Preface

One of the major goals of genome studies is to determine the roles of all the genes involved in biological processes. To reach this goal, it is necessary to monitor, qualitatively and quantitatively, all gene products dynamically present in the cells. One way of achieving this is to monitor the transcripts that reflect directly the activities of gene expression in the genome.

Technological developments since the beginning of modern molecular biology have played key roles in transcript identification, detection, and quantitation. Such developments include the 1) hybridization-based methods such as northern blotting, "Dot-blot" techniques, RNA *in situ* hybridization, and RNase protection; 2) PCR amplification-based methods such as quantative and semi-quantitative RT-PCR, and real-time PCR; and 3) gel display-based methods such as Differential Display. To cope with the rapid progress of genome sciences, new technologies, such as large-scale EST collection, cDNA and oligonucleotide microarray, SAGE and MPSS, have been developed specifically for large-scale, high-throughput, and whole-genome level transcript analysis.

SAGE has its unique advantages for transcript analysis. By identifying a short tag sequence from a transcript and collecting a large number of tags from transcripts, SAGE provides a quantitative measure of gene expression at the genome level. Due to the cancatermerization of multiple tags for single sequencing reaction, SAGE provides an order of magnitude higher sensitivity than EST for transcript identification. This significantly overcomes the barrier of transcript redundancy and results in the detection of more low abundant transcripts without the need of subtraction /normalization. As an open system, SAGE detects all types of transcripts including high abundant transcripts, low abundant transcripts, known transcripts, alternative spliced transcripts, antisense transcripts, and novel transcripts. Since its invention, SAGE has been increasingly applied to many studies and data from these have provided valuable information for our understanding of biology and medicine. Wider applications of SAGE are expected in near future.

Like any other technology, SAGE also has its limitations. These include the technical difficulties for SAGE library construction, the high cost of tag sequence collection, the lower-specificity for SAGE tags to represent the original genes, the statistical challenging task of mining SAGE data, and complicated tag-to-gene annotation. Understanding the basic features of SAGE will provide a rationale for maximizing the advantages and minimizing the disadvantages of SAGE experiments.

This book reports the latest progress in the SAGE field. Written by scientists who are experts in the SAGE, the book contains chapters on a wide variety of SAGE issues including the evaluation of SAGE, the improvement and further development of the SAGE methodology, the statistical analysis and mathematical modeling of SAGE data, and the application of SAGE in various biological and medical topics. I expect that this book will provide useful and systematic information for people who are using SAGE, who are planning to use SAGE, and who are not going to use SAGE but want to know more about SAGE.

San Ming Wang
October 2004