HHS Public Access

Author manuscript

Technometrics. Author manuscript; available in PMC 2016 July 13.

Published in final edited form as:

Technometrics. 2015 ; 57(2): 234–244. doi:10.1080/00401706.2014.915890.

Robust Optimization of Biological Protocols

Patrick Flaherty and

Biomedical Engineering Department, Worcester Polytechnic Institute

Ronald W. Davis

Department of Biochemistry and Genetics and Stanford Genome Technology Center, Stanford University

Abstract

When conducting high-throughput biological experiments, it is often necessary to develop a protocol that is both inexpensive and robust. Standard approaches are either not cost-effective or arrive at an optimized protocol that is sensitive to experimental variations. We show here a novel approach that directly minimizes the cost of the protocol while ensuring the protocol is robust to experimental variation. Our approach uses a risk-averse conditional value-at-risk criterion in a robust parameter design framework. We demonstrate this approach on a polymerase chain reaction protocol and show that our improved protocol is less expensive than the standard protocol and more robust than a protocol optimized without consideration of experimental variation.

Keywords

Experimental Design; Analysis of Designed Experiments; Robust Parameter Design; Response Surface Methods; Quality Control / Process Improvement

1 Introduction

Comprehensive and coordinated experimental efforts are increasingly being used to collect large amounts of data to address difficult biological questions while realizing economies of scale to keep the cost per sample low. Typically, these large scale efforts start with a pilot stage during which protocols are improved and preliminary data are collected and then move to a production phase where fixed protocols and procedures are repeated on many samples. The Cancer Genome Atlas (McLendon et al. 2008) and the 1000 Genomes Project (1000 Genomes Project Consortium 2010) are two examples of such large scale efforts. The pilot-production stage model is not exclusive to large scale projects, and is common in research and development in both academic and industrial laboratories for the development of pharmaceutical production processes, clinical biomarker assays and next-generation sequencing protocols.

SUPPLEMENTARY MATERIAL

Supplementary Materials and Methods: Experimental protocols, derivation of (7) and model checking analysis. (pdf file)

PCR data set: Data set and software used for the example in Section 3.2 (also available at the authors' website): http://genomics.wpi.edu/rdoe. (.zip file)

In the laboratory, the most common approach to optimizing a protocol is a one-at-a-time design/analysis where each factor in the process is adjusted individually until a protocol that meets the particular needs of the experimentalist is found (Roux 2009). Though it is well known that one-at-a-time designs are inefficient they are still widely used. Part of the reason for their continued use is familiarity and simplicity to simultaneously reduce cost, minimize failures, and provide insights into the process being optimized. However, better experimental procedures are possible.

Statistical design of experiments (DOE) was pioneered for use in agricultural yield improvement (Yates 1935). Later, DOE methods were improved and updated for use in industrial manufacturing to identify and improve processes and ensure the production of products that are robust to environmental variations (Box and Jones 1990). In particular, fractional factorial and split-plot designs proved to be among the most useful experimental designs for manufacturing applications (Michaels 1964). As computational tools have matured, DOE has been used to obtain data for complex models using response function modeling.

Response function modeling (RFM) aims to quantitatively model a response as a function of control and noise factors (Wu and Hamada 2009, Ch. 11). Control factors are those that can be set during the production phase to optimize the system, while noise factors are hard to control during the production phase though they may be adjustable during pilot experiments. Noise factors are not easily measurable during the production phase (Wu and Hamada 2009, p. 576). The aim of RFM is to characterize the process and account for the response variation in terms of control and noise factors.

Once a quantitative model of the system has been obtained by RFM, robust parameter design (RPD) provides a way of choosing settings of control factors so that the influence of the noise factors on the response is minimized (Wu and Hamada 2009, p. 511). The approach we describe uses robust optimization methods to accomplish robust parameter design.

Robust optimization (RO) methods are useful in real-world decision environments where the data contain noise, where the optimal solution is difficult to implement exactly and where small perturbations in the optimal solution yield infeasible solutions (Ben-Tal and Nemirovski 2002). The uncertainty set of the original problem is reformulated using convex analysis to form a robust counterpart that is computationally tractable to solve, insensitive to small perturbations and implementable in practice. RO methods have been successfully used for antenna design, truss topology design, dynamic system control and linear regression (Ben-Tal and Nemirovski 2002; el Ghaoui et al. 1998). Recently, robust optimization has been combined with novel measures of risk in portfolio optimization problems (Artzner et al. 1999). We show that the asymmetric coherent risk measures that have been developed and proven useful for those applications are also useful in the protocol optimization methods (Rockafellar 2007).

The general aim of this study is to obtain control factor settings that minimize cost subject to probabilistic constraints on performance across a range of noise levels using coherent risk

measures in a robust optimization framework. In Section 2, we describe our three-stage approach of experimental design, modeling and robust optimization. Then, we demonstrate the method on a polymerase chain reaction protocol in Section 3. We present experimental design, statistical model and optimization results in Sections 3.1, 3.2, 3.3, respectively. Additionally, in Section 3.3, we provide independent experimental validation of our approach. We show that this approach results in a protocol that has minimal cost and is robust to process variations.

2 Approach

Our approach combines statistical response function modeling (RFM) and robust optimization (RO) in the context of robust parameter design (RPD) to obtain an improved protocol. Though we describe our approach as a sequence of steps in this paper, in practice one would iterate through each of the steps (Figure 1). An initial experimental design is used to obtain a set of factors that are used in subsequent experimental design and modeling iterations. The model and variation estimates are then used in a robust risk optimization framework to improve the protocol. The optimized operating conditions are finally checked by independent validation experiments.

2.1 Experiment Design

We first denote and classify the factors that influence the response in our application to biological protocols. Control factors, x, are controllable during the experimental phase and set for the production phase. Noise factors, z, are controllable during the experimental phase, but not during the production phase. Finally, noise factors, w, are not controllable during the during the experimental phase or production phase. The effect of noise factors on the response must be controlled by adjusting the control factors as usual in the robust parameter design framework.

We first ran a screening experiment focusing on main effects to eliminate potentially unimportant factors. We then designed a fractional factorial experiment to explore the response space of the subset of factors identified in a screening experiment. The fractional factorial design was then augmented with a center point to assess curvature in the response model. Finally, we designed and conducted a center-face composite arranged experiment to estimate quadratic effects.

We chose this staged strategy to allow for the adjustment of the experimental plan as more information became available from previous rounds. However, this strategy can lead to inefficiencies or missed significant factors. A two-level screening design may miss important quadratic effects that are identifiable for higher-level designs. If quadratic effects are anticipated a 3-level fractional factorial design or 2-level design with well chosen center points has more capacity to identify quadratic effects and may be better at the screening stage than a two-level design. It is challenging to select an optimal strategy on a fixed experimental budget when uncertainty is high at the outset. We present the approach we used with the understanding that as with all experimental strategies a more efficient experimental strategy may have been employed in hindsight.

2.2 Model Fitting

We used a mixed effects model to estimate the factor effects and variance components in order to understand the protocol as a system and to predict its behavior under novel conditions. In a model of the form,

$$g(x, z, w, e) = f(x, z, \beta) + w^T u + e, \quad (1)$$

β terms were modeled as fixed effects and {*u*, *e*}, were modeled as random effects (Robinson 1991). Modeling *u* and *e* as random effects essentially means that we will be interested in estimating the variance parameters associated with these random variables. The variable *w* is an indicator variable for the noise factor. We fit a model with all main and interaction effects among the fixed effects and identified outlying and influential observations using measures based on the residuals, prediction matrix, volume of the confidence ellipsoid and influence function (Chatterjee and Hadi 1986). We are primarily interested in eliminating data due to various forms of execution error; situations caused by a discrepancy between what the experimental arrangement called for and what was actually done (Anscombe and Guttman 1960). Replications in the experimental design aid greatly in identifying such non-reproducible observations. We refit the complete model to the non-outlying data and selected a parsimonious model by dropping model terms with insignificant regression coefficient t-statistics until the Bayesian Information Criterion (BIC) of the reduced model increased (Schwarz 1978; Hansen and Yu 2001). Having fit a parsimonious model for the fixed effects, we dropped random effects terms from the model that have a standard deviation that is not significantly greater than zero (Gelman 2005). The model was considered adequate for use in an optimization program when the estimated variance of the response is at least three-fold greater than the variance of the residual error (Box et al. 2005). We checked the quality of the model fit by leave-one-out cross validation.

In addition to using REML to fit the model, we also used a Bayesian approach to combine the model selection and parameter estimation step. The details of that analysis are provided in Supplementary Section 2. We found that the model form and coefficients generally agree with the REML estimates and we carry the REML estimates forward in this analysis.

