Usage Report for Fiscal Year 2019

Project Title:

Non-coding RNA structure

Name: ○Michiel de Hoon (1)
Laboratory at RIKEN:
(1) Center for Integrative Medical Sciences, Laboratory for Applied Computational Genomics

---

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

The Laboratory for Applied Computational Genomics is leading the bioinformatics analysis in the sixth edition of the Functional Annotation of the Mammalian Genome (FANTOM) project. As part of FANTOM6, we are studying the secondary structure of non-coding RNA using next-generation sequencing data that provide information on base-pairing interactions within the same RNA molecule and in between different RNA molecules. Additionally, as part of the previous edition of FANTOM (FANTOM5), we performed a comparative analysis of the coding and non-coding transcriptome in vertebrate genome to elucidate the key components of the transcriptional regulatory network that determine cellular identity.

2. Specific usage status of the system and calculation method

Usage status as reported by the listcpu command on hokusai is shown in the table below.

Resource unit	Limit (h)	Used (h)	Used (%)
gwmpc	3,027,456.0	231,213.8	7.6%
gwacsg	63,072.0	59,007.1	93.6%
gwacsl	10,512.0	0.0	0.0%
bwmpc	2,943,360.0	0.1	0.0%

3. Result

In human, more than 400 different cell types have been defined (Vickaryous and Hall, 2006), and many more are likely to be discovered by single-cell studies such as the Human Cell Atlas (Regev et al., 2017). While (with a few exceptions) different cell types have the same genome information content, their morphology, function, behavior, and life cycle is extremely diverse. These differences between cell types is thought to be due to differences in the transcriptional program running in each cell type, under the influence of transcription factors binding to regulatory sequence elements found in the DNA near promoters and enhancers.

While regulatory sequence elements can be identified by DNA sequence motif analysis, this is challenging because the sequence motifs are short, and therefore it is difficult to distinguish true regulatory sites from non-regulatory sites. We therefore used genome sequence conservation between organisms to identify candidate regulatory sites that were strongly conserved during evolution, which are more likely to be true regulatory DNA sites.

We used the GreatWave massively parallel computer (gwpmc) to perform pairwise alignments by the lastz (Harris, 2007) program of vertebrate genomes. Recently, we have also included recent genomes of primates in the alignments, which are particularly useful as they are highly similar to the human genome and can be used to identify DNA regulatory sites that are primate-specific. We then used the vertebrate genome alignments and analyzed FANTOM5 CAGE expression data to see the patterns of conservation and evolution of transcription and its regulation. Motif activity analysis (Suzuki et al., 2009) of the CAGE expression profiles together with the predicted DNA regulatory sites revealed that both promoters and enhancers tend to be regulated by the same transcription factor in different vertebrate species (figure), indicating the existence of

a core transcriptional regulatory program that has been conserved through evolution.

4. Conclusion

The computational resources provided by HOKUSAI allowed us to perform a systematic evolutionary analysis of transcriptional regulation in vertebrates, identifying a core regulatory network involved in specifying cellular identity.

5. Schedule and prospect for the future

The manuscript describing this analysis and its conclusions (Alam et al., 2020) is currently under revision and will be resubmitted before the end of the first quarter of 2020.

6. If no job was executed, specify the reason.

Not applicable.

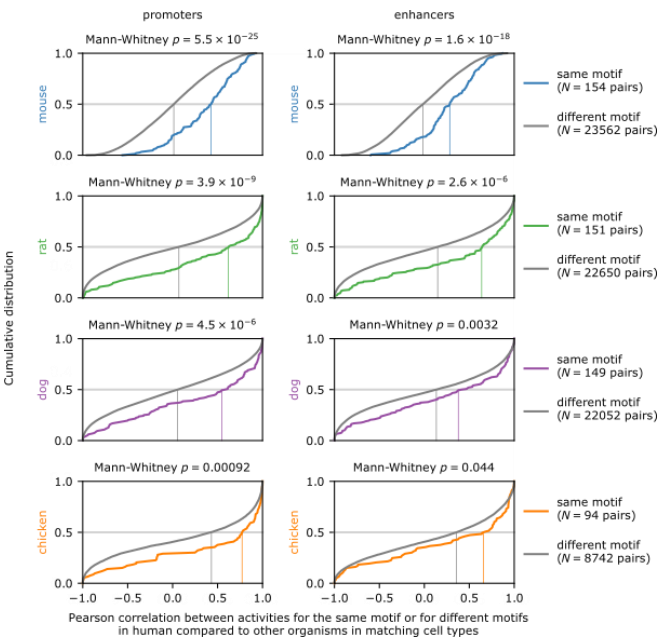

Figure. Analysis of FANTOM5 CAGE expression data and computationally identified DNA regulatory sites showed that both promoters and enhancers are regulated by the same transcription factors in different vertebrate species.

References

Alam, T., Agrawal, S., Severin, J., et al. 2020. Comparative transcriptomics of primary cells in vertebrates. Currently in revision for resubmission.

Harris, R.S. (2007) Improved pairwise alignment of genomic DNA. Ph.D. Thesis, The Pennsylvania State University.

Regev, A., Teichmann, S.A., Lander, E.S., et al. 2017. The human cell atlas. eLife 6.

FANTOM Consortium, Suzuki, H., Forrest, A.R.R.F., Van Nimwegen, E., et al. (2009) The transcriptional network that controls growth arrest and differentiation in a human myeloid leukemia cell line. Nature Genetics 41: 553-562.

Vickaryous, M.K. and Hall, B.K. 2006. Human cell type diversity, evolution, development, and classification with special reference to cells derived from the neural crest. Biological Reviews of the Cambridge Philosophical Society 81(3), pp. 425–455.