different t-statistics associated with each interaction effect from the ANOVA model.

5.4.1 Stoepel

Because all of the growth curves in the Stoepel data set are fairly sigmoidal in shape, we should expect a high degree of concordance between each method of estimating strain fitness. Methods relying on exponential growth should see approximate exponential growth in the regions they are calculated, and the four parameter logistic model is coercible, with some hand-holding, to the exponential and sigmoidal growth curves seen. The AUC and

AULC, being model-free definitions of strain fitness, see no computational hiccups in their implementation.

First, I plot a scatterplot matrix of the standardized interaction effects (t-statistics) produced for each interaction effect by each method. Pearson correlations (P) and Spearman correlations (S) are computed and placed in the top-left corner of each panel, to help assess the strength of the linear relationship between any two methods. A 45^{o} line is plotted in each panel to further guide the reader in assessing the relative concordance between each method.

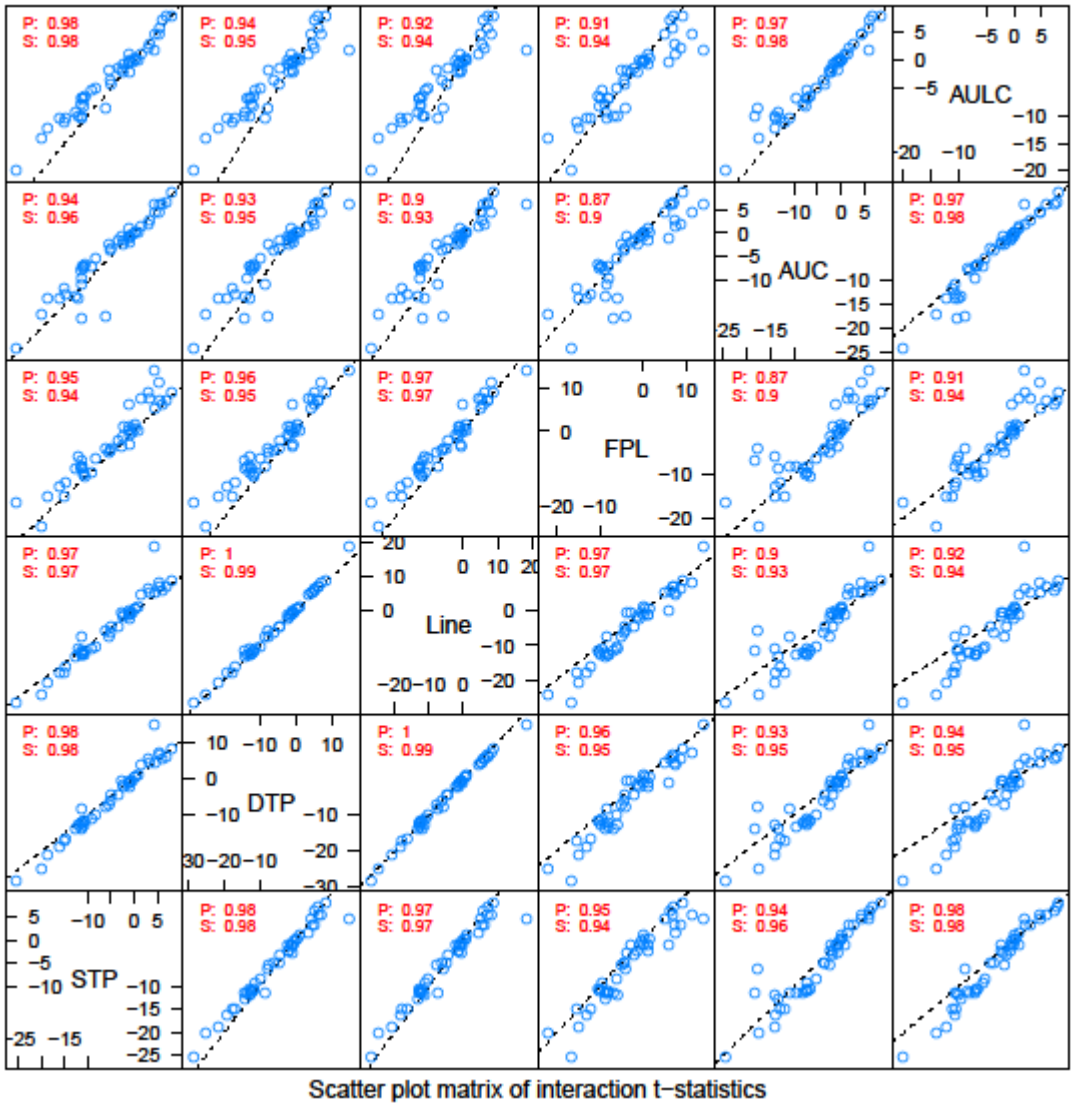

Figure 5.1 Stoepel – scatterplot matrix of t-statistics

The level of concordance here is very high across all methods; however, from the plot we cannot ascertain whether or not one method is performing better relative to the others. Regardless, the degree of agreement here is quite large. Next, a dot plot of t-statistics is presented, with significant interaction effects plotted in orange, relative to a Bonferroni corrected 0.05 cut-off. Dotted lines are drawn between each pair of genes to help visualize the degree of concordance between methods. The number of significant interaction effects is presented for each method as well.

