bioRxiv preprint doi: https://doi.org/10.1101/222562; this version posted November 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

CREAM: Clustering of genomic REgions Analysis Method

Seyed Ali Madani Tonekaboni[1,2,$], Parisa Mazrooei[1,2], Victor Kofia[1], Benjamin Haibe-Kains[1,2,3,4,$], Mathieu Lupien[1,2,4,$]

[1]Princess Margaret Cancer Centre, Toronto, Ontario M5G 1L7, Canada

[2]Department of Medical Biophysics, University of Toronto, Toronto, Ontario M5G 1L7, Canada

[3]Department of Computer Science, University of Toronto, Toronto, Ontario M5T 3A1, Canada

[4]Ontario Institute for Cancer Research, Toronto, Ontario M5G 1L7, Canada

[$]Corresponding author: Seyed Ali Madani Tonekaboni: ali.madanitonekaboni@mail.utoronto.ca; Benjamin Haibe-Kains: benjamin.haibe.kains@utoronto.ca; Mathieu Lupien: mlupien@uhnres.utoronto.ca

Keywords: Cis-Regulatory Element, Clustering, Cluster of cis-regulatory elements, ENCODE, DNase I, Chromatin accessibility, CORE, Transcription factor, Cell identity, Chromatin, Super-enhancer, ROSE, Cancer, Transcriptional regulation, Essentiality.

bioRxiv preprint doi: https://doi.org/10.1101/222562; this version posted November 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

ABSTRACT

Cellular identity relies on cell type-specific gene expression profiles controlled by cis-regulatory elements (CREs), such as promoters, enhancers and anchors of chromatin interactions. CREs are unevenly distributed across the genome, giving rise to distinct subsets such as individual CREs and Clusters Of cis-Regulatory Elements (COREs), also known as super-enhancers. Identifying COREs is a challenge due to technical and biological features that entail variability in the distribution of distances between CREs within a given dataset. To address this issue, we developed a new unsupervised machine learning approach termed Clustering of genomic REgions Analysis Method (CREAM). We demonstrate that COREs identified by CREAM are predictive of cell identity, consists of CREs strongly bound by master transcription factors according to ChIP-seq signal intensity and are proximal to highly expressed genes. We further show that COREs identified by CREAM are preferentially found near genes essential for cell growth. Overall, CREAM offers an improved method compared to the state-of-the-art to identify COREs of biological function. CREAM is available as an open source R package (https://CRAN.R-project.org/package=CREAM) to identify COREs from cis-regulatory annotation datasets from any biological samples.

bioRxiv preprint doi: https://doi.org/10.1101/222562; this version posted November 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

BACKGROUND

Over 98% of the human genome consists of sequences lying outside of gene coding regions that harbor functional features, including cis-regulatory elements (CREs), important in defining cellular identity [1]. CREs such as enhancers, promoters and anchors of chromatin interactions, are predicted to cover 20-40% of the noncoding genomic landscape [2]. CREs define cell type identity by establishing lineage-specific gene expression profiles [3–5]. Current methods to annotate CREs in any given biological sample include ChIP-seq for histone modifications (e.g., H3K27ac, H3K4me3, H3K4me1) [3,4,6], for chromatin binding protein (e.g., MED1, P300) [6,7] or through chromatin accessibility assays (e.g., DNase-seq, ATAC-seq) [8,9].

Clusters Of cis-Regulatory Elements (COREs) were recently introduced as a subset of CREs based on different parameters including close proximity to each other [7,10–12]. COREs are significantly associated to cell identity and are bound with higher intensity by transcription factors than individual CREs [7,11,13]. Furthermore, inherited risk-associated loci preferentially map to COREs from disease related cell types [10,14–16]. Finally, COREs found in cancer cells lie proximal to oncogenic driver genes [17–19]. Together, these features showcase the utility of classifying CREs into individual CREs versus COREs.

Recent work assessed the role of COREs as a collection of individual CREs proximal to each other as opposed to a community of synergizing CREs [20–22]. Partial redundancy between effect of individual CREs versus a super-enhancer/CORE on regulating expression of genes in embryonic stem cells was observed [20] as well as low synergy between the individual CREs within COREs [23]. Whether COREs provide an added value over individual CREs to gene expression is still debated. Conclusions may be confounded by the simplistic approach commonly used to identify COREs. For instance, the distance between CREs is a critical feature that distinguishes COREs from individual CREs. Available methods to identify COREs dismiss

bioRxiv preprint doi: https://doi.org/10.1101/222562; this version posted November 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

the variability in the distribution of distances between CREs that stems from technical and biological features unique to each CRE dataset. Instead, arbitrary thresholds are considered including 1) a fixed stitching distance limit between CREs (such as 12.5 [7] or 20 [12] kilobases) to report them within a CORE, 2) a fixed cutoff in the ChIP-seq signal intensity from the assay used to identify CREs to separate COREs from individual CREs [7], or 3) reporting an individual CRE with high signal intensity as a CORE [7,24]. To address these limitations, we developed a new methodology termed CREAM (Clustering of genomic REgions Analysis Method) (Fig. 1). CREAM is an unsupervised machine learning approach that takes into account the distribution of distances between CREs in a given biological sample.

Benchmarking CREAM against Rank Ordering of Super-Enhancers (ROSE) [11], the current standard method to call COREs, we demonstrate that CREAM identifies COREs predictive of cell identity, proximal to highly expressed genes and associated with high intensity transcription factor binding. We further demonstrate the utility of COREs identified by CREAM as chromatin regions associated with genes essential for the growth of cancer cells.

bioRxiv preprint doi: https://doi.org/10.1101/222562; this version posted November 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

