illumina®

Considerations for Designing a Successful TruSeq® Targeted RNA Expression Experiment

Introduction

TruSeq Targeted RNA Expression (TREx) assays offer MiSeq® owners an entirely new way to profile gene expression. With TREx, you can obtain very accurate and reproducible data across a large dynamic range, allowing study of up to 15,000 gene expression results using 50 bp sequencing reads to target specific exons, splice junctions, cSNPs, and gene fusions. Using DesignStudio, you can simply upload a list of targets identified by RNA-Seq or expression array experiments and choose from over 400,000 pre-designed assays based on RefSeq gene models in human, rat, and mouse. Quickly build out your custom panel of up to 1,000 assays, add new targets to a previously ordered custom panel, or choose one of many fixed panels.

The purpose of this technical note is to explain the concept of working with a fixed number of sequencing reads, or a "read budget", when designing experiments measuring very high or low expressing genes. We will also address the concepts of normalization, multiplexing, and sample pooling in order for you to design the best possible TruSeq Targeted RNA Expression experiments.

Read Budget: Definition

The concept of working within a read budget is not technically daunting. The current output of the MiSeq system is ~15 million total sequencing reads[1], and this number, or read budget, is shared by all the assays in the study. Assays are defined as the probeset measuring the expression of a single region of interest, or targeted part of a gene. If each assay uses up 1,000 reads on average, then each MiSeq run can support 15,000 data points. TREx sample prep kits allow 12-1,000 assays per sample and 1–384 samples per MiSeq run. A common way of reaching this number of data points would be to run 96 indexed samples with 150 multiplexed assays in each sample. This is roughly equivalent to running 15,000 qPCR reactions, or 40 individual 384-well plates. To use your read budget efficiently, a good practice is to include genes that are low to moderately high expressors. For example, if you have RNA-Seq data, these would have expression values of 1,000 reads per kilobase per million mapped reads (RPKM) or lower. Alternatively, you can use data in the Illumina Human Body Map 2.0 Project[2] to determine if your genes of interest are high expressors. The main concept of budgeting reads is to maximize your study by sharing reads across all assays and not allowing it to be dominated by a few very highly expressed genes.

However, experimental evidence has shown that even including an assay targeting a very highly expressed transcript still results in a good correlation of reads compared to running the panel without the transcript (Figure 1). In other words, the TREx amplification process is robust, and both low and high expressed genes are amplified quantitatively.

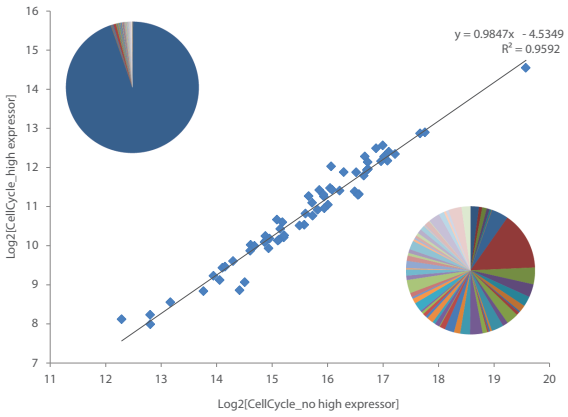
Read Budget and Read Depth

A TruSeq Targeted RNA Panel consisting of 12 assays designed against target transcripts in Universal Human Reference RNA (UHR) was tested at different plexity, to measure the effect of two different reads/assay ratios (Figure 2). The panel was designed to cover a wide dynamic range of expression (3.9 Log10 units in this experiment) and consisted of 3 groups of 4 assays targeting low, medium, or high expression targets. mRNA sequencing of UHR RNA put the selected targets in the <1, ~30, and ~1000 fragments per kilobase of transcript per million mapped reads (FPKM) range, respectively.

At high read depth (greater than 20,000 reads/assay), all assays targeting medium and high expression targets performed well, with only 1 out of 4 assays targeting low expression transcripts producing average raw counts <10 (equivalent to -2.27 Log2, Figure 2). This dropout corresponds to an expressor of very low abundance, 0.13 FPKMs by mRNA sequencing, and would only be measurable at higher read depth or in an experiment were dynamic range is reduced.

At lower read depth (less than 1,000 reads/assay), all assays targeting both medium and high expressors still performed well. The effect of increasing plexity is only evident with low expression targets. In this case, 3 out of 4 assays targeting the low expression transcripts fall below the threshold of 10 raw counts

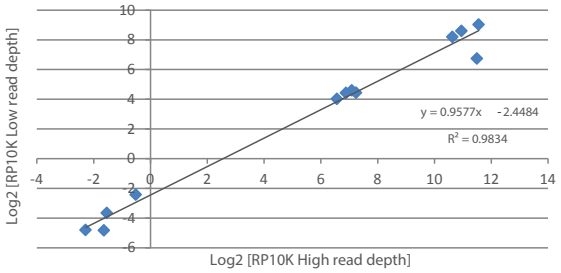
Figure 1: Correlation with the Addition of a High Expressor



Effect when a panel is supplemented with a highly expressed transcript, which takes up a majority of the reads (dark blue, top pie chart) compared to the same experiment that does not include this transcript (bottom pie chart), showing that there is good correlation of both experiments.

