ChIP-Seq peaks classification

Oleg Shpynov
oleg.shpynov@jetbrains.com

July 14, 2016

Contents

1 Introduction

The goal of this work is to answer the question: *How to distinguish narrow and broad ChIP-Seq peaks?*

2 Methods and tools

We used ENCODE dataset GSE26320 to estimate ChIP-Seq peak length distribution. It is was shown that different ChIP-Seq modification has different biological function and demonstrate different patterns across genome. **TODO**
In order to deal with epigenomic data in human readable predicate way it is quite natural to have the ability to classify ChIP-Seq peaks by distribution, shape, intensity, length, etc.
To address this question we performed a simple experiment, which allowed us to compare peaks characteristics provided by different peak callers. We've used gold-standard peak callers: MACS2, SICER as well as our own peak caller ZINBRA.

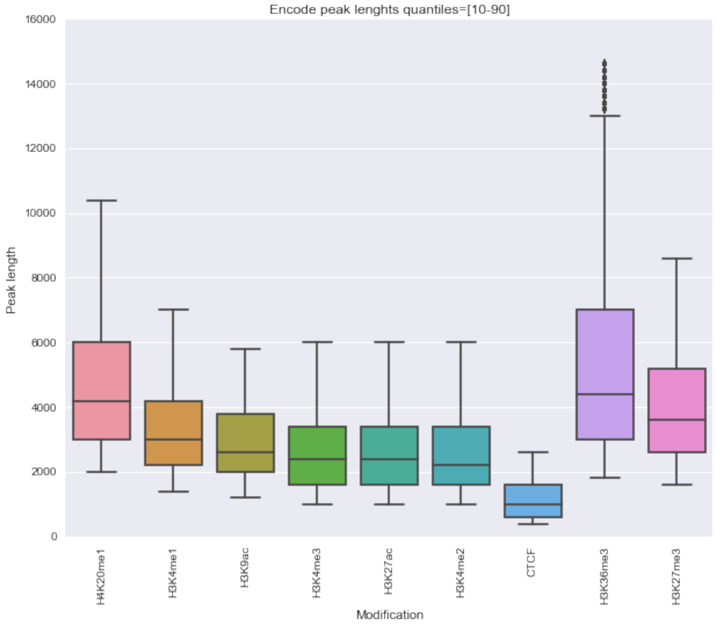
2.1 Experiment design

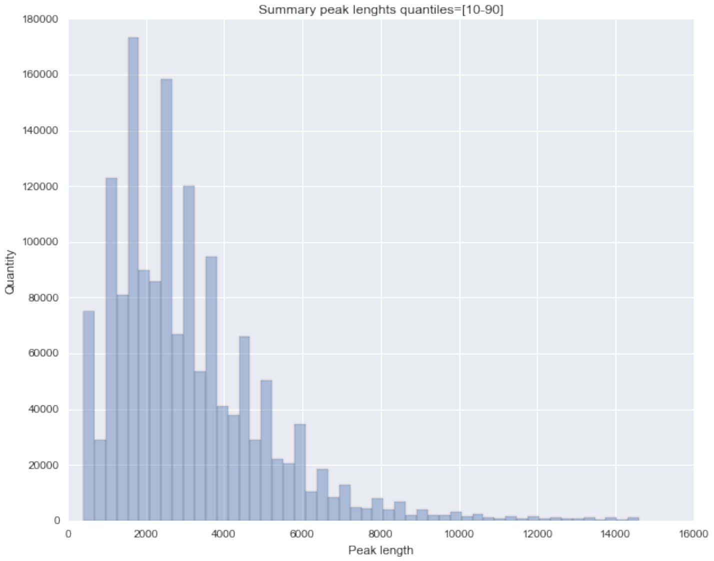
- Download GSE26320 dataset
- Align all reads to hg18 using bowtie with default params
- Launch MACS2, SICER and ZINBRA to produce peaks
- Aggregate peaks, produced for different ChIP-Seq tracks by distinct histone modifications
- Filter out outliers by processing only 10-90 percentile of all values
- Analyze peaks length distribution for different histone modifications
- Formulate criterion to classify peaks into narrow and broad classes

For each tool we plot aggregated peak lengths distribution and boxplots for each histone mark.