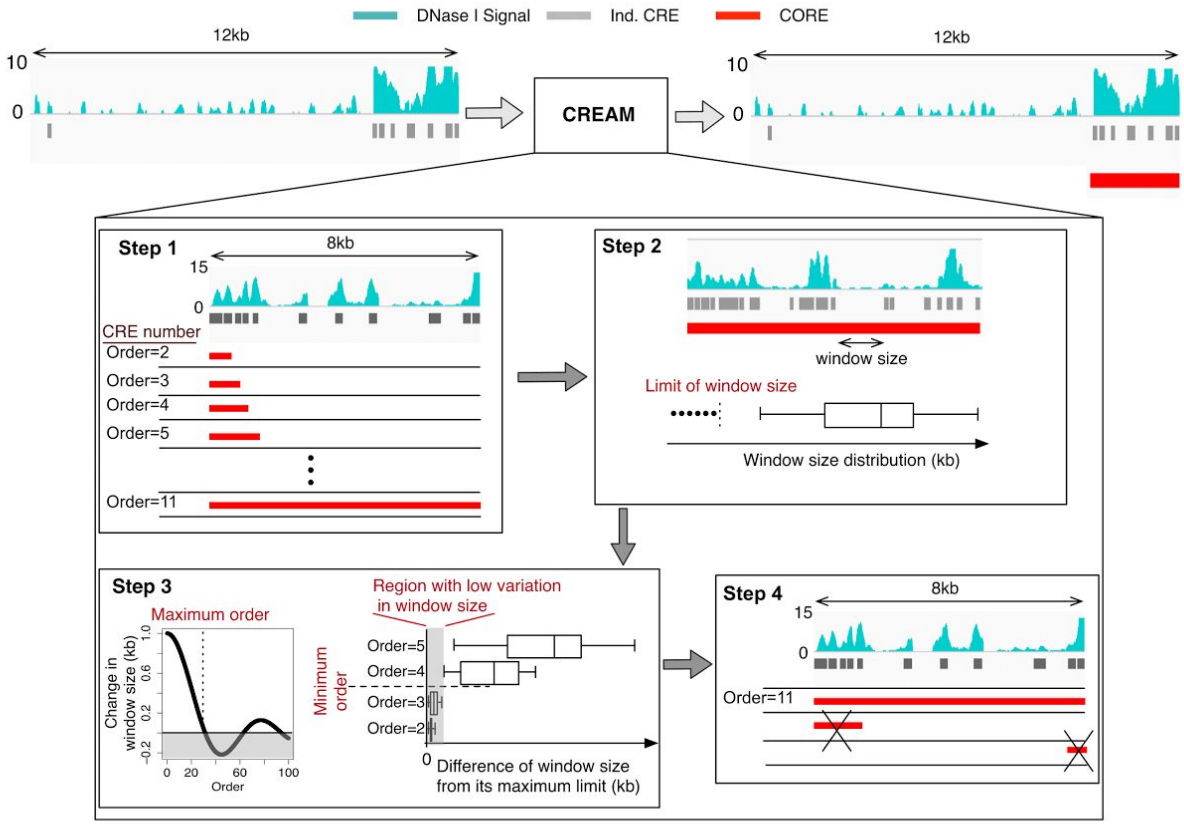


Figure 1. Schematic representation of the four main steps of Clustering of genomic REgions Analysis Method (CREAM): *Step 1) CREAM identifies all clusters of 2, 3, 4 and more neighboring CREs. The total number of CREs in a cluster defines its "Order"; Step 2) Identification of the maximum window size (MWS) between two neighboring CREs in clusters for each Order. The MWS corresponds to the greatest distance between two neighboring CREs in a given cluster; Step 3) identification of maximum and minimum Order limits of COREs from a given dataset; Step 4) CORE reporting according to the criteria set in step 3 from the highest to the lowest Order.*

RESULTS

We compared the number and width of the COREs identified by CREAM and ROSE

bioRxiv preprint doi: https://doi.org/10.1101/222562; this version posted November 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

using GM12878 and K562 cell lines. We focused on these cell lines because of their extensive characterization by the ENCODE project for DNA-protein interactions, namely ChIP-seq profiles for over 80 transcription factors. This provides a unique opportunity to assess the biological relevance of CORE identified in each cell line. CREAM identified a total of 1,694 and 4,968 CORE in GM12878 and K562 cell lines, respectively, based on their DNase-seq defined CREs. These CORE account for 14.6% and 17.2% of all CREs reported by the DNase-seq profiles from these cells. In contrast, ROSE identifies 2,490 and 2,527 CORE in GM12878 and K562, respectively. These account for 31% and 30% of the CREs detected in GM12878 and K562 cell lines. To determine if CREAM identifies new CORE or simply subdivides those reported by ROSE, we assessed the exclusivity of the CORE identified by CREAM and ROSE in GM12878 or K562 cell lines. The CREAM-identified CORE have 85% and 49% shared genomic regions with ROSE-identified CORE in GM12878 and K562 cell lines, respectively (Fig. 2A). However, shared genomic regions between ROSE and CREAM-identified CORE account only for 14% and 8% of the total identified CORE by ROSE in GM12878 and K562 cell lines, respectively. Hence, while many CORE are identified by both methods, CREAM and ROSE differ sufficiently that a number of CORE are uniquely identified by each method. Moreover, ROSE-identified CORE occupy significantly larger genomic regions (average 138 kb width) than those identified by CREAM (average 5kb width) (Fig. 2B). We therefore compared CORE identified by CREAM and ROSE according to a series of biological features previously shown to discriminate CORE from individual CREs.

DNase I hypersensitive signal is elevated within CORE

CORE are reported to associate with higher levels of binding for a wide range of chromatin binding proteins [11]. Our results show that CORE identified by CREAM have 2 to 5

bioRxiv preprint doi: https://doi.org/10.1101/222562; this version posted November 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

fold higher average DNaseI hypersensitivity signal per base pair (bp) compared to individual CREs (Fig. 2C). This is in contrast to COREs identified by ROSE, which show an equivalent DNaseI hypersensitivity to individual CREs in K562 cells and less than a 1.5 fold increased in GM12878 cells (Fig. 2C). The distinct behavior between CREAM and ROSE-identified COREs could be due to their size difference that translates in more base pairs free of CRE (CRE-free gaps) in ROSE-identified COREs (102 mbp and 208 mbp in GM12878 and K562 cells, respectively) compared to CREAM-identified COREs (14.7 mbp and 18.7 mbp in GM12878 and K562 cells, respectively) (Fig. 2D). This stems from a permissive <12.5kb distance between CREs criteria in ROSE. In contrast, a learned maximum distance limit between CREs criteria is used by CREAM resulting in smaller average of maximum distances (<1.7 kb in GM12878 and <1 kb in K562 cells for their respective DNase-seq delineated CREs).

bioRxiv preprint doi: https://doi.org/10.1101/222562; this version posted November 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