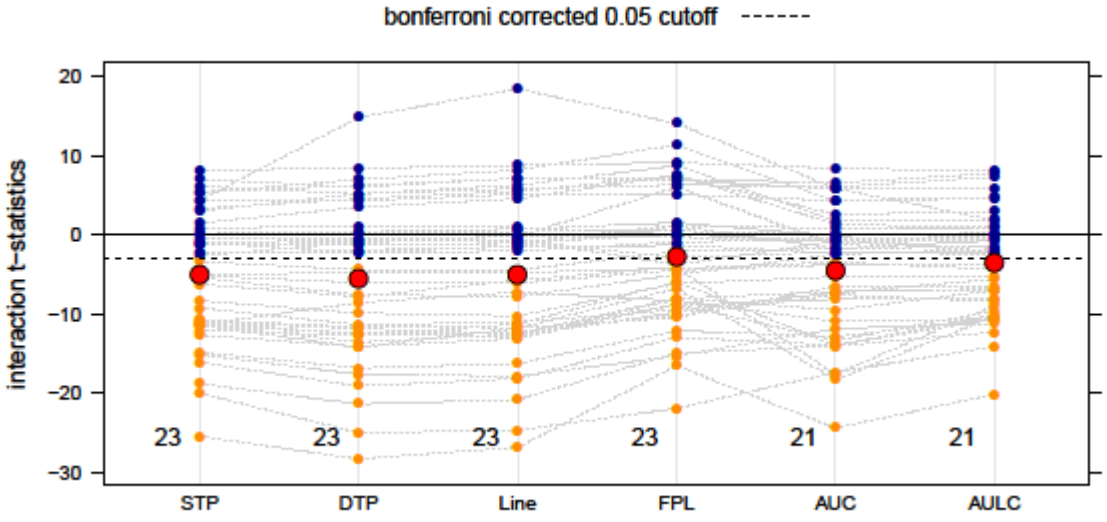

Figure 5.2 Stoepel – dot plot of t-statistics

The dot plot of t-statistics tells the same story: with this set of growth curves, in regards to statistical significance, the performance of each method seems to be roughly the same. There is a bit of variation from method to method; however, the overall variability is similar among methods. Interestingly, we observe the same number of significant interaction effects for each of the four methods predicated on exponential growth (23), while we lose two of those interactions under the AUC and AULC methods (21). It is worth noting that the list of significant interaction effects from each of STP, DTP, Line and FPL are entirely the same in this situation, despite minor shuffling in ranks throughout.

A residual plot for each method is used to assess the ANOVA assumptions. A smoother is overlain in red to assess for any trend in residuals. Although there seems to be a small amount of heteroscedasticity in the residuals, there is certainly nothing to be alarmed about.

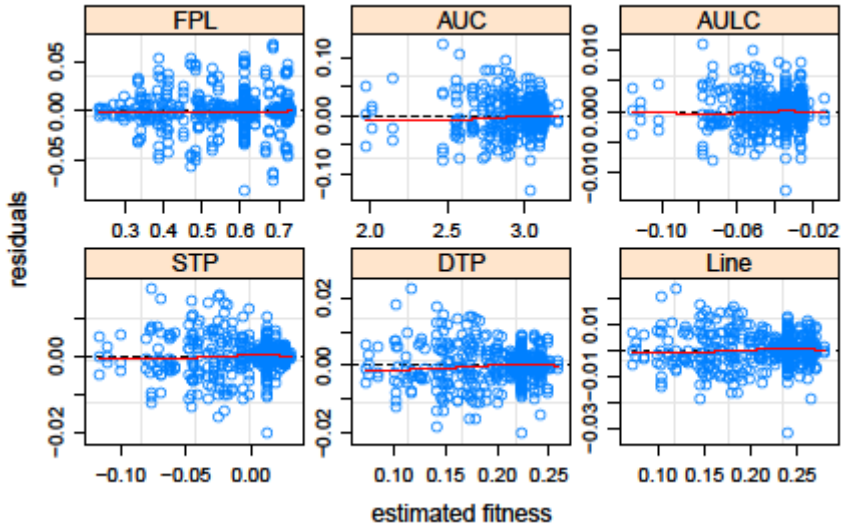

Figure 5.3 Stoepel – residual plots

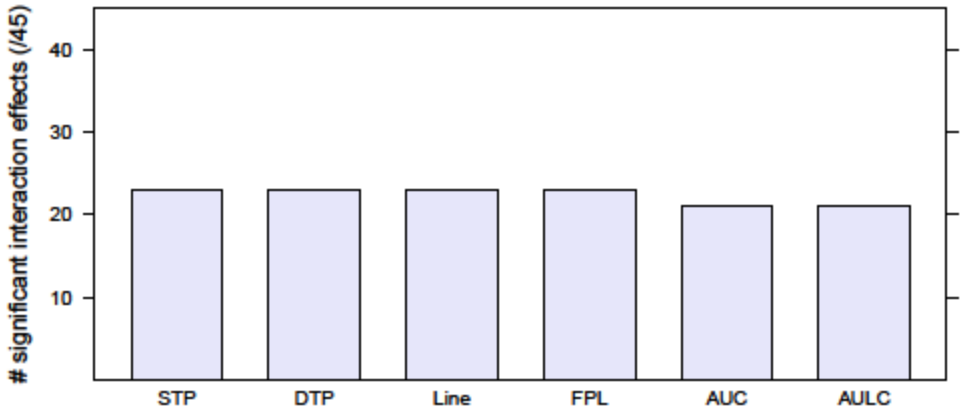

Figure 5.4 Stoepel – barchart of # significant interaction effects by method

As a curiosity, there seems to be three significant interaction effects as assessed by the AUC which disagree with the other four (exponential) methods of estimating strain fitness.