2.3 Robust Optimization

Our aim is to produce a protocol that, when implemented, is inexpensive and robust to experimental variations. We use a convex risk optimization framework to select a setting of control factors such that the per reaction cost of the protocol is minimized while providing for a margin of safety against failure due to experimental variation. We cast the problem as one of minimizing cost subject to a lower-bound constraint on the protocol performance,

$$\text{minimize} g_0(x)$$

$$\text{subject to} g(x, z, w, e) \geq t \quad (2)$$

$$x \in \mathscr{S}.$$

where $g_0(x) = c^T x$ is the per reaction cost of the protocol with cost vector c and factor levels vector $x \in \mathscr{S}$. The constraint $g(x, z, w, e) \geq t$ ensures that the protocol performance, as predicted by the model, is at least as high as some threshold t. We have a stochastic optimization problem because $g(x, z, w, e)$ is random due to the randomness in the noise factors z, w, and e.

Several classical approaches to dealing with the randomness in the constraint include: "guessing the future" "worst-case analysis" "relying on expectations" "standard deviation units as safety margins" and "probability of compliance." In "guessing the future" values for z, w, and e are simply set and the problem is then treated as a deterministic one. For "worst-case analysis" the minimal value of $g(x, z, w, e)$ across z, w, and e is used for each x.

In order to compare these approaches for handling uncertainty in the optimization problem we use the concept of coherency of a risk measure. Consider risk measure $(R) : \mathscr{L}^2 \mapsto (-\infty, \infty]$ that is a functional of a random variable that quantifies the risk of loss. Such functionals that satisfy convexity, monotonicity, closedness and positive homogeneity axioms (Artzner et al. 1999) are considered coherent measures of risk in the basic sense and have been shown to (1) preserve convexity of the deterministic function, (2) preserve certainty and (3) be scale insensitive (Rockafellar 2007). Without coherency, we may loose the convexity of the original problem, or we may be left with a solution that is exceedingly fragile to small perturbations. The risk measures induced by all of the classical approaches fail one or more of these axioms.

Another potential solution that has been used for similar problems is to swap the constraint and objective in (2) and pose a larger-the-better optimization problem. In the larger-the-better robust design framework, one maximizes $E[g(x, z, w, e)^2]$ over x such that the cost $g_0(x) < t$ for some threshold t. Here we must have some idea of what an acceptable cost is for the protocol, which may not be available. But more importantly, recasting the objective in this way combines the expected value of the response and the variance of the response into one functional and thus only exerts indirect control over the worst-case events. In our original problem, we are not interested in a better response beyond a certain threshold. Instead, we are interested in ensuring that the response exceeds the threshold with high probability. Recasting the problem as a larger-the-better optimization may solve that objective, but only indirectly.

Instead of these classical approaches, we employ the conditional value-at-risk (CVaR) functional because it is coherent and it solves the original problem statement directly (Rockafellar 2007). The definition of CVaR is

$$\overline{g}(x) = \mathrm{CVaR}_\alpha[\underline{g}(x)] = \frac{1}{\alpha}\int_0^\alpha \max\left\{s | F_{\underline{g}(x)}(s) \leq \gamma\right\} d_\gamma \quad (3)$$

where γ is dummy variable for integration and $\underline{g}(x) \triangleq g(x, z, w, e)$ to simplify notation. We see that CVaR at level α measures the expected value of $\underline{g}(x)$ in the α-tail.

By ensuring that the CVaR exceeds some threshold *t*, we ensure that with high probability 1 − α the response will be at least *t*. Indeed, we guarantee that the expected value in the α-tail is a least *t*. Our CVaR definition (3) differs slightly from the one provided by Rockafellar (2007) because theirs was constructed to ensure loss in a portfolio does not go above a certain threshold whereas ours ensures the yield of the protocol does not go below a certain threshold.

In some applications it is more useful, for interpretation purposes, to consider the value-at-risk (VaR) or α-quantile functional, and we use that measure as well because of its intuitive accessibility. VaR is the percentile constraint functional common in stochastic optimizatio

$$\overline{g}(x)=\mathrm{VaR}_{\alpha}[\underline{g}(x)]=\max\left\{s|F_{\underline{g}(x)}(s)\leq\alpha\right\}. \quad (4)$$

Figure 2 shows a simplified diagram of the constraint $\bar{g}(x) \geq t$ where the cumulative distribution function $F_{g(x)}(z)$ for a Normal(0,1) random variable is shown with level α. The value-at-risk constraint $\mathrm{VaR}_{\alpha}[g(x)]$ in (4) is the α–percentile of the distribution. The conditional value-at-risk constraint $\mathrm{CVaR}_{\alpha}[g(x)]$ in (3) is the expected value of the lower α-tail of the distribution. Requiring that $\mathrm{CVaR}_{\alpha}[g(x)] \geq t$ is more conservative than $\mathrm{VaR}_{\alpha}[g(x)] \geq t$ because the conditional value-at-risk considers where the mass of the distribution lies along *s* in the tail while the value-at-risk criterion only considers that there is α mass in the tail without regard to where it lies along *s*.

The general robust counterpart to (2) is then

$$\mathrm{minimize} g_0(x)$$

$$\mathrm{subject\ \ to} \overline{g}(x) \geq t. \quad (5)$$

In other applications of this methodology it may be useful to endow a noise variable with a prior and form the marginal cumulative distribution function $F_{g(x)}$. Samples may be obtained from $F_{g(x)}$ and the sample mean in the α-tail of the empirical cumulative distribution function may be used as an approximation of the true distribution. This approach was used by Palmquist et al. (1999) in their analysis of stock market returns from simulation data.

In special cases, such as if a noise variable enters as a linear term in $g(x)$ and has a Gaussian distribution, we may write $F_{g(x)}$ analytically. A simple example of the closed-form solution can be constructed by considering the model $g(x, z, e) = 1 - 0.2x - 0.5xz + e$ where $z \sim \mathcal{N}(0, 1)$ and $e \sim \mathcal{N}(0, 1)$ are independent. Then we have $g(x) \sim \mathcal{N}(1 - 0.2x, 0.25x^2 + 1)$. We can write the CVaR risk function as $\overline{g}(x)=1-0.2x+\sqrt{0.25x^2+1}\Phi^{-1}(\alpha)$ where $\Phi(\cdot)$ is the standard normal cdf.

3 Polymerase Chain Reaction Protocol

Polymerase chain reaction (PCR) is a common and indispensable molecular biology technique used to amplify the total number of molecules of a fragment of DNA (Mullis and Faloona 1987; Schochetman et al. 1988) and verify that a sequence of interest is present in a DNA sample. Thus, our objective is to minimize the per reaction cost while maintaining an adequate yield so the PCR product band is clearly visible on an agarose gel. The process with control and noise factors identified is shown in Figure 3.

Taq DNA polymerase, an enzyme that is able to duplicate a single strand of DNA, is made to operate repeatedly yielding a geometric increase in the number of copies of the original DNA molecule. The Taq polymerase is combined with other components: MgCl, dNTPs, oligonucleotide primers and template DNA. The temperature of this mixture is cycled from 95°C to 55°C to 72°C and a new copy of the template DNA is synthesized for each original molecule.

The difference between the volume of the constituent ingredients and the total volume is made up with sterile water. The amount of water varies with the total volume of the other components but is always greater than 50% of the total volume. Because water serves as a slack variable, the level of a control factor does not need to be adjusted to compensate for an increase or decrease in another control factor in the mixture.

The amplified DNA product is visualized by running the product through an agarose gel by electrophoresis. This separates the DNA on the basis of size since longer fragments move more slowly through the gelatin matrix. The DNA is then stained by immersing the gel in a water bath containing ethidium bromide which intercalates in the DNA polymer and fluoresces at UV wavelengths.

The agarose gel is scanned by a laser and the intensity of each pixel in the band region of interest (ROI) is measured as a 16-bit integer. The background intensity (BG) of a ROI is the median intensity of pixels on the border of the ROI. The signal-to-noise ratio (SNR) is the ratio of the average intensity in the ROI to the median background intensity for all of the ROIs on the gel image. This SNR could also be considered a contrast ratio,

$$y=\mathrm{SNR}=\frac{\frac{1}{|\mathrm{ROI}|}\sum_{q\in \mathrm{ROI}} q_i}{\mathrm{median}[\mathrm{BG}]}, \quad (6)$$

where ROI is the set of pixels in the region of interest, q is a pixel intensity, and BG is a vector of background intensities near the ROI on the gel.

We divide the experimental noise factors into two groups: those associated with the template DNA, z, and those associated with the experimental batch and gel staining w. Template length and template concentration can be controlled during the experimentation phase, but are not during the production phase. Biological replicates, which are completely independent experiments, are captured in the factor, w_2. Technical replicate, which are splits within a biological replicate, are captured in w_1. Multiple runs are stained together in ethidium bromide on a single gel slab and the staining batch is captured in the noise factor w_3.

