SAND2021-13373

SANDIA REPORT
SAND20XX-XXXX
Printed September 2021

Platform for Single-Cell Dual RNA Sequencing of Host-Pathogen Interactions

Ramdane Harouaka, Anna Fisher, Ryan Wyllie, Benjamin David, Paul Jensen

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico
87185 and Livermore,
California 94550

Issued by Sandia National Laboratories, operated for the United States Department of Energy by National Technology & Engineering Solutions of Sandia, LLC.

NOTICE: This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government, nor any agency thereof, nor any of their employees, nor any of their contractors, subcontractors, or their employees, make any warranty, express or implied, or assume any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represent that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government, any agency thereof, or any of their contractors or subcontractors. The views and opinions expressed herein do not necessarily state or reflect those of the United States Government, any agency thereof, or any of their contractors.

Printed in the United States of America. This report has been reproduced directly from the best available copy.

Available to DOE and DOE contractors from
U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Telephone:	(865) 576-8401
Facsimile:	(865) 576-5728
E-Mail:	reports@osti.gov
Online ordering:	http://www.osti.gov/scitech

Available to the public from
U.S. Department of Commerce
National Technical Information Service
5301 Shawnee Rd
Alexandria, VA 22312

Telephone:	(800) 553-6847
Facsimile:	(703) 605-6900
E-Mail:	orders@ntis.gov
Online order:	https://classic.ntis.gov/help/order-methods/

National Nuclear Security Administration

ABSTRACT

The aim of this project was to advance single cell RNA-Seq methods toward the establishment of a platform that may be used to simultaneously interrogate the gene expression profiles of mammalian host cells and bacterial pathogens. Existing genetic sequencing methods that measure bulk groups of cells do not account for the heterogeneity of cell-microbe interactions that occur within a complex environment, have limited efficiency, and cannot simultaneously interrogate bacterial sequences. In order to overcome these challenges, separate biochemistry workflows were developed based on a Not-So-Random hexamer priming strategy or libraries of targeted molecular probes. Computational tools were developed to facilitate these methods, and feasibility was demonstrated for single-cell RNA-Seq for both bacterial and mammalian transcriptomes. This work supports cross-agency national priorities on addressing the threat of biological pathogens, and understanding the role of the microbiome in modulating immunity and susceptibility to infection.

This page left blank

CONTENTS

LIST OF FIGURES

This page left blank

ACRONYMS AND DEFINITIONS

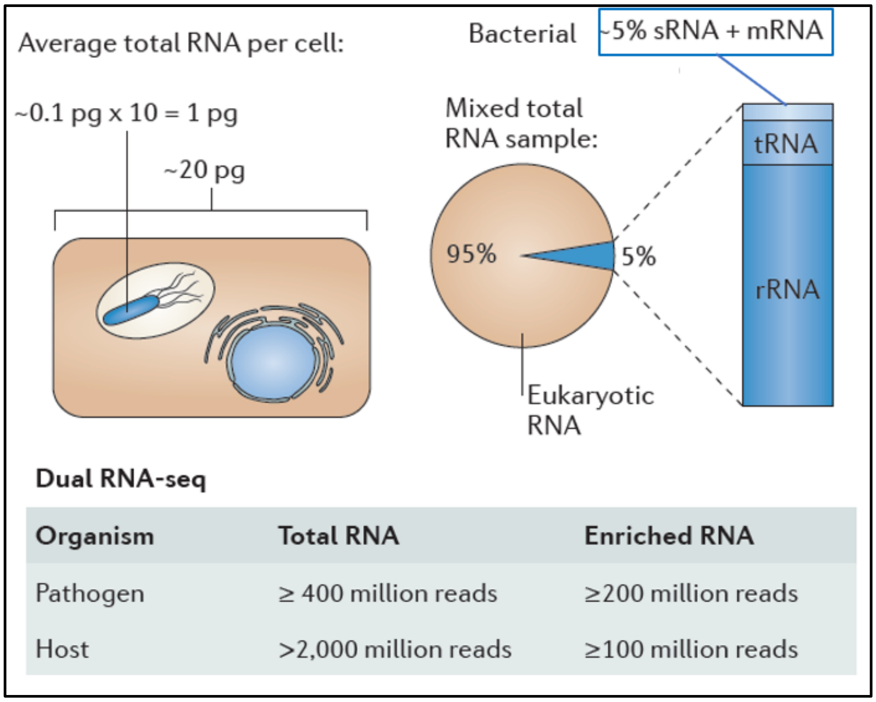
Abbreviation	Definition
DNA	Deoxyribonucleic acid
cDNA	Complementary DNA
dsDNA	Double-stranded DNA
RNA	Ribonucleic acid
rRNA	Ribosomal RNA
tRNA	Transfer RNA
mRNA	Messenger RNA
RNA-Seq	RNA sequencing
scRNA-Seq	single-cell RNA-Seq
DualRNA-Seq	dual single-cell RNA-seq
NSR primers	Not-So-Random primers
BSA	Bovine serum albumin
PEG	Polyethylene glycol
MDP	Markov Decision Process
DULQ	distribution uniformity, lower quartile
PCR	Polymerase Chain Reaction

1. INTRODUCTION

The microbiome is a diverse population of trillions of microorganisms that coexist and interact with the human body and its environment. Recent advances in genomic techniques have led to the discovery of microbiomes dynamically mediating susceptibility to infections, therapeutic efficacy, and progression of diseases including neurological disorders such as Parkinson's disease and autism [1-3]. In 2016 the White House announced a National Microbiome Initiative leading a group of 21 government agencies (including DOD, NIH and FDA) to form a Strategic Plan whose aims include "supporting interdisciplinary research" and "developing platform technologies" to advance microbiome research towards healthcare, food security, and environmental restoration [4].

The goal of this project was to develop methods for single-cell RNA-seq (scRNA-seq) for simultaneous transcriptional profiling of both a bacterium and a mammalian cell, known hereafter as "dual single-cell RNA-seq", or simply DualRNA-seq. The major challenge for DualRNA-seq is the highly skewed ratios of bacterial to mammalian RNA when a single mammalian cell is lysed with a single, much smaller, bacterial cell. Bulk RNA isolated from host/pathogen co-cultures will have 10-1000 times more host RNA than pathogen RNA. To see transcriptional changes in both organisms, previous work has 1.) used enormous amounts of sequencing until the pathogen's transcriptome can be resolved; or 2.) depleted the host RNA, usually by negative selection against the poly-A tails of eukaryotic transcripts. Neither strategy is amenable to single cell experiments. Preparing RNAseq from single cells already requires high sequencing depth and a single library preparation, so DualRNA-seq methods are too costly or technically infeasible.

Additionally, the mRNA desired for DualRNA-seq is less than 10% of both the bacterial or mammalian RNA pools, with the rest of the pools made up of the highly abundant rRNA and tRNA (Figure 1). Typically, rRNA is depleted by passing total RNA through columns that selectively bind rRNA. For mammalian RNA, the poly-A tails can be used to select mRNA directly. However, neither of these techniques can be applied to single cells. Previously the study of combined host and microbe transcriptomes was only possible by extraction of RNA from bulk samples, which averages interactions between cells across time and population heterogeneity, completely obscuring information needed to understand stability or progression of signatures that differentiate commensal (ie. microbiome) versus pathogenic activity. Single-cell dual RNA-Seq of host and pathogen cells has been proposed as a solution for retaining information linked to heterogeneity and increasing the limit of detection for rare pathogen transcripts [5]. The recent development of barcoding strategies combined with droplet-based microfluidics has enabled high-throughput processing of tens of thousands of single cells for RNA-Seq 6, 7. However, these emerging research solutions and commercially available systems (ie. 10x Chromium) fundamentally rely on barcoding off of poly-adenylated messenger RNA molecules from eukaryotic cells, and therefore cannot effectively capture RNA from bacteria 8. Furthermore, droplet-based microfluidic solutions are not compatible with maintaining cell viability after partitioning, and do not allow additional characterization of chemistries using combined techniques (eg. imaging).



Organism	Total RNA	Enriched RNA
Pathogen	≥ 400 million reads	≥200 million reads
Host	>2,000 million reads	≥100 million reads

Figure 1: Low amounts of bacterial RNA and an overabundance of rRNA complicate DualRNA-Seq. [5]

This project developed two methods to address the above problems. Our proposed innovation was to adapt a very recently published isothermal RNA-Seq library preparation protocol employing a subset of random hexamer primers (Not-So-Random primers) to avoid ribosomal RNA amplification. This approach has been demonstrated to allow total RNA-Seq (including long non-coding RNA and enhancer RNA) from mammalian cells 9. We hypothesized that this method would also allow amplification of the entire transcriptomes from both species of a cell-bacterium pair, which is unprecedented. Due to its isothermal nature, this protocol is compatible with easily fabricated open microwell formats that allow microfiltration strategies we have previously developed 10 to selectively capture viable single cell-microbe pairings. We improved computational methods to develop pools of Not-So-Random (NSR) primers to selectively amplify mRNA from both bacteria and the host cells while avoiding rRNA and tRNA. We also developed Splintlock-seq, a gene expression profiling technique that used molecular probes to target only the genes of interest in a sample. Both methods can address the small ratio of bacterial to mammalian mRNA by increasing the amount of probes or NSR primers that target the bacterial genes.

1.1. No-So-Random Primers

Highly abundant transcripts from rRNA and tRNA genes constitute up to 95% of the RNA in the cell [5]. If these transcripts are not removed before sequencing, they can vastly inflate the sequencing cost needed to quantify the abundance of mRNA. In bulk RNA from eukaryotes, mRNA transcripts can be enriched by polyT selection; however, prokaryotic mRNAs are not polyadenylated, and the highly abundant rRNA transcripts must instead be removed by physical capture with silica columns or magnetic beads [11-13]. Column- and bead-based separations are not possible in single-cell RNA-seq studies where libraries are prepared from picograms of RNA in droplets or microwells.

Some RNA-seq protocols use random hexamers to prime reverse transcription. The hexamers bind randomly across the transcriptome, so libraries made from total RNA will be dominated by rRNA. Recently, NSR primers have been used to selectively amplify non-rRNA sequences [14]. An NSR pool contains only the hexamers that are not found in the rRNA genes, so rRNA transcripts are not primed for reverse transcription and subsequent amplification. NSR pools are used to selectively prime reverse transcription in when preparing sequencing libraries. The original NSR pools were designed to avoid hybridization to rRNA transcripts [14].