5.4.1.1 Which are the Mutants Lacking Concordance?

Looking at AUC vs. FPL, the least amount of concordance between methods is seen between mutants with YJL127C, YML094W and YDR359C knocked out. These are all genes where the double mutant is highly sick, and the carrying capacity is not, if barely, observed. AUC has ranked these interaction effects much higher than FPL has. It seems that the AUC is affording more statistical significance, or lower fitness estimates, to sicker mutants; especially those where both the single- and double-mutants appear sick, relative to wild type and *ctf8*.

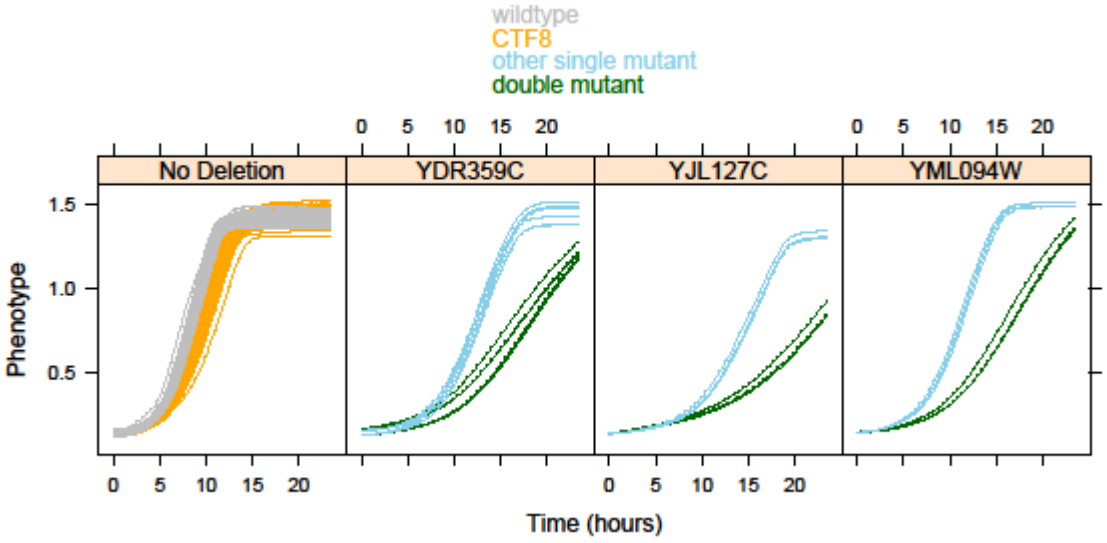

Figure 5.5 Stoepel – genes ranked differently by AUC, FPL

	YDR359C	YJL127C	YML094W
t-stat rank by AUC	2	3	6
t-stat rank by FPL	16	22	18

Table 5.1 Stoepel – genes ranked differently by AUC, FPL

5.4.2 McLellan

The McLellan data sets are different from the Stoepel data set in that the growth curves assessed now depart from sigmoidal growth enough that fitting of the four parameter logistic model is no longer possible. The other methods described in estimating *r* and hence strain

fitness should still prove workable; however, with the assumption of exponential / logistic growth now not met over the entire body of growth curves, we are not so sure whether those methods will still produce sensible estimates of strain fitness, or remain capable of capturing a large number of interaction effects.

First, I present scatterplot matrices of the t-statistics. I first present the t-statistics obtained from fitting the plates run at 26^{o}C.