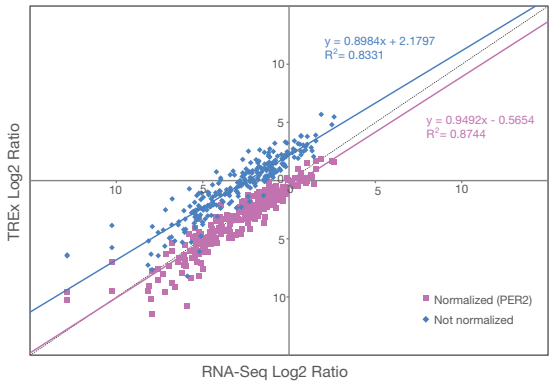
AGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAAC
TCAACGTACCGTAACGAACGTATCATTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGATAACAGTAACACACTTCTGTTAACCTT
CGACGAAAAGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACC
AACGTACCATTAAGAGCTACCGTGCAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGATAA
AGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAAC
GATTACTTGATCCACTGATTCAACGTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAGCAACGACG
CGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGAAAAGAATGATAACAGTA

Figure 2: Effect of Read Depth on Assay Drop-Out



For both high (>20,000 reads/assay) and low (<1,000 reads/assay) read depth, raw data points were normalized by total counts per sample, converted to RP10K units and log2 transformed. At high read depth, only one assay falls below the threshold of 10 reads, while at low read depth, three assays fall below this threshold.

Figure 3: Normalizing with an Invariant Gene



TREx expression ratios normalized to an example of an invariant gene, PER2 (purple) and not normalized (blue) were compared to RNA-Seq expression ratios. Normalization with this gene results in better correlation.

Normalization

Normalization is an essential step in gene expression analysis. In microarray data analysis, normalization adjusts for systematic biases to allow comparisons of expression levels between and within samples. For RNA-Seq, normalization helps correct for gene length, count distribution between samples, and differences in sequencing depth. The most common method for normalizing RNA-Seq data is by total counts, where counts are divided by the transcript length and multiplied by millions of mapped reads[3].

Expression ratio accuracy relies on the use of an internal gene by which the TruSeq Targeted RNA data is normalized by sample. This allows for cross-sample and run comparison by adjusting the data to the counts read in the sample and the internal gene. Good candidates for normalization include genes from RNA-Seq data that are relatively unchanged across samples. As an example. the adjustment of expression ratios using such a gene (PER2) is shown in Figure 3. Invariant expression of genes used for normalization should be verified across all samples being studied.

Conclusions

Although it is a slight departure from other RNA expression methods that routinely include highly expressed housekeeping genes, the read budget concept provides guidelines to maximize study design and take full advantage of MiSeq output. To most efficiently use your read budget for TruSeq Targeted RNA Expression, it is not recommended to use very highly expressed transcripts. A good practice is to include genes that are low to moderate expressors. Assuming an average of 1,000 counts per assay allows for up to 15,000 reactions per run, and sufficient dynamic range to cover most applications.

For experiments requiring a very large dynamic range, decrease the number of samples per run to ensure low expressors are detected. See the TruSeq Targeted RNA Expression User Guide[4] for further sample pooling guidelines.

Internal gene-based normalization is recommended for accurate relative quantification, as long as this gene expression is invariant. This step can be performed manually across all samples. See the MiSeq Reporter User Guide[5] for further instructions on how to normalization by total counts is performed.

References

1. http://www.illumina.com/systems/miseq/performance_specifications.ilmn
2. http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE30611
3. Mortizavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature Methods 5;621–28.
4. http://support.illumina.com/downloads/truseq_targeted_rna_expression_guide_15034665.ilmn
5. http://support.illumina.com/documents/documentation/Software_Documentation/MiSeqReporter/MiSeqReporter_UserGuide_15028784_J.pdf

Illumina, Inc. • 9885 Towne Centre Drive, San Diego, CA 92121 USA • 1.800.809.4566 toll-free • 1.858.202.4566 tel • techsupport@illumina.com • illumina.com

FOR RESEARCH USE ONLY

© 2013 Illumina, Inc. All rights reserved.
Illumina, illumina*Dx*, BeadArray, BeadXpress, cBot, CSPro, DASL, Eco, Genetic Energy, GAIIx, Genome Analyzer, GenomeStudio, GoldenGate, HiScan, HiSeq, Infinium, iSelect, MiSeq, Nextera, Sentrix, Solexa, TruSeq, the pumpkin orange color, and the Genetic Energy streaming bases design are trademarks or registered trademarks of Illumina, Inc. All other brands and names contained herein are the property of their respective owners. Pub. No. 470-2013-001 Current as of 19 April 2013.

illumina®

AGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAAC
TCAACGTACCGTAACGAACGTATCATTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGATAACAGTAACACACTTCTGTTAACCTT
CGACGAAAAGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACC
AACGTACCATTAAGAGCTACCGTGCAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGATAA
AGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAAC
GATTACTTGATCCACTGATTCAACGTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAGCAACGACG
CGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGAAAAGAATGATAACAGTA