CAP-ANALYSIS
GENE EXPRESSION
(CAGE)
THE SCIENCE OF DECODING GENE TRANSCRIPTION
CAP-ANALYSIS
GENE EXPRESSION
(CAGE)

THE SCIENCE OF DECODING GENE TRANSCRIPTION

EDITOR
PIERO CARNINCI
RIKEN, Japan

PAN STANFORD PUBLISHING
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
## Contents

**Preface**

1. Cap Analysis Gene Expression (CAGE) .................................................. 1

2. Tagging Transcription Starting Sites with CAGE .............................. 7
   2.1 The Output of the Genome is Complex ........................................ 7
   2.2 Mapping 5’ Ends: From ESTs to Tagging Technologies ................... 9
   2.3 Linking Core Promoters to Genomic Elements .............................. 12
   2.4 cDNA Ends or the Whole Sequence? ........................................... 13
   2.5 Identification of Functional Elements in the Genome .................... 15
   2.6 Technology Evolution, Same Lessons? ....................................... 17

3. Construction of CAGE Libraries .......................................................... 21
   3.1 Introduction ................................................................................. 21
   3.2 Stage 1: Synthesis of First-Strand cDNA ................................. 22
      3.2.1 Synthesis of First-Strand cDNA .......................................... 22
      3.2.2 CTAB/urea Purification ..................................................... 23
   3.3 Stage 2: Oxidation/Biotinylation ................................................. 24
      3.3.1 Oxidation ............................................................................. 24
      3.3.2 Biotinylation ........................................................................ 24
      3.3.3 RNase I Treatment Removal of Biotinylated Cap when cDNAs do not Reach the 5’ end .............................................................. 25
   3.4 Stage 3: Capture-Release ............................................................... 25
      3.4.1 Capture and Subsequent Release of 5’-Completed cDNAs .......... 26
   3.5 Stage 4: Single Strand Linker Ligation ......................................... 27
      3.5.1 Single Strand Linker Ligation ............................................... 28
      3.5.2 S400 Spin Column .............................................................. 28
   3.6 Stage 5: the Second Strand cDNA Synthesis ............................. 29
   3.7 Stage 6: Preparing CAGE Tags ...................................................... 30
      3.7.1 MmeI Digestion .................................................................. 30
      3.7.2 2nd Linker Ligation ............................................................ 31
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.7.3</td>
<td>Purification with Magnetic Beads</td>
<td>31</td>
</tr>
<tr>
<td>3.7.4</td>
<td>Purification by G50 Column</td>
<td>32</td>
</tr>
<tr>
<td>3.8</td>
<td>Stage: 7 Amplification of CAGE Tags</td>
<td>32</td>
</tr>
<tr>
<td>3.8.1</td>
<td>1st PCR Amplification</td>
<td>33</td>
</tr>
<tr>
<td>3.8.2</td>
<td>PAGE Purification</td>
<td>34</td>
</tr>
<tr>
<td>3.8.3</td>
<td>2nd PCR Amplification</td>
<td>34</td>
</tr>
<tr>
<td>3.8.4</td>
<td>Purification with QIAGEN MinElute Column</td>
<td>35</td>
</tr>
<tr>
<td>3.9</td>
<td>Stage 8: Restriction</td>
<td>35</td>
</tr>
<tr>
<td>3.9.1</td>
<td>Restriction with XmaJI</td>
<td>36</td>
</tr>
<tr>
<td>3.9.2</td>
<td>Removal of the Linkers Tips</td>
<td>36</td>
</tr>
<tr>
<td>3.9.3</td>
<td>PAGE Purification</td>
<td>36</td>
</tr>
<tr>
<td>3.10</td>
<td>Stage 9: Concatenation</td>
<td>37</td>
</tr>
<tr>
<td>3.10.1</td>
<td>Appendix</td>
<td>38</td>
</tr>
<tr>
<td>4.</td>
<td>Transcriptome and Genome Characterization Using Massively Parallel Paired End Tag (PET) Sequencing Analysis</td>
<td>41</td>
</tr>
<tr>
<td>4.1</td>
<td>Introduction</td>
<td>42</td>
</tr>
<tr>
<td>4.2</td>
<td>The Development of Pair end diTag (PET) Analysis</td>
<td>43</td>
</tr>
<tr>
<td>4.3</td>
<td>GIS-PET for Transcriptome Analysis</td>
<td>45</td>
</tr>
<tr>
<td>4.4</td>
<td>ChIP-PET for Whole Genome Mapping of Transcription Factor Binding Sites and Epigenetic Modifications</td>
<td>48</td>
</tr>
<tr>
<td>4.5</td>
<td>ChIA-PET for Whole Genome Identification of Long Range interactions</td>
<td>52</td>
</tr>
<tr>
<td>4.6</td>
<td>Perspective</td>
<td>55</td>
</tr>
<tr>
<td>5.</td>
<td>New Era of Genome-Wide Gene Expression Analysis</td>
<td>61</td>
</tr>
<tr>
<td>5.1</td>
<td>Introduction</td>
<td>62</td>
</tr>
<tr>
<td>5.2</td>
<td>Tagging Technologies for Genome-Wide Analysis</td>
<td>62</td>
</tr>
<tr>
<td>5.3</td>
<td>Principles of Next Generation Sequencing Technologies</td>
<td>63</td>
</tr>
<tr>
<td>5.3.1</td>
<td>Genome Sequencer 20/FLX System (Roche Diagnostics/454 Life Sciences)</td>
<td>63</td>
</tr>
<tr>
<td>5.4</td>
<td>Genome Analyzer (Illumina/Solexa)</td>
<td>65</td>
</tr>
<tr>
<td>5.5</td>
<td>SOLiD System (Applied Biosystems)</td>
<td>67</td>
</tr>
<tr>
<td>5.6</td>
<td>Advantages of Next Generation Sequencing Technologies over Conventional Sequencing Technology on Tagging Technologies</td>
<td>69</td>
</tr>
</tbody>
</table>
5.7 From Static Analysis to Dynamic Analysis ..... 70
5.8 CAGE Method and Next Generation Sequencing Technologies ................................ 72
5.9 Conclusions and Outlook .......................... 73

6. Computational Tools to Analyze CAGE — Introduction to PART II 79

7. Extraction and Quality Control of CAGE Tags 83
  7.1 Overview ..................................... 83
  7.2 Using Read Qualities and Read Properties, Pre- and Post-Extraction .................... 84
  7.2.1 Background: Types of Sequencing Errors .............................................. 84
  7.3 Procedures Before Tag Extraction ............... 85
  7.3.1 Ambiguous Base Calls .......................... 85
  7.3.2 Unusual Read Lengths ........................ 85
  7.3.3 Low Average Read Quality Score .............. 86
  7.3.4 Presence of Sequencing Errors ................. 86
  7.4 Using QC Values After Tag Extraction .......... 87
  7.5 Origin of Sequence Errors ..................... 87
  7.6 Using Sequence Errors to Estimate CAGE Quality .......................... 88
  7.7 A Simple CAGE Tag Extraction Method .......... 88

8. Setting CAGE Tags in a Genomic Context 93
  8.1 Mapping Pipelines for Sequence Tag Technologies .................................. 93
  8.2 A Mapping Pipeline for CAGE ........................ 96
  8.3 Benchmarking with a Sample Dataset .......... 97

9. Using CAGE Data for Quantitative Expression 101
  9.1 High Throughput Expression Platforms .......... 101
  9.2 Comparing CAGE to Other Measures of Gene Expression .......................... 103
  9.3 Platform Normalization .......................... 103
  9.3.1 Microarray Normalization ...................... 104
  9.3.2 qRT-PCR Normalization ....................... 104
  9.3.3 CAGE Normalization ......................... 105
  9.4 Replication ................................... 105
  9.5 Gene Models and Complex Loci ................... 106
  9.5.1 Complex Transcription ....................... 106
  9.5.2 Gene Expression vs. Transcript Expression 107
Contents

9.6 Construction of CAGE Promoters and Calculation of Gene Expression Levels 107
9.7 Comparison of CAGE Expression between Technical Replicates 109
9.8 Comparison of CAGE Expression from Biological Replicates 110
9.9 Comparison of CAGE Expression Between Different Time Points Within a Single Time-Course 111
9.10 Comparison of CAGE Expression Profiling to qRT-PCR Expression Measurements 113
9.11 Comparison of CAGE Expression Profiling to Microarray Measurements 114
9.12 Present/Absent Calls 114
9.13 Discussion 116

10. Databases for CAGE Visualization and Analysis 123
10.1 Introduction 123
10.2 Transcription Maps and Activity 124
10.3 Public Databases 126
10.4 Genomic View of In-House Data 128
10.5 For Expression Analyses 132
10.6 Discussion 133

11. Computational Methods to Identify Transcription Factor Binding Sites Using CAGE Information 137
11.1 Introduction 138
11.2 Schema of the Methodology Process 139
11.3 Initial Links of TF with the Affected Genes 142
11.3.1 Mapping of TFBSs to Promoters 142
11.3.2 Determining Enriched TFBSs 144
11.3.3 Score for Confidence of the Predicted TF→TFBS→TSS/Promoter→Gene Association 145
11.4 Correlation of CAGE Tag Counts of Genes and TFs 146
11.5 Ranking TF→TFBS→TSS/Promoter→GENE Association: The Effective Use of CAGE Tags 147
11.6 Verification of Results 149
11.7 Reconstruction of TRNs 150
12. Transcription Regulatory Networks Analysis
   Using CAGE 153
   12.1 CAGE Data for Network Reconstruction . . . . . . 154
   12.2 Methodology . . . . . . . . . . . . . . . . . . . . 156
       12.2.1 Step 1 of the Process . . . . . . . . . . . . . 157
       12.2.2 Step 2 of the Process . . . . . . . . . . . . . 157
       12.2.3 Step 3 of the Process . . . . . . . . . . . . . 158
       12.2.4 Step 4 of the Process . . . . . . . . . . . . . 158
   12.3 Gene Expression Data Complementary to CAGE
       for Network Reconstruction . . . . . . . . . . . . . 158
   12.4 Using Physical Interactions . . . . . . . . . . . . . 159
   12.5 TRNs Reconstruction . . . . . . . . . . . . . . . . . 161
   12.6 Using Pathway Information . . . . . . . . . . . . . 162
   12.7 Validation of the Reconstructed Networks . . . . . 162