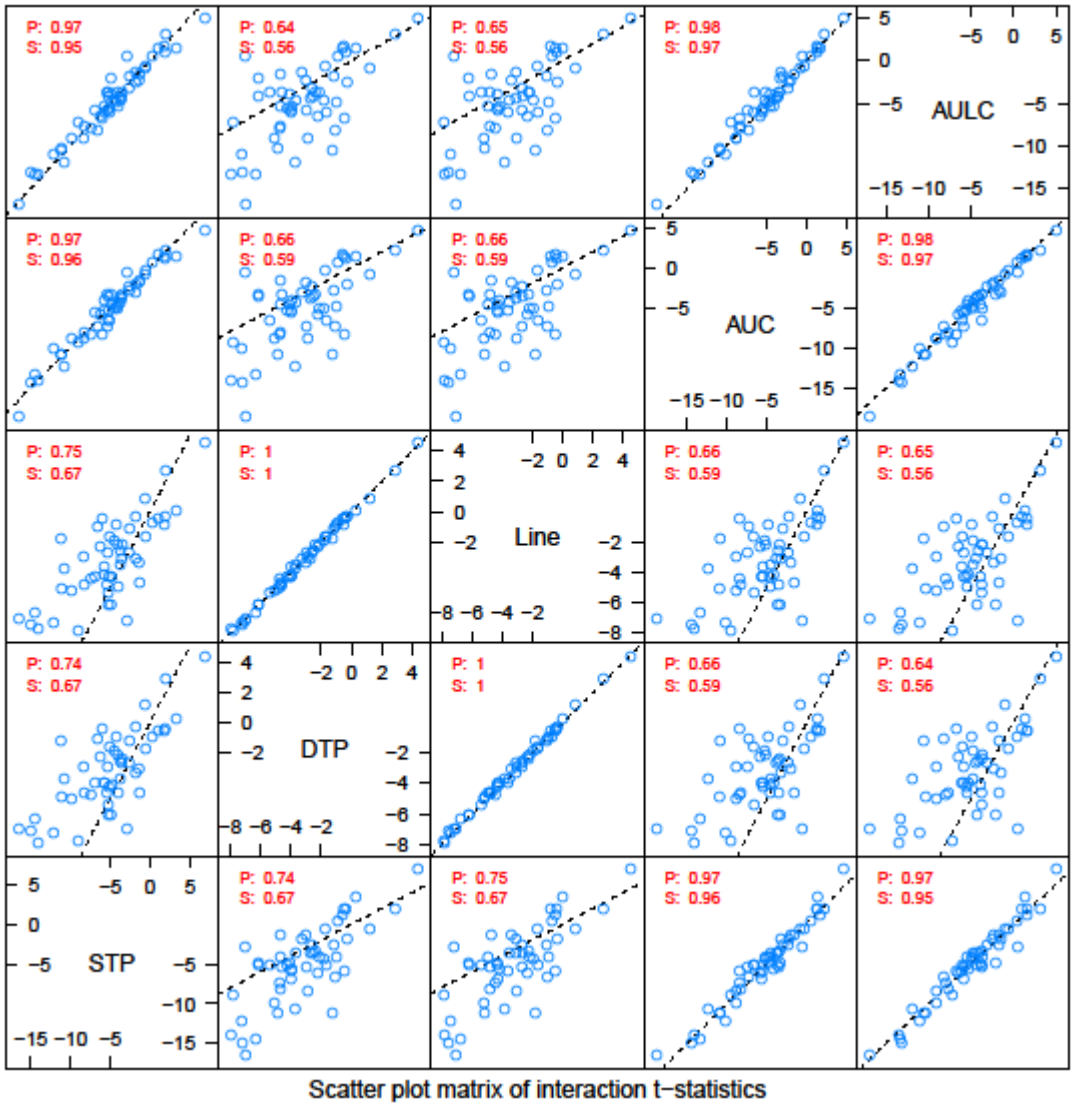

Figure 5.6 McLellan 26°C – scatterplot matrix of t-statistics

Interestingly, we see a large amount of concordance between STP, AUC and AULC methods, while we see a second class of concordance between the DTP and Line methods. However, the amount of agreement between these two groups is quite small.

Next, I present a scatterplot matrix of t-statistics for the data run at 30°C.