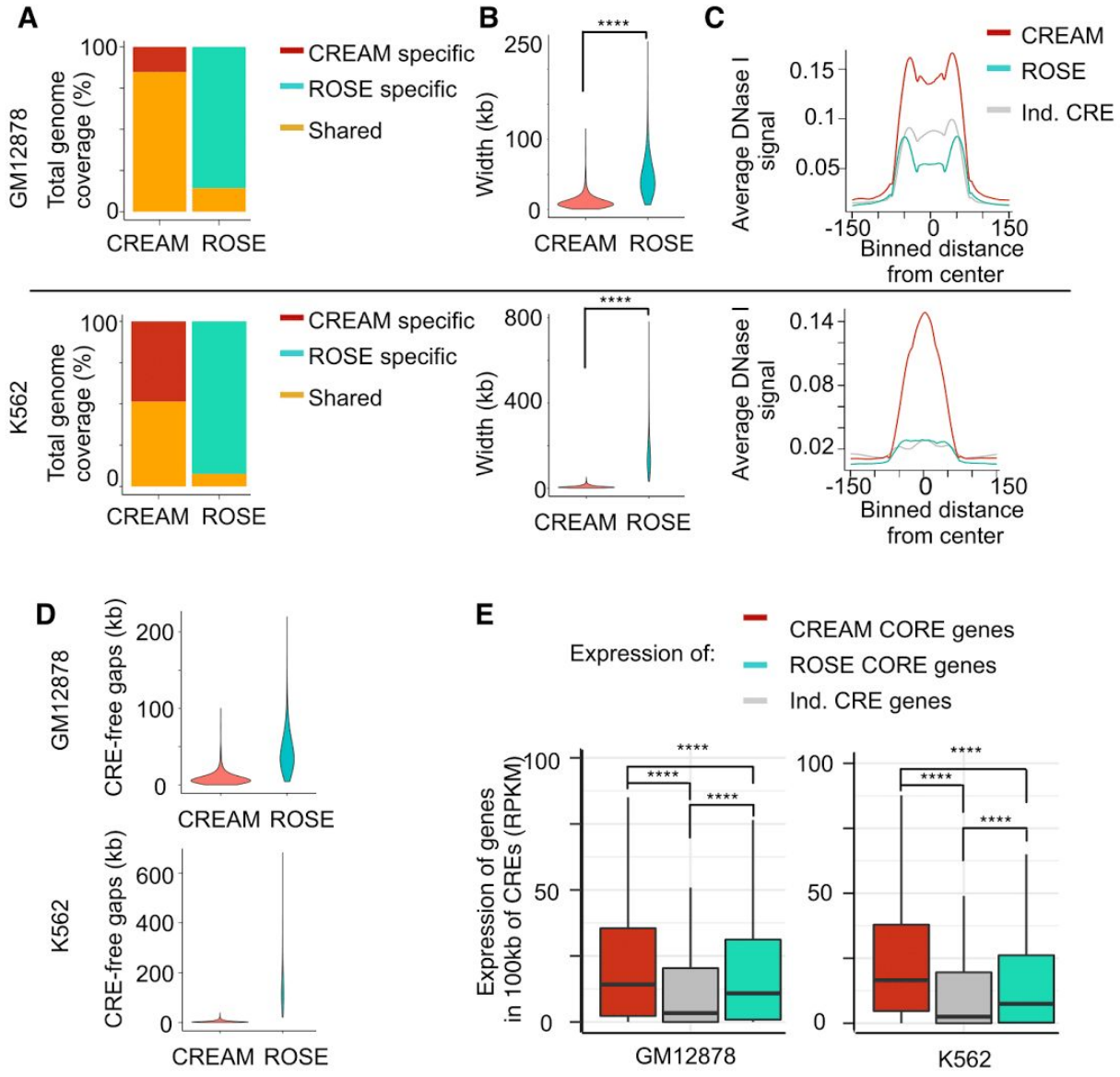


Figure 2. Comparison of genomic characteristics of the COREs identified by CREAM versus ROSE in GM12878 and K562 cell lines. ***(A)*** *Percentage of the identified COREs by CREAM and ROSE which are unique or shared regarding the genomic coverage. (**B**) Distribution of CORE widths. (**C**) Enrichment of DNase I signal profile in individual CREs and COREs. E*ach CORE (or individual CREs) plus its flanking regions are binned into 300 binned regions in total (100 bins each). *(**D**) Distribution of CRE-free gaps within COREs. (**E**) Transcription of associated genes to individual CREs and COREs. (**F**) Enrichment of number of*

bioRxiv preprint doi: https://doi.org/10.1101/222562; this version posted November 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

*essential genes among the associated genes to individual CREs and COREs in K562 cell line. (**G**) Transcription level of essential genes which are associated with individual CREs and COREs.*

CREAM-identified COREs are proximal to highly expressed genes.

In agreement with previous reports [7,11], COREs identified by ROSE are proximal to genes expressed at higher levels than those near individual CREs (Fig. 2E). This also applies to COREs identified by CREAM in both GM12878 (>4 fold difference) and K562 cell lines (>2.5 fold difference)(Fig. 2E). Noteworthy, genes proximal to CREAM-identified COREs have a 1.5 fold higher expression levels compared to genes proximal to ROSE-called COREs ($p < 0.001$) (Fig. 2E). We further assessed expression of genes in proximity of COREs specific to CREAM and ROSE. Expression of genes in proximity of CREAM-specific COREs were significantly higher than genes in proximity of ROSE-specific COREs in both GM12878 and K562 cell lines ($p<0.001$) (Supplementary Fig. 1). Moreover, comparing the percentage of COREs which overlap with promoters, exons, introns, and intergenic regions reveals a very similar distributions for COREs identified by CREAM or ROSE in GM12878 and K562 cell lines. (Supplementary Fig. 2). Taken together, our results show that COREs identified by CREAM share similarities with those identified by ROSE in term of genomic distribution but are associated with stronger differences in gene expression compared to individual CREs.

bioRxiv preprint doi: https://doi.org/10.1101/222562; this version posted November 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

