Application of Synthetic External RNA Controls in a Targeted Hybrid Capture Assay

Application of Synthetic External RNA Controls in a Targeted Hybrid Capture Assay

Erika Roth

A Thesis in the Field of Biotechnology

for the Degree of Master of Liberal Arts in Extension Studies

Harvard University

March 2021

Abstract

It is common practice for laboratories that run any type of diagnostic assay or molecular test at even moderate frequency to employ a well characterized in-process control sample to verify and validate assay specifications and performance. ERCC spike in mixes are a widely used set of external synthetic RNA transcripts designed to be added to in-process control RNA samples. These specific spike in mixes are most often utilized in gene expression profiling assays such as total RNA seq and microarrays to verify sensitivity, limit of detection, and reproducibility. This study investigates the feasibility of using ERCC spike in mixes to perform similar measurements in a targeted hybrid capture assay instead of whole transcriptome sequencing or microarrays. Manufacturer instructions do not provide guidance for this less common application, empirical knowledge and several rounds of testing were employed to determine optimal conditions to enable translation to a hybrid capture assay. Sensitivity, reproducibility, and sequencing read distribution were analyzed across a range of spike in mix concentrations and captured with targeted panels of varying size. Subsequent results indicate that when capturing ERCC spike in mixes with a large sized panel, the ERCC spike in mix should remain more concentrated to preserve ability to detect ERCC transcripts present at low abundance in the mix. When capturing with a smaller targeted panel, the ERCC spike in mix should be diluted to a lower concentration to avoid consuming an exorbitant amount of sequencing reads and therefore taking away valuable coverage of other targets in the

panel. A linear relationship was demonstrated between spike in mix concentration, percent of sequencing reads consumed, and successful detection of ERCC transcripts present at low molecular abundance. Expression correlation between identical sample replicates demonstrated high reproducibility at every dilution of spike in mix, including the least concentrated twenty-fold dilution. The results of this study demonstrate a unique and promising application of synthetic spike in transcripts in a hybrid capture assay.

Dedication

This effort is dedicated to the individual patients and families affected by cancer diseases, and the brilliant scientists and clinicians working hard to make that number smaller by each day.

Acknowledgements

This work would not have been possible without the support of many people. Thank you to Dr. Daniela Munafo for graciously stepping in as my thesis director and providing support and guidance throughout this process. Thank you also to Dr. Christine Malbouef for facilitating the many important discussions about this work. Both of your support, advice, and encouragement has been invaluable for the completion of this project and I am extremely appreciative of all the extra hours you so generously spent on this effort.

Thank you to my thesis advisor Dr. Steven Denkin for your guidance and direction over the past few years as this project has come to fruition. Thank you to my friends, family, and colleagues who have provided unfailing moral support throughout this journey.

Table of Contents

List of Tables

List of Figures

Chapter I.

Introduction

Understanding the Importance of Genetics

In 1869, swiss chemist Fredrich Miescher first identified a material referred to as 'nuclein' inside human white blood cells. Originally planning to characterize leukocytes (white blood cells), Miescher encountered a substance in cell nuclei that had properties very different from the cellular proteins catalogued at this point in time. This material had significantly higher phosphorous content and exhibited resistance to proteolysis (Dahm, 2008). This mid-1800's discovery went largely unremarked at the time, until 1910 when Albrecht Kossel was finally awarded the Nobel Prize in Physiology or Medicine for determining the chemical components of nucleic acids, now widely known as the building blocks of DNA and RNA (Jones, 1953). This discovery added to the framework of genetic inheritance established by Gregor Mendel and Charles Darwin decades earlier. An overall trend occurred in the field of genetics; many discoveries were made in the 1800s, but it was not until the 1900s and beyond that technology allowed researchers to better understand the function and mechanisms behind these molecules and structures. In the mid-1950s, Rosalind Franklin's work in X-ray crystallography lead to the discovery of DNA fibers, which proved paramount in enabling James Watson and Francis Crick experiments that revealed the double helix structure of DNA. Although this discovery would not have been made possible without Rosalind Franklin's initial work, she received no credit for their discovery and died a few years later after a short battle with

cancer, likely exacerbated by time spent using carcinogenic material to perform her research work (Maddox, 2003). This discovery of the double helix structure triggered a frenzy of genetic research, which has remained a burgeoning field to this day. With the advent of new technologies, a myriad of information has been generated and the concepts are understood much more in depth, but the initial findings have remained as follows. DNA is a double stranded molecule; the two strands are connected by hydrogen bonds, and on each side of the bond is a nucleotide that pairs with only one other type of nucleotide. There are four types of nucleotides (chemical bases) that make up the DNA that holds genetic code: adenine (A), cytosine (C), guanine (G), and thymine (T). RNA molecules have an uracil (U) nucleotide instead of thymine. The DNA double helix is anti-parallel, meaning that the 5' end of one strand pairs with the 3' end of its complementary strand (Várnai & Zakrzewska, 2004). This enables replication, the process necessary for life to occur.

DNA Sequencing Technology, Then vs Now

In the early 1950's, a British biochemist by the name of Frederick Sanger was the first person to determine the complete amino acid sequence of the two polypeptide chains of insulin, initially using bovine insulin for his testing. From here, Sanger then explored methods aimed at determining the sequences of RNA and DNA molecules. In 1977, Sanger and colleagues introduced a new approach of DNA sequencing that would become known as 'Sanger' Sequencing. The Sanger technique used a chain-termination method with DNA primers, polymerase, ssDNA template, deoxyribonucleotides (dNTPs), and dideoxy ribonucleotides (ddNTPs) which were modified to terminate DNA template extension (Men et al., 2008).

In 1987, Applied Biosystems introduced the first automated sequencer, which utilized capillary electrophoresis to determine nucleic acid sequences. These capillary sequencing machines and Sanger sequencing methodology were later used as the main tools to complete the sequencing of the Human Genome Project in 2003.

At present, the most well-known sequencing technologies are Sanger Sequencing and Next-Generation Sequencing (NGS). NGS is similar to the Sanger method in that the template strand is fragmented and the bases in each fragment are identified by a signal when the fragments bind to the template strand. However, the older Sanger method sequences a single DNA fragment at a time, whereas the newer NGS method sequences millions of fragments during the same run (Lundin et al., 2010). Sanger sequencing is a fast, cost effective method used to sequence a very small number of targets whereas NGS is a much more robust, high throughput method to generate hundreds of times more data from the same amount of input DNA (Fuller et al., 2009). Where the Sanger method can only sequence one fragment at a time, NGS is 'massively parallel' meaning that millions of fragments can be sequenced all in one sequencing run.

There are several types of recently developed platforms for sequencing. Some technologies, such as the PacBio Sequel or Oxford Nanopore, can sequence very long fragments up to tens of thousands of base pairs in read length (Jain et al., 2016). The most widely used NGS platform is Illumina sequencing, which generates millions of shorter reads with rapid turn around time (Loman et al., 2012). This method uses a chemistry called sequencing by synthesis (SBS) to amplify and read the individual bases contained on segments of DNA. In the first step of sequencing, the double stranded molecules of the sample are denatured into single strands that can cluster and bind to the flow cell for

the instrument to read (Shendure et al., 2011). DNA polymerase is used to extend and amplify the clusters of DNA that are bound to the sequencing flow cell. The nucleotides that bind to the DNA template are chemically modified and contain a fluorescent tag, which emits a unique signal to indicate which base is present (A, T, C, or G). After fluorescing, the base is cleaved so the next base can bind, fluoresce, and be cleaved again. For paired end sequencing, after the forward strand is read by the machine, the reads are washed away, and the process repeats for the reverse strand. This cycle continues over and over, resulting in base-by-base sequencing reads with high accuracy and low error rates. Ultimately this method enables the generation and collection of high-fidelity data from hundreds or thousands of genes, allowing for greater power of detection for low frequency variants or low abundance molecules (Grada & Weinbrecht, 2013).

Library Preparation for Sequencing

Historically, to obtain the genetic code of any organism, nucleic acid (DNA or RNA) must be extracted from the source, purified, and processed through a series of steps within the laboratory. The end result of this multi-step process is a chemically modified sample in a form that enables the genetic code to be read by a sequencing instrument. Before DNA sequencing become widely used and in high demand an older, antiquated method of library preparation called whole-genome shotgun sequencing was used (Ventner et al., 2001). In this method, the DNA is extracted and purified, then fragmented and cloned into a universal cloning vector, such as *E. coli*. These smaller fragments of DNA have portions of their code that overlap with neighboring segments and after

sequencing completes can be re-assembled into a whole genome by using this overlap information (Rothberg & Leamon, 2008). The assembled genomic structure is called a contig, short for contiguous sequence (Weber & Meyers, 1997). This antiquated approach is referred to as shotgun sequencing because it mimics the rapid, random pattern of a shotgun. This technique worked acceptably well in the early days of DNA sequencing, but more sophisticated strategies for library preparation have been developed and are now used.

Though there are a variety of technical approaches and slight nuances or exceptions to each, the general principles used to create 'libraries' of molecular information remain the same. After nucleic acid extraction the first step of sample manipulation for whole genome and targeted sequencing is to break apart the segments of DNA into smaller fragments. The commonly used Illumina sequencing technology utilizes read lengths much shorter than the length of a genome, so in order to sequence all of the genetic information comprised in a genome the extracted material must be separated into shorter fragments compatible with sequencer read length. Sample fragmentation can be achieved by using acoustic waves to shear the sample, through enzymatic or chemical reactions, or the tried and true method of gel-based fragment size selection.

After the sample has been broken into smaller fragments that fall within the desired size range, the fragment ends are 'repaired' through an enzymatic process. The rough ends of the molecules are blunted and phosphorylated at the 5' ends with polynucleotide kinase and DNA polymerase. Following end repair, an A-base is added to the 3' ends of the molecules using a mix of enzymes. After 3' adenylation, adaptors are

ligated onto each end of the molecule. These adaptors must be specifically designed to be compatible with the sequencing machine, reagents, and flow cell to facilitate clonal amplification on the flow cell and ensure each sample is readable uniquely identifiable once loaded onto the sequencer (Van Dijk et. al, 2014).

One exception to the general overview of dsDNA library preparation described above is observed with RNA sample preparation. RNA differs from DNA in its single stranded nature and the substitution of uracil nucleotide instead of thymine. Due to the single stranded composition of RNA, it tends to be more fragile than DNA molecules and requires additional upfront manipulation before library preparation. For most NGS methods, total RNA must first be isolated from tissue or cells, purified, then different parts of the RNA are selected depending on downstream application. For example, most total RNA preparation includes an enzymatic step to remove globin transcripts, which make up a large percentage of the total RNA but hold no transcriptional value in downstream analysis. Another commonly used technique employs Oligo dT magnetic beads to select the poly-A tailed mRNA from total RNA.

After isolation and purification, chemical fragmentation is used to create the desired size of single stranded RNA molecules. In the following step, primers are added to enable extension by DNA polymerases or reverse transcriptase enzymes in subsequent reactions. In common practice random primers are used to give unbiased coverage to all regions of the RNA, which generates cDNA of varying lengths. After primer addition, the first strand is synthesized by adding reverse transcriptase to an RNA template in order to bind to the 3’ end and facilitate extension. The second strand is created by using DNA polymerase enzymes and a similar method of primer binding and extension. The end

result of cDNA synthesis is generation of double stranded molecules made from the single stranded RNA template. This double stranded molecule is now ready for downstream library preparation the same as a DNA sample.

In most DNA and RNA library preparation workflows, the final step of library construction is a selective amplification to make more copies of the fragments present and increase sample yield. Increasing the of the number of unique molecules that have been modified and adaptor ligated will give each sample the best possible chance at generating valuable sequencing data. This amplification occurs through a process referred to as polymerase chain reaction (PCR), which uses varying temperature cycles to initiate reactions with primers, enzymes, and oligonucleotides such as dNTPs. Both a forward and a reverse primer are used to amplifying the selected fragments and must be complementary to the template DNA. The forward primer initiates the start of PCR amplification as it binds to the start codon or region of the template DNA, whereas the reverse primer binds to the stop site of the complementary strand.

In the first step of PCR, the double stranded DNA is denatured into a single strand by heating at a high temperature. Once the sample is denatured, the reaction temperature is lowered, and primers anneal to their complementary site. Next, the temperature is raised slightly to initiate polymerase binding to the strands and extending across the fragment by adding complementary nucleotides to each template strand. This generates a double stranded copy of the original fragment. The PCR amplification cycle of denaturation, annealing, and extension is repeated several times, resulting in many copies of the original DNA molecule (Quail et. al, 2012). Outside of the library construction application, this technique can be performed on genomic, plasmid, or cDNA.

The success of the library construction process through sample yield is verified or quality checked (QC) before sequencing or downstream enrichment. A variety of methods are available for determining fragment size and relative concentration of a sample. Absolute quantitative analysis can be performed using a Real-Time PCR (qPCR) method to determine the amount of material that would be amplifiable on a sequencing instrument (Wang et al., 2009). Similar to the amplification method described above, qPCR utilizes complementary primers that bind to target sites in order to generate more copies of the molecules in solution from a small volume aliquot of the stock sample. This is used to quantitate the specific molecules present by measuring the amount of cycles or time needed to cross a set abundance threshold (C_T). Whereas PCR amplification increases the number of molecules in a library and results in an enriched product with increased yield, the qPCR method is used to quantify and does not affect the actual molecules present in a sample.

Targeted Enrichment and Hybrid Capture

As DNA sequencing technology progressed it became evident that there were many future applications beyond just sequencing the whole genome of an organism or individual. Whole genome sequencing remains a reliable method to provide a broad picture of genetic makeup. However, given its large size and complexity, sequencing a whole genome to deep coverage is extremely costly and time consuming. This led to the development of workflows that target specific regions of the genome with known or unknown significance, allowing a small percentage of the whole genome to be sequenced at a much higher coverage much more efficiently.

