Preface

Gene expression profiling using microarrays has been increasingly brought to task for the simple goal of ensuring that data from a particular tissue and a particular platform match the same tissue run on another platform in another lab. Expression arrays are becoming extremely valuable in the clinic and, as such, are being closely monitored by the Food and Drug Administration (FDA). MammaPrint uses the Agilent microarray system, which has been FDA approved. The National Institute for Standards has created the Microarray Quality Control Consortium in order to identify those factors that tend to throw lab-to-lab and cross-platform concordance off. They have made remarkable progress, and papers and new insight continue to pour out of their initial investment. The External RNA Consortium is purposed to create standards and spike-ins to allow proper measurements of array performance no matter what the location or environmental conditions are.

These efforts are moving the expression microarray from the experimental lab into the position of a trusted and widespread technological commodity. At that point it becomes a tool, but before that happens, the concept of data preprocessing and cross-platform concordance must be thoroughly addressed. Both topics have received much attention, but a line in the sand must be drawn beyond which we should assume that proper normalization will be used to minimize noise, bias, and cross-platform discordance, and to improve accuracy of biological interpretation.

Normalization and array preprocessing encompass a variety of data manipulations, but in this book we will look at those mathematical formulae that remove known and intrinsic biases in order to level the playing field—leaving expression data to stand on their own without accommodating the platform manufacturer, the fluorescent dye used, the varying quality of RNA, the laboratory, scanner, or any of a number of potential nuisance factors. In this book, many of the most respected authors in the field present their work for the reader to use and judge. In addition, several scientists have chosen to present their novel techniques for next-generation technologies such as the single nucleotide polymorphism (SNP) chip, the exon array, the comparative genomic hybridization (CGH) array, and the protein array.

Some useful references for the reader to engage in his or her own application of normalization include:

DNMAD (Diagnosis and Normalization for Microarray Data): http://dnmad.bioinfo.cnio.es/

Global normalization of two-color microarray data with loess: http://nbc11.biologie.uni-kl.de/framed/left/menu/auto/right/microarray/loess.shtml

SNOMAD: http://pevsnerlab.kennedykrieger.org/snomadinput.html

Terry Speed's "Normalization of Microarray Data—How to Do It!": http://www.maths.lth.se/bioinformatics/calendar/20010409/

along axons on actin and microtubules. This was a compelling data set, especially when examined using the appropriate statistics. I wished to extract insight using highly repetitive observations. This was the realm of molecular (nonclinical trials-based) biostatistics.

It was at this time that Motorola Life Sciences was growing their CodeLink microarray business. I joined their group and took advantage of Motorola University's Blackbelt Program to apply Six Sigma statistical methods, which were then strictly associated with semiconductor manufacturing. Mixing Six Sigma ideas with biological data was revolutionary. The first few meetings of Cambridge Healthtech's microarray data analysis conferences were filled with statisticians arguing about confidence, false positives, and how best to normalize cDNA data. Affymetrix was just making their early U95 chips; Agilent had split from Hewlett Packard and was optimizing the ink-jet printing technology to spot cDNA molecules. As commercial expression arrays grew in demand, more emphasis was placed on quality control, replication, and experimental design. I enjoyed the business of creating and analyzing new microarray technology, but I was drawn to a new field: translational genomics. Here was a field that was using microarrays to their fullest extent; TGen, the Translational Genomics Research Institute, and Jeff Trent were leading the charge.

Imagine a clinical trial where you could measure improvement in health between patients whose doctor did or did not have access to information from an expression microarray. This was the type of clinical research that led to Agendia's array-based breast cancer detection system, MammaPrint. Something dramatic happened: Normalization of expression data could have a direct effect on an individual's health. No longer was mathematics abstract; a mistake in data processing could ruin the outcome of a cancer treatment. This is arguably the current state of events, and it was at this point I began seeking the next big, highly parallel biomedical technology. Antibody arrays, glycoarrays, protein and peptide chips, methylation arrays, aptomer, CGH, exon, and SNP chips are emerging quickly. With them come more mathematical challenges. The Biodesign Institute at Arizona State University prides itself on tackling these challenges on the way to creating disruptive paradigms in medicine and health care. This is where I pursue my research; it is here that new bioinformatic solutions are being created to interpret highly parallel measurements of biomolecules.

To those who are interested in pursuing this field, I would leave some advice: Learn mathematics while you learn biology. Develop models that predict biological behavior. Think about modular software design even as you think about biological behavior. The newest biomedical technologies are only as good as the computational support behind them. These are the new integrative sciences that will propel science into the next millennium, and today's bioinformatics students will be creating tomorrow's medical advances.

Phillip Stafford
Tempe, Arizona