SLAS Technology 28 (2023) 16–21

Contents lists available at ScienceDirect

SLAS Technology

journal homepage: www.elsevier.com/locate/slast

Improving an rRNA depletion protocol with statistical design of experiments

Benjamin M. David[a,1], Paul A. Jensen[a,b,c,1,*]

[a] Department of Bioengineering, University of Illinois Urbana-Champaign, 1406 W Green St, Urbana, IL, 61801, United States
[b] Department of Microbiology, University of Illinois Urbana-Champaign, 601 S Goodwin Av, Urbana, IL, 61801, United States
[c] Carl Woese Institute of Genomic Biology, University of Illinois Urbana-Champaign, 1206 W Gregory Dr, Urbana, IL, 61801, United States

ARTICLE INFO

Keywords:
RNA-seq
Design of experiments
Process improvement
NGS library preparation

ABSTRACT

In prokaryotic RNA-seq library preparation, rRNA depletion is required to remove highly abundant rRNA transcripts from total RNA. rRNA is so abundant that small improvements in depletion efficiency lead to large increases in mRNA sequencing coverage. The current gold-standard method for rRNA depletion makes rRNA depletion the most expensive step in prokaryotic RNA-seq library preparation. A variety of commercial and home-made methods exist to lower the cost or increase the efficiency of rRNA removal. Many of these techniques are suboptimal when applied to new species of bacteria or when the protocol or reagents need to be changed. Re-optimizing a protocol by trial-and-error is an expensive and laborious process. Systematic frameworks like the statistical design of experiments (DOE) can efficiently improve processes by exploring the quantitative relationship between multiple factors. DOE allows experimenters to find factor interactions that may not be apparent when factors are studied in isolation.

We used DOE to optimize an rRNA depletion protocol by updating reagents and identifying factors that maximize rRNA removal and minimize cost. The optimized protocol more efficiently removes rRNA, uses fewer reagents, and is less expensive than the original protocol. Our optimization required only 36 experiments and identified two significant interactions among three protocol factors. Overall, our approach demonstrates the utility of a rational, DOE framework for improving complex molecular biology protocols.

1. Introduction

RNA-seq uses high-throughput DNA sequencing to measure mRNA transcript abundance and quantify gene expression. Over 80% of the total RNA in a prokaryotic cell is rRNA that must be removed from the sample during RNA-seq library preparation to minimize the sequencing depth required to measure mRNA abundance [1]. Removing rRNA is the largest single expense when preparing an RNA-seq library, accounting for 43% of the total cost [2]. A variety of methods—both commercial and home-made—have been developed to lower cost, improve convenience, or increase the efficiency of rRNA removal [3–7].

Beginning in the 1950's, the field of quality engineering discovered that nearly any process can be improved through the statistical design of experiments (DOE) [8]. Statistical DOE is a series of techniques used to minimize the resources required to explain and eliminate variation in experimental processes [9]. Even processes that operate well can benefit from a quantitative exploration of the relationship between process variables. DOE process optimization frameworks like Response Surface Methodology (RSM) allow experimenters to vary multiple factors simultaneously and search for optimized operating conditions using a mathematical model [10]. More recently, DOE and RSM have been used to improve molecular biology protocols [11–13].

RSM can improve molecular biology protocols in three ways. First, the original protocol developers may have stopped tuning the protocol when it worked well enough for their needs [14]. RSM can find changes to the protocols that may not have been considered by the developers. Second, reagents used in the original protocol may not be available later or might be manufactured in a different way. RSM can adjust the protocols to account for changing reagents, equipment, or techniques. Finally, the protocol developers may not have considered the effects of some factors. Analyzing all factors in a protocol leads to a combinatorial explosion of variants, so scientists often focus on only a few factors or ignore interactions between factors [15]. RSM and statistical DOE allow developers to efficiently measure combinatorial effects with relatively few experiments.

* Corresponding author.
E-mail addresses: pjens@umich.edu, pjens@illinois.edu (P.A. Jensen).
[1] Present address: Department of Biomedical Engineering, University of Michigan.

https://doi.org/10.1016/j.slast.2022.09.004
Received 15 March 2022; Received in revised form 17 August 2022; Accepted 14 September 2022

We used RSM and DOE to optimize a previously developed rRNA depletion protocol that uses complementary RNA probes labeled with biotin to remove hybridized rRNA with streptavadin-coated magnetic beads [3]. Compared to commercially available alternatives, this rRNA depletion strategy is both less expensive and automation friendly, which makes it ideal for large-scale transcriptomic experiments [5,7]. The protocol was originally developed to remove rRNA from total RNA isolated from environmental communities before RNA-seq. The protocol designers optimized rRNA depletion efficiency by designing probes that target rRNAs from multiple taxa in an environmental community. In this work, we optimized the rRNA depletion protocol by using updated reagents and finding reagent stoichiometry settings that maximize depletion efficiency and minimize cost in a single bacterial species.

