Transitioning from Microarrays to mRNA-Seq

By providing HiSeq® 2000 mRNA-Seq data in two formats, Expression Analysis enables its customers to compare sequencing results with older array data, while gaining insight into the entire transcriptome.

Introduction

Gene expression microarrays have been commercially available for more than 10 years, providing researchers with biologically meaningful results that have enhanced our understanding of disease progression and supported advances in therapeutic development. These microarrays use probes to simultaneously analyze the expression of thousands of genes, with each probe targeting a unique sequence within an mRNA transcript of interest. The result is a snapshot of actively expressed genes and transcripts (transcriptome) at a given point in time. While microarrays are effective for identifying the expression of known genes and transcripts, the snapshot of the transcriptome they provide is incomplete. They cannot detect previously unidentified genes or transcripts, and will thus miss changes in their expression that may be associated with a phenotype of interest.

As the cost of sequencing has gone down, researchers have begun to turn to mRNA sequencing (mRNA-Seq) for gene expression analysis. Offering more comprehensive coverage through single-read or paired-end sequencing, this approach enables rapid profiling and deep investigation of the transcriptome. Yet, with more than a decade of experience using microarrays and often little experience with sequencing, it's hard for some scientists to make the switch. While they need the more informative, higher quality data that mRNA-Seq provides, they also want the ability to compare the sequencing data with earlier microarray results.

Expression Analysis, an Illumina Certified Service Provider (CSPro®) based in Durham, North Carolina, developed a service solution that addresses these concerns. By developing tools that increase the accessibility of the mRNA-Seq data for sequencing neophytes and make it easier to compare sequencing data with microarray results from previous studies, Expression Analysis has uniquely positioned itself to better serve the mRNA-Seq market. iCommunity spoke with Steve McPhail, President and Chief Executive Officer, Wendell Jones, Ph.D. Vice President, Statistics and Bioinformatics, and Joel Parker, Ph.D., Principal Scientist, Statistics and Bioinformatics at Expression Analysis to learn how Illumina sequencing systems and reagents are supporting their service offering.

(Pictured left to right) Steve McPhail, President and Chief Executive Officer, Wendell Jones, Ph.D., Vice President, Statistics and Bioinformatics, and Joel Parker, Ph.D., Principal Scientist, Statistics and Bioinformatics at Expression Analysis.

Q: What types of services does Expression Analysis offer?

Steve McPhail (SM): We started out offering genotyping and gene expression testing using microarrays and brought sequencing on board in 2009 with an Illumina Genome Analyzer™ sequencer. We've since added another Genome Analyzer and three HiSeq 2000 systems, and now offer a variety of sequencing services, including small RNA sequencing, exome sequencing, candidate gene resequencing, and most recently, mRNA sequencing. The HiSeq 2000 system provides us with a way to generate more data cost-effectively.

Wendell Jones (WJ): In partnering with clients to complete a project, we advise them about different approaches to solve the research challenge. From a client's perspective, there's always the question of what platform they should use. We provide the benefits, tradeoffs, and costs of each approach and they decide the direction they'd like to take. It can be a fairly complex equation with a lot of intangibles, such as will they be able to analyze the data in the same way they have in the past or can they compare these results with data that they've generated previously. This has certainly been the case with mRNA-Seq.

Q: Why did you decide to add mRNA-Seq services?

SM: We'd been thinking about adding mRNA-Seq for just over a year because we felt that it offered very compelling benefits over microarray-based technologies in conducting transcriptome studies. mRNA-Seq offers improved specificity, so it's better at detecting transcripts, and specifically isoforms, than microarrays. It's also more sensitive in detecting differential expression and offers increased dynamic range. One of the things that all clients are concerned about with microarrays is batch processing bias, something we don't see to the same degree with sequencing. Also, it's well known that when we see expression on a microarray platform, the expression signal tends to be very compressed compared to qPCR. We see much less compression taking place in mRNA-Seq. We also see significantly improved reproducibility of data sets in mRNA-Seq based on a variety of sequencing parameters.

WJ: The fact that mRNA-Seq can detect single nucleotide variants and small insertions and deletions is huge. It's impossible to see indels with a microarray. mRNA-Seq is also unbiased with respect to potential content, so no predefined transcriptome information is required. That is, there is not a requirement to design and manufacture probes to create a result. mRNA-Seq enables us to capture a very broad range of what is going on in the sample, rather than only look into areas we already know are important.

Q: How have your customers responded to moving to mRNA-Seq for transcriptome studies?

SM: I think there were several reasons why our clients were initially slow to evaluate mRNA-Seq. Number one was cost. Until very recently, the cost of mRNA-Seq was still significantly higher than the cost of running expression profiling on microarrays. Illumina solved that with the introduction of HiSeq 2000 and its TruSeq™ v3 reagents. Since most of our customers were more familiar with microarrays, we also felt that many of them were waiting to see more gene expression studies performed with mRNA-Seq.

Q: Is that why you conducted an mRNA-Seq vs. array comparison study?

SM: We wanted to help our clients understand the benefits of mRNA-Seq over microarrays for transcriptome analysis, so we decided to conduct real-world experiments and give our clients access to these data sets so they could begin to integrate them into their research. We felt the data would be persuasive in showing how mRNA-Seq could provide additional information about the biology of what they were studying.

WJ: Instead of reference samples, we used samples with biological diversity. We simulated a real biological experiment that people send us every week, performed the analysis on microarrays and the HiSeq 2000 system, and compared the results.

Joel Parker (JM): The design of the study involved 15 breast cancer cell lines, with five unique lines representing each of three cancer subtypes. We created three independent library preparations of each line using the Illumina TruSeq protocol. Forty-five samples were multiplexed into seven pools and run on two HiSeq flow cells. The 15 samples were also assayed on Affymetrix and Illumina microarrays and we compared the performance based on repeatability, sensitivity, specificity of detection, abundance estimation, fold-change, and differential expression. The results demonstrated the uniformity and quality or the data produced with mRNA-Seq. Replicates run at different times demonstrated a negligible decrease in reproducibility as opposed to microarrays, where batch effects due to technical variation, sample handling, and process variances often negatively impact reproducibility.

mRNA-Seq detected more of the content specific to Affymetrix and Illumina microarrays than either of the microarray platforms on the same samples. It also detected genes, isoforms, and differential expression not detected by either one, with approximately one-third of the SNVs detected by mRNA-Seq not known to dbSNP, the public domain archive of all established genetic variation. We found that 25M reads are necessary for repeatability equivalent to or better than the microarray platforms, while 10M may be sufficient for equivalent or improved detection and differential expression.

SM: We generated a pretty compelling data set, initially presenting it in a September webinar[1]. It's been a popular presentation, with over 200 downloads and more than 125 people viewing the webinar. We're now sharing this data with customers.

Q: Were you surprised by these results?

WJ: We expected the sequencer to perform better than arrays on several different fronts and were happy to see that confirmed. While we weren't surprised that we would obtain more information from sequencing, we were amazed when the higher resolution of mRNA-Seq showed events occurring that we had not anticipated.

JP: mRNA-Seq identified variations that we couldn't see in the array data. In one cell line, we were able to confirm a known heterozygous mutation in GATA3, a transcription factor involved in growth control and the maintenance of differentiated epithelial cells. Research has shown that GATA3 variants may contribute to tumorigenesis in certain breast tumors[2].

Unlike microarrays, mRNA-Seq enabled us to see novel expression, specifically isoform switching between two breast cancer stages. There are instances where a gene's isoform expression increases in one cell subtype and decreases in another, while another isoform of the same gene behaves oppositely. For example, CD44 is a cell surface protein that modulates cellular signaling in differentiation, growth, and apoptosis and has been linked with metastasis. The mRNA-Seq data clearly showed increased luminal expression of one variant of CD44 and decreased claudin expression, while another variant of the same gene was just the opposite. When we first saw the switch, we thought it was interesting. Later, we found that another researcher had recently independently published documentation of the same event, and found that it was essential for epithelial-mesenchymal transition and breast cancer progression[3].

"While we weren't surprised that we would obtain more information from sequencing, we were amazed when the higher resolution of mRNA-Seq showed events occurring that we had not anticipated."

We also identified EGFR isoform switching in our samples. While this growth factor has been found to undergo isoform switching in colorectal cancer, this event had not been seen in breast cancer. Further research is needed to understand its potential role in the progression of the disease.

Q: What other hurdles are there for clients wanting to switch from gene expression microarrays to mRNA-Seq?

SM: Mainly the fact that computational and analytical tools have not been readily available to assist researchers in analyzing the sequencing data. Most of the gene expression informatics infrastructure has been built around handling array-based data sets, so customer pipelines are not yet in shape to handle RNA-Seq data sets. Up until now, they've also had no way to compare their new mRNA-Seq data with their old microarray data.

Q: Part of your mRNA-Seq offering is a bioinformatics output that mimics array reports. Can you talk more about your deliverables and the software you developed to do that?

WJ: We recognized that we'd need to help some of our clients overcome the inertia of switching platforms, especially their concerns about wanting to compare their new data with their old data. So we offer the typical sequencing output you'd expect, with reads

and alignments, as well as quantitation of the transcriptome—in an easy-to-read table output—to a much finer degree than you could have with microarrays.