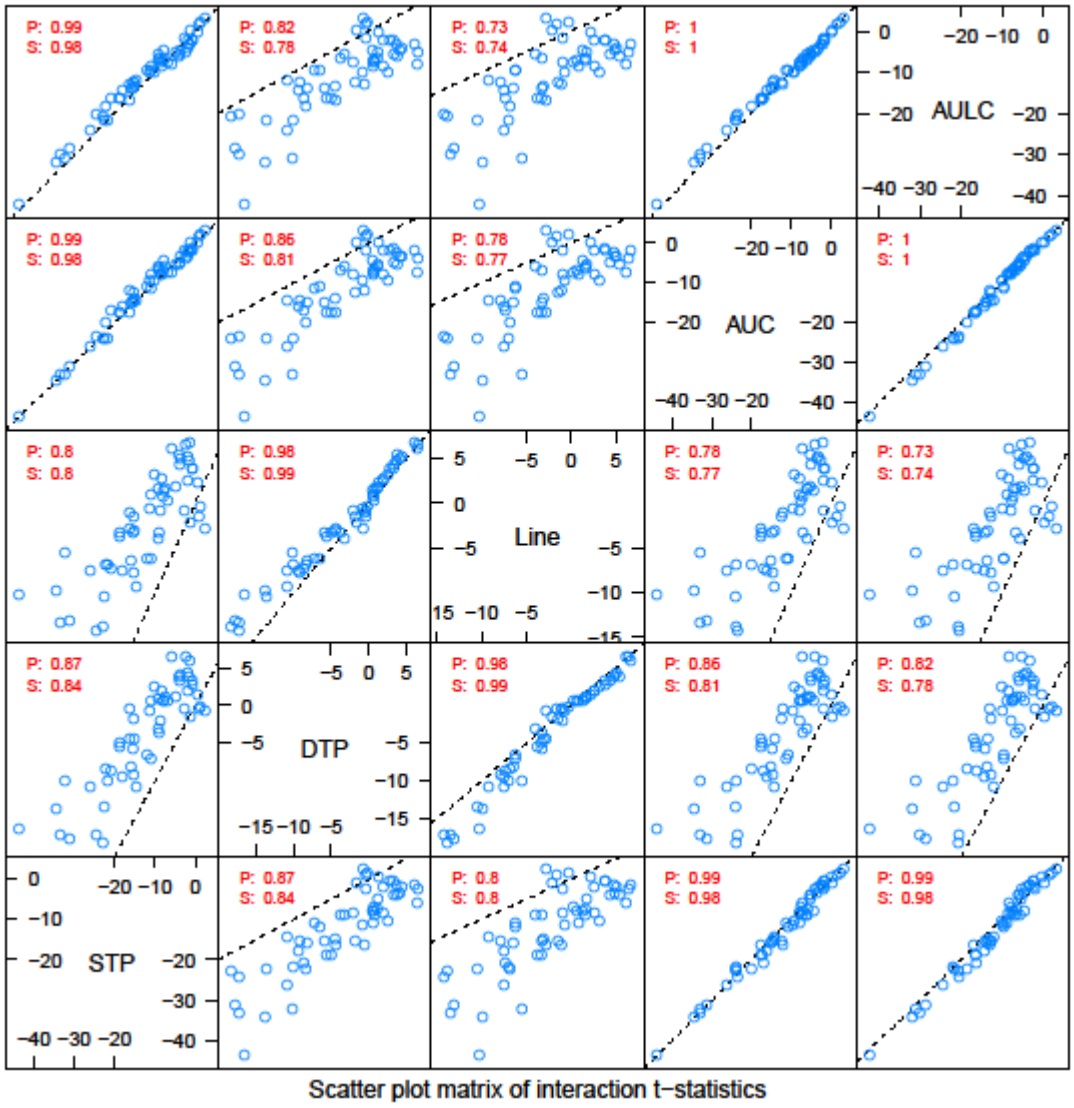

Figure 5.7 McLellan 30°C – scatterplot matrix of t-statistics

We see the same story once again for this set of growth curves. It seems that DTP and Line are over-invoking the exponential assumption, which draws their estimates of strain fitness away from the other methods in a systematic way.

With the two concordance classes fairly well established from the figure, I next present dot plots of t-statistics by method.

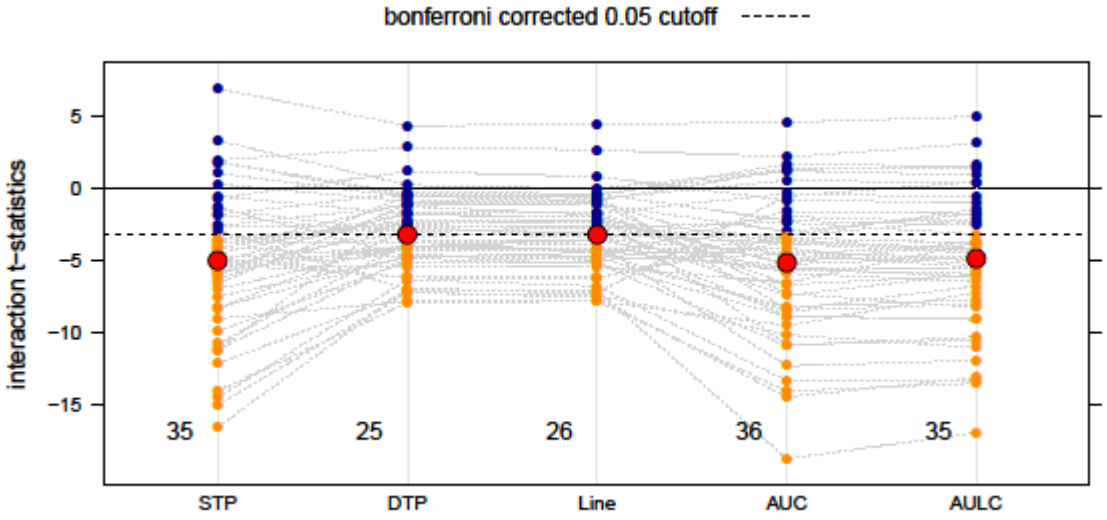

Figure 5.8 McLellan 26°C – dot plot of t-statistics

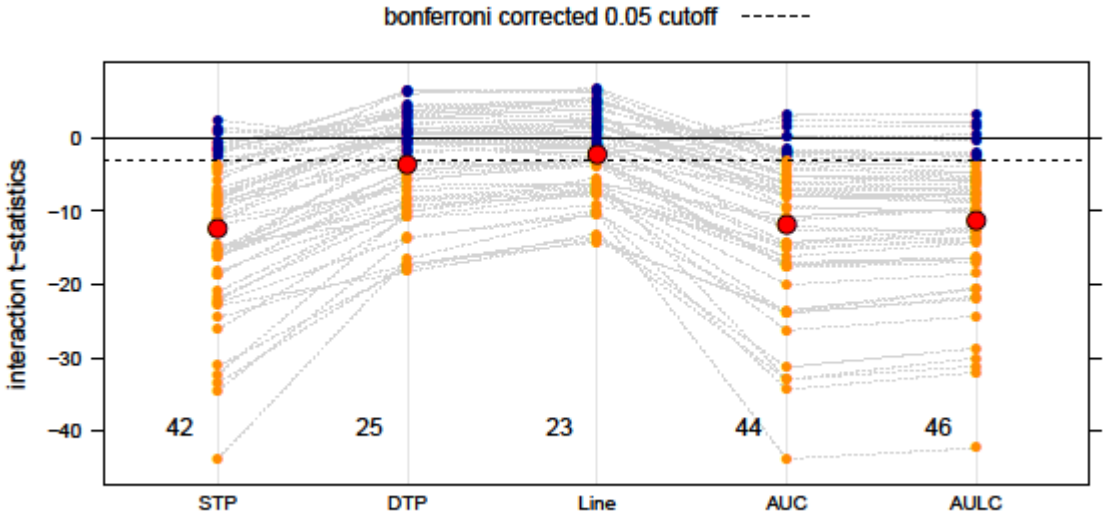

Figure 5.9 McLellan 30°C – dot plot of t-statistics

It is quite interesting that the single time point measure of strain fitness is so concordant with AUC and AULC over this set of growth curves – although we may certainly expect the quantities to be related, I would not expect such a scenario to arise in general. That said, this degree of concordance is seen in both the Stoepel and McLellan data sets. We see both a larger amount of variability, as well as more significant interactions, for STP, AUC and AULC methods, relative to DTP and Line.