2. Methods

2.1. RNA and DNA extraction

Streptococcus mutans strain UA159 was grown in THY (Todd Hewitt Broth + 0.5% yeast extract, Millipore-Sigma) at 37 °C and 5% CO_2 and harvested during mid-log phase growth. DNA was extracted using the DNeasy® UltraClean Microbial Kit (QIAGEN) and stored at −20 °C. For RNA extraction, cultures were centrifuged, aspirated, and re-suspended in TRIZOL reagent (Ambion) and $\approx 200\ \mu L$ of 0.1 mm diameter silica beads (BioSpec). Samples were homogenized for 3 min in two 90 s intervals at 1600 rpm (OHAUS homogenizer), then heated at 65 °C for 5 min. RNA extraction and purification were performed using the DirectZol® RNA Minprep Plus Kit (Zymo). Extracted RNA was stored at −80 °C.

2.2. Probe design and synthesis

A detailed protocol for probe design and rRNA depletion is provided in Supplementary Material 5. Primers (IDT) for amplifying probe templates targeted the *S. mutans* rRNA genes (Supplementary Table 4) and included a T7 promoter sequence at 5' end of the reverse primer. The DNA templates were PCR amplified using Q5 DNA Polymerase (NEB), purified using the GeneJet PCR Purification Kit (Thermo), and size verified by gel electrophoresis. *In vitro* transcription (IVT) was performed with the HiScribe™ T7 RNA Synthesis Kit (NEB). For probes in the original protocol, the biotin-labeled UTP concentration was 50% (Supplementary Material 3). The optimized probes contained 20% biotin-labeled UTP for the 16s and 23s probes and 50% for the 5s rRNA. Following IVT, the probes were purified with the Monarch® RNA Cleanup Kit (500 µg, NEB).

2.3. rRNA depletion

rRNA probes were hybridized to total RNA at 70 °C for 5 min, followed by a ramp down to 25 °C by 5 °C increments for 1 min each. Meanwhile, regular (NEB S1420S) or hydrophilic (NEB S1421S) streptavadin magnetic beads were aliqoted and washed in 0.1 N NaOH followed by two washes in 1× SSC buffer (Invitrogen). Hybridized RNA and probes were diluted to the original bead volume using a 1× SSC and 20% formamide solution and incubated at room temperature for 3 min prior to bead capture. RNA and probe hybrids were bound to the washed beads for 10 min at room temperature and separated using a magnetic rack. The rRNA depleted supernatant was purified with the Monarch®RNA Cleanup Kit (10µg) before downstream analysis.

2.4. qPCR assay for rRNA depletion

qPCR primers (IDT) were designed to amplify both the rRNAs and remaining probes (Supplementary Table 5). The *S. mutans* housekeeping gene *ldh* was amplified as a reference. PCRs were performed using the Luna® Universal One Step RT-qPCR Kit (NEB) in a ThermoFisher Quantstudio 3 instrument. Cycle threshold (C_t) values were measured for the rRNA targets and *ldh* gene in both depleted and undepleted samples. Each experiment used total RNA isolated from the sample to control for biological variation. rRNA and probe abundance was calculated using the equation

$$\%\ \text{abundance} = 2^{\Delta C_t\{\text{depleted}\} - \Delta C_t\{\text{undepleted}\}} \times 100\%,$$

where

$$\Delta C_t\{\text{depleted}\} = C_t\{\text{target, depleted}\} - C_t\{ldh, \text{depleted}\}$$
$$\Delta C_t\{\text{undepleted}\} = C_t\{\text{target, undepleted}\} - C_t\{ldh, \text{undepleted}\}.$$

The average abundance of the 16s and 23s rRNAs was used as the response for our RSM experiments. Protocol benchmarking experiments used the abundance of the 16s and 23s rRNAs individually as a response.

Table 1
Factor ranges for our 3 factor RSM design were selected to satisfy operational constraints and search a design space suggested by pilot experiments.

Factor	$-\alpha$	−1	0	+1	$+\alpha$
Probe (ng)	200	322	500	678	800
RNA (ng)	100	161	250	339	400
Beads (µl)	50	70	100	130	150

Table 2
A 3 factor Central Composite Design to optimize probe, RNA, and bead levels. A CCD with 8 factorial points is rotatable when $\alpha = \sqrt[4]{8} \approx 1.682$. The % abundance of rRNA and probe was measured by qPCR. A second-order linear model was fit to the data and used to find optimal protocol settings.