This project developed a new computational framework, called OligoRL, to solve combinatorial oligonucleotide optimization problems such as designing NSR primer pools. The OligoRL framework improves NSR selection by avoiding the avoid brute-force selection of NSR primers and instead design pools of hexamers with optimal coverage and uniformity. The oligos in brute-force pools are scored individually and may not represent the best overall pool when combined. Current NSR pools are designed only to maximize the number of binding sites in the transcript, leading to skewed coverage of transcripts.

1.2. Splintlock-seq

Probe-based library preparation techniques are also designed to avoid rRNA. RASL-seq uses SplintR ligase and pairs of DNA probes to quantitatively profile gene expression [15]. In this approach, pairs of DNA probes hybridize adjacent to one another on the target mRNA and are ligated together by SplintR ligase [15]. Subsequent amplification and tagging can be accomplished by making use of known sequences in the probe tails. However, RASL-seq suffers from high levels of noise at low target RNA concentrations and requires manual probe design. The latter requirement has limited the scope of RASL-seq studies to a small subset of the transcriptome. Our approach improves upon RASL-seq library prep by building an automated, computational pipeline capable of designing thousands of probes to target the entire transcriptome. We also employ padlock probes, whose architecture improves assay sensitivity as well as specificity and allows for unbiased signal amplification.

Chorella virus DNA ligase, also known as SplintR ligase, catalyzes the formation of phosphodiester linkages between adjacent resides of single stranded DNA molecules hybridized on RNA [16]. SplintR ligase exhibits minimal bias for nucleotide identity at the ligation junction while demonstrating a high degree of sensitivity to ligation junction mismatches [17]. These qualities have led to the use of SplintR ligase in highly specific detection assays targeting miRNA isoforms [18], SNPs [17], and splice variants [19]. Assays employing SplintR ligase have demonstrated remarkable sensitivity. One recent study showed that SplintR ligase could be used in conjunction with qPCR to specifically detect attogram quantities of a specific miRNA from within 10 ng of total RNA [18].

A padlock probe is a single stranded, linear DNA oligonucleotide that contains sequence complementary to a target molecule at the 5’ and 3’ ends, with a constant non-hybridizing backbone in between. When a padlock probe binds to its target, the two ends hybridize directly adjacent to one another. The ends can then be ligated together using an enzyme such as SplintR ligase, creating a circular single-stranded DNA (ssDNA) molecule in the process. Padlock probe-based detection of nucleic acids exhibits greater targeting specificity and sensitivity than paired probe strategies due to the cooperative annealing dynamics of the two probe ends. Furthermore, circularized ssDNA molecules can be efficiently used as template for isothermal, rolling circle amplification with Phi29 DNA polymerase. Phi29 DNA polymerase exhibits very low bias, high fidelity, extremely high processivity, and strong multiple strand displacement activity [20]. By adding a primer targeting the padlock backbone, Phi29 DNA polymerase can be used to initiate a rolling circle amplification, selectively amplifying intramolecular ligation products. This improves signal-to-noise ratios as non-specific intermolecular ligation products and unreacted probes will not be amplified. The degree to which rolling circle amplification boosts signal is remarkable. At 30 °C, Phi29 DNA polymerase is capable of incorporating 2280 nt/min. The enzyme's average processivity is greater than 70 kb, indicating that a single 90 nt circularized padlock probe can be used as template to produce a concatamer containing over 775 copies in a half an hour [21].

The experimental workflow of the proposed Splintlock-seq library prep is outlined in Figure 2. A pool of padlock probes generated by our automated design pipeline will be synthesized and hybridized to isolated total RNA from the condition of interest. If a target mRNA is present, the ends of a respective padlock probe will be ligated together by SplintR ligase, generating a circularized ssDNA molecule. Signal amplification and enrichment will then be simultaneously carried out using Phi29-mediated rolling circle amplification initiated from a primer annealed to the padlock backbone. In this way, only intramolecular ligation products undergo amplification. The rolling circle amplification product is then used as template for a low-cycle number PCR with primers that add custom, dual-indexed Illumina adapters. These adapters enable the pooling of dozens of individual libraries to further reduce per library sequencing costs. Libraries will then be purified using AMPure beads and pooled in an equimolar fashion. The pooled libraries are then sequenced and used for differential gene expression analyses.

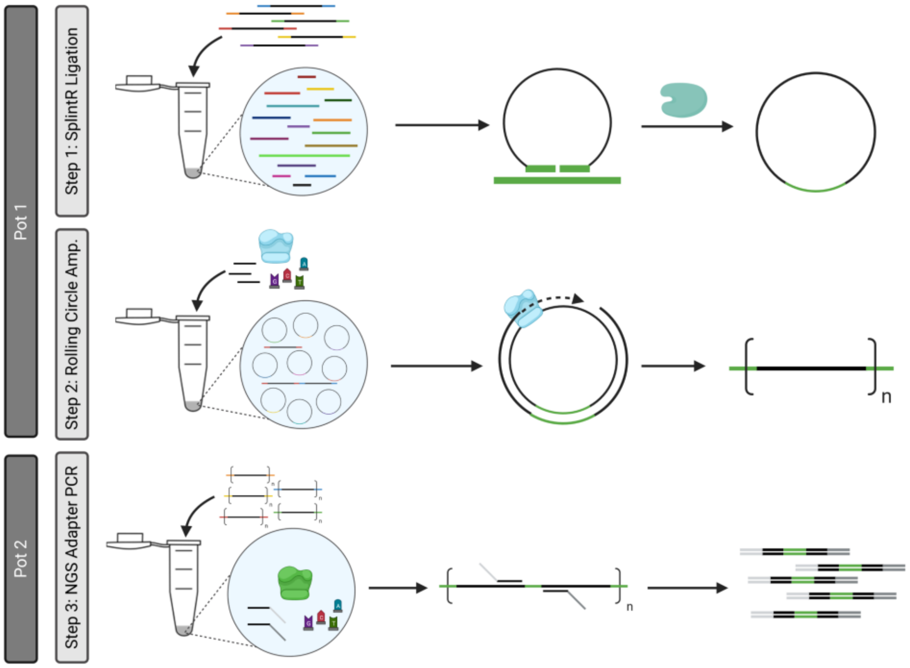


Figure 2: The Splintlock-seq pipeline.

We also developed an automated computational pipeline for the design of padlock probes. This pipeline was used to create an initial probe pool targeting all annotated transcripts in the genome. The probe pool was then assessed for ligation on rRNA and redesigned to remove non-specific probes.

2. RESULTS

2.1. Optimization of Biochemistry for Whole Transcriptome Amplification

The initial steps for optimization involved adapting the Smart-seq method [22] for use with murine lung epithelial LA-4 cells to establish a benchmark for cDNA quantification and amplification. Establishing this benchmark would allow us to evaluate reaction efficiency while optimizing the RamDA-seq method, as well as explore different methods for increasing sensitivity. Each Smart-seq optimization experiment started with LA-4 cell culture and lysis, then continued with reverse transcription and amplification. After amplification, the samples were cleaned using Ampure XP beads. Finally, quantification and/or qRT-PCR were performed to evaluate the effects of each optimization on cDNA yield and amplification.

2.1.1. Lysis Buffer

With the goal of improving RNA stability and cDNA yield, we tested the addition of 1 mg/mL bovine serum albumin (BSA) to our lysis buffer. BSA has been found to help stabilize RNA and protect it from degradation[23]. We found that BSA decreased the efficiency of Ampure XP bead cleanup and produced a 5-fold decrease in cDNA yield (Figure 3), so we proceeded with lysis buffers that did not contain BSA for future experiments.

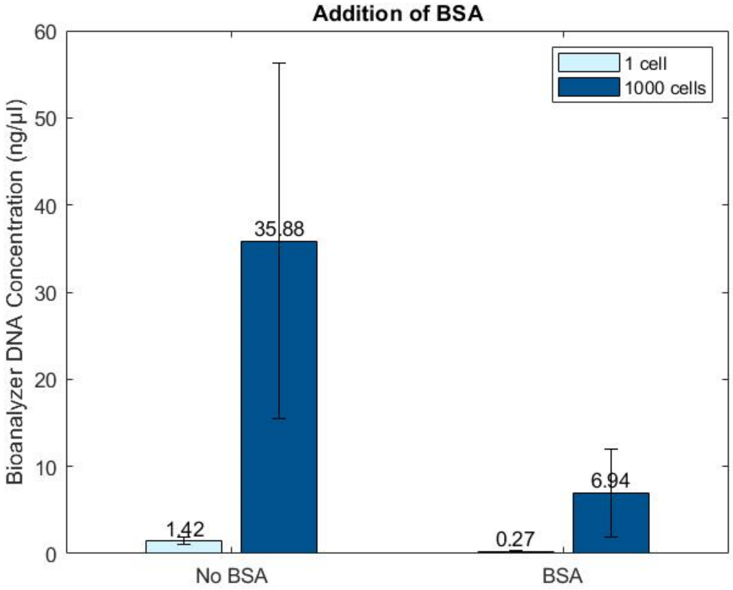


Figure 3. cDNA yield of Smart-seq samples prepared with and without the addition of BSA to the lysis buffer. All samples contained LA-4 lysate.

2.1.2. Reverse Transcription & PCR Amplification

We tested the addition of two different hydrogels during the reverse transcription and PCR reactions with the goal of increasing reaction efficiency to improve the sensitivity of Smart-seq. We first tested the addition of polyethylene glycol (PEG) 8000 to achieve a molecular crowding effect. In the molecular crowding effect, hydrogel molecules take up space in the aqueous solution, resulting in a lower effective reaction volume[24]. We tested PEG concentrations between 0 and 15% (Figure 4) and found that PEG did increase cDNA yield.