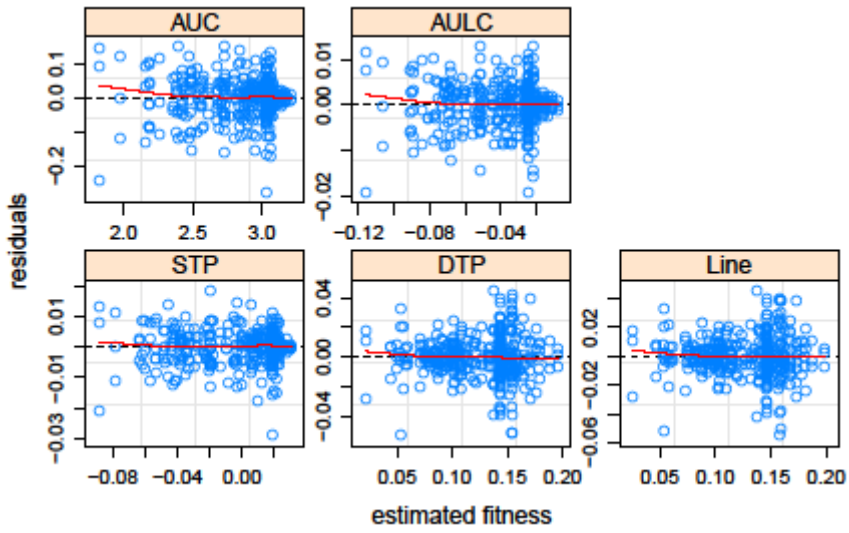

Figure 5.10 McLellan 26°C – residual plots

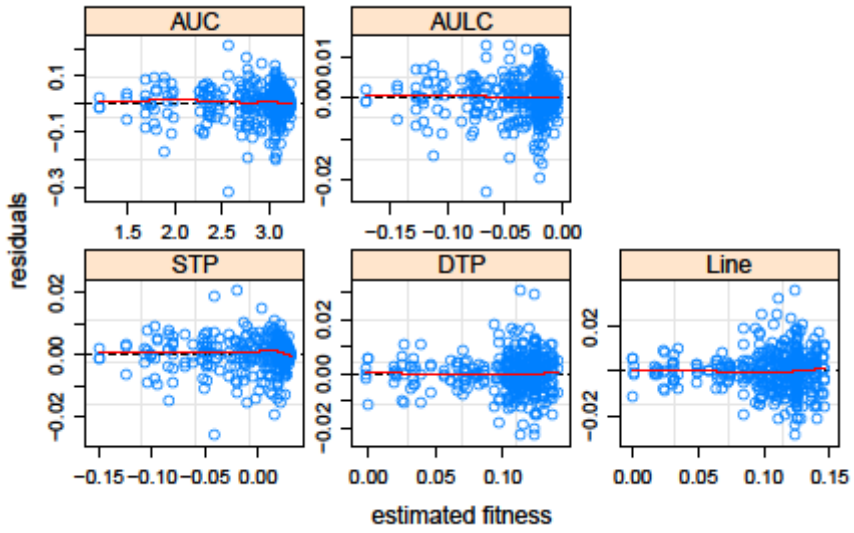

Figure 5.11 McLellan 30°C – residual plots

Residual plots are used to assess the ANOVA assumptions. Although nothing is particularly alarming, there is perhaps some heteroscedasticity present, with the variability in residuals being smaller for the larger fitted values obtained.

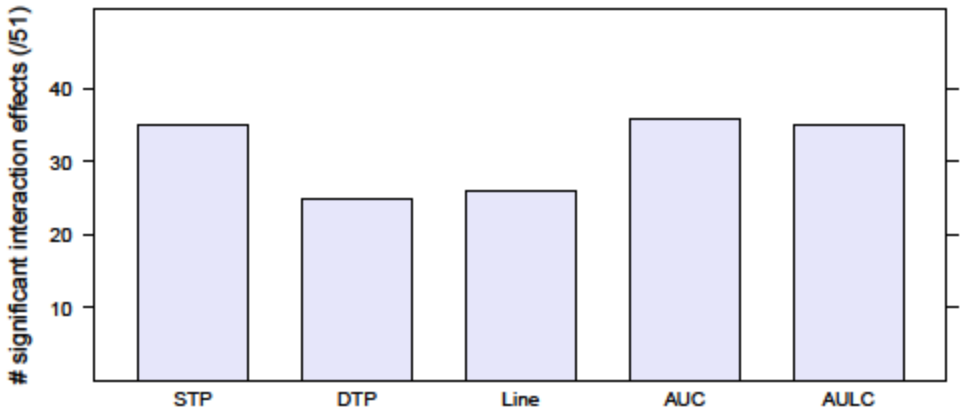

Figure 5.12 McLellan 26°C – barchart of # significant interaction effects by method

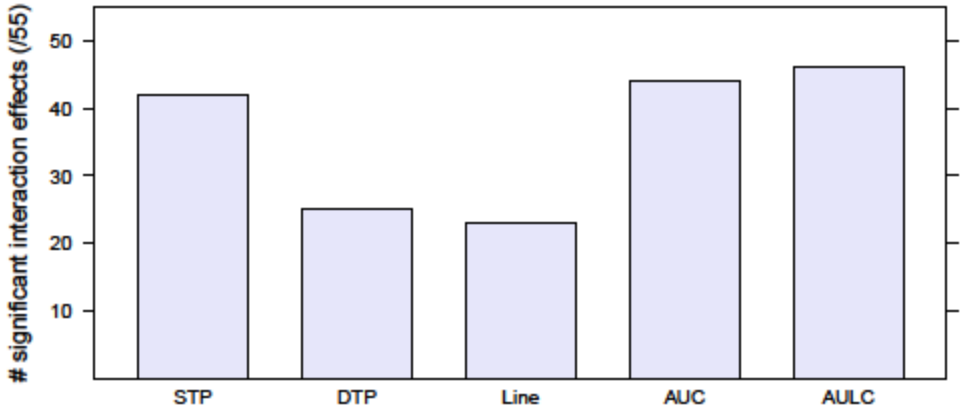

Figure 5.13 McLellan 30°C – barchart of # significant interaction effects by method