13. Gene-Expression Ontologies and Tag-Based
    Expression Profiling 169
   13.1 Introduction . . . . . . . . . . . . . . . . . . . . . 169
   13.2 Annotating Gene Expression . . . . . . . . . . . . 173
   13.3 Using Ontologies to Integrate Expression Infor-
       mation . . . . . . . . . . . . . . . . . . . . . . . . 174

14. Lessons Learned from Genomic CAGE 179
   14.1 Introduction . . . . . . . . . . . . . . . . . . . . . 179
   14.2 The Classic View on Transcription Start Sites and
       Core Promoters . . . . . . . . . . . . . . . . . . . 180
   14.3 CAGE-Based Views of Transcription Start Sites . . 182
       14.3.1 Transcription Start Site Landscapes . . . . . 182
       14.3.2 Biological Functions of TSS Distributions . . 186
       14.3.3 Alternative TSSs . . . . . . . . . . . . . . . 187
       14.3.4 Promoters at Unexpected Locations . . . . . 187
   14.4 Probing Biological Mechanisms Using CAGE . . . 190
       14.4.1 Measuring the Effect of TATA-TSS Spacing 190
       14.4.2 Dynamic TSS Selection . . . . . . . . . . . . 191
       14.4.3 Evolutionary Turnover of Core Promoters 191
       14.4.4 The TSS Initiation Code and Prediction of
           TSS Propensity . . . . . . . . . . . . . . . . . 192

15. Future Challenges in CAGE Analysis 197
   15.1 What are we Measuring? . . . . . . . . . . . . . . 197
Contents

15.1 Transcriptome Content — Not TSS Usage 197
15.1.1 Transcriptome Content — Not TSS Usage 197
15.1.2 Unknown Population Size . . . . . . . . . 198
15.1.3 Averaging out the Differences . . . . . . . 199
15.2 How Close to “The Truth” are We? . . . . . . . . 199
15.2.1 Sampling Depth and Saturation . . . . . . 199
15.2.2 Methodological vs Biological Noise . . . . 204
15.2.3 Clustering CAGE Tags in Meaningful Way 205
16. Comparative Genomics and Mammalian Promoter Evolution 209
16.1 Introduction . . . . . . . . . . . . . . . . . . . . . 210
16.2 Resources for Comparative Genomic Analysis . 212
16.2.1 Precise Definition of Transcription Start Sites by CAGE . . . . . . . . . . . . . . . 212
16.2.2 CAGE as a Quantitative Measure of Expression . . . . . . . . . . . . . . . . . 214
16.2.3 Alignments for Comparative Genomics . 214
16.2.4 Identifying Spatial Constraints and Patterns . . . . . . . . . . . . . . . . . . . 215
16.3 Genome Wide Trends in Mammalian Promoter Evolution . . . . . . . . . . . . . . . 217
16.4 Promoters Represent an Unusual Genomic Environment . . . . . . . . . . . . . . . . 218
16.4.1 Sequence Composition . . . . . . . . . . . . 218
16.4.2 Chromatin Structure at Promoters . . . . . 219
16.4.3 Promoters and Meiotic Recombination . 220
16.5 Integration of Population Genetic Data with Comparative Genomics . . . . . . . . . 221
16.6 Concluding Remarks . . . . . . . . . . . . . . . . . 222
17. The Impact of CAGE Data on Understanding Macrophage Transcriptional Biology 227
17.1 Introduction . . . . . . . . . . . . . . . . . . . . . 227
17.2 Transcription start site and promoter characteristics revealed by CAGE . . . . . . . . . . . . . . . 229
17.3 Transcriptional Complexity: Sense-Antisense Transcription and Non-Coding RNA . . . . . . . . . . . . . . . . . . . . . . . 235
17.4 Construction of Macrophage Transcriptional Networks . . . . . . . . . . . . . . . . . 236
Contents

17.5 What does CAGE Data Offer for Traditional Studies of Promoter Regulation? ........... 237
17.6 Conclusion ........................................ 240

Color Index ............................................. 245
Index ................................................... 265