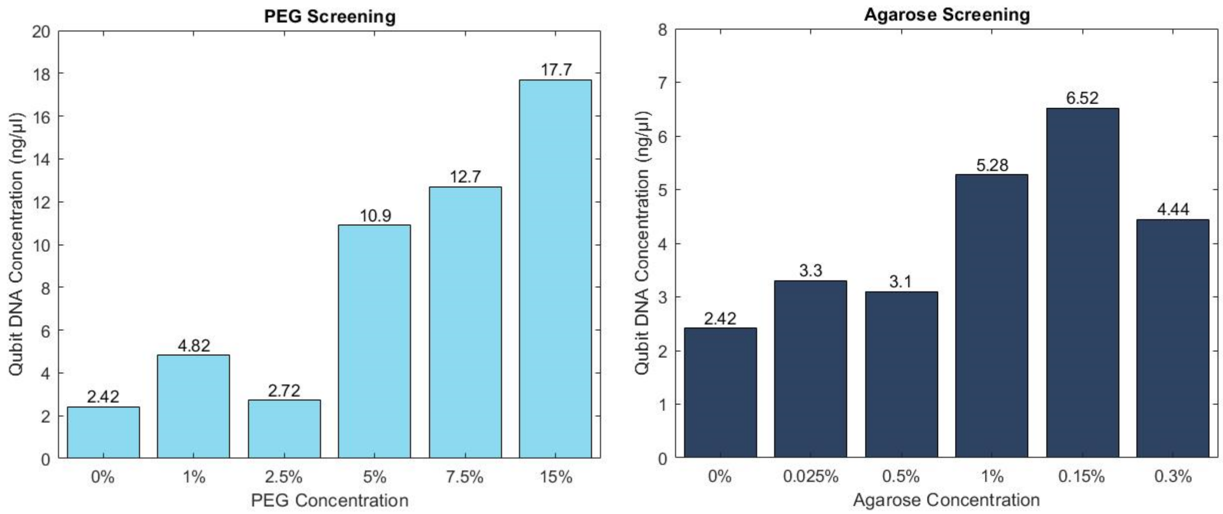


Figure 4. cDNA yield of samples containing concentrations from 0-15% of PEG (left) and concentrations from 0-0.3% of agarose (right).

Next, we tested the addition of agarose, a hydrogel known to have a molecular confinement effect. Molecular confinement, in contrast to molecular crowding, is where the hydrogel transforms the solution into more of a solid state, creating small pores for the reactions to occur in. This pushes the reacting molecules closer together, increasing reaction efficiency. In addition, agarose has been found to improve enzyme stability, which also helps increase reaction efficiency[25]. We tested agarose concentrations between 0 and 0.3% (Figure 4) and found that agarose increased cDNA yield at lower concentrations than PEG.

Finally, we tested different combinations of PEG and agarose in the hopes of achieving both molecular crowding and confinement effects simultaneously (Figure 5)**.** We concluded that adding PEG and agarose did increase cDNA yield. We ultimately selected the 5% PEG and 0.1% agarose condition to use in future experiments, as it was a good balance between increasing reaction efficiency while still being easy to pipette accurately. After performing a replicate experiment with a *t*-test (p=0.0253), we concluded that the difference in cDNA yield between samples with PEG and agarose versus samples with neither was statistically significant (Figure 5).

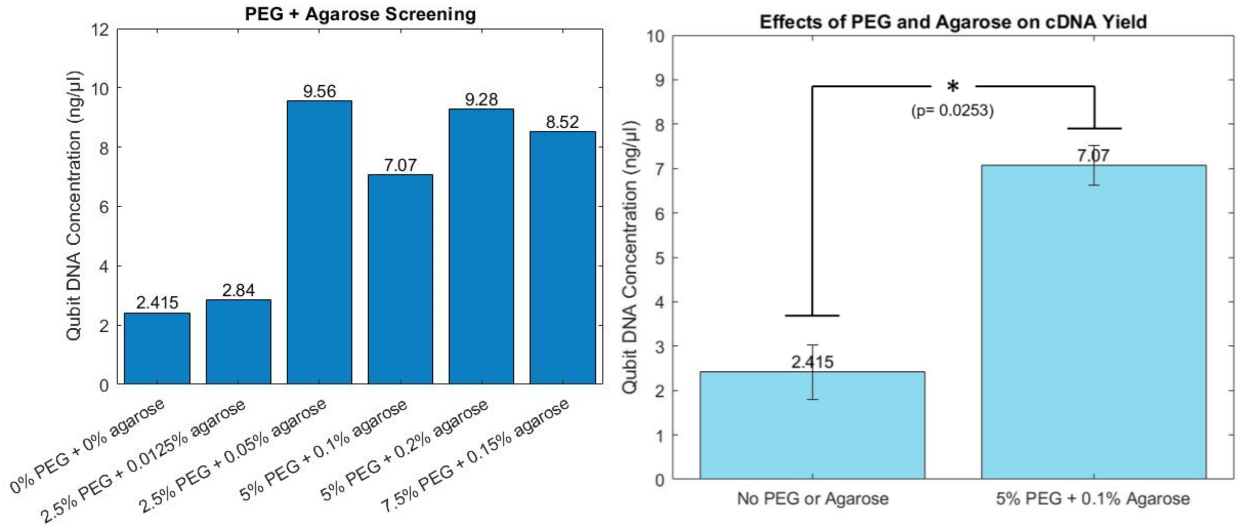


Figure 5. Screening of different combinations of PEG and agarose (left). Comparison of cDNA yield between samples with 5% PEG and 0.1% agarose and samples with neither PEG nor agarose (right). The difference in cDNA yield was statistically significant (p=0.0253).

2.1.3. qRT-PCR Validation

To confirm that our optimized method had minimal amplification bias after 18 PCR cycles, we performed Real-Time Quantitative Reverse Transcription PCR (qRT-PCR) experiments using both lysate samples and single-cell samples. In both experiments, we included a bulk lysate control that was produced by performing reverse transcription on purified RNA stock. We normalized the expression levels of our samples against this control sample for each experiment. Relative expression levels were calculated by finding the average C(t) value of each sample, taking the reciprocal, and normalizing against the control sample.

We measured the expression of six different genes: Tubb5, Trim28, Sdha, Tfrc, Tbp, and Eef1b2. In our experiments, we expected all of these genes to be housekeeping genes in LA-4 cells except for the Trim28 and Tfrc genes. The Tubb5 gene codes for tubulin, which is involved in mitosis and the cell cycle. Trim28 is a transcription mediation factor, while Eef1b2 is a translation mediation factor. Sdha is involved in cellular respiration, and Tbp codes for the TATA binding protein. Finally, the Tfrc gene codes for the transferrin receptor, which is involved in the uptake of iron into the cell. We ran two different qRT-PCR experiments with our Smart-seq samples: one with lysate samples and one with single-cell samples (Figure 6). The goal of the lysate experiment was to determine whether our optimized Smart-seq method was resulting in amplification bias, while the goal of the single-cell experiment was to see if our method could resolve the heterogeneity between individual cells.

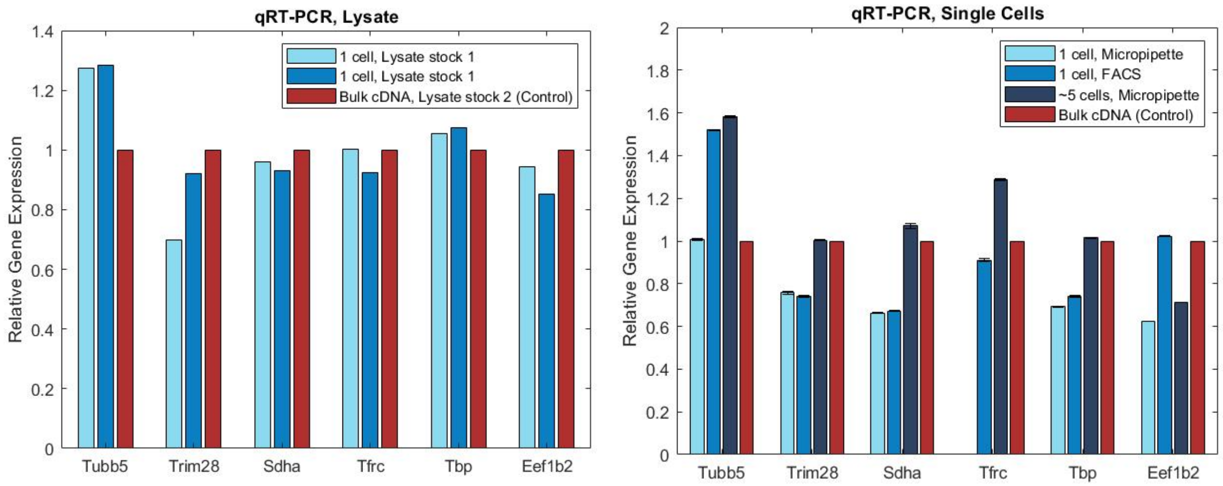


Figure 6. Gene expression of lysate samples (left) and single cells (right) prepared with our optimized Smart-seq method. We expected all six of these genes to be housekeeping genes except for Trim28 and Tfrc.

Lysate experiment included two 1 cell samples taken from lysate stock 1, making them technical replicates. The third sample was a bulk cDNA sample taken from a purified LA-4 RNA stock, making it a biological replicate with the other two samples. If minimal amplification bias was present, we would expect all housekeeping genes to have similar expression levels. We saw similar expression levels for all genes except Tubb5 and Trim28. The differences in Tubb5 expression levels are likely related to differences in cell cycle stage at the time of lysis, as tubulin involved in mitosis. Differences in Trim28 expression were expected, as it is not a housekeeping gene. We concluded that our optimized Smart-seq method had minimal amplification bias, as we saw less than 10% technical variation and less than 15% biological variation in our lysate samples.

Single-cell experiment included four samples: one cell selected with a micropipette, one cell sorted using FACS, ~5 cells selected with a micropipette, and the same bulk cDNA control from the lysate experiment. We found that our 5 cell sample expression levels more closely matched the bulk cDNA control, while our single cell samples closely matched each other but deviated from the bulk cDNA control. We concluded that our optimized Smart-seq method could resolve the heterogeneity between individual cells, as it was sensitive enough to reveal differences in gene expression found in different single-cell samples.

2.2. RamDA-seq

Though Smart-seq is advantageous due to its full transcriptome coverage, one of its main drawbacks is that it cannot be used to study bacterial RNA. Smart-seq uses an oligo (dT) to prime for reverse transcription, which binds to the poly (A) tail of each mRNA strand. However, most bacterial RNA strands do not have poly (A) tails[26]. To address this problem, the RamDA-seq protocol[27] uses not-so-random primers (NSRs), which are hexamers that bind at random points along the RNA strand.