A master mix cocktail contains all the PCR reaction components except the DNA template. The total cost of the master mix can be divided into fixed and variable costs. Buffer and $MgCl_2$ are supplied with Taq, so they are considered fixed costs. Primers could be considered either a variable or fixed cost because they are usually supplied in ample quantity to perform many reactions. We consider it a variable cost here. The component costs of a standard 50μL reaction are shown in Table 2. Taq DNA polymerase is by far the most expensive component, and we expect an improved protocol will minimize the amount of Taq required.

3.1 Experiment Design

Our experiment design used three stages: First, a screening experiment was conducted to identify important factors affecting PCR yield. Then a fractional factorial experiment of higher resolution was conducted to fit a response surface. Finally, validation experiments were conducted at the optimized control factor levels to verify that the yield was sufficient to visualize on a gel with high confidence.

For the screening experiment, we designed a 2^{6-3}_{III} fractional factorial arrangement to identify factors listed in Table 1 having a significant effect on yield. Figure 4 shows a least squares regression analysis of two replicates of the arrangement. Figure 4A shows the fitted values and the observed values indicating that there is significant variation due to the control factors and a much lower intra-run variation - the experiments are reproducible. Model coefficients (shown in Figure 4B) indicate that the factors x_1, x_3, x_4, and z_1 have a significant effect on the response. Though the dNTP concentration, x_2, did not have a statistically significant effect, the main effect is aliased with the $M \times P$ and $T \times L$ interaction effects which we believe may be significant due to prior experience working with this system (Roux 2009). We include these two interaction effects in the next experimental stage to obtain more data on them. Figure 4C and Figure 4D show that the quality of the fit is sufficient for this stage of the experimental plan and we move to the next stage of experimentation with these factors.

To collect data for fitting a response surface model, we designed a $2^{5-1}_{IV}+2$ fractional factorial arrangement. This notation indicates that we used a basic 2^{5-1}_{IV} arrangement and augmented that with 2 center points at each of the levels of the template length noise factor, L. Upon initial examination of the data, we identified a peak in the PCR yield around the center point for the $MgCl_2$ factor, so a center-face composite design was conducted to augment the $2^{5-1}_{IV}+2$ fractional factorial design data. The center-face composite design (runs 19–34) fixes levels of all factors but one of the master mix components to the average of the high and low levels in the fractional factorial design. These 16 runs are divided into 8 runs with the long template and 8 runs with the short template. Within each subdivision, there are 4 pairs of experiments where one of the 4 master mix components is set to the high and low levels and the remaining components are set to the center value.

The experiment was conducted by setting up the reaction mixture runs prescribed by the arrangement, then the reaction was split in half for technical replication. Each technical

replicate was assayed on separate agarose gels for each of the three experimental block replicates. The noise factors (experiment, run, stain) were recorded and entered into the statistical model to estimate the variance components of the random effects.

Figure 5 shows the PCR yield results on the agarose gels and the quantified yield for the second stage experiment. Figure 5A shows the yield of the PCR reaction on an agarose gel. Runs 1–18 comprise the fractional factorial part of the design with two center points at the extreme template lengths and runs 19–36 comprise the center-face composite design. The fractional factorial design has two technical replicates nested within three biological replicates for each run. The center-face composite design also has two technical replicates, but only one biological replicate.

Intensity measurements shown in Figure 5A are converted to SNR values in Figure 5B according to (6). Due to the replication structure, the fractional factorial runs each have six data points and the center-face composite runs each have 2 data points. A signal-to-noise ratio of six or greater (denoted by the horizontal dashed line) is clearly visible on the gel. The majority of factor combinations are consistently visible, while some reactions mixtures produce little if any product. Outliers are labeled as filled circles.

Figure 5C shows the fractional factorial experimental arrangement. The fractional factorial design is shown in the left matrix (runs 1–18) and the center-face composite design is shown in the right matrix (runs 19–36). The runs for each design are shown in the same sequence on the gel lanes in 5A.

The gel staining effect is clearly visible and is significant in the statistical model. The experimental block replicates are reproducible in general though some runs have specific reaction yields that are distant from the average. Our arrangement identifies factor combinations that consistently fail to achieve a threshold for visibility on the gel as well as combinations that consistently succeed.

The estimated main effect of each of the control factors is shown in Figure 6. The low level for $MgCl_2$ produces very little yield and the medium and high levels produce approximately equal yield. Furthermore, for all levels of $MgCl_2$, the reaction produces higher yield for the 1000bp template length compared to the 3000bp length. A similar pattern is observed for the other control factors. To more fully understand the interdependencies amongst the control factors, we turn to fitting a response surface model. If our objective were to simply obtain a robust reaction mixture it would be sufficient to select a factor setting from the runs and stop here. However, since we aim to achieve an inexpensive robust mixture, we use this data to estimate a model for the yield and then optimize over the cost.

3.2 Statistical Model

To construct a statistical model for the protocol, we followed the procedure outlined in Section 2.2. Briey, we fit a full model with all main effects for the control factors, x, and noise factors, z; pairwise interactions among the control factors; pairwise interactions between the control factors and noise factors, z; and quadratic effects for the control factors. Then we subjected that full model to several tests for influential observations and removed

those identified data points. Third, we identified a parsimonious model by stepwise model selection and a Bayesian model selection method. We applied these two independent model selection procedures and identified the same model structure giving us confidence that the model identified is not particular to one model selection procedure. Finally, we checked the fit of the parsimonious model to the data by leave-one-out cross validation.

The full mixed effects model (1) was fit by maximum likelihood with the control and noise factors, z, were entered as fixed effects. The effects u_3, u_2, and u_1 associated with staining (w_3), biological replicate (w_2) and technical replicate (w_1) were entered as random effects because we are interested in the variance components associated with these factors. Again, w_1, w_2 and w_3 are indicator variables. The technical replicate is nested within the biological replicate in the experimental arrangement and entered into the mixed effects model as such.

Outliers were identified by a variety of methods. Chatterjee and Hadi (1986) review 11 influence measures that can be broadly categorized into five groups: residual-based measures, prediction matrix-based measures, confidence ellipsoid-based measures, influence function-based measures, and partial influence-based measures. We computed these influence measures with a significance cutoff of $\alpha = 0.001$ where necessary. Figure 7 shows the standardized and studentized residuals for the data set used to estimate the model. We called an outlier if it was identified in two or more of four categories: residual, prediction matrix, confidence ellipsoid volume, influence function. The results for all of the individual influence tests are shown in Supplementary Section 2.3. In total, six data points out of 144 measurements were called outliers (shown as filled circles in Figure 5).

After removing outliers, we selected a parsimonious model by minimizing BIC for fixed effects and eliminating random effects with near zero standard deviation. The final model for the PCR protocol was fit by restricted maximum likelihood (REML). The estimated fixed effects, $\hat{\beta}$ are shown in Figure 8A. The coefficients indicate that increasing $MgCl_2$ (x_1) or Taq (x_3) increases yield and longer templates yield less product holding other factors constant. The main effect of $MgCl_2$ is the largest, consistent with anecdotal evidence, and has a maximum for the response due to the quadratic term. One might expect that increasing the amount of the raw material for DNA, dNTPs (x_2), would increase yield. So the observation that increasing dNTP concentration decreases yield is initially surprising. However, it has been shown that dNTP chelates $MgCl_2$ and actually decreases the yield by decreasing the amount of available $MgCl_2$ which in turn has a large main effect (Henegariu et al. 1997). The eight significant interaction terms illustrate interdependence of the factors in a biochemical reaction. The effect of the interaction terms are illustrated by the curvature of the contours in the pairwise marginal plots of the response in Figure 8B. Each plot is a marginal of fixed effect factors shown in the corresponding row and column. In particular, a rapid decline in the yield is evident as $MgCl_2$ decreases, but the yield is relatively unchanged for a wide range of Taq and primer concentrations holding the other factors fixed.

The random effects u_1 and u_3 were retained in the final model, and their standard deviations were estimated to be 0.42 and 0.26 respectively. The cause of the u_1 effect is likely due to pipetting variation; a multichannel pipettor was used to split the reaction material for the

technical replicate as well as to load the agarose gel. The significance of this effect illustrates the importance of recording as much of the experimental process and using statistical analysis to interpret which effects to disregard or retain in the model.

To check the fit of the model, we performed a leave-one-out cross-validation. That procedure gives a root mean squared prediction error of 0.96, which compares favorably to the total random effect standard deviation of 0.76 obtained by summing the variances due to random effects and the residual variance.

3.3 Optimization Problem

The protocol was optimized using the value-at-risk criterion and the conditional value-at-risk criterion. We compare the solutions under these formulations to solutions without margin-of-safety constraints and the standard protocol.