3 SICER

SICER is a tool, which supports peak calling for both narrow and broad peaks out of the box, so that we used it as reference.

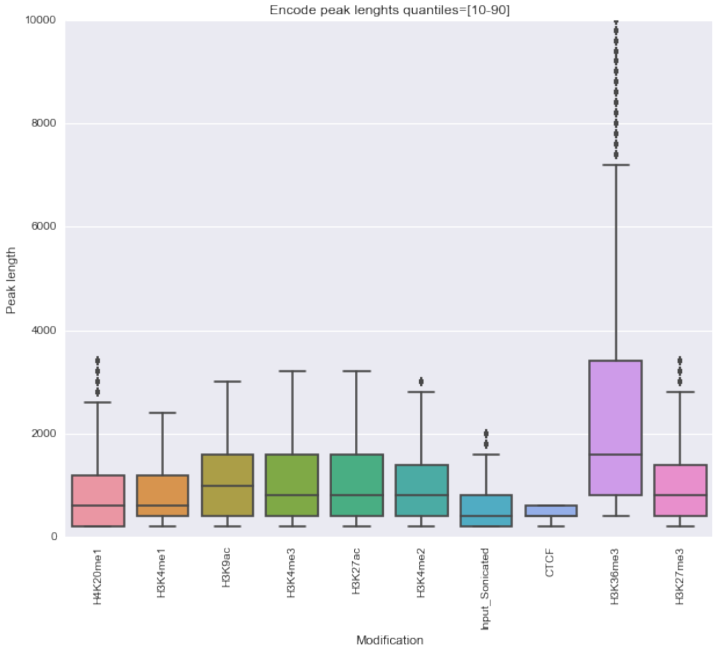


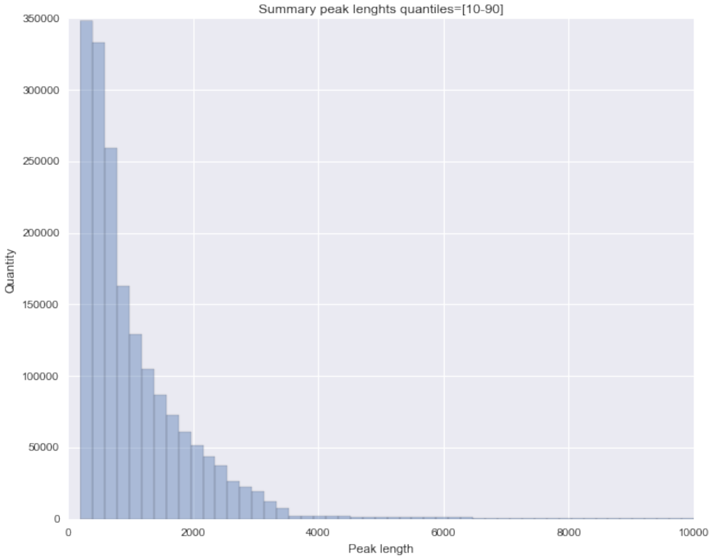


4 ZINBRA

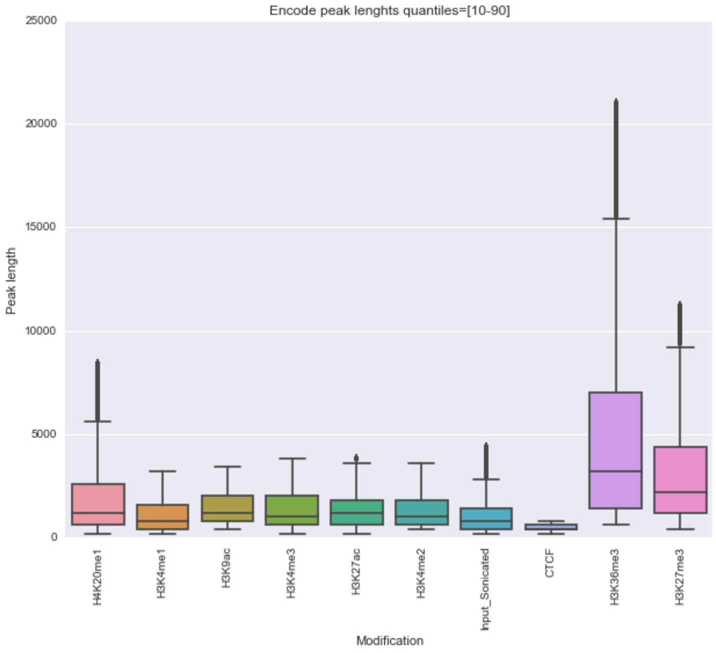
ZINBRA uses single statistical model for both narrow and broad peaks, producing enrichment states for binned(default bin size = 200, which corresponds to typical nucleosome size) genome, concatenating adjacent enriched bins into wider peaks.