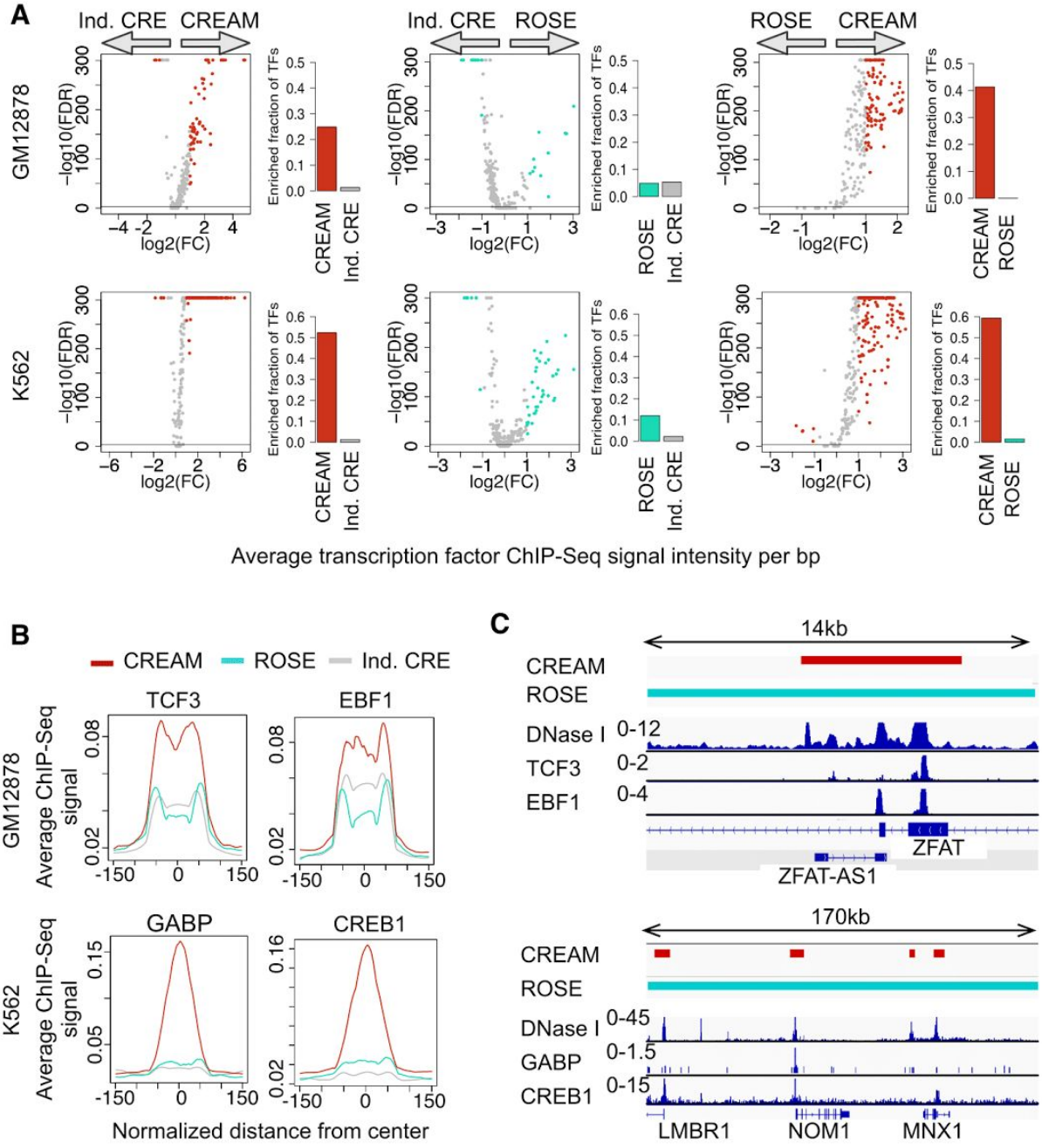


Figure 3. Transcription factor (TF) binding enrichment in identified individual CREs and COREs for GM12878 and K562 cell lines. A) *Comparison of enrichment of TF binding intensity in the identified COREs by CREAM or ROSE or the individual CREs. Volcano plots represent -log$_{10}$(FDR) versus log$_2$(fold change [FC]) of comparison of signal intensities*

bioRxiv preprint doi: https://doi.org/10.1101/222562; this version posted November 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

comparing COREs and individual CREs (each dot is one transcription factor). The barplots also shows how many transcription factors have higher signal intensity with FDR<0.001 and log_2(FC)>1 comparing COREs and CREs. Fold change (FC) is defined as division of average signal per base pair in COREs versus CREs or CREAM COREs versus ROSE COREs. ***B****) Normalized signal intensity for TCF3 and EBF1 as examples of master TFs in GM12878 [25] and for GABP and CREB1 as examples of master TFs for K562 [26–28].* ***C****) Examples of genomic regions which are identified by both CREAM and ROSE (with different coverage) and enriched for the illustrated TFs in panel (****B****).*

CREAM identifies COREs bound by master transcription factors.

Transcription factors bind to CREs to modulate the expression of cell-type specific gene expression patterns [29,30]. COREs were previously found to associate with strong transcription factor binding intensity based on ChIP-seq signal [11]. Hence, we assessed transcription factors binding intensities within COREs using the extensive characterization of transcription factor binding profiles performed by the ENCODE project in GM12878 and K562 cells [31]. We find that more than 25% of transcription factors show binding intensity significantly higher over CREAM-identified COREs compared to individual CREs in both GM12878 and K562 cell lines (FC > 2; FDR < 0.001) (Fig. 3A). In contrast, less than 15% of all transcription factors bind with higher intensity in ROSE-identified COREs compared to individual CREs in both GM12878 and K562 cell lines (FDR < 0.001, FC>2; Fig. 3A).

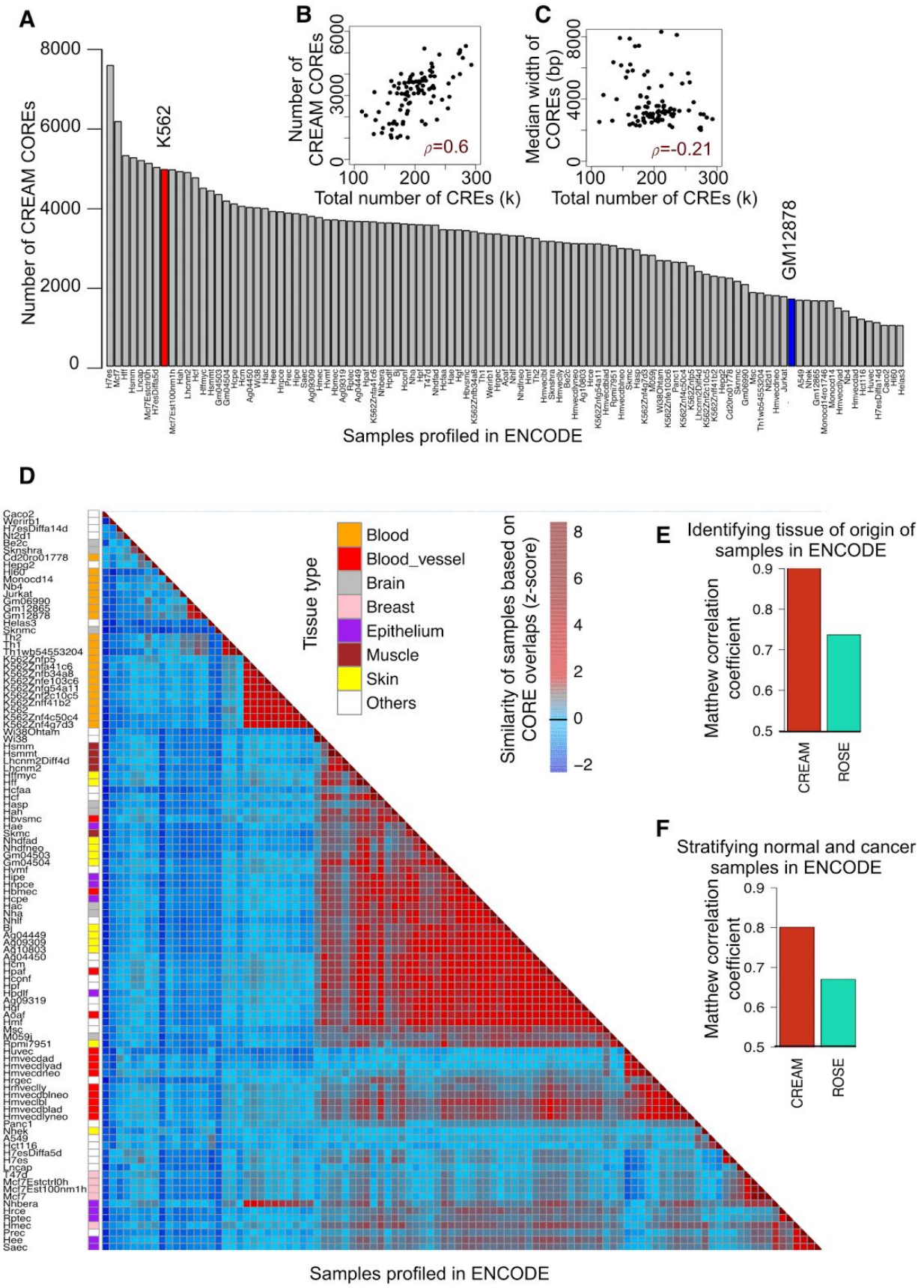
Difference in transcription factor binding intensity at CREAM versus ROSE-identified COREs is showcased by the master transcription factors TCF3 and EBF1 [25] in GM12878 cells and GABP and CREB1 [26–28] in K562 cells. Indeed, over a 2 fold difference in binding intensity of TCF3 and EBF1 is observed for CREAM-identified COREs compared to individual

bioRxiv preprint doi: https://doi.org/10.1101/222562; this version posted November 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