When optimizing the master mix, we are uncertain as to the particular value of the template length factor, z_1. A conservative approach is to choose values of the control factors such that the yield is great enough in the worst case over template length. So, we require that the protocol yield enough product for all values of template length, z_1 with high probability. This allows us to minimize (1) over z_1 and eliminate z_1 from the optimization problem.

Recall that our model of the process is quadratic in x and z. Ignoring the random variables u and e for now, we can write the process model as

$$\begin{aligned} f(x) &= \min_z f(x,z) \\ &= \min_z (A + x^T B_1 + z^T B_2 + x^T C_{11} x + x^T C_{12} z + z^T C_{22} z) \\ &= \min_z \left(\begin{bmatrix} 1 & x^T & z^T \end{bmatrix} \begin{bmatrix} A & \frac{1}{2}B_1 & \frac{1}{2}B_2 \\ \frac{1}{2}B_1^T & C_{11} & \frac{1}{2}C_{12}^T \\ \frac{1}{2}B_2^T & \frac{1}{2}C_{12} & C22 \end{bmatrix} \begin{bmatrix} 1 \\ x \\ z \end{bmatrix} \right), \end{aligned} \quad (7)$$

where we have recast the fixed effects terms in (1) to a standard optimization form by partitioning $\hat{\beta}$ into $\{A, B_1, B_2, C_{11}, C_{12}, C_{22}\}$. Since the model does not have a quadratic effect for template length ($C_{22} = 0$), the objective function involving z is affine. The optimization over z is then solvable by linear programming (see Supplementary Section 2.1). Furthermore, since $C_{11} \preccurlyeq 0$, the model is a concave function of the decision variables, x.

Now, reintroducing the random variables for u and e, the process model is

$$\underline{g}(x) = \min_z f(x,z) + w^T u + e, \quad (8)$$

Where

$$
\begin{aligned}
f(x,z) = {} & 4.02 + 1.19x_1 - 0.79x_2 + 0.23x_3 + 0.31x_4 - 0.38z_1 - 1.61x_1^2 + 0.60x_1x_2 \\
& - 0.20x_1x_3 \\
& - 0.25x_1x_4 \\
& - 0.23x_2x_3 \\
& - 0.22x_2x_4 \\
& - 0.38x_3x_4 \\
& + 0.30x_3z_1 + 0.22x_4z_1,
\end{aligned}
$$

and $u+e \sim \mathcal{N}(0, \sigma_1^2+\sigma_3^2+\sigma_e^2)$ models the variation in the measured yield on the agarose gel. Recall, the variance terms σ_1^2 and σ_3^2 correspond to the technical replicate w_1 and stain batch w_3. The estimates of the variance components are $\hat{\sigma_1} = 0.26$, $\hat{\sigma_3} = 0.42$, and $\hat{\sigma_e} = 0.58$.

We use (8) in the robust optimization framework (5) to solve for robust parameters x for the process. A margin-of-safety level $\alpha = 10^{-3}$ was used for both the value-at-risk criterion (4) and the conditional value-at-risk criterion (3). We obtained the deterministic form of the optimization problem to compare these robust solutions to a process optimized without robust constraints.

A comparison of the optimal solutions is shown in Figure 9. A table of the robust and non-robust parameters is shown in Figure 9A. All of the optimized solutions contain more $MgCl_2$, less dNTPs, and less Taq than the standard formulation. The primer concentration is greater in the two RO formulations. The standard formulation is most expensive at \$0.86 per 50μl reaction and the deterministic solution is least expensive at \$0.62. The RO solutions are in between those extremes with the more conservative CVaR solution costing more than the VaR formulation. However, the per reaction cost reduction comes at the price of limited robustness to experimental variations. The deterministic reaction is expected to fail to meet the yield requirement 50% of the time for a 3kb template (Figure 9B). Considering a geometric rate of success for repeated reactions, the long term cost per reaction is \$0.8998, \$1.242, \$0.6783 and \$0.7885 for the standard, deterministic, VaR and CVaR formulations respectively. The apparent cost savings of the deterministic formulation are illusory due to the fragility of the solution to experimental variation.

Figure 9C shows a sensitivity analysis varying each control factor one-at-a-time around the CVaR solution point. The sensitivity analysis is shown for two settings of the template length (z_1) for the short (1kb, solid line) and long (3kb, dashed line) template with ±1*s.d.* intervals (light and dark gray respectively). The solution is insensitive to large scale variations in any one component. Figure 9D shows an independent experimental validation of the four optimization solutions at a 1kb and 3kb template length. The PCR yield was measured at each of the four optimized solutions for two template length levels with two replicates for a total of 16 validation experiments. The deterministic reaction failed two out of two times for the long template length while the standard, VaR and CVaR formulations worked in all trials (the measured value is greater than the detection threshold shown as a dotted line). To ensure the formulation was not particular to the specific genomic region amplified, we selected a different 3kb region in the genome for validation. These

observations are consistent with our model-based expectation that the naive optimization of the biological protocol without a margin of safety can lead to overly optimistic solutions that fail when implemented in practice.

4 Discussion

Improving high throughput biological protocols requires attention to both the cost and the fragility of the final protocol. By exploring the factor space efficiently with a designed experiment, modeling the systems accurately and exploiting that understanding using robust optimization methods, we are able to optimize the protocol and ensure a margin-of-safety against failure of the protocol.

While we have focused on product yield, other applications may have constraints on fidelity and specificity of the product. The gel image in Figure 5A shows some product bands are more smeared than others which may indicate less specific amplification. By modeling each response of interest and incorporating constraints on those in the optimization program, we can obtain a robust solution that accounts for multiple response factors.

In this paper, we fixed the parameters of the model in the optimization problem. A full Bayesian procedure would incorporate parameter uncertainty through estimated distributions for the coefficients and would likely give better estimates of the variation in the predicted response distribution. In future work we aim to address the issue of incorporating parameter uncertainty in the risk optimization framework in a way that preserves the computational tractability of the RO program.

We considered the worst case PCR yield over template length in a bounded interval in (7). An alternative approach would be to incorporate the distribution over z in response variation in the CVaR risk functional. The outcome of the optimization would give us a statement about the confidence on the PCR yield over random variation due to noise factors as well as template length. However, in a typical use case, the experimentalist knows the template length before doing the PCR. Indeed, primer sequences were designed specifically to ank the template sequence. In contrast, staining effect u_3 and the other noise factors are not even measurable prior to the PCR experiment. So we found it more practical to consider the worst case over z and incorporate the other noise effects into the CVaR risk framework.

We have assumed a Gaussian distribution for the model error term, but the RO program provides a flexible framework to incorporate any error distribution including empirical distributions obtained from simulation or experimental data. In the Gaussian case, the CVaR constraint reduces to a margin-of-safety based on standard deviations which is in general not coherent, so we have described our approach using the general case so that it can be used more broadly in applications.

We have used the DOE data to empirically model the protocol as a process and predict the outcome of different factor settings in order to optimize the protocol. This approach has additionally pointed towards the underlying biological mechanisms of the enzymatic action of Taq polymerase. Specifically, we observed that increasing dNTP concentration decreased the yield due to $MgCl_2$ chelation. We expect that this systematic empirical modeling

approach will generally inform valuable directions for research into the underlying mechanistic causes of surprising observations.

As the unit cost of translational medicine assays (medical assays that make use of basic genetic research to optimize patient care) continue to decline, it is becoming feasible to collect more data on more patients. A robust protocol is necessary to ensure reproducible results and consistent data collection. A margin-of-safety against failure is provided for in the production protocol by optimizing the protocol for both cost-efficiency and obustness using the DOE-RO method.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENTS

This work was supported by National Institutes of Health [T32 CA1211940 to P.F., P01 HG000205 to P.F. and R.W.D.]. We would like to thank Joe Horecka and Angela Chu for the generous donation of the yeast strain JHY222, primer sequences to amplify genomic loci and experimental advice. The authors are grateful to the Editor, the Associate Editor, and referee for their helpful comments that led to substantial improvements to this article. The authors thank Sarah Moore, Nancy Zhang, and Janine Mok for their careful reading and helpful comments on early versions of this manuscript.