Run	Probes	RNA	Beads	Abundance (%)	Predicted (%)
1	−1	−1	−1	69.2	86.4
2	1	−1	−1	1360	1270
3	−1	1	−1	24.4	23.8
4	1	1	−1	560.	574
5	−1	−1	1	23.4	20.8
6	1	−1	1	647	660.
7	−1	1	1	10.0	106
8	1	1	1	114	110.
9	0	0	0	115	159
10	0	0	0	183	159
11	0	0	0	176	159
12	$-\alpha$	0	0	4.31	-55.3
13	α	0	0	904	945
14	0	$-\alpha$	0	472	511
15	0	α	0	53.2	-3.86
16	0	0	$-\alpha$	524	561
17	0	0	α	169	115

2.5. Modeling and statistical analysis

Analysis was performed using R version 3.6.2 and the `rsm` package [16]. We provide Supplementary Data and Code that reproduce our modeling and statistical analysis. Data in Table 2 were used to fit a linear model for rRNA and probe abundance that included all first-order, two-way interaction, and pure quadratic terms:

$$\begin{aligned}\%\ \text{abundance} = {} & \beta_0 + \beta_1\,[\text{probe}] + \beta_{11}\,[\text{probe}]^2 \\ & + \beta_2\,[\text{RNA}] + \beta_{22}\,[\text{RNA}]^2 \\ & + \beta_3\,[\text{beads}] + \beta_{33}\,[\text{beads}]^2 \\ & + \beta_{12}\,[\text{probe}][\text{RNA}] + \beta_{13}\,[\text{probe}][\text{beads}] \\ & + \beta_{23}\,[\text{RNA}][\text{beads}] \\ & + \text{residual}.\end{aligned}$$

Factor significance in the model was assessed by a *t*-test. Lack of fit was assessed by an *F*-test. The model's residual standard error was calcu-

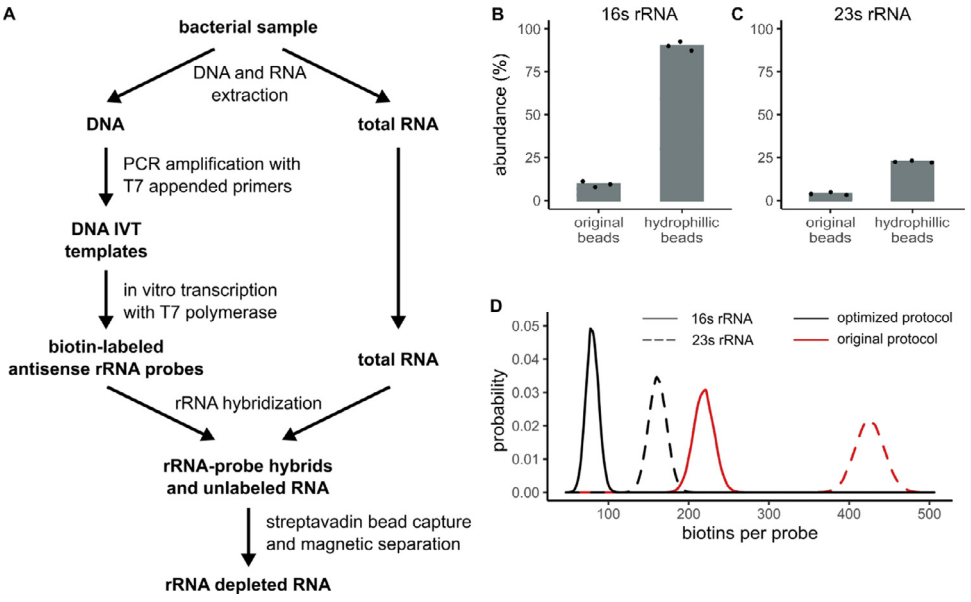

Fig. 1. Biotin-labeled antisense rRNA probes remove rRNA from total RNA. **A.** Genomic DNA and total RNA are extracted from a bacterial sample. rRNA genes are amplified by PCR with a T7 promoter sequence included in the reverse primer. Antisense rRNA probes are synthesized in an *in vitro* transcription reaction that includes biotin-labeled nucleotides. The rRNA probes are hybridized to rRNA in the total RNA and captured by strepavadin-coated magnetic beads. **B,C.** replacing the original streptavadin magnetic beads with the hydrophilic type increases the abundance of rRNA and leftover probe in the remaining total RNA as measured by qPCR. The gray bars represent the mean of 3 replicates per condition. **D.** Antisense rRNA probes in the original protocol contain biotin-labeled CTP and UTP. Our optimized protocol reduces the overall biotin concentration by removing biotin-labeled CTP and decreasing the concentration of UTP.

lated using the equation,

$$\text{RSE} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-p}}$$

where y_i and $\hat{y}_i$ are the measured and predicted rRNA and probe abundance, respectively, $n = 17$ is the number of experiments used to fit the model, and $p = 10$ is the number of parameters in the model.

A ridge analysis of the model's response surface was performed using the `ridge` function from the `rsm` package. Starting from the center of the design space, small steps are taken along the path of steepest response descent out to a predetermined distance away from the center in coded units. Experimenters can pick a location along the path to begin a new round of RSM at the new location.

3. Results

3.1. rRNA removal by hybridization

Figure 1A depicts the rRNA removal protocol described by Stewart et al. [3]. The 16s and 23s rRNA genes are amplified by PCR with a T7 promoter added to the reverse primer. In vitro transcription (IVT) creates RNA probes that are complementary to the 16s and 23s rRNAs. The IVT reaction includes biotin-labeled cytosine and uracil ribonucleotides, adding biotin along the backbone of each probe. The antisense probes are mixed with the total RNA and hybridize to the rRNA. The hybrids bind to streptavidin-coated magnetic beads and are separated from the other RNA. An RNA-seq library is prepared from the rRNA depleted samples.