CREs in GM12878 cell (Fig. 3B). This is exemplified over the CORE proximal to the ZFAT gene in GM12878 cells (Fig. 3C). Similarly, over a 3 fold difference in GABP and CREB1 binding intensity is observed over COREs compared to individual CREs in K562 cells (Fig. 3B) and exemplified at the 7q36 locus harboring a series of COREs bound strongly by GABP and CREB 1 in K562 cells (Fig. 3C).

The binding intensity of transcription factors over COREs was calculated as the average ChIP-seq signal within each CORE. We assessed if the difference between enrichment of transcription factor binding intensity within CREAM- and ROSE-identified COREs is not merely due to the difference in their burden of CRE-free gaps. We calculated the transcription factor binding intensity excluding the CRE-free gaps within ROSE-identified COREs (Supplementary Fig. 3). More than 25% of the transcription factors have significantly higher binding intensity within CREAM- identified COREs compared to the signal over the CREs in COREs identified by ROSE ($FC > 2$; $FDR < 0.001$).

bioRxiv preprint doi: https://doi.org/10.1101/222562; this version posted November 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.



bioRxiv preprint doi: https://doi.org/10.1101/222562; this version posted November 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Figure 4. Specificity of COREs to the phenotype of the cell lines in ENCODE. (A) *Distribution of number of COREs identified for the available DNase I samples in ENCODE.* ***(B)*** *Positive correlation of number of COREs with number of individual CRE in the samples. (**C**) Relation between median width of the COREs and number of individual CRE. (**D**) Heatmap of similarities of the ENCODE samples based on Jaccard index of overlap of their corresponding identified COREs by CREAM. (**E**) Matthew correlation coefficient (MCC) for classification of ENCODE samples as normal or cancerous using CREAM-versus ROSE-identified COREs. (**F**) Matthew correlation coefficient for classification of ENCODE samples based on their tissue of origin using CREAM- versus ROSE-identified COREs.*

Generalizability of CORE identification across cell and tissue types

ROSE-identified COREs were reported to discriminate cell types [7]. We therefore assessed the predictive value of CREAM-identified COREs to discriminate cellular identity. Running CREAM on the DNase-seq defined CREs from 102 cell lines provided by the ENCODE project [31] reveals between 1,022 to 7,597 COREs per cell line (Fig. 4A). The number of COREs correlates with the total number of CREs identified in each cell line (Fig. 4B). However, the average width of COREs across cell lines shows low correlation with the total number of CREs (|Spearman correlation| $\rho < 0.25$; Fig. 4C). Hence, CORE widths are specific to each biological sample irrespective of the total number of CREs.

To test whether COREs can discriminate cells with respect to their tissue source and their malignant status, we clustered the ENCODE cell lines based on their CREAM- and ROSE-identified COREs (Fig. 4D). Predicting each cell line based on its nearest neighbor, we could classify tissue source with high accuracy using CREAM-identified COREs (Matthew correlation coefficient [MCC] of 0.90 for tissues with ≥ 5 cell lines; Fig. 4E). However,

bioRxiv preprint doi: https://doi.org/10.1101/222562; this version posted November 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

ROSE-identified COREs yielded substantially lower predictive value (MCC of 0.74; Fig. 4E). Similarly, CREAM-identified COREs were more discriminative of non-malignant versus malignant cell lines than ROSE-identified COREs (MCC of 0.80 versus 0.67 for CREAM and ROSE, respectively; Fig. 4F).

CREAM-identified COREs are proximal to essential genes.

COREs are reported to lie in proximity to genes essential for self-renewal and pluripotency of stem cells, respectively [32]. A CRISPR/Cas9 gene essentiality screen was recently reported in K562 cells by *Wang et al. (2015)* [33]. Merging these genomic screening data with CORE identification from K562 cells reveals a significant enrichment of gene essential for growth proximal to CREAM-identified COREs (FDR < 1e-4; Fig. 5A). BCR is the top essential gene in proximity of CREAM-identified COREs. Oncogenic BCR-ABL gene fusion plays an essential role in pathogenesis of Chronic Myelogenous Leukemia which is the tumor of origin of K562 cell line [34].

In contrast, genes proximal to individual CREs or ROSE-identified COREs are not enriched with essential genes (CRE: FDR=0.26; ROSE-identified CORE: FDR=0.92; Fig. 5B). Moreover, expression of genes essential for growth in K562 proximal to CREAM-identified COREs is significantly higher than expression of the essential genes associated with individual CREs or ROSE-identified COREs ($p < 0.001$; Fig. 5C). Hence, CREAM identifies COREs associated with essential genes in K562 cell line. To further assess the specificity of COREs' association with essential genes we extended our analysis to essentiality score from other model cell lines tested by *Wang et al. (2015)* [33]. Essentiality score of genes proximal to K562 CREAM-identified COREs in KBM-7, Jiyoye, and Raja cell lines were significantly less negative

bioRxiv preprint doi: https://doi.org/10.1101/222562; this version posted November 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

than for the genes proximal to the COREs in K562 cells (FDR<0.001; Fig. 5D). This supports the cell type-specific nature of COREs and their association with essential genes.