By sequencing smaller regions at much higher coverage, researchers are able to

gain a deeper more comprehensive view of how any given individual's genetic makeup correlates with known alternations that underlie disease (Meyer & Kirchner, 2010). Within the past decade, a newer method has been developed to target smaller more specific regions of interest within the genome, providing a more efficient and cost-effective approach for research and diagnostic goals (Rizzo & Buck, 2012). It is widely recognized that the protein coding region of the genome accounts for less than 2% of the whole human genome (Gnirke et al., 2009). Trying to make inferences about protein coding or expression from whole genome data would be challenging and futile.

One of the first techniques developed to sequence smaller target regions utilized oligonucleotide bases (i.e. PCR primers) complementary to the region of interest. The sequence of the target region must be understood in order to design the complementary primers for this method to work efficiently. This technique aims to enrich and amplify only those specific target regions, but as with any PCR reaction there are duplication, bias, and errors (Fuller et al., 2009).

This targeted amplification method can also be coupled with microarray technology to enrich regions of interest within the genome. Multiplex amplification utilizes microarrays to synthesize oligonucleotide probes that can then be cleaved and amplified via polymerase chain reaction. (Head et al., 2014).

Another common method is microarray capture, which hybridizes oligonucleotide probes matching the region of interest to an array chip. This method is limited by the number of amplicons that can be multiplexed at one time. Another limitation of microarray capture is fragment size, which does not align with the median size (120bp) of human protein coding exons (Gnirke et al., 2009). Directly amplifying specific targets or

amplification via hybridization on an array have both been shown to generate a large amount of non-specific amplification products or off target material (Krishnakumar et. al, 2008). This conflicts with the original intent of narrowing down the genome size to only sequence regions of interest.

A more recently developed technique, referred to as 'hybrid capture' or 'hybrid selection', utilizes the efficient kinetics of in-solution hybridization coupled with the robustness of microarray oligonucleotide synthesis (Gnirke et al., 2009). The oligonucleotide probes are designed to be complementary to the target region of interest, by using in vitro transcription with biotin conjugated uridine triphosphate (biotin-UTP) a single stranded RNA molecule is created that will bind to the targets present in the mass of adaptor ligated human genome DNA. During the following 'capture' portion of this process, streptavidin beads are used to bind the biotin-UTP probes and their complementary target region of the sample (Gnirke et al., 2009). After sample-bead binding, a series of washes are performed to remove any off-target material left over from the hybridization, while the target sample fragments conjugated to beads are retained through magnetism. In theory, after the capture and washes, all that remains is the small fragments of the target regions of interest. Because most of the genomic material has been washed away (e.g. 98% of the whole genome is considered 'off' target if capturing the human whole exome), PCR amplification is performed on the final product to enrich the targets captured and provide enough experimental yield to be sequenced successfully (Lundin et al., 2010). This hybrid capture method has become so robust that it can be used to successfully capture large regions of the genome (e.g. whole exome or whole transcriptome) but also can be used for very small, focused captures with panels targeting

a subset of genes or the small genomes of organisms such as bacteria or viruses. This dynamic range of capture panel and target size can be accommodated in a standard wet lab hybrid capture workflow with very little manipulation or modification required.

Gene Expression Profiling

Targeted enrichment of smaller regions of the genome as highlighted above has enabled researchers and clinicians to gain a more comprehensive understanding of the biological and cellular processes underlying a disease state. This technique applies to both DNA and RNA libraries. One method to gain a higher level of insight has been developed by sequencing RNA to determine which genes are being expressed at a given timepoint, referred to as gene expression profiling. Understanding how expression levels change with time or experimental condition (i.e. treated vs non treated) provides useful data to make clinical decisions.

Gene expression profiling can be performed by using a variety of platforms and techniques. In recent years the lower throughput methods for gene expression studies such as northern blots and RT-qPCR have been replaced by RNA seq and microarrays. This allows generation of robust, sequencing ready RNA libraries that can be sequenced together in one sequencing run. Where microarray technology tends to be faster and cheaper, transcriptome sequencing (RNA seq) enables gene expression data that is much more accurate and robust than conventional microarray analysis (Park et al., 2019). Using NGS methods to sequence RNA enables a lower limit of detection with higher resolution

of differentially expressed genes. This data has a wide range of useful applications, such as fusion detection, splicing events, and gene expression profiling (Trapnell et al., 2012). This last application is perhaps the most valuable and exciting, since the evaluation of differentially expressed genes can shed light upon the function of a cell or protein cascade at any given time. This application of gene expression data is valuable for both research scientists and clinicians.

High Throughput Scale Up for NGS

With the advent of massively parallel DNA sequencing many companies have invested a large amount of time and energy into scaling up methods to construct sample libraries compatible for Next Generation Sequencing (Mardis, 2008). Whereas the first human genome sequence took over ten years to complete and cost over one billion dollars; today a patient can have their genome sequenced in less than a week for about one thousand dollars (Collins, 2003).

As a result, a bottleneck has emerged in the field of Next Generation Sequencing. While sequencing capacity has greatly increased, protocols for sample library preparation remain a time consuming and expensive limiting step. One solution to this bottleneck is incorporation of robotic liquid handlers to replace the manual processes performed by technicians in the laboratory (Farias-Hesson et al., 2010). Automated systems such as the Agilent Bravo Liquid Handler enable batches of 96 samples to be processed in the amount of time it would take a technician to manually prepare 8 or 12 samples (Holmberg et al., 2013).

In common practice a manufacturer or vendor will supply reagents in kits that

contain all necessary enzymes and chemicals needed to prepare an extracted nucleic acid sample so it can be loaded on a flow cell and read by the sequencer instrument. Depending on the sample batch size and volume of reagents, a technician may prepare samples by hand at the bench, which tends to be slower and more error prone. If the sample batch size is considerably larger, laboratories can use liquid handling robots to process hundreds of samples at one time.

One way this exponential increase in sample batch size is achieved is by using magnetic beads during the enzymatic library preparation process. This technique, referred to as solid phase reversible immobilization (SPRI) has enabled scientists to vastly scale up the number of samples processed within a shorter amount of time (Fisher et al., 2011). The SPRI method utilizes magnetic beads coated in carboxyl molecules added to solution with the double stranded DNA molecule. These beads are magnetic only in a magnetic field (paramagnetic), which prevents the beads from falling out of solution.

Depending chemical composition of buffers used, the magnetic beads will either bind the DNA or the DNA will precipitate out and be released from the magnetic beads. In the presence of a crowding agent such as polyethylene glycol (PEG) and sodium chloride (NaCl), the beads will reversibly bind DNA. The crowding agent pushes out the molecules of water present in solution, which then causes the negatively charged DNA to bind to the carboxyl groups on the surface of the beads. Presence and amount of sodium ions in a given buffer will change the charge of the solution, which leads to the release of bound dsDNA molecules (DeAngelis et al., 1995). The ratio of beads to DNA volume is critical given the immobilization is dependent on concentration of crowding agent and salt in the reaction.

By adjusting these conditions, the DNA sample in solution will be either bound to the beads or precipitated off the beads. Being able to control this magnetic binding enables purification of the DNA molecules from enzymes and other reagents used in the library construction process (Head et al., 2014). For most bead-based library preparation methods, the magnetic beads are added to the sample following fragmentation and remain in the vessel throughout the sample preparation workflow. Using the same beads for each step decreases the number of liquid transfer steps, and subsequent material loss are greatly reduced, allowing for higher sample yield and better sequencing results.

As library preparation methods scale up and the number of samples processed in each batch increases, quality checks become even more important (Fisher et al., 2011). At high throughput, one lab processing error could affect hundreds of samples. This highlights a need for a reliable and reproducible control sample to serve as a reference to distinguish between potential errors in wet lab processing versus sample-to-sample performance variability.

Process Controls for NGS

Process control samples, or process match controls, are an essential part of ensuring a wet lab test meets important criteria regarding sample quality. In almost every type of Next Gen Sequencing assay, process control samples are used as surrogates for patient specimens, processed as if the sample came from an individual patient, and subject to each step of the assay in the same regard as all other experimental samples. This method is used to monitor the overall performance of the entire process, both by individual batch and through historical assay data (Mardis, 2008). Molecular tests such as DNA sequencing usually include positive and negative controls. When an assay is being

developed, it is common and almost expected that sensitivity controls with known expected values will be used to demonstrate that a target is still detectable even at low levels of analyte (Ansorge, 2009).

In the past few decades enormous resources have been dedicated to developing process match controls. Analysis of sequencing data is challenged by the complex nature of human genetic variation as well as confounding errors inherently introduced during sample preparation and sequencing itself (Linnarsson, 2010). Biological and/or synthetic controls are used as a baseline reference measurement to determine origins of variation. These control samples are intended to verify assay performance at relevant analytical and clinical decision points. For example, multiple myeloma is a cancer that affects plasma cells and is characterized by a several step transformation of normal to malignant cells. Within the past decade, transcriptomic studies and gene expression profiling have been used to shed light upon this complex process, resulting in a more accurate diagnosis and better therapeutic options for treatment (Szalat et al., 2016). These studies depend on reliable and reproducible controls to serve as a baseline reference for analyzing changes in cellular composition and identifying significance in the levels of gene expression measured. Another application of controls in the clinical setting is demonstrated by biomarker use for drug development. In almost every situation, a clinical biomarker must first be validated with a set of well characterized and thoroughly understood control samples before any interferences can be responsibly made about treatment of an individual patient (Lesko et al., 2001).

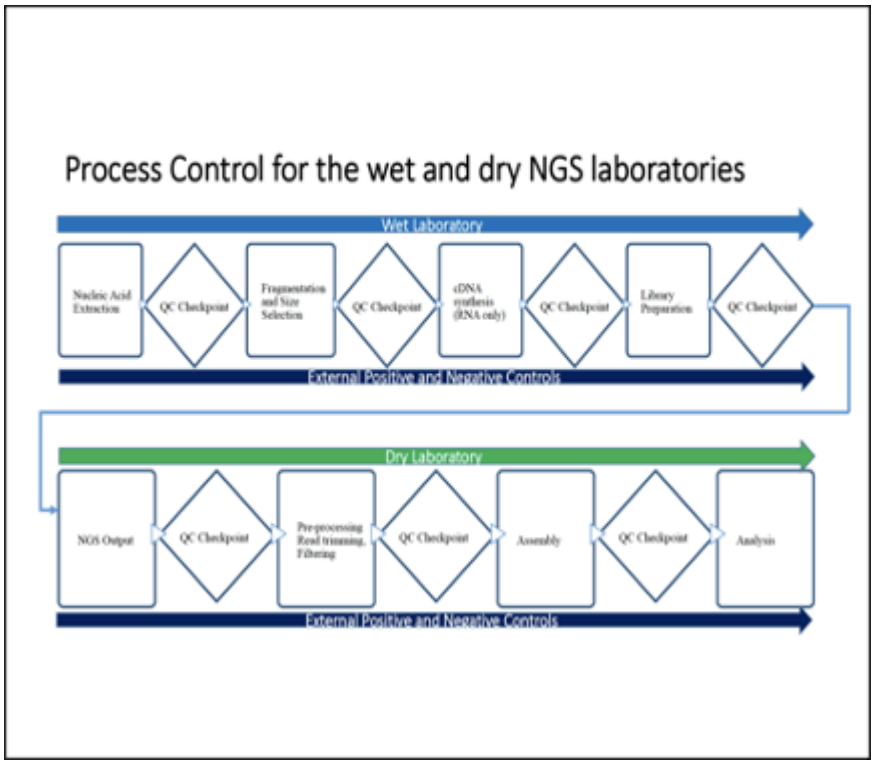

Figure 1: Application of In Process Controls.

FDA overview of process control applications for both wet and dry laboratory.

Next generation sequencing has three main areas in which control samples are utilized. The first is during nucleic acid extraction. Controls ensure that the extraction was performed successfully, and the sample meets criteria to continue preparation for sequencing. The second area is during sample library preparation itself. Due to the many complex enzymatic steps in this process it is critical to every step was executed correctly and any samples not meeting criteria will fail due to a sample related issue as opposed to an error in processing (McKenna et al., 2010). Lastly, a sequencing control (e.g. Illumina PhiX) is often included in the pool of samples as a final step before the samples are

loaded on to the sequencer (Rizzo & Buck, 2012). This illustrates another reason why controls are so important for Next Generation Sequencing assays due to the exorbitant cost of running these tests, both financially and in terms of turnaround time.

The type of samples used for process controls are available in a wide range of material and isoforms; from widely characterized reference cell lines (i.e. NA12878 or material from Master Cell bank) to synthetic spike in controls such as sequins or RNA transcripts designed by the External RNA Control Consortium (ERCC) (Blackburn et al., 2019).

Reference cell lines, as mentioned above, have become increasingly valuable and informative as Next Generation Sequencing technology has exploded in the past few decades. In 2001, the International HapMap Project was launched with the goal of establishing a haplotype map ('HapMap') of the human genome. Haplotype maps shed light on which regions of an individual's genome tend to be inherited together, and how this phenomenon varies by country and ethnicity (Gibbs, et al., 2003). While a haplotype refers to a set of genetic variations inherited together, a genotype refers to an individual or cluster of traits inherited together; i.e. the combination of genes at a given locus (Manolio et al., 2008). By understanding the patterns of genomic variations in humans, researchers can make inferences on a broad scale about genetic links to a multitude of diseases around the globe (Liu et al., 2004). As part of the project, all HapMap data are widely accessible to the public through an online database.

Phase I HapMap was completed in 2005, comprised of over one million single nucleotide polymorphisms (SNPs) generated from hundreds of individual participants. These individuals were sourced from four distinct and genetically diverse populations:

Beijing, China; Ibadan, Nigeria; Utah, U.S.A (with European ancestry); and Tokyo, Japan (Manolio et al., 2008). These data shed light upon how genetic sequences vary by region or continent.