Reverse transcriptase then binds to the NSRs and produces a cDNA strand, which is nicked by DNase I and then pulled off and protected by the T4g32 protein[27]. We optimized the RamDA-seq method as it would allow us to be able to sequence both mammalian and bacterial RNA at the same time, enabling the study of both host and pathogen expression simultaneously.

To ensure that we could perform RamDA-seq successfully, we performed preliminary experiments following the original RamDA-seq protocol that included only RNA denaturation and reverse transcription. However, our samples consistently had cDNA yields of less than 0.5 ng/µl when quantified with the Qubit fluorometer, which was too low to proceed with the rest of the RamDA-seq protocol. We then ran a series of experiments changing different aspects of the reverse transcription reaction in an attempt to produce a higher cDNA yield. We compared T4g32 proteins purchased from both Roche and New England Biolabs (NEB), since Hayashi et al found T4g32 protein purchased from NEB to be more stable[27]. We tested increasing the time at 37ºC from 30 minutes to either 60 minutes or 120 minutes, as it was shown to increase cDNA amplification[27]. We also tested 2x and 0.1x DNase I conditions, as well as the addition of 50% PEG to achieve a molecular crowding effect. However, all of these test conditions continued to produce low cDNA yields, even in samples with 100 cells' worth of lysate.

Our next experiments also included the second-strand synthesis step of the RamDA-seq protocol. Performing second-strand synthesis would allow us to more accurately quantify our cDNA yield for each sample using the Qubit dsDNA HS kit, since the cDNA would be double-stranded. We also switched from using LA-4 lysate to using purified LA-4 RNA, to eliminate the possibility of genomic DNA contamination.

After continuing to see low cDNA yields in our samples even after adding second-strand synthesis, we performed qRT-PCR to confirm that the cDNA produced with the RamDA-seq protocol was being amplified sufficiently, and that common housekeeping genes could be detected. We measured the expression of the Tubb5, Sdha, B2m, and Eef1b2 genes. The Tubb5 gene codes for tubulin, which is involved in mitosis and the cell cycle. Sdha is involved in cellular respiration, while B2m is involved in antigen presentation. Finally, Eef1b2 codes for a translation elongation factor. Though we did see expression of these four genes in some of our samples, we also observed some gene dropout (Figure 7).

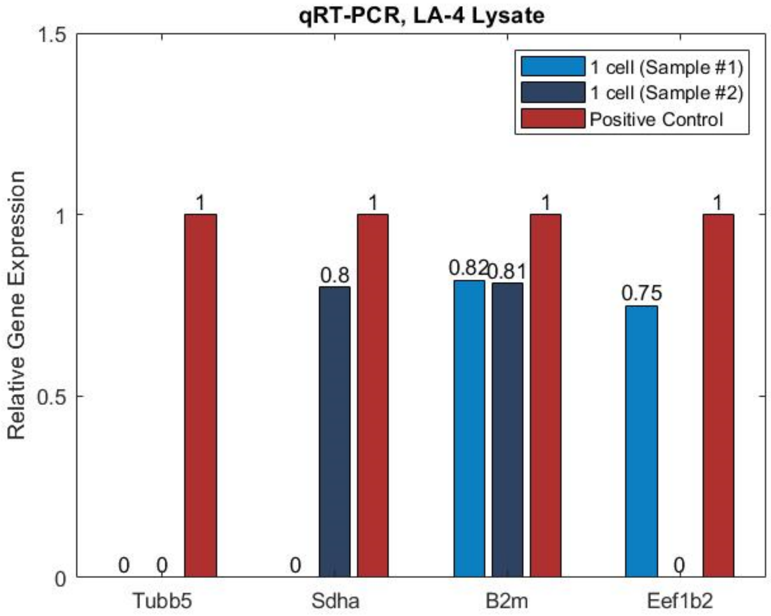


Figure 7. Relative gene expression of 1 cell LA-4 lysate samples prepared with RamDA-seq compared to positive control. Positive control sample contained ~50 cells worth of RNA that underwent the Smart-seq process up to but not including PCR.

To troubleshoot this problem, we ran an experiment to test our oligo (dTs) and NSRs to confirm that both worked individually and together. We ran samples with only oligo (dT), only NSRs, both oligo (dT) and NSRs (Figure 8) and compared the cDNA yield of each condition. We ran duplicate samples of both 1000 cell/µl LA-4 lysate and 20 ng/µl purified LA-4 RNA and concluded that both the oligo (dTs) and NSRs worked. However, we did see some inconsistencies in cDNA yield, as we did not see any cDNA yield in our lysate samples containing both oligo (dT) and NSR primers.

As previously mentioned, we again tested increasing amounts of time at 37ºC during reverse transcription. In this experiment, we tested 30, 60, 120, and 240 minutes at 37ºC (Figure 8) and found that while cDNA yield did appear to increase slightly with time, the difference was not statistically significant when compared with the 30 minute condition (t-test, $p>0.05$). We decided to use the 60 minute condition in future experiments in order to maximize amplification without significantly affecting the length of the experiment.

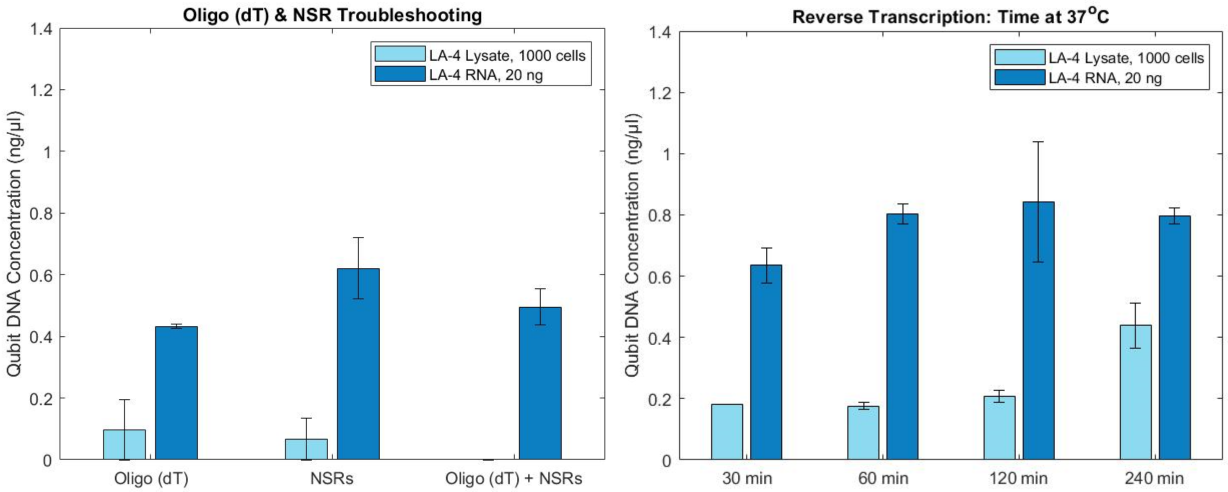


Figure 8. Testing both our oligo (dT) primers and our NSR primers to ensure that both can be used for reverse transcription (left). Testing increasing amounts of time at 37ºC during reverse transcription (right).

After procuring a fresh set of reagents, we tested different concentrations of DNase I and T4g32 protein. We found that decreasing the concentration of both DNase I and T4g32 resulted in amplification and significantly increased cDNA yield, while increasing DNase I and T4g32 concentrations decreased cDNA yield (Figure 9). We saw approximately an 18-fold increase in cDNA yield between the 1x DNase/T4g32 and 0.1x DNase/T4g32 conditions.

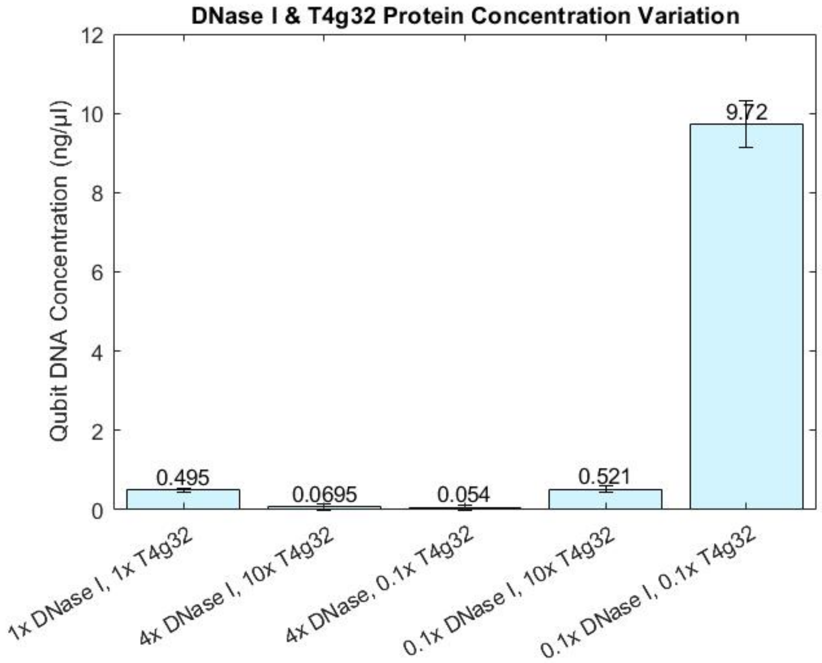


Figure 9. Varying concentrations of DNase I and T4g32 protein compared to original RamDA-seq protocol (1x DNase/1x T4g32 condition).

Having identified a condition that produced a substantial cDNA yield, we began testing other combinations of DNase and T4g32 concentrations to improve sensitivity. We found that conditions with 0.2x DNase had approximately a four-fold increase in cDNA yield compared with samples that had 0.05x DNase (Figure 10). We considered this condition to be a suitable candidate for further optimization.