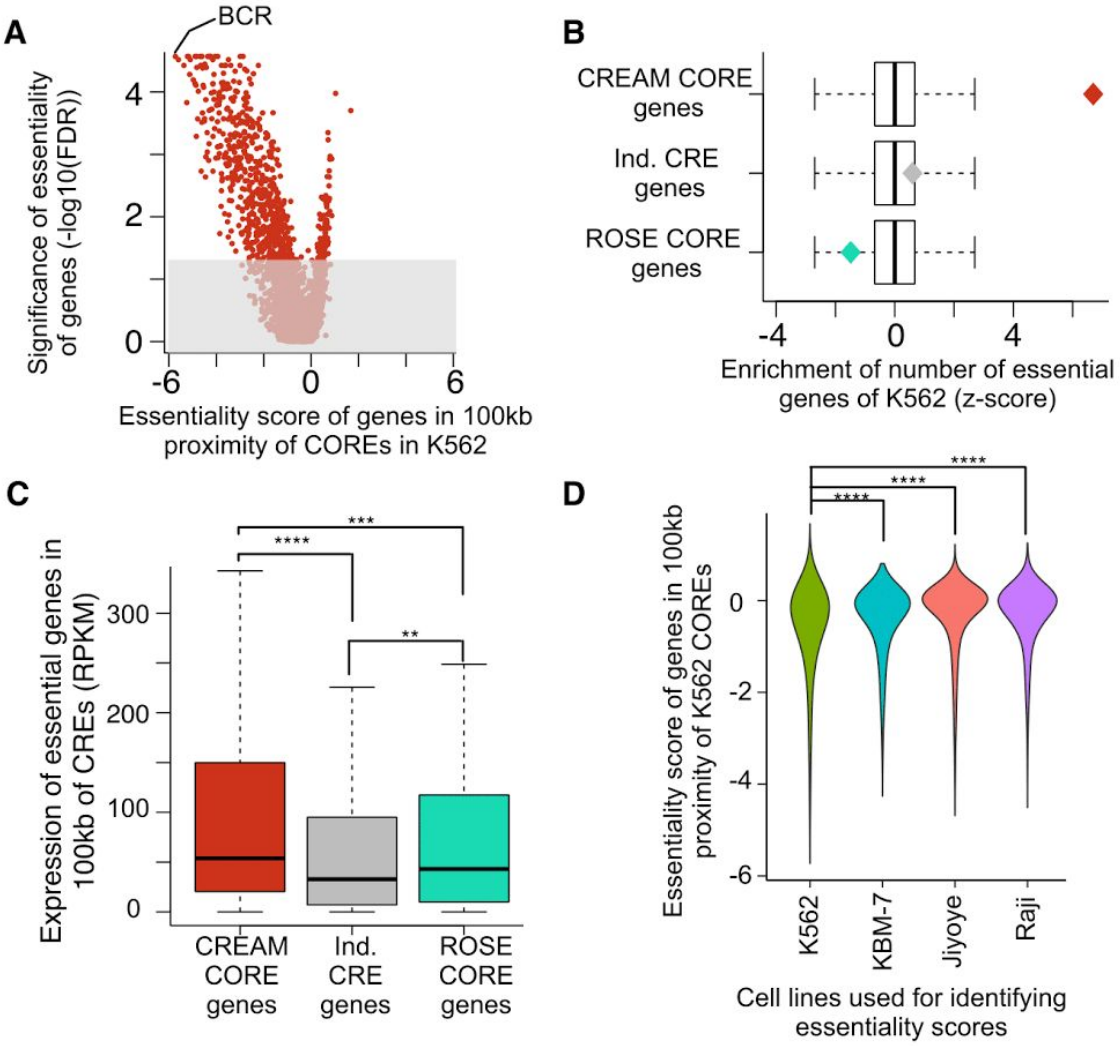


Figure 5. Essentiality of genes in proximity of COREs in K562 cell lines. ***(A)*** *Volcano plot of significance (FDR) and effect size (essentiality score) of genes in proximity of COREs in K562 cell line.* ***(B)*** *Enrichment of number of essential genes among the associated genes to individual CREs and COREs in K562 cell line. (**C**) Transcription level of essential genes which*

bioRxiv preprint doi: https://doi.org/10.1101/222562; this version posted November 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

are associated with individual CREs and COREs. ***(D)*** *Comparing essentiality score of genes in 100kb proximity of COREs (of K562) in K562, KBM-7, Jiyoye, and Raji cell line to check specificity of essentialities to K562.*

CONCLUSIONS

State-of-the-art approach for CORE calling (ROSE) dismiss the variability in the distribution of distances between CREs, unique to each CRE dataset, hence limited by considering a fixed threshold of 12.5 to 20 kilobases in distance between CREs within COREs, and a fixed cutoff in the ChIP-seq signal intensity from the assay used to identify CREs to separate COREs from individual CREs [11,12]. To overcome these limitations, we developed CREAM as an unsupervised machine learning method providing a systematic approach for identifying COREs.

Here, we show that CREAM identifies COREs that, (*i*) have higher transcription factor binding intensity with respect to individual CREs, (*ii*) associated with identity of normal and cancer cell lines, and (*iii*) have significantly higher probability of being essential for growth of the cells compared to the rest of epigenetic landscape. Hence, CREAM can open a new avenue of research for personalized therapeutic identification in clinical cancer setting. Taken all together, we show that CREAM can be used to further characterize cis-regulatory landscapes of cells.

METHODS

CREAM

CREAM uses genome-wide maps of cis-regulatory elements (CREs) in the tissue or cell type of interest, such as those generated from chromatin-based assays including DNase-seq,

bioRxiv preprint doi: https://doi.org/10.1101/222562; this version posted November 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

ATAC-seq or ChIP-seq. CREs can be identified from these profiles by peak calling tools such as MACS [35]. The called individual CREs then will be used as input of CREAM. Hence, CREAM does not need the signal intensity files (bam, fastq) as input. CREAM considers proximity of the CREs within each sample to adjust parameters of inclusion of CREs into a CORE in the following steps (Fig. 1):

Step 1: Clustering of individual CREs throughout genome. CREAM initially groups neighboring individual CREs throughout the genome. Each group (or cluster) can have different number of individual CREs. Then it categorizes the clusters based on their included CRE numbers. We defined Order (***O***) for each cluster as its included CRE number. In the next steps, CREAM identifies maximum allowed distance between individual CREs for COREs of a given ***O***.

Step 2: Maximum window size identification. We defined maximum window size (***MWS***) as the maximum distance between individual CREs included in a CORE. For each Order, CREAM builds a distribution of window sizes, as the maximum distance between individual CREs in each CORE, in all clusters of that Order within the genome. Afterward, ***MWS*** will be identified as follows

$$\boldsymbol{MWS} = Q1(\log(\boldsymbol{WS})) - 1.5 * IQ(\log(\boldsymbol{WS}))$$

where ***MWS*** is the maximum distance between neighboring individual CREs within a CORE. Q1(log(***WS***)) and IQ(log(***WS***)) are the first quartile and interquartile of distribution of window sizes (Fig. 1).

Step 3: Maximum Order identification. After determining ***MWS*** for each Order of COREs, CREAM identifies maximum ***O*** (***O***max) for the given sample. By increasing ***O*** of COREs, the individual CREs should be allowed to have further distance from each other as a result of gain of information within the clusters. Hence, starting from COREs of ***O***=2, the ***O*** increases up to a

bioRxiv preprint doi: https://doi.org/10.1101/222562; this version posted November 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

plateau at which an increase of ***O*** does not result an increase in ***MWS***. This threshold is considered as maximum ***O*** (***O***max) for COREs within the given sample.

Step 4: CORE calling. CREAM starts to identify COREs from ***O***max down to ***O***=2. For each ***O***, it calls clusters with window size less than ***MWS*** as COREs. As a result, many COREs with lower ***O***s are clustered within COREs with higher ***O***s. Therefore, remaining lower ***O*** COREs, for example ***O***=2 or 3, have individual CREs with distance close to ***MWS*** (Fig. 1). These clusters could have been identified as COREs because of the initial distribution of ***MWS*** derived mainly by COREs of the same ***O*** which are clustered in COREs of higher ***O***s. Hence, CREAM eliminate these low ***O*** COREs as follows.