References

1. 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. Nature. 2010; 467:1061–1073. [PubMed: 20981092]
2. Anscombe FJ, Guttman I. Rejection of Outliers. Technometrics. 1960; 2:123–147.
3. Artzner P, Delbaen F, Eber JM. Coherent measures of risk. Mathematical Finance. 1999; 9:203–228.
4. Ben-Tal A, Nemirovski A. Robust optimization–methodology and applications. Mathematical Programming. 2002; 92:453–480.
5. Box G, Jones S. Designing Products That Are Robust to the Environment. Tech. rep. 1990
6. Box, GE.; Hunter, JS.; Hunter, W. Statistics for Experimenters: Design, Innovation, and Discovery. 2nd ed.. Wiley-Interscience; 2005.
7. Chatterjee S, Hadi AS. Inuential observations, high leverage points, and outliers in linear regression. Statistical Science. 1986; 1:379–393.
8. el Ghaoui L, Oustry F, Lebret H. Robust solutions to uncertain semidefinite programs. SIAM journal of optimization. 1998; 9:33–52.
9. Gelman A. Analysis of variance: why it is more important than ever. The annals of statistics. 2005; 33:1–31.
10. Hansen MH, Ya B. Model selection and the principle of minimum description length. Journal of the American Statistical Association. 2001; 96:746–774.
11. Henegariu O, Heerema NA, Dlouhy SR, Vance GH, Vogt PH. Biotechniques. Tech. rep. 1997 UNITED STATES.
12. McLendon R, Friedman A, Bigner D, Van Meir EG, Brat DJ, et al. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature. 2008; 455:1061–1068. [PubMed: 18772890]
13. Michaels SE. The usefulness of experimental designs. Applied Statistics. 1964; 13:221–235.
14. Mullis KB, Faloona FA. Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction. Methods in enzymology. 1987; 155:335–350. [PubMed: 3431465]

15. Palmquist J, Uryasev S, Krokhmal P. Portfolio Optimization with Conditional Value-at-risk Objective and Constraints. Tech. Rep. 1999:99–14.
16. Robinson GK. That BLUP is a good thing: The estimation of random effects. Statistical Science. 1991; 6:15–32.
17. Rockafellar RT. Coherent approaches to risk in optimization under uncertainty. Tutorials in operations research: OR tools and applications: glimpses of future technologies. 2007
18. Roux KH. Optimization and Troubleshooting in PCR. Cold Spring Harbor Protocols. 2009:1–7.
19. Schochetman G, Ou C, Jones W. Polymerase chain reaction. The Journal of infectious diseases. 1988; 158:1154–1157. [PubMed: 2461996]
20. Schwarz G. Estimating the dimension of a model. The annals of statistics. 1978; 6:461–464.
21. Wu, CFJ.; Hamada, MS. Experiments: Planning, Analysis, and Optimization (Wiley Series in Probability and Statistics). 2nd ed.. Wiley; 2009.
22. Yates F. Complex experiments. Supplement to the Journal of the Royal Statistical Society. 1935; 2:181–247.

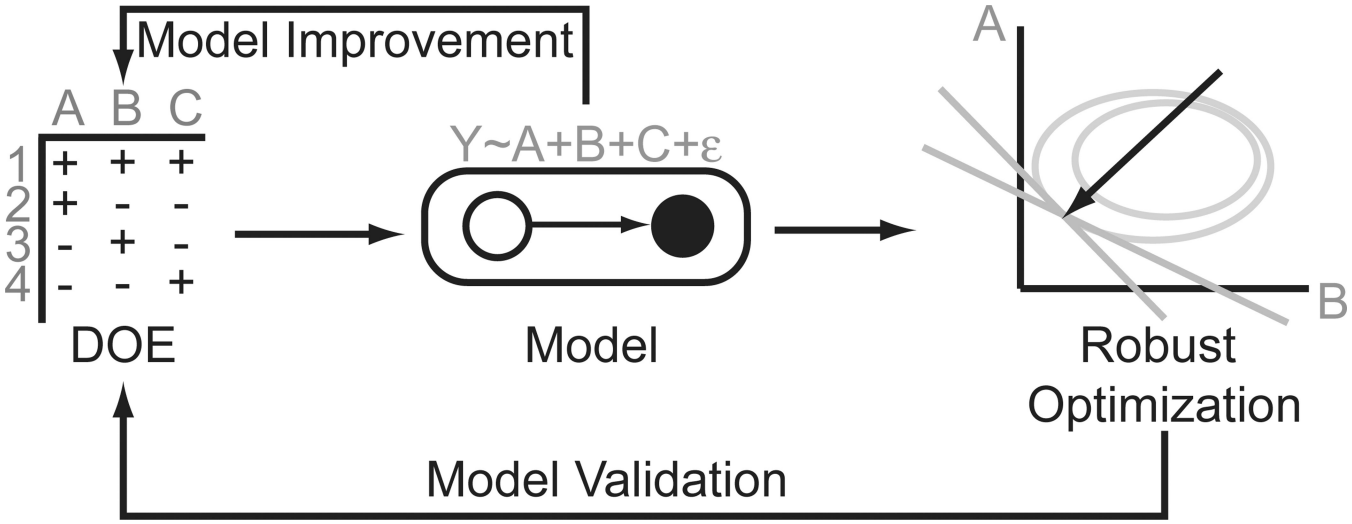


Figure 1.
The DOE-RO approach to robust parameter design.

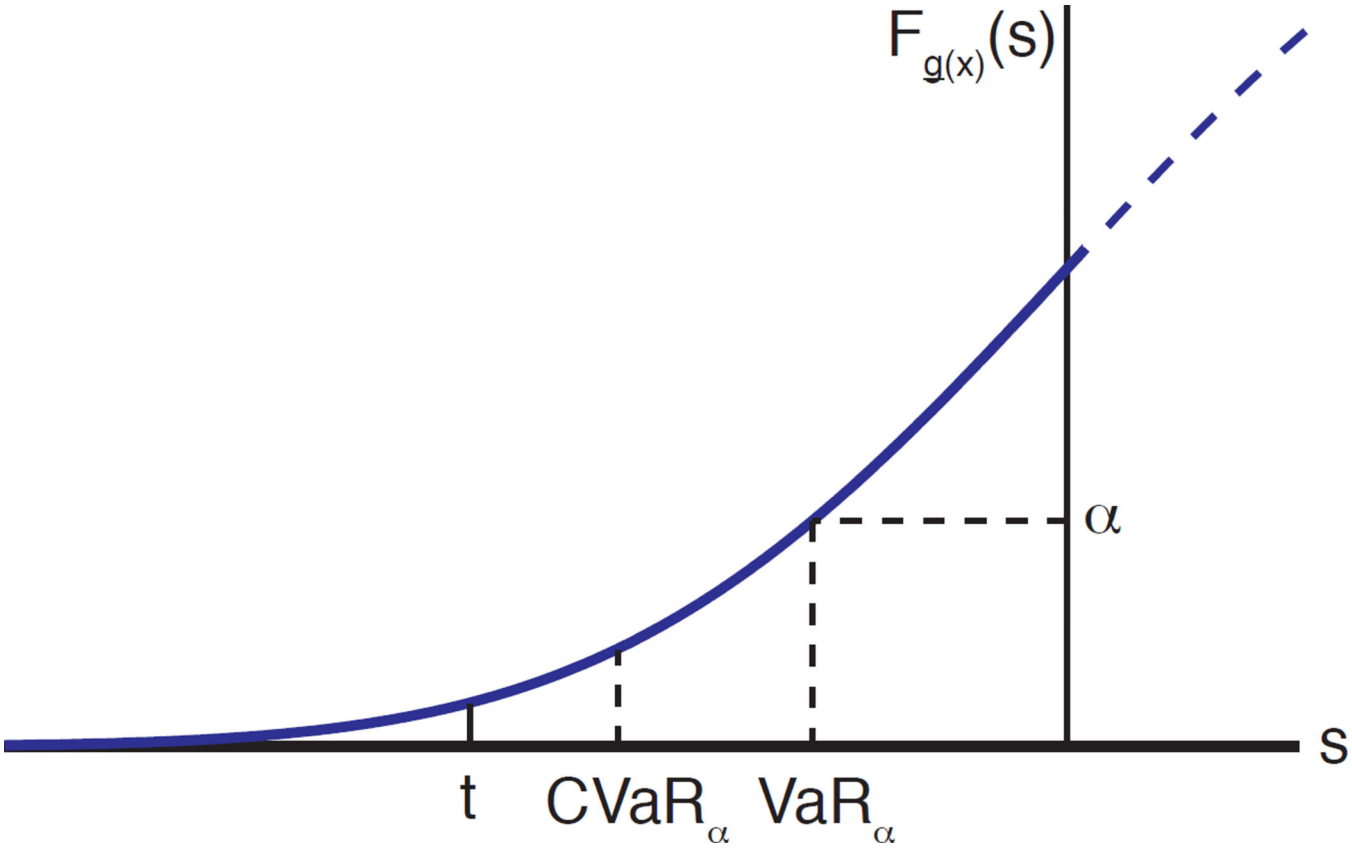


Figure 2.
Comparison of Conditional Value-at-Risk and Value-at-Risk probabilistic constraints. VaR_α is the α-percentile of the distribution and $CVaR_\alpha$ is the expected value of the α-tail of the distribution.