The second phase of the HapMap project added over two million SNPs to the original database of the hundreds of individuals generated in the first phase to further support its utility. Enabled by this comprehensive and widely available database, HapMap cell lines and specimens are often used for NGS assay validation and as a process match control during sample processing (Buchanan et al., 2012). Because HapMap cell lines such as NA12878 are so well characterized and widely used it is well suited as a reference specimen to establish baseline accuracy for confounding factors such as insertions, deletions, or single nucleotide variants (SNVs) (Manolio et al., 2008). Another common application of in process controls is as a reference in RNA gene expression profiling assays. At present, there are several different types of refence RNA control materials commercially available as mixes of tissues or cell lines from several manufacturers. Common reference RNA mixes include Universal Human Reference RNA (UHRR) and Human Brain Reference RNA (HBRR). Universal Human Reference RNA is comprised of ten human cell lines, each individual cell line adds its own unique set of genes and expression levels to the UHRR mix as a whole (Novoradovskaya et al., 2001).

RNA controls in gene expression profiling assays can also be used as a normalization tool. For example, in microarray gene expression profiling, the signal of a probe is measured as the ratio of experimental RNA to UHRR targets (as opposed to absolute signal intensity). This decreases variability in the raw data by normalizing signal

output, giving more accurate readings and decreasing sample CV (Novoradovskaya et al., 2004). These widely characterized controls are also used in RNA seq and NGS experiments as a reference for gene expression levels. The rigorous quality standards necessary for a universal RNA reference control enables a well characterized composition of RNA at different expression levels that remains consistent and reproducible across replicates and across batches. This reproducibility in a process control is crucial for NGS assays in which sequencing artifacts and process errors can easily confound data.

Overview of ERCC Spike-Ins

As RNA sequencing has become more popular and in greater demand, it was clear that a common set of RNA based controls were needed to standardize gene expression measurements. In 2003, the External RNA Control Consortium hosted by National Institute of Standards and Technology (NIST) addressed the lack of standardized RNA controls for gene expression profiling assays and dedicated a large effort towards meeting this need (Devonshire et al., 2010). To date, the ERCC has established two sets of 92 synthetic transcripts designed to mimic eukaryotic mRNA sequences, with each set in a mixture ranging 6-fold in abundance. The transcripts are 250 to 2,000 nucleotides in length and are traceable through manufacturing to the NIST plasmid reference material. The spike in mixes are designed to be added to isolated RNA samples before downstream processing. These synthetic transcripts can be used to evaluate many different technologies for gene expression measurement such as microarray, real-time quantitative PCR (RT-qPCR), and next generation RNA sequencing (Baker et al., 2005). In addition to the evaluation of gene expression platforms, ERCC spike in mixes can also be used as

a quality control for library amplification workflows (Jiang et al., 2011).

Each ERCC set of 92 transcripts is comprised of four smaller sub-pools. Analyzing the fold-change measurement between these mixes provides robust and reproducible expression data. For example, the dynamic range of transcript abundance in each ERCC sub-pool and the individual transcript ratio across sub pools enables calculation of known relative differences between the pools across a large range of dynamic abundance (Devonshire et al., 2010). This design facilitates signal response assessment of individual ERCC transcripts at various folds of abundance, e.g. 1, 3, 5-fold increases in concentration. Additionally, performing a pairwise comparison of the transcript abundance in each pool enables a ratio-based evaluation of the dynamic range (Devonshire et al., 2011).

Applications of ERCC Spike In Mixes

Ambion ® ERCC RNA Spike-In control mixes are used as an external synthetic control to assess gene expression measurements and sensitivity to detect transcripts at low abundance across different platforms. The ERCC spike in mixes come in two formulations, referred to as mix 1 and mix 2. Each mix contains 92 transcripts spanning a 10^6 fold concentration range (Lemire et al., 2011; Figure 2). ERCC spike in mixes are most commonly used in microarray and RNA-Seq for gene expression profiling. A notable difference between using ERCC spike ins for microarray versus RNA seq is highlighted in the manufacturer guidelines, which strongly recommends that microarray assays should include probes complementary to the ERCC transcripts in order to fully capture and utilize ERCC transcript data. Depending on the sample type and purpose of

the assay, these probes may have to be modified to achieve the most efficient performance.

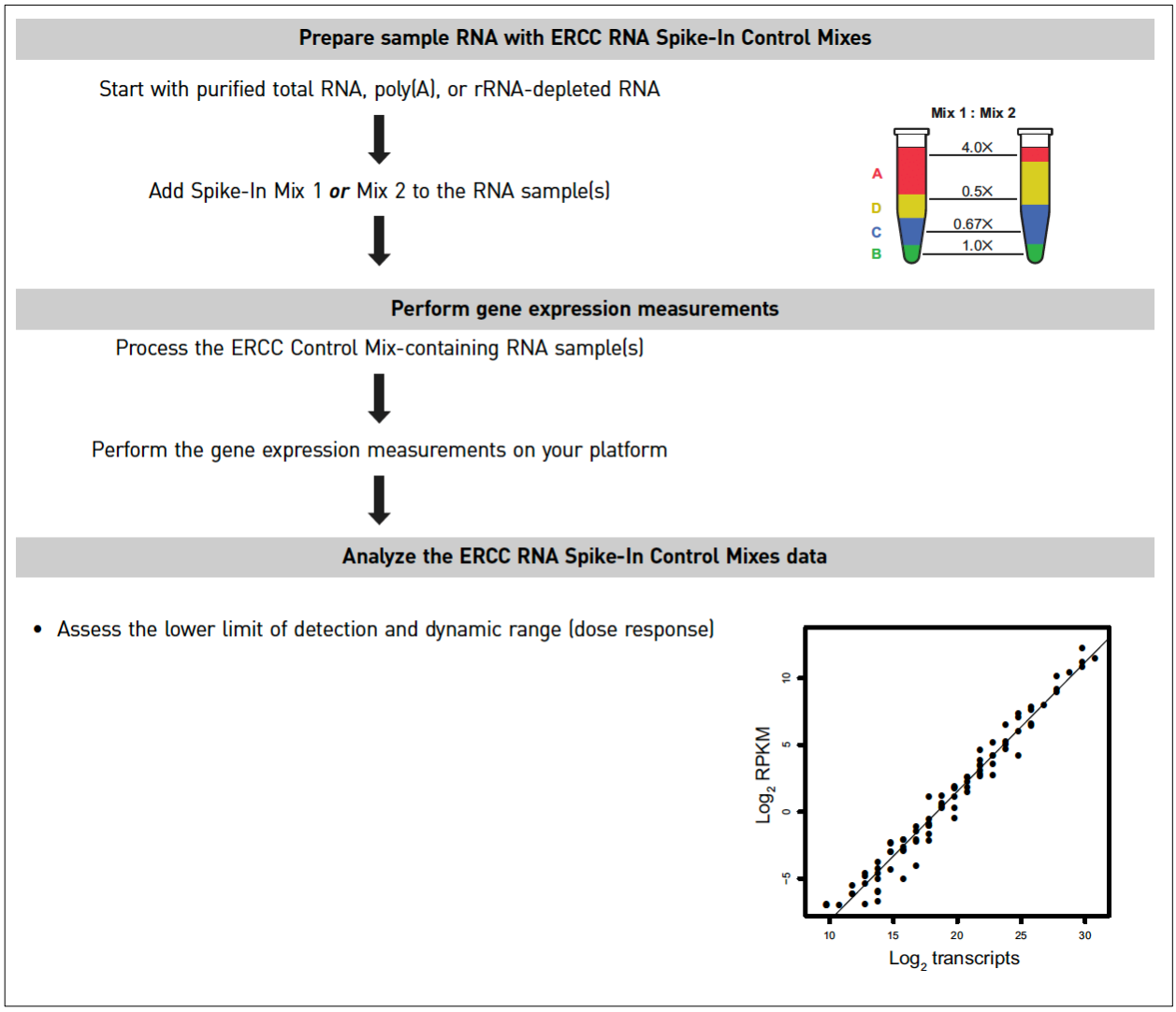

Figure 2. Application of ERCC Spike in Control.

In vitro application of ERCC transcripts as depicted by Ambion® user guide https://www.thermofisher.com/order/catalog/product/4456740?us&en#/4456740?us&en

Synthetic ERCC transcripts are well suited for expression measurement assays because they provide a reliable and reproducible reference value, whereas RNA transcript detection and gene expression levels are highly variable by nature. Whether or not a given gene is even expressed, let alone detected, can change based on tissue type, cell count, time point, and a multitude of other factors. These dynamic ratios of transcript abundance are a way to measure biological activity in a given system, e.g. gene expression profiling (Munro et. al, 2014). The

The ERCC RNA spike in mixes are designed to be added to the following type of samples: purified total RNA, poly(A) RNA, and rRNA-depleted RNA. Manufacturer guidelines recommend adding either Spike In Mix 1 or Spike In Mix 2 to RNA samples for dynamic range and lower limit of detection. To measure fold change response, manufacturer guidelines recommend adding Mix 1 and Mix 2 to separate samples (e.g. Add Mix 1 to the control sample and add Mix 2 to the treated sample). In standard application the concentration of ERCC spike in mix must be adjusted to determine optimal conditions based on experimental criteria (Garalde et al., 2018; Table 1). Manufacturer guidelines recommend performing serial dilutions with nuclease free water to reach the target spike-in mix concentration based on downstream application.

Dilution	Spike-In Mix†	Nuclease-free Water
1:10	1 µL undiluted	9 µL
1:100	1 µL of 1:10	9 µL
1:1000	1 µL of 1:100	9 µL
1:10,000	1 µL of 1:1000	9 µL

† ERCC RNA Spike-In Mix 1, ExFold Spike-In Mix 1, or ExFold Spike-In Mix 2.

Table 1. Guidelines for Preparing Spike-In Dilution.

These guidelines instruct the user to select a Spike-In Mix dilution ratio based on the starting RNA sample mass (Table 2). The manufacturer recommends increasing the concentration of Spike-In Mix at higher amounts of starting sample RNA for both total RNA and Poly(A) selected RNA.

Amount of sample RNA	Volume of Spike-In Mix 1 or Mix 2 (dilution)†	
	Total RNA	Poly(A) RNA
20 ng	4 µL (1:10000)	2 µL (1:100)
50 ng	1 µL (1:1000)	5 µL (1:100)
100 ng	2 µL (1:1000)	1 µL (1:10)
500 ng	1 µL (1:100)	5 µL (1:10)
1000 ng	2 µL (1:100)	–
5000 ng	1 µL (1:10)	–

Table 2. Ratio of Suggested RNA Input to ERCC dilution.

In addition to identifying differentially expressed transcripts as a reliable reference for gene expression levels, measuring ERCC abundance ratios can provide insight into the lower limit of detection. The limit of detection in an assay is the lowest

quantity of analyte that can be detected against a blank measurement with a predetermined level of confidence. Using ERCC spike in transcripts to establish the limit of detection in an RNA Seq assay or other gene expression profiling assays is often achieved by plotting signal abundance to compare the expected expression or detection of the ERCC transcripts at a known concentration versus the observed values generated by the assay. This is often depicted as a dose response curve with observed ERCC counts or expression per transcript on one axis and ERCC concentration (attomole) on the opposing axis.

This research will use commercially available pre-extracted human reference RNA material comprised from a mix of five different tissue types. This aids in the provision of an accurate representation of the different RNA species present in differing tissues or cells (Islam et al., 2014). Following library construction, the control RNA containing ERCC spike in transcripts will go through target enrichment using hybrid capture methodology. Complementary oligonucleotide probes will be used to bind to the target regions of interest in these samples. Another set of probes will be used to capture the ERCC transcripts spiked into each sample, enabling the detection and quantitation of each individual ERCC transcript present at a known molecular abundance. Analysis of sequencing data will be performed on the raw data, which will involve many steps such as removal of duplicate molecules and regions that are not of interest, e.g. ribosomal RNA (Zhao et al., 2014). The goal of this study is to determine the optimal conditions to enable application of ERCC spike in mixes in a hybrid capture assay instead of the commonly used application in total RNA Seq. This will be achieved by adjusting spike-

in concentration and capture panel size while analyzing sensitivity, reproducibility, and lower limit of detection of the ERCC transcripts.

Experimental Design and Expected Outcome

This study will investigate the relationship between ERCC spike in concentration, hybrid capture panel size, and sequencing reads consumed by ERCC transcripts. This will be accomplished by a series of experiments testing various dilutions of ERCC transcripts with panels of the following sizes: 30 Mb, 1.5 Mb, and 650 kb. The first experiment of this study will test ERCC spike in transcripts captured alongside a large panel. Based on these results, further work will be performed to test ERCC spike in transcripts enriched via hybrid capture alongside two smaller sized panels. It is expected that several modifications will be made to the ERCC spike in mix dilution and probe set in order to maintain parameters for sensitivity and lower limit of detection across the two different panel sizes.

This study will use a series of experiments to address three main objectives:

- Determine whether ERCC spike in transcript mixes can be applied to a capture based assay in alignment with application for RNA seq method
- Understand how the relationship between panel size and read consumption affects sensitivity of transcript detection
- Establish optimal concentration in HC assay that achieves and maintains high sensitivity and reproducibility

We hypothesize that the same concentration of ERCC spike in transcripts

captured with a smaller panel will result in the ERCC transcripts consuming more sequencing reads compared to capture with a larger panel. If the ERCC transcripts receive high sequencing coverage this will take away coverage from more important target regions in the pane, which is a detrimental effect given that adequate sequencing coverage target regions is crucial in the development of a successful assay. The concentration of ERCC transcripts will need to be further diluted to accommodate for a smaller panel size. The ideal concentration of ERCC transcripts will enable accurate detection without taking up too many valuable sequencing reads. To investigate this relationship across several experiments we will generate plots that illustrate the read distribution across targets, including reads covering ERCC transcripts. We expect to see a tradeoff between concentration and detection limit, as well as decreased sensitivity with increasing dilution factor.

Chapter II.

Materials and Methods

Preparation of Control RNA:

A commercially available and widely used reference RNA control sample was needed for this study. The ideal sample is comprised of different cell lines or tissue types given that RNA transcripts differ in abundance between cell lines or tissues, and the material used should mimic a biological sample as closely as possible. It is expected that a well characterized control mix will be reliable and reproducible across varying experimental conditions. A total RNA control sample comprised of 5 tissue types met the criteria above and was selected for use in this study. The mix of total RNA purchased for this study is designed specifically to be a reliable process control in gene expression profiling and other RNA based assays.