Step 5: Minimum Order identification. COREs that contain individual CREs with distance close to ***MWS*** can be identified as COREs due to the high skewness in the initial distribution of ***MWS.*** To avoid reporting these COREs, CREAM filters out the clusters with (***O*** < ***O***min) which does not follow monotonic increase of maximum distance between individual CREs versus **O** (Fig. 1).

ROSE

ROSE clusters the neighboring individual CREs in a given sample if they have distance less than 12.5kb. It subsequently identifies the signal overlap on the clusters and sorts the identified clusters based on their signal intensity. It then stratifies the clusters based on the inflation point in the sorted clusters and call the clusters with signal intensity higher than the inflation point as super-enhancers (or COREs). This method is comprehensively explained in Whyte *et al.* (2013) [11]. We ran ROSE using the default parameters.

Genomic overlap of COREs

bioRxiv preprint doi: https://doi.org/10.1101/222562; this version posted November 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Bedtools (version 2.23.0) is used to identify unique and shared genomic coverage between CREAM and ROSE-identified COREs.

Comparison of CORES identified by CREAM and ROSE and single enhancers

First, signals (either DNase I hypersensitivity or ChIP-seq) over the identified COREs (or individual CREs) and 1kb flanking regions of them were extracted from the BAM files. Then each CORE (or individual CREs) is binned to 100 binned regions with equal size. Each left and right flanking region is also divided to 100 bins with equal size. Hence, in total 300 bins are obtained for each CORE plus its flanking regions. We then scale the signal in these regions to the library size for the mapped reads. Finally, a Savitzky-Golay filter is applied to remove high frequency noise from the data while preserving the original shape of the data [36,37].

Association with genes

A gene is considered associated with a CRE or a CORE if found within a ∓100kb window from each other.

Gene expression

RNA sequencing profile of GM12878 and K562 cells lines, available in ENCODE database [31], are used to identify expression of genes in proximity of individual CREs and COREs.

Transcription factor binding enrichment

Bedgraph files of ChIP-Seq profiles of transcription factors are overlapped with the identified COREs and individual CREs in GM12878 and K562 using bedtools (version 2.23.0).

bioRxiv preprint doi: https://doi.org/10.1101/222562; this version posted November 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

The resulting signal were summed over all the individual CREs or COREs and then normalized to the total genomic coverage of individual CREs or COREs, respectively. These normalized transcription factor binding intensities are used for comparing TF binding intensity in individual CREs and COREs (Fig. 3).

Sample similarity

Similarity between two samples in ENCODE is identified based on Jaccard index for the commonality of their identified COREs throughout the genome. Then this Jaccard index is used as the similarity statistics in a 1-nearest-neighbor classification approach. We assess performance of the classification using leave-one-out cross validation. In this classification scheme, we considered phenotype of the closest sample to an out of pool sample as its phenotype.

Association with essential genes

Number of genes which are in ∓100kb proximity of COREs and are essential in K562 are identified [33]. This number is then compared with number of essential genes in 10,000 randomly selected (permuted) genes, among the genes included in the essentiality screen. This comparison is used to identify FDR and z-score regarding the significance of enrichment of essential genes among genes in ∓100kb proximity of COREs identified for K562 cell line.

Pathway enrichment analysis

ConsensusPathDB is used to implement pathway enrichment analysis [38]. Protein complex-based gene sets is used as query gene sets.

bioRxiv preprint doi: https://doi.org/10.1101/222562; this version posted November 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Research Reproducibility

CREAM is now available as an open source R package (https://CRAN.R-project.org/package=CREAM).

List of abbreviations

Abbreviation	Stand for
CRE	Cis-Regulatory Element
CORE	Cluster Of cis-Regulatory Element
CREAM	Clustering of genomic REgions Analysis Method
TF	Transcription factor
MWS	Maximum Window Size
O	Order
FC	Fold Change
FDR	False Discovery Rate
MCC	Matthew Correlation Coefficient
TF	Transcription Factor

bioRxiv preprint doi: https://doi.org/10.1101/222562; this version posted November 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

DECLARATIONS

Author contributions

S.A.M.T. developed CREAM. S.A.M.T. and V.K. prepared CREAM R package. S.A.M.T. and P.M. performed the analysis and interpreted the results. S.A.M.T., P.M., B.H-K, and M.L. conceived the design of the study. S.A.M.T., P.M., B.H-K, and M.L. wrote the manuscript. B.H-K and M.L. supervised the study.

Funding

This study was conducted with the support of the Terry Fox Research Institute, Canadian Cancer Research Society and the Ontario Institute for Cancer Research through funding provided by the Government of Ontario. We acknowledge the Princess Margaret Bioinformatics group for providing the infrastructure assisting us with analysis presented here. This work was supported by the Princess Margaret Cancer Foundation (M.L. and B.H.K.). M.L. holds an Investigator Award from the Ontario Institute for Cancer Research, a CIHR New Investigator Award and a Movember Rising Star Award from Prostate Cancer Canada and is proudly funded by the Movember Foundation (grant #RS2014-04). S.A.M.T was supported by Connaught International Scholarships for Doctoral Students. P. M. was supported by the Canadian Institutes of Health Research Scholarship for Doctoral Students. B.H.K is supported by the Gattuso-Slaight Personalized Cancer Medicine Fund at Princess Margaret Cancer Centre and the Canadian Institutes of Health Research.

Acknowledgment

DNase I sequencing profile of HeLa cell line is used in this research. Henrietta Lacks, and the HeLa cell line that was established from her tumor cells without her knowledge or consent in 1951, have made significant contributions to scientific progress and advances in human health. We are grateful to Henrietta Lacks, now deceased, and to her surviving family members for their contributions to biomedical research. We also acknowledge the ENCODE Consortium and the ENCODE production laboratories that generated the data sets provided by the ENCODE Data Coordination Center used in this manuscript.

Competing financial interests

The authors declare no competing financial interests.

bioRxiv preprint doi: https://doi.org/10.1101/222562; this version posted November 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

REFERENCES

1. Zhou S, Treloar AE, Lupien M. Emergence of the Noncoding Cancer Genome: A Target of Genetic and Epigenetic Alterations. Cancer Discov. 2016;6:1215–29.

2. Kellis M, Wold B, Snyder MP, Bernstein BE, Kundaje A, Marinov GK, et al. Defining functional DNA elements in the human genome. Proc. Natl. Acad. Sci. U. S. A. 2014;111:6131–8.

3. Lupien M, Eeckhoute J, Meyer CA, Wang Q, Zhang Y, Li W, et al. FoxA1 translates epigenetic signatures into enhancer-driven lineage-specific transcription. Cell. Elsevier Ltd; 2008;132:958–70.