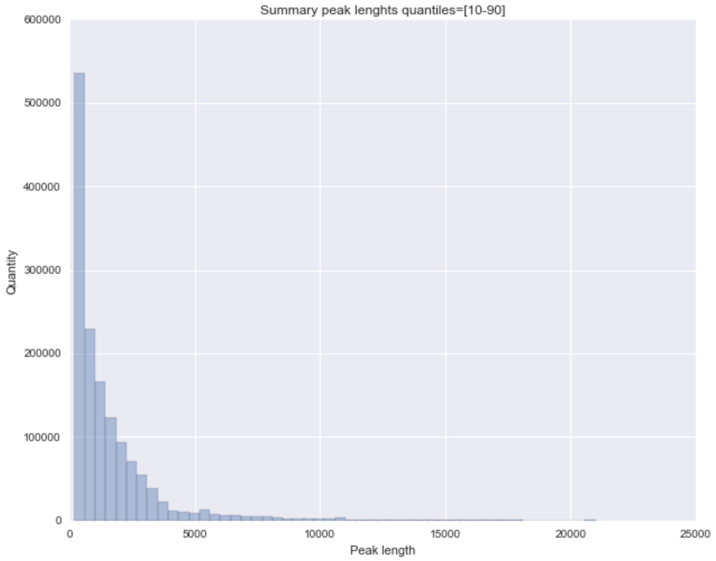
4.1 FDR = 0.0001



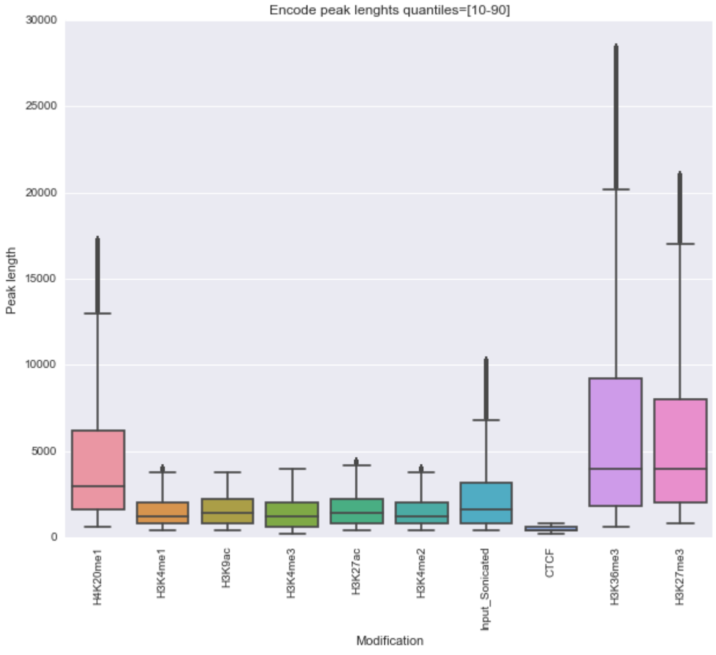


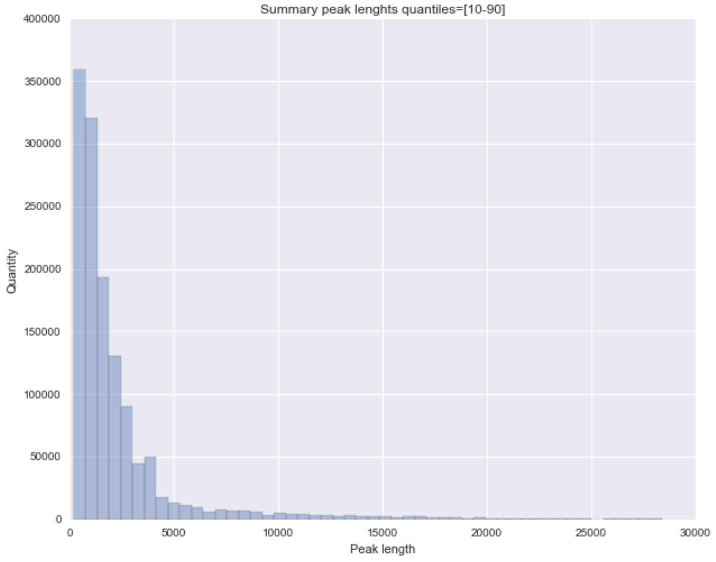
4.2 FDR = 0.001





4.3 FDR = 0.01

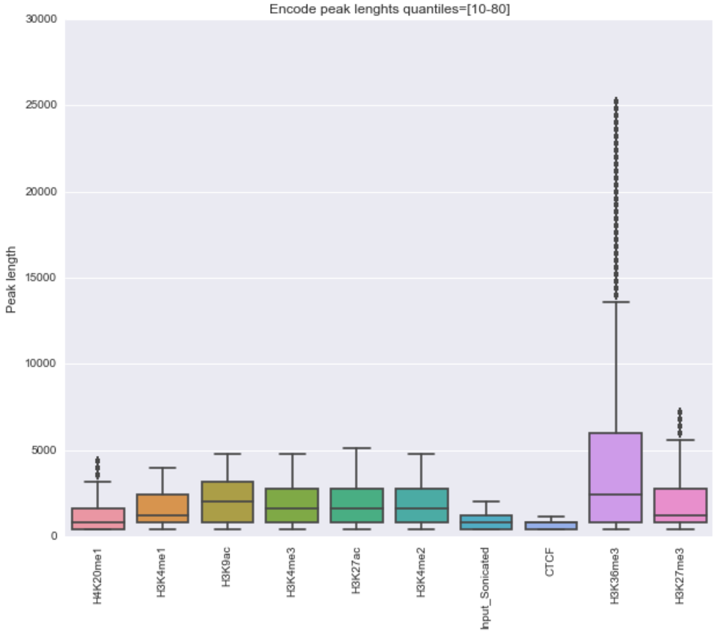


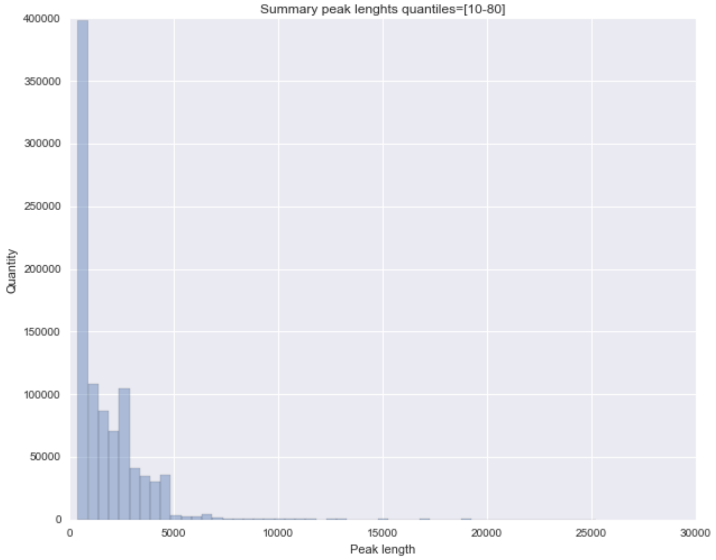


5 ZINBRA with extended peaks

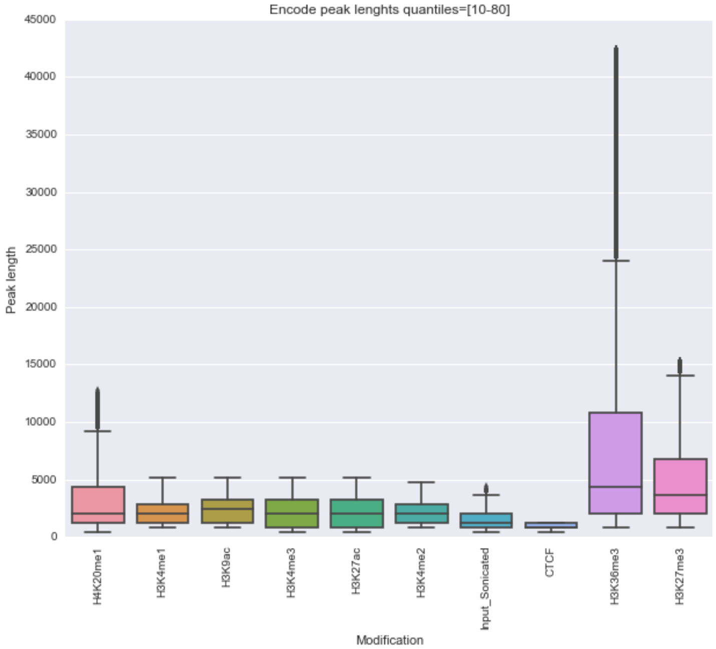
Motivation: in ChIP-Seq enrichment predicates, decision on whether given locus is enriched or not is made according to the fraction of enriched bins within. So that we decided to extend enriched regions (adjacent enriched bins) in following manner: extend each region as much as possible, so that fraction of enriched bins is $\geq$ threshold $k = 0.5$.

