CREME is a one-function-python-tool, part of the genepy package. We built this tool noticing the lack of easy to use, publicly available Chip Replicate Merging tool article. It allows users to merge any set of Chip-seq replicates in a fast manner. CREME can merge replicates from different groups and different methods. CREME works without requiring the raw sequencing files and as a post processing step to the MACS2 peak calling algorithm. It works with 1 to many sets of replicates for each pulled protein/mark. It takes as inputs bed files and bigwig tracks for each replicate and outputs a merged bed file, together with a set of merging quality metrics and of putative bad quality replicates and bad quality proteins/marks.

CREME first computes a consensus, considering any peaks at most 150 bp from another peak, to be in overlap. We have noticed that changing this parameter from 0 to 150 decreased the total number of peaks found by only 8% (see figure S17A).

Non overlapping peaks are kept in the consensus. When we have an overlap we take the mean of signals and the product of p-values across overlapping replicates.

Then, CREME will look at their overlap (see figure S17B) and select the one that has the best overlap score:

$$O_{score}(A) = \sum ifrom0tom \sum Kincomb(i,G) i * \sum jfrom0tonAND(A[j], ...K[j])$$

Where:

- $G$ is a binary matrix of size (row/col) $m * n$ of $m$ replicates with $n$ consensus peaks and a value of 1 if replicate $m_i$ has a peak on consensus peak $n_i$.
- $comb(i,G)$ is a list of all possible matrices made from taking $i$ elements (row) from matrix $G$ without replacement.
- $AND$ is a binary operation returning 1 if all passed elements are 1 else 0.

The best scoring sample not labelled as bad quality replicate will be selected as the **main replicate**.

Bad replicates are user provided annotations, e.g. found by visual inspection of bigwig tracks, a threshold on FRiP scores or any other method (figure S17C).

If we find that the second best scoring replicate and the **main replicate** have both less than 30% of their peaks in common we mark that protein/mark as **failed** and only return the **main replicate**.

We then look for new peaks.

The process of calling peaks is loosely based on MACS2's peak calling algorithm:

We compute a KL divergence between two poisson distributions. One is representing the distribution of signals from a bigwig file under our region of interest. The other is representing the same signal under the entire chromosome of that region.If that distance is above a threshold (here, 8) we validate the region as being a peak.

For each additional replicate, we will now look for new peaks. First, if we find that the second best replicate and the first best replicate have both less than 30% of their peaks in common we **discard** that protein/mark and only return the main replicate.

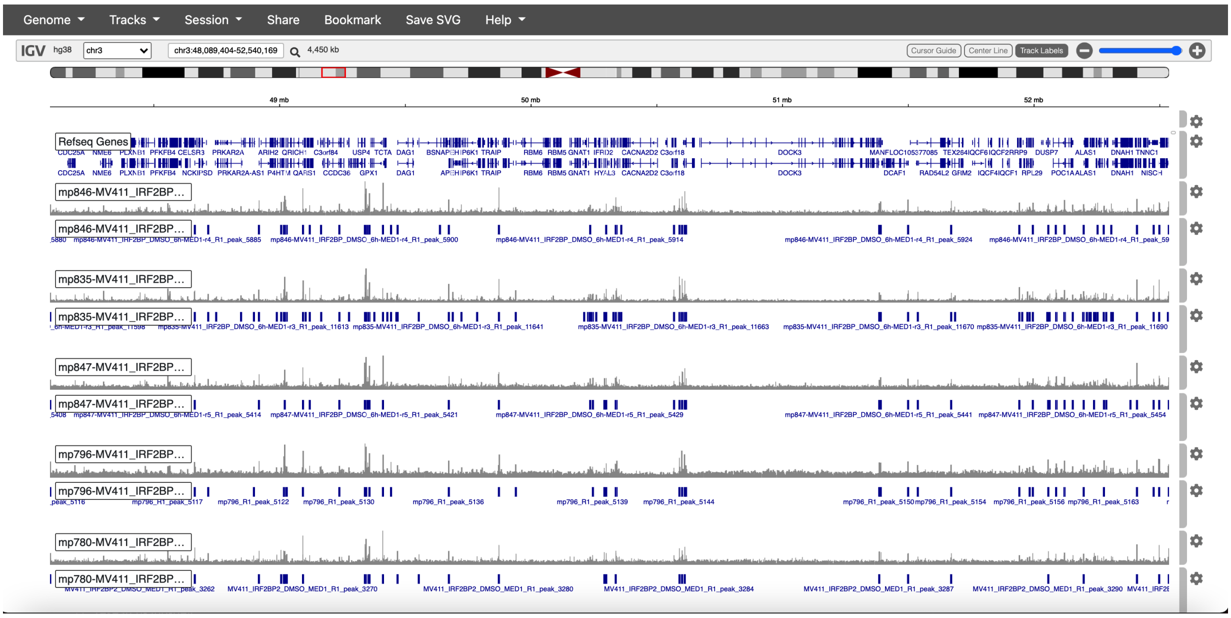
Then, taking peak loci in the **main replicate**, we will try to call them under the second best replicate using its bigwig and a lower threshold than MACS2. We then do the same for peaks in that replicate that were not in the **main replicate**. If after calling new peaks we get less than 30% overlap in both replicates, we discard the replicate.

