is to integrate information from sequencing, epigenetics, and expression to identify functional and regulatory genes that are associated with melanoma. Families with two or more 1st degree relatives with melanoma were considered at high risk and were investigated in this study. Initially, sequencing data of families with 3 or more relatives with the disease were examined and shared DNA variants were selected for further examination. Genetic databases and annotation tools were used to identify genes based on their known gene function and regulation, pathways, and variant conservation. Gene browsers were also used to identify any histone markers, DNA methylation sites, and other epigenetic indicators. Based on our candidate genes, there is a possibility of genetic heterogeneity, in which multiple genes may be responsible for disease susceptibility. Selected candidate genes will undergo fine mapping to further investigate the region and replication in additional families and population studies of melanoma.

11

What Does Take to Identify the Signal From the Noise in Molecular Profiling of Tumors?

N. Eric Olson, Hugh Arnold, Todd Smith

PerkinElmer

Cancer is a complex, heterogeneous disease that is driven by continually evolving genomic changes. Our current efforts to identify and the cure or demise of patients has utilized snap shots of DNA, RNA, proteins, and/or protein/nucleic acid interactions among numerous assays. For example, sequencing genomes or exomes distinguishes germline variants from somatic mutations as one step toward identifying nucleotide changes that are truly driving mutations. However, these assays identify very large numbers of variants and substantially reducing the noise requires considering the potential impact of variants (missense, non-sense, synonomous), quality of the call, prevalence of mutations in tumor versus normal cells, and whether a gene carrying a mutation is even expressed. Consequently, molecular profiling of tumors, benefits from data obtained from several different kinds of DNA sequencing-based assays. Using data from paired tumor and normal samples we will show an example workflow that combines exome and transcriptome sequencing to identify putative driver mutations that display high signal for being impactful in cancer.

12

Quantitative Analysis of Shotgun Metagenomic Data with the Real Time Genomics Platform

Brian Hilbush, John G. Cleary, Richard Littin, Len Trigg, Sean Irvine

Real Time Genomics, Inc., San Bruno, CA

Real Time Genomics has developed high performance tools for quantitative analysis of metagenomic experiments on the Illumina, Ion Torrent and Roche 454 platforms. The search algorithms and processing pipeline were created around requirements for analysis of shotgun metagenomics samples from the Human Microbiome Project (HMP). With the configurable metagenomics pipeline, data sets can be analyzed in three separate arms for 1) functional or metabolic profiling using translated nucleotide searches of annotated databases such as KEGG with *mapx* in conjunction with the gene/metabolic pathway analysis program HUMAnN (HMP and Huttenhower lab), 2) taxonomic profiling with searches of reference genome databases with map followed by quantitative analysis of species abundance and sample composition from alignment data (SAM files) using RTG species, and 3) analysis of differences in bacterial community structure with a RTG's similarity matrix generation tool, used in combination with multivariate analysis (PCA or hierarchical clustering algorithms (SVD, singular value decomposition)). We will present analysis of HMP project data to demonstrate the utility of the pipeline for a series of metagenomic investigations.

13

Search Strategies for Glycopeptide Identification

Marshall Bern[1], Christopher Becker[1], Wilfred Tang[1], Yong J. Kil[1], Xiaoke Yin[2], Manuel Mayr[2], Kay-Hooi Khoo[3], Rosa Viner[4]

[1]*Protein Metrics,* [2]*King's College, U. of London,* [3]*Academia Sinica, Taiwan,* [4]*Thermo Fisher Scientific*

Byonic is a new proteomics search engine that can identify peptides carrying N- and O-linked glycans. Byonic offers a number of ways to search for glycopeptides, including preset glycan tables and manually entered glycan masses, and the search strategy affects the quality and quantity of spectrum assignments. Here we show how a progression of searches, from wider to narrower in both proteins and glycans, can improve sensitivity and specificity for glycopeptide identification.

We obtained data from the following samples: Glycophorin-A, PSA, human blood serum enriched for glycoproteins, and secreted proteins from human endothelial cells. All data were acquired on various Thermo Orbitrap instruments and included both HCD and ETD fragmentation. We first searched the data with a full human protein database with contaminants and decoys, and later with smaller databases produced by Byonic's "focused database" option. We started with Byonic's preset glycan search, which allows only one glycan per peptide, and then, guided by prior search results, augmented or replaced these tables with user-defined glycan modifications with appropriate limits on each type of modification.

We found that focused protein databases containing 10 – 200 proteins greatly improve the sensitivity of glycopeptide search relative to full-database searches. We found a database of likely glycoproteins, determined by PNG-ase release of N-glycans in O18 water, helpful for identifying glycopeptides carrying single N-linked glycans in the endothelial secretome. Focused glycan lists also improve sensitivity, and make possible still more complex searches. We have identified glycopeptides carrying up to two N-glycans, one N-glycan and one O-glycan, and up to four O-glycans, with only minor ambiguities in modification placement and mass distribution. More complex searches, for example, five or more O-glycans, will require improvements in completeness of fragmentation and computational methods.