An emerging landscape of transcriptome complexity

Piero Carninci
RIKEN Center for Integrative Medical Sciences

Gene regulatory elements are key to understand health and diseases. Mapping 5'-ends of RNAs is the key to understand the gene regulation as they identify promoters. In order to comprehensively understand regulatory elements, we developed the Cap Analysis of Gene Expression (CAGE) technology, which enables to identify transcription start sites (TSSs) and quantitatively measure their activity throughout the genome at high-throughput. In the RIKEN Functional Annotation of the Mammalian Genome 5 (FANTOM5) project, we created a very broad map of the promoterome and regulatory networks by simultaneously mapped mRNAs and lncRNAs TSSs and measured their expression at each different promoters with CAGE, on a comprehensive panel of human and mouse primary cells and other tissues. The study revealed the existence of 223,428 and 162,264 promoters and 65,423 and 44,459 enhancers, in human and mouse respectively, which are often tissue specific (Forrest et al. Nature 507, 462, 2014, Andersson et al. Nature 507, 455, 2014). The project also explored complexity of genome activation hierarchy in which transcription initiates with enhancer followed by promoter and then other genes (Arner et al. Science 347, 1010, 2015). Classification of lncRNAs revealed that most intergenic lncRNAs are derived from enhancer-like regions rather than classic promoters and GWAS trait-associated SNPs enriched at lncRNA loci were specifically expressed in cell-types relevant to the specific diseases, suggesting their roles in diseases (Hon et al. Nature 543, 199, 2017). Ongoing FANTOM6 is aiming at creating the broadest database of functional lncRNAs, as a valuable resource in the community.

Furthermore, we are pursuing a strategic collaboration with the International Human Cell Atlas (HCA) project with our newly developed single cell CAGE. The HCA is aiming at the creation of a comprehensive map of all human cell types and states at single cell level. The new types of data, images and models based on this precise reference of individual cells in the population would generate completely different perspectives of biology and bring new technologies for diagnosis and treatments of diseases.