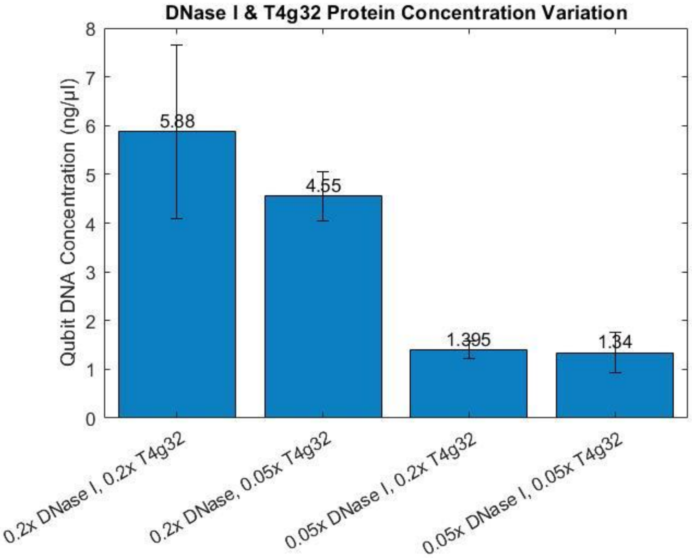


Figure 10. cDNA yields of samples prepared with different concentrations of DNase I and T4g32.

We then ran preliminary experiments applying our optimized RamDA-seq method to samples containing purified Burkholderia RNA to validate that our method will work with bacterial RNA, despite the not-so-random primers being optimized for use with mouse cells. We found that though there was a significant difference in cDNA yield between the LA-4 and Burkholderia samples, our optimized RamDA-seq method was able to produce a substantial cDNA yield when starting from bacterial RNA (Figure 11). We also prepared samples that contained 10 pg of purified LA-4 RNA and 10 pg purified Burkholderia thailandensis RNA to test if our method would work with a combination of mammalian and bacterial RNA. While one of our samples had a dsDNA concentration of 4.66 ng/µl, the other had a yield that was too low to detect with a Qubit fluorometer. We hope to continue optimizing the RamDA-seq method to improve consistency between samples.

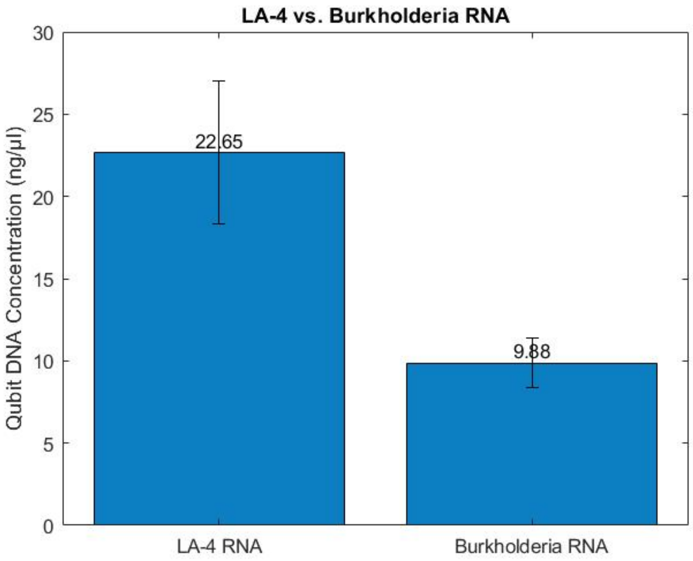


Figure 11. cDNA yields of LA-4 RNA samples and Burkholderia RNA samples, starting from 20 ng of purified RNA.

2.3. The OligoRL Framework

OligoRL formulates the oligo design problem as a Markov Decision Process (MDP) [28]. The MDP describes how an agent in a state s_i selects an action a_i that moves the agent to a new state s_{i+1}. The transition between states is accompanied by a reward r_i. The agent's goal is to select actions that maximize the sum of all the rewards. Our problem is to build an oligo of length L by selecting degenerate base codes at each position. The oligo codes are selected sequentially beginning at the 5' end. An agent in state s_i has selected the first $i - 1$ oligo codes, so the agent begins at state s_1, when zero oligo codes have been selected, and finishes at state s_{L+1}. The state defines not only how many but also which codes have been selected. An agent that has selected codes ACG is in a different state than an agent that has selected codes ACT.

Once in state s_i, the agent selects the code to place at position i. This selection corresponds to the action a_i, which is drawn from the set of possible codes $A(s_i)$. The set of allowed codes is state-dependent—the codes selected for the prior positions $1 \ldots i - 1$ can change the codes available to the agent at position i. Each available code $a_i \in A(s_i)$ has an associated reward $r_i(a_i)$. This reward depends on the entire oligo up to and including position i. The final reward $r_L(a_L)$ is based on the entire oligo.

We do not make any assumptions about the reward functions. For example, the reward for an oligo can be based on aligning the oligo to a genome and counting the number or quality of the hits. It is also possible to set all but the final rewards to zero, delaying the reward calculation until the entire oligo has been selected. Furthermore, the reward function can be applied to either a single oligo or an entire oligo pool. The flexibility of the reward function underlies the generality of our approach, but it also requires us to solve the oligo selection problem by simulation.

We use a rollout algorithm to choose the best code at each position. Rollout is a reinforcement learning (RL) technique used to solve large MDPs by simulating trajectories using a computer model

[29, 30]. In state s_i we begin by considering the first code $a_1 \in A(s_i)$. We simulate ahead to the end of the oligo, choosing codes randomly and summing the rewards. By averaging the rewards from many random trajectories, all beginning with action a_1, we can estimate the average reward the agent will experience when code a_1 is selected. We compute this reward-to-go estimate for all other actions available at state s_i. The action we ultimately choose in state s_i corresponds to the maximum reward from the rollout simulations. After the code is selected, we move to the next position (state s_{i+1}) and repeat the rollout process starting at the new state.

We used OligoRL to design pools of Not-So-Random (NSR) primers. Using OligoRL, we found smaller NSR pools with increased uniformity across all mRNAs in a representative organism. This final example demonstrates "black-box" reward functions that map the NSR primers to transcriptomes and calculate the uniformity of an NSR pool. Neither of these reward functions can be expressed as algebraic constraints on the OligoRL problem.

Our goal is to find an optimal pool of Not-So-Random (NSR) primers that 1.) avoid rRNA, tRNA, or transcripts from any unwanted genes, 2.) bind to every gene in a target set at least once, 3.) uniformly cover the transcripts from targeted genes, and 4.) use the smallest number of oligos necessary to meet objectives 1–3.

Current workflows for designing NSR hexamer primers start with a pool of all 4,906 possible hexamers and remove hexamers that appear in the undesired transcripts. The remaining hexamers are aligned to the rest of the transcriptome. We developed a multifaceted reward function that scores NSR primer pools using five criteria:

1. *Specificity*. Each NSR primer is compared to hexamers in rRNA and tRNA genes, sequencing adapters, and the other NSR primers. Any NSR candidate that contains these sequences receives a reward of zero.
2. *Gene count*. The agent receives a reward for any gene hit at least once by an oligo in the pool.
3. *Total hits*. The agent is rewarded for maximizing the total number of hits across the transcriptome.
4. *Intergene uniformity*. The agent is rewarded for placing the same number of hits on each gene.
5. *Intragene uniformity*. The agent is rewarded for uniformly distributing hits across the length of each gene.

Inter- and intragene uniformity are quantified by the distribution uniformity, lower quartile (DULQ) score (Figure 12A) [31]:

$$\mathrm{DULQ} = \frac{\mathrm{mean(lower\ quartile)}}{\mathrm{mean(sample)}}.$$

The DULQ is bounded between zero (all hits at a single location) and one (perfect uniformity). Only genes with at least one hit are used to calculate the DULQ. The total reward is the weighted sum of the individual criteria:

$$\begin{aligned}\text{reward} \quad = \; & \beta_{\text{gene}}\,\{\text{gene count}\} \\ & +\beta_{\text{hits}}\,\{\text{total hits}\} \\ & +\beta_{\text{inter}}\,\{\text{intergene uniformity}\} \\ & +\beta_{\text{intra}}\,\{\text{intragene uniformity}\}.\end{aligned}$$

Users can change the weights to emphasize certain criteria when designing primer pools.

We tested the performance of our RL-guided NSR primer design program, called NSR-RL, by designing primer pools using varying weights in the reward function. We compared the NSR-RL pools to primer pools designed using a standard brute-force approach. Both pools targeted the 1.76 Mb *Streptococcus mutans* transcriptome. Changing the reward function weights prioritizes different design criteria. For example, if we are only interested in designing a pool that hits every gene at least once, we can do so by zeroing out the other terms in the reward function. NSR-RL can design a pool that hits every gene using only 10 oligos. A brute-force approach requires 453 oligos to hit every gene.

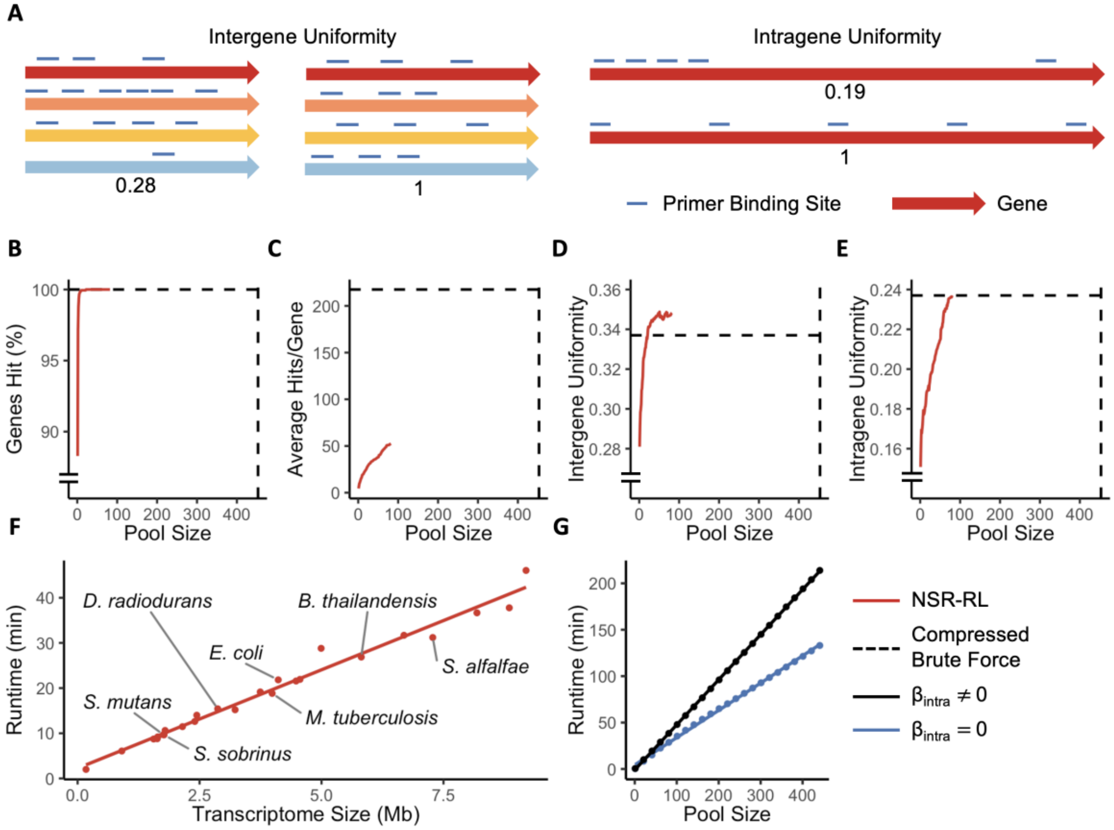


Figure 12: NSR-RL creates hexamer pools using a multivariate reward function. **A.** Intergene uniformity measures the distribution of the hits per gene. Intragene uniformity measures the distribution of hits across the length of each gene. Both uniformity scores range from $[0,1]$. NSR hexamer libraries produced by NSR-RL were compared to a pool of 453 hexamers produced by a standard brute-force approach. The libraries were compared across four criteria: the number of unique genes hit at least once (**B**), the total number of hits (**C**), intergene uniformity (**D**), and intragene uniformity (**E**). The dashed black lines show the performance of the brute-force pool, and the solid red lines show the performance of the NSR-RL pool as each hexamer is added to the pool.