Prior to optimization by DOE and RSM, we exchanged the original streptavadin magnetic beads for a new hydrophilic type. The hydrophilic beads are recommended for use with nucleic acids because they exhibit a lower rate of non-specific nucleic acid binding; however, they have a lower overall binding capacity than the original beads [17](NEB, Personal Communication). We tested the effect of switching the streptavadin bead type by first designing probes and depleting rRNA from total RNA samples from the bacterium *Streptococcus mutans* and then measuring the abundance of rRNA and probe remaining by qPCR. Both the remaining rRNA and probe were measured simultaneously because any remaining probe will be sequenced along with the rest of the total RNA sample, negating the benefit of rRNA depletion; therefore, we minimized both the amount of remaining rRNA and probe. Following the original protocol and using the original magnetic beads, we observed an rRNA and probe abundance of 9.7% and 3.9% for the 16s and 23s rRNAs, respectively (Fig. 1B,C). (The abundance is relative to the concentration of rRNA in the samples before depletion.) As expected, when using the hydrophilic beads, the measured abundance increased to 90.3% and 22.5%. Because our assay measures both leftover probe and rRNA, these results suggested that excess probe remained in the samples.

We hypothesized that probes remained in the sample because the streptavidin on the hydrophilic beads was saturated with biotin. The original protocol recommends adding 800 ng each of the 16s and 23s probes to 400 ng of total RNA, creating a 2-fold excess of probe to rRNA. Additionally, 50% of the UTPs in the IVT reaction had biotin labels, adding an average of 219 and 425 biotins per molecule of 16s and 23s probe, respectively. (To limit the number of reagents, we chose to not include CTP and double the concentration of UTP from 25% to 50% to match the original protocol (Supplementary Material 3).) To reduce the number of biotins per probe in our optimized protocol, we used only 20% of labeled UTPs. The lower concentration adds an average of 80 biotins per molecule in the shorter 16s probe with less than 1 in $5x10^{38}$ probes having no biotin modification (Fig. 1D). Unfortunately, decreasing the number of biotins did not improve the amount of probe captured by the beads, and much of the probe remained in the sample. However, decreasing the concentration of labeled UTP and removing the CTP reduced the cost of the probe synthesis reaction by 55% (Supplementary Material 4.2).

Varying biotin concentration alone failed to reproduce the results in the original manuscript. Rather than testing additional factors one at a time, we switched to RSM to identify factors and interactions that improve rRNA depletion.

3.2. RSM optimization

Optimization by RSM has three stages (Fig. 2A) [10,18]. First, an experimental design is selected with each experiment (a *run*) having a unique combination of settings for all factors. Varying factors simultaneously makes RSM designs efficient, and the design is chosen so the effects of individual factors and their interactions can be estimated. Second, the experiments are carried out and the *response*—in our case the abundance of rRNA and probe measured by qPCR—is recorded for each run. A linear model is fit to the experimental data to estimate main (first-order) effects, two way interactions, and pure quadratic effects for each factor. Finally, the model directs the search for new factor settings that

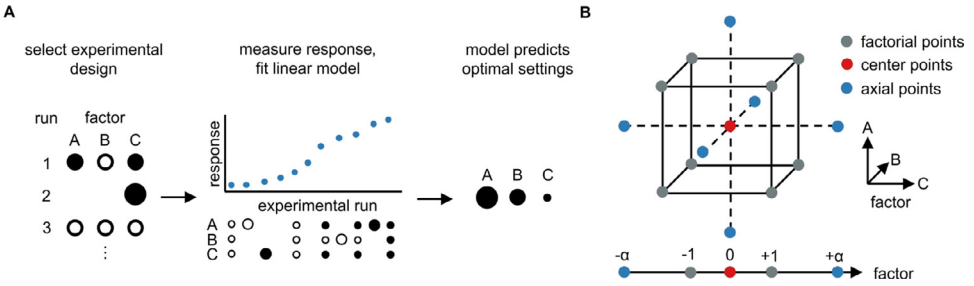

Fig. 2. Response surface methodology (RSM) is used for process optimization. **A.** RSM is performed in three stages. First, a multi-factorial experimental design is selected. Second, the experiment is carried out, the response is measured, and a mathematical model is fit to the response surface. Third, the model guides the search for optimal settings. **B.** A 3-factor rotatable central composite design (CCD) was chosen to assess first-order, two-way interaction, and quadratic effects. The CCD has a factorial core to map first-order and two-way interaction effects, and center and axial points to measure quadratic effects. For a rotatable CCD, the coded level for the axial points (α) is set at $\sqrt[4]{F}$, where F is the number of factorial points.

optimize the response. The new factor settings are tested, and the experimenter can repeat the RSM process to further improve the process in a new design space.

