PREreview from the Computational Biology & Gene Regulation group at NCMM

Anthony Mathelier[1]

[1]University of Oslo

April 28, 2020

This is a preprint review from our group's journal club. We reviewed the following manuscript:

CREAM: Clustering of genomic REgions Analysis Method

by Seyed Ali Madani Tonekaboni, Parisa Mazrooei, Victor Kofia, Benjamin Haibe-Kains, Mathieu Lupien

doi: https://doi.org/10.1101/222562

Review

Tonekaboni et al. presents CREAM, a new unsupervised method to predict clusters of pre-defined genomic regions. The method is innovative and will be of high interest to the research community. While the manuscript is well written overall and enjoyable to read, the description of the method should be improved. Moreover, the terminology used is confusing or wrong in several places and particular attention should be given to this. The authors applied CREAM to DNase hypersensitive sites and obtained clusters of cis-regulatory elements (CORES) that they compared to predictions from the ROSE method, a previous method developed to predict clusters of enhancers (aka super-enhancers). While the results are of biological relevance and show the usefulness of the method, the comparison with ROSE seems unfair to us. Specific details about our concerns are listed below.

Major concerns:

1. We encourage the authors to clarify the usage of the term CORE in the manuscript. Specifically, the abstract reads that COREs are aka super-enhancers while the background section presents super-enhancers as an example of COREs, while COREs can also be clusters of open chromatin regions. The lack of a clear definition of COREs make the reading of the manuscript confusing. We would suggest the others to avoid referring to COREs as super-enhancers in the abstract since COREs are not restricted to super-enhancers. Regarding terminology, the authors used the term transcription factor (TF) for a collection of transcriptional regulators (TRs), not only TFs. This is a recurrent issue in the field. Please refer to the nice work done in Lambert et al., 2018, Cell, which provides an extensive list of TFs; you will see that TRs such as SMC1, RAD21, and Pol2 are definitely not TFs.

2. Along the same lines, the authors used CREAM to predict clusters of open chromatin regions as COREs and compare the results to predictions from ROSE on the same data, which was specifically developed to predict only a subset of COREs: super-enhancers. This is extremely important since the vast majority (>80%) of the identified COREs by CREAM and ROSE are located at promoters (Supp. Fig. 2). We feel that this makes the comparison a bit unfair. In the manuscript describing ROSE [13], it has been shown that one should use Med1 ChIP-seq to define super-enhancers, or H3K27ac marks when Med1 is not available. It raises the following concerns:

a) The authors used ROSE to predict clusters of open regions while it has not been designed for this.

b) Using open chromatin regions will provide a mixture of cis-regulatory elements with distinct function (e.g. enhancers, promoters, insulators, silencers, chromatin conformation anchors).

We would recommend that the authors apply CREAM and ROSE to H3K27ac data for a more fair comparison. The authors should then emphasize the fact that CREAM is not restricted to the identification of clusters of enhancers but can be applied to the clusterization of any pre-defined genomic elements.

3. In the background section, the authors list 3 limitations associated previous methods developed to identify COREs, (i) fixed stitching distance between CREs, (ii) fixed cutoff in ChIP-seq signal, and (iii) identification of individual CREs as potential COREs. The manuscript reads that CREAM addresses all 3 limitations while previous methods do not. Regarding (i), the most up-to-date version of ROSE (ROSE2) does not assume a pre-defined stitching distance anymore as it computes automatically this distance. Moreover, CREAM also computes automatically a threshold (MWS) to consider CREs in the same CORE.

4. The description of the algorithm used in CREAM could be improved. First, the methods section lists 5 steps while Fig. 1 provides 4 steps. Second, we did not understand steps 3 to 5 and how they are illustrated in Fig. 1. We hope this section could be rewritten for more clarity. Finally, we believe it would be nice to start the Results section by an overview of the CREAM method referring to Fig. 1 to use the reading of the manuscript. That would also remove the reference to Fig. 1 in the background section as it is already a result. Lastly on this point, the authors should explicitly write that the identification of the MWS is based on the search for window sizes that represent outliers in the distribution of the sizes (since the formula Q1 - 1.5 * IQ is specific to the identification of outliers).

6. It is concerning to us to see that ROSE-identified COREs get a lower DNase signal for GM12878. Indeed, ROSE COREs should have been defined based on higher signal intensity. By the way, the text reads that it is an increase of signal for ROSE COREs when compared to individual COREs, where it should read that it is a decrease. Also, the comparison of signal intensity should be performed by excluding CRE-free regions as it was nicely done for the binding intensity of TRs.

7. The authors highlight some TRs as enriched at COREs and situated at TAD boundaries. When doing such analyses from ChIP-seq peaks, one has to keep in mind that binding motifs associated to CTCF, ZNF143, and others (called zinger motifs) are significantly enriched in the majority of ChIP-seq data sets from ENCODE (see Hunt and Wasserman, 2014, Genome Biol). This suggests that regions containing zinger motifs are not TF-specific and might reflect chromatin organization for instance. Hence, the authors should consider this aspect in their analyses and at least discuss it in their manuscript as some of the signal that they observed can be a reflection of this phenomenon.

Minor comments:

1. The authors argue that they chose GM12878 and K562 cell lines for in depth analyses due to the large amount of data available for these cell lines. It is due to the fact that both are ENCODE Tier 1 cell lines. We would recommend to also include H1-hESC as it is the third Tier 1 cell line from ENCODE.

2. In the Methods section, the authors should provide links (URLs) to *all* the data that they used to ease reproducibility instead of simply referring to the main ENCODE paper.

3. In the overlap comparison between ROSE and CREAM, the authors should point out the fact that the overlaps do not provide a one to one relationship as a CORE from ROSE for instance can overlap multiple COREs from CREAM.

4. Please provide the regions identified by CREAM as BED files.

5. How did the authors select the distance of +/- 100kb to assign COREs to genes? How much would the results change with this distance threshold?

6. Fig. 5E is missing the x-axis to ease the reading of the figure.

7. There is a typo for the Matthew*s* correlation coefficient (MCC).

8. Can the authors describe how they computed MCCs for individual COREs (Fig. 6D).

9. The author used random genomic regions to show the enrichment of COREs of CTCF, RAD21, SMC3, and ZNF143. We suggest to match the %GC composition of the original regions for this analysis as it might impact their genomic distributions.

10. To reinforce the functional importance of CREAM-derived COREs vs ROSE-derived ones, we suggest to look at the enrichment of these regions for GWAS SNPs and a GO functional enrichment analysis of nearby genes. These analyses are of course not mandatory but might help to further highlight the functional importance of CREAM-specific COREs.

11. In the identification of master TF you mentioned 'more than 25% of transcription factors show binding intensity significantly higher', please mention the number (or at least the total number of TFs analyzed).

12. Fig. 5D: the figure shows the median, the text refers to the mean.

13. Fig. 5E: Explicitly refer to the Jaccard similarity in the axis legend.

14. No information is provided in the Methods about how the genomic distribution of COREs in Supp. Fig. 2 has been performed (what is the set and sizes of promoters used for instance).

15. How is Supp. Fig. 5 showing the independence between the difference in size of individual CREs vs COREs and the enrichment of COREs at TAD boundaries?

16. No caption is provided for the supplementary table and figures.

17. Could you please provide the supplementary table in a .csv or alike format instead of within a PDF?

18. Typo (Enrichment of COREs in TAD boundaries): many times you said 'enrichment' when you should say 'enriched'.

19. Enrichment of COREs in TAD boundaries: you mention that COREs are significantly enriched, but you didn't provide a p-value (or another value) to support such statement.