Pre-Processing of Quantitative Phenotypes from High Throughput Studies

by

STEVE KANTERS

B.Sc., McGill University, 2004

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

December 2006

Abstract

High throughput phenotypic experiments include both deletion sets and RNAi experiments. They are genome wide and require much physical space. As a result, multiple plates are often required in order to cover the whole genome. The use of multiple plates leads to systematic plate-wise experimental artefact, which impede statistical inference. In this paper, current pre-processing methodology will be reviewed. Their fundamental principle is to align a common feature shared by all plates. From this very principle, we propose an improved method which simultaneously estimates all parameters required for the pre-processing transformation.

Some of the alignment features popular today implicitly assume conditions which are often not met in practice. We discuss the various choices of features to align. Specifically, the upper quantiles and the mean of the left tail trimmings of each plate's data distribution are features which are always available and simple to obtain. Moreover, they are robust to non-randomization of genes to plates. Their use will be motivated through simulation and applied to real data. Applications to real data will be used to demonstrate superiority over current methods as well as to discuss choices in transformation types.

Contents

List of Figures

Acknowledgements

First and foremost, I would like to thank Jenny Bryan for her guidance, support and patience. Hopefully, my emotional roller coaster throughout this program did not rub off on others. I would also like to thank Elizabeth Conibear from the Center for Molecular Medicine and Therapeutics for giving me the opportunity to work on such an interesting project. My experience as a graduate student in the Department of Statistics at UBC has by far exceeded my expectations. I feel like a statistician - a scientist - something I did not feel after my undergraduate studies.

I would like thank my parents, Ron and Renee, for their life long love and support. They have given me the proper tools to lead the life I choose. Dad, you've shown me that life is an adventure. One which too many people take for granted. Maman, tu m'a donner un outil essentielle a la vie: la communication.

Many students in the department helped me along the way. Thank you Justin for helping improve my computing skills and for entertaining coffee breaks. Thanks Jeff for helping me with my mathematics when I needed it and speaking French. Mike Marin, you've become more than a fellow student - you're a true friend. I enjoyed the many conversations we've had which I think were beneficial to both of us. Finally thanks to Robin Steenweg and Jon Morrison who both continue to inspire me greatly in the game we call life.

STEVE KANTERS

The University of British Columbia
December 2006

Chapter 1

Introduction

Genetics has greatly developed in the last century and a half. Traditionally, experiments started with a phenotype and tried to identify the mutation which lead to it. This is known as forward genetics. Today, select organisms have most of their genome mapped. Accordingly, one can cause or observe changes in genetic sequence and witness the induced change in phenotype. This procedure, in the opposite direction of classical genetics, is called reverse genetics. By perturbing a specific gene and evaluating its effects on a phenotype of interest, we can better understand each gene's functionality. Such experiments carried out on a genome-wide basis are called high throughput phenotypic (HTP) experiments. Choosing a specific phenotype may lead to the proper identification of genes composing a pathway of interest. Furthermore, the introduction of double deletion sets in HTP experiments has allowed scientists to identify key synthetic interactions [1]. Unlike other genomic-based approaches such as microarrays and proteomics, gene perturbation provides a direct link from gene to function and, to date, offers the best tool for realizing the full potential of the genome project [3].

In all high throughput genome-wide experiments, data are very much affected by the biological phenomenon being investigated and by the experimental processes involved. Thus, analyzing the raw data to draw conclusions on the underlying biological mechanisms may be extremely misleading. Data pre-processing is the process of removing as much experimental artefact as possible while affecting the biological information as little as possible. In microarrays it is a topic which is well documented in the literature [13],[5],[8].

There are key differences between HTP and microarray experiments. Consequently, normalization methods developed for microarrays are of limited usefulness here. These experiments are young and relatively unknown to the statistical community. The goal of this report is to develop a pre-processing method for HTP experiments which properly reflects the biology involved and efficiently alleviates the experimental artefacts.

Five sections will be used to present and justify our proposed normalization approach. The first will describe HTP experiments. The second will discuss current approaches for normalizing these and microarray experiments. The ensuing section will present our proposed method which will be followed by applications to real data. Finally, the fifth section will be composed of concluding remarks.

Chapter 2

HTP Experiments

Proper discussion of high throughput phenotypic data pre-processing requires a better understanding of the experiments. In this section, key aspects of these experiments will be described and, in the process, essential vocabulary will be developed.

Individual HTP experiments differ in many respects. The primary distinctions are the organism of study, the gene perturbing strategy, the experimental format and the phenotype of interest. Organisms studied in this manner are *S. cerevisiae*, *Drosophila melanogaster*, *C. elegans* and various mammalian cell cultures (most commonly mice and human). There are numerous reasons for studying these particular organisms. For example, a significant proportion of genes in *S. cerevisiae* are homologous to human genes implicated in human disease [12]. Moreover, *C. elegans* is the simplest organism to have biological systems, such as a nervous system and digestive system, akin to those found in humans [11]. Most importantly, though, these are the organisms for which a substantial portion of the genome is known.

There are two classes of gene perturbation: gene deletion and gene inhibition. Gene deletion is permanent. The gene is removed from start codon to stop codon and usually replaced with a cassette. The cassette contains many regions which it shares with the targeted gene both upstream and downstream to allow proper insertion. It usually contains a kanamycin resistance gene and unique molecular bar-codes. The latter was done for the genome-wide *S. cerevisiae* deletion set [9]. Figure 2.1 illustrates this process. Gene inhibition is temporary. This is accomplished through RNA interference (RNAi). Interfering RNA molecules will bind to the mRNA expressed by a targeted gene and neutralize it, hence silencing the gene as depicted in figure 2.2. Once the gene inhibiting substance is exhausted, the cell behaviour returns to normal. RNAi is not yet perfected. Presently, there are still issues with efficiency and specificity. The targeted gene is not always completely silenced and other genes may be silenced simultaneously. If genes were televisions, deletion would be the mute option and RNA interference would be turning down the volume. Hence, gene deletion is more successful than inhibition. For convenience, throughout this report, we will often use **mutant** to describe a specimen with one or more deleted or inhibited genes even though technically the latter is not a mutant. It is also common to refer to gene deletions as **knock-outs** and inhibited genes as **knock-downs**.

These gene perturbing strategies are employed on a genome-wide scale. Dealing with genome-wide collections of mutants requires much physical space. There are three main experimental **formats**: multi-plate, living cell microarrays and pooled cells. Figure 2.3 depicts the three formats. The multi-plate format is used for all types of gene perturbations and organisms. For RNAi experiments, microtitre plates are most commonly used. Hence in the RNAi community, the multi-plate format is often referred to as the microtitre or multi-well plate format [2]. For deletion sets, other types of plates are also used. For example, yeast deletion sets are often grown on agar plates. Agar plates are sterile Petri dishes that contain agar and nutrients. They lend themselves particularly well to yeast since it is a great medium for growth, the most common phenotype for deletion sets, and allow other interesting phenotypes to be measured, such as invasiveness [4]. Rarely are plates capable of hosting more than 1536 individual cell colonies or cultures. Genome-wide experiments deal with thousands of mutants thus requiring a collection of plates. A **plate-set** is the collection of plates used to perform one experimental replicate for all mutants. It is this format which is of interest to this report.

Within the multi-plate format there are many experimental set-ups. Technical replicates of a mutant are colonies or wells with the same gene perturbation within a plate-set. In general, if there are technical replicates, they are contained on the same plate. Consequently, the mutants found on two separate plates within a plate-set are (almost) completely different. This has important consequences for data pre-processing. Fortunately, most, if not all, experiments have common elements found on multiple plates. These are generally referred to as **controls**. In essence, a control is anything which is shared on two or more plates in the set and should have a similar phenotype on each plate. Wild type cells placed on each plate are a typical control. Wild types have not been altered in any way, so they are providing information in the form of a reference phenotype. This has the convenient feature of allowing the identification of gene perturbations which result in an enhanced, diminished or unchanged phenotype. Another common control is to have mutants with known phenotypic behaviour spanning over the range of phenotypes across plates. For example, suppose a growth experiment on a yeast deletion set is conducted. A plausible choice of controls would be to have a slow growing yeast, an average growing yeast and a fast growing yeast on each plate. It is also common for **blanks** to be present on each plate, but despite their consistent behaviour they are different from controls. They don't provide as much information about the plate effect as biological controls do. We will expand on the topic of the plate effects in the following sections where this will become evident. Blanks are nonetheless useful for quality control at the pre-processing stage.

Once the plate-set has been prepared, phenotyping may begin. The mutants are set in a fixed environment for a fixed period of time. A specific stress or treatment, such as an enzyme or drug, may be introduced. Essentially, the mutants are exposed to specific conditions. The phenotype of interest is often growth, even when exposed to treatment. Furthermore, the study of one phenotype does not preclude the study of others, even with the same plate-set. In particular, in deletion sets, growth assays are often the first phenotype studied because other phenotypes may have to be adjusted for the observed growth.

The final experimental stage is phenotype quantification. Of interest is the exact measure of the phenotype. For example, in the case of a growth assay, the information of interest is the number of cells resulting from the growth period. These experiments are on a genome-wide level, so it is impractical for phenotypes to be measured exactly. As a compromise, technologies have been developed which estimate the phenotype. For simple assays, the plates may be scanned to capture an image and the resulting spot size or brightness for each mutant is measured. For more complex phenotypes cell engineering may be required to ease the task of quantification. Fluorescent or luminescent markers may be inserted to allow plate readers to quantify phenotypes. These are usually the product of **reporter genes** inserted in the cell. The raw image resulting from the scan is then processed using image quantification software. Figure 2.4 shows a raw image of a single agar plate from a growth assay using a yeast deletion set along with its processed image produced by such software. In the end, a numeric quantity is assigned to each mutant which hopefully reflects the phenotype.

Statisticians' motto is to never throw away data. Some experimenters choose to evaluate the results visually as healthy/not healthy. Such an approach removes the need for normalization, yet simplifying potentially continuous data to a categorical or binary form is equivalent to throwing away a portion of the data. Furthermore, it introduces issues of replicability and increased error. Firstly, different people will categorize differently. In fact, due to the subjective nature of these measurements, the same person may categorize differently on two separate occasions. Secondly, the extra data manipulation required here may lead to human error. Fortunately with the advent of image quantification technologies, this strategy is becoming less common. Phenotype quantification gives access to all information provided by the experiment and gives rise to a need for data pre-processing.

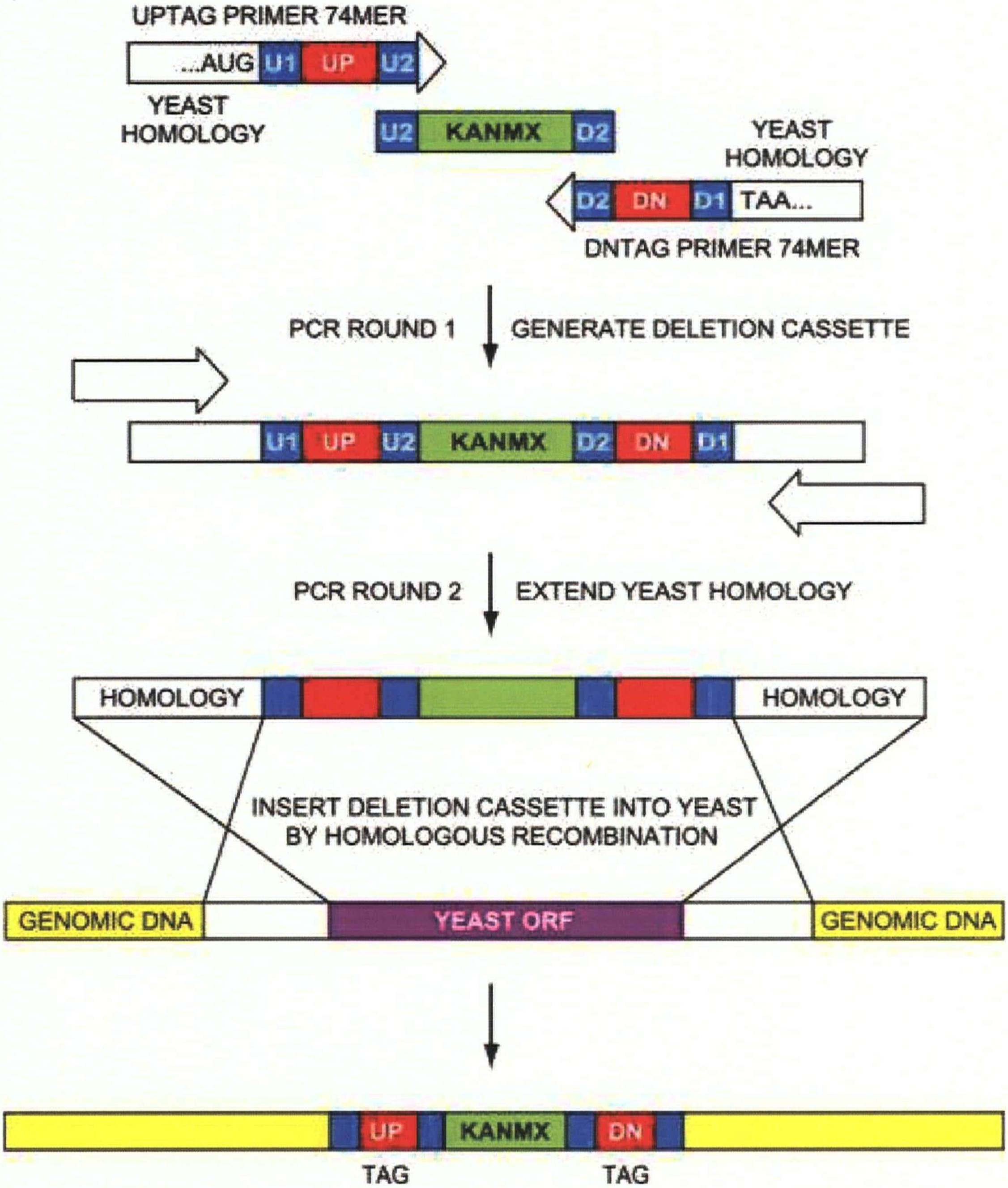


Figure 2.1: The following image was obtained from Robert G. Eason's article on the characterization of synthetic DNA bar codes [6], copyright(2004). It depicts both the construction of the cassette and its insertion into the location of the deleted gene. The unique identifying bar codes are essential for hybridizing purposes.

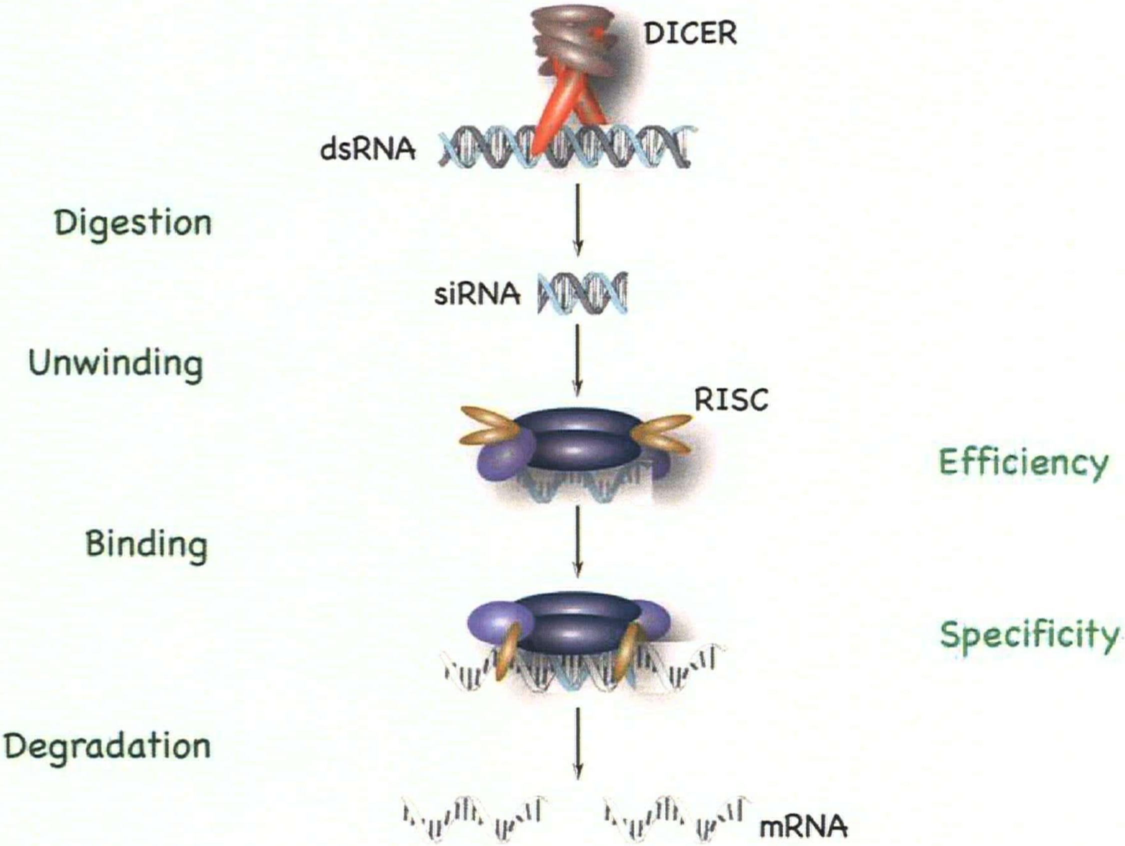


Figure 2.2: The above figure was obtained from: Henschel A., Buchholz F. and Habermann B.,DEQOR: A Web-Based Tool for the design and Quality Control of siRNAs.*Nucleic Acids Research*, **32**:W113-W120,2004, by permission of Oxford University Press. It depicts the multiple stages involved in gene inhibition via RNAi.