3.2.1. Selection of factor levels

Compared to other DOE methodologies, RSM requires a relatively large number of runs per factor and varies each factor over three to five levels. When selecting an experimental design, experimenters first need to define appropriate and testable ranges for each factor [19]. In some cases, the testable range may be defined by operational constraints, such as a minimum or maximum working volume. In others, an appropriate range needs to be defined by prior knowledge of the system or determined empirically.

Our RSM design included three factors: probe mass, total RNA mass, and hydrophilic streptavidin magnetic bead solution volume. To find an appropriate range for the probe mass, we performed a 13 run pilot experiment that that varied the concentration of each probe independently over five levels spanning 400–1600 ng per probe (Supplementary Table 1). We measured the abundance of rRNA and probe remaining in the total RNA sample, and we observed that runs with the lowest overall probe concentration (1000–1400 ng combined) performed best. These results suggested the optimal level of probes may be lower than we previously expected, and we subsequently varied the probe mass in our RSM design between 200–800 ng per probe. We also fixed the ratio of 16s and 23s probes at 1:1 by mass for all runs.

Both the total RNA mass and bead volume ranges were determined by operational constraints. We require at least 100 ng of total RNA to prepare an RNA-seq library and used the level in the original protocol (400 ng) as an upper limit. Finally, the volume of beads must be no less than the combined volume of the probes and total RNA to ensure proper capture, therefore, the concentration of the probes and the well size of our 96-well plates constrained the bead volume to between 50 and 150 µl.

3.2.2. The central composite design for RSM

We selected a rotatable central composite design (CCD) [9,18,20] with five levels for each factor, coded as $-\alpha, -1, 0, +1$, and $+\alpha$ (Fig. 2B). The CCD consists of a core of factorial points to estimate first-order effects and interactions between factors along with a set of axial and center points to estimate quadratic factor effects. The CCD provides optimal estimates of factor interactions and quadratic effects. A model trained using data from a CCD is most precise at the center of the CCD (when each factor is at or near its "0" level), and its variance increases as one moves away from the center. A CCD is said to be rotatable if the change in the model's variance is independent of the direction one moves away from the center, and a CCD can be made rotatable by setting the value of α equal to $\sqrt[4]{F}$, where F is the number of factorial points ($F = 8$ in our case). Table 1 shows the levels of each factor in our CCD.

The final, unreplicated, three-factor design includes 17 runs (Table 2). We measured rRNA and probe abundance by qPCR as the response. Figure 3A shows the abundance of each run plotted in descending order of response. This run ordering confirmed our observations that lower levels of probe reduces rRNA and probe abundance.

3.2.3. Second-order linear model to predict rRNA and probe abundance

We fit a linear model to the response data to quantify the relationship between rRNA and probe abundance and the three factors. The fitted model (showing only significant effects) is

$$\begin{aligned}\%\ \text{abundance} = {} & 159 + 297\,[\text{probe}] + 101\,[\text{probe}]^2 \\ & - 153\,[\text{RNA}] \\ & - 133\,[\text{beads}] \\ & - 159\,[\text{probe}][\text{RNA}] - 137\,[\text{probe}][\text{beads}].\end{aligned}$$

The model uses coded factors with levels -1.68, -1, 0, $+1$, and $+1.68$. The model is nonlinear with significant main effects, two way interactions, and pure quadratic terms ($p < 0.05$ by t-test), and no significant lack of fit ($p > 0.19$ by F-test) (Supplementary Table 2). The model's adjusted R^2 is 0.96, and the residual standard error is 69.9%. The residual standard error appears to be driven by points with either an extremely high or low abundance. A Normal-QQ plot indicates that the model residuals are approximately normally distributed (Supplementary Figure 1).

As mentioned above, adding more probe increases the abundance of rRNA. Because our qPCR assay measures both rRNA and probe, the increased abundance may be caused by leftover probe that is not removed by the beads. The significant quadratic effect creates a nonlinear relationship between probe concentration and rRNA depletion (Supplementary Figure 2). For example, the model predicts that increasing the probe mass from 678 ng (the +1 level) to 800 ng (the +1.68 level) while holding other factors at the 0 level would increase the measured rRNA and probe abundance by 69%. We believe the nonlinear effects of adding probe are due to bead saturation because of the significant interaction between the probes and beads. Adding more beads is predicted to decrease the abundance of rRNA and probe while offsetting the detrimental effects of high probe levels.

The most interesting result from the model is the effect of total RNA. We expected that increasing total RNA would increase the percent abundance, but the model predicts the opposite—more RNA increases the relative depletion. There is a strong interaction between RNA and probe levels, making it difficult to interpret the effects of RNA directly. The levels of total RNA, probe, and beads must be considered simultaneously when optimizing the protocol. Changing any one of the levels without adjusting the others can easily overwhelm the probe's ability to bind rRNA and the beads' ability to remove all the probes.

3.2.4. Model-guided search for optimal protocol settings

We used numerical optimization to search for an improved protocol. A quadratic model can have a single optimum or a saddle point that indicates there is a tradeoff between different factors. Eigenanalysis of our model indicates a saddle point, so there is not a single best set of factor levels that maximize rRNA depletion. Moreover, the saddle point lies outside of the region explored by our experiments, and caution must

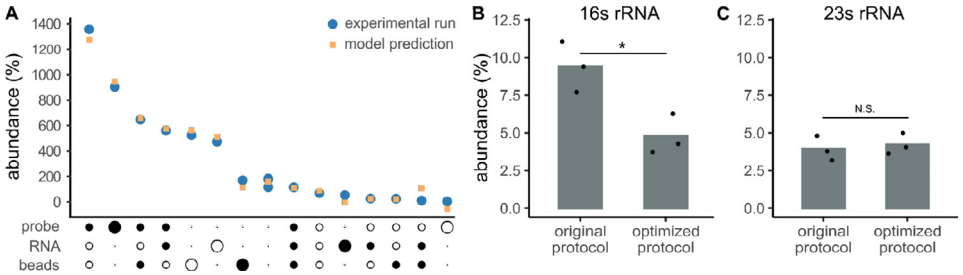

Fig. 3. A second-order linear model predicts how rRNA depletion depends on the amount of probe, total RNA, and streptavadin beads. The response measures both rRNA and unbound probe remaining in the sample. **A.** A factor-and-response plot shows the average 16s and 23s rRNA and probe abundance (blue circle) along with the model prediction (orange square) for each experimental run. Factor settings are encoded on the horizontal axis with their size proportional to the absolute value of their coded level. An unfilled circle indicates a negative coded value, while a filled circle indicates a positive coded value. **B,C.** The final optimized protocol removes more 16s rRNA ($p = 0.02$ by ANOVA) and requires fewer reagents than the original protocol. The optimized protocol removes up to 95% of the 16s and 23s rRNAs. The 16s and 23s rRNA abundances are plotted individually instead of an average as in previous experiments. The gray bars represent the mean of 3 replicates for each protocol.

Table 3
Original and optimzied rRNA depletion reagent settings.

Reagent	Original Protocol	Optimized Protocol
total RNA	400 ng	400 ng
16s rRNA probe	800 ng	200 ng
23s rRNA probe	800 ng	250 ng
% biotin-labeled UTP	50%	20%
streptavadin bead type	original	hydrophilic
streptavadin bead volume	100 µl (400 µg)	100 µl (400 µg)

be used when extrapolating so far outside the data used to fit the model. Instead, RSM practitioners recommend using a ridge analysis to follow the best direction calculated by the model for future experimentation and perform a second round of RSM at the new factor settings [21,22].

Ridge analysis directed us toward the run with the lowest probe levels (200 ng). We began a second round of RSM centered at this new point by varying the amount of total RNA and the volume of beads over a narrowed range. The central composite design can be split into two blocks: the factorial runs and the axial runs. Experimenters can first run the factorial block and use it as a screen. Factors with small or insignificant effect sizes can be removed from the later axial block to reduce the number of runs. The 6 run factorial block of our subsequent CCD indicated that runs at the center point performed best, so there was no need to continue with the axial block (Supplementary Table 3). Increasing the total RNA or reducing the volume of beads increased the percent abundance, and increasing the volume of beads did not improve depletion efficiency.

Our percent abundance measurement represents the average of the 16s and 23s rRNAs. In experiments where probes were added in equal mass, we observed that the 16s probes outperformed the 23s probes (< 5% vs. 10%). We hypothesized that since the lower molecular weight 16s probe was in a higher *molar* abundance, the probes were better able to capture the 16s rRNA. To compensate for this difference, we increased the mass of the 23s probe from 200 ng to 250 ng. Indeed, this improved 23s rRNA capture without over-saturating the beads.

3.2.5. Optimal protocol settings

The final, optimal conditions for 16s and 23s rRNA depletion appear in Table 3. Compared to the original protocol in Stewart, *et al*, our optimized protocol is more efficient at removing rRNA (4.7% vs. 9.4% abundance for 16s, 4.2% vs. 3.9% for 23s) (Fig. 3B,C). We confirmed the rRNA depletion in our optimized protocol with capillary electrophoresis (Supplementary Figure 3).

3.3. Cost analysis

The reagent costs for our optimized rRNA depletion assay appear in Supplementary Material 4. Note that costs may vary by region, scale, or supplier. The streptavidin-coated magnetic beads are the largest single expense, accounting for 62% of the total reagent cost. We tested if the volume of beads could be reduced without sacrificing the rRNA depletion efficiency. As reported in Supplementary Figure 4, the volume of beads can be reduced from 100 µl to 75 µl without loss of efficiency. Unless reducing costs is a primary concern, we recommend using 100 ul to ensure full removal of rRNA and the probes. We also used the excess of beads when adding antisense probes for the 5s rRNA, as discussed in the following section. Even with 100 µl of beads, our optimized protocol costs 14% less than the original protocol.

