
Preface

Completion of the sequence of the human genome represents an unparalleled achievement in the history of biology. The project has produced nearly complete, highly accurate, and comprehensive sequences of genomes of several organisms including human, mouse, drosophila, and yeast. Furthermore, the development of high-throughput technologies has led to an explosion of projects to sequence the genomes of additional organisms including rat, chimp, dog, bee, chicken, and the list is expanding.

The nearly completed draft of genomic sequences from numerous species has opened a new era of research in biology and in biomedical sciences. In keeping with the interdisciplinary nature of the new scientific era, the chapters in *Gene Mapping, Discovery, and Expression: Methods and Protocols* recapitulate the necessity of integration of experimental and computational tools for solving important research problems. The general underlying theme of this volume is DNA sequence-based technologies. At one level, the book highlights the importance of databases, genome-browsers, and web-based tools for data access and analysis. More specifically, sequencing projects routinely deposit their data in publicly available databases including GenBank, at the National Center of Biotechnology (NCBI) in the United States; EMBL, maintained by the European Bioinformatics Institute; and DDBJ, the DNA Data Bank of Japan. Currently, several browsers offer facile access to numerous genomic DNA sequences for gene mapping and data retrieval. These include the map-view at NCBI; the genome browser at the University of California at Santa Cruz, UCSC; and the browser maintained by Ensembl. All three browsers offer sophisticated tools for gene mapping and localization on genomic DNA.

For beginners in the field, through a specific example, one chapter provides a step-by-step procedure for localization, creating a map, and a graphical representation of genes of interest using the genome browser at UCSC. Since the drafts of the genomic sequences provide primarily a reference for studies of gene organization, additional methods are needed for understanding the complexity and dynamic nature of chromosomes. Significantly, segmental duplications are a common feature of many mammalian genomes. Therefore, *Gene Mapping, Discovery, and Expression: Methods and Protocols* provides a computational protocol for identifying and mapping recent segmental and gene duplications. Another chapter offers a step-by-step procedure for identifying paralogous genes, using the genome browser at UCSC.

To examine local variations in specific regions of chromosomes experimentally, a chapter provides a novel method, Quantitative DNA Fiber Mapping, that relies on fluorescent *in situ* hybridization (FISH) to identify, delineate, and characterize selected, often small, DNA sequences along a larger piece of the human genome. In another experimental contribution, a chapter describes a sensitive and specific method, Primed *in situ* labeling, that can be used for localization of single copy genes and sequences too small for detection by conventional FISH.

Novel DNA sequence-based strategies include methods for the discovery and mapping of the functional elements and the “codes” in DNA that regulate the expression of genes. The completed sequence of the human genome and the genomic sequences of model organisms offer a rich source of data for addressing this problem. A fundamental and powerful method is based on comparing the sequences from different species to identify the conserved functional elements. A chapter in this volume describes the VISTA family of computational tools, created to assist researchers in aligning DNA sequences for locating the genomic DNA regions that are highly conserved. Another chapter aims at using sequence conservation as a guide for identifying the elements that may regulate the expression of genes. This chapter describes how to use publicly available servers (Galaxy, the UCSC Table Browser, and GALA) to find genomic sequences whose alignments show properties associated with *cis*-regulatory modules and conserved transcription factor binding sites. Furthermore, this volume describes additional versatile and web-based tools for promoter, regulatory region, and expression analyses. These tools include CORG “COMparative Regulatory Genomics” and BEARR “Batch Extraction and Analysis of *cis*-Regulatory Regions.”

DNA sequence-based technologies include other strategies that could help with the identification of regulatory signals and potential protein binding elements in the regulatory regions of genes. For example, a chapter describes how a database of 9-mers from promoter regions of human protein-coding genes could be accessed via the web for the discovery of the lexical characteristics of potential regulatory motifs in human genomic DNA. These characteristics could help with predicting and classifying regulatory *cis*-elements according to the genes that they control.

Cis-elements can control the expression of genes in an allele-specific fashion. The analysis of allele-specific gene expression is of interest in the study of genomic imprinting. Significantly, there is growing awareness that differences in allelic expression could be widespread among autosomal non-imprinted genes. A chapter in *Gene Mapping, Discovery, and Expression: Methods and Protocols* provides protocols for *in vivo* analysis of allelic-specific gene

expression. These include analysis of the relative allelic abundance of transcribed RNA, and of transcription factor recruitment and Pol II loading by chromatin immunoprecipitation. Another chapter describes miRNAs expression vectors containing human RNA polymerase II or III promoters for studies of the control of gene expression.

In this new scientific era, gene expression is extensively studied using microarray technologies. Two chapters describe how to use web-based tools for accessing and analyzing the microarray data. One chapter describes Gene Expression Omnibus (GEO) developed at NCBI. GEO has emerged as a leading fully public repository for gene expression data. The chapter describes how to use Web-based interfaces, applications, and graphics to effectively explore, visualize and interpret the hundreds of microarray studies and millions of gene expression patterns stored in GEO. Another chapter describes the resources at the Stanford Microarray Database (SMD). This database offers a large amount of data for public use. The chapter describes how to use the primary tools for searching, browsing, retrieving, and analyzing data available at SMD. Furthermore, researchers, educators, and students may find SMD a very useful repository of a large quantity of publicly available data that together with analysis tools, could be used for exploratory, unsupervised analysis and discovery.

Another level of sequence-based technologies depends on how best to analyze the structural organization of chromosomes, evaluate the sequence specificity of transcription factors, and isolate and identify the components of the protein complexes formed with DNA. More specifically, in cells, the chromosomal DNA is associated with proteins to form complexes referred to as chromatin. A major group of chromosomal proteins, the histones, functions in the compaction of DNA by forming nucleosomes. Another major group corresponds to transcription factors, which control the expression of genes through protein–DNA and protein–protein interactions. Evidence supports major roles for the underlying DNA sequence on the relative arrangement of proteins along the chromosomes. Two chapters in this volume provide DNA sequence-based methods for probing chromatin structure. One chapter describes a step-by-step procedure for detecting and analyzing nucleosome ladders on unique DNA sequences. Another offers a non-invasive method of assaying relative DNA accessibility in yeast chromatin without disrupting DNA–protein interactions.

The DNA sequence specificities of transcription factors are key components of the *cis* regulatory networks. However, despite their importance, the DNA binding specificities of many transcription factors remain unknown. Furthermore, methods routinely used for characterizing protein binding sites are not scalable and are time-consuming. These issues are problematic because complete, accurate, and reliable datasets of transcription factor binding elements

are needed for localizing the regulatory regions of genes. This volume offers two chapters on novel DNA microarray-based technologies for rapid, high-throughput in vitro characterization of the DNA sequence specificities of transcription factors.

Lastly, several chapters in *Gene Mapping, Discovery, and Expression: Methods and Protocols* offer non-invasive technologies for the isolation of transcription factor complexes formed with specific DNA sequences used as bait. Identification of the components of large protein–DNA complexes is an important step in elucidating the mechanisms by which gene expression is controlled. Two chapters describe the use of powerful methods based on mass spectrometry for identification of proteins in the complexes formed with DNA. These methods can lead to the discovery of novel transcription factors with important roles in the control of gene expression.

Minou Bina