NSR-RL hit every target with increased intergene uniformity and equivalent intragene uniformity with only 100 oligos. **F.** NSR-RL's runtime was measured for pools designed to target bacteria with transcriptomes between 0.17 Mb and 9.2 Mb in size. **G.** Quantifying intragene uniformity requires calculating the gaps between all hits on each transcript. Consequently, the runtime of NSR-RL decreases when intragene uniformity is removed from the reward function by setting the associated weight $\beta_{\text{intra}} = 0$.

Rather than minimize the number of oligos, we can use NSR-RL to design a fixed-size pool with improved coverage or uniformity. We used NSR-RL to design a pool containing 100 oligos with nonzero weights for all four criteria in the reward function. The resulting pool exceeded the performance of the compressed brute-force pool (Figure 12B-E). The NSR-RL pool hit every gene in the *S. mutans* transcriptome after only 22 oligos. The NSR-RL pool also placed an average of 993 hits per oligo while the brute-force pool placed an average of 910 hits per oligo. Note that it is impossible to generate more total hits than the brute-force designed pool since the brute-force pool includes all hexamers that are not found in the rRNA or other "unallowable" genes. While the NSR-RL pools contain fewer total hits, the hits are distributed more evenly across the transcriptome as measured by intergene uniformity. Interestingly, we observed that the intergene uniformity score quickly approached a maximum but then oscillated near this value as new oligos were added to the pool. The oscillations indicate that NSR-RL added new oligos that improved the scores of other terms at the expense of intergene uniformity, and vice-versa. The NSR-RL pool's intragene uniformity matched the performance of the brute-force pool. Users can tune the reward function's weights to produce NSR primer pools that prioritize either the number of genes hit, total hits, or uniformity. In addition, users can easily add terms to the reward function or create a custom reward to design specialized pools.

NSR-RL's runtime increases linearly with the size of the problem. We generated NSR pools containing 30 hexamers for an assortment of bacterial transcriptomes ranging between 0.17–9.2 Mb in size. We observed that the algorithm's runtime scaled linearly with each species' transcriptome size (Figure 12F). The NSR-RL runtime also increases linearly with the number of oligos in the final pool. The amount of computation required depends heavily on the structure of the reward function. In particular, calculating the intragene uniformity score requires measuring the hit positions of every simulated oligo and calculating the gap distances between each hit position along the length of every gene. Pools designed with reward functions that include intragene uniformity took approximately 50% longer to generate (Figure 12G). We implemented a bypass to skip these calculations if the user is not interested in intragene uniformity, i.e. when the user sets $\beta_{\text{intra}} = 0$.

2.4. Splintlock-seq

Proof of concept experiments using a single synthesized padlock probe and a single RNA target were conducted to determine the feasibility of the approach and to begin biochemical optimization. Protocols for SplintR ligation and rolling circle amplification were developed which resulted in template-specific signal generation.

An automated, computational pipeline for the design of padlock probes was built using the R programming language (Figure 13). The padlock probe architecture consists of a common 60 nt backbone with 15 nt sequences at the 5' and 3' ends that hybridize to a 30 nt stretch on the target mRNA. Briefly, a Genbank file is read into the program from the NCBI and sequences for rRNA, tRNA, and mRNA are extracted. The reverse compliment sequence of each mRNA is hashed into 30-mers to generate all possible candidate padlock probe target sequences. If the respective 30-mer is unique across the annotated transcriptome, it is next filtered against a database containing all 15-mer sequences found in the reverse compliment of rRNA and tRNA transcripts using a fuzzy matching algorithm. If both of the 15-mer halves of a given 30-mer cannot be matched to any rRNA/tRNA 15-mer given a specified mismatch tolerance, the candidate sequence passes the filter. An analogous fuzzy matching filter is then utilized to ensure there is no interaction between padlock probe ends and the common padlock probe backbone. The candidate 30-mer is then filtered with a set of user-specified criteria to ensure it is compatible with the Splintlock-seq experimental pipeline. These filters select for 30-mer sequences with no predicted secondary structures, a lack of polynucleotide tracts, and favorable annealing characteristics. Next, candidate sequences are binned according to their binding position on the target transcript and a greedy multi-objective optimization algorithm is utilized to select for the "best" probe in each bin. In this way, an initial pool of 12,007 probes targeting 99% (1882/1900) of the annotated ORFs in the genome of *S. mutans*, with a median coverage of 6 probes per ORF, was designed.

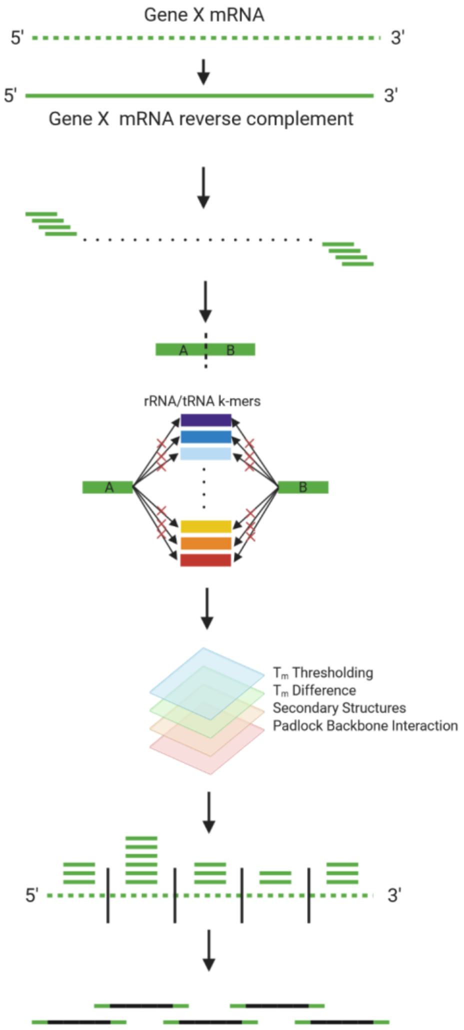


Figure 13: Computational pipeline for identifying Splintlock-seq padlock probes across an entire genome.

The initial pool of candidate probes was synthesized using CustomArray electrochemical DNA microarray technology. Due to the low per oligo yields typically associated with massively parallel synthesis platforms, an experimental pipeline outlined by Murgha et al. [31] was utilized to amplify the pool in a strand specific manner and generate an ssDNA oligo pool with sufficient quantity for Splintlock-seq experiments. SplintR ligation and rolling circle amplification reactions were re-optimized using the new padlock pool on total RNA. Conditions were identified in which RNA-specific signal could be reliably generated at levels over 1000-fold higher than no-RNA conditions. In order to assess the specificity of the padlock probe pool for mRNA, Splintlock-seq library preps were performed with total RNA, rRNA-depleted RNA, a synthetic mix of rRNA generated by IVT, or no RNA serving as template. Libraries were prepped, pooled, and sent off for next generation sequencing on an Illumina NovaSeq 6000. Over 496 million reads were obtained, demultiplexed, and aligned to the *S. mutans* genome in a strand specific manner. Read indices were then mapped to probe sites and analyzed for relative read abundance by template group (Figure 14). Over 12,000 probes were designed and synthesized to specifically target mRNA and avoid ligation on rRNA or tRNA. However, a group of 900 probes demonstrated highly efficient ligation on the synthetic rRNA mix or without any template. These probes were removed before further characterization of the Splintlock-seq library prep using a second round of sequencing. The second set of probes showed a reduced number of high-abundance probes that bound to rRNA transcripts.

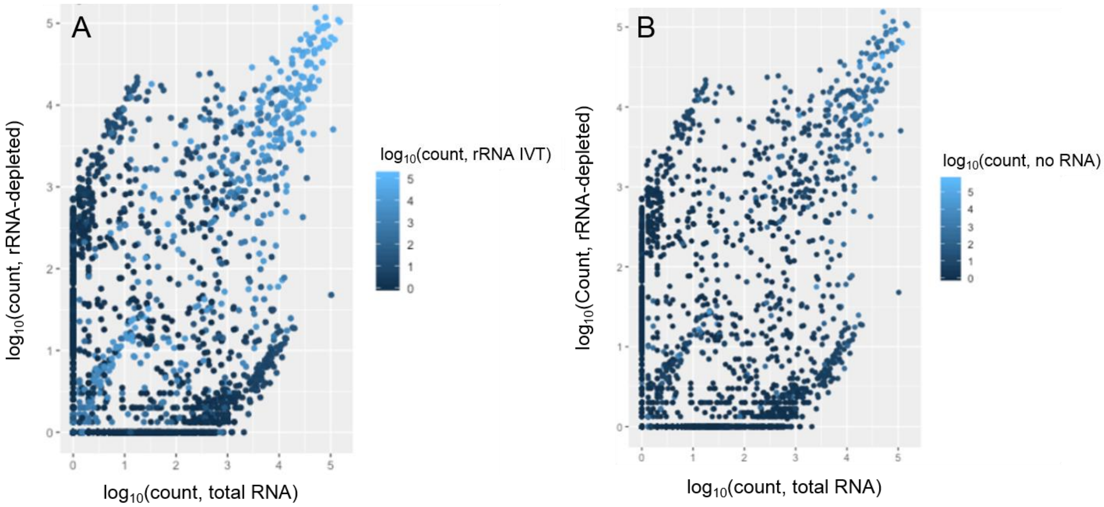


Figure 14: Probe counts averaged by gene show significant off-target effects when rRNA is present. These probes were removed from future Splintlock-seq pools.

