Leveraging the Old with the New: Exploring and Integrating Historic Microarray Studies with Next Generation Sequencing for Multiple Myeloma

Michael A. Bauer, PhD[1], Erich A. Peterson, PhD[1], Shweta S. Chavan, PhD[1], Daniel Pollock, BS[1], Christoph Heuck, MD[2] and Donald Johann, MD, MS[1]

[1]Myeloma Institute for Research and Therapy, University of Arkansas for Medical Sciences, Little Rock, AR
[2]University for Arkansas for Medical Sciences, Little Rock, AR

Remarkable advances in our ability to study the molecular profiles of human cancers by Next Generation Sequencing (NGS) methods are in rapid progress. The promise of scientific discovery that is offered by NGS also includes the challenge concerning the interpretation of large and non-trivial experimental results. This task is potentially further complicated when a multitude of molecular profiling modalities are available, with the goal of a more integrative type analysis of the cancer biology.

As technology advances, it is critical to leverage what has been gained from historic approaches (e.g., microarrays) to new approaches (NGS). Microarray transcriptome analyses have resulted in important advances in biomedicine. For instance in multiple myeloma (MM), microarray approaches led to the development of a prognostic 70 gene risk score as well as an effective disease subtyping via cluster assignment. Both enabled an improved molecular understanding of MM, and have provided translational information for the purposes of more advanced patient management. Many researchers are now transitioning to NGS DNA approaches for the detection of sequence variants, as well as RNA-seq, also due to its discovery based nature, improved sensitivity, and dynamic range. Additionally, RNA-seq allows for the analysis of gene isoforms, splice variants, novel gene fusions, etc. There is now a pressing need to associate and integrate microarray and NGS data via advanced bioinformatic approaches.

Utilizing a model-view-controller (MVC) approach, a suite of custom software tools have been developed to integrate NGS experimental data along with microarray and Affymetix probe set-IDs, and gene annotation information from a variety of public sources, e.g., Affy HG-U133 Plus 2 release 33, Ensembl gene annotations and cross references for GRCh37, and NCBI gene annotations. The approach employs a variety of strategies to integrate, annotate, and associate

microarray and NGS datasets. Output from the Tuxedo suite and GATK can be directly integrated, and/or associated with microarray and Affymetrix probe set data, as well as necessary gene identifiers and/or symbols from a variety of sources. Strategies are employed to maximize the integration and annotation processes across molecular profiling modalities. To further enhance a cancer biology focus, novel gene sets (e.g., MM 70 risk score) can be specified. Finally, the tool can be directly interfaced into a NGS pipeline. As a proof of concept, a custom MM gene list composed of CCND1, CCND3, DKK1, FGFR3, MAF, MAFB, NFKB, and WHSC1 were applied to NGS data sets derived from the MM cell lines RMPI-8226 and H929, which were obtained from ATCC. All NGS experiments were run on an Illumina HiSeq-2000 using 101 bp PE sequencing. Agilent SureSelect was used for whole exome sample prep and Illumina TrueSeq mRNA kit v2 with poly-A selection for RNA-seq.Integrative visualizations and renderings of MM molecular biology will be illustrated.

A novel bioinformatic approach to aid in the facilitation of both annotation and association of historic microarray data in conjunction with richer NGS data is now assisting with the study of MM cancer biology.