A Tale of Two RNA Library Preparation Kits

A critical comparison between two popular TruSeq™ RNA Library Prep Kits reveals new information of interest to researchers conducting RNA sequencing studies.

Introduction

Raffaele Calogero is an Associate Professor in the Department of Molecular Biotechnology and Health Science at the University of Torino. He leads the Bioinformatics and Genomics Unit, a five-member research group focused on mining genomic and transcriptomic data to identify biomarkers and investigate the molecular basis of cancer and other multifactorial diseases. The group designs its own software apps and uses BaseSpace™ Sequence Hub to analyze RNA sequencing (RNA-Seq) data and to provide sequencing and expert bioinformatics support services to other research groups.

As a specialist in genomic and transcriptomic data analysis, Professor Calogero is interested in how RNA data are generated. Seeking new ways to streamline laboratory operations, he recently performed a comparison of the TruSeq RNA Access Library Prep Kit* and the TruSeq RNA Library Prep Kit. Sequencing was performed on the NextSeq™ 500 System†, with data streaming to BaseSpace Sequence Hub for data analysis using open-source software.

iCommunity spoke with Professor Calogero about how the results of the study could inform preparation of RNA libraries for variant detection, fusion detection, and circular RNA analysis.

Q: What is the research focus of your group?
Raffaele Calogero (RC): Our research is focused on oncology and biomarker discovery. We also have a few projects focused on rare disease–related biomarker discovery for drug response and patient stratification. For example, we have a study in progress to identify the genes that are involved in resistance to the ALK inhibitor crizotinib. We are also involved in a leukemia biomarker characterization study looking at extracellular vesicular RNA.

The principal method we are using in all these studies is differential gene expression analysis. We also use isoform differential expression analysis, fusion detection, and circular RNA detection.

Q: What sequencing systems and data analysis software do you use for RNA-Seq studies?
RC: We use the NextSeq 500 System for RNA-Seq studies. The NextSeq 500 System is the ideal size for us. It gives us the flexibility to set up our experiments in a dynamic way. It is constantly in use throughout the week.

For data analysis, we use open-source software, mainly R[1] and Python,[2] and we design scripts for data analysis. Because we work mainly in oncology, we prepare DNA and RNA data according to best practices for the Broad Institute Genome Analysis Toolkit (GATK)[3] and use its MuTect[4] software for variant calling. We aggregate MuTect results using the Wellcome Centre's Platypus[5] variant caller.

Q: What prompted you to conduct an RNA library prep protocol comparison study?
RC: We were looking for a way to improve laboratory efficiency by consolidating library preparation around a single method. The manner in which the TruSeq RNA Access Library Prep Kit targets coding RNA is analogous to exome sequencing. We thought that the RNA Access kit might work as well or better for variant calling than the TruSeq RNA Sample Prep Kit, which targets polyadenylated (polyA) RNA species. The scope of the comparison study expanded to also evaluate the suitability of RNA Access data for fusion detection, which is normally performed on polyA capture data, as well as circular RNA analysis, which usually requires total RNA library prep.

Q: How did you carry out the TruSeq RNA Access vs. TruSeq RNA Sample Prep data comparison study?
RC: Dr. Gary Schroth's lab at Illumina provided us with breast adenocarcinoma (MCF7) cell line RNA sequence data produced using the TruSeq RNA Access Sample Prep Kit, the TruSeq RNA Sample Prep Kit, and total RNA preparation. We received excellent coverage data that allowed us to perform our comparison study efficiently. We carried out the data analysis using open-source software, including STAR[6] mapping in two-step mode, according to GATK best practices. We did not use any in-house bioinformatics tools because we wanted researchers to be able to test our method.

Raffaele Calogero is an Associate Professor in the Department of Molecular Biotechnology and Health Science at the University of Torino.

*Currently known as TruSeq RNA Exome

†The NextSeq 500 System is no longer available and has been replaced by the NextSeq 550 System.

Q: What were the results of the comparison study?
RC: We discovered that we can detect more variants with RNA Access than with the polyA TruSeq RNA Sample Prep data at low input read levels of 20–25 million reads per sample. At higher read levels, the difference in the number of variants detected in each of the data libraries diminishes until it reaches zero at around 100 million reads.

We also found that the amount of off-targeting is much lower in the RNA Access data than in the polyA data. In the polyA data, we saw many reads that localize to intergenic regions, but we did not see the same thing with the RNA Access data. That's because the RNA Access method is designed using coding exon sequences, so it is more efficient than polyA in detecting variants at relatively low input reads.

> "RNA Access enables researchers to consolidate library preparation for variant calling, fusion detection, and circular RNA identification around a single method."

Q: What should researchers keep in mind when considering which RNA library prep kit to use?
RC: RNA Access is going to be more efficient than polyA in detecting variants in a standard gene-level analysis format of 20–25 million reads per sample. However, polyA is clearly the better choice if researchers are looking for variants outside of the coding exon because those areas are not specifically covered by RNA Access.

Researchers should also consider the high level of library prep flexibility that is offered by the RNA Access method. For example, when we searched for fusion data, we found the same number of fusion transcripts in the RNA Access and polyA data, even at low input levels. This means that it is possible to use RNA Access to search for fusion genes in degraded samples that are not compatible with polyA capture.