To enable customers to compare their data, we also provide sequencing data in a format that mimics the CEL files many Affymetrix clients use for downstream analysis of microarrays. CEL files are the standard format for basic Affymetrix array data and reflect hybridization data for all the probes on an array. Numerous pipelines and statistical packages such as R/Bioconductor have adopted this format as a standard input to facilitate downstream analysis. We created a 'translated' CEL file that's in the same CEL format they're used to, but reflects sequencing data rather than hybridization data. Even though the richness of the sequencing data set is not in the CEL file, it will be easy to see that a portion of the mRNA-Seq results are better. For example, you'll see greater detection and dynamic range. These files allow our clients to compare data that they've generated previously with mRNA-Seq data, creating a bridge into the richer data set that they obtain via sequencing.

Q: Since they aren't used to working with sequencing data, how do you help customers decipher mRNA-Seq results?

WJ: We provide summarized quantitation of transcripts that appear in a spreadsheet-like format, just like they're used to seeing, with transcripts in each row, samples in each column, and quantitative values for the transcripts in each sample. To help dig deeper into the data set, we provide alignment files that can be used in genome or transcriptome viewers to examine alignments, point mutations, or small indels that could be relevant to the biology of what is being studied. Our alignment tools were designed to provide better unambiguous alignments, and deliver better detection and repeatability than when using standard tools like Bowtie or TopHat individually.

> "The HiSeq 2000 system generates the number of reads to make mRNA-Seq cost-competitive with microarrays."

SM: We have an agreement with Golden Helix to offer a cloud-based service for the computational heavy lifting and processing of the mRNA-Seq intermediate data. The Golden Helix application manages the data on an as-needed basis, allowing researchers to see information at any place in the genome without downloading all the data at once.

WJ: In this way, researchers are getting the full richness of the sequencing data set, with the ability to see what's occurring at a base level, and if they want to, to view their entire sequencing data set. They can also zoom out and easily look at the overall experiment, just like they do with microarrays.

Q: Do you anticipate increasing demand for your mRNA-Seq services?

WJ: I think the data are very compelling and we're doing a good job of eliminating perceived barriers by providing researchers with a data set that mimics their microarray data set formats and lets them view their output data in viewers on their desktop, and store, process, and analyze the intermediate data in the cloud.

I see the transition from microarrays to mRNA-Seq as paralleling the shift in data demands that we initially saw with microarrays. For a long time, people wanted the microarray data and the images. As they became more comfortable with the image processing quality, they realized all they needed were the final data. The same thing may be true for mRNA-Seq, especially those who are looking at it as a microarray replacement. They won't necessarily need to have all the sequencing data; they could look at the quantification output, identify the novel information, and use that to move their research forward. We also archive the data for a period of time, so if they need to look at it later on they can.

Q: Why did you choose Illumina sequencers, and specifically the HiSeq 2000 system, for your mRNA-Seq services?

SM: We chose Illumina sequencers because your company is the market leader. We spoke with a number of your customers who were very happy with the performance of their systems.

WJ: Especially for mRNA-Seq, just about every measure regarding performance (including many quality metrics) is going to be driven by the number of reads that you can generate on the sequencer. So if you have a technology that can generate a very large number of reads, that gives you a competitive advantage performing mRNA-Seq. The HiSeq 2000 system generates the number of reads to make mRNA-Seq cost-competitive with microarrays.

Q: Do you see the use of microarrays becoming limited to niche applications?

SM: I see the use of arrays for discovery-based applications falling off pretty substantially. I think arrays will still have a place in the market for development-based applications and potentially for clinical applications.

WJ: For clinical applications, it depends on how quickly things become CLIA-approved. Microarrays have a head start because they've been around longer, but turnaround time is also important in a clinical setting. Systems like the Illumina MiSeq® have a fast turnaround, so for a small number of samples in the clinic, sequencers may offer an advantage even over microarrays.

References:

1. http://www.expressionanalysis.com/webinars/
2. Usary J, Llaca V, Karaca G, Presswala S, Karaca M, et al. (2004) Mutation of GATA3 in human breast tumors. Oncogene 23:7669–7678.
3. Brown RL, Reinke LM, Damerow MS, Perez D, Chodosh LA, et al. (2011) CD44 splice isoform switching in human and mouse epithelium is essential for epithelial-mesenchymal transition and breast cancer progression. J Clin Invest 121: 1064–1074.

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

Pub. No. 070-2011-034 Current as of 12 December 2011

illumina®

AGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAAC
TCAACGTACCGTAACGAACGTATCATTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGATAACAGTAACACACTTCTGTTAACCTT
CGACGAAAAGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACC
AACGTACCATTAAGAGCTACCGTGCAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGATAA
AGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAAC
GATTACTTGATCCACTGATTCAACGTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAGCAACGAGC
CGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGAAAAGAATGATAACAGTA