

**ATS Post-Graduate Course**  
**Demystifying Microarrays**  
**Mark W Geraci, MD**

The overarching theme of the presentation is to present, in a global fashion, the currently used applications of microarrays. The past decade has witnessed an enormous growth in the utilization of microarray technologies to interrogate the genome. With the completion of the sequencing of the genome, and the identification of variation of the human genome with the Hap-Map project, array-based technologies can perform on new functionalities.

The presentation progresses through the three major functional utilizations of microarrays: 1) gene expression; 2) regulation of transcription; and 3) comparing genomes.

**Gene Expression**

Gene expression has been the earliest adopter of array-based technologies. Historical presentation of spotted arrays and two-color competitive hybridization are presented. More traditional gene expression analysis utilizes a three-prime bias amplification strategy. Description of the technologies for traditional expression is presented. The work by Dr. David Erle demonstrates the utility of traditional gene expression algorithms in a large study published this past year. Disease-specific gene expression profiles can be elicited through the examination of multiple murine models of lung disease. Moreover, Dr. Pan-Chyr Yang has recently published the utility of traditional gene expressions signatures for determining the clinical outcome in non-small cell lung cancer.

More recently, Affymetrix has embarked upon exon-level analysis. In this strategy, 1.4 million exons are assayed on a single microarray. The strategy behind this level of analysis required the optimization of unbiased amplification. This technique is termed whole transcriptome amplification (WTA). Examples of gene-level as well as exon-level analysis are presented utilizing samples from the lungs of patients with pulmonary arterial hypertension. Importantly, recent work by Miller et al. has demonstrated that exon-level analysis most highly correlates with proteomic data. Previously, significant discrepancies existed between microarray data and data derived from proteomic analysis. The investigators are able to demonstrate that, with exon-level analysis, protein expressions are much more highly correlated. With this important finding in mind, a brief introduction of proteomics is offered. Dr. Lorraine Ware has been involved in proteomic analysis in lung diseases extensively, and work is highlighted regarding the NHLBI clinical proteomics programs.

Ultimately, the most unbiased approach for transcript expression occurs from the use of tiling arrays. Tiling arrays make utilization of high-density probe construction covering the entire genome. By this investigative strategy, Gingeras et al., were able to demonstrate (in a landmark article published in *Science*) that significant transcriptional activity occurs in numerous regions not previously predicted by traditional molecular

biology. Intronic expression accounts for approximately one-third of all expression and intergenic expression accounts for approximately two-thirds of all transcriptional activity. With tiling arrays come significant new possibilities, including functionalities of non-coding RNAs, transcribed fragments, and transcripts of unknown function (TUFs). These transcriptional units yet to be investigated with great detail scientifically.

### **Regulation of Transcription**

Microarrays are now capable of assaying the regulation of gene transcription. All major commercial vendors have the capability for measuring methylation sensitive aspects of gene transcription, including Affymetrix, Agilent, and Illumina. Dr. David Schwartz will speak later in the conference evaluating genome-wide DNA methylation changes by methylation specific digital karyotyping among other means. Important work has recently been published by Irizarry et al., utilizing a novel strategy involving comprehensive array-based analysis for relative methylation. This new assay, entitled "Comprehensive High Through-put Arrays For Relative Methylation" (CHARM), demonstrates that while many investigators have focused on CpG islands as regulators for gene transcription, areas within 2000 base pairs of the CpG islands (termed methylation shores) may be much more important in determining gene transcription regulation. These important findings have been shown in human colon cancer and are likely to be utilized with increasing frequency for an unbiased examination of the methylome. At best, we can ascertain that we are only beginning to understand the impact of methylation in transcript regulation, and these newer tools and strategies are highly likely to impact our understanding of this means of regulation.

MicroRNAs (miRs) represent important regulators of gene transcription. Dr. Patrick Nana-Sinkam recently reviewed the topic *vis a vis* Lung Diseases. Dr. Avi Spira utilized an integrated approach of assessment of microRNAs as well as traditional gene expression measurements to determine the concordance between microRNA expression and the expression level of predicted targets. This highly integrated approach (by examining both microRNA expression as well as traditional gene expression) is an optimal experimental paradigm, and likely to be adopted by many investigators interested in microRNA targets.

Regulation of transcription can also occur by the precise examination of nucleosome positioning. Nucleosome positioning can be determined by chromatin immunoprecipitation (ChIP) and measurements of chromatin immunoprecipitation can be performed on microarrays - therefore the ChIP-on-chip technique. In addition, chromatin IP can be assessed by the newer high through-put sequencing technologies thus termed ChIP-Seq.

### **Comparing Genomes**

With the evolution and completed sequence of the human genome, the haplotype mapping project determined human variation (HapMap Project). Hybridization strategies can be utilized to examine single nucleotide polymorphisms (SNPs) between individuals. This strategy enables investigators to perform Genome-Wide Association Studies (GWAS) with high fidelity. Moreover, quantitative assessment can be utilized to

determine copy number variations (CNVs) in addition to single nucleotide polymorphic variations. Both of these approaches have considerable power to determine the genetic etiology of certain diseases and risks for diseases. The major platforms utilized are Affymetrix and Illumina. Affymetrix utilizes predigested genomic DNA in a hybridization strategy for allele-specific determination of SNPs. Illumina uses a slightly distinct primer extension technology for SNP analysis. Dr. Carol Ober will speak later in the forum for her extensive work using SNPs and has recently published an article in the *New England Journal of Medicine* examining candidate genes in lung diseases for asthma and determinants of lung function. Copy number variations, in essence a form of polymorphic variations between individuals, represent another genetic variation that can be assayed by SNP arrays. Indeed, segmental copy number variations have recently been shown to shape tissue transcriptomes. They are of utmost important in disease pathogenesis. Indeed, some of the highest risk factors for autism have been linked CNV felt to be most associated with diseases. We present data from our own laboratory utilizing SNP arrays on isolated endothelial cells from patients with pulmonary artery hypertension. A significant percentage of the cells show chromosomal disruption in selected chromosomes. Confirmation has been performed by fluorescent *in situ* hybridization (FISH). Moreover, application of FISH technologies to the lung tissue from explanted lungs of patients with PAH confirms, unquestionably, the presence of chromosomal copy number aberrations in patients with this disease.

Next generation sequencing offers the potential and hope for quantization of expression, broad coverage of sequence variation, and the hope of the “\$1,000 genome.” A brief description of some of these strategies to be utilized is presented with the hope that within a decade, the “thousand dollar genome” will be a reality.

### **Systems Biology Analysis**

Putting together the wealth of data that comes from all of the array-based technologies, and in the future will come from high through-put sequencing, can be extraordinarily challenging. Dr. Naftali Kaminski will speak later in the forum regarding a functional and regulatory map of asthma and some of the excellent work he has preformed in informatics from traditional expression arrays. Our group has been utilizing traditional expression arrays for the predictive capability of sensitivity to the EGFR inhibitors for the treatment of non-small cell lung cancer. The characteristic signature for EGFR sensitivity also prospectively predicts cell line sensitivity. Moreover, in cancers of the head and neck, which do not harbor EGFR mutations, the signature for EGFR sensitivity is maintained. Dr. Eric Schadt will conclude the seminar with a broad perspective of drug development and a summary of the overarching themes presented throughout this forum.