The total RNA mix was received in several tubes containing 25 µL RNA per tube at a concentration of 1µg/µL. This material had to be modified before addition of ERCC transcripts and downstream processing in order to ensure the control RNA mimicked a biological sample as closely as possible. Before starting any bench work with the RNA sample or ERCC spike in mixes, the space was decontaminated using bleach and disinfected with 70% EtOH. New, unopened pipette tips and reagents were used at every step to minimize possibility of contamination and preserve integrity of the RNA material.

RNA Dilution: First, the control mix had to be diluted from the stock concentration it was received at in order to fall within a range suitable for ERCC spike in

mix addition and assay input parameters. Nuclease free water was used to dilute the control RNA sample using the manufacturer listed starting concentration. A bulk dilution of RNA sample was made, which was then aliquoted into several smaller volume replicates to generate enough material for any repeat experiments and avoid unnecessary freeze thaw cycles of the RNA material.

Fragmentation: Ambion® RNA fragmentation reagents and manufacturer guidelines were used to fragment the total RNA sample into smaller segments to facilitate downstream sample manipulation and library construction.

- 2 µL of 10X Fragmentation Buffer was added to 18 µL of control RNA sample.
- The sample and buffer were mixed, centrifuged briefly, then incubated at 70°C for 15 minutes using an Eppendorf Mastercycle thermal cycler.
- After incubation, 2 µL of Stop Solution was added to the RNA sample.

Sample Clean up: After fragmentation, a filter based purification was performed on the sample to remove enzymes and other compounds that might degrade or compromise sample integrity. Zymo® RNA Clean and Concentrator Kits were used as follows.

- 100 µL RNA Binding Buffer was added to each sample and pipette mixed upon addition.
- 100 µL 100% ethanol was added to the solution above and mixed.
- This solution was transferred to the Zymo-Spin™ Column and centrifuged for 1 minute. Flow-through supernatant was discarded after centrifugation.
- 400 µL RNA Prep Buffer was added to the column and centrifuged for 30 seconds. Flow-through was discarded after centrifugation.

- 400 µL RNA Wash Buffer was added to the column and centrifuged for 30 seconds. Flow-through was discarded after centrifugation.
- 400 µL RNA Wash Buffer was added to the column and centrifuged for 2 minutes to ensure complete removal of the wash buffer.
- The column was then transferred into an RNase-free tube. 100 µL nuclease free water was added directly to the column matrix and centrifuged for 30 seconds.
- After dilution, fragmentation, clean up and QC, the total RNA material is now ready for addition of ERCC spike in mix and downstream library preparation.

Preparation of ERCC spike in transcript mix:

ERCC Spike-In Mixes are synthetic RNA transcripts that serve as a reference for measuring dynamic range, lower limit of detection, fold-change response, and gene expression of an RNA assay or platform. The ERCC spike in mixes are shipped ready to be diluted and added to RNA samples before executing an experiment, with no additional manipulation required. Manufacturer guidelines advise that the spike in mixes are added to the following type of samples: purified total RNA, poly(A) RNA, and rRNA-depleted RNA.

1	2	3
100 ng Total RNA ERCC spike-in mix at 0.25X (n=2)	100 ng Total RNA ERCC spike-in mix at 0.5X (n=2)	100 ng Total RNA ERCC spike-in mix at 1X (n=2)

Table 3. General experimental design. RNA Sample input remained constant at 100 ng.

This study used 100 nanograms (ng) of total RNA for all samples regardless of ERCC transcript concentration, in contrast to the manufacturer guidelines. The input mass of total RNA sample was kept the same across all experiments to serve as a control as the panel size and ERCC dilution varied. At least two technical replicates per condition were included in each experiment (Table 3). Given that the manufacturer guidelines were designed for using spike in mixes in microarrays and RNA seq, there was no information about spike in mix concentration for hybrid capture with targeted panels. Dilution factor was determined empirically based on the assumption that the dilution guidelines for RNA seq would be too concentrated for targeted sequencing and would blow out sequencing read consumption.

Serial dilutions with nuclease free water as outlined in Table 1 were performed to reach each target ERCC dilution factor. Per manufacturer guidelines, a fresh dilution of ERCCs was made for each new batch of total RNA samples using nuclease free water. The diluted ERCC transcript mix was added directly to the total RNA samples before any additional processing or sample manipulation occurred.

Library Preparation and Hybrid Capture:

After addition of ERCC transcripts, the total RNA control material was converted from single stranded molecules to double stranded cDNA molecules. This was achieved by using reverse transcriptase enzyme with an RNA template and complementary primer. Once the RNA molecules were successfully converted to double stranded cDNA, standard preparation methods were used to generate NGS compatible libraries. The libraries went through a quality check to determine success of sample preparation and amount of sample yield before hybrid capture target enrichment. Hybrid capture was performed with panels of varying sizes (30, 1.5, and 0.65 megabases) and each panel also contained probes designed to target the ERCC transcripts. After capture and subsequent batch quality check, samples were pooled together and loaded onto a sequencing instrument to generate sequencing data.

Analysis of Raw Sequencing Data:

Amount of sequencing reads, or coverage, consumed by ERCC transcripts will be a key metric throughout the study. More reads taken up by ERCC targets will result in less coverage and decreased sensitivity for the other targets that each panel was designed to capture. The unfiltered data from the sequencing run will give a raw read count for each sample, which is then broken down to number of sequencing reads mapping to each target region (e.g. genes within a panel or ERCC transcripts). Using the total number of sequencing reads for any given sample, the percentage of reads for each target region is calculated. This method will be used to determine the percentage of reads absorbed by

ERCC transcripts, a metric to compare across experimental conditions and assess performance.

Ambion ® ERCC RNA Spike-In Control Mixes User Guide:

Instructions for normalizing by dilution: *'Before data analysis, it is often convenient to transform the Mix 1 and Mix 2 concentration values to reflect the dilution scheme used. For example, if 2 µL of a 1:100 dilution of Spike-In Mix was added to 1 µg of total RNA, multiply by 0.02 to give new concentration values, expressed as attomoles of ERCC transcript/1 µg total RNA. The concentration values can be expressed in terms of absolute number by conversion of moles to molecules with Avogadro's number (NA; 6.02214179 ✕1023 mol–1).'*

The following guidelines from Ambion ® ERCC RNA Spike-In Control Mixes User Guide were followed to perform data analysis: 'The dynamic range can be measured as the difference between the highest and lowest concentration of ERCC transcript detected in each sample. Some platforms, such as microarrays, have a fairly restricted linear range. In such cases the dynamic range can be defined by the lower limit of detection (LLoD) and the region of signal saturation. NGS platforms do not exhibit a region of saturation, so the dynamic range can be determined by observing the concentration difference between the highest-concentration ERCC transcript detected and the low limit of detection (LLoD). The LLoD is used as a measure of sensitivity, defined as the lowest molar amount of ERCC transcript detected in each sample.

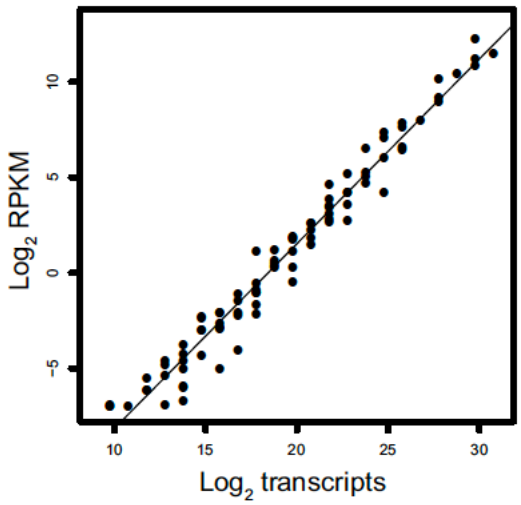

Figure 3. Example dose response curve.

Plot taken from user guide as a method to measure sensitivity and limit of detection. Y-axis corresponds to observed expression measurements for the ERCC transcripts (RPKM). X-axis corresponds to expected transcript abundance (i.e. known molar concentration or amount).

After normalizing and filtering the expression data, the expression signal for each ERCC transcript was plotted against the known molar concentration or abundance. The expression signal measurements are plotted on the Y axis and are usually measured in terms of sequencing reads covering the transcripts; i.e. Reads Per Kilobase per Million mapped reads (RPKM) or Transcripts Per Million mapped reads (TPM). A linear regression was used to determine the best fit line as demonstrated in the dose-response curve above. Manufacturer guidelines define the lower limit of detection as the X-axis value where the regression line crosses the threshold (Figure 3). According to these guidelines, the threshold for defining lower limit of detection may be determined

empirically or arbitrarily.

Reproducibility was measured by comparing the expression values for each individual transcript between two sample replicates of the same dilution and panel size. Expression values for one replicate were plotted on the x-axis of a linear plot, and expression values for the second replicate were plotted on the y-axis. R2 values were used to measure correlation and overall reproducibility across dilution factors.

Chapter III.

Results

Optimization of ERCC dilution for a broad 30Mb RNA expression panel

In order to determine the optimal ERCC spike-in amounts for a large 30Mb expression panel, three different ERCC dilutions were tested in a hybrid capture workflow with the large panel. As expected, the higher concentration of the ERCC dilution spiked-in, the lower the sequencing reads consumed by ERCC transcripts. The number of unique ERCC transcripts detected also decreases with higher dilutions (Table 4).

ERCC Dilution Factor	Panel Size	Number ERCC Transcripts Detected	Percent ERCC Transcripts Detected	Percent Sequencing Reads Consumed
1X	30Mb	73	79.34	0.3
0.5X	30Mb	84	91.31	0.5
0.25X	30Mb	86	93.48	0.9

Table 4: 0.25X, 0.5X, and 1X ERCC Dilutions captured with a 30 Mb panel

At standard sequencing coverage for an RNA expression panel, the majority of ERCC transcripts were detected successfully (79.34% of ERCC transcripts detected at highest dilution). However, not all ERCC transcripts were detected, even at the lowest dilution (86 ERCC transcripts detected at 0.25X dilutions, Table 4). To maximize RNA transcript target coverage, the read consumption by ERCC transcripts was limited to less than 1% total sequencing reads. This trade-off for read consumption resulted ERCC transcript dropouts (86 out of 92 ERCC transcripts detected). This drop out can be

attributed to the competition of the large size of the expression panel with overwhelming number of targets consuming the majority of sequencing coverage (Figure 4).

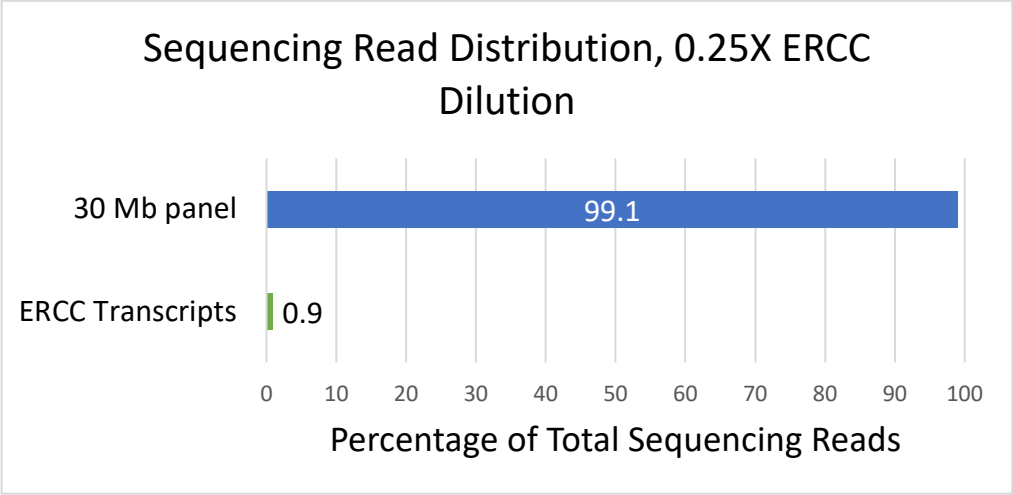

Figure 4: Sequencing read distribution for 0.25X ERCC dilution.

Percentage of sequencing reads consumed by ERCC and RNA expression targets captured with the 30 Mb panel.

Out of three spike-in mix dilutions tested with the 30 Mb panel, none demonstrated the ability to detect all 92 ERCC transcripts in the mix. At the highest concentration of spike-in mix only 86 out of 92 transcripts were detected. Because ERCC transcripts are present in the mix at different molar concentrations, the lower abundance ERCC transcripts are the most likely to be missed.

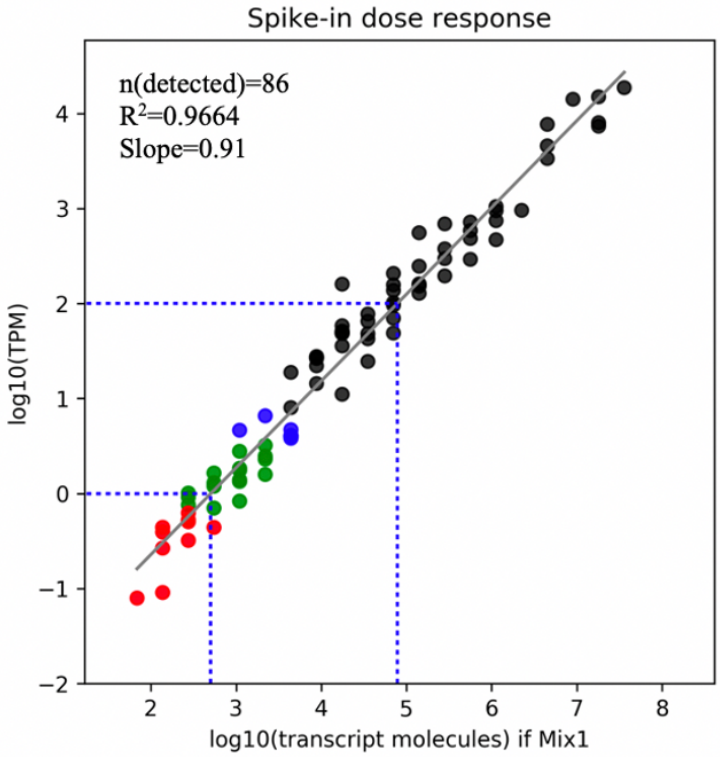

Figure 5: Dose-response curve for 0.25X ERCC dilution.