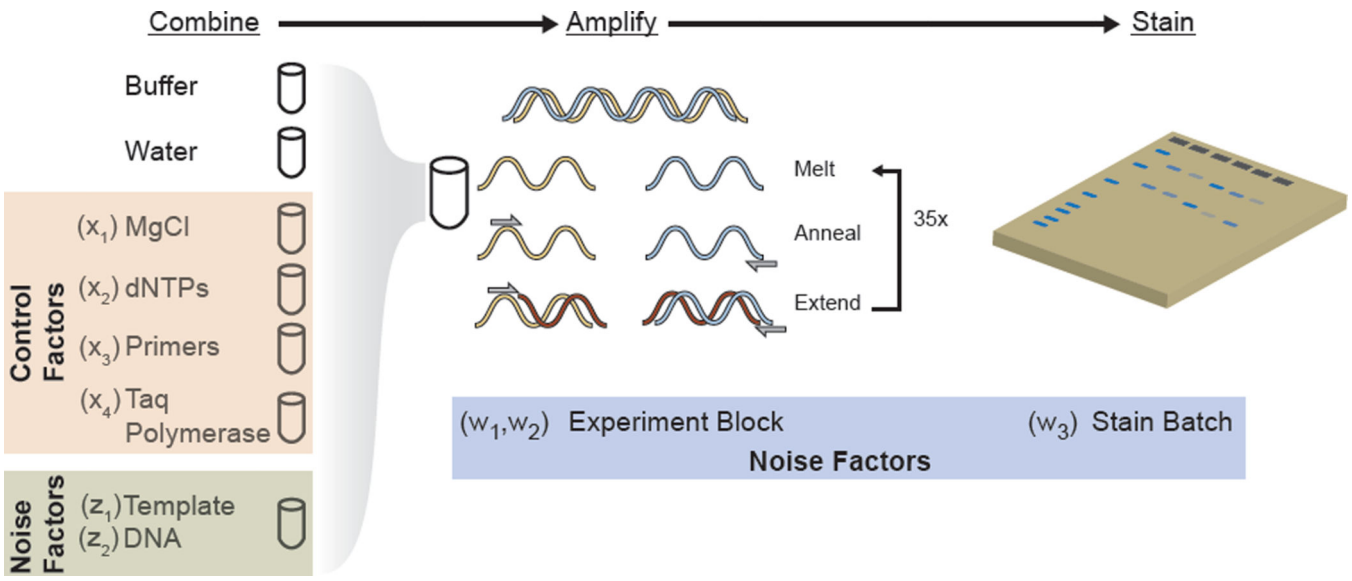


Figure 3.
PCR Process Flow

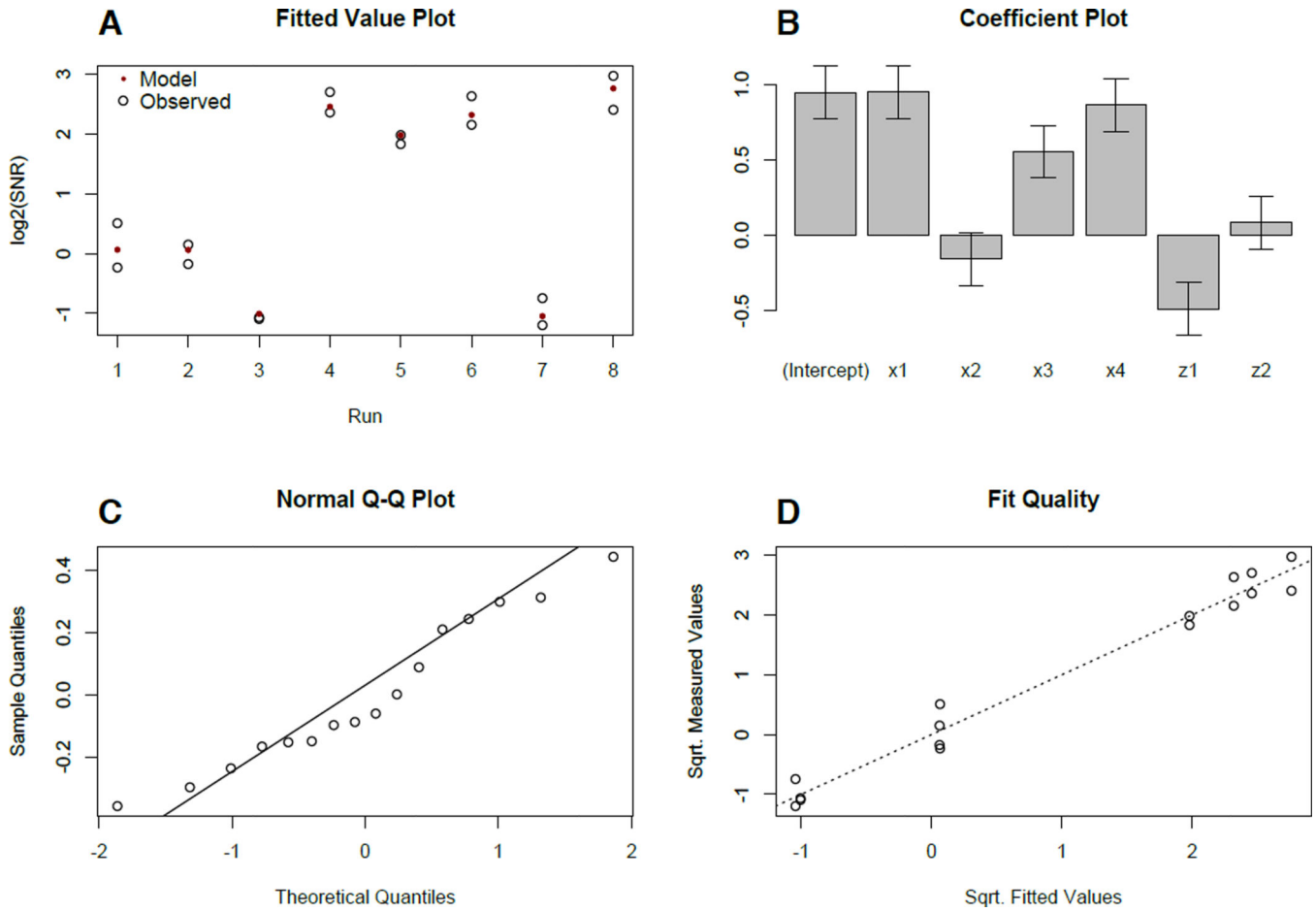
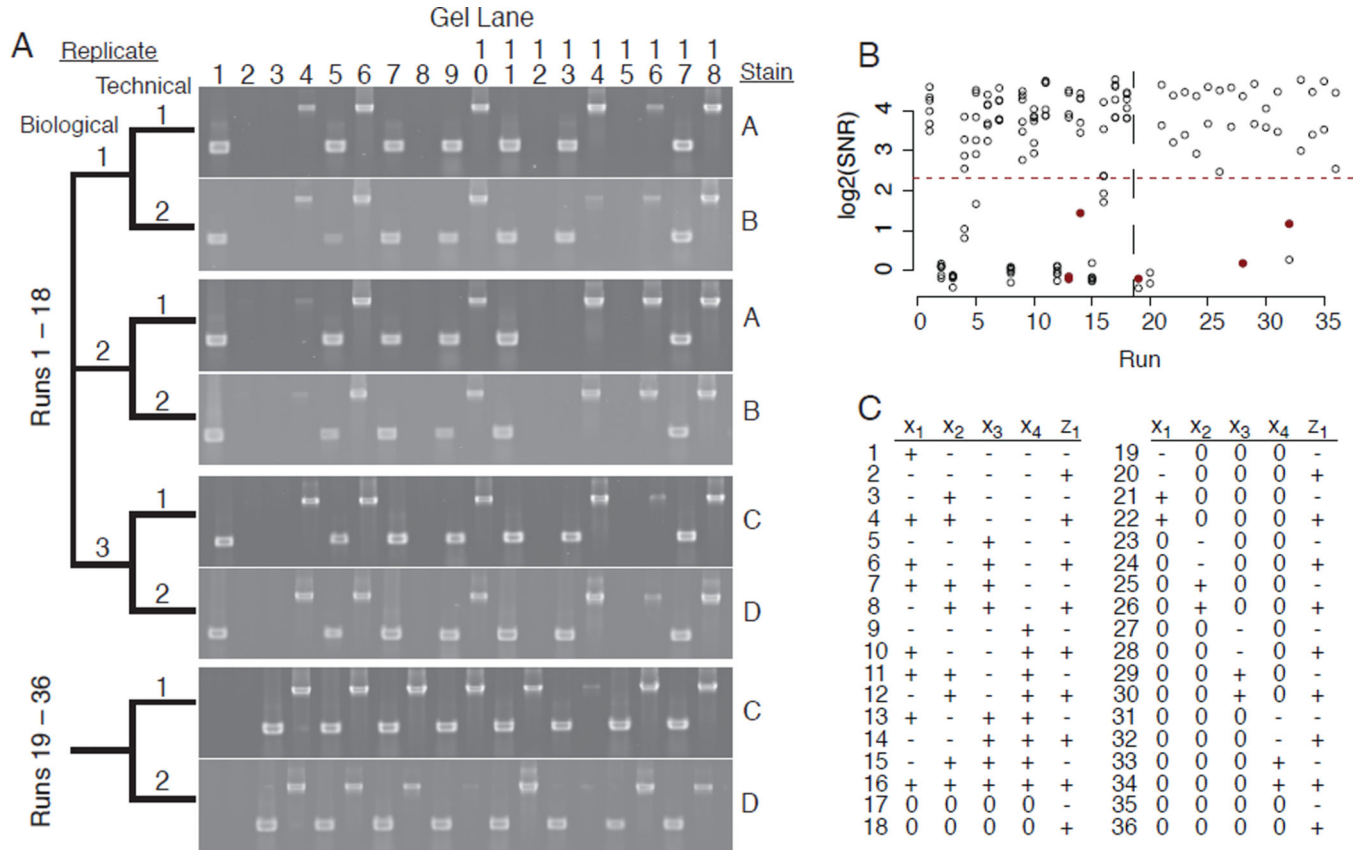


Figure 4.

