Usage Report for Fiscal Year 2018

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

Our laboratory applies computational methods to analyze genome and transcriptome data and elucidate regulatory interactions, including coding and non-coding RNA. As part of FANTOM5, we perform comparative studies to understand patterns of conservation of regulation of RNA transcription by comparing deep sequencing data of the human transcriptome to that in other organisms. In FANTOM6, we focus on the structure of non-coding RNA and the connection to its functional role in gene regulation.

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	3,010,867.2	577,110.0	19.2%
gwacsg	63,072.0	62,177.2	98.6%
gwacsl	10,512.0	0.0	0.0%
bwmpc	2,927,232.0	430.2	0.0%

3. Result

Gene regulation is orchestrated by the binding of transcription factors near the starting site of gene, as well as at distal regulatory sites, known as enhancers, by recognizing a sequence motif in the DNA. Previous work in FANTOM5 (Andersson et al. 2014) has shown that active enhancers are characterized by the transcription of non-coding RNAs known as enhancer RNAs.

As the sequence motifs are typically short, it is challenging to distinguish true transcription factor binding sites from non-regulatory sites with a spurious similarity to the motif. Additional information such as sequence conservation between organisms is therefore oftentimes taken into account to increase the prediction accuracy of motif detection software.

For our comparative study, we used the GreatWave massively parallel computer (gwpmc) to perform pairwise genome alignments of the rat, dog, and chicken genome against 29 other species. Using Jim Kent's UCSC genome browser bioinformatics utilities, genome sequences were split into segments and aligned using lastz. Alignment coordinates were corrected using blastz-normalizeLav and converted to .psl format using lavToPsl. Alignments were chained using axtChain and further processed using chainAntiRepeat, chainMergeSort, chainPreNet, chainNet, netSyntenic, and netFilter. The best alignments were extracted using netToAxt, followed by axtSort and axtToMaf to generate a .maf (multiple alignment format) file.

Next, we used the GreatWave massively parallel computer (gwpmc) to perform genome-wide transcription factor binding site predictions. For human and mouse, we used the genome-wide alignments provided by the University of California, Santa Cruz; for rat, dog, and chicken, we used our own genome alignments described above. We extracted the alignments for human, macaque, mouse, rat, dog, horse, cow, opossum, and chicken from the multiple genome alignments, divided them into segments, and ran the T-Coffee (Notredame et al. 2000) multiple sequence aligner version 9.01 on each segment. On each segment, we ran MotEvo (Arnold et al. 2012) for the 190 motifs in SwissRegulon (Pachkov et al. 2013); the MotEvo software identifies candidate transcription factor binding sites by searching for conserved motifs in the genome sequence; the SwissRegulon database

maintains a set of transcription factor motifs appropriate for the MotEvo software.
Additionally, we used the GreatWave Application Computing Server with GPU (gwacsg) to perform exploratory molecular dynamics analysis runs for non-coding RNA structure elucidation.

4. Conclusion

The genome-wide predictions of transcription factor binding sites for human, mouse, rat, dog, and chicken were used to analyze the conservation and evolution of transcriptome data obtained in FANTOM5. This analysis showed that both gene promoters and enhancers tend to be activated by the same transcription factors in human, mouse, rat, dog, and chicken, revealing a strong conservation of the core regulatory network between primary cells in different organisms. This has important implications for single-cell transcriptome studies such as those undertaken as part of the Human Cell Atlas (HCA), in which cell types are classified and novel cell types are identified based on their transcriptomic signatures.

5. Schedule and prospect for the future

We are currently preparing a manuscript summarizing our analysis, which we plan to submit within this fiscal year. The multiple genome alignments and genome-wide transcription factor binding site predictions produced in this project will be released to the scientific community as supplementary materials to the manuscript.
After publication, our focus will further shift to the analysis of the structure of RNA.

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

Not applicable.

References

Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., Suzuki, T., Ntini, E., Arner, E., Valen, E., Li, K., Schwarzfischer, L., Glatz, D., Raithel, J., Lilje, B., Rapin, N., Bagger, F.O. and Sandelin, A. 2014. An atlas of active enhancers across human cell types and tissues. Nature 507(7493), pp. 455–461.

Arnold, P., Erb, I., Pachkov, M., Molina, N. and Van Nimwegen, E. 2012. MotEvo: integrated Bayesian probabilistic methods for inferring regulatory sites and motifs on multiple alignments of DNA sequences. Bioinformatics 28(4), pp. 487–494.

Notredame, C., Higgins, D.G. and Heringa, J. 2000. T-Coffee: A novel method for fast and accurate multiple sequence alignment. Journal of Molecular Biology 302(1), pp. 205–217.

Pachkov, M., Balwierz, P.J., Arnold, P., Ozonov, E. and Van Nimwegen, E. 2013. SwissRegulon, a database of genome-wide annotations of regulatory sites: recent updates. Nucleic Acids Research 41(Database issue), pp. D214-20.