

# Contents

|                                                                                        |    |
|----------------------------------------------------------------------------------------|----|
| <i>Preface</i>                                                                         | v  |
| 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 |

|        |                                                                                                                                       |    |
|--------|---------------------------------------------------------------------------------------------------------------------------------------|----|
| 3.7.3  | Purification with Magnetic Beads . . . . .                                                                                            | 31 |
| 3.7.4  | Purification by G50 Column . . . . .                                                                                                  | 32 |
| 3.8    | Stage: 7 Amplification of CAGE Tags . . . . .                                                                                         | 32 |
| 3.8.1  | 1st PCR Amplification . . . . .                                                                                                       | 33 |
| 3.8.2  | PAGE Purification . . . . .                                                                                                           | 34 |
| 3.8.3  | 2nd PCR Amplification . . . . .                                                                                                       | 34 |
| 3.8.4  | Purification with QIAGEN MinElute<br>Column . . . . .                                                                                 | 35 |
| 3.9    | Stage 8: Restriction . . . . .                                                                                                        | 35 |
| 3.9.1  | Restriction with XmaII . . . . .                                                                                                      | 36 |
| 3.9.2  | Removal of the Linkers Tips . . . . .                                                                                                 | 36 |
| 3.9.3  | PAGE Purification . . . . .                                                                                                           | 36 |
| 3.10   | Stage 9: Concatenation . . . . .                                                                                                      | 37 |
| 3.10.1 | Appendix . . . . .                                                                                                                    | 38 |
| 4.     | Transcriptome and Genome Characterization<br>Using Massively Parallel Paired End Tag<br>(PET) Sequencing Analysis . . . . .           | 41 |
| 4.1    | Introduction . . . . .                                                                                                                | 42 |
| 4.2    | The Development of Pair end diTag (PET) Analysis . . . . .                                                                            | 43 |
| 4.3    | GIS-PET for Transcriptome Analysis . . . . .                                                                                          | 45 |
| 4.4    | ChIP-PET for Whole Genome Mapping of<br>Transcription Factor Binding Sites and Epigenetic<br>Modifications . . . . .                  | 48 |
| 4.5    | ChIA-PET for Whole Genome Identification of<br>Long Range interactions . . . . .                                                      | 52 |
| 4.6    | Perspective . . . . .                                                                                                                 | 55 |
| 5.     | New Era of Genome-Wide Gene Expression Analysis . . . . .                                                                             | 61 |
| 5.1    | Introduction . . . . .                                                                                                                | 62 |
| 5.2    | Tagging Technologies for Genome-Wide Analysis . . . . .                                                                               | 62 |
| 5.3    | Principles of Next Generation Sequencing<br>Technologies . . . . .                                                                    | 63 |
| 5.3.1  | Genome Sequencer 20/FLX System<br>(Roche Diagnostics/454 Life Sciences) . . . . .                                                     | 63 |
| 5.4    | Genome Analyzer (Illumina/Solexa) . . . . .                                                                                           | 65 |
| 5.5    | SOLiD System (Applied Biosystems) . . . . .                                                                                           | 67 |
| 5.6    | Advantages of Next Generation Sequencing<br>Technologies over Conventional Sequencing<br>Technology on Tagging Technologies . . . . . | 69 |

|       |                                                                                 |     |
|-------|---------------------------------------------------------------------------------|-----|
| 5.7   | From Static Analysis to Dynamic Analysis . . . . .                              | 70  |
| 5.8   | CAGE Method and Next Generation Sequencing<br>Technologies . . . . .            | 72  |
| 5.9   | Conclusions and Outlook . . . . .                                               | 73  |
| 6.    | Computational Tools to Analyze CAGE —<br>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,<br>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<br>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 |

|        |                                                                                                       |     |
|--------|-------------------------------------------------------------------------------------------------------|-----|
| 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 Information                  | 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 |

|        |                                                                              |     |
|--------|------------------------------------------------------------------------------|-----|
| 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 |

|                                                                                          |     |
|------------------------------------------------------------------------------------------|-----|
| 17.5 What does CAGE Data Offer for Traditional Studies of Promoter Regulation? . . . . . | 237 |
| 17.6 Conclusion . . . . .                                                                | 240 |
| <i>Color Index</i> . . . . .                                                             | 245 |
| <i>Index</i> . . . . .                                                                   | 265 |

## Y Hayashizaki

Osaka Science Center, RIKEN Yokohama Institute, Yokohama, Japan

E-mail: yoshihide@osc.riken.jp

Cap Analysis of Gene Expression (CAGE) targeting to identify promoters by our group, is used to perform large-scale analysis of promoters. Our group is currently engaged in large-scale CAGE experiments and has developed various CAGE technologies for this analysis. In addition, we have developed a high-throughput CAGE selection method known as the 5' RAGE method (see Chapter 20) that is one of the fundamental technologies for CAGE. In the CAGE, the 5'-CAP of mRNA is identified using this high-throughput RNA selection method, or 5' RAGE, step. From the 5' end of mRNA, which is a product of transcription, 20-base pairs are selected as tags using Type II restriction enzymes, and the tag sequence is determined by sequencing. These tags are then classified. We have recently developed a technique for analyzing these tags on a large scale. The obtained tags can identify promoters in the sequencing reads of organism genomes to the human genome, and by analyzing the expression frequency of the tags mapped to each promoter site, we can measure the transcription activity of promoters. Thus, the CAGE has been the main method for identifying promoters and determining their activity at the genome-wide level.

For the analysis of promoters in the genome, we are also constructing the gene network that comprises the promoters like a component at a molecular level. In this chapter, we present our recent functional annotation of Mammalian CAGE data in the genome (see Chapter 20).

Cap Analysis of Gene Expression (CAGE) and 5' RAGE: The 5' RAGE method

Transcriptome edited by P. Carninci

Copyright © 2010 by Pan Stanford Publishing Pte. Ltd.

www.panstanford.com

978-981-4241-34-2