Screening experiment model fit. (A) Observed and estimated outcomes for an 8 run screening design. (B) Coefficients for main effects model. (C) Normal Q-Q showing error distribution. (D) Fitted vs Measured SNR for screening experiment data.



	x_1	x_2	x_3	x_4	z_1		x_1	x_2	x_3	x_4	z_1
1	+	-	-	-	-	19	-	0	0	0	-
2	-	-	-	-	+	20	-	0	0	0	+
3	-	+	-	-	-	21	+	0	0	0	-
4	+	+	-	-	+	22	+	0	0	0	+
5	-	-	+	-	-	23	0	-	0	0	-
6	+	-	+	-	+	24	0	-	0	0	+
7	+	+	+	-	-	25	0	+	0	0	-
8	-	+	+	-	+	26	0	+	0	0	+
9	-	-	-	+	-	27	0	0	-	0	-
10	+	-	-	+	+	28	0	0	-	0	+
11	+	+	-	+	-	29	0	0	+	0	-
12	-	+	-	+	+	30	0	0	+	0	+
13	+	-	+	+	-	31	0	0	0	-	-
14	-	-	+	+	+	32	0	0	0	-	+
15	-	+	+	+	-	33	0	0	0	+	-
16	+	+	+	+	+	34	0	0	0	+	+
17	0	0	0	0	-	35	0	0	0	0	-
18	0	0	0	0	+	36	0	0	0	0	+

Figure 5.

Response surface experimental design and measurements. (A) Agarose gel band intensities measure the amount of PCR product for each experimental run. Runs 1–18 each have 2 technical replicates nested in 3 biological replicates and runs 19–36 each have 2 technical replicates. The blocking structure for the gel staining step (A–D) is shown to the right of the gel image. (B) Intensity values are normalized and quantified by signal-to-noise ratio (SNR) for the 36 runs. (C) Experimental design for the response surface model.

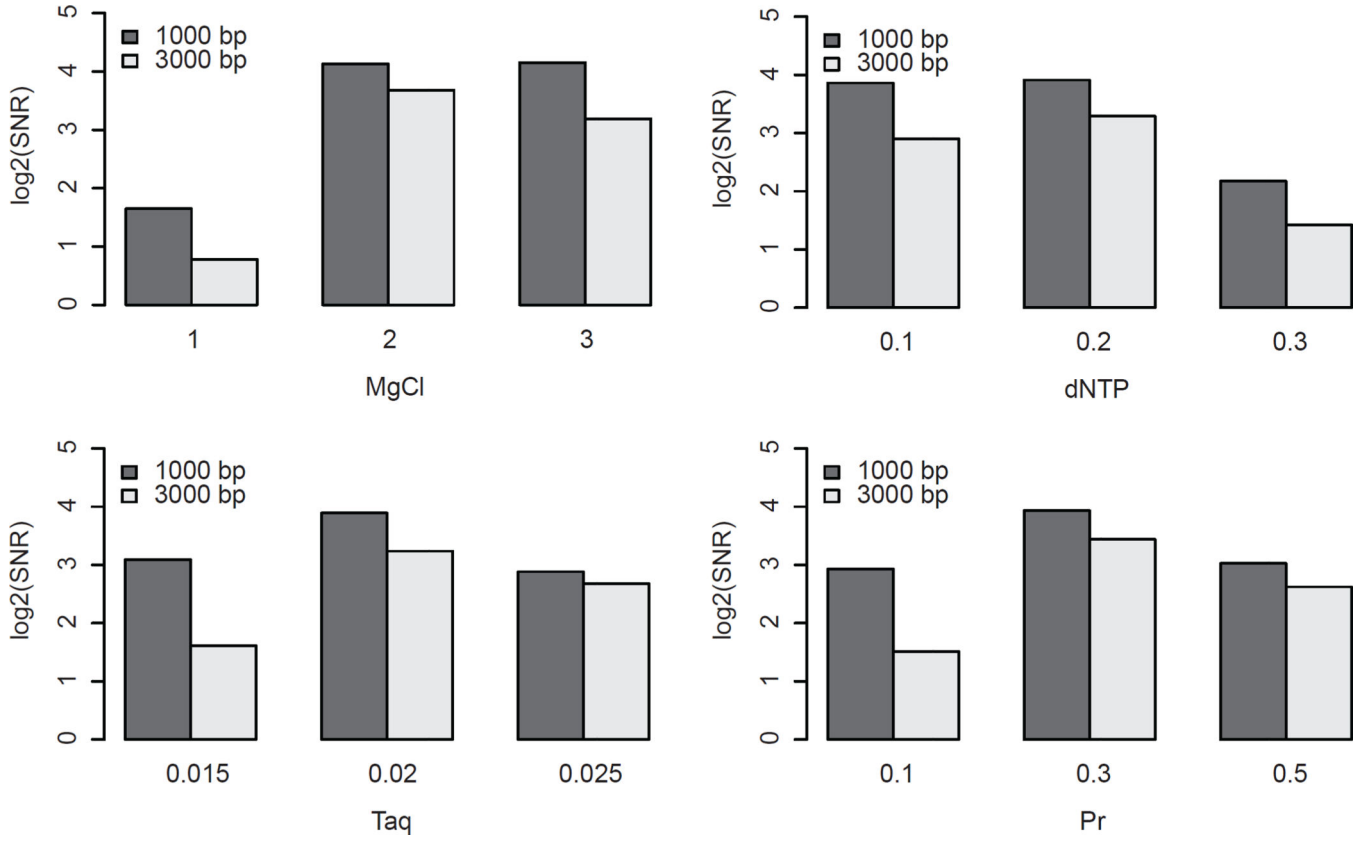


Figure 6.

Control factor main effects by template length. Yield as measured by SNR is higher across all factor levels for shorter (1000bp) template length.

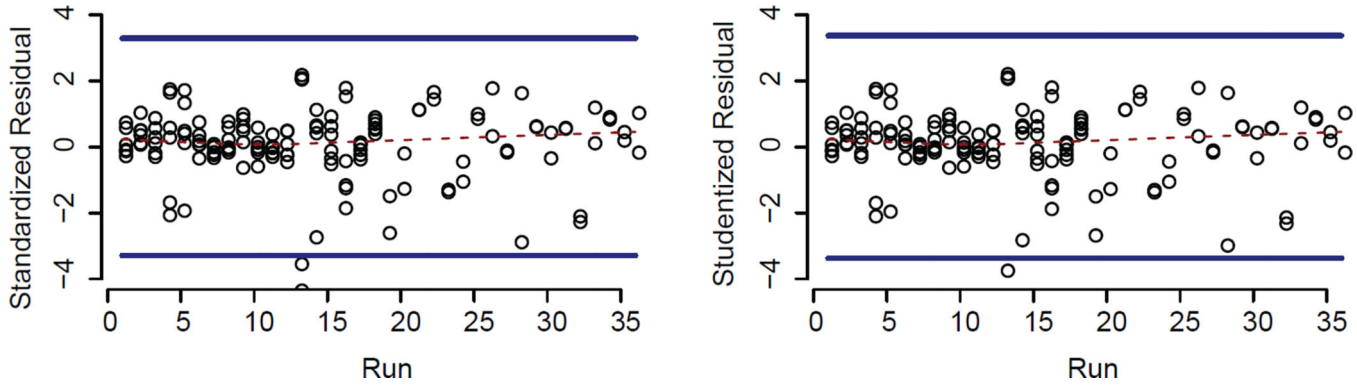


Figure 7.

Residual-based influence measures to identify outliers. The line indicates a outlier threshold for the particular measure. Only six points are called outliers out of a sample size of 144.