When faced with a set of growth curves in which a number of curves showcase non-sigmoidal growth, STP, AUC and AULC are able to identify a large amount of significant interaction effects. AUC and AULC have nearly identical performance with this set of curves.

A final set of interacting and non-interacting gene pairs from the McLellan 30°C data set are presented, with AUC used as a definition of strain fitness. WT denotes wild type and NI denotes the expected, or neutral, interaction. We see that the *cdc20-2, scc1* double mutant is nearly dead, even though its corresponding single mutants are not sick. As a contrast, the

kar3, *scc1* double mutant is not sick, nor are its corresponding single mutants. Even though the *cdc20-2*, *scc1* interaction is quite obvious, we should note that the non-sigmoidal / non-exponential shape of the double mutant would cause any parametric approach to fail, once again hinting that a non-parametric definition of strain fitness might be preferred.

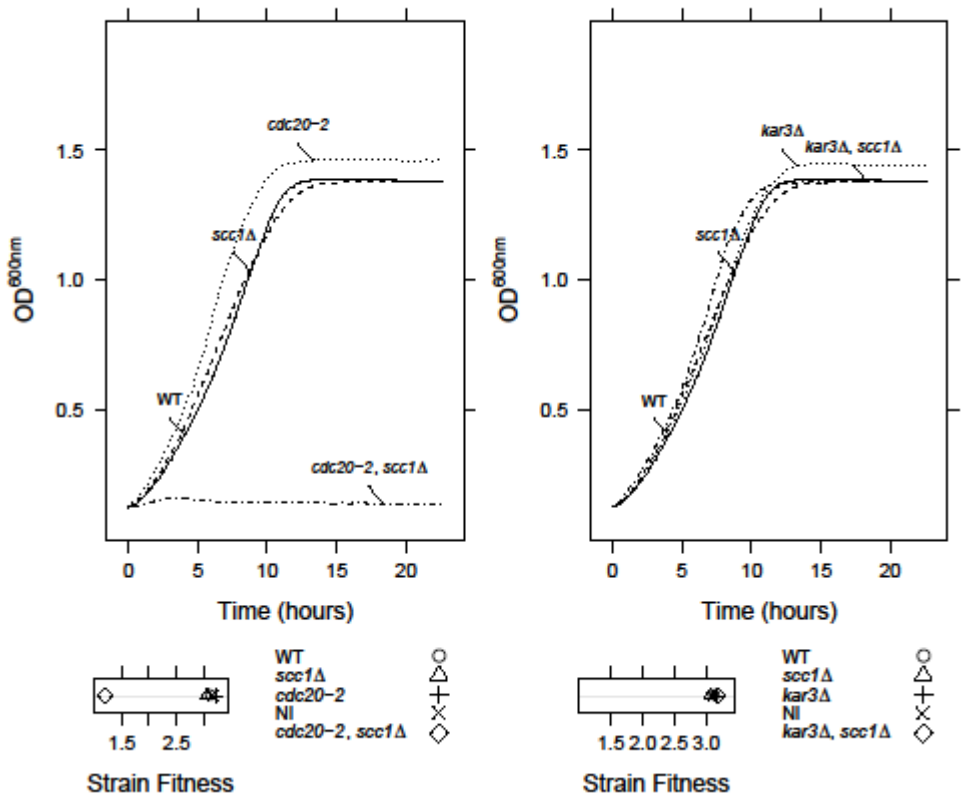

Figure 5.14 Interacting, non-interacting gene pairs from McLellan 30°C data set

5.5 Conclusions from Empirical Studies

Exploration of the Stoepel data set seems to suggest that, when faced with growth curves that express exponential / sigmoidal growth quite uniformly, one may be able to identify a few more significant interaction effects when using a method that exploits the exponential / sigmoidal nature of the growth curves. In this data set, AUC and AULC still performed well, but were slightly outperformed by other methods. Interestingly, even simple methods like STP, DTP and Line were able to capture the same interaction effects as FPL, suggesting that

the gain in fitting more complex models might be small, and almost certainly not worth the extra hassle in coercing such models to fit.

If we have a mixture of sigmoidal and non-sigmoidal growth curves as in the McLellan data set, the AUC and AULC gain more utility. These non-parametric definitions of strain fitness are able to identify far more interaction effects relative to the DTP and Line methods. Furthermore, because DTP and Line methods rely quite strongly on the assumption of exponential growth, it would be quite difficult to justify such methodology in the face on non-exponential growth, as r is an inadequate definition of strain fitness for such curves. However, the AUC/AULC will still assign higher fitness scores to strains which grow fast and to high carrying capacities, hence remaining valid definitions of strain fitness. Interestingly, a single time point calculation of strain fitness also seems to perform just as well as AUC and AULC for this set of growth curves; however, an image of the entire growth curve is still required for my implementation to pick a suitably good time point. It is worth recording that, for each of the sets of data, the time point used in the STP method was taken to be near the midpoint of the study: at the 12 hour point for Stoepel, 14 hours for McLellan 26^{o}C and 13 hours for McLellan 30^{o}C.