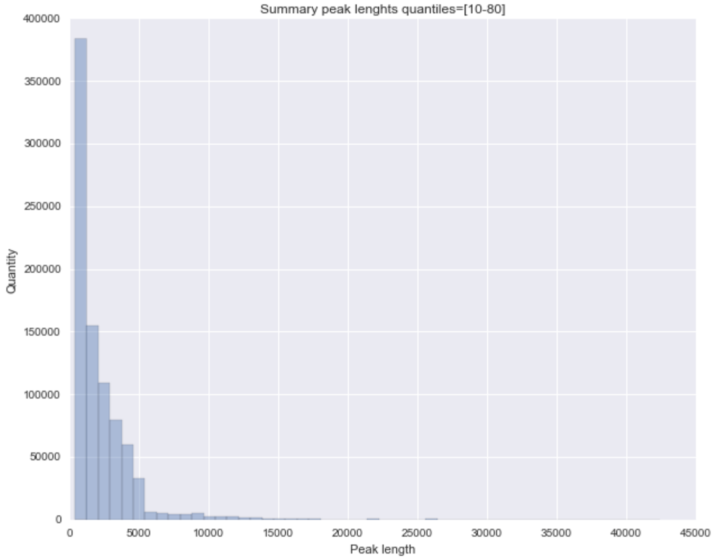
5.1 Extended peaks FDR = 0.0001 K = 0.5





5.2 Extended peaks FDR = 0.001 K = 0.5

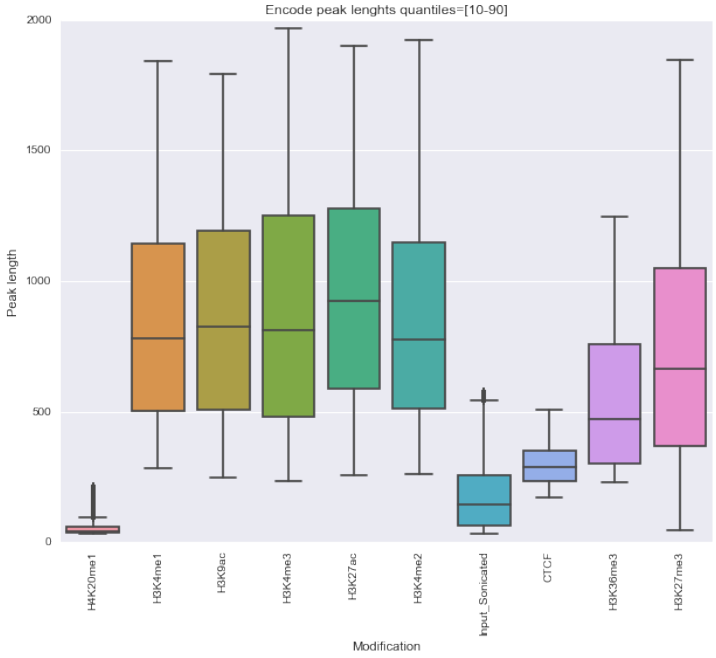


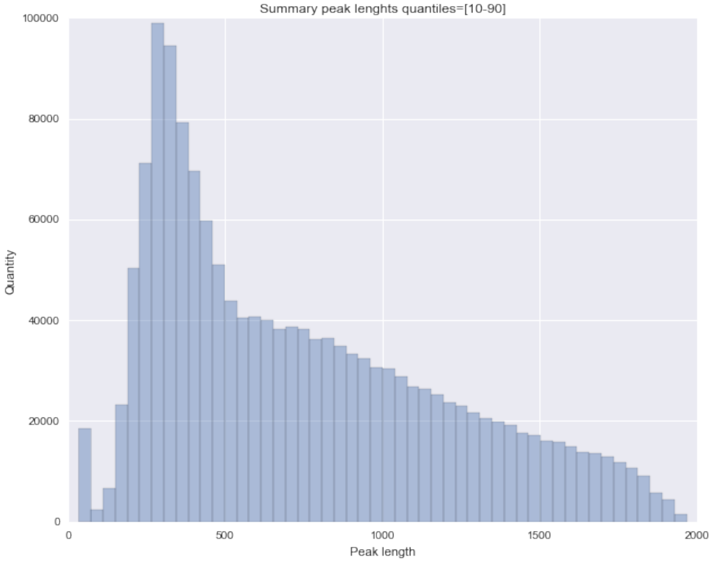