3. DISCUSSION

This project demonstrated the feasibility of applying single-cell RNA-seq as a tool for amplifying and analyzing both mammalian and bacterial trancriptomes. The Smart-seq method was optimized for use with LA-4 cells and established benchmarks for cDNA quantification and amplification. We observed a technical variation between single cell samples of only 9.18% and a biological variation of 12.11% across a panel of selected housekeeping and reference genes. We also made significant progress optimizing the RamDA-seq method for use with purified LA-4 and Burkholderia RNA. We selected the 0.1x DNAse/0.1x T4g32 and 0.2x DNAse/0.2x T4g32 condition as candidates for further optimization and will continue optimizing these conditions towards increasing sensitivity and decreasing amplification bias. While both of these conditions produced promising results, the success rate of the process was between 60-70%, indicating that further optimization would be desirable. As a future direcion, we hope to improve the consistency of the RamDA-seq method and apply it to single LA-4 and Burkholderia cells to study each cell's gene expression and better understand the interactions between them.

OligoRL uses true "black-box" reward functions. The quality of a candidate oligo pool can be measured using simple algebraic expressions (like degeneracy of the pool) or complex calculations performed by external software packages (such as genome-wide sequence aligners). NSR-RL has a complex, multifactorial reward function, and calculating rewards makes up the majority of the algorithm's runtime. Researchers with computationally intensive reward functions may consider approximating the reward with a simpler function. Performing more rollout simulations with a less accurate reward may yield better solutions than fewer simulations with better reward estimates.

OligoRL works best when finding optimal solutions from a large set of valid solutions. When the pool of valid solutions shrinks, the nature of the design problem shifts from finding optimal solutions to finding valid solutions that satisfy the problem's constraints. Rollout, and therefore OligoRL, performs better at optimization than constraint satisfaction. When valid solutions are difficult to find, OligoRL explores many dead-end solutions with poor rewards. For example, instructing NSR-RL maximize total hits leads to states where there are only a few valid hexamers left. In this scenario, OligoRL randomly samples many hexamers but often fails to find the few valid ones. The invalid simulations do not provide useful information to the agent since all invalid actions appear equally poor. Conversely, when nearly all solutions are valid, OligoRL quickly determines good actions for each state since every simulation provides information about an action.

NSR-RL finds sets of oligos with differing degeneracy. Some wet-lab protocols suggest oligo pools with equimolar concentrations, so experimenters should be careful to mix the oligos in proportion to their degeneracy. The added mixing complexity is a trade-off for the savings gained when using these tools.

Our iterative approach to Splintlock-seq development identified and removed several probes with affinity toward rRNA transcripts. Our results highlight the difficulty of a purely bioinformatic approach to identifying "good" probes that target only mRNA. Future work should combine both a bioinformatic pipeline and experimental profiling to identify the optimal probe set.

Similarly, the reaction conditions for a Splintlock-seq experiment could be improved. Several factors affect the library prep, and a multifactorial design is required to assess the quality of the final libraries. Our experience shows that library quality is nonlinear with respect to input factors, so simple first-order linear models may not be adequate.

The Splintlock-seq computational pipeline was tested using the transcriptome of *S. mutans*. We also used the pipeline to generate probe sets for *S. sobrinus* and *Burkholderia thailandensis*. For both other species, our pipeline was able to final probe sets that hit nearly every gene while missing rRNA sequences. We do not anticipate any difficulty applying these methods to other bacteria of interest.

4. METHODS

Cell Culture

LA-4 cells, which are a murine lung epithelial cell line, were cultured in F-12K medium with 15% fetal bovine serum and 1% penicillin-streptomycin at 37°C in 5% CO_2. Single cells were either picked up with a micropipette under a microscope or sorted using fluorescence-activated cell sorting (FACS). Premade lysate was prepared by centrifuging a cell suspension of known concentration at 150 g for 5 minutes and resuspending the pellet in lysis buffer. Lysis buffer was composed of 0.3% NP-40 surfactant in DI water.

Smart-seq

Premade lysate was diluted from 1000 cells/µl down to 1 cell/µl for our optimization experiments. Either premade lysate or a single cell was added to each tube of a set of strip tubes. 1 µl oligo dT primer and 1 µl 40 mM dNTPs were added to each tube, and the samples were denatured at 72°C for 3 minutes and 42°C for 2 minutes. 6 µl of reverse transcription mix, containing 2 µl 5x Maxima Buffer, 0.25 µl RNase OUT, 0.75 µl template-switching oligo (TSO), 0.5 µl Maxima -H Reverse Transcriptase, and 2.5 µl nuclease-free water, was added to each sample. The reverse transcription reaction started at 42°C for 90 min, then continued with 10 cycles of (50°C for 2 min, 42°C for 2 min) and ended with a final inactivation step of 70^{c}C for 15 min. 10 µl of PCR mix, containing 4 µl 5x GC Buffer, 0.4 µl dNTPs, 2 µl 1 µM IS PCR Oligo, 0.2 µl Phusion polymerase, and 3.4 µl RNase-free water, was added to each sample[32]. Each sample was cleaned using Ampure XP beads following the Smart-seq2 protocol[32], using a bead to sample ratio of 0.8:1. cDNA yield was quantified using either a 2100 Agilent Bioanalyzer with the High Sensitivity DNA kit or a Qubit 2.0 Fluorometer with the High Sensitivity dsDNA kit.

RamDA-seq

RamDA-seq biochemistry processes and whole transcriptome amplification were adapted from the protocol established by Hayashi et al [9]. 1 µl of either purified RNA or premade lysate was added to each tube of a set of strip tubes and denatured at 70°C for 90 seconds. 2 µl of reverse transcription mix, containing 0.6 µl 5x PrimeScript Buffer, 0.2 µl DNase I, 0.06 µl 10 µM Oligo (dT), 0.8 µl 10 µM 1st-NSR primers, 0.2 µl 1 mg/mL NEB (or Roche) T4g32 protein, and 0.15 µl PrimeScript enzyme mix, was added to each sample. The reverse transcription reaction started at 25°C for 10 min, continued at 30°C for 10 min, 37°C for 30 min, and 50°C for 5 min, then ended with a final inactivation step of 94^{c}C for 5 min. 2 µl of second-strand synthesis mix, containing 0.5 µl 10x NEBuffer2, 0.5 µl 10 mM dNTPs (2.5 mM each), 0.4 µl 100 µM 2nd-NSR primers, 0.45 µl RNase-free water, and 0.15 µl Klenow Fragment, was added to each sample. The second-strand synthesis reaction started with 16°C for 60 min and ended with 70°C for 10 min. Each sample was cleaned using Ampure XP beads following the Smart-seq2 protocol[32] using a bead to sample ratio of 1:1. Each sample's cDNA yield was quantified using either a NanoDrop 1000 instrument or a Qubit fluorometer with the High Sensitivity dsDNA kit.

qRT-PCR

10 μl SYBR Green PCR Master Mix (Applied Biosystems) was added to each well of a 96-well plate. 0.4 μl of 10 μM forward primer and 0.4 μl of 10 μM reverse primer (Appendix A) were then added to each well, along with 8.2 μl of RNase-free water. Finally, 1 μl of sample was added to each well. The plate was then incubated at 95°C for 10 min, followed by 40 cycles of 95°C for 15 seconds, 60°C for 1 min, and a plate read, and ending with a 5 second melt curve step starting at 65°C and increasing to 95°C in increments of 0.5°C.

Computational Methods

OligoRL and all simulation codes are available as a Julia package at http://jensenlab.net/tools. Simulations were run using Julia version 1.2.8 on a 16-core 3.2 GHz AMD Threadripper processor with 48 Gb of RAM.

The rollout algorithm used in OligoRL can be parallelized at either the action or simulation level. For example, when simulating the reward for a single base, each simulation can be executed in parallel by a separate thread. This study used Julia's multithreading tools to perform parallel computations on a multicore processor. The code can also be configured for a cluster computing environment where parallel simulations execute on separate machines.

The Splintlock-seq probe design pipeline was written in the R programming language using the Bioconductor library. All sequencing data analysis was performed using custom R scripts using bowtie[33] for sequence alignment.

NSR-RL Algorithm

NSR-RL designs Not-So-Random primer pools for RNA-seq library preparation and other multiplex genomic assays. Users supply two sequence files containing 1.) "targeted" transcripts that should be targeted by the NSR primers, and 2.) "unallowed" transcripts to avoid, e.g. transcripts from rRNA and tRNA genes. The user also specifies the number of NSR primers to create and the length of the primers (the default is hexamers).

NSR-RL builds oligos using rollout with dynamic action spaces. Candidate oligos are assigned a reward of zero if they hit any unallowed transcript. Palindromic candidates are also assigned a reward of zero since palindromic reverse transcription primers may self-anneal during amplification. Non-palindromic candidate oligos that miss the unallowed transcripts are scored by the multifaceted reward function. The first three terms in the reward are calculated by counting the number of times the oligo hits each targeted sequence. First, the gene count term is the number of genes that are hit at least once. Second, the total hits term is sum of all hits across the transcriptome. Third, the intergene uniformity score is calculated using the DULQ score of all of the hit counts. Calculating the fourth term in the reward, intragene uniformity, requires the gaps between hits to calculate the DULQ for each transcript. Transcripts with a more uniform gap distance distribution will score higher than transcripts with different sized gaps. The overall intragene uniformity score is the average DULQ across all transcripts. We multiply the inter- and intragene uniformity scores by the number of targets, n_{targets}, to place these rewards on a similar scale as the other terms. The target count and uniformity terms range from 0 to n_{targets}, while the total hits is term is unbounded. Each term in the reward function has an associated weight β, and the weights can be changed to tune the pools empirically.

After NSR-RL finishes an oligo, the oligo and its reverse complement are added to the list of unallowed sequences to prevent avoid repeats or selecting oligos that could form dimers when the libraries are amplified.