Chapter 6: Conclusion

Growth curve experiments are a powerful tool for learning about genetic interaction; however, analysis of a growth curve data set is not trivial. I have explored the most common growth models used when defining strain fitness (exponential and logistic), and related different definitions of strain fitness to these models. However, because in a growth curve experiment we are often faced with a mixture of both sigmoidal and non-sigmoidal growth curves (notwithstanding computational difficulties in fitting the logistic growth model to an entire data set's worth of growth curves), a more flexible non-parametric modeling strategy may be required. I have proposed AUC and AULC as definitions of strain fitness that are both conceptually sound and easy to implement under both well-behaved exponential and logistic growth curves (Stoepel) and more difficult non-sigmoidal growth curves (McLellan).

Interestingly, a single time point implementation seems to find a middle ground in terms of statistical significance, and its ease of implementation may also be very attractive to researchers. The empirical studies also suggest that, were someone to form a single time point analysis as I have, the time point should be chosen taken 12 to 14 hours after the beginning of the experiment. This could help to save time relative to the canonical 24-hour growth curve experiment.

Future research should focus on confirmation that the methodology described in this thesis is capable of identifying true interaction effects with a reasonable amount of sensitivity while also controlling the false positive rate at some user-specified level. Because the gene pairs seen throughout this thesis were expected to react, one metric used in verification of the methodology's performance was the number of significant interaction effects found. But one cannot truly claim that one method is more powerful than another until it is established that both have the Type I error rate under control. One might compare these methods over a set of known interacting gene pairs, assessed in a pool of non-interacting gene pairs, to properly assess false positive and false negative rates. Although simulation studies in which the number of true interaction effects is known could be performed, it would be difficult to simulate growth curves that did not follow a parametric model easily, and thus the results from such a study may not be as biologically relevant.

References

Addinall, S. G., Holstein, E., Lawless, C., Yu, M., Chapman, K., Banks, A. P., . . . Lydall, D. (2011). Quantitative fitness analysis shows that NMD proteins and many other protein complexes suppress or enhance distinct telomere cap defects. *PLoS Genetics, 7*(4), e1001362.

Diggle, P., Heagerty, P., Liang, K., & Zeger, S. (2002). *Analysis of longitudinal data* (2nd ed.) Oxford University Press, USA.

Doostzadeh, J., Davis, R. W., Giaever, G. N., Nislow, C., & Langston, J. W. (2007). Chemical genomic profiling for identifying intracellular targets of toxicants producing parkinson's disease. *Toxicological Sciences, 95*(1), 182-187. doi:10.1093/toxsci/kfl131

Hartman, J., & Tippery, N. (2004). Systematic quantification of gene interactions by phenotypic array analysis. *Genome Biology, 5*(7), R49. doi:10.1186/gb-2004-5-7-r49

Hastie, T., & Tibshirani, R. (1990). *Generalized additive models.* Chapman & Hall, London.

Kahm, M., Hasenbrink, G., Lichtenberg-Frat'e, H., Ludwig, J., & Kschischo, M. (2010). Grofit: Fitting biological growth curves with R. *Journal of Statistical Software, 33*(7), 1-21.

Kennedy, M. A., Kabbani, N., Lambert, J., Swayne, L. A., Ahmed, F., Figeys, D., . . . Baetz, K. (2011). Srf1 is a novel regulator of phospholipase D activity and is essential to buffer the toxic effects of C16:0 platelet activating factor. *PLoS Genetics, 7*(2), e1001299.

Lee, W., St.Onge, R.,P., Proctor, M., Flaherty, P., Jordan, M. I., Arkin, A. P., . . . Giaever, G. (2005). Genome-wide requirements for resistance to functionally distinct DNA-damaging agents. *PLoS Genetics, 1*(2), e24.

Mani, R., St Onge, R. P., Hartman, J. L.,4th, Giaever, G., & Roth, F. P. (2008). Defining genetic interaction. *Proceedings of the National Academy of Sciences of the United States of America, 105*(9), 3461-3466. doi:10.1073/pnas.0712255105

McLellan, J., O'Neil, N., Tarailo, S., Stoepel, J., Bryan, J., Rose, A., & Hieter, P. (2009). Synthetic lethal genetic interactions that decrease somatic cell proliferation in caenorhabditis elegans identify the alternative RFC CTF18 as a candidate cancer drug target. *Molecular Biology of the Cell, 20*(24), 5306-5313. doi:10.1091/mbc.E09-08-0699

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., & R Development Core Team (2011). *nlme: Linear and nonlinear mixed effects models*. R package version 3.1-100.

R Development Core Team (2011). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Shah, N., Laws, R., Wardman, B., Zhao, L., & Hartman, J. (2007). Accurate, precise modeling of cell proliferation kinetics from time-lapse imaging and automated image analysis of agar yeast culture arrays. *BMC Systems Biology, 1*(1), 3. doi:10.1186/1752-0509-1-3

St Onge, R. P., Mani, R., Oh, J., Proctor, M., Fung, E., Davis, R. W., . . . Giaever, G. (2007). Systematic pathway analysis using high-resolution fitness profiling of combinatorial gene deletions. *Nature Genetics, 39*(2), 199-206. doi:10.1038/ng1948