Captured with 30 Mb panel. Y axis values correspond to expression levels detected for each transcript; X axis corresponds to abundance of molecules per transcript. Marker colors correspond to read count per transcript; red = 1-5, green = 6-20, blue = 21-100, black = >100 reads.

The dose response curves generated for this study, as shown in Figure 5, the markers representing each transcript were colored based on the read count per transcript. Red represents the lowest read count (less than 5 reads), black represents transcripts with over 100 reads. This serves as a visual reminder that more noise in the scatterplot is expected at the lower end of the distribution due to random integer counting differences, i.e. less reads, more variation.

The 0.25X ERCC transcript dilution factor consumed minimal sequencing reads but maintained sensitivity to detect the majority of ERCC transcripts present in the mix. These results are only directly applicable to the 30 Mb panel tested with the three ERCC dilutions (Table 4), how these findings may translate to a panel of different size is unclear.

Evaluation of large panel ERCC dilution with focused 1.5 Mb panel

To assess how the findings above translate to a much smaller targeted panel, a 0.5X ERCC dilution was tested in capture with a 1.5 Mb panel (Table 5). The median dilution factor from the 30 Mb panel evaluation was tested as an initial first pass to better understand the relationship between panel size and concentration of ERCC transcripts.

ERCC Dilution Factor	Panel Size	Number ERCC Transcripts Detected	Percent ERCC Transcripts Detected	Percent Sequencing Reads Consumed
0.5X	1.5 Mb	91	98.91	34.19

Table 5: 0.5X ERCC dilution captured with a 1.5 Mb panel.

At standard sequencing coverage for small targeted panels, the majority of ERCC transcripts were detected successfully (91/92 ERCC transcripts detected). The single transcript not detected is present at the lowest abundance in the mix and is rarely detected in standard capture based assays (Curion et al., 2020). The 0.5X dilution of ERCC transcripts was well suited for capture with a 30 Mb panel based on the amount of sequencing reads consumed while maintaining sensitivity to detect ERCC transcripts. This dilution factor demonstrated an optimal tradeoff between the amount of sequencing reads consumed and the number of ERCC transcripts detected. These findings were directly applicable to the 30 Mb panel tested only, when the same ERCC dilution was tested with a smaller panel, the amount of sequencing coverage consumed by ERCC transcripts took up more than a third of the total sequencing reads (Figure 6).

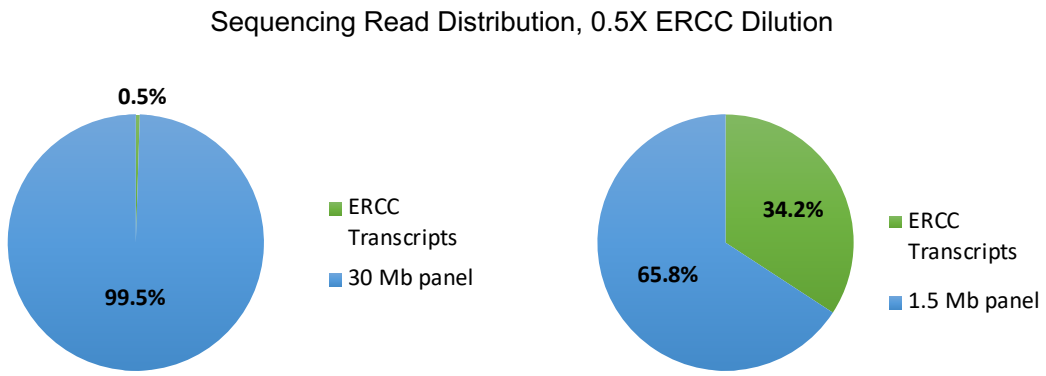

Figure 6: Sequencing reads consumed by ERCC transcripts.

0.5X dilution with 30 Mb panel (left) versus 1.5 Mb panel (right).

The 0.5X ERCC dilution captured with a 1.5 Mb panel took up almost 60x more sequencing reads than the same dilution factor with a 30 Mb panel. These results made it strikingly clear that a dilution optimized for a capture with a large panel does not readily translate to capture with a smaller targeted panel. The increased number of reads consumed by ERCC transcripts at this dilution with a 1.5 Mb panel is likely to be problematic when put into practice. Given that over one third of total sequencing reads are covering ERCC transcripts, less than two thirds of the remaining sequencing coverage is spread across all of the targets encompassed in the 1.5 Mb panel.

Lower Limit of Detection, 0.5X Dilution

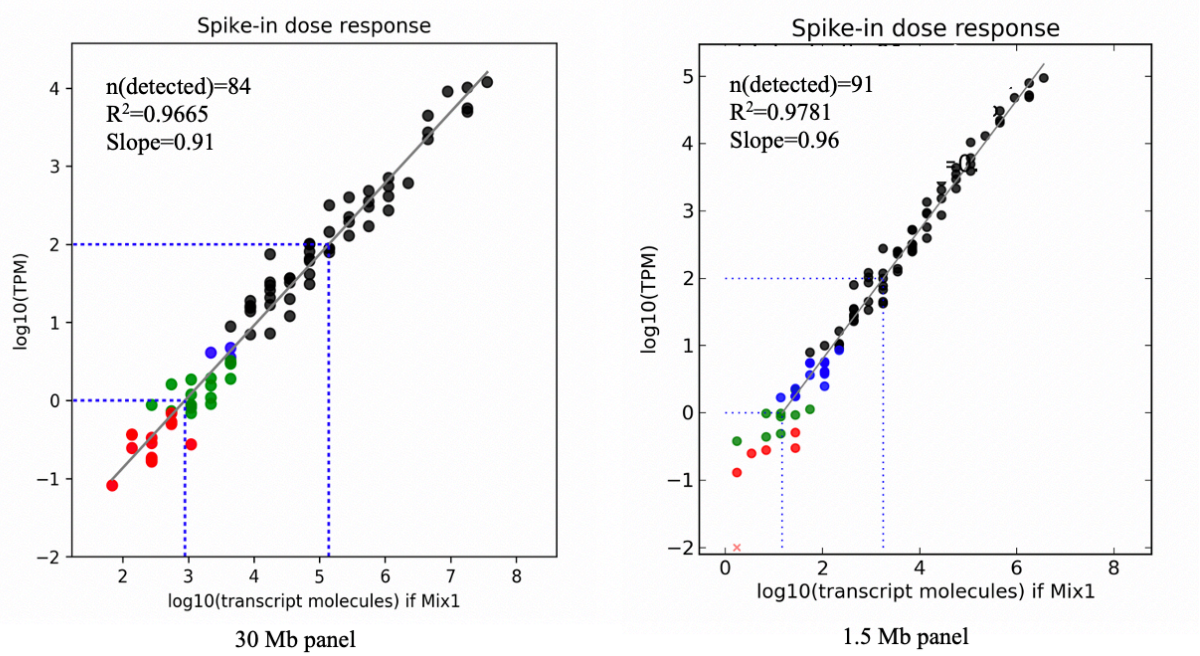

Figure 7: Dose response curves for 0.5X ERCC transcript dilution.

Captured with 30 Mb (left) and 1.5 Mb (right) panels. Y-axis corresponds to observed expression levels for each ERCC transcript (TPM), X-axis corresponds to expected number of transcript molecules present in mix. Marker colors correspond to read count per transcript; red = 1-5, green = 6-20, blue = 21-100, black = >100 reads.

The 0.5X ERCC dilution captured with the small panel that received exponentially more sequencing reads had a higher amount of ERCC transcripts detected than the same dilution captured with a 30 Mb panel (Figure 7). These results confirm the theory that a 0.5X ERCC transcript dilution is better suited for capture with large panel containing many targets. Even though the same dilution with a 1.5 Mb panel received ~60x more sequencing reads, the same dilution in a 30 Mb capture was still able to detect

~92% of the number of transcripts compared to the 1.5 Mb capture but with a fraction of the sequencing coverage (84 transcripts detected with 30 Mb capture versus 91 transcripts detected with 1.5 Mb capture, both at same 0.5X ERCC dilution).

Optimization of ERCC transcript concentration for small targeted panels

Given that the 0.5X dilution optimized for capture with a 30 Mb panel was much too concentrated for capture with a smaller 1.5 Mb panel (determined by sequencing read distribution as illustrated in Figure 7), a titration of ERCC transcript dilutions was performed using two targeted panels of differing size (1.5 Mb vs 650 kb) to fully understand the relationship between each variable depicted in Table 6.

Capture Panel Size: 650 kb

ERCC Dilution Factor	Panel Size	Number ERCC Transcripts Detected	Percent ERCC Transcripts Detected	Percent Sequencing Reads Consumed
1X	650kb	83	90.22	9.65
2X	650kb	77	83.70	5.24
4X	650kb	67	72.83	2.38
10X	650kb	63	68.48	0.97
20X	650kb	55	59.78	0.55

Capture Panel Size: 1.5 Mb

ERCC Dilution Factor	Panel Size	Number ERCC Transcripts Detected	Percent ERCC Transcripts Detected	Percent Sequencing Reads Consumed
1X	1.5Mb	79	85.87	6.58
2X	1.5Mb	75	81.52	3.83
4X	1.5Mb	67	72.83	1.80
10X	1.5Mb	62	67.39	0.75
20X	1.5Mb	54	59.78	0.33

Table 6: ERCC transcripts detected and reads consumed captured with 1.5 Mb and 650 kb small targeted panels.

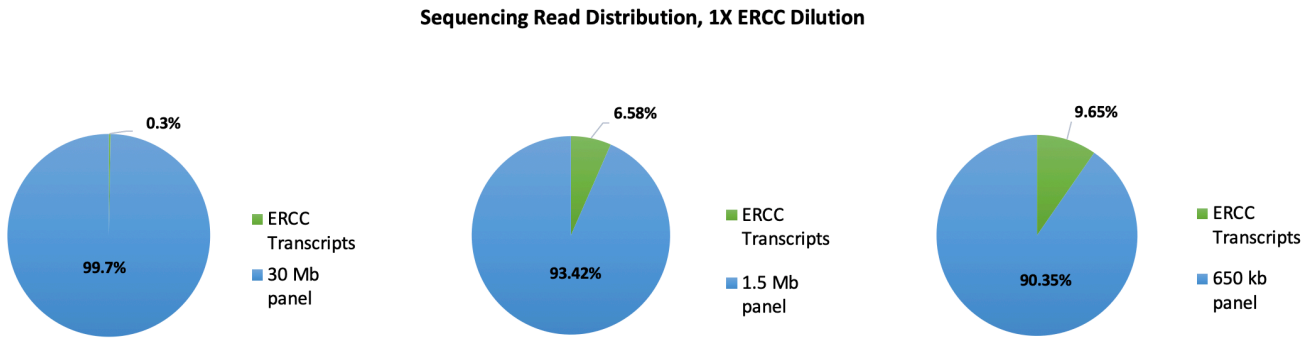

Figure 8: Sequencing reads consumed by ERCC transcripts at 1X dilution.

Captured with 30 Mb (left), 1.5 Mb (center), and 650 kb (right) panels. Green area of pie chart corresponds to percent of sequencing reads consumed by ERCC transcripts. Blue area represents percent of reads consumed by all other targets in the panel (left to right: 30 Mb, 1.5 Mb, and 650 kb

As illustrated above, while ERCC transcript dilution factor remaining constant (1X), amount of sequencing reads consumed by ERCC transcripts increases as panel size decreases. This trend is explained by a central tenet of targeted sequencing; smaller panels with less regions targeted will receive more sequencing coverage under the same sequencing run conditions than larger panels with more targets to cover. To put it simply, it all comes down to a matter of stoichiometry. Though this general concept can be explained in simple terms, applying the underlying principles can be time consuming and costly given the expense of Next Generation Sequencing reagents.