6 MACS2

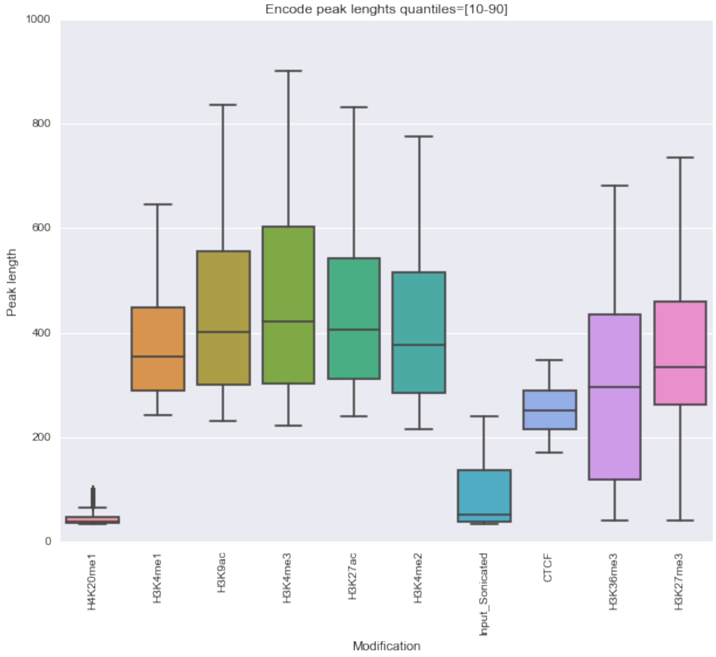
MACS2 by default deals with narrow peaks, however it has –*broad* command line options to process broad peaks as well.

