Preface

About the time when the draft of the human genome sequence had first appeared, I kept asking myself if there is a systematic, scalable approach to decoding regulatory elements. I also wondered "how can we now understand the network between transcription factors and the genes that they activate and can we have a technology that can be applied to any biological question and sample"?

Even then it was possible to map expressed sequencing tags (ESTs) to the genome, to locate promoter regions and somehow correlate this with expression, but it was too expensive to do it routinely. I had a crazy idea one Sunday afternoon after hours of drawing on a notebook: what if I could chop and concatenate the 5′ ends deriving from full-length cDNA and sequence more than 10 or 20 of these short tags in a single sequencing run? Despite the technical challenges it seemed to make sense: by aligning short tags to the genome, it would become possible to detect all the active promoters and then the expression of transcription factors. I concluded that, if more libraries than the number of existing transcription factors and regulators are made, we could theoretically detect the network between these proteins and the regulated genes. Thus cap-analysis gene expression (CAGE) was conceived and I thought it was worth giving it a try.

After a couple of years of overcoming technical challenges, my team could create a protocol for CAGE with a method of profiling gene expression by producing and sequencing 20-nt short sequencing tags corresponding to the beginning of the RNAs. CAGE technology was readily employed in the Fantom3 project. This revolutionized our understanding of the genome, because we unexpectedly found that the genome produces a much larger variety of RNAs than earlier reported. The genome sequence alone gave us enough information to explain its functions, but with CAGE it became possible to promptly identify novel mRNAs, non-coding RNAs and their promoters, as we did in FANTOM, ENCODE and a growing number of other projects.

These projects have challenged the dogma: contrary to expectations, most genes produce multiple mRNAs and non-coding RNAs starting from multiple promoters. Ultimately, CAGE analysis can elucidate the relationship between the mRNAs and the promoters that control their expression in order to decipher the networks that regulate gene expression and the transcription factors.

Moreover with CAGE, we can comprehensively identify the exact locations of the genome from which the mRNAs originate, identify core promoters, and simultaneously quantify RNA expression levels. Therefore, CAGE has been broadly adopted to infer transcriptional networks because it provides the tools to understand the molecular mechanisms underlying gene expression. CAGE becomes even more valuable when it is used in conjunction with next-generation sequencing technologies, which make CAGE cheaper and more informative than current microarrays.

This book is a guide for current and potential users of CAGE technology who wish to reveal molecular mechanisms in CAGE experiments. The book includes protocols and a guide to the bioinformatics analysis of CAGE datasets, including the design of software and tools for constructing web resources or using existing genome browsers to customize data. I hope that the chapters will be particularly useful to those who are not yet specialists in the field, and provide them with a guide for setting up CAGE technology and/or analysis in their laboratories. This book also provides examples of applications written by the first group of scientists to use CAGE technology for promoter identification, genome annotation, identification of novel RNAs and reconstruction of models of transcriptional control and networks, which may help the readers extracting additional biological insights from the published data.

In conclusion, CAGE technology offers a revolutionary approach for a growing number of scientists beyond early users of genome sequencing centers. This book introduces CAGE technology and its analysis to a broad readership with interests in expression analysis, transcriptional control, marker identification, molecular diagnostics, analysis of networks and RNA biogenesis. I hope the scientists, postdocs, students, technicians and all other readers will expand these approaches to a variety of biological problems using different models and bring forth exciting results to enrich our knowledge of biological systems. Exciting times for scientific discoveries are ahead.

My final thoughts are for the people that have been working with me over years at RIKEN and in the FANTOM consortium and all other collaborators. There are many technicians that have diligently developed experimental conditions and others that have carefully analyzed larger and larger datasets. I am the most grateful to the scientists that have inspired the analysis and interpretation of the CAGE data: their contributions have been essential and the process of discovery has been excitement and fun. Nothing could have been done in isolation, and more excitement lies ahead from the analysis of rich datasets inherent in CAGE libraries.

Piero Carninci
Omics Science Center, RIKEN, Yokohama Institute