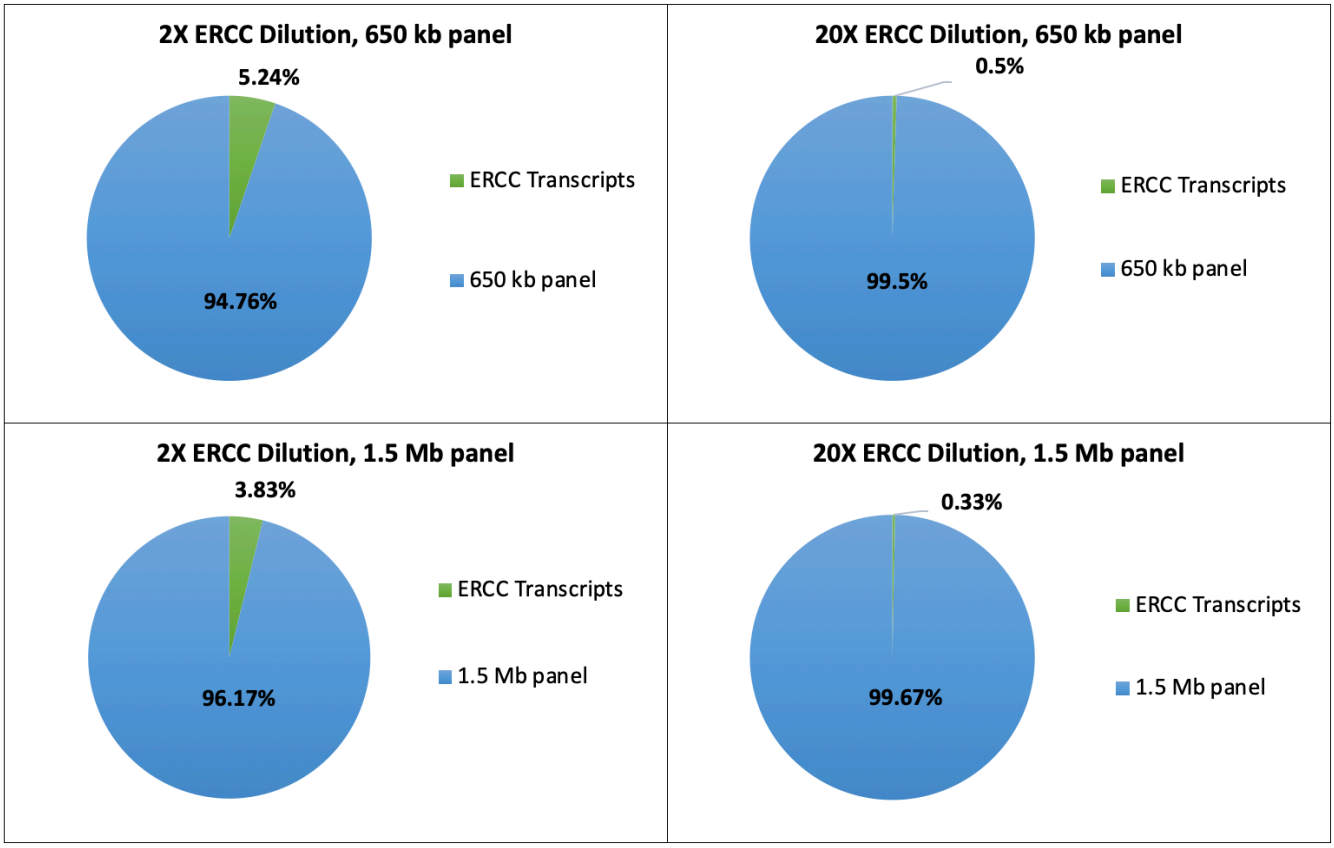

Figure 9: Percentage of sequencing reads consumed by ERCC transcripts at 2X and 20X dilutions.

Captured with 1.5 Mb and 650 kb panels. For each pie chart, green area represents the percentage of sequencing reads consumed by ERCC transcripts. Blue area represents percentage of sequencing reads consumed by all other targets in the two panels tested.

In alignment with previous results, as ERCC transcript concentration decreases so does sequencing reads consumed and number of transcripts detected (Table 6). Figure 9 illustrates the consistent linear relationship between dilution factor and sequencing coverage of ERCC transcripts, the dilution 10x higher consumes almost exactly ten times more sequencing reads. Further evidence that smaller panels receive

higher sequencing coverage as a general trend is witnessed when comparing the percentage of reads for ERCC transcripts in the 1.5 Mb panel versus the 650 kb panel.

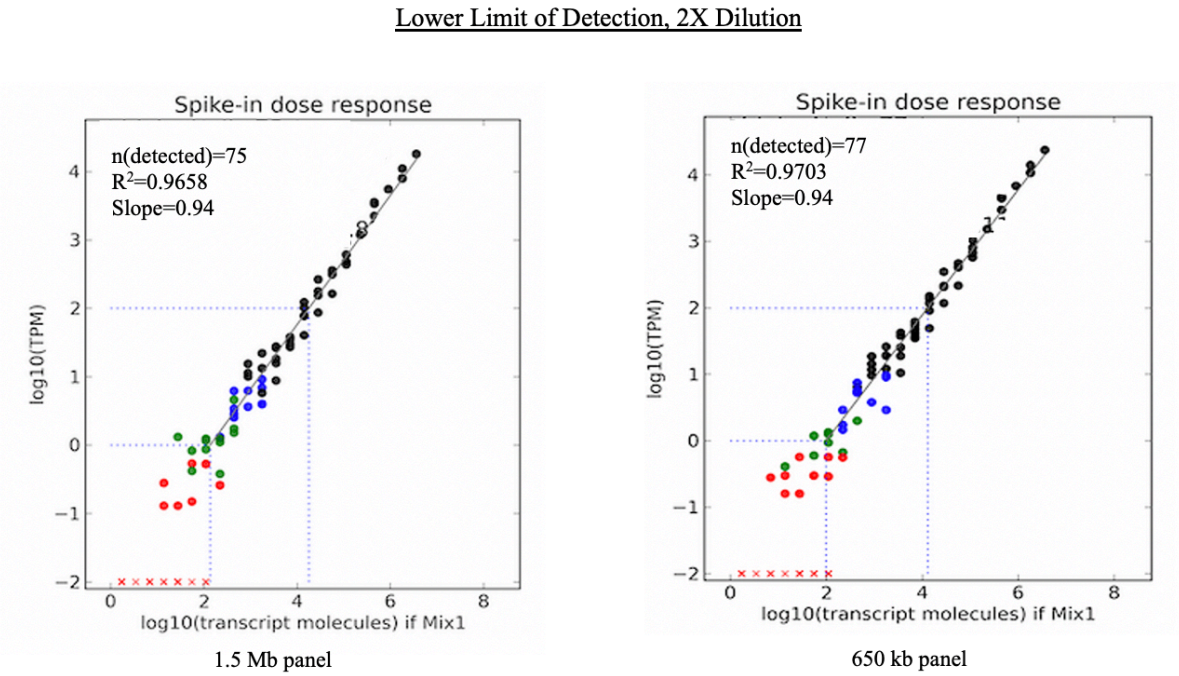

Figure 10: Dose response curves for 2X dilution.

2X ERCC transcript dilution captured with 1.5 Mb panel (right) and 650 kb panel (left). Y-axis corresponds to observed expression levels for each ERCC transcript (TPM), X-axis corresponds to expected number of transcript molecules present in mix. Marker colors correspond to read count per transcript; red = 1-5, green = 6-20, blue = 21-100, black = >100 reads

Similar to the data shown in Figure 5, a smaller panel with increased sequencing coverage is able to detect a higher number of ERCC transcripts when compared to a larger panel at the same ERCC dilution. However, even with increased sequencing reads for the smaller panel, the same low abundance transcripts drop out and are not detected with either panel.

Reproducibility Across ERCC Transcript Dilutions

Reproducibility across ERCC transcript dilutions was investigated by plotting expression levels for two identical sample replicates of the same ERCC transcript dilution factor and captured with the same panel. Interestingly, even at high dilutions the ERCC transcript expression correlation remained consistent.

ERCC Dilution	Panel Size	Number Transcripts Detected	Expression Correlation R^2
1X	650kb	83	0.999
2X	650kb	77	0.997
4X	650kb	67	0.999
10X	650kb	63	0.999
20X	650kb	55	0.994
1X	1.5Mb	79	0.992
2X	1.5Mb	75	0.997
4X	1.5Mb	67	0.995
10X	1.5Mb	62	0.998
20X	1.5Mb	54	0.999

Table 7: Expression correlation.
Measured by comparing replicates of the same ERCC transcript dilution and panel size.

Expression correlation between replicates of each condition was calculated for each dilution factor. R^2 values show high correlation across sample replicates, even at very low concentration of ERCC transcripts. The plots below demonstrate reproducibility even at the least concentrated levels of ERCC transcripts (20X dilution).

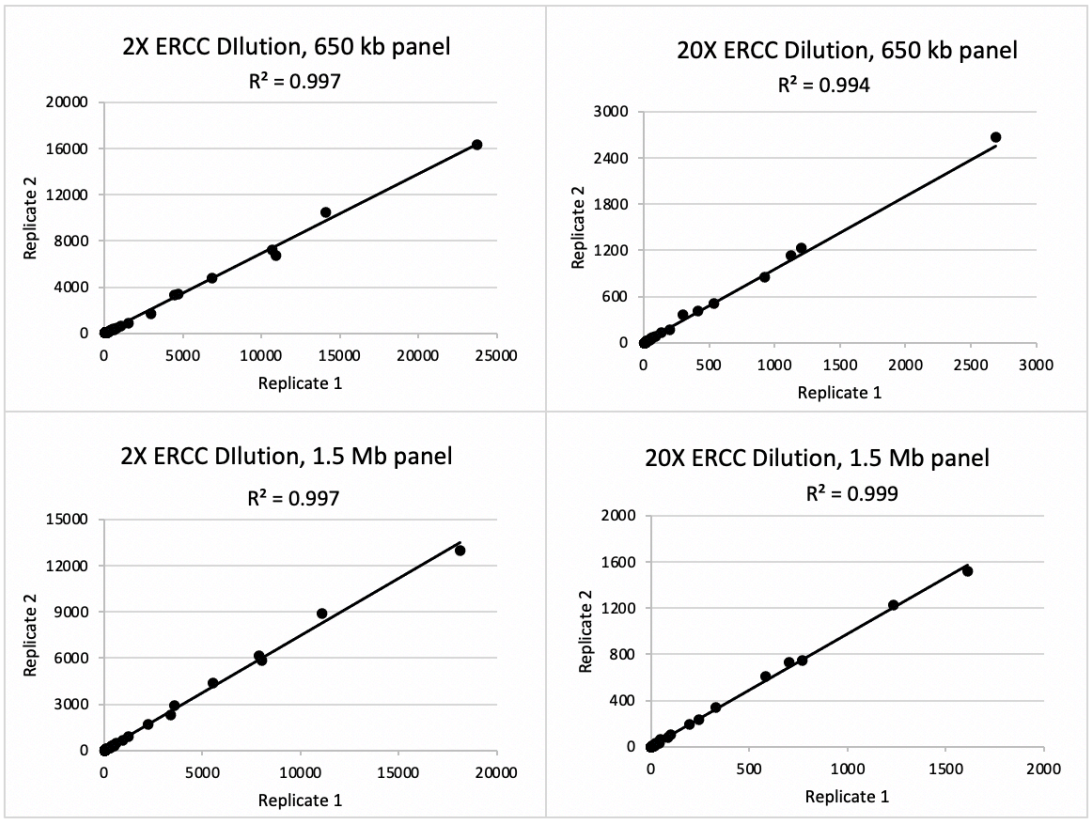

Figure 11: Correlation of ERCC transcript expression between identical sample replicates.

Summary: Applying External RNA Controls in Targeted Hybrid Capture Assay

Initial data indicated that the 0.5X ERCC dilution was optimal for capture with a 30 Mb panel. 85 out of 92 transcripts were detected with the 0.5X dilution compared to 79 out of 92 transcripts detected with the 1X dilution. This dilution consumed <1% of total sequencing reads, leaving the vast majority of sequencing reads allocated to other targets captured in the panel. This distribution of sequencing reads ensures that the regions of the genome targeted by the capture panel receive enough coverage without

costly over sequencing to outcompete the coverage of ERCC transcripts. Adequate sequencing coverage is critical to successful target enrichment.

When the 0.5X ERCC dilution was subsequently tested with a smaller capture panel (1.5 Mb), 91 out of 92 transcripts were detected. However, the ERCC transcripts consumed approximately 40% of sequencing reads, leaving slightly more than half of total sequencing reads to all other targets in the smaller panel. This distribution is far from ideal for achieving adequate sequencing coverage of the targets in the capture panel.

To gain a better understanding of how panel size and ERCC transcript concentration affects read consumption in a capture based assay, a titration of ERCC transcript dilutions was performed. These dilutions were tested using a 650 kb panel and a 1.5 Mb panel. In general, the amount of sequencing coverage and number of transcripts detected decreased as dilution factor increased. Expression correlation within sample replicates was assessed to gauge reproducibility, the results were consistent even across the higher dilutions in which the ERCC transcripts were much less concentrated.

Chapter IV.

Discussion

The global DNA sequencing market has grown rapidly and is expected to reach 26 million dollars by 2025. Technological advances in the past quarter century have enabled DNA sequencing methods and applications to expand exponentially, to the benefit of researchers and clinicians alike. As sequencing technology has become more sophisticated and efficient the need for a standardized in process control has become evident. These process control samples are often used to mimic a biological sample, serving as a quality control and reliable reference to ensure a batch of patient samples was prepared for sequencing correctly and without any confounding artifacts. Ultimately, in-process controls ensure the sequencing data generated for all samples in a given subset are the best quality possible and of high integrity. This allows for confidence in data analysis, reporting, and clinical implications.

Process match control samples have many different applications depending on the type and purpose of a given assay. Both DNA and RNA process controls are commonly used, as well as biological and synthetic source material. Ideally a control is used as a standardized reference data point when assessing a batch of samples, meaning that the less variability in control performance the better. In some cases, a control sample from a biological source is used in conjunction with synthetic spike in material, usually to determine that an assay as a whole is meeting performance criteria (e.g. synthetic spike ins for contamination detection). This type of control is also useful when measuring data that by nature tends to have high variability in outcome. For example, RNA expression data varies widely depending on a multitude of factors. To mitigate this issue a common

set of external RNA controls was developed by a consortium of organizations, hosted by the National Institute of Standards and Technology (NIST). These controls were developed to evaluate different platforms such as real-time qPCR, next generation RNA sequencing platforms, microarray systems, and can even be used as an in-process control for quality checking library amplification processes (Lemire et al., 2011). One of the earlier studies that investigated the use of ERCC spike in transcripts identified a wide range of successful measurements but also highlighted a main shortfall of RNA seq in this application. High sequencing coverage of transcripts is crucial for analyzing and understanding data generated with the spike in mixes. The entire length of the transcript must cover by the sequencing reads, which necessitates a large number of sequencing reads for these targets to be covered when sequencing the whole transcriptome (Jiang et al., 2011). This ends up being a costly and time consuming process in order to generate useful data.

The synthetic transcripts are commercially available as ERCC RNA spike in mixes and are commonly used for platform agnostic quantitation in gene expression experiments (Baker et al., 2005; Curion et. al, 2020; Zhao et al., 2014). The RNA spike in mixes can also be used during the development and validation of an RNA based assay to provide standardized and reproducible data for assay sensitivity, limit of detection, dynamic range, and differential gene expression. ERCC spike in transcripts are well suited for characterizing lower limit of detection because each of the 92 transcripts in the mix exists at a known abundance and concentration. Determining the number of molecules or transcripts detected at a given ERCC transcript concentration can provide valuable insight into how sensitive an assay would perform at detecting very low levels

of a molecule.

