RICC Usage Report for Fiscal Year 2012

Project Title:

Analysis of next generation sequencing data

Name: Timo Lassmann*, Akira Hasegawa*, Hiroko Ohmiya**

Laboratory:

*** Large-Scale Genome Analysis System Development Collaboration Unit,**
Industry Collaboration Group, RIKEN Omics Science Center, RIKEN Yokohama Institute
****DNAFORM Inc.**

1. Background and purpose of the project, relationship of the project with other projects.

The production of specific mRNAs by RNA polymerase II is regulated in most phases of homeostasis, growth, differentiation, and development in eukaryotes. Therefore, understanding the mechanisms of transcription initiation will enable us to interpret the functional elements of the genome and consequently reveal the molecular constituents of cells and tissues. Our goal is constructing the pipeline to automatically analysis high throughput next generation sequencing data sets.

2. Specific usage status of the system and calculation method

The parametric clustering method can define transcription start sites (TSSs) clusters with an arbitrary size and identify clusters within other clusters. Because it is obvious that clusters have complicated structures, the algorithm is more useful than the previously proposed methods. However, the method omits clusters with less than a particular tag count to reduce running time, which discards clusters with a small size but high density. Therefore, we adopted tags per million (TPM) as a threshold for clustering instead of the total tag count of a cluster. Additionally, we calculated hierarchical stability to evaluate the reproducibility of each cluster with the IDR method.

3. Result

It is demonstrated that our pipeline performed appropriate parametric clustering of CAGE data sets, which were not massively sequenced or abundantly sequenced. Therefore, we are confident in the versatility of the pipeline, regardless of the sequencing platform.

4. Conclusion

The preparation of CAGE libraries is relatively simple, and we developed a convenient pipeline for the use of this technology. Nucleotide codes, distribution, intensity, and structure of hierarchical clusters for transcription initiation obtained from CAGE technology and our pipeline provide us with clues to understand transcriptional regulatory networks.

5. Schedule and prospect for the future

We are going to develop a pipeline to uncover gene regulatory networks based on the information provided by our previous pipeline.