3.4. Designing probes for the 5s rRNA

The original protocol in Stewart, *et al.* includes probes for the 16s and 23s rRNAs, but not the 5s rRNA. We created antisense RNA probes for the 5s rRNA and measured depletion by qPCR. Initially, the depletion of 5s rRNA was poor (60–80% abundance), implying either poor hybridization to the rRNA or probes left uncaptured to the beads. We suspected that the 27 uracils in the 116 bp 5s rRNA created too few sites for potential biotin labeling. Indeed, when using 20% labeled UTPs (the level used for the 16s and 23s probes) approximately 7% of the probes would have less than three biotin modifications. Increasing the concentration of labeled uracils on the 5s rRNA probes from 20% to 50% improved the depletion of the 5s rRNA to levels similar to the 16s and 23s rRNAs.

The final protocol for rRNA depletion with 5s probes includes 200 ng 16s probe, 250 ng 23s probe, 50 ng 5s probe, 400 ng total RNA, and 100 µl beads. A step-by-step protocol is available in Supplementary Material 5.

4. Discussion

This work used statistical DOE and an RSM framework to optimize an rRNA depletion protocol. Our study demonstrated three benefits of protocol re-optimization by RSM. First, there may be room for improvement in the original protocol. Our optimized rRNA depletion protocol removes more rRNA than the original protocol and uses fewer reagents. Both increasing the efficiency and decreasing the cost of rRNA depletion improve mRNA sequencing workflows. Second, reagents can change over time. The streptavadin beads used by Stewart, *et al.* have a higher binding capacity but also a higher rate of non-specific binding than the new hydrophilic bead type. Switching to the hydrophilic beads required changes in probe levels, biotin concentration, and total RNA in the reaction. Third, the original protocol developers may not have characterized important factor interactions. RSM identifies interactions that are missed by one-at-a-time optimization. Interactions make it difficult to predict the relationship between factors and the response. For example, the optimal concentration of probe depends on the amount of beads, and we cannot explain the beneficial interaction between probe and RNA concentration. Our quantitative model allowed us to find improved conditions despite the nonlinear interactions between factors.

Optimizing other protocols may require studying the effects of more factors than we examined in this work. While adding additional factors may lead to more robust optimization, each additional factor nearly

doubles the number of experiments required for the central composite design we chose for this study. Often, a balance must be struck between model complexity and experimental costs. We recommend a fractional-factorial or Placket-Burman screening design for studies with a large number of factors to reduce the number of experiments needed for optimization by RSM.

We do not expect that probes designed for a different bacterium will significantly change the performance of the optimized protocol. We analyzed the rRNA sequences of 25 organisms of varying phylogeny whose mRNA GC content varied from 18–75%. The rRNA sequences exhibited a high degree of similarity and had much less variation in GC content than their respective mRNAs (Supplementary Figures 5,6). However, the rRNA depletion protocol could be re-optimized for other organisms, particularly if a researcher plans to do extensive sequencing on a particular organism. We provide guidelines for designing probes in the Supplementary Material and recommend using the approach outlined in this study to re-optimize the protocol if necessary.

Users who design probes for rRNAs or other transcripts may need to empirically determine the optimal ratio of labeled to unlabeled nucleotides. It is the total number of biotin modifications in a probe, not the ratio of labeled to unlabeled nucleotides, that matters. Our experiments indicated that for large transcripts such as the 16s and 23s rRNAs, a large excess of biotin labels unnecessarily increased cost. Conversely, having too few modifications resulted in inefficient capture, as was the case for our original 5s rRNA probes. A referee also noted that they split their larger 16s and 23s probes into multiple shorter sequences. This change may improve the IVT probe yield and quality without significantly impacting rRNA hybridization or capture.

Small increases in rRNA depletion efficiency can significantly increase mRNA abundance in sequencing libraries. Because rRNAs represent >80% of the total RNA in a bacterial cell, they dominate sequencing libraries unless they are depleted. Our optimized protocol removes >95% of the rRNA, so non-rRNA transcripts make up 85% of the remaining transcripts. There is still room for improvement. If rRNA depletion efficiency increased to 97.5%, non-rRNA transcripts would make up 92% of the sample, an 8% increase over our optimized protocol. Additionally, one could consider designing probes that target tRNAs or other highly abundant transcripts to further increase the proportion of mRNA in the library [23,24]. Any improvements to mRNA enrichment improve the depth and lower the effective cost of downstream sequencing.

DOE and RSM can efficiently optimize protocols, with few runs per factor. These techniques allow protocol developers to consider more factors and produce better protocols. DOE and RSM can also quantify the relationship between factors and the response, which is helpful when making rational, data-driven decisions about design tradeoffs. We hope this work increases the use of DOE and RSM in molecular and synthetic biology.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We thank Bill Metcalf for introducing us to the original rRNA depletion protocol and Walden Li for independently testing the optimized protocol. This work was supported by the National Institutes of Health grant GM138210. BMD is supported in part by an Illinois Distinguished Fellowship.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at 10.1016/j.slast.2022.09.004.