This study investigates the feasibility of using synthetic ERCC spike in transcript mixes with targeted panels in a hybrid capture assay, an 'off label' application given that manufacturer guidelines only recommend using these mixes in RNA seq or microarray assays. Using synthetic transcripts, as opposed to biological controls such as housekeeping genes, provides a reliable reference for precise measurements of sensitivity and reproducibility. Synthetic RNA also facilitates easier detection of contamination and decreases the chance of misalignment when the fragments are read by a sequencer.

The experiments to address the main goal of this study started with a broad, more generalized experiment and translated these results to a smaller focused panel. The optimal ERCC dilution for use with a 30 Mb hybrid capture panel was determined by testing different ERCC dilutions and comparing number of transcripts detected versus percent of sequencing reads consumed. It was hypothesized that the ERCC dilution optimized for a 30 Mb panel capture would be too concentrated for a 1.5 Mb panel capture. The ERCCs encompass a much smaller region compared to a 30 Mb panel versus a 1.5 Mb panel. This was confirmed by demonstrating that the 0.5X dilution of ERCC transcripts were taking up far too many sequencing reads, resulting in less coverage of other important targets in the panel. The smaller capture panels used in this study targeted regions that are difficult to detect at low sequencing coverage. Given this sensitivity, minimizing the amount of reads consumed by ERCC transcripts is even more important, and target dropouts are a potential concern.

The results of the initial experiments in this study provided insight into an approach to better understand the relationship between ERCC spike in concentration,

panel size, and sequencing coverage when using these spike-in mixes in a hybrid capture assay. The subsequent titration of ERCC spike in dilutions aimed to find a satisfactory compromise between number of transcripts detected and amount of sequencing reads consumed by these transcripts. The corresponding data confirmed an inverse relationship between ERCC dilution factor and number of transcripts detected. At higher ERCC spike in dilution factor (i.e. lower concentration), fewer sequencing reads were consumed by the transcripts. As expected, the more dilute ERCC spike in mixes took up less reads, but at the cost of linearity and sensitivity to detect low abundance ERCC transcripts. Dropout of the same lowest abundance transcripts has been historically demonstrated in previous studies using RNA seq for transcript detection (Jiang et al., 2011).

For the application of ERCC spike in mixes, the lower limit of detection (LLoD) is generally described as the lowest analyte concentration that can be reliably detected against the limit of blank at which detection is feasible. The LLoD is a measure of sensitivity as determined by the lowest molar amount of ERCC transcript able to be detected in each sample. Given the nature of NGS data and because the platform does not exhibit a region of saturation, the dynamic range of the ERCC transcript concentrations can be determined using the concentration difference between the highest concentration transcript detected and the lower limit of detection.

In the dose response curves generated for each dilution (Figures 5, 7, and 10), the x-axis value where the regression line crosses the threshold represents the concentration value corresponding to lower limit of detection. This study tested ERCC RNA spike in mixes at varying dilution factors with the hypothesis that decrease in ERCC transcript concentration will have a detrimental impact on limit of detection, or sensitivity to detect

transcripts at low molecular abundance. In general, and per the manufacturer guidelines, the values for determining detection limit are defined by the user. This threshold may be determined empirically or arbitrarily based on other studies, primary literature, etc. For example, differential expression analysis uses high stringency thresholds to increase the accuracy of expression calls. Alternatively, lower stringency thresholds (e.g. 1 mapped read) can be used to increase sensitivity of fusion detection and splicing discovery. In the dose response curves generated for this study, the markers representing each transcript were different colors based on the read count per transcript. Red represents the lowest read count (less than 5 reads), black represents transcripts with over 100 reads. This is a visual reminder that more noise in the scatterplot is expected at the lower end of the distribution due to random integer counting differences, i.e. less reads, more variation.

The expression correlation between replicate samples of the same ERCC dilution remained high even as the transcript concentration decreased, as demonstrated by Figure 11. This indicates the results are reproducible across a wide range of ERCC transcript concentration. Other studies using ERCC spike in transcripts with different technologies have tested performance of ERCC transcripts across different sample types. One study found that percentage of reads consumed by ERCC transcripts in low quality (e.g. FFPE) samples was much higher than reads consumed in higher quality cell lines. Interestingly, the read consumption from internal expression controls was comparable across all samples, regardless of quality (Reeser et al., 2017). Given that our study used high quality RNA only, follow up work from the results presented in this study could investigate how the findings translate across different sample types.

ERCC spike in mixes were designed for use in microarray technologies and whole

transcriptome RNA sequencing. When comparing the two assays, both have benefits and limitations depending on the intended application of the data generated. For example, previous work has shown that RNA seq and microarray based models perform similarly in prediction of clinical endpoints (Zhang et al., 2015) but RNA seq enables profiling of the whole transcriptome, whereas microarrays can only profile predefined targets. In general, RNA Seq provides more robust and powerful data than microarray assays. Confident measurements of accuracy, reproducibility, detection limits and dynamic range are crucial for generating usable data for both platforms and can be confirmed by using ERCC spike in mixes (Jiang et al., 2011).

Several studies have used ERCC spike in transcripts as a means to compare gene expression measurements across RNA seq and microarray technologies (Devonshire et al., 2010; Kamel et al., 2017; Zhang et al., 2015). This work has demonstrated that RNA seq provides an incomplete characterization of the transcriptome, especially for low expressed genes and transcripts. This highlights a need for a more targeted approach to sequence low expressed genes and transcripts of interests at higher sequencing coverage, while maintaining linearity of the assay. Very recent studies have explored a more targeted approach to RNA sequencing via hybridization to an array and found a 250-fold enrichment of target genes as well as detection of 10% more additional genes (Curion et al., 2020). ERCC transcript detection also improved by over 60-fold compare to standard whole transcriptome RNA seq.

Using ERCC spike in transcripts to measure sensitivity, specificity, and reproducibility in a targeted RNA hybrid capture workflow, as described by the work of this study, is a less common practice. Our approach to limit amount of sequencing reads

consumed by ERCC transcripts was achieved by titrating and optimizing the amount of ERCC transcripts added to the reaction. Limiting read consumption became even more important as the size of targeted panel used became smaller. Other methods to limit read consumption include only targeting a small percentage of the 92 ERCC transcripts instead of all transcripts present in the mix (Reeser et al., 2017).

Previous studies have used ERCC transcripts to compare the sensitivity and range of microarray versus RNA seq assays. One study demonstrated that the microarray platform has higher sensitivity at the lower end of the dynamic range compared to RNA seq data, resulting in higher dynamic range for microarray compared to RNA seq when averaging over 100M reads per sample (Munro et al., 2014). The data presented in this study were generated by capping sequencing reads at 30 million read pairs for each sample replicate to normalize and reduce bias by outliers with high coverage. The same study also compared sensitivity by taking into account both limit of detection rate (LODR) and false discovery rate (FDR). By using these parameters, the researchers demonstrated an ability to detect small changes in ERCC transcript spike in concentration between both universal human reference RNA and human brain reference RNA, two widely characterized controls for measuring gene expression. Similarly, our study analyzed limit of detection as the ability to detect ERCC transcripts at very low transcript abundance or molecular concentration across varying size of targeted panels.

Comparisons of microarrays versus RNA seq have highlighted the difference between nominal and normalized ERCC ratios and variability in measurement of low expressed genes (Pine et al., 2016; Uygun et al., 2016; Wang et al., 2009). This work demonstrates that both platforms perform efficiently when measuring genes expressed at

high and average levels, while gene expression microarrays provide more stable and precise measurements of low expressed genes compared to RNA seq. The results presented in this study demonstrate a method of using NGS technology to sequence target regions of RNA to capture and measure low expressed genes with high specificity and sensitivity. Targeted sequencing of RNA is proven to be much more efficient than the commonly used method of whole transcriptome sequencing. The dynamic range of the whole transcriptome presents many challenges such as difficulty discovering rare transcripts. The most abundant or highly expressed RNAs in the transcriptome will consume the majority of sequencing reads, leaving minimal coverage for low expressors or other regions of interest (Jiang et al., 2011).

As highlighted above, the use of ERCC transcripts for detection and sensitivity measurements translates to a range of applications beyond just RNA Seq and microarrays. It remains unclear if other types of spike in material would also translate linearly to other applications such as capture based assays. A newer product, referred to as Spike-In RNA variants (SIRVs) is gaining popularity for characterization of RNA isoforms in RNA seq workflows. These synthetic spike ins are added to reference RNA samples to provide molecular information about variants, splicing, and gene fusions (Paul et al., 2016). Future work could test the application of this type of spike-in with a capture based assay.

Using ERCC spike in mixes at different concentrations with different sizes of capture panel generated the most meaningful data not just for the goal of this study, but also has exciting implications in other assays or future applications. If the relationship between read consumption and ERCC transcript concentration is linear and reproducible

as indicated by this study, other targeted panel hybrid capture assays could save time and money by using this as a guideline for approximate dilution of ERCC transcripts, avoiding a costly titration of several different dilutions.

The results from this study highlight the tradeoff between amount of sequencing reads (i.e. coverage) versus ability to detect molecules at low abundance (i.e. sensitivity). This limiting factor is almost universal across platforms that use DNA sequencing technology to generate insightful data at the molecular level. At higher ERCC transcript concentrations sensitivity to detect targets in smaller capture panels is compromised. If future work or follow up experiments investigate a way to mitigate this tradeoff, the application would be universal and widely beneficial for targeted sequencing with custom panels.

References

Armbruster, D. A., & Pry, T. (2008). Limit of blank, limit of detection and limit of quantitation. *The clinical biochemist reviews*, *29*(Suppl 1), S49.

Ansorge, W. J. (2009). Next-generation DNA sequencing techniques. *New biotechnology*, *25*(4), 195-203.

Baker, S. C., Bauer, S. R., Beyer, R. P., Brenton, J. D., Bromley, B., Burrill, J., ... & Foy, C. (2005). The external RNA controls consortium: a progress report. *Nature methods*, *2*(10), 731.

Bell, J. T., Pai, A. A., Pickrell, J. K., Gaffney, D. J., Pique-Regi, R., Degner, J. F., ... & Pritchard, J. K. (2011). DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. *Genome biology*, *12*(1), R10.

Blackburn, J., Wong, T., Madala, B. S., Barker, C., Hardwick, S. A., Reis, A. L., ... & Mercer, T. R. (2019). Use of synthetic DNA spike-in controls (sequins) for human genome sequencing. *Nature Protocols*, *14*(7), 2119.

Borodina, T., Adjaye, J., & Sultan, M. (2011). A strand-specific library preparation protocol for RNA sequencing. In *Methods in enzymology* (Vol. 500, pp. 79-98). Academic Press.

Buchanan, C. C., Torstenson, E. S., Bush, W. S., & Ritchie, M. D. (2012). A comparison of cataloged variation between International HapMap Consortium and 1000 Genomes Project data. *Journal of the American Medical Informatics Association*, *19*(2), 289-294.

Cabili, M.N. et al. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev.* 25, 1915–1927 (2011).

Chen, G., Li, R., Shi, L., Qi, J., Hu, P., Luo, J., ... & Shi, T. (2011). Revealing the missing expressed genes beyond the human reference genome by RNA-Seq. *BMC genomics*, *12*(1), 590.

Chen, K., Hu, Z., Xia, Z., Zhao, D., Li, W., & Tyler, J. K. (2016). The overlooked fact: fundamental need for spike-in control for virtually all genome-wide analyses. *Molecular and cellular biology*, *36*(5), 662-667.

Ciriello, G., Miller, M. L., Aksoy, B. A., Senbabaoglu, Y., Schultz, N., & Sander, C. (2013). Emerging landscape of oncogenic signatures across human cancers. *Nature genetics*, *45*(10), 1127.

Collins, F. S., Morgan, M., & Patrinos, A. (2003). The Human Genome Project: lessons

from large-scale biology. *Science*, *300*(5617), 286-290.

Curion, F., Handel, A. E., Attar, M., Gallone, G., Bowden, R., Cader, M. Z., & Clark, M. B. (2020). Targeted RNA sequencing enhances gene expression profiling of ultra-low input samples. *RNA biology*, 1-13.

DeAngelis, M. M., Wang, D. G., & Hawkins, T. L. (1995). Solid-phase reversible immobilization for the isolation of PCR products. *Nucleic acids research*, *23*(22), 4742.

Desai A.N., Jere A. (2015) Next-Generation Sequencing for Cancer Biomarker Discovery. In: Wu W., Choudhry H. *Next Generation Sequencing in Cancer Research*, (2). Springer, Cham

Devonshire, A. S., Elaswarapu, R., & Foy, C. A. (2010). Evaluation of external RNA controls for the standardisation of gene expression biomarker measurements. *BMC genomics*, *11*(1), 662.

Devonshire, A. S., Elaswarapu, R., & Foy, C. A. (2011). Applicability of RNA standards for evaluating RT-qPCR assays and platforms. *BMC genomics*, *12*(1), 118.

Farias-Hesson, E., Erikson, J., Atkins, A., Shen, P., Davis, R. W., Scharfe, C., & Pourmand, N. (2010). Semi-automated library preparation for high-throughput DNA sequencing platforms. *BioMed Research International*, *2010*.

Fisher, S., Barry, A., Abreu, J., Minie, B., Nolan, J., Delorey, T. M., ... & Berlin, A. M. (2011). A scalable, fully automated process for construction of sequence-ready human exome targeted capture libraries. *Genome biology*, *12*(1), R1.

Fuller, C. W., Middendorf, L. R., Benner, S. A., Church, G. M., Harris, T., Huang, X., ... & Vezenov, D. V. (2009). The challenges of sequencing by synthesis. *Nature biotechnology*, *27*(11), 1013.