4. Heintzman ND, Hon GC, Hawkins RD, Kheradpour P, Stark A, Harp LF, et al. Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature. Nature Publishing Group; 2009;459:108–12.

5. Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, et al. Mapping and analysis of chromatin state dynamics in nine human cell types. Nature. Nature Publishing Group; 2011;473:43–9.

6. Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, et al. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nat. Genet. Nature Publishing Group; 2007;39:311–8.

7. Hnisz D, Abraham BJ, Lee TI, Lau A, Saint-André V, Sigova AA, et al. Super-enhancers in the control of cell identity and disease. Cell. 2013;155:934–47.

8. Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, et al. The accessible chromatin landscape of the human genome. Nature. Nature Publishing Group; 2012;489:75–82.

9. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat. Methods. Nature Publishing Group; 2013;1–8.

10. Parker SCJ, Stitzel ML, Taylor DL, Orozco JM, Erdos MR, Akiyama JA, et al. Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. Proc. Natl. Acad. Sci. U. S. A. 2013;110:17921–6.

11. Whyte WA, Orlando DA, Hnisz D, Abraham BJ, Lin CY, Kagey MH, et al. Master transcription factors and mediator establish super-enhancers at key cell identity genes. Cell. 2013;153:307–19.

12. Gaulton KJ, Nammo T, Pasquali L, Simon JM, Giresi PG, Fogarty MP, et al. A map of open chromatin in human pancreatic islets. Nature Publishing Group. Nature Publishing Group; 2010;42:255–9.

13. Dowen JM, Fan ZP, Hnisz D, Ren G, Abraham BJ, Zhang LN, et al. Control of cell identity genes occurs in insulated neighborhoods in mammalian chromosomes. Cell. 2014;159:374–87.

14. Vahedi G, Kanno Y, Furumoto Y, Jiang K, Parker SCJ, Erdos MR, et al. Super-enhancers

bioRxiv preprint doi: https://doi.org/10.1101/222562; this version posted November 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

delineate disease-associated regulatory nodes in T cells. Nature. 2015;520:558–62.

15. Corradin O, Cohen AJ, Luppino JM, Bayles IM, Schumacher FR, Scacheri PC. Modeling disease risk through analysis of physical interactions between genetic variants within chromatin regulatory circuitry. Nat. Genet. 2016;48:1313–20.

16. Pasquali L, Gaulton KJ, Rodríguez-Seguí SA, Mularoni L, Miguel-Escalada I, Akerman İ, et al. Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants. Nat. Genet. 2014;46:136–43.

17. Lovén J, Hoke HA, Lin CY, Lau A, Orlando DA, Vakoc CR, et al. Selective inhibition of tumor oncogenes by disruption of super-enhancers. Cell. 2013;153:320–34.

18. Northcott PA, Lee C, Zichner T, Stütz AM, Erkek S, Kawauchi D, et al. Enhancer hijacking activates GFI1 family oncogenes in medulloblastoma. Nature. 2014;511:428–34.

19. Chipumuro E, Marco E, Christensen CL, Kwiatkowski N, Zhang T, Hatheway CM, et al. CDK7 inhibition suppresses super-enhancer-linked oncogenic transcription in MYCN-driven cancer. Cell. 2014;159:1126–39.

20. Moorthy SD, Davidson S, Shchuka VM, Singh G, Malek-Gilani N, Langroudi L, et al. Enhancers and super-enhancers have an equivalent regulatory role in embryonic stem cells through regulation of single or multiple genes. Genome Res. 2017;27:246–58.

21. Hay D, Hughes JR, Babbs C, Davies JOJ, Graham BJ, Hanssen L, et al. Genetic dissection of the α-globin super-enhancer in vivo. Nat. Genet. 2016;48:895–903.

22. Shin HY, Willi M, HyunYoo K, Zeng X, Wang C, Metser G, et al. Hierarchy within the mammary STAT5-driven Wap super-enhancer. Nat. Genet. 2016;48:904–11.

23. Dukler N, Gulko B, Huang Y-F, Siepel A. Is a super-enhancer greater than the sum of its parts? Nat. Genet. 2016;49:2–3.

24. Pott S, Lieb JD. What are super-enhancers? Nat. Genet. Nature Publishing Group; 2014;47:ng.3167.

25. Somasundaram R, Prasad MAJ, Ungerbäck J, Sigvardsson M. Transcription factor networks in B-cell differentiation link development to acute lymphoid leukemia. Blood. 2015;126:144–52.

26. Erkeland SJ, Valkhof M, Heijmans-Antonissen C, Delwel R, Valk PJM, Hermans MHA, et al. The gene encoding the transcriptional regulator Yin Yang 1 (YY1) is a myeloid transforming gene interfering with neutrophilic differentiation. Blood. 2003;101:1111–7.

27. Yang Z-F, Zhang H, Ma L, Peng C, Chen Y, Wang J, et al. GABP transcription factor is required for development of chronic myelogenous leukemia via its control of PRKD2. Proc. Natl. Acad. Sci. U. S. A. 2013;110:2312–7.

28. Shankar DB, Cheng JC, Kinjo K, Federman N, Moore TB, Gill A, et al. The role of CREB as a proto-oncogene in hematopoiesis and in acute myeloid leukemia. Cancer Cell. 2005;7:351–62.

bioRxiv preprint doi: https://doi.org/10.1101/222562; this version posted November 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

29. Bulger M, Groudine M. Functional and mechanistic diversity of distal transcription enhancers. Cell. 2011;144:327–39.

30. Maston GA, Evans SK, Green MR. Transcriptional regulatory elements in the human genome. Annu. Rev. Genomics Hum. Genet. 2006;7:29–59.

31. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489:57–74.

32. Di Micco R, Fontanals-Cirera B, Low V, Ntziachristos P, Yuen SK, Lovell CD, et al. Control of embryonic stem cell identity by BRD4-dependent transcriptional elongation of super-enhancer-associated pluripotency genes. Cell Rep. 2014;9:234–47.

33. Wang T, Birsoy K, Hughes NW, Krupczak KM, Post Y, Wei JJ, et al. Identification and characterization of essential genes in the human genome. Science. 2015;350:1096–101.

34. Ren R. Mechanisms of BCR--ABL in the pathogenesis of chronic myelogenous leukaemia. Nat. Rev. Cancer. Nature Publishing Group; 2005;5:172–83.

35. Feng J, Liu T, Zhang Y. Using MACS to identify peaks from ChIP-Seq data. Curr. Protoc. Bioinformatics. 2011;Chapter 2:Unit 2.14.

36. Savitzky A, Golay MJE. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. Anal. Chem. American Chemical Society; 1964;36:1627–39.

37. Press WH. Numerical Recipes in C: The Art of Scientific Computing. Cambridge University Press; 1992.

38. Kamburov A, Stelzl U, Lehrach H, Herwig R. The ConsensusPathDB interaction database: 2013 update. Nucleic Acids Res. 2013;41:D793–800.