HL116

Comparison of CAGE and RNA-seq transcriptome profiling using clonally amplified and single-molecule nextgeneration sequencing

Hideya Kawaji 1,2,3,5, Marina Lizio 2,3, Masayoshi Itoh 1,2,3, Mutsumi Kanamori, Katayama 2, Ai Kaiho 2, Hiromi Nishiyori-Sueki 2,3, Jay W. Shin 2,3, Miki Kojima, Ishiyama 2,3, Mitsuoki Kawano 2, Mitsuyoshi Murata 2,3, Noriko Ninomiya-Fukuda 2, Sachi Ishikawa-Kato 2,3, Sayaka Nagao-Sato 2, Shohei Noma 2,3, Yoshihide, Hayashizaki 1,2, Alistair R.R. Forrest 2,3, Piero Carninci 2,3,5

1 RIKEN Preventive Medicine and Diagnosis Innovation Program, Japan

2 RIKEN Omics Science Center, RIKEN Yokohama Institute, Japan

3 RIKEN Center for Life Science Technologies (CLST), Division of Genomic Technologies (DGT), Japan

Abstract

CAGE (cap analysis gene expression) and RNA-seq are two major technologies used to identify transcript abundances as well as structures. They measure expression by sequencing from either the 5′ end of capped molecules (CAGE) or tags randomly distributed along the length of a transcript (RNA-seq). Library protocols for clonally amplified (Illumina, SOLiD, 454 Life Sciences [Roche], Ion Torrent), secondgeneration sequencing platforms typically employ PCR preamplification prior to clonal amplification, while third-generation, single-molecule sequencers can sequence unamplified libraries. Although these transcriptome profiling platforms have been demonstrated to be individually reproducible, no systematic comparison has been carried out between them. Here we compare CAGE, using both second- and third-generation sequencers, and RNA-seq, using a second-generation sequencer based on a panel of RNA mixtures from two human cell lines to examine power in the discrimination of biological states, detection of differentially expressed genes, linearity of measurements, and quantification reproducibility. We found that the quantified levels of gene expression are largely comparable across platforms and conclude that CAGE and RNA-seq are complementary technologies that can be used to improve incomplete gene models. We also found systematic bias in the secondand third-generation platforms, which is likely due to steps such as linker ligation, cleavage by restriction enzymes, and PCR amplification. This study provides a perspective on the performance of these platforms, which will be a baseline in the design of further experiments to tackle complex transcriptomes uncovered in a wide range of cell types. The CAGE protocol based on single-molecule sequencer is employed in FANTOM5, a large scale transcriptome study, and this study also provide a technical baseline to understand the data set in depth.