References

[1] Westermann AJ, Gorski SA, Vogel J. Dual RNA-seq of pathogen and host. Nat Rev Microbiol 2012;10(9):618–30. doi:10.1038/nrmicro2852.
[2] UIUC Roy J Carver Biotechnology Center. Pricing. 2022. https://biotech.illinois.edu/htdna/pricing.
[3] Stewart FJ, Ottesen EA, DeLong EF. Development and quantitative analyses of a universal rRNA-subtraction protocol for microbial metatranscriptomics. ISME J 2010;4(7):896–907. doi:10.1038/ISMEJ.2010.18.
[4] Culviner PH, Guegler CK, Laub MT. A simple, cost-effective, and robust method for rRNA depletion in rna-sequencing studies. mBio 2020;11(2). doi:10.1128/MBIO.00010-20.
[5] Benes V, Blake J, Doyle K. Ribo-Zero gold kit: improved RNA-seq results after removal of cytoplasmic and mitochondrial ribosomal RNA. Nat Methods 2011;8(11) iii–iv. doi:10.1038/nmeth.f.352.
[6] Choe D, Szubin R, Poudel S, Sastry A, Song Y, Lee Y, et al. RiboRid: a low cost, advanced, and ultra-efficient method to remove ribosomal RNA for bacterial transcriptomics. PLOS Genet 2021;17(9):e1009821. doi:10.1371/JOURNAL.PGEN.1009821.
[7] Herbert ZT, Kershner JP, Butty VL, Thimmapuram J, Choudhari S, Alekseyev YO, Fan J, Podnar JW, Wilcox E, Gipson J, Gillaspy A, Jepsen K, Bon-Durant SS, Morris K, Berkeley M, LeClerc A, Simpson SD, Sommerville G, Grimmett L, Adams M, Levine SS. Cross-site comparison of ribosomal depletion kits for Illumina RNAseq library construction. BMC Genomics 2018;19(1):1–10. doi:10.1186/S12864-018-4585-1/TABLES/1.
[8] Fisher RA. Design of experiments. Oliver and Boyd; 1935. ISBN 978-0028446905
[9] Lawson J. Design and analysis of experiments with R. Chapman and Hall; 2014. ISBN 978-1439868133
[10] Box GEP, Draper NR. Response surfaces, mixtures, and ridge analyses. Wiley-Interscience; 2007. ISBN 978-0-470-05357-7
[11] Singleton C, Gilman J, Rollit J, Zhang K, Parker DA, Love J. A design of experiments approach for the rapid formulation of a chemically defined medium for metabolic profiling of industrially important microbes. PLOS ONE 2019;14(6):e0218208. doi:10.1371/JOURNAL.PONE.0218208.
[12] Onyeogaziri FC, Papaneophytou C. A general guide for the optimization of enzyme assay conditions using the design of experiments approach. SLAS Discov. 2019;24(5):587–96. doi:10.1177/2472555219830084.
[13] Flaherty P, Davis RW. Robust optimization of biological protocols. Technometrics 2015;57(2):234. doi:10.1080/00401706.2014.915890.
[14] Roux KH. Optimization and troubleshooting in PCR. Cold Spring Harb Protoc 2009;2009(4):pdb.ip66. doi:10.1101/PDB.IP66.
[15] Barnes WM. Long and accurate PCR. Cold Spring Harb Protoc 2006;2006(1). doi:10.1101/PDB.PROT4094. pdb.prot4094
[16] Lenth RV. Response-surface methods in R, using RSM. J Stat Softw 2010;32(7):1–17. doi:10.18637/JSS.V032.I07.
[17] New England BioLabs. Hydrophilic streptavidin magnetic beads | NEB. 2022. https://www.neb.com/products/s1421-hydrophilic-streptavidin-magnetic-beads#Quality, Safety & Legal.
[18] Myers RH, Montgomery DC, Anderson-Cook CM. Response surface methodology: process and product optimization using designed experiments. 4th ed. Wiley; 2016. ISBN 978-1-118-91601-8
[19] Box GEP, Hunter JS, Hunter WG. Statistics for experimenters: design, innovation, and discovery. 2nd ed. Wiley; 2005. ISBN 978-0-471-71813-0
[20] Box GEP, Wilson K. On the experimental attainment of optimum conditions on JSTOR. J R Stat Soc 1951;13(1). https://www.jstor.org/stable/2983966
[21] Hoerl AE. Optimum solution of many variables equations. Chem Eng Prog 1959;55:67–78.
[22] Draper NR. 'Ridge analysis' of response surfaces. Technometrics 1963;5:469–79.
[23] Van Goethem A, Yigit N, Everaert C, Moreno-Smith M, Mus LM, Barbieri E, et al. Depletion of tRNA-halves enables effective small RNA sequencing of low-input murine serum samples. Sci Rep 2016;6(1):1–11. doi:10.1038/srep37876.
[24] Engelhardt F, Tomasch J, Häussler S. Organism-specific depletion of highly abundant RNA species from bacterial total RNA. Access Microbiol 2020;2(10). doi:10.1099/ACMI.0.000159.