Clustering reveals ubiquitous heterogeneity and asymmetry of genomic signals at functional elements.

Anshul Kundaje[1], Sofia Kyriazopoulou-Panagiotopoulou[1], Max W. Libbrecht[1], Cheryl L. Smith[2], Debasish Raha[3], Elliott E. Winters[4], Steven M. Johnson[4], Michael Snyder[5], Serafim Batzoglou[2], Arend Sidow[5]
[1] Stanford CS, [2] Stanford Pathology, [3] Yale Molecular, Cellular, and Developmental Biology, [4] Brigham Young University Department of Microbiology and Molecular Biology, [5] Stanford Genetics

ABSTRACT

The advent of high-throughput DNA sequencing has enabled the measurement of many types of chemical phenomena in the human genome [2]. These methods have enabled the location of both punctate phenomena, such as transcription factor binding sites (TFBSs), and broad phenomena, such as the location of chemical modification of the histone proteins that wrap DNA. Experiments measuring broad phenomena are typically processed into real-valued signal tracks defined across every base pair.

A popular and highly effective method for visualizing and quantifying the relationship between a punctate phenomenon and a genomic signal is the so-called aggregation plot (AP). In a typical AP, the signal around several predefined anchor sites (such as TFBSs) is averaged for each position within a window around the sites. If the signal behaves similarly across the anchor sites, the AP will reveal these common signal patterns.

For an AP to display signals that are asymmetric about the anchor, the alignment of features has to be robust, and some other data is utilized to provide the correct orientation. Features that can be aligned, but for which there exists no external information regarding their orientation (e.g. TFBSs), can produce APs with strong but obligatorily symmetric signals. We developed a hierarchical agglomerative clustering strategy for this data which may reverse the orientation of sites in order to merge clusters that are mirror images of one another. This analysis revealed that asymmetries of chromatin signals are a pervasive feature at TFBSs, not only within genes but, surprisingly, equally strongly at gene-distal sites [1].

BODY

Aggregation plots cannot represent asymmetrical signals around sets of unoriented sites. Clustering is necessary for such analysis.

REFERENCES

[1] A. Kundaje, S. Kyriazopoulou-Panagiotopoulou, M. Libbrecht, C. L. Smith, D. Raha, E. E. Winters, S. M. Johnson, M. Snyder, S. Batzoglou, and A. Sidow. Ubiquitous heterogeneity and asymmetry of the chromatin environment at regulatory elements. *Genome Research*, 22(9):1735–1747, 2012.

[2] The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489:54–74, 2012.

Volume 2 of Tiny Transactions on Computer Science