Garalde, D. R., Snell, E. A., Jachimowicz, D., Sipos, B., Lloyd, J. H., Bruce, M., ... & Jordan, M. (2018). Highly parallel direct RNA sequencing on an array of nanopores. *Nature methods*, *15*(3), 201.

Gibbs, R. A., Belmont, J. W., Hardenbol, P., Willis, T. D., Yu, F. L., Yang, H. M., ... & Tam, P. K. H. (2003). The international HapMap project.

Gnirke, A., Melnikov, A., Maguire, J., Rogov, P., LeProust, E. M., Brockman, W., ... & Gabriel, S. (2009). Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nature biotechnology*, *27*(2), 182.

Goh, G., & Choi, M. (2012). Application of whole exome sequencing to identify disease-causing variants in inherited human diseases. *Genomics & informatics*, *10*(4), 214.

Grada, A., & Weinbrecht, K. (2013). Next-generation sequencing: methodology and application. *The Journal of investigative dermatology*, *133*(8), e11.

Griffith, M. et al. Alternative expression analysis by RNA sequencing. *Nat. Methods* 7, 843–847 (2010).

Gusev, A., Lowe, J. K., Stoffel, M., Daly, M. J., Altshuler, D., Breslow, J. L., ... & Pe'er, I. (2009). Whole population, genome-wide mapping of hidden relatedness. *Genome research*, *19*(2), 318-326.

Haas, B.J., Dobin, A., Li, B. *et al.* Accuracy assessment of fusion transcript detection via read-mapping and de novo fusion transcript assembly-based methods. *Genome Biol* 20, 213 (2019).

Head, S. R., Komori, H. K., LaMere, S. A., Whisenant, T., Van Nieuwerburgh, F., Salomon, D. R., & Ordoukhanian, P. (2014). Library construction for next-generation sequencing: overviews and challenges. *Biotechniques*, *56*(2), 61-77.

Heyer, E.E., Deveson, I.W., Wooi, D. *et al.* Diagnosis of fusion genes using targeted RNA sequencing. *Nat Commun* 10, 1388 (2019). https://doi.org/10.1038/s41467-019-09374-9

Holmberg, R. C., Gindlesperger, A., Stokes, T., Brady, D., Thakore, N., Belgrader, P., ... & Chandler, D. P. (2013). High-throughput, automated extraction of DNA and RNA from clinical samples using TruTip technology on common liquid handling robots. *JoVE (Journal of Visualized Experiments)*, (76), e50356.

International HapMap Consortium. (2003). The international HapMap project. *Nature*, *426*(6968), 789.

Islam, S., Zeisel, A., Joost, S., La Manno, G., Zajac, P., Kasper, M., ... & Linnarsson, S. (2014). Quantitative single-cell RNA-seq with unique molecular identifiers. *Nature methods*, *11*(2), 163.

Jain, M., Olsen, H. E., Paten, B., & Akeson, M. (2016). The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome biology*, *17*(1), 239.

Jiang, H. & Wong, W.H. Statistical inferences for isoform expression in RNA-seq. *Bioinformatics* 25, 1026–1032 (2009).

Jiang, L., Schlesinger, F., Davis, C. A., Zhang, Y., Li, R., Salit, M., ... & Oliver, B. (2011). Synthetic spike-in standards for RNA-seq experiments. *Genome research*, *21*(9), 1543-1551.

Jones, M. E. (1953). Albrecht Kossel, a biographical sketch. The Yale journal of biology and medicine, 26(1), 80.

Kamel, H. F. M., & Al-Amodi, H. S. A. B. (2017). Exploitation of gene expression and cancer biomarkers in paving the path to era of personalized medicine. *Genomics, proteomics & bioinformatics*, *15*(4), 220-235.

Krishnakumar, S., Zheng, J., Wilhelmy, J., Faham, M., Mindrinos, M., & Davis, R. (2008). A comprehensive assay for targeted multiplex amplification of human DNA sequences. *Proceedings of the National Academy of Sciences*, *105*(27), 9296-9301.

Kumar, R., Ichihashi, Y., Kimura, S., Chitwood, D. H., Headland, L. R., Peng, J., ... & Sinha, N. R. (2012). A high-throughput method for Illumina RNA-Seq library preparation. *Frontiers in plant science*, *3*, 202.

Kumar, S., Vo, A., Qin, F. *et al.* Comparative assessment of methods for the fusion transcripts detection from RNA-Seq data. *Sci Rep* 6, 21597 (2016).

Lemire, A., Lea, K., Batten, D., Gu, S. J., Whitley, P., Bramlett, K., & Qu, L. (2011). Development of ERCC RNA spike-in control mixes. *Journal of biomolecular techniques: JBT*, *22*(Suppl), S46.

Lesko, L. J., & Atkinson Jr, A. J. (2001). Use of biomarkers and surrogate endpoints in drug development and regulatory decision making: criteria, validation, strategies. *Annual review of pharmacology and toxicology*, *41*(1), 347-366.

Li, B., Ruotti, V., Stewart, R.M., Thomson, J.A. & Dewey, C.N. RNA-seq gene expression estimation with read mapping uncertainty. *Bioinformatics* 26, 493–500 (2010).

Linnarsson, S. (2010). Recent advances in DNA sequencing methods–general principles of sample preparation. *Experimental cell research*, *316*(8), 1339-1343.

Liu, T., Johnson, J. A., Casella, G., & Wu, R. (2004). Sequencing complex diseases with HapMap. *Genetics*, *168*(1), 503-511.

Loman, N. J., Misra, R. V., Dallman, T. J., Constantinidou, C., Gharbia, S. E., Wain, J., & Pallen, M. J. (2012). Performance comparison of benchtop high-throughput sequencing platforms. *Nature biotechnology*, *30*(5), 434-439.

Luca, D., Ringquist, S., Klei, L., Lee, A. B., Gieger, C., Wichmann, H. E., ... & Devlin, B. (2008). On the use of general control samples for genome-wide association studies: genetic matching highlights causal variants. *The American Journal of Human Genetics*, *82*(2), 453-463.

Lundin, S., Stranneheim, H., Pettersson, E., Klevebring, D., & Lundeberg, J. (2010). Increased throughput by parallelization of library preparation for massive sequencing. PloS one, 5(4), e10029.

Maddox, B. (2003). The double helix and the'wronged heroine'. *Nature*, *421*(6921), 407-408.

Manolio, T. A., Brooks, L. D., & Collins, F. S. (2008). A HapMap harvest of insights into the genetics of common disease. *The Journal of clinical investigation*, *118*(5), 1590-1605.

Mardis, E. R. (2008). Next-generation DNA sequencing methods. *Annu. Rev. Genomics Hum. Genet.*, *9*, 387-402.

Mardis, E.R. The impact of next-generation sequencing technology on genetics. *Trends Genet.* 24, 133–141 (2008).

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., ... & DePristo, M. A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research*, *20*(9), 1297-1303.

McVean, G., Spencer, C. C., & Chaix, R. (2005). Perspectives on human genetic variation from the HapMap Project. *PLoS genetics*, *1*(4).

Men, A. E., Wilson, P., Siemering, K., & Forrest, S. (2008). Sanger DNA sequencing. *Next Generation Genome Sequencing: Towards Personalized Medicine*, 1-11.

Meyer, M., & Kircher, M. (2010). Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harbor Protocols*, *2010*(6), pdb-prot5448.

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-seq. *Nat. Methods* 5, 621–628 (2008).

Munro, S. A., Lund, S. P., Pine, P. S., Binder, H., Clevert, D. A., Conesa, A., ... & Jafari, N. (2014). Assessing technical performance in differential gene expression experiments with external spike-in RNA control ratio mixtures. *Nature communications*, *5*(1), 1-10.

Novoradovskaya, N., Chin, N., Payette, T., Pergamenschikov, A., Fero, M., Botstein, D., & Braman, J. (2001). Using universal human reference RNA in microarray gene expression studies. *Nature Genetics*, *27*(4), 76-76.

Novoradovskaya, N., Whitfield, M. L., Basehore, L. S., Novoradovsky, A., Pesich, R., Usary, J., Karaca, M., Wong, W. K., Aprelikova, O., Fero, M., Perou, C. M., Botstein, D., & Braman, J. (2004). Universal Reference RNA as a standard for microarray

experiments. *BMC genomics*, *5*(1), 20.

Park, Y. S., Kim, S., Park, D. G., Kim, D. H., Yoon, K. W., Shin, W., & Han, K. (2019). Comparison of library construction kits for mRNA sequencing in the Illumina platform. *Genes & genomics*, 1-8.

Paul, L., Kubala, P., Horner, G., Ante, M., Holländer, I., Alexander, S., & Reda, T. (2016). SIRVs: Spike-In RNA Variants as external isoform controls in RNA-sequencing. *bioRxiv*, 080747.

Pickrell, J. K., Pai, A. A., Gilad, Y., & Pritchard, J. K. (2010). Noisy splicing drives mRNA isoform diversity in human cells. *PLoS genetics*, *6*(12).

Pine, P.S., Munro, S.A., Parsons, J.R. *et al.* Evaluation of the External RNA Controls Consortium (ERCC) reference material using a modified Latin square design. *BMC Biotechnol* 16, 54 (2016).

Quail, M. A., Otto, T. D., Gu, Y., Harris, S. R., Skelly, T. F., McQuillan, J. A., ... & Oyola, S. O. (2012). Optimal enzymes for amplifying sequencing libraries. *Nature methods*, *9*(1), 10-11.

Rizzo, J. M., & Buck, M. J. (2012). Key principles and clinical applications of "next-generation" DNA sequencing. *Cancer prevention research*, *5*(7), 887-900.

Reeser, J. W., Martin, D., Miya, J., Kautto, E. A., Lyon, E., Zhu, E., ... & Parks, H. (2017). Validation of a targeted RNA sequencing assay for kinase fusion detection in solid tumors. *The Journal of Molecular Diagnostics*, *19*(5), 682-696.

Rohland, N., & Reich, D. (2012). Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. *Genome research*, *22*(5), 939-946.

Rothberg, J. M., & Leamon, J. H. (2008). The development and impact of 454 sequencing. *Nature biotechnology*, *26*(10), 1117-1124.

Schumacher, T. N., & Schreiber, R. D. (2015). Neoantigens in cancer immunotherapy. *Science*, *348*(6230), 69-74.

Shendure, J. A., Porreca, G. J., Church, G. M., Gardner, A. F., Hendrickson, C. L., Kieleczawa, J., & Slatko, B. E. (2011). Overview of DNA sequencing strategies. *Current Protocols in Molecular Biology*, *96*(1), 7-1.

Shigemizu, D., Momozawa, Y., Abe, T., Morizono, T., Boroevich, K. A., Takata, S., ... & Tsunoda, T. (2015). Performance comparison of four commercial human whole-exome capture platforms. *Scientific reports*, *5*, 12742.

Singer, V. L., Jones, L. J., Yue, S. T., & Haugland, R. P. (1997). Characterization of PicoGreen reagent and development of a fluorescence-based solution assay for double-stranded DNA quantitation. *Analytical biochemistry*, *249*(2), 228-238.

Stranneheim, H., Werne, B., Sherwood, E., & Lundeberg, J. (2011). Scalable transcriptome preparation for massive parallel sequencing. *PloS one*, *6*(7), e21910.

Szalat, R., Avet-Loiseau, H., & Munshi, N. C. (2016). Gene expression profile in clinical practice. *Clinical cancer research: an official journal of the American Association for Cancer Research*, *22*(22), 5434.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., ... & Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature protocols*, *7*(3), 562.

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., Van Baren, M. J., ... & Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology*, *28*(5), 511.

Uygun, S., Peng, C., Lehti-Shiu, M. D., Last, R. L., & Shiu, S. H. (2016). Utility and limitations of using gene expression data to identify functional associations. *PLoS computational biology*, *12*(12), e1005244.

Van Dijk, E. L., Jaszczyszyn, Y., & Thermes, C. (2014). Library preparation methods for next-generation sequencing: tone down the bias. *Experimental cell research*, *322*(1), 12-20.

Van't Veer, L. J., Dai, H., Van De Vijver, M. J., He, Y. D., Hart, A. A., Mao, M., ... & Schreiber, G. J. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *nature*, *415*(6871), 530.

Várnai, P., & Zakrzewska, K. (2004). DNA and its counterions: a molecular dynamics study. *Nucleic acids research*, *32*(14), 4269-4280.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., ... & Gocayne, J. D. (2001). The sequence of the human genome. *science*, *291*(5507), 1304-1351.

Vu, T., Deng, W., Trac, Q. *et al.* A fast detection of fusion genes from paired-end RNA-seq data. *BMC Genomics* 19, 786 (2018).

Wang, L., Si, Y., Dedow, L. K., Shao, Y., Liu, P., and Brutnell, T. P. (2011a). A low-cost library construction protocol and data analysis pipeline for illumina-based strand-specific multiplex RNA-seq. *PLoS ONE* 6, e26426.

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews genetics*, *10*(1), 57.

Weber, J. L., & Myers, E. W. (1997). Human whole-genome shotgun sequencing. *Genome research*, *7*(5), 401-409.

Zhang, W., Yu, Y., Hertwig, F., Thierry-Mieg, J., Zhang, W., Thierry-Mieg, D., ... & Deng, Y. (2015). Comparison of RNA-seq and microarray-based models for clinical endpoint prediction. *Genome biology*, *16*(1), 1-12.

Zhao, W., He, X., Hoadley, K. A., Parker, J. S., Hayes, D. N., & Perou, C. M. (2014). Comparison of RNA-Seq by poly (A) capture, ribosomal RNA depletion, and DNA microarray for expression profiling. *BMC genomics*, *15*(1), 1-11.

User Guides:

Ambion® by *Life Technologies*™. ERCC ExFold RNA Spike-In Mixes. Catalog Number 4456740, 4456739. Publication Number 4455352. Revision D.

Ambion® RNA Fragmentation Reagents. Catalog #AM8740.

Zymo Research. RNA Clean & Concentrator™-25. Catalog No. R1017, R1018. Version 2.0.7.