Another useful finding was that RNA Access enabled circular RNA identification, which typically requires a total RNA prep. So, RNA Access enables researchers to consolidate library preparation for variant calling, fusion detection, and circular RNA identification around a single method. While RNA Access has the disadvantage of being a little more expensive than polyA, it offers more flexibility in the type of RNA that it can detected. Unlike polyA, it also provides a way of normalizing RNA samples of varying quality.

Q: Are there unique situations that researchers should understand before using these RNA library prep methods?
RC: In rare situations, a coding and a noncoding gene might be localized on the same strand. They are sharing part of the common region, but the exon and introns are not overlapping completely. In that case, we might assign reads to a specific exon that belongs to the coding and noncoding region of the same strand. We might be unable to assign the reads correctly to the coding or noncoding regions unless we rely on inference by looking at the surrounding sequence.

Although we found the same number of fusion transcripts in the RNA Access and polyA data, the ability to detect any specific fusion transcript was RNA library prep–dependent. Because MCF7 is very well studied, we collected all the validated fusion events that were published. We used JAFFA[7] to search for them in the RNA Access and polyA data. We started with technical replicates of the RNA, so the only difference was the library prep. Some fusions were detected in both data sets. However, other fusions were present in only one data set or the other. It's difficult to say whether one RNA library prep method is better than the other in detecting fusion transcripts. I think they are comparable.

Q: What are the next steps in your research?
RC: We are preparing a paper on our TruSeq RNA Access vs TruSeq RNA library prep comparison to submit for publication. We hope that others will soon have the opportunity to examine the study in detail.

For our crizotinib inhibitor study, we're working with RNA-Seq, exome, and microRNA data generated from the same samples. We're using RNA Access to see what is changing during the conversion of lymphoma cell lines from crizotinib-sensitive to crizontinib-resistant. Using RNA Access allows us to correlate the expressed variants with exome-level data and to determine which ones are affecting functional proteins.

For our leukemia biomarker characterization study, we have data on hundreds of acute lymphocytic (ALL), acute myelogenous (AML), and chronic lymphocytic (CLL) leukemia samples, as well as other leukemia samples. We are looking for potential relationships between the RNA extracellular transcriptomes and patient clinical histories.

> "BaseSpace Apps make complex analysis easy for less-experienced bioinformaticians and enable them to trace the analysis steps that were executed."

Q: How can other researchers take advantage of your team's knowledge and expertise in bioinformatics?
RC: BaseSpace Sequence Hub provides an effective way for us to share our bioinformatics experience. We have one BaseSpace App for miRNA analysis that is already in BaseSpace Sequence Hub, and two others that are under release. Currently, there isn't an app for circular RNA detection. However, we have developed one that embeds CIRI[8] software and submitted it to BaseSpace Sequence Hub for publication.

We designed the BaseSpace apps to enable biologists who are not bioinformatics experts to duplicate the analyses we have developed in our lab. BaseSpace apps make complex analysis easy for less-experienced bioinformaticians and enable them to trace the analysis steps that were executed. Another advantage is that with BaseSpace apps, there is no need to build a local infrastructure. Users have access to the required computing resources for the experiments they are running.

We also use BaseSpace Sequence Hub as a teaching tool in the genomic and transcriptomic data analysis courses that we have offered to biologists in Italy and Singapore, as well as in Germany at the European Molecular Biology Laboratory (EMBL). One of the most interesting features of BaseSpace Sequence Hub is its intuitive interface. When I ask wet-lab scientists to use it, they are not sidetracked by having to write analysis scripts. Instead, they can concentrate on understanding the analysis steps. With BaseSpace Sequence Hub, researchers can be more focused on the biological reason for what they are doing rather than how they are doing it.

Learn more about the Illumina products and systems mentioned in this article:

BaseSpace Sequence Hub, www.illumina.com/products/by-type/informatics-products/basespace-sequence-hub.html

NextSeq 550 System, www.illumina.com/systems/sequencing-platforms/nextseq.html

TruSeq RNA Access Library Prep Kit (currently known as TruSeq RNA Exome), www.illumina.com/products/by-type/sequencing-kits/library-prep-kits/truseq-rna-access.html

References

1. The R Project for Statistical Computing. www.r-project.org/. Accessed November 10, 2017.
2. Python, www.python.org/, Accessed November 10, 2017.
3. The Broad Institute. Genome Analysis Toolkit. software.broadinstitute.org/gatk/. Accessed November 10, 2017.
4. MuTect1. Genome Analysis Toolkit. software.broadinstitute.org/gatk/download/mutect. Accessed November 10, 2017.
5. Platypus: A Haplotype-Based Variant Caller for Next Generation Sequence Data. Wellcome Centre Human Genetics. www.well.ox.ac.uk/platypus. Accessed November 10, 2017.
6. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013;29:15–21.
7. GitHub. Oshlack/JAFFA. github.com/Oshlack/JAFFA/wiki. Accessed November 10, 2017.
8. Gao Y, Wang J, Zhao F. CIRI: an efficient and unbiased algorithm for *de novo* circular RNA identification. *Genome Biol*. 2015;16:4 doi: 10.1186/s13059-014-0571-3.

Illumina, Inc. • 1.800.809.4566 toll-free (US) • +1.858.202.4566 tel • techsupport@illumina.com • www.illumina.com