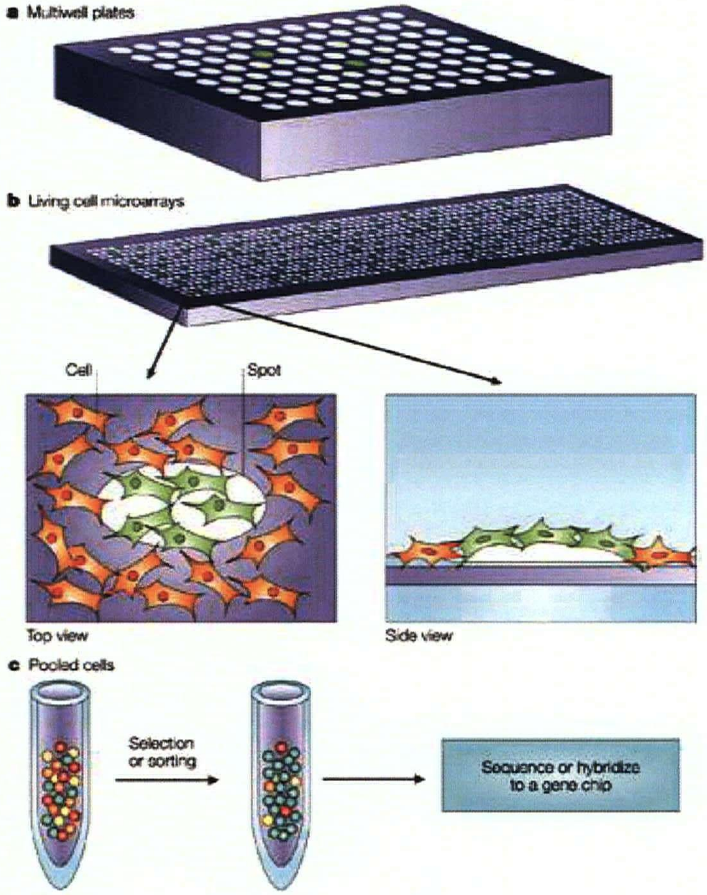


Figure 2.3: **Experimental Formats a:** Multi-well plates come in various sizes (typically 96, 384 and 1536 wells), thus the complete experiment would comprise of many such plates. **b:** Living cell microarrays are large arrays with thousands of spots printed on them. An inhibiting solution targeting a precise gene is placed at each spot. Cells are then distributed across the entire array. This format does not lend itself well to whole organisms such as the worm due to space constraints. In contrast to pooled cells, living cell microarrays are confined to gene inhibition because they rely on inhibiting solutions. **c:** Pooled cells begin with an equal share of each mutant. After a set period of time or exposure to some stress, the collective DNA is analyzed using microarrays. This requires each mutant to be identifiable. Most commonly, identification is obtained through bar codes. Thus, this format is mostly confined to deletion mutants. Adapted by permission from Macmillan Publishers Ltd: Nature Reviews Genetics, copyright (2004) [2].

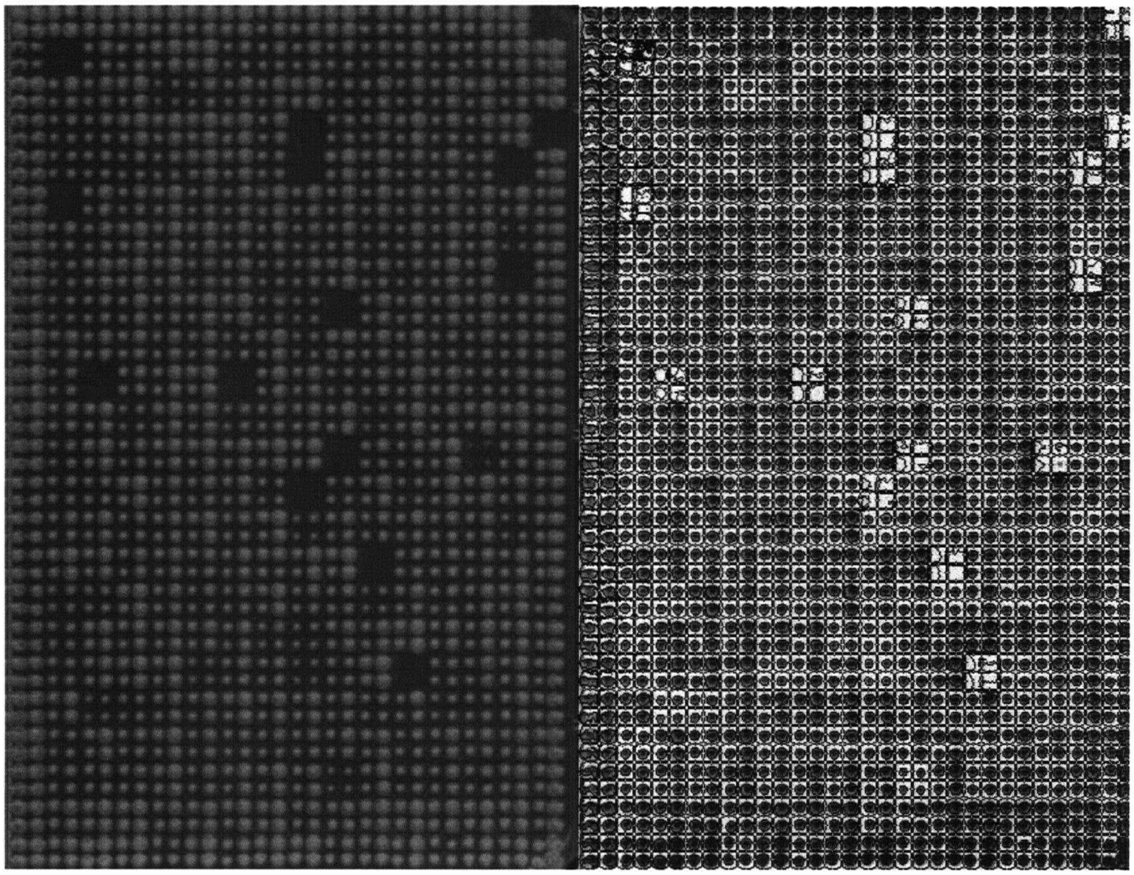


Figure 2.4: The image on the left is a scan of an actual plate used in a growth assay. The image on the right is a resulting quantified image obtained using the Grid Grinder software. Along with this image comes a numerical value for each individual colony. These images are provided courtesy of the Conibear Lab.

Chapter 3

Foundations for HTP Data Pre-Processing

Multi-plate high throughput phenotypic experiments always involve multiple steps. Inevitably, the handling of each plate will be different to some degree. The discrepancies may come from various sources, such as the temperature of treatment or the distribution of nutrients in the agar. Undoubtedly, the experimental processes will affect the data in a plate-wise manner. The experimental artefacts introduced to the data in this fashion are referred to as the plate effect. HTP experiments usually involve multiple conditions just as in microarrays. Plate-wise experimental artefacts exist both within and across plate-sets. On one hand, plate-set pre-processing - the removal of experimental artefacts across plates within a plate-set - presents a novel problem. On the other hand, pre-processing plate-sets across experimental conditions may be achieved using existing microarray methodology. Throughout the remainder of this paper, plate effect will pertain to the plate-wise distortion within a plate-set only.

Data arising from multi-plate HTP experiments are used in a fundamentally different manner than microarray data. In microarray experiments, comparisons of a gene across different conditions are direct and inter-gene comparisons are indirect. Comparisons between genes are done based on their behaviour across conditions. For example, genes may be clustered according to their up-regulation and down-regulation across conditions. Pre-processing methods in microarrays are typically concerned with removing experimental artefacts with regards to a gene across all samples. Doing so allows for both types of comparisons to be meaningful. HTP experiments, on the other hand, are very interested in direct inter-mutant comparisons within plate-sets and across conditions. These differences are depicted in Figure 3.1. We can hope to make meaningful comparisons of mutants within a plate-set under the presence of random biological and/or technical variability, but they will be compromised by the presence of systematic, plate-wise experimental artefacts. The complete removal of the plate effect is an ideal which is (almost) never achieved in practice. The goal of pre-processing is to minimize the magnitude of the plate effect. Existence of a plate effect and a genuine desire to make direct comparisons within a plate-set establish a need to develop a pre-processing method for plate-sets in HTP experiment which is the purpose of this paper. The phenotype of a particular mutant may be further affected by its location on the plate. These experimental artefacts will not be dealt with in this paper, but it is suspected that the conceptual infrastructure developed here will be generalizable to other sources of experimental artefacts such as location.

There is a simple blueprint to deriving a data pre-processing method. First, a distortion model deemed appropriate for data produced by the class of experiments at hand is developed. Second, biological assumptions plausible for these types of data, and which simultaneously allow for estimation of the parameters in the proposed distortion model, are established. From these a natural solution is formulated: the pre-processing method. A good pre-processing method will remove as much experimental artefacts as possible and remove as little gene perturbation effect as possible. Having established our goal, the remainder of this chapter will explore current methods where the above blueprint will be used extensively. In order to do this properly, certain data characteristics will be shown and a distortion model will be proposed.

3.1 Phenotype Distribution

Perturbation of a gene, under a specific condition, can lead to four different outcomes with regards to a given phenotype. The phenotype may be enhanced, diminished, remain unchanged or become undefined because the mutant is nonviable. Since the mutant collections on each plate differ, the phenotype distributions also differ. If mutants were randomly allocated to plates in the set, the discrepancy with regards to phenotype distribution on each plate would solely be a result of sampling and not be substantial. Unfortunately, in practice, the allocation is not performed randomly and often much larger proportions of weak mutants (mutants with gene perturbations inducing diminished phenotype or mutants which are nonviable) are allocated to a few select plates within the set. For example, it is not uncommon for most plates to have a small proportion of weak mutants, say approximately 10%, and the few remaining plates to have a large proportion of weak mutants, say approximately 60%. Nonetheless, these distributions do have common features. There are always mutants from all four categories (unaffected, diminished, enhanced and lethal) on each plate. Therefore, the distributions of phenotypes on each plate have approximately the same range. The mutants which incur phenotypic lethality (nonviable mutants) and blanks are structural zeroes. Consequently, the distributions are often bimodal with one mode corresponding to the structural zeroes. The distribution may be multi-modal with each mode corresponding to a category. Figure 3.3 depicts smoothed histograms of the raw observed phenotypes from two plates from the same plate-set. It shows the bimodal shape of phenotype distributions and the consequences of non-random mutant allocation. The non-random mutant allocation to plates, approximately equal range and bimodality are important features of the data which should be acknowledged when deriving a pre-processing method.

3.2 Distortion Model

It is accepted by most scientists that the effect of gene perturbation is multiplicative in nature. Likewise, the effects of most experimental processes are also multiplicative. Figure 3.2 depicts the distribution of phenotypes on separate plates within a plate-set. None of the plates have observations close to zero where the phenotypes of nonviable mutants and blanks are expected to be. If the plate effect were solely multiplicative their phenotypes would be almost unaffected. Therefore, the plate effect also has an additive component. These box plots not only support the notion of an additive component to the plate-effect, but also serve as motivation for pre-processing by displaying the importance of the systematic experimental artefacts (*i.e.* not removing the plate effect will be extremely misleading).

Recall that the quantity recorded in the phenotype quantification stage of the experiment is not a measurement of the true phenotype, but a mapping of the phenotype to a numeric quantity. There is no serious interest in recovering the underlying phenotype measurement. Therefore throughout the rest of this paper, true phenotype will refer to the numeric quantity obtained from the mapping which would be observed if the *exact* protocol were followed. It will be denoted by the variable z and the observed quantity will be denoted by x.

The distortion model needs to be simple and sufficient. Thus, the distortion model may be expressed as a linear multiplicative-additive model,

$$x_{gp} = \alpha_p + \beta_p z_{gp}, \tag{3.1}$$

where g is the gene being perturbed and p the plate. Here no assumptions are made about the source of the location and scale parameters of the plate effect. The multiplicative component, β_p, is a combination of the effect of all experimental sources of multiplicative distortion. Similarly, the additive component, α_p, engulfs all sources of additive effect.

3.3 Current Approaches to Data Pre-Processing

In this section, the most common methods of data pre-processing in HTP experiments will be presented as well as select methods used for microarray data. It will be demonstrated that the latter methods are not valid for the data at hand, but certain concepts found within them may be useful for HTP data pre-processing.

3.3.1 HTP Experiments

The plate effect as specified by the distortion model in (3.1) is most easily dealt with in two steps. The first step is to remove the additive effect. It is the typical phenotype recorded when the true phenotype is zero. There are various methods of estimating this effect. One is to subtract, from all observed phenotypes on each plate, the plate's minimal observed value. This can also be done within smaller geographical confines of the plate. For example, the plate may be divided into quadrants and so on. At the most extreme level, these quantities are estimated locally for each mutant colony or well. The latter requires local estimates to be provided by the image quantification software and is only valid for plates which don't use wells, such as agar plates. Removing the estimated additive effect, $\hat{\alpha}_p$, provides adjusted observations. Most approaches currently used in HTP experiments use this first step. For sake of simplicity, let $x_{gp}^{(new)} = x_{gp}^{(old)} - \hat{\alpha}_p$ refer to the adjusted observations for the remainder of subsection 3.3.1. The plate effect for these adjusted observations may be expressed as,

$$x_{gp} = \beta_p z_{gp} \tag{3.2}$$

where the subscripts are as previously defined. The problem is now reduced to a multiplicative model.

Belief in a common location of the distribution of true phenotypes on each plate may not enable the multiplicative components of the plate effect to be estimated, but it does allow for meaningful comparisons to be made. Let $m_p = \bar{x}_{gp}$ be the mean of the adjusted observed phenotypes on plate p and let $m = \bar{z}_{gp}$ be the mean of the distribution of phenotypes having followed the *exact* experimental protocol. Hence, the value of m is unknown. The variables m and m_p are instances of random variables with common expectations in the absence of a plate effect. Thus, the plate effect is estimated by $\hat{\beta}_p = \frac{m_p}{m}$. Given m, pre-processing can be accomplished by dividing observations on plate p by $\hat{\beta}_p$. Pre-processed phenotypes will be denoted by y, so this may be expressed as $y_{gp} = x_{gp}/\hat{\beta}_p$. Due to the multiplicative nature of the effects of gene perturbation, comparisons are typically done by way of ratios. Hence comparing two mutants pre-processed in this fashion leads to

$$\frac{y_{gp}}{y_{g'p'}} = \frac{\hat{\beta}_{p'} x_{gp}}{\hat{\beta}_p x_{g'p'}} = \frac{\beta_p}{\hat{\beta}_p} \frac{\hat{\beta}_{p'}}{\beta_{p'}} \frac{z_{gp}}{z_{g'p'}}$$

where hopefully $\frac{\beta_p}{\hat{\beta}_p}\frac{\hat{\beta}_{p'}}{\beta_{p'}} \approx 1$. If the estimated plate effect is taken to be $\frac{m_p}{m}$,

$$\begin{aligned}
\frac{\hat{\beta}_{p'} x_{gp}}{\hat{\beta}_p x_{g'p'}} &= \frac{\frac{m_{p'}}{m} x_{gp}}{\frac{m_p}{m} x_{g'p'}} \\
&= \frac{m_{p'} x_{gp}}{m_p x_{g'p'}} = \frac{x_{gp}/m_p}{x_{g'p'}/m_{p'}}
\end{aligned}$$

it becomes apparent that the value of m is irrelevant. Thus, by simply dividing the adjusted observations from each plate by the plate's mean, the plates are normalized to each other and meaningful comparisons may be made. The mean may be replaced by other measures of location, most commonly the median. This approach, referred to as mean or median plate centering, is the most commonly used in current HTP experiment analyses.

An alternative approach is to take advantage of controls conveniently placed on every plate in the set. This method was applied by B. L. Drees *et al.* [4]. Keeping to the same distortion model as in equation 3.2, rather than using the belief in a common location as a means to meaningful comparisons, it is the fact that the controls have a common phenotype that is used. Define w_p to be the observed phenotype of the control on plate p. Under this constraint (common phenotype for controls) and model, the plate effect may be estimated by

$$\hat{\beta}_p = \frac{w_p}{median_p(w_p)}. \tag{3.3}$$

The adjusted data from plate p are normalized by dividing them by their respective estimated plate effect $\hat{\beta}_p$. Similarly to median plate centering, it is equivalent to divide all x_{gp} by w_p. In essence, all that is needed once the observations has been adjusted for additive effect is to divide by a phenotype which is believed to be approximately equal across all plates. The exception to this rule are blanks. These should already be approximately equal after the removal of additive effects. Hence, aligning these would provide no insight to the multiplicative component of the plate effect.

Both approaches are honest attempts at removing the plate effect, but they may not be the most effective ones. To begin with, estimations of the additive component as described above are crude. This in turn may affect model 3.2. If the model for which we are estimating the parameters, in this case β_p, is wrong to begin with, then the so called normalized data will still be heavily distorted. Moreover, mean/median centering assumes location to be the same, but this does not comply with the distribution of phenotypes previously described. Disproportions in the number of weak mutants on different plates within the set, such as that depicted in figure 3.3, will assuredly lead to different plate locations. Evidently, aligning a mutant which is a structural zero on one plate with a mutant whose phenotype is unchanged on another does not respect the underlying biology - an undesirable quality to any preprocessing method. At the very least, it is much wiser to use the median rather than the mean, but regardless both these are poor measures of location when dealing with distributions which are not unimodal. Using control phenotypes to estimate $\hat{\beta}_p$ is an improvement on median/mean plate centering. In principle this is a really good idea, however it does have a weakness. Controls are often replicated in small numbers and any given colony may be strongly affected by location phenomena. Therefore, controls have substantial variances and division by variables with large variance lead to bad variance properties for the result. Moreover, quality control should be performed to ensure the control's behaviour is consistent enough to be useful. If technical replicates are included in the experiment, then variance of a control may be obtained on each plate. By comparing the average of these variances to the variance of control phenotypes across plates in the set, quality of the controls may be assessed.

3.3.2 Microarrays

A variety of normalization approaches have been proposed for microarray experiments. Such diversity is due in part to progress and different types of microarray chips. For example, data generated by high density oligonucleotide microarray technology differ from those obtained using cDNA microarrays allowing variant normalization approaches to be employed. Two classes of normalization here are channel to channel normalization in cDNA experiments and normalizing many profiles to each other. The latter is used in high density oligonucleotide microarrays and in cDNA experiments which use many arrays. The pre-processing strategies developed for normalizing many profiles are the most useful to our cause. As previously mentioned the biological effect is multiplicative in nature. For this reason, it is common practice to *log* transform the data to render the effect linear. This, in turn, allows more traditional statistical methods to be applied. Two methods still popular today are quantile normalization [5] and variance stabilizing normalization [8].

In microarrays, each probe is present on each array. It is assumed that the majority of genes are not differentially expressed across conditions being studied. For the genes that are, change occurs in both directions. In other words, there is both up-regulation of some genes and down regulation of others. These changes generally off-set each other, so it is reasonable to assume equal intensity distributions on each array across an array set. It is this assumption which is used to generate quantile normalization. Recall that an array set in the microarray context is very different from a plate-set in the HTP context. A plate-set is like a single microarray. To compare two datasets with regards to distribution, one can construct a quantile-quantile (Q-Q) plot. A perfectly diagonal line on a Q-Q plot indicates equal distribution, up to a location, of the two datasets. The Q-Q plot serves as motivation for a transformation that can be used to render the distribution of phenotypes on both plates equal. Two arrays in an array set will almost never have the same distribution. In practice the exact distribution of each plate is unavailable. Rather than use quantiles as a basis for a transformation, the G order statistics are used. Let $x_{(g)p}$ be the g^{th} order statistic of observed phenotypes on plate $p = 1, 2$. Then the mapping

$$y_{(g)p} \mapsto \frac{x_{(g)1} + x_{(g)2}}{2} \qquad \forall p, g \tag{3.4}$$

will force the distribution of both arrays to be the same. This non-linear transformation easily extends to array sets of size P. An example using multiple arrays is shown in figure 3.4. Unfortunately, assuming the equal phenotype distributions on each plate within a set, in the HTP context is not reasonable. Here the mutants composing each plate are entirely different and often not randomly allocated. However, the next chapter will discuss how exploiting quantiles may be of use in the quest to HTP data pre-processing.

Huber *et al* suggest an alternative pre-processing method for microarray data which both normalizes and stabilizes the variance all at once, which is aptly named variance stabilizing normalization [8]. It is based on a multiplicative-additive distortion model as in equation 3.1 combined with a multiplicative-additive error model suggested by Rocke, D.M. and Durbin, B.P. [10]. The resulting model is

$$x_{gp} = a_p + \beta_p z_{gp} e^{\eta_{gp}} + \bar{\nu}_{gp} \qquad \eta_{gp} \sim N(0, \sigma_\eta)\ _{i.i.d.}, \bar{\nu}_{gp} \sim N(0, \sigma_\nu)\ _{i.i.d.}. \tag{3.5}$$

This implies a quadratic mean-variance dependency and leads to the following pre-processing transformation

$$y_{gp} = arsinh\Big(\frac{x_{gp} - a_p}{b_p}\Big) \tag{3.6}$$

where b_p is a function of β_p and σ_ν. Thus the goal here is to estimate the coefficients a_p and

b_p as best as possible to eliminate the plate effect and (a topic which is not addressed here) to minimize the structural relationship between the mean and the variance.

The assumption used to derive this normalization method is that the majority of genes are not differentially expressed. The data is trimmed to include only genes believed to be non-differentially expressed. Naturally, the means of non-differentially expressed genes should not change across arrays. Thus, by mean centering the trimmed data, maximum likelihood estimates for the parameters can be obtained. The result is a normalization approach which estimates a_p and b_p simultaneously. Thus, rather than use crude estimates for the additive parameters, both the additive and multiplicative parameters of the distortion model are estimated using biological assumptions.

To some the inverse hyperbolic sine function may seem foreign, but it is closely related to the *log* all while having advantages over it: $\text{arsinh}(x) = \log(x + \sqrt{x^2 + 1})$. Therefore, using this normalizing method removes the need to *log* transform the data as discussed earlier.

Variance stabilizing normalization may not be directly applied to HTP data because key to the maximum likelihood estimation of the parameters is the presence of the same genes on each array. Unfortunately, plates within a set do not share mutants in HTP experiments.

Over the course of this chapter, currently employed HTP pre-processing methods were presented along with two microarray normalization methods. The fundamental principles driving current methodolgy are good, but there is room for improvement. Methods used on microarray data may not be directly applied to HTP data, but do offer tools which can help deal with the issues of non-random allocation and parameter estimation strategies.

author = B.M.Bolstad et al., title = A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Variance and Bias, journal = Bioinformatics, year = 2003, volume = 19, number = 2, pages = 185-193

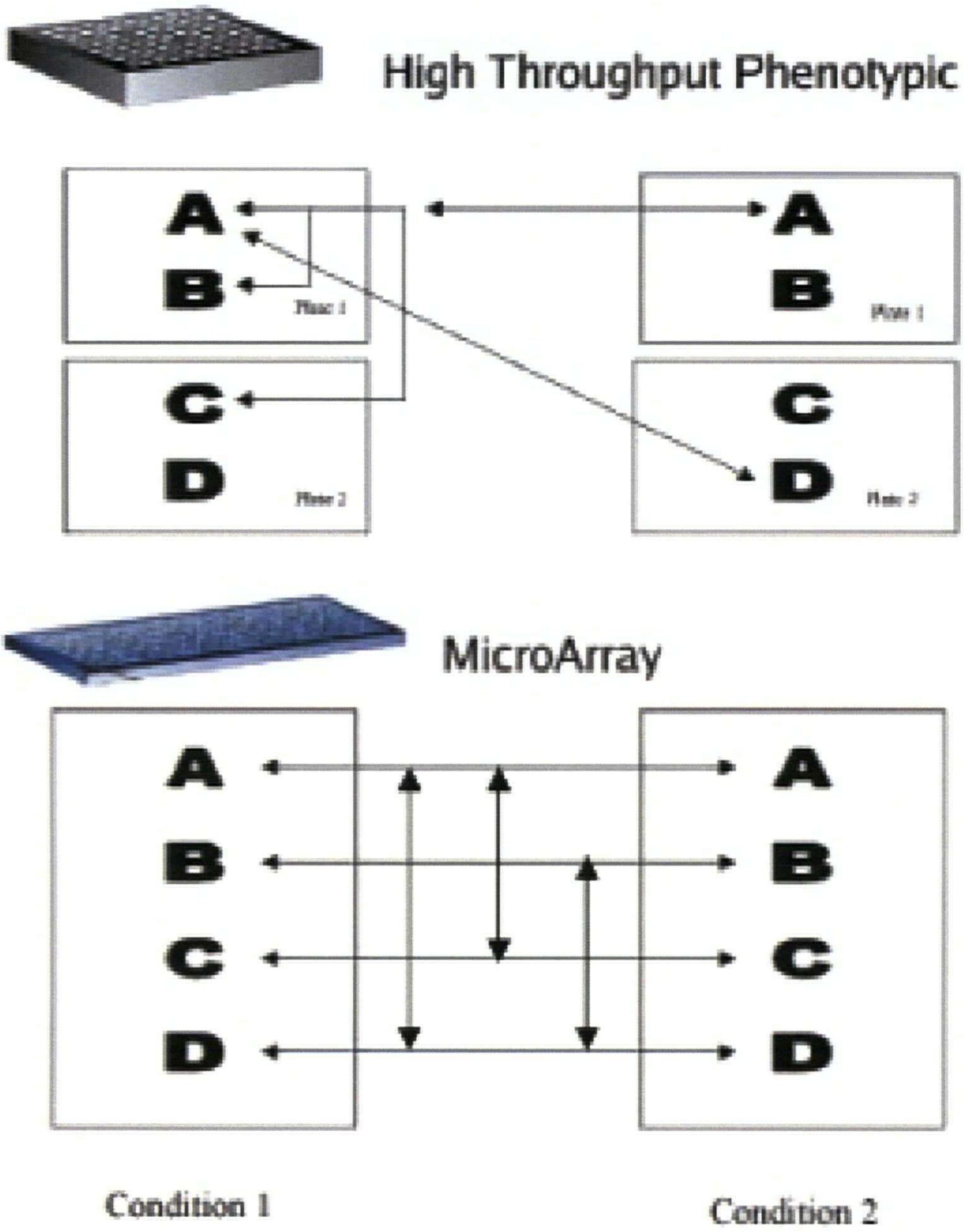


Figure 3.1: This diagram is a simple comparison of HTP and microarray experiments. In HTP experiments, plate-sets are required for each condition. Pre-processing of plate-sets is done independently for each condition.

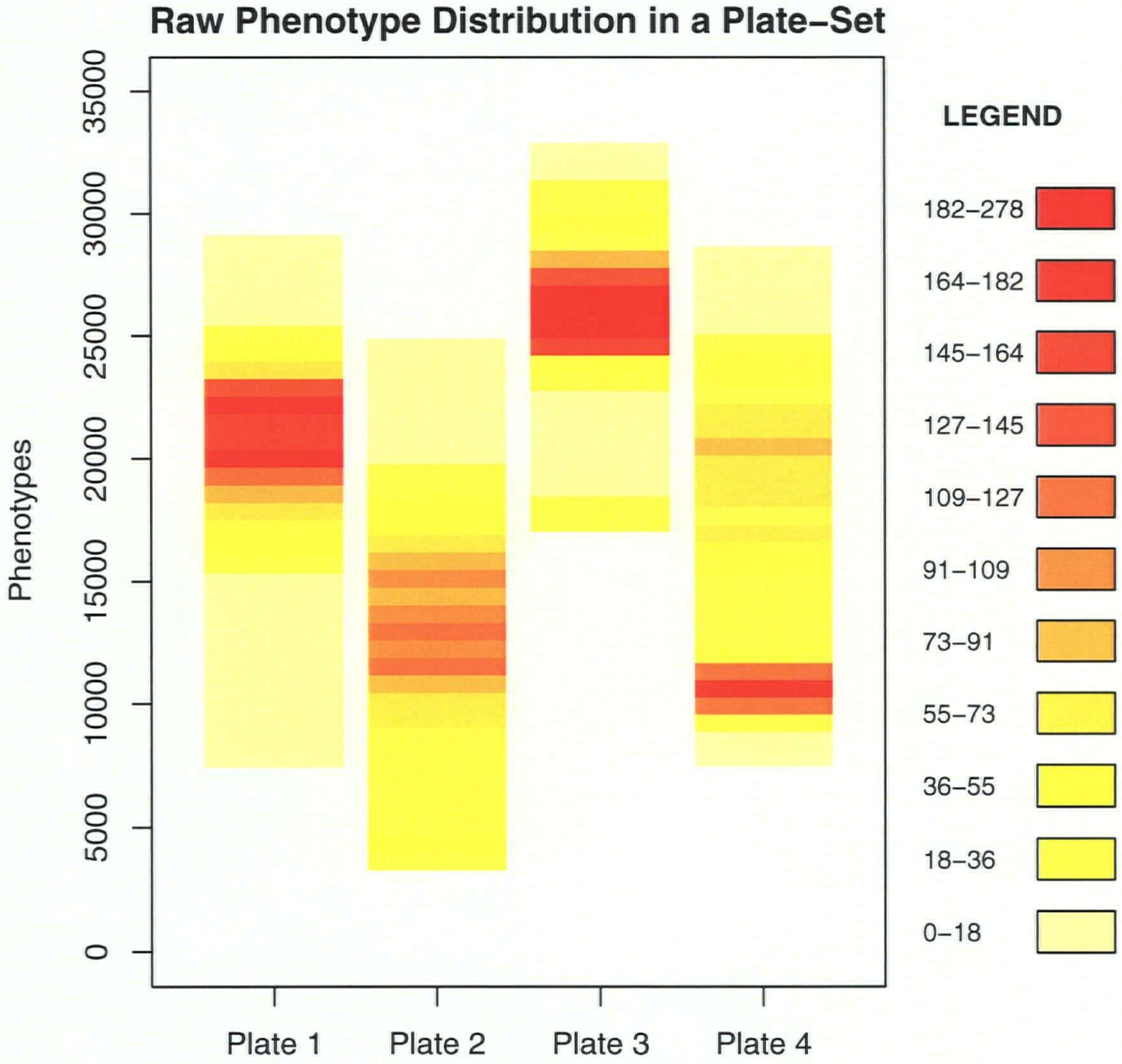


Figure 3.2: These data come from a plate-set containing more plates, but only four are included here for convenience. In these modified box plots, each box covers the entire phenotype range and uses a color scheme to present the approximate data distribution.

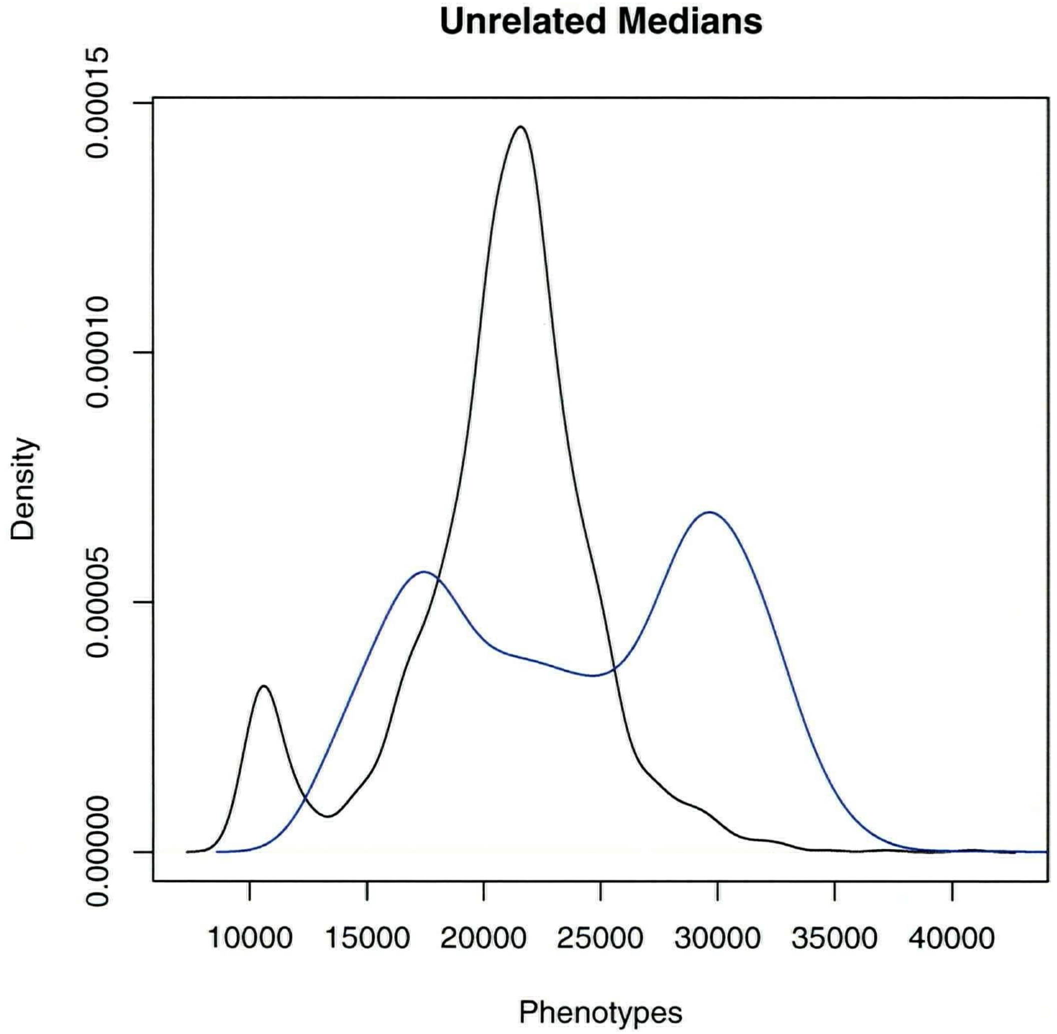


Figure 3.3: These are smoothed histograms of the observed raw phenotypes from two plates from the same plate-set. The plate depicted in blue contains many more weak mutants than the other plate and this was a conscious choice.

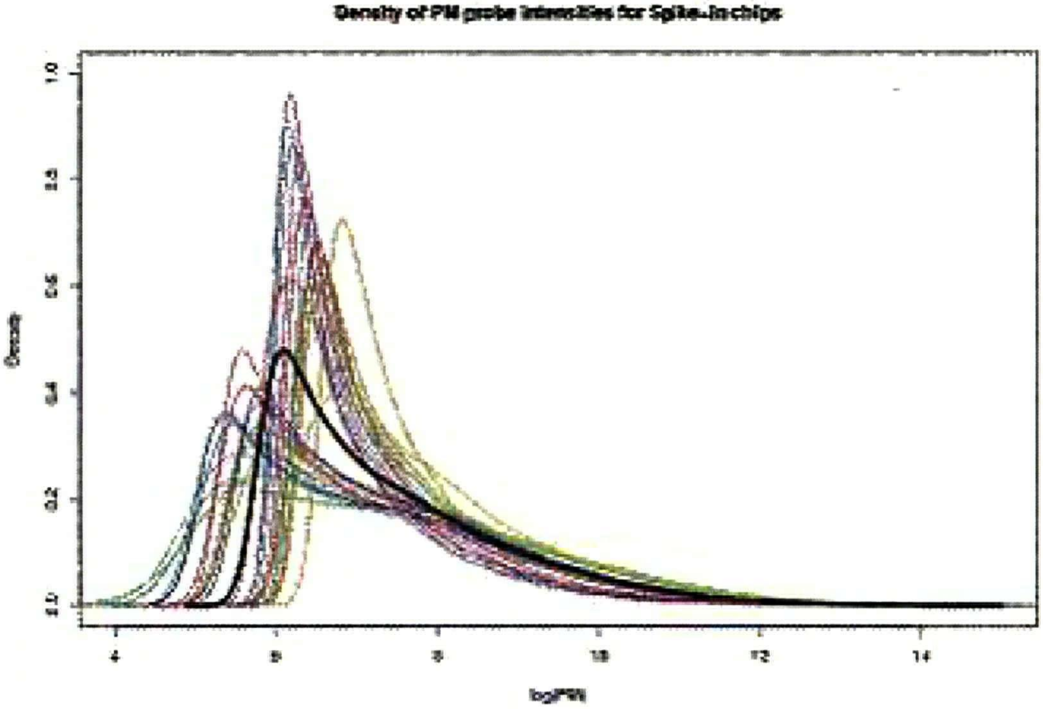


Figure 3.4: A plot of the densities for 27 arrays. The density after quantile normalization is shown in bold black. The above figure was obtained from: B.M.Bolstad et al.,A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Variance and Bias. *Bioinformatics*,**19**(2):185-193, 2003 [5].

Chapter 4

Recommended Pre-Processing Method

The previous chapter considered, among others, the assumption that phenotype distributions on plates within a set share a common mean or median. The event of non-random allocation of mutants to plates renders this assumption implausible. Furthermore, distributions with more than one mode are not well characterized by such a measure of centrality. A broad theme identified in the previous chapter is the idea of aligning a collection of datasets on one or more features believed to be constant across the collection, prior to the introduction of experimental artefacts.

A distortion model has been been proposed in (3.1). The criticism in the previous chapter was with regards to the methods, not the model. We will also use this model and simple extensions of it. Furthermore, the concept of feature alignment makes sense and will also be used here. Thus, the task at hand is to establish proper features for alignment.

4.1 Features

There are two classes of features: internal and external.

4.1.1 External Features

This class of features was introduced when phenotypes on each plate were divided by each plate's wild type phenotypes. External features are those which arise from the design of the experiment. They are external to the phenotype distribution on plates. Controls and blanks are external features. Their presence is not required for the experiment to run, thus they will not always be available as a pre-processing tool. By definition, in the absence of a plate effect, the phenotype of each external feature would be approximately equal across plates. There is no questioning the biological validity of their approximate equality, but, as previously mentioned, quality control should carried out before aligning these features.

4.1.2 Internal Features

Internal features are those which pertain to the phenotype distributions. Both median centering and quantile normalization use internal features. In the first case the feature is the median of the phenotype distributions on each plate and in the second the features are all quantiles of the plate distributions. Their invalidity as features to be aligned in HTP experiments has been established. Yet, there is a middle ground in which certain quantiles, and the mean up to certain quantiles, tend to be approximately equal in the absence of experimental artefacts in the HTP context.

Internal features considered thus far have been susceptible to departures from mutant randomization. For a better understanding of how the data are affected by a lack of randomization, consider a simple simulation which compares two plate-sets: one constructed to represent randomized mutant allocation which is implicitly assumed by median centering and the other to represent non-randomized mutant allocation which reflects the reality of the yeast deletion set. Real data are used to construct a fictitious population of possible phenotypes. They arise from a single agar plate, with 1536 yeast colonies. Figure 4.1 shows a smoothed histogram of this population.

For sake of simplicity, only two categories of data are considered: healthy and unhealthy. The cut-off value of 7500 was used to distinguish healthy mutants (those with phenotypes greater than 7500) and unhealthy mutants. It is based on a rough visual estimate which assumes the two left modes are composed of phenotypes from unhealthy mutants. According to this cut-off, approximately 30% of the mutants are unhealthy.

Both plate-sets consist of two plates, each containing 300 mutants. In the random plate-set, plates are simply populated by sampling with replacement from the population, thus representing random allocation. The pathological plate-set simulates the allocation often seen in practice. One plate receives 25% of its mutants from the unhealthy subpopulation and the other plate recieves 50% of its mutants from that subpopulation. Figure 4.2 displays particular results using both sampling schemes. These represent plate-sets which have not been distorted by the plate effect. They are the data which we desire to recover after pre-processing.

The simulation consisted of 500 iterations. In each iteration, an instance of both plate-sets was populated. Within each set, for a range of quantiles, the observed quantile value was recorded for plate 1 and plate 2. The results are depicted in figure 4.3 which are fancy scatter plots. They show the 500 observed quantile differences at each quantile using a smoothed color density representation. The yellow lines are the smoothed mean differences.

The random plate-set suggests that most quantiles are approximately equal when mutants are randomly allocated to plates. The only exception being quantiles in the extreme tails, as these contain outliers. Stability in the lower tail persists to more extreme quantiles than in the upper tail because outliers in the lower tail are solely due to error. The results presented in figure 4.1 don't suggest large discrepancies in the lower tail, but the smallest measured quantile is at 1%. Caution with regards to outliers should be taken at both ends. Here the quantiles in both tails which are not too extreme align just as well as the median. In fact some of them are better behaved than the median (e.g. the 85% quantile). In general, under random allocation the medians differ less across plates than do the extreme quantiles, but the difference tends to be small.

The pathological plate-set tells a different story. Here the median differs by a substantial amount across the set - it is the most unequal quantile. In fact, their values correspond to phenotypes of different biological categories: healthy and unhealthy. On the other hand, the quantiles in the tails seem to be much less affected by the disproportionality of healthy and unhealthy mutants across the set, specifically, quantiles roughly between 1% and 5% and between 85% and 95%. Define the tails of the distribution which are not too extreme to be affected by outliers as the almost-tails. The simulation suggests that quantiles in the almost-tails will be approximately equal across the set and that this is robust with regards to departures from randomization of the mutants.

One expects the range of underlying phenotypes on each plate to be roughly the same. However range may be severely altered by outliers. Thus, the idea is to use a compromise between range and center with more emphasis placed on range. Therefore, the middle ground between assuming the equality of medians and assuming equality of all quantiles is to assume equality of quantiles in the almost-tails. Quantiles in the almost-tails of phenotype distributions are internal features which are appropriate for HTP experiments due to their robustness to modest departures in randomization.

Another example of internal assumptions is the approximate equality of the modes in the phenotype distributions. In particular, the left modes, when these exist, often correspond to phenotypes of blanks and nonviable mutants, so their alignment would make much sense. It isn't clear whether other modes are approximately equal. The upside to aligning the left modes is that they are the least affected by error and hence have the smallest variance. The measurement errors are believed to be more than a simple mean zero additive error, but blanks and nonviable mutants are only affected by this error. Ultimately, it is a simple task to compute quantiles, while it is often not simple to identify modes of a distribution. Therefore, we place a heavy emphasis on quantiles in our pre-processing approach.

There exists cases where approximate equality of quantiles in the lower tail may be questionable. In cases where the left mode corresponds to non-viable mutants and blanks only, the discrepancy between phenotypes of viable and non-viable mutants may be large. This, in turn, may reduce the region of lower quantiles over which approximate equality across the plate-set holds and moreover, very much increase the difference between lower quantiles which are not in the region of approximate equality. An example of such a scenario will be expanded upon in the next chapter. As an alternative, rather than assume the approximate equality of the quantiles in the lower tail, assume the approximate equality of the average phenotype up to a given quantile. The quantile becomes an upper bound for the values which are averaged over. This is analogous to mean centering, but restricted to the left tail. Essentially, the claim is that the expected value of the lowest $r\%$ of the phenotype distributions on each plate across the set should be approximately equal for value of r that are small (but not vanishly small). In Huber *et al*'s work, mean centering was carried out using trimmed data. Here we suggest the opposite, to align the mean of the "trimmings" in the left tail. Hence this internal feature is called the mean of left tail trimmings. Figure 4.4 depicts the observed differences of the mean left tail trimmings on both plate-sets and compares the results with those of quantiles.

There do not seem to be any differences in the approximate equality of these internal features within the random plate-set context. However, there is an important difference observed in the pathological plate-set. The approximate equality of the mean left tail trimmings spans a much larger range of quantiles than does the approximate equality of the quantiles themselves. The instability due to non-randomized allocation is simply delayed by averaging over the trimmings. The mean up to a given quantile is always available and the region of approximate equality continues to be the almost-tails. Thus, mean left tail trimmings preserve the simplicity found in quantiles and will be favored as the internal feature of choice in the left tail. The choice of alignment features will be revisited in the next chapter.

4.2 Proposed Pre-processing Method

Pre-processing involves transforming the raw data to remove systematic experimental artefacts. Having specified a model for the plate effect, a natural transformation arises,

$$f_p(x) = \frac{x - \alpha_p}{\beta_p} \tag{4.1}$$

where α_p and β_p are as in (3.1). Thus far we've referred to α_p and β_p as the additive and multiplicative components, respectively, of the plate effect. Within the context of the transformation in (4.1), they will, respectively, be referred to as the location and scale parameters. The novelty of Huber *et al*'s normalization method which we wish to incorporate into our method is the simultaneous estimation of both the location and scale parameters based on biological assumptions. The assumption proposed is the approximate equality of select features of the data. Two classes of such features which respect the underlying biology of the experiments have been established. Using this transformation on a plate-set containing P plates leads to $2P$ unknown parameters. Simply aligning one feature across the plate-set will not be sufficient to solve for the unknown parameters. The alignment of two or more features, though, does allow for simultaneous parameter estimation. However, the estimation method will differ if only two features are aligned rather than three or more. These two cases will be considered separately.

In order to avoid over-parameterization, constraints need to be set. There are a few natural choices here. One plate may be set as a reference plate. Another option is to constrain the quantiles to equal their averages which is analogous to what is done in quantile normalization. In the end, the choice of constraints is not critical. Results obtained using one constraint may be linearly transformed to conform to another constraint using the very same slope and intercept for each datum. Using a reference plate introduces artefacts with respect to plate-sets under other conditions, but pre-processing across conditions will remove these. By setting restraints, the transformation does not use proper estimations of α_p and β_p as suggested in (4.1). Therefore, we reformulate the transformation as $f_p(x) = \frac{x-a_p}{b_p}$. Without loss of generality, set $a_1 = 0$ and $b_1 = 1$ (*i.e.* set plate 1 as the reference plate). Hence, $a_p = \alpha_p - \alpha_1$ and there remains $2P - 2$ unknown parameters.

4.2.1 Exact alignment of 2 features under the linear model

We begin with the case of two features. Let q_{jp} denote feature j on plate p. Having set plate 1 as a reference plate the following $2P - 2$ equations may be formulated based on the application of the transformation in (4.1):

$$f_1(q_{j1}) = f_p(q_{jp}) \qquad \text{for } j = 1,2 \qquad \forall p = 2, ..., P. \tag{4.2}$$

There are $2P - 2$ equations and an equal amount of unknowns, therefore an exact solution exists. In fact, when using transformation (4.1), a direct solution of the unknown parameters is only possible when aligning two features. For each individual plate, simple algebra may be used to solve for its location and scale parameters.

$$a_p = \frac{q_{2p}q_{11} - q_{21}q_{1p}}{q_{11} - q_{21}} \tag{4.3}$$

$$b_p = \frac{q_{1p}}{q_{11}} - \frac{q_{2p}q_{11} - q_{21}q_{1p}}{q_{11}(q_{11} - q_{21})} \tag{4.4}$$

Applying the transformation in equation (4.1) using estimates as described in (4.3) and (4.4) will force features 1 and 2 one each plate to be equal to those of the reference plate.

4.2.2 Approximate alignment of more than two features through optimization

When there are more alignment features than there are parameters, a direct solution for the unknown parameters is not available. In such circumstances, under this distortion model, the ability to align all selected features exactly across the plate-set is lost. Alternatively, optimization may be used to estimate the parameters. Suppose there are Q features to align. Let $\mathbf{a} = (0, a_2, ..., a_P)$ be the vector of location parameters, $\mathbf{b} = (1, b_2, ..., b_P)$ be the vector of scale parameters and $\mathbf{q}_j = (q_{j1}, q_{j2}, ..., q_{jP})$ be the vectors of features for $j = 1, ..., Q$. Without the possibility of an exact alignment of the features, the next best solution is to minimize their variance across the set. This leads to the following objective function:

$$\underset{\mathbf{a},\mathbf{b}}{\operatorname{argmin}}\; g(\mathbf{a},\mathbf{b}) = \sum_{j=1}^{Q} var\Big(\frac{\mathbf{q}_j - \mathbf{a}}{\mathbf{b}}\Big) \tag{4.5}$$

Various optimization algorithms may be implemented to minimize the objective function and obtain the parameter estimates of interest. The Newton-Raphson method requires the inversion of the Hessian matrix which is problematic as it often becomes singular. Therefore, it is recommended to use optimization methods which only use the exact calculation of first derivatives. There are two classes of such methods: conjugate gradient methods and quasi-Newton methods. They differ in their strategy to numerically estimate the Hessian. Conjugate gradient methods require less storage than quasi-Newton methods require, but the latter are generally less fragile [7]. Therefore, since these will never be large problems, it is recommended to use quasi-Newton methods in general, though both methods are appropriate.

Regardless of the class of method chosen, the gradient is required for their proper implementation. It is a vector of length $2P-2$ with $P-1$ entries which are the partial derivatives of the objective function $g(\mathbf{a},\mathbf{b})$ with respect to each location parameter (excluding $a_1 = 0$) and $P-1$ entries which are the partial derivatives with respect to each scale parameter (excluding $b_1 = 1$). Let $\bar{q}_j = \frac{1}{P}\sum_{p=1}^{P} f_p(q_{jp})$ be the average j^{th} feature across the the transformed plate-set, then the following equations may be used to construct the gradient.

$$\frac{\partial}{\partial a_k} g(\mathbf{a},\mathbf{b}) = \frac{-2}{b_k(P-1)} \sum_{j=1}^{Q} (f_k(q_{jk}) - \bar{q}_j) \qquad \text{for } k = 2,3,...,P \tag{4.6}$$

$$\frac{\partial}{\partial b_k} g(\mathbf{a},\mathbf{b}) = \frac{-2}{b_k^2(P-1)} \sum_{j=1}^{Q} (f_k(q_{jk}) - \bar{q}_j)(q_{jk} - a_k) \qquad \text{for } k = 2,3,...,P \tag{4.7}$$

The complete derivation of these results are given in Appendix A.

Once the optimization method has been selected and the first derivatives specified, the last requirement is a choice of initial parameters. Initial parameters which are too far from the minimum may lead the algorithm to converge to a local minimum or to travel in the wrong direction and fail to converge within the specified number of iterations. Furthermore, the choice of initial parameters will affect the number of iterations required to converge. There are various ways to obtain good initial parameters. One way is to first use the difference between the minimum observation on each plate and that of the reference plate to initiate the location parameters: $a_{p0} = min(\mathbf{x}_p) - min(\mathbf{x}_1)$. Here, $\mathbf{x}_p$ is the vector of phenotypes on plate p. Then initiate the scale parameters by taking the ratio of the range of each plate over that of the reference plate. In other words, first subtract the obtained location parameter guesses from each plate: $y_{gp} = x_{gp} - a_{p0}$. This is analogous to the removal of the additive effect discussed in the current methods section. Then initiate the scale parameter as $b_{p0} = \frac{max(\mathbf{y}_p) - min(\mathbf{y}_p)}{max(\mathbf{y}_1) - min(\mathbf{y}_1)}$.

4.2.3 Exact alignment of more than two features using a piecewise distortion model

Within the context of aligning three or more features, there is an alternative to optimization. A simple extension to the distortion model allows for exact alignment of the features. The distortion model could be described as piecewise linear. Perhaps the plate effect is not exactly the same at different magnitudes of the phenotype. Using linear spline interpolation with knots at the features will allow these to be equal across the set, rather than approximately equal. Also, different slopes and intercepts between each feature would help account for any change in plate effect due to phenotype. Figure 4.5 depicts the difference between the linear transformation as proposed in (4.1) and linear spline interpolation.

The transformation called upon by linear splines is as follows. Suppose there are Q features. Then there are Q knots (x_i, y_i) which the transformation is to pass through. The linear function between each of these knots is:

$$S_i = y_i + \frac{y_{i+1} - y_i}{x_{i+1} - x_i}(x - x_i) \qquad x \in [x_i, x_{i+1}]. \tag{4.8}$$

Thus the function is only defined for phenotypes between the selected features. The linear mappings between the first and second features and between the two upper features may be extended to map all observations. Alternatively, all values below x_1 may be mapped to y_1 and all those above x_Q may be mapped to y_Q.

Constraints are once again called for. The candidates are the same as before, either align to a reference plate or align to the average value of each feature. In this case, the choice cannot be changed afterward using a global linear transformation, therefore the choice of constraint will have a heavier impact on the results. For this reason, it seems more reasonable to map each feature to the average value of that feature rather than to the value observed on a reference plate.

There are two advantages to using linear splines over using linear transformations and optimization. First solving the latter is more computationally expensive: it requires more memory and time. Secondly, modest departures from linearity in the plate effect will be compensated for using this approach. Hence, it offers more flexibility.

The method we propose uses biological assumptions to simultaneously estimate the additive and multiplicative components of the plate effect. The assumptions used to achieve this are the approximate equality of two or more features. The approximate equality of quantiles in the almost-tails is an example of internal assumptions which are robust to departures from random mutant allocation to plates. Quantiles are not the only features which can be used. For aligning more than two features, two methods have been proposed. They are both valid and have their pros and cons. A more in depth comparison of the two methods will be conducted in the next chapter.

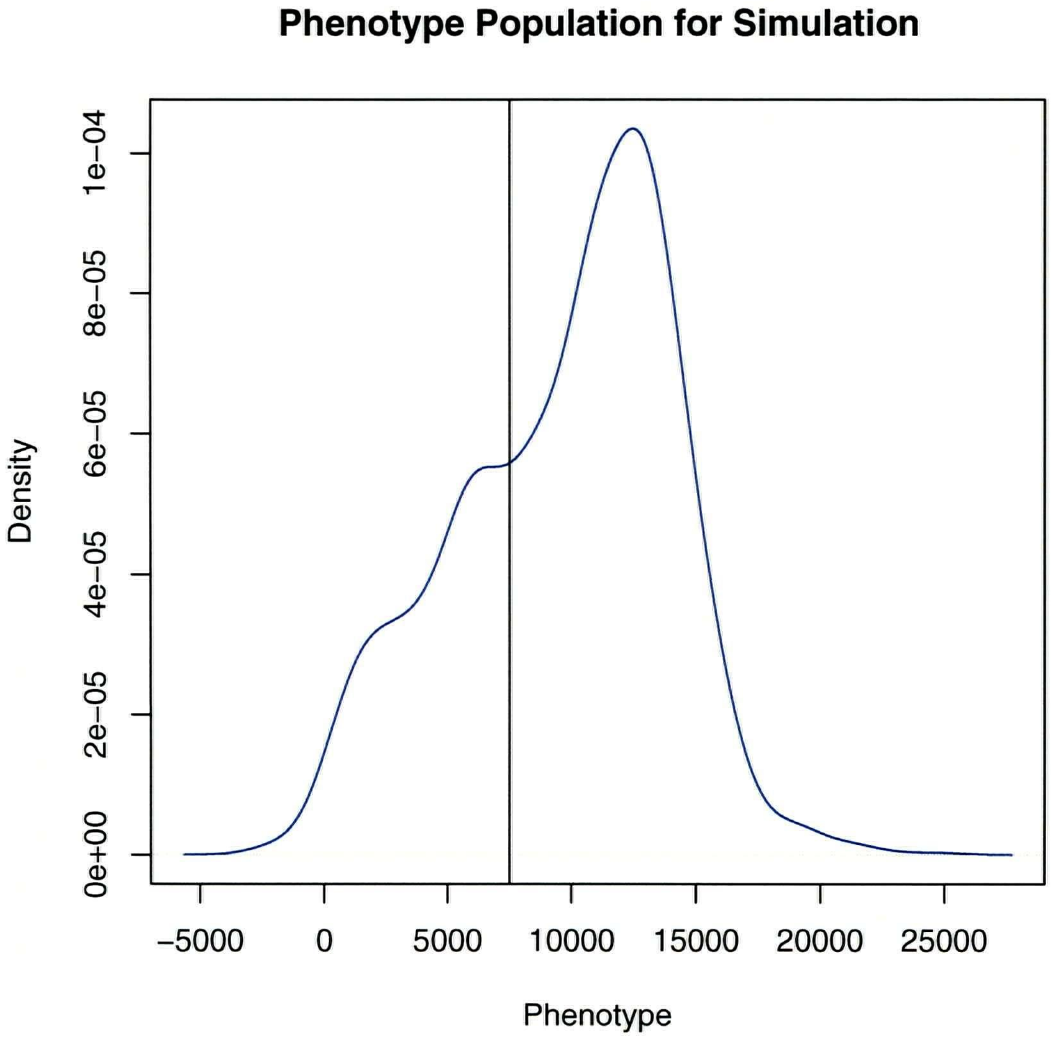


Figure 4.1: The approximate density of the phenotype population used for the simulation. The cut-off is depicted by the black line at 7500. The heavy left tail suggests that a fair number of mutants are affected by the gene perturbation, which this cut-off value tries to capture.

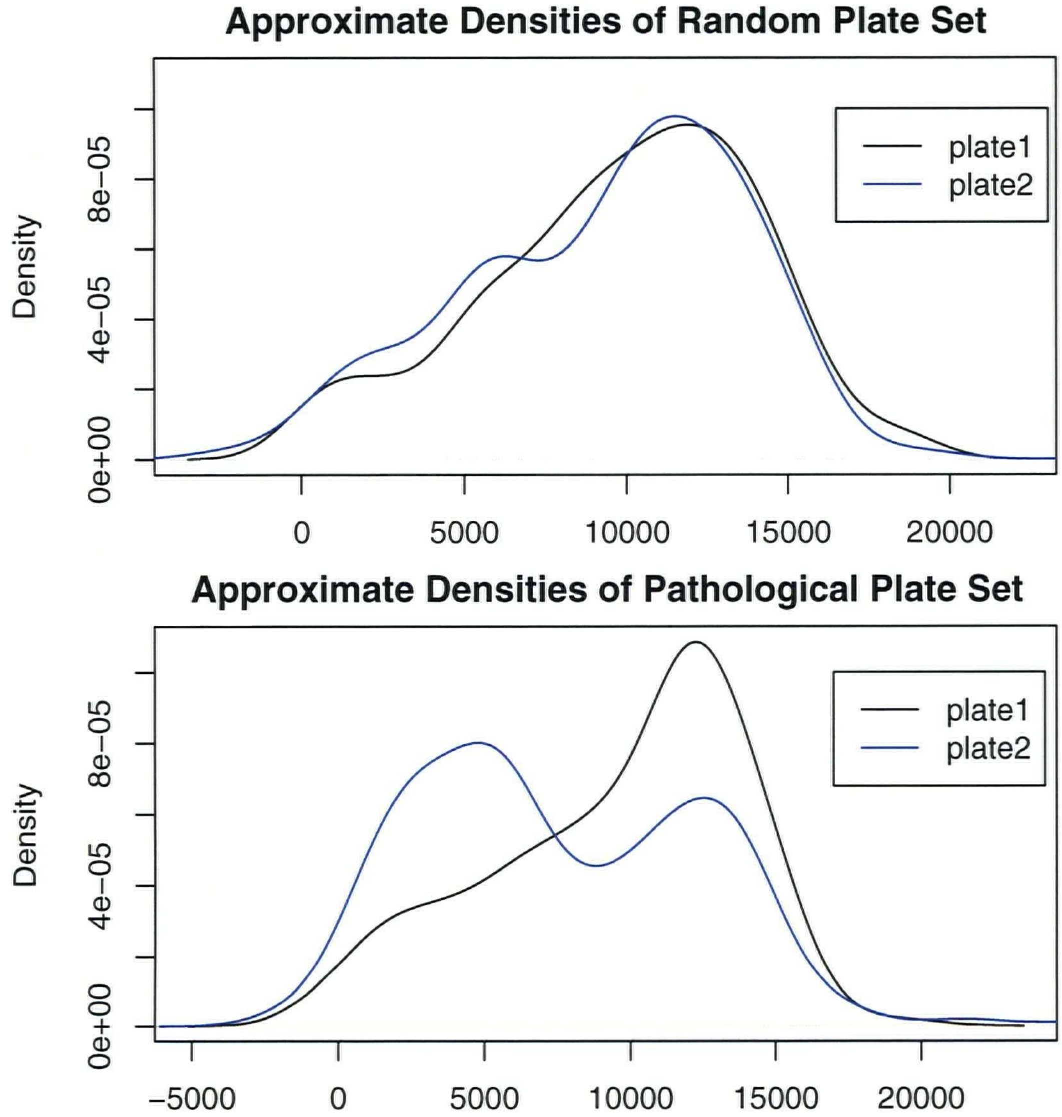


Figure 4.2: Plates generated in a random fashion still present different phenotype distributions since the mutant collections on each plate are different. However this difference is minimal compared to that which arises when a large collection of weak mutants are purposely contained on one plate.

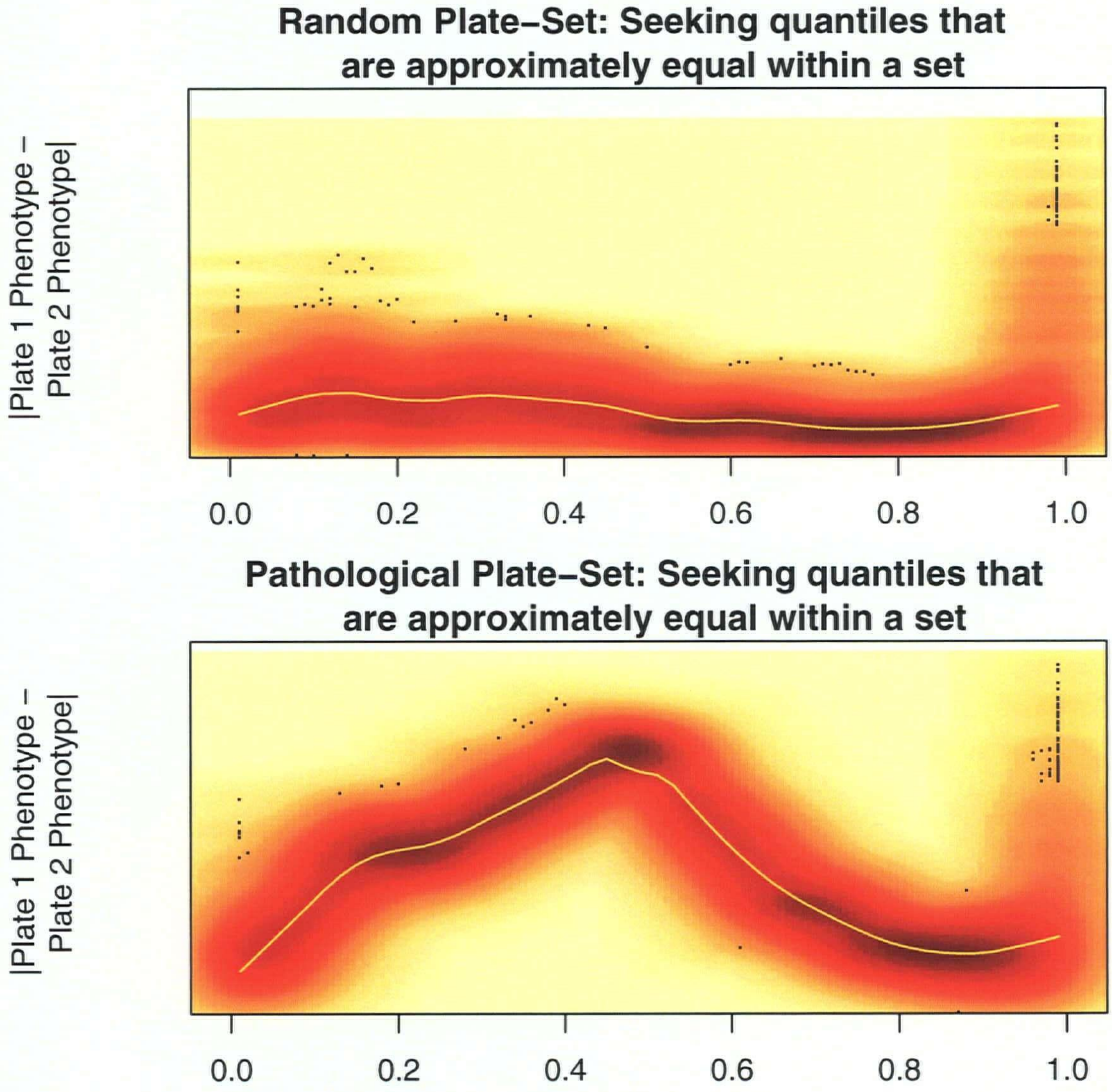


Figure 4.3: The actual values are not shown in the y-axis, but the limits are kept constant for both graphs and both start at 0.

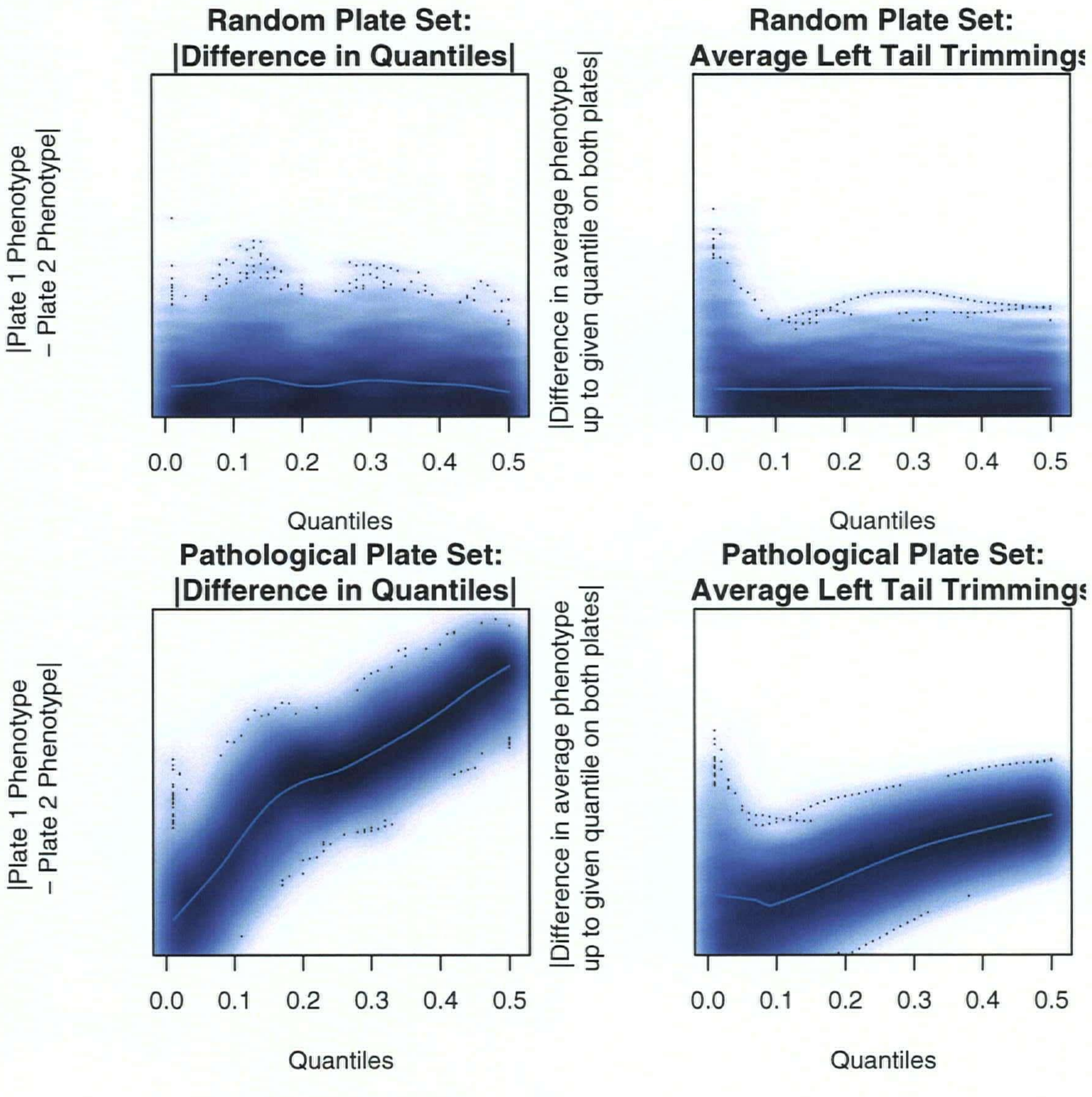


Figure 4.4: Again the range of the y-axis is kept constant and all start at 0.

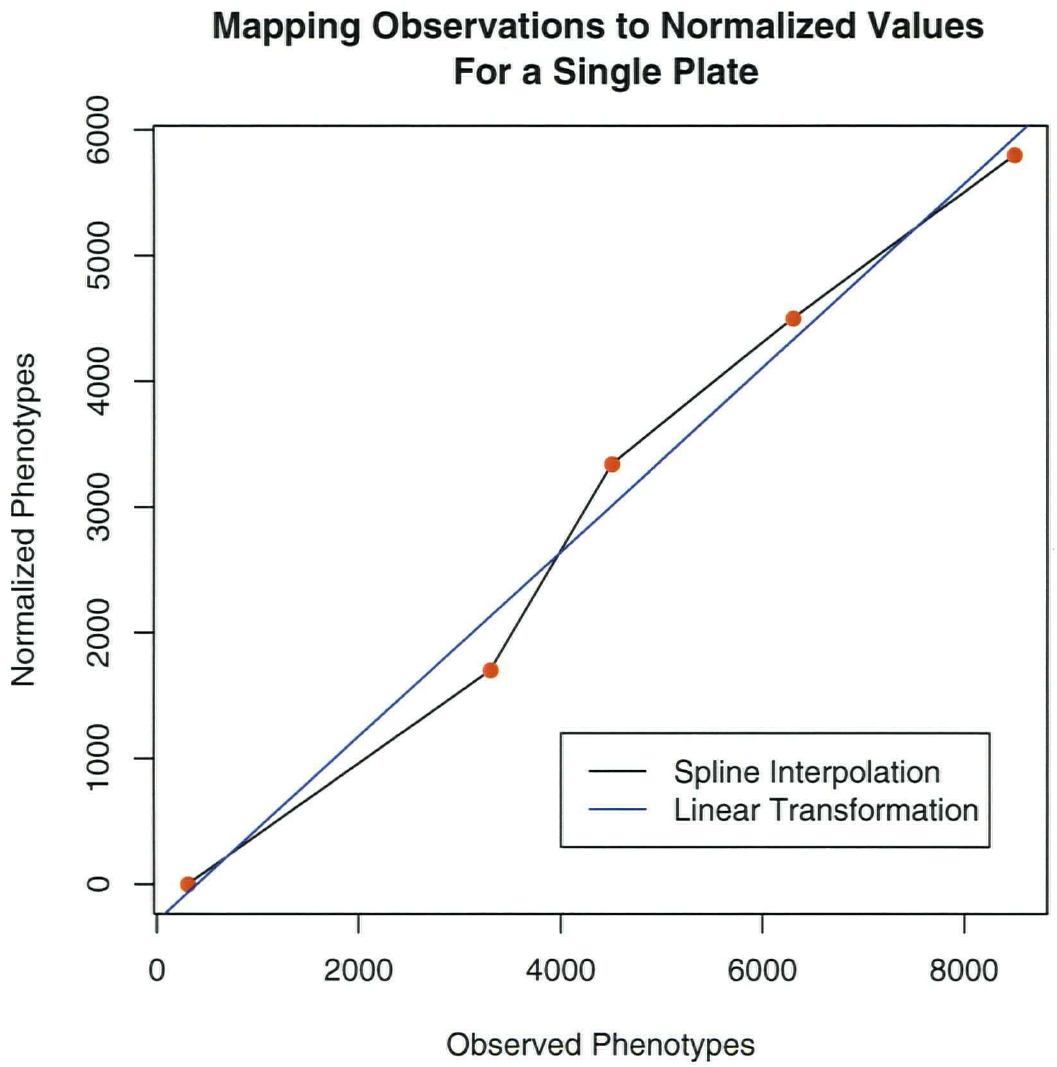


Figure 4.5: The red dots are ordinal pairs composed of the observed features and the values they should be mapped to. Linear spline interpolation allows for these features to be mapped exactly to the desired location. A linear transformation cannot do this, so the line which minimizes the distance between the mapped features and their desired value is chosen instead.

Chapter 5

Application to Real Data

In this chapter, our recommended pre-processing method will be applied to real data. The many knobs resulting from the methods flexibility, in both the class of features to be aligned and the alignment method, will be explored. To properly do so requires an understanding of the data used throughout the chapter.

5.1 Data

The data arise from growth assays using yeast deletion sets and are provided by the Conibear lab. Each plate-set is composed of 14 agar plates. Each agar plate contains 1536 yeast colonies. There are 384 distinct mutants per plate with four technical replicates of each (384 x 4 = 1536). Of the 384 mutants, only five are not deletion mutants. One is a blank and four are controls. The controls are yeasts which have different growth rates expected to cover the spectrum of possible phenotypes. The location effect is quite strong in this data. Appendix B describes the nature of the edge effect along with the interim solution. It is worth noting because the controls in this experiment are near the edge.

The mutants are not randomly allocated to plates. On twelve of the fourteen plates, yeast colonies are placed on the agar plates according to their location on purchased plates which is roughly based on their order on the chromosome. The remaining two plates contain mutants of interest (many of them of the weak variety or susceptible to food competition) with additional blanks.

5.2 Recommended Method versus Median Centering

We begin by comparing the old method of median centering to the recommended method to see if there is any apparent improvement. Figure 5.1 contains three plots and, for simplicity, only 4 of the 14 plates are shown. The first plot depicts smoothed histograms of observed raw phenotypes on the four selected plates. It shows the presence of a plate effect as well as the presence of non-random mutant allocation (as seen by a much greater proportion of mutants in the left mode of one plate). The plate whose density is depicted by the blue line, plate 13, is one of two plates which contains mutants of interest as described above.

The second plot in this figure depicts smoothed histograms of preprocessed phenotypes obtained by median centering. Specifically, the additive effect was first removed by subtracting from each observation its plate's minimal phenotype: Then the resulting adjusted phenotypes were divided by the plate medians. There appears to be an improvement over the raw data. The plate depicted in green is now aligned with the others, but there still appears to be issues with plate 13 (blue line). The majority of its non-blanks have a pre-processed phenotype larger than those reported on the other three plates resulting in a range which is one and half times larger than that of the other three plates. Another problem is the misalignment of the left modes which should align as they pertain to blanks and nonviable mutants.

The third plot depicts results from aligning both the blanks and the 90% quantiles. Since only two features were aligned, an exact solution was available, so there was no need to choose between linear splines and optimization. Visually the results are quite pleasing. Firstly, the non-blanks on plate 13 have adjusted phenotypes within the range of the other plates and, secondly, the left modes align well.

These are real data, so it is impossible to evaluate these methods based on the true plate effects as they are unknown. However, the presence of blanks and controls on all plates can be used as an assessment tool to complement visual inspection of the smoothed phenotype histograms. Figure 5.2 contains three diagnostic plots to complement the previous figure. Each plot displays the median phenotypes of the four controls and blanks on the selected plates. Controls one and two have four technical replicates and controls three and four have two (due to the edge effect). This allows for an analysis of variance to be carried out. Unfortunately, the blanks are on corners, so they don't have technical replicates. The p-value in the title of each plot corresponds to the existence of a plate effect in a two way analysis of variance (controls and plates as explanatory variables with no interactions). The phenotypes were transformed in order to achieve homoscedasticity. Looking at these p-values, it is apparent that the recommended method has removed much more of the plate effect.

The diagnostic plot for median centering supports the argument above: having a range which is one and half times larger than the range of other plates is problematic. This can be seen by the normalized values of controls 1, 3 and 4 of plate 13, with control 3 being of the chart. The diagnostic plot for the recommended method of feature alignment shows a clear improvement on both the raw data and median centering. The blanks, of course, are perfectly aligned and the controls are better behaved. It is apparent on this plot that no transformation would allow both the blanks and control 2 to be perfectly aligned. Results would have been different had we chosen to align control 2 instead of blanks. This begs the question: which features should we align?

5.3 Choice of Feature

5.3.1 Observing internal feature behaviour induced through the alignment of external features

In chapter 4, two classes of features to align across plates were defined: internal and external. By design, external features - controls and blanks - should be approximately equal across all plates. Thus, the natural question to ask is whether the alignment of external features will induce approximate equality of quantiles. To complement chapter 4's simulation, the alignment of external features is applied to real data and the resulting quantiles and mean left tail trimmings on each plate are compared in similar fashion. This particular data-set does not have severe problems in proportions of healthy and unhealthy mutants as can be seen in the first plot in figure 5.3. It shows the smoothed histograms of the phenotypes on each of its plates after alignment of controls. The distance between the left mode and the remaining phenotypes is large as was discussed in chapter 4.

Dealing with real data, it is not possible to obtain 500 values for each quantile, thus using a smoothed scatter plot as in the simulation is not useful. Instead, the variance of each quantile across the 13 plates is presented. In the top right hand corner plot, the variance of quantiles over the entire range is shown. The quantiles in the upper almost-tail align well, but the lower quantiles are problematic as was stated in chapter 4. There is a small region over which the quantiles don't differ too much, but quantiles slightly outside this region do not align well at all. The last plot depicts the variance of mean left tail trimmings. The region over which this feature is stable across the set is much larger and the increase in its discrepancy outside that region is much less severe. This further supports the superiority of mean left tail trimmings over quantiles in the lower almost-tail. It also suggests that the alignment of external features also induces the approximate equality of the proposed internal features. In the next section we will further compare internal and external features.

5.3.2 Internal and External Features in a Case Study

Up to now controls and blanks have been used as a means of evaluation. Controls can equally be assessed via blanks and internal features. Consider two data sets A and B. Figures 5.4 and 5.5 show the results of pre-processing each data set by minimizing the variance of the four controls across plates via a linear transformation. Each figure contains three sections. The first gives the settings used for pre-processing and the p-value resulting from a kruskal wallis test for the presence of a plate effect. The second shows the smoothed histograms of the pre-processed data and the third shows the corresponding values of select features for each plate. The alignment is not perfect as is readily observable, but real data is always noisy. In the first figure, the p-values do not suggest that there is a plate effect. Furthermore, the upper quantiles align with about equal variance to controls 1 and 3, and the two smaller mean left tail trimmings align with approximately the same variance as the upper quantiles if we ignore the one plate which seems problematic for the controls as well. This is encouraging and supports our claims. The second figure suggests that this complimentary behaviour between internal and external features isn't always observed. The low p-value for control 1 suggests heavy contradictions between controls 1 and 3. In the diagnostic plot (the bottom section), the internal features and the blanks do not seem approximately equal with the alignment of controls. The results from the alignment of controls in this scenario are not desirable.

Similarly to the results of median centering shown earlier, the pre-processed phenotypes from some plates are almost all higher than those from others. Though central behaviour of plate distributions are expected to differ, the biology does not dictate such a severe difference. Hence, figure 5.5 doesn't contradict the approximate equality of internal features via the alignment of controls as much as it questions the reliability of controls. Inevitably, having only 4 or 2 replicates of each control may give way to substantial variance and considerably limit the usefulness of the controls.

There are benefits to be had by approaching the problem the other way round by aligning internal features and observing the resulting alignment of controls. Figure 5.6 considers Set A. Again the results of pre-processing, this time with internal features, are visually pleasing. The controls seem for the most part to be well behaved, although not enough to suggest that there is no plate effect (see kruskal-wallis p-values). Figure 5.7 depicts the pre-processing of Set B using internal features. Not surprisingly the controls in turn do not align well, but the smoothed histograms suggest superior results here. The plates are well aligned as are the left modes. Furthermore, the blanks are also approximately equal. The controls in the same range of phenotypes as the blanks do not align well at all. What distinguishes the blanks from the controls is that they are only affected by additive effects and errors. Thus they are more reliable than controls when technical replicates are limited.

In this brief case study, controls were effective pre-processing features for Set A, but not for Set B. Simultaneously, internal features were effective pre-processing features in both data sets. The phenotype of each control is based on a limited number of technical replicates. They are susceptible to various sources of error and experimental effects, such as edge effect, which in turn may severely increase their variance. The biology dictates that their values should be approximately equal across the plate-set, but their high variance and low replication count hinders their usefulness here. Quantiles and mean left tail trimmings on the other hand are based on all the plate's phenotypes. This renders them more robust and thus more reliable as pre-processing features. Regardless of which feature is used, quality control should always be carried out as was done in all four cases here.

5.4 Optimization versus increasing the number of parameters in the model

Two different paths, which both stem from the exact alignment of two features, may be taken when aligning more than two features. Either the distortion model can be kept linear, leading to parameter estimation via optimization, or it can be slightly extended allowing for linear splines to be used in favor of a linear transformation. Solely based on the model, it is more desirable to use the simple extension to the distortion model which allows for slight departures from linearity to take place. It is a more flexible approach. On the basis of implementation, linear splines are much less expensive than optimization, both in memory and time. However, when using linear splines, it is important for the knots to be in order so that the resulting transformation is injective. When using controls as features, this is not always possible. For example, on one plate from Set B, control 4 expresses a lower phenotype than control 2, yet the average value of control 4 across the plate-set is higher than the average value of control 2. Therefore, using linear splines with controls may not always be possible. Quantiles and other internal features on the other hand must be properly ordered and lend themselves nicely to linear splines. Lastly, when using splines, there is the danger of over fitting. Figure 5.8 differs from figure 5.7 only in its estimation method: optimization versus linear splines with knots at the features. The p-values do not change significantly suggesting that there is no over-fitting. This seems to hold consistently when using three or four features at a time. Thus, linear splines are the suggested alignment method and optimization of parameters for a linear transformation should be used when linear splines are not tractable.

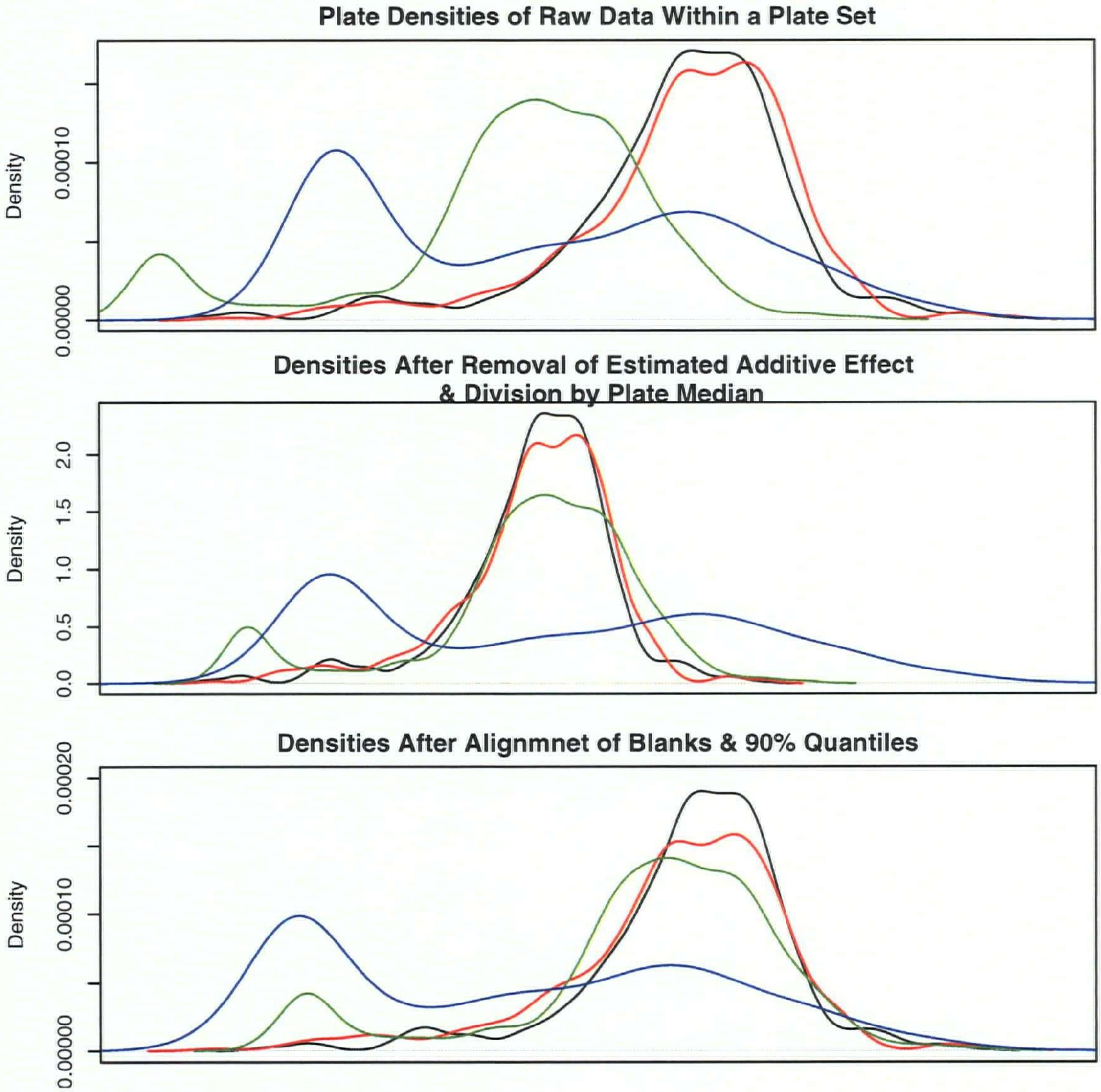


Figure 5.1: For simplicity, we only depict the density of four plates. The plate whose density is depicted by the blue line is one whose mutants were not randomly allocated.

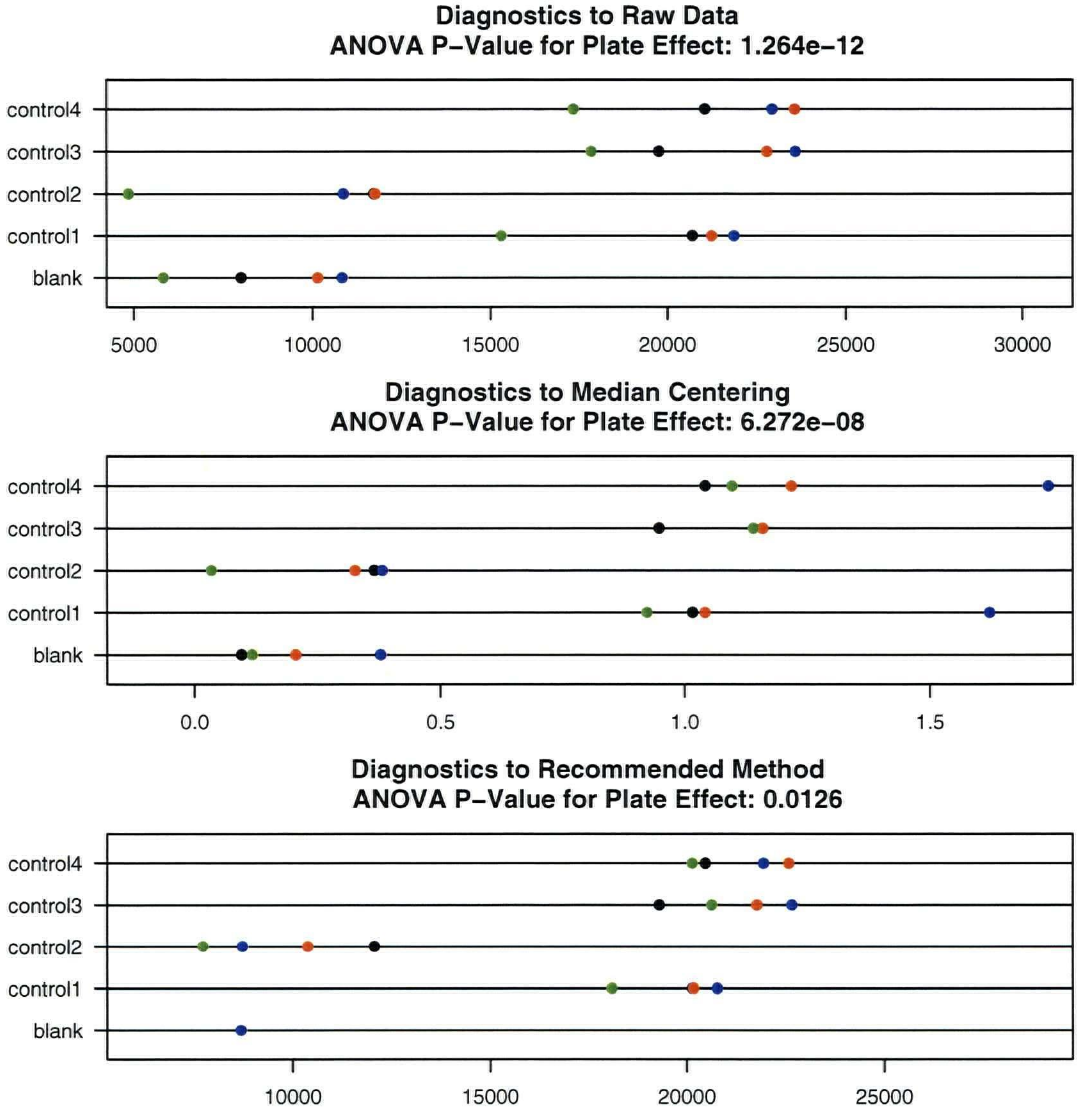


Figure 5.2: These plots compliment the previous ones. They depict the values of controls and blanks across the (reduced) plate-set considered in the two normalization approaches shown in (5.1).

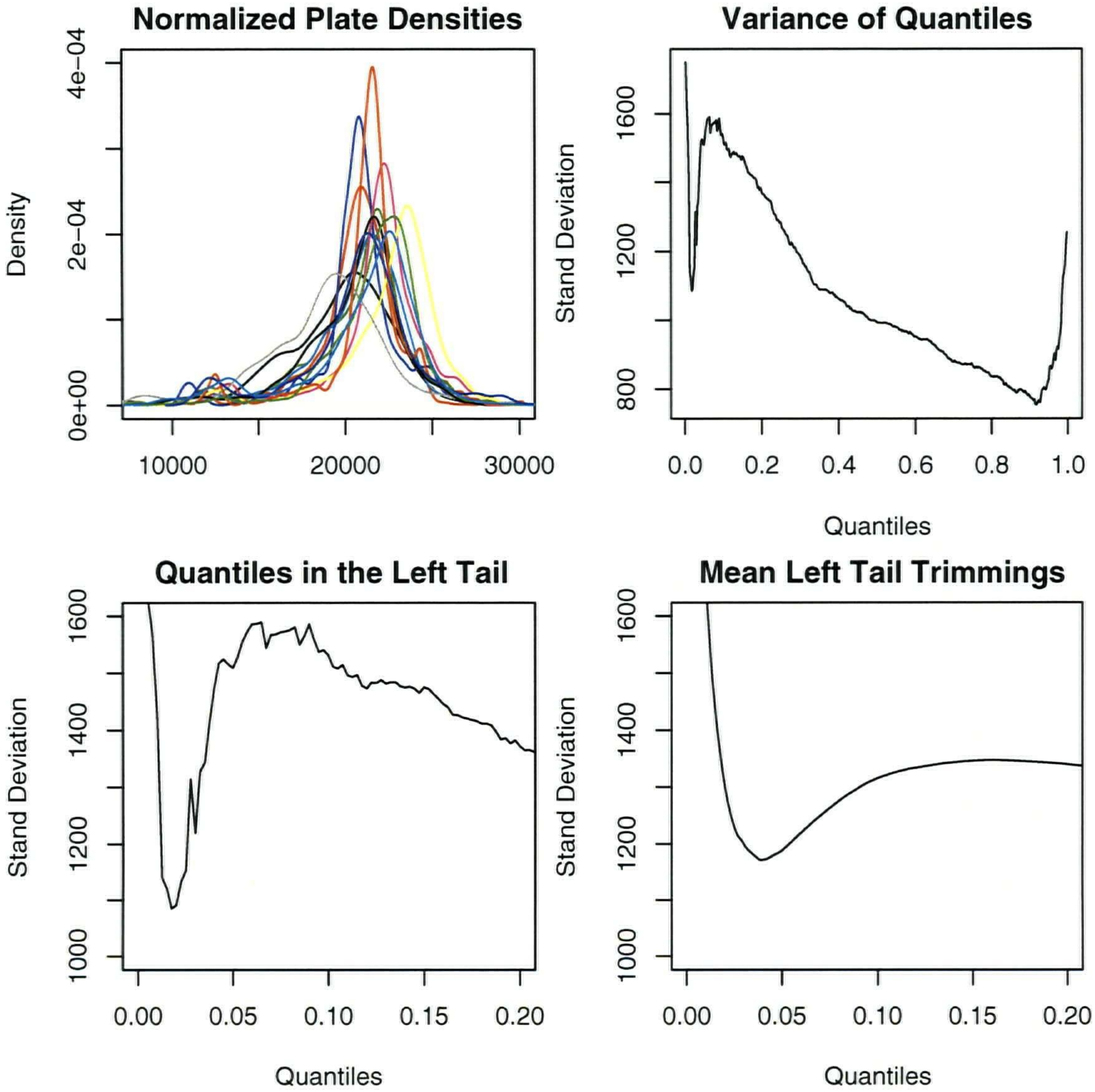

Normalized Plate Densities
Density
0e+00
2e-04
4e-04
10000
20000
30000
Variance of Quantiles
Stand Deviation
800
1200
1600
0.0
0.2
0.4
0.6
0.8
1.0
Quantiles
Quantiles in the Left Tail
Stand Deviation
1000
1200
1400
1600
0.00
0.05
0.10
0.15
0.20
Quantiles
Mean Left Tail Trimmings
Stand Deviation
1000
1200
1400
1600
0.00
0.05
0.10
0.15
0.20
Quantiles

Figure 5.3:

Features Aligned:	control1, control2, control3, control4	**Kruskal P-value for Control 1:**	0.0988
Method of Estimation:	Optimization using Quasi Newton	**Kruskal P-value for Control 2:**	0.1446
Data Set used:	Set A	**Kruskal P-value for Control 3:**	0.3684
Plates Used:	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13	**Kruskal P-value for Control 4:**	0.631
Edge Removed: TRUE	**Background Corrected:** FALSE		

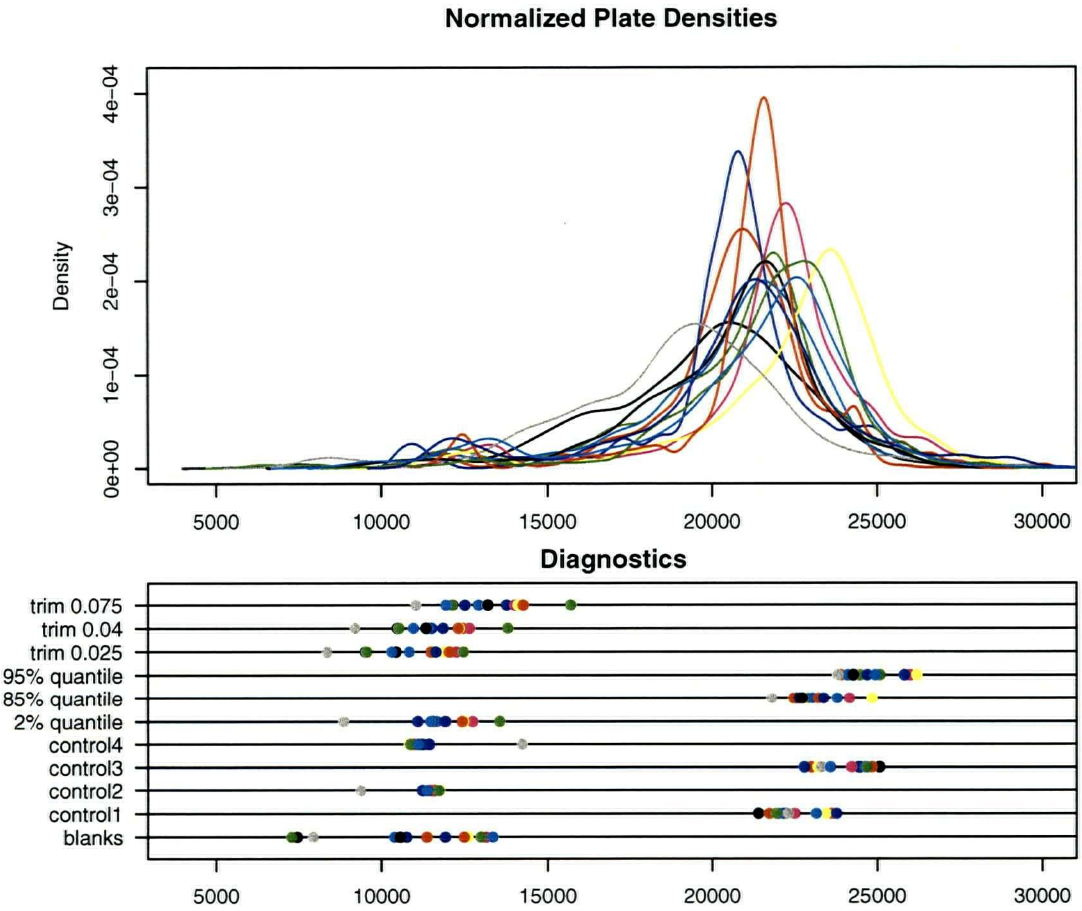


Figure 5.4:

Features Aligned:	control1, control2, control3, control4	**Kruskal P-value for Control 1:**	8e-04
Method of Estimation:	Optimization using Quasi Newton	**Kruskal P-value for Control 2:**	0.1413
Data Set used:	Set B	**Kruskal P-value for Control 3:**	0.291
Plates Used:	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13	**Kruskal P-value for Control 4:**	0.5035
Edge Removed: TRUE	**Background Corrected:** FALSE		

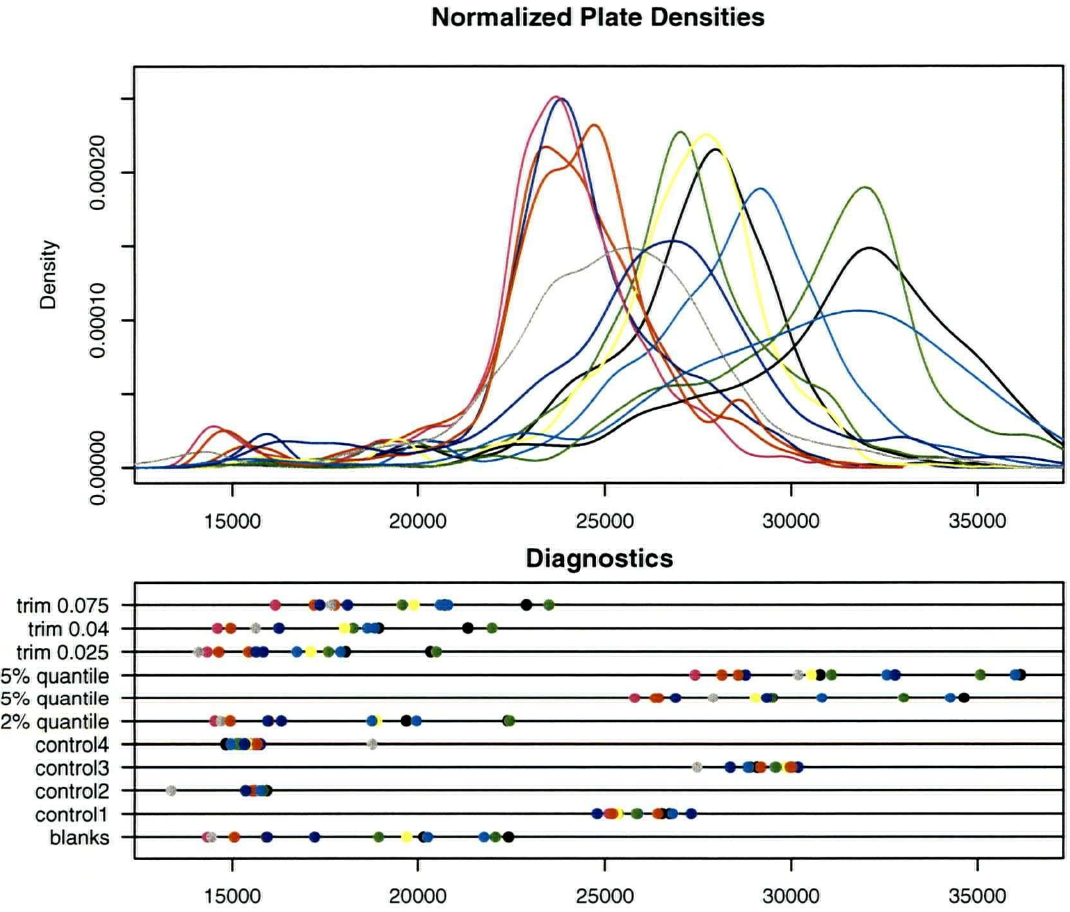


Figure 5.5:

Features Aligned:	Trim 0.025/0.04 and Quantiles 85/95%	**Kruskal P−value for Control 1:**	0.0038
Method of Estimation:	Optimization using Quasi Newton	**Kruskal P−value for Control 2:**	0
Data Set used:	Set A	**Kruskal P−value for Control 3:**	0.1086
Plates Used:	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13	**Kruskal P−value for Control 4:**	0.0476
Edge Removed: TRUE	**Background Corrected:** FALSE		

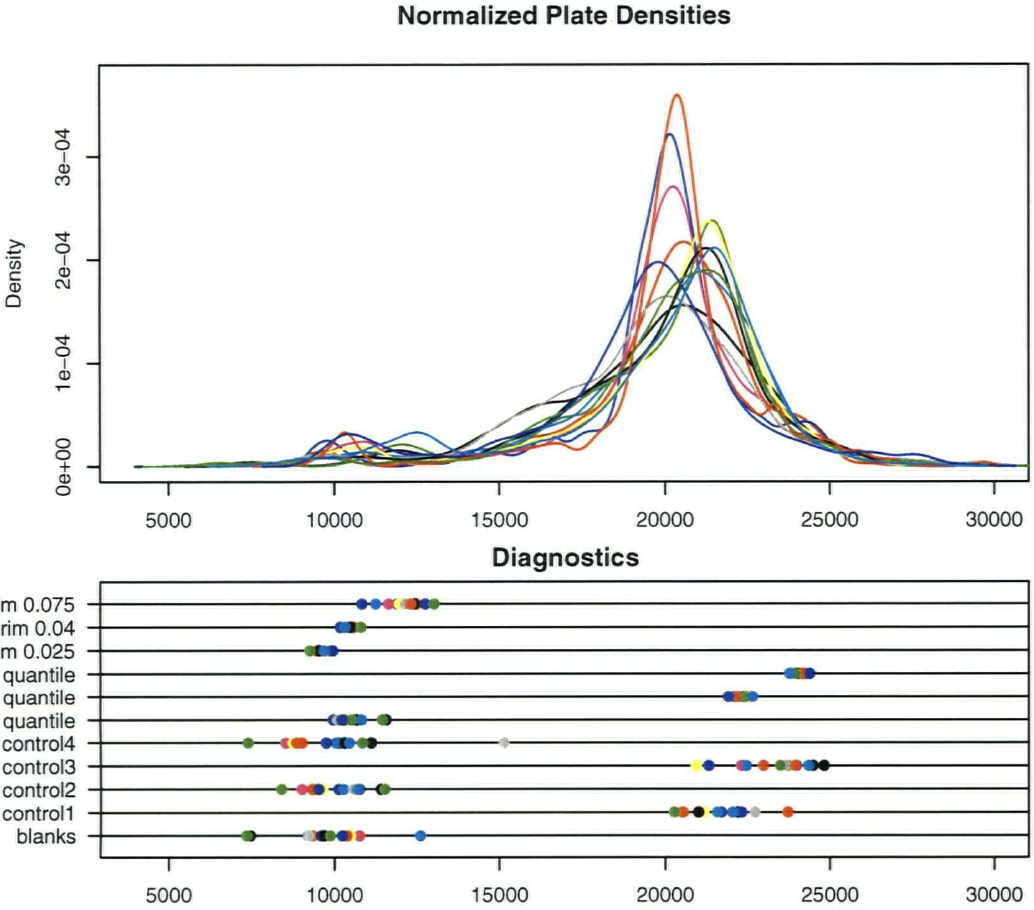


Figure 5.6:

Features Aligned:	Trim 0.025/0.04 and Quantiles 85/95%	**Kruskal P-value for Control 1:**	0
Method of Estimation:	Optimization using Quasi Newton	**Kruskal P-value for Control 2:**	0
Data Set used:	Set B	**Kruskal P-value for Control 3:**	0.0231
Plates Used:	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13	**Kruskal P-value for Control 4:**	0.0198
Edge Removed: TRUE	**Background Corrected:** FALSE		

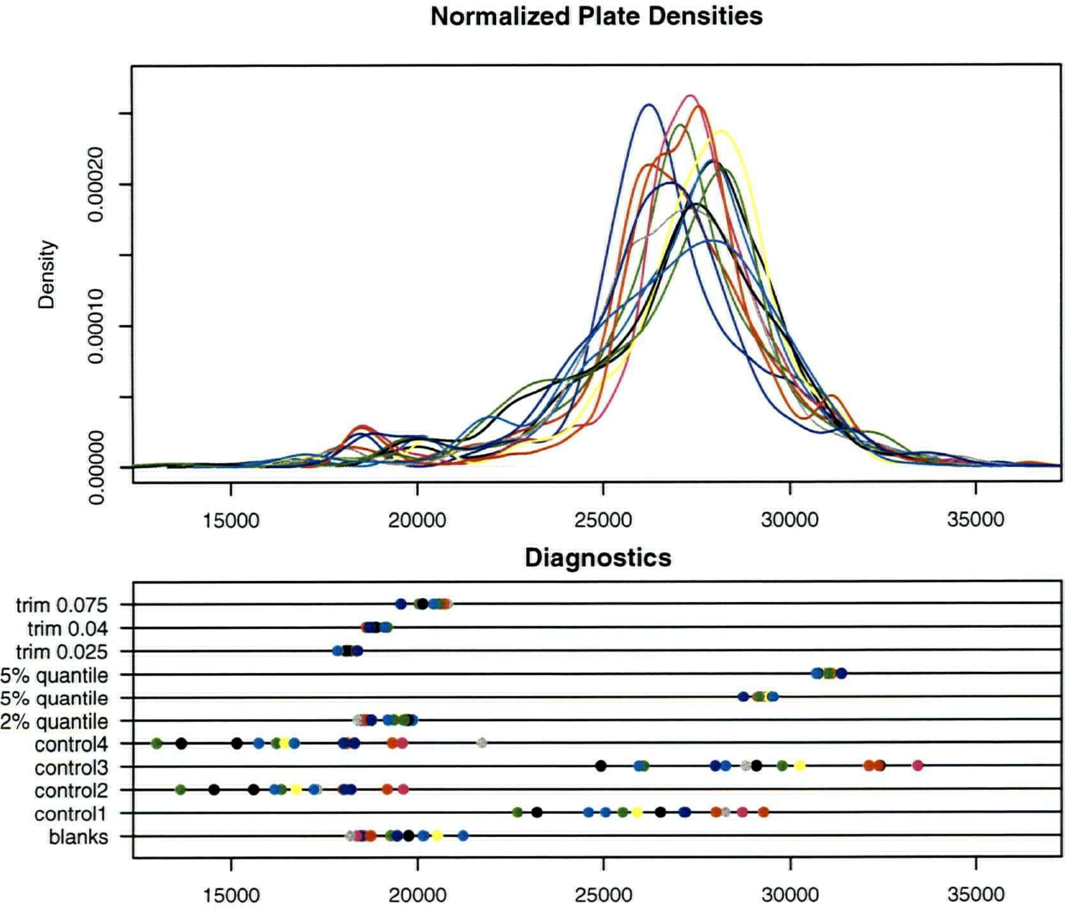


Figure 5.7:

Features Aligned:	Trim 0.025/0.04 and Quantiles 85/95%	Kruskal P-value for Control 1:	0.003
Method of Estimation:	Linear splines with knots @ features	Kruskal P-value for Control 2:	0
Data Set used:	Set A	Kruskal P-value for Control 3:	0.0891
Plates Used:	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13	Kruskal P-value for Control 4:	0.0609
Edge Removed: TRUE	Background Corrected: FALSE		

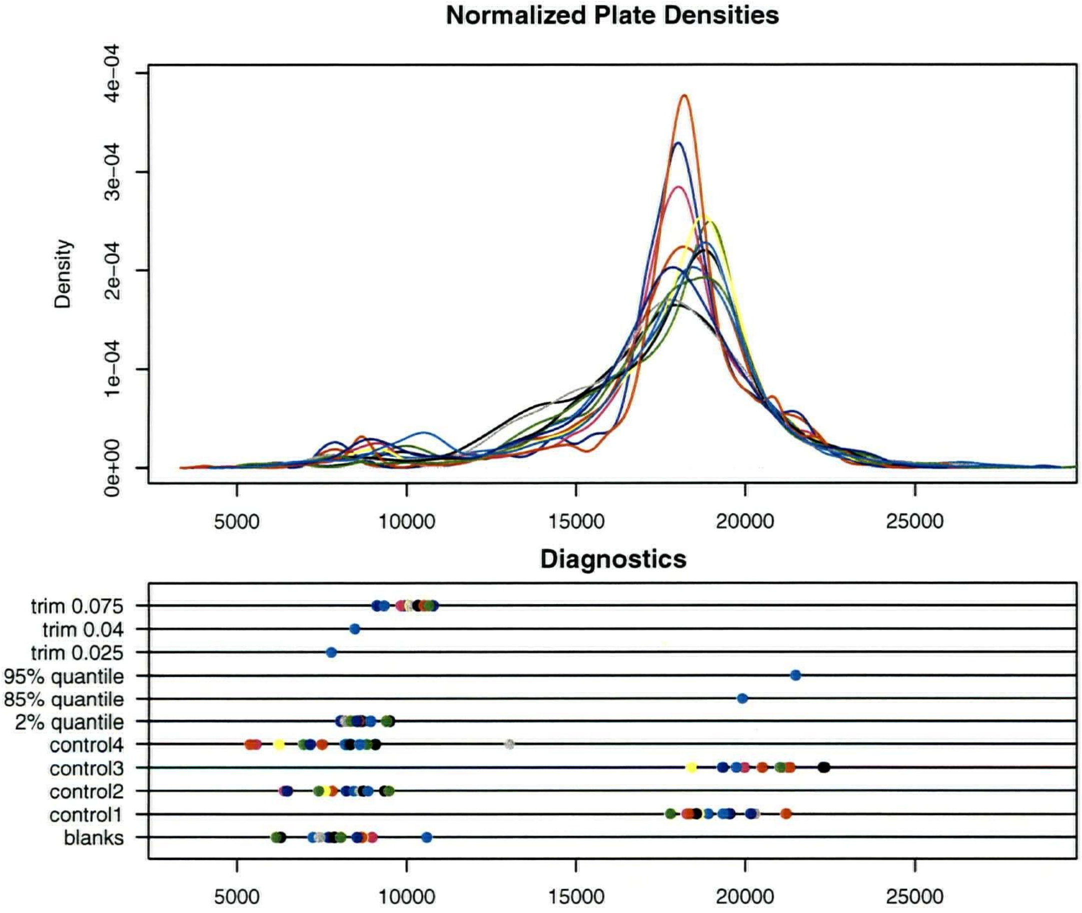


Figure 5.8:

Chapter 6

Conclusion and Future Work

There are many levels at which experimental artefacts distort high throughput phenotypic experiments: location within a plate, plate-wise within a plate-set and across conditions. In this report, we have concentrated on the plate-wise variety. Current methods used for pre-processing were presented and their issues discussed. A class of alternative pre-processing methods which allow for all transformation parameters to be based on biological assumptions were proposed. The biological assumptions required here are the alignment of features across the plate-set. The alignment of two features leads to an exact solution while, otherwise, two routes may be taken. A simple extension to the linear distortion model which increases the number of transformation parameters to allow for linear splines is generally favored over minimization of feature variance across the set. The former has computational advantages as well as the capability to adapt to slight departures from linearity. However, in order to use linear splines, the features must retain the same phenotype order on each plate. Furthermore, in the event that a large number of features are selected for alignment, say five or more, using linear splines may lead to over-fitting. In such situations, optimization should be employed.

We have also discussed the choice of features to align. Two classes were defined: external and internal. While external features should align across the plate-set by design, low replication may lead to limited pre-processing effectiveness as was suggested in section 5.3.2. Internal features, on the other hand, are based on all the data from each plate which saves them from this shortcoming. Often, in practice, mutant are non-randomly allocated to plates which in turn may have a heavy influence on various internal features such as the plate's median phenotype. We have presented the upper almost-tail quantiles and the mean left tail trimmings as internal features which are approximately equal across the plate-set under randomized allocation and robust to modest departures from randomization. These are always available, regardless of the experimental design, and simple to obtain.

There is still room for improvement. It would be desirable to use a model which combines the distortion model with an error model, as was done by Huber *et al* in the microarray context, as a basis for a pre-processing transformation. Consequently, any mean variance relationship resulting from the plate effect interacting with the random errors would be removed using this strategy. Pre-processing across conditions can be achieved using existing methodology developed by the microarray community such as quantile normalization. However, pre-processing with regards to plate location effect is presently done very crudely. Improvement here is the most pressing matter.

Bibliography

[1] et al. A. H. Tong, G. Lesage. Global mapping of the yeast genetic interaction network. *Science*, 303(5659):808–13, 2004.

[2] A.E.Carpenter and D.M.Sabatini. Systematic genome-wide screens of gene function. *Nature Reviews*, 5:11–22, 2004.

[3] S. Armknecht and M. Boutros et al. High-throughput rna interference screens in drosophila tissue culture cells. *Methods Enzymol*, 392:55–73, 2005.

[4] B. L. Drees and V. Thorsson et al. Derivation of genetic interaction networks from quantitative phenotype data. *Genome Biol*, 6(4):R38, 2005.

[5] B.M.Bolstad et al. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–193, 2003.

[6] Robert G. Eason et al. Characterization of synthetic dna bar codes in saccharomyces cerevisiae gene-deletion strains. *Proc Natl Acad Sci U S A.*, 101(30):11046–11051, 2004.

[7] William H. Press et al. *Numerical Recipes in C: The Art of Scientific Computing.* Cambridge University Press, 40 West 20th st., New York, NY 10011-4211, USA, 1988.

[8] Wolfgang Huber et al. Parameter estimation for the calibration and variance stabilization of microarray data. *Statistical Applications in Genetics and Molecular Biology*, 2(1):1–22, 2003.

[9] G. Giaever and A. M. Chu et al. Functional profiling of the saccharomyces cerevisiae genome. *Nature*, 418(6896):387–91, 2002.

[10] David M. Rocke and Blythe Durbin. A model for measurement error for gene expression analysis. *Journal of Computational Biology*, 8:557–569, 2001.

[11] C. Sachse and E. Krausz et al. High-throughput rna interference strategies for target discovery and validation by using synthetic short interfering rnas: functional genomics investigations of biological pathways. *Methods Enzymol*, 392:242–77, 2005.

[12] E. A. Winzeler and D. D. Shoemaker et al. Functional characterization of the s. cerevisiae genome by gene deletion and parallel analysis. *Science*, 285(5429):901–6, 1999.

[13] Luu P. Speed T.P. Yang Y.H., Dudoit S. Normalization for cdna microarray data. Technical report, University of California, Berkeley, 2001.

Appendix A

Derivation of Gradient

The derivative of a sum is the sum of the derivatives, so we begin with a single term from the sum. Let $\bar{q}_j = \frac{1}{P}\sum_{p=1}^{P} f_p(q_{jp})$, then,

$$
\begin{aligned}
\frac{\partial}{\partial a_k} var\Big(\frac{\mathbf{q}_j - \mathbf{a}}{\mathbf{b}}\Big) &= \frac{\partial}{\partial a_k}\Big(\frac{1}{(P-1)}\sum_{p=1}^{P}(f_p(q_{jp}) - \bar{q}_j)^2\Big) \\
&= \frac{1}{(P-1)}\sum_{p=1}^{P} 2(f_p(q_{jp}) - \bar{q}_j)\frac{\partial}{\partial a_k}(f_p(q_{jp}) - \bar{q}_j) \\
&= \frac{1}{(P-1)}\sum_{p \neq k} 2(f_p(q_{jp}) - \bar{q}_j)(\frac{1}{Pb_k}\Big) + \frac{2}{(P-1)}\Big(-\frac{1}{b_k} + \frac{1}{Pb_k}\Big)(f_k(q_{jk}) - \bar{q}_j) \\
&= \frac{1}{(P-1)}\sum_{p=1}^{P} 2(f_p(q_{jp}) - \bar{q}_j)\Big(\frac{1}{Pb_k}\Big) + \frac{2}{(P-1)} Bigl(-\frac{1}{b_k}\Big)(f_k(q_{jk}) - \bar{q}_j) \\
&= \frac{2}{(P-1)}\Big(-\frac{1}{b_k}\Big)(f_k(q_{jk}) - \bar{q}_j) \\
&= \frac{-2}{b_k(P-1)}(f_k(q_{jk}) - \bar{q}_j)
\end{aligned}
$$

Therefore, by adding these terms up we obtain,

$$
\frac{\partial}{\partial a_k} g(\mathbf{a}, \mathbf{b}) = \frac{-2}{b_k(P-1)}\sum_{j=1}^{Q}(f_k(q_{jk}) - \bar{q}_j).
$$

The process is very similar for partial derivatives of the scale parameters.

$$\begin{aligned}
\frac{\partial}{\partial b_k} var\Big(\frac{\mathbf{q}_j - \mathbf{a}}{\mathbf{b}}\Big) &= \frac{\partial}{\partial b_k}\Big(\frac{1}{(P-1)}\sum_{p=1}^{P}(f_p(q_{jp}) - \bar{q}_j)^2\Big) \\
&= \frac{1}{(P-1)}\sum_{p=1}^{P} 2(f_p(q_{jp}) - \bar{q}_j)\frac{\partial}{\partial b_k}(f_p(q_{jp}) - \bar{q}_j) \\
&= \frac{1}{(P-1)}\sum_{p\neq k} 2(f_p(q_{jp}) - \bar{q}_j)\Big(\frac{f_k(q_{jk})}{Pb_k}\Big) + \frac{2}{(P-1)}(f_p(q_{jp}) - \bar{q}_j)\Big(-\frac{(f_k(q_{jk})}{b_k} + \frac{(f_k(q_{jk})}{Pb_k}\Big) \\
&= \frac{1}{(P-1)}\sum_{p=1}^{P} 2(f_p(q_{jp}) - \bar{q}_j)\frac{f_k(q_{jk})}{Pb_k} + \frac{2}{(P-1)}(f_k(q_{jk}) - \bar{q}_j)\Big(-\frac{(f_k(q_{jk})}{b_k}\Big) \\
&= \frac{2}{(P-1)}\Big(-\frac{(f_k(q_{jk})}{b_k}\Big)(f_k(q_{jk}) - \bar{q}_j) \\
&= \frac{-2}{b_k^2(P-1)}(f_k(q_{jk}) - \bar{q}_j)(q_{jk} - a_k)
\end{aligned}$$

Which yields the equation

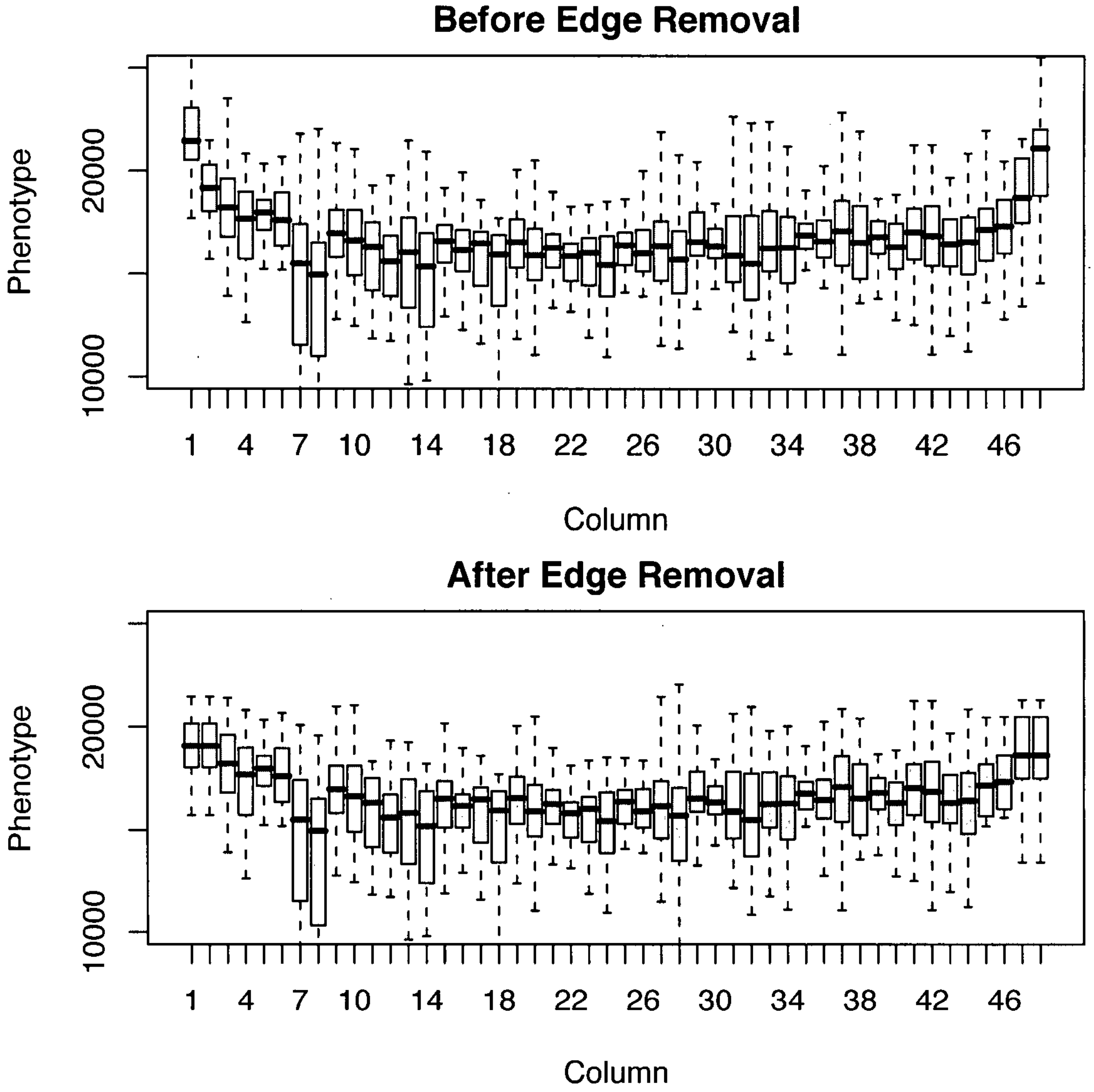
$$\frac{\partial}{\partial b_k} g(\mathbf{a}, \mathbf{b}) = \frac{-2}{b_k^2(P-1)}\sum_{j=1}^{Q}(f_k(q_{jk}) - \bar{q}_j)(q_{jk} - a_k)$$

Appendix B

The Edge Effect

Colonies near the edge produce systematically higher phenotypes. This effect is dubbed the edge effect. There are two explanations for this. Firstly these colonies have less neighbours and hence less competition for nutrients (distributed in the agar) allowing them to grow faster. Secondly the agar plates are slightly curved at the edges which bends the light from the scanner and results in brighter recorded intensities (*i.e.* stronger phenotypes are reported). Figure 2.4 shows an image of a particular agar plate obtained via scanner (left). The larger size of colonies near the edges, indicating a higher growth rate, is immediately observable as is the relative location of the technical replicates. These are allocated in two by two squares directly next to each other. Similarly, the controls are contained in a two by two square on the edge. Thus they are contained within the first four rows from the edge.

As previously mentioned, the location effect is outside the realm of this paper, nevertheless it cannot be entirely ignored here. The interim solution is to remove the data from the edges (first and last columns and rows). Figure B.1 shows the distribution of phenotypes within columns of a single plate before and after this manipulation. It is quite evident that while there is an improvement, the effect is not entirely removed and that it spans over the first few rows/columns. Fortunately for the controls, their location is the same from plate to plate. It is expected that the edge effect will cancel out, but it may be argued that the standard error will be increased because the median value is taken over two observations rather than 4.



Before Edge Removal

After Edge Removal

Figure B.1: Evidently, the observations on the first and last columns are those most affected by the location effect, but it is also evident that the edge effect is not restricted to these two columns. The data is eventually collapsed to the median of the four technical replicates, thus it is equivalent to replace the values on the edge with the values from the neighbouring row or column which explains why the first two and last two columns are equally distributed in the second plot.