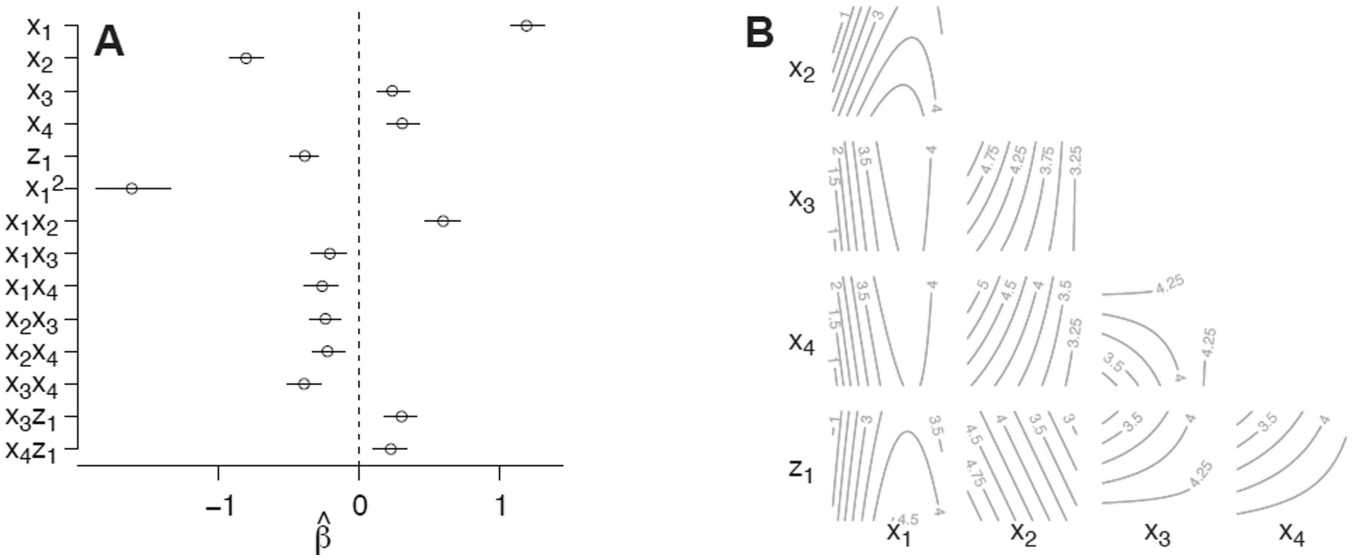


Figure 8.

Response surface model coefficients and predictions for control factors *x* and noise factors *z*. (A) Coefficients for the linear model with quadratic and two-way interactions. (B) Model yield isoclines show that yield is sensitive to some combinations of factors changes and robust to other factors.

A

	Standard	Deterministic	VaR	CVaR
MgCl(mM)	2.00	2.08	2.15	2.11
dNTP(mM)	0.20	0.10	0.10	0.10
Taq(U/ul)	0.020	0.015	0.016	0.019
Pr(uM)	0.40	0.10	0.50	0.50
Cost ($/50ul)	0.8623	0.6209	0.6776	0.7882
Pr(fail | TL=3kb)	0.0417	0.5000	0.0010	0.0004

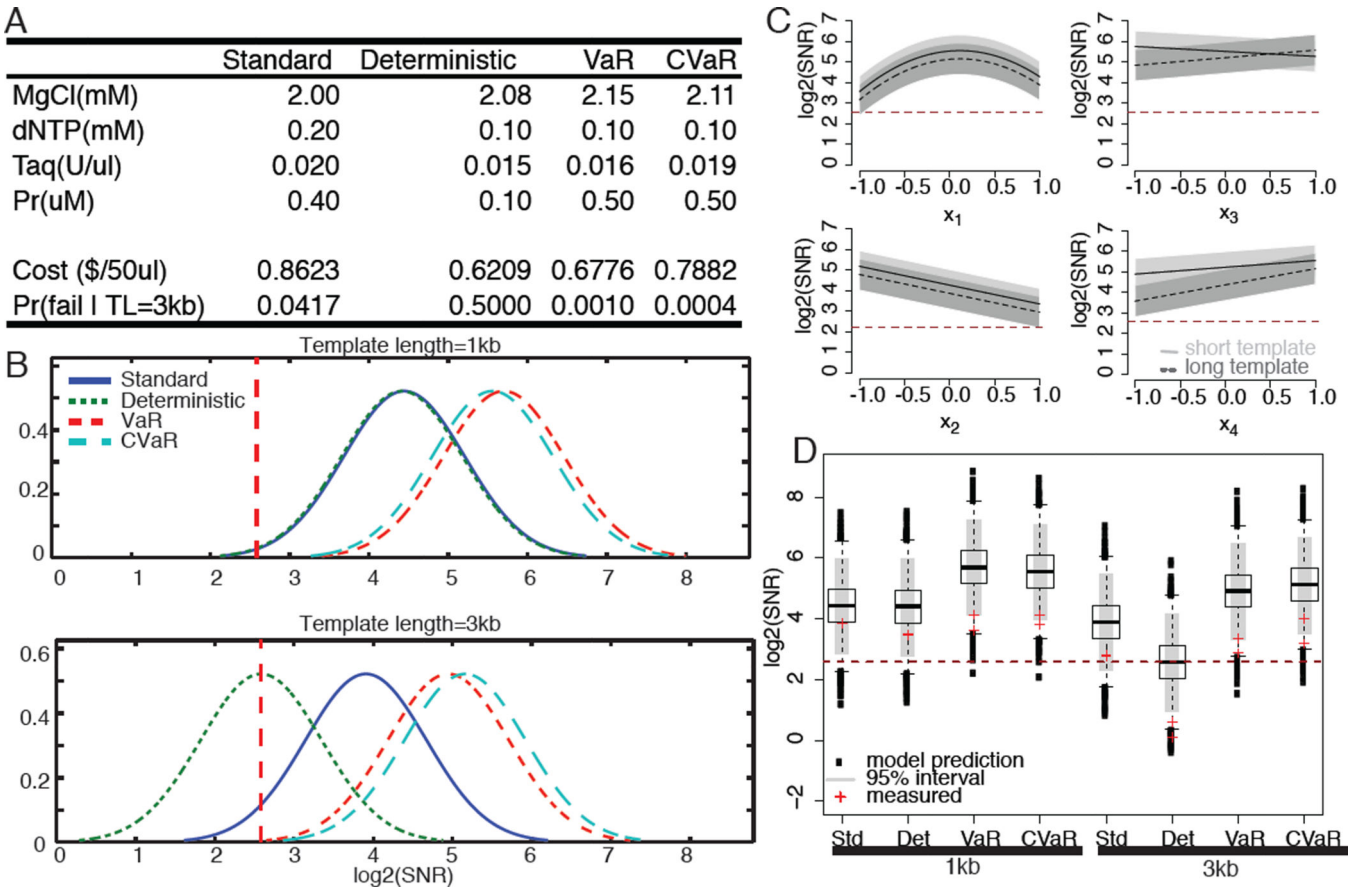


Figure 9.

Robust optimization solutions and validation experiments. (A) Optimal levels of the PCR master mix components for deterministic, VaR and CVaR optimization strategies compared to the standard manufacturer's recommended solution. (B) Probability distribution functions for the predicted response for the four master mix solutions. The minimum yield that is visible on an agarose gel is shown as a dashed vertical line. (C) Sensitivity analysis of the yield to large variations in each master mix factor for short and long template length scenarios. (D) Predicted response and independent validation experimental data shows that the model fit is accurate and the deterministic optimization solution has a high failure rate for longer template length.

Table 1

Experimental factor levels

Factors	Low	High	Coding
M: $MgCl_2$(mM)	1	3	$x_1 = (M - 2)$
N: dNTP (mM)	0.1	0.3	$x_2 = (N - 0.2)/0.1$
T: Taq (U/μL)	0.015	0.025	$x_3 = (T - 0.02)/0.005$
P: Primer (each μM)	0.1	0.5	$x_4 = (P - 0.3)/0.2$
L: Template length (bp)	1000	3000	$z_1 = (L - 2000)/1000$
C: Template concentration (ng/μL)	0.1	10	$z_2 = \log_{10}(C)$

Table 2

Typical PCR reaction cost structure

	unit cost	unit vol.	unit conc.	final conc.	50μL rxn cost
10× Buffer	—	1.5ml	10×	1×	—
$MgCl_2$	—	1ml	50mM	2.0mM	—
dNTP	$233	2.5ml	10mM	0.2mM	9.3¢
Taq	$191	50μL	$5\frac{U}{\mu L}$	$0.020\frac{U}{\mu L}$	76.4¢
Fwd. Primer	$3.8	300μL	100μM	0.4μM	0.25¢
Rev. Primer	$3.8	300μL	100μM	0.4μM	0.25¢
Total					**$0.862**