NSR-RL was benchmarked by creating 100 degenerate hexamers targeting the transcriptome of the 1076 Mb transcriptome of *Streptococcus mutans* strain UA159 (Figure 12B–E). Unless otherwise specified, the reward weights were $\beta_{\text{gene}} = 1$, $\beta_{\text{hits}} = 10^{-4}$, $\beta_{\text{inter}} = 1$, and $\beta_{\text{intra}} = 1$. To compare NSR-RL runtime with transcriptome size (Figure 12F), 30 degenerate hexamers were designed to target the transcriptomes of 25 species of bacteria.

Splintlock-seq

A pool of padlock probes generated by our automated design pipeline was synthesized and hybridized to isolated total RNA from the condition of interest. If a target mRNA is present, the ends of a respective padlock probe was ligated together by SplintR ligase, generating a circularized ssDNA molecule. Signal amplification and enrichment was then simultaneously carried out using Phi29-mediated rolling circle amplification initiated from a primer annealed to the padlock backbone. In this way, only intramolecular ligation products undergo amplification. The rolling circle amplification product is then used as template for a low-cycle number PCR with primers that add custom, dual-indexed Illumina adapters. These adapters enable the pooling of dozens of individual libraries to further reduce per library sequencing costs. Libraries were purified using AMPure beads and pooled in an equimolar fashion. The pooled libraries were sequenced on a NovaSeq 6000 instrument at the Biotechnology Core Facility at the University of Illinois.

REFERENCES

1. Ursell, L. K.; Metcalf, J. L.; Parfrey, L. W.; Knight, R., Defining the human microbiome. *Nutrition reviews* **2012,** *70 Suppl 1* (Suppl 1), S38-S44.

2. Knight, R.; Callewaert, C.; Marotz, C.; Hyde, E. R.; Debelius, J. W.; McDonald, D.; Sogin, M. L., The Microbiome and Human Biology. *Annual review of genomics and human genetics* **2017,** *18*, 65-86.

3. Gilbert, J. A.; Blaser, M. J.; Caporaso, J. G.; Jansson, J. K.; Lynch, S. V.; Knight, R., Current understanding of the human microbiome. *Nature medicine* **2018,** *24* (4), 392-400.

4. Whitehouse.gov https://www.whitehouse.gov/the-press-office/2016/05/12/fact-sheet-announcing-national-microbiome-initiative.

5. Westermann, A. J.; Gorski, S. A.; Vogel, J., Dual RNA-seq of pathogen and host. *Nature reviews. Microbiology* **2012,** *10* (9), 618-30.

6. Klein, A. M.; Macosko, E., InDrops and Drop-seq technologies for single-cell sequencing. *Lab on a chip* **2017,** *17* (15), 2540-2541.

7. Hwang, B.; Lee, J. H.; Bang, D., Single-cell RNA sequencing technologies and bioinformatics pipelines. *Experimental & molecular medicine* **2018,** *50* (8), 96.

8. 10xGenomics https://kb.10xgenomics.com/hc/en-us/articles/218170643-Can-gene-expression-in-microbial-samples-be-profiled-.

9. Hayashi, T.; Ozaki, H.; Sasagawa, Y.; Umeda, M.; Danno, H.; Nikaido, I., Single-cell full-length total RNA sequencing uncovers dynamics of recursive splicing and enhancer RNAs. *Nature communications* **2018,** *9* (1), 619.

10. Harouaka, R. A.; Zhou, M. D.; Yeh, Y. T.; Khan, W. J.; Das, A.; Liu, X.; Christ, C. C.; Dicker, D. T.; Baney, T. S.; Kaifi, J. T.; Belani, C. P.; Truica, C. I.; El-Deiry, W. S.; Allerton, J. P.; Zheng, S. Y., Flexible micro spring array device for high-throughput enrichment of viable circulating tumor cells. *Clinical chemistry* **2014,** *60* (2), 323-33.

11. Sooknanan, R.; Hitchen, J.; Radek, A.; Pease, J. J. J. o. B. T. J., Superior rRNA removal for RNA-Seq library preparation. **2012,** *23* (Suppl), S57.

12. Stewart, F. J.; Ottesen, E. A.; DeLong, E. F. J. T. I. j., Development and quantitative analyses of a universal rRNA-subtraction protocol for microbial metatranscriptomics. **2010,** *4* (7), 896-907.

13. Culviner, P. H.; Guegler, C. K.; Laub, M. T. J. M., A simple, cost-effective, and robust method for rRNA depletion in RNA-sequencing studies. **2020,** *11* (2), e00010-20.

14. Armour, C. D.; Castle, J. C.; Chen, R.; Babak, T.; Loerch, P.; Jackson, S.; Shah, J. K.; Dey, J.; Rohl, C. A.; Johnson, J. M.; Raymond, C. K., Digital transcriptome profiling using selective hexamer priming for cDNA synthesis. *Nature methods* **2009,** *6* (9), 647-9.

15. Li, H.; Qiu, J.; Fu, X. D., RASL-seq for massively parallel and quantitative analysis of gene expression. *Current protocols in molecular biology* **2012,** *Chapter 4*, Unit 4.13.1-9.

16. Sriskanda, V.; Shuman, S. J. N. a. r., Specificity and fidelity of strand joining by Chlorella virus DNA ligase. **1998,** *26* (15), 3536-3541.

17. Krzywkowski, T.; Nilsson, M. J. N. a. r., Fidelity of RNA templated end-joining by chlorella virus DNA ligase and a novel iLock assay with improved direct RNA detection accuracy. **2017,** *45* (18), e161-e161.

18. Jin, J.; Vaud, S.; Zhelkovsky, A. M.; Posfai, J.; McReynolds, L. A. J. N. a. r., Sensitive and specific miRNA detection method using SplintR Ligase. **2016,** *44* (13), e116-e116.

19. Roy, C. K.; Olson, S.; Graveley, B. R.; Zamore, P. D.; Moore, M. J. J. E., Assessing long-distance RNA sequence connectivity via RNA-templated DNA–DNA ligation. **2015,** *4*, e03700.

20. Blanco, L.; Bernad, A.; Lázaro, J. M.; Martín, G.; Garmendia, C.; Salas, M. J. J. o. B. C., Highly efficient DNA synthesis by the phage ɸ 29 DNA polymerase: Symmetrical mode of DNA replication. **1989,** *264* (15), 8935-8940.
21. Soengas, M. a. S.; Gutiérrez, C.; Salas, M. J. J. o. m. b., Helix-destabilizing activity of φ29 single-stranded DNA binding protein: effect on the elongation rate during strand displacement DNA replication. **1995,** *253* (4), 517-529.
22. Picelli, S.; Björklund, Å. K.; Faridani, O. R.; Sagasser, S.; Winberg, G.; Sandberg, R., Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nature Methods* **2013,** *10* (11), 1096-1098.
23. Svec, D.; Andersson, D.; Pekny, M.; Sjöback, R.; Kubista, M.; Ståhlberg, A., Direct cell lysis for single-cell gene expression profiling. *Front Oncol* **2013,** *3*, 274-274.
24. Bagnoli, J. W.; Ziegenhain, C.; Janjic, A.; Wange, L. E.; Vieth, B.; Parekh, S.; Geuder, J.; Hellmann, I.; Enard, W., mcSCRB-seq: sensitive and powerful single-cell RNA sequencing. *bioRxiv* **2017**, 188367.
25. Ross, M. L.; Kunkel, J.; Long, S.; Asuri, P., Combined Effects of Confinement and Macromolecular Crowding on Protein Stability. *International Journal of Molecular Sciences* **2020,** *21* (22).
26. Sarkar, N., POLYADENYLATION OF mRNA IN PROKARYOTES. *Annual Review of Biochemistry* **1997,** *66* (1), 173-197.
27. Hayashi, T.; Ozaki, H.; Sasagawa, Y.; Umeda, M.; Danno, H.; Nikaido, I., Single-cell full-length total RNA sequencing uncovers dynamics of recursive splicing and enhancer RNAs. *Nature Communications* **2018,** *9* (1), 619.
28. Bellman, R. J. J. o. m.; mechanics, A Markovian decision process. **1957,** *6* (5), 679-684.
29. Bertsekas, D., *Reinforcement learning and optimal control*. Athena Scientific: 2019.
30. Bertsekas, D., *Rollout, policy iteration, and distributed reinforcement learning*. Athena Scientific: 2021.
31. Burt, C.; Styles, S. W., *Drip and micro irrigation for trees, vines, and row crops: Design and management (with special sections on SDI)*. Irrigation Training and Research Center, Bioresource and Agricultural …: 1999.
32. Picelli, S.; Faridani, O. R.; Björklund, Å. K.; Winberg, G.; Sagasser, S.; Sandberg, R., Full-length RNA-seq from single cells using Smart-seq2. *Nature Protocols* **2014,** *9* (1), 171-181.
33. Langmead, B.; Trapnell, C.; Pop, M.; Salzberg, S. L. J. G. b., Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. **2009,** *10* (3), 1-10.

APPENDIX A. QRT-PCR PRIMER SEQUENCES

Gene	Sequence
Tubb5	CAGTCTGAGACCGGCCCAG
Trim 28	ACCAGCTCAGGCTTGGAGGT
Sdha	GCTTACCTGCGTTTCCCCTC
B2m	CAGTCGTCAGCATGGCTCG
Eef1b2	CCTTCGCCATGGGATTCG
Tbp	CCCCCTCTGCACTGAAATCA
Tfrc	TGCTAATGAGACCCACAGATACTGG

Reverse Primers	
Gene	Sequence
Tubb5	TGTGCACGATTTCCCTCATG
Trim 28	ACACGGCAGATAGTGGCACTG
Sdha	CTGGCGCAACTCAATCCCT
B2m	AGCATACAGGCCGGTCAGTG
Eef1b2	CGCCAGGTAATCGTTGAGCA
Tbp	GTAGCAGCACAGAGCAAGCAA
Tfrc	AGCTCATATTATTTGGATTGTGGCA

DISTRIBUTION

Email—Internal

Name	Org.	Sandia Email Address
Victoria VanderNoot	08621	vavande@sandia.gov
Technical Library	01977	sanddocs@sandia.gov

This page left blank

This page left blank

Sandia
National
Laboratories

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc. for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.