We add the values for all new found peaks and continue iteratively for each following replicate (figure S17D, figure S17E). The final merged bed file is comprised of all peaks of the **main replicates** (with newfound peaks) and all peaks of other replicates if present in at least 2 replicates. The overlapping peaks are aggregated in their intensity, their p-values and boundaries.

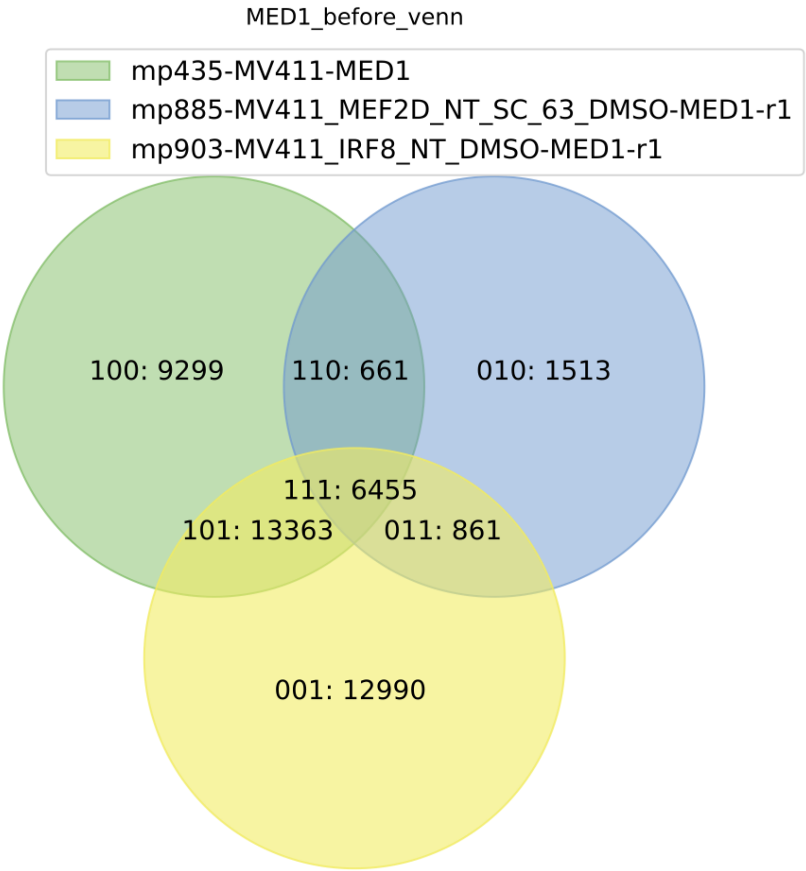
The pipeline also outputs a set of putative bad quality replicates and bad quality proteins/marks based on this overlap & peak calling procedure.

We think that CREME can fit as a quick and simple utility for aggregating any set of replicate ChIP-seq, even with a coarse knowledge of their provenance and no raw sequencing data. We have reliably used it on hundreds of replicates of variable quality, from different sequencing platforms and different labs to greatly improve the accuracy and quality of our subsequent data discovery pipeline. CREME is easy to modify and understand.

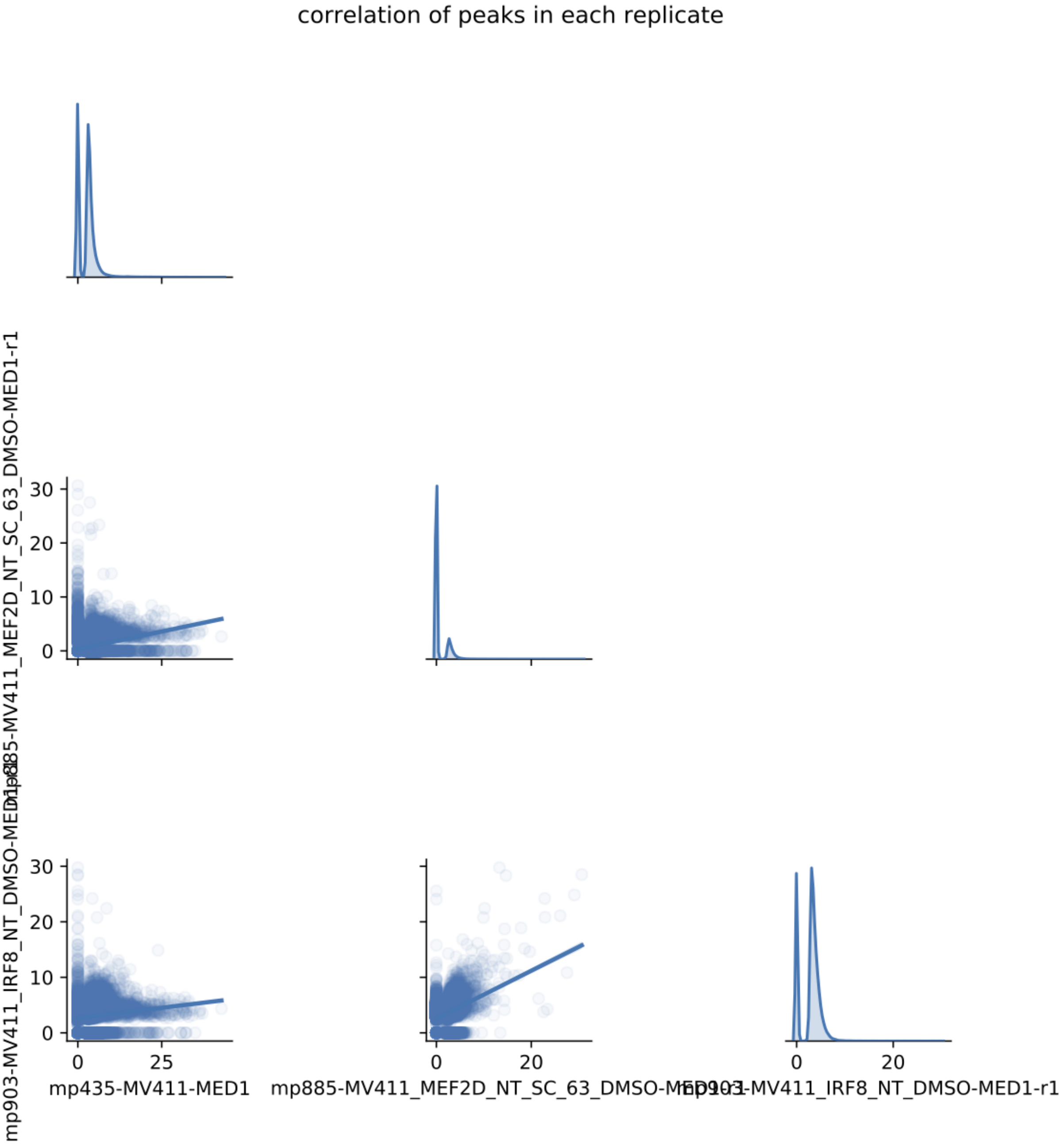
- 17A:


Genome
Tracks
Session
Share
Bookmark
Save SVG
Help
IGV
hg38
chr3
chr3:48,089,404-52,540,169
4,450 kb
Cursor Guide
Center Line
Track Labels
49 mb
50 mb
51 mb
52 mb
Refseq Genes
CDC25A NME6 PLXNB1 PFKFB4 CELSR3 PRKAR2A ARIH2 QRICH1 C3orf84 USP4 TCTA DAG1 BSNAP IP6K1 TRAIP RBM6 RBM5 GNAT1 IFRD2 CACNA2D2 C3orf18 DOCK3 MANF LOC105377085 TEX264 IQCF6 IQCF2 RRP9 DUSP7 ALAS1 DNAH1 TNNC1
CDC25A NME6 PLXNB1 PFKFB4 NCKIPSD PRKAR2A-AS1 P4HTM QARS1 CCDC36 GPX1 DAG1 AP IP6K1 TRAIP RBM6 RBM5 GNAT1 HYAL3 CACNA2D2 C3orf18 DOCK3 DCAF1 RAD54L2 GRM2 IQCF4 IQCF1 RPL29 POC1A ALAS1 DNAH1 NISCH
mp846-MV411_IRF2BP...
mp846-MV411_IRF2BP...
mp846-MV411_IRF2BP_DMSO_6h-MED1-r4_R1_peak_5885
mp846-MV411_IRF2BP_DMSO_6h-MED1-r4_R1_peak_5900
mp846-MV411_IRF2BP_DMSO_6h-MED1-r4_R1_peak_5914
mp846-MV411_IRF2BP_DMSO_6h-MED1-r4_R1_peak_5924
mp835-MV411_IRF2BP...
mp835-MV411_IRF2BP...
mp835-MV411_IRF2BP_DMSO_6h-MED1-r3_R1_peak_11613
mp835-MV411_IRF2BP_DMSO_6h-MED1-r3_R1_peak_11641
mp835-MV411_IRF2BP_DMSO_6h-MED1-r3_R1_peak_11663
mp835-MV411_IRF2BP_DMSO_6h-MED1-r3_R1_peak_11670
mp835-MV411_IRF2BP_DMSO_6h-MED1-r3_R1_peak_11690
mp847-MV411_IRF2BP...
mp847-MV411_IRF2BP...
mp847-MV411_IRF2BP_DMSO_6h-MED1-r5_R1_peak_5414
mp847-MV411_IRF2BP_DMSO_6h-MED1-r5_R1_peak_5421
mp847-MV411_IRF2BP_DMSO_6h-MED1-r5_R1_peak_5429
mp847-MV411_IRF2BP_DMSO_6h-MED1-r5_R1_peak_5441
mp847-MV411_IRF2BP_DMSO_6h-MED1-r5_R1_peak_5454
mp796-MV411_IRF2BP...
mp796-MV411_IRF2BP...
mp796_R1_peak_5117
mp796_R1_peak_5122
mp796_R1_peak_5130
mp796_R1_peak_5136
mp796_R1_peak_5139
mp796_R1_peak_5144
mp796_R1_peak_5150
mp796_R1_peak_5154
mp796_R1_peak_5156
mp796_R1_peak_5163
mp780-MV411_IRF2BP...
mp780-MV411_IRF2BP...
MV411_IRF2BP2_DMSO_MED1_R1_peak_3262
MV411_IRF2BP2_DMSO_MED1_R1_peak_3270
MV411_IRF2BP2_DMSO_MED1_R1_peak_3280
MV411_IRF2BP2_DMSO_MED1_R1_peak_3284
MV411_IRF2BP2_DMSO_MED1_R1_peak_3287
MV411_IRF2BP2_DMSO_MED1_R1_peak_3290

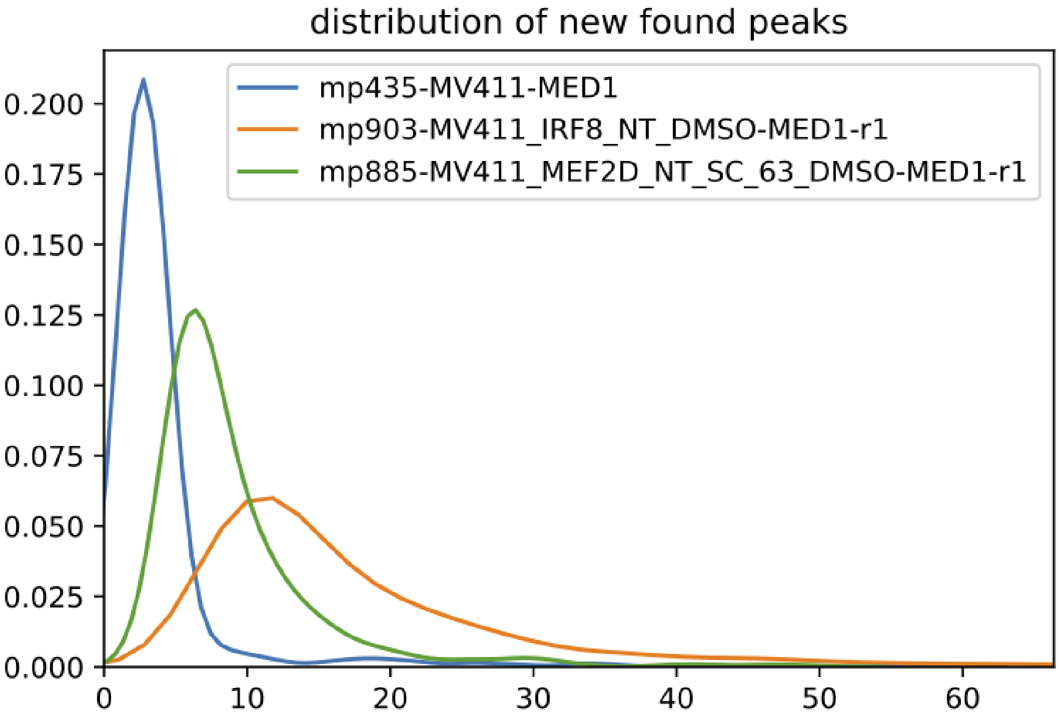
- 17B:


MED1_before_venn
mp435-MV411-MED1
mp885-MV411_MEF2D_NT_SC_63_DMSO-MED1-r1
mp903-MV411_IRF8_NT_DMSO-MED1-r1
100: 9299
110: 661
010: 1513
111: 6455
101: 13363
011: 861
001: 12990

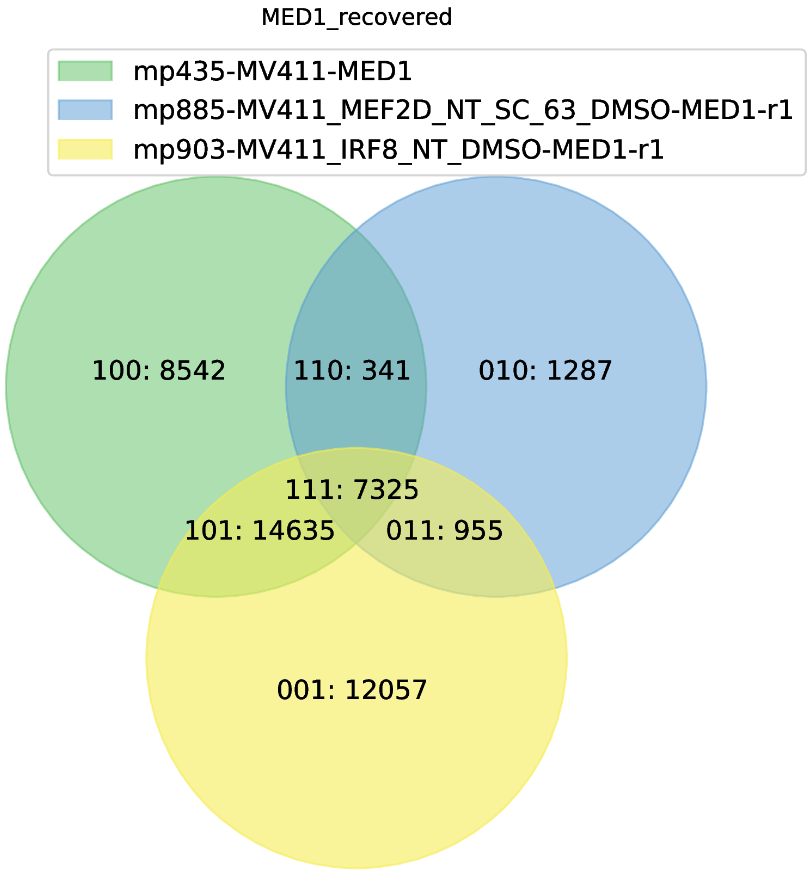
- 17C:


correlation of peaks in each replicate
mp885-MV411_MEF2D_NT_SC_63_DMSO-MED1-r1
mp903-MV411_IRF8_NT_DMSO-MED1-r1
0
10
20
30
0
25
20
mp435-MV411-MED1
mp885-MV411_MEF2D_NT_SC_63_DMSO-MED1-r1
mp903-MV411_IRF8_NT_DMSO-MED1-r1

- 17D:


distribution of new found peaks
mp435-MV411-MED1
mp903-MV411_IRF8_NT_DMSO-MED1-r1
mp885-MV411_MEF2D_NT_SC_63_DMSO-MED1-r1
0.200
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
0
10
20
30
40
50
60

- 17E:



- A: An IGV plot of mediator binding across replicates over a random loci in MV411. We can see regions that match and regions that could match, given looser peak calling thresholds.
- B: A venn diagram of the number of overlapping peaks across replicates.
- C: In addition to the venn diagram, correlation between each replicate's peak signals is computed and displayed to the user.
- D: A density plot over the distribution of intensity values of the newfound peaks of each replicate.
- E: Same as *B* but after calling new peaks in each replicate.

We will note the existence of similar tools such as PePr, genoGAM, multiGPS, MSPC, code, sierra Platinum, which are all both more complex, not implemented in python, requiring the raw sequencing data and often difficult to easily use and modify.

More information about CREME is available in the supplementary material and at our github explainer readme.