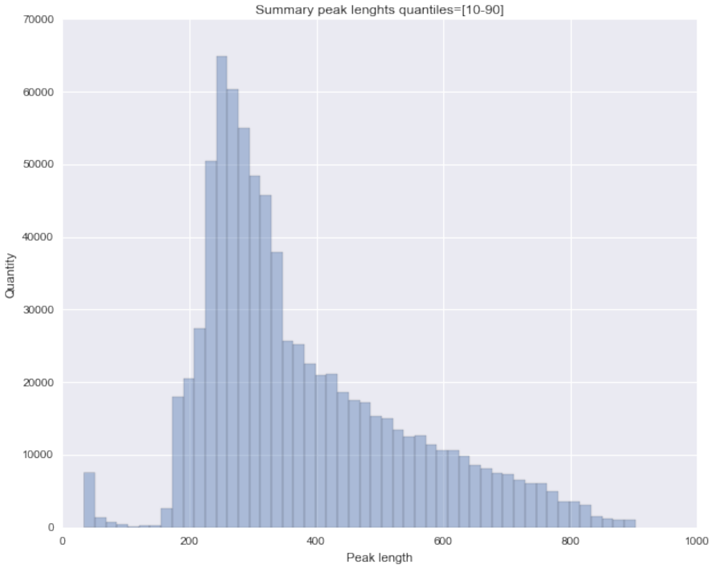
6.1 MACS2 Broad



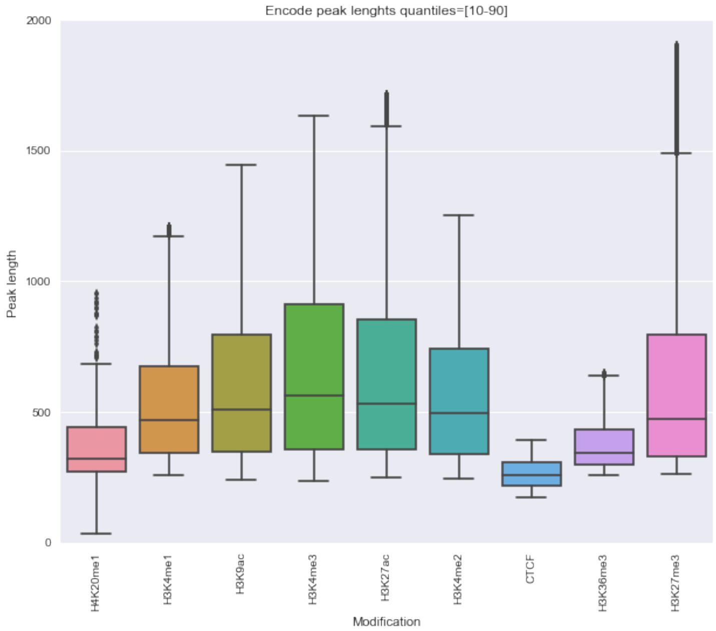


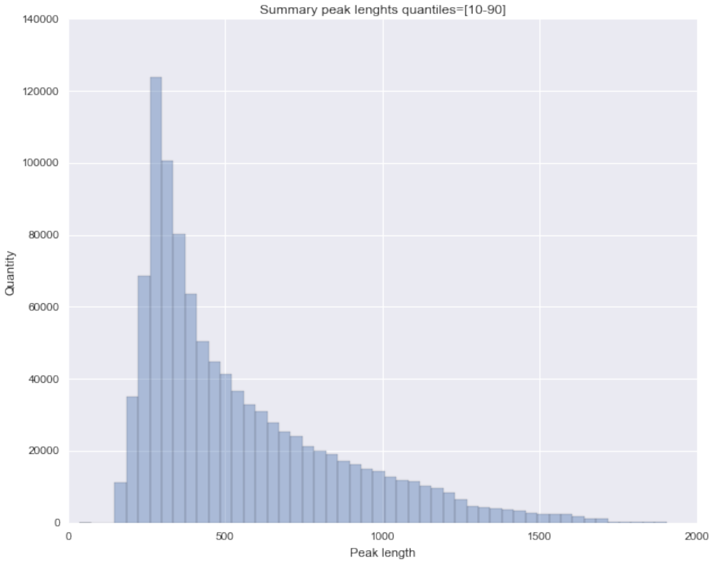
6.2 MACS2 Narrow





6.3 MACS2 Narrow with control





7 Discussion

The evidence shows that MACS2 can hardly be used to estimate both narrow and broad peaks. Even MACS2 in –*broad* mode fails to classify H3K36me3 as broad peaks, since it is known to be associated with gene body.
Comparing SICER and ZINBRA seems tricky, since FDR correction plays crucial role in peak length distribution.
ZINBRA with $fdr = 10^{-3}$ results in the closest distribution to SICER. Moreover we can see, that peak extension doesn't make sense for $fdr = 10^{-3}$.
Also threshold 3kbp looks natural to separate narrow and broad peaks.

Conclusion:

- ZINBRA is in general consistent with MACS2 in terms of narrow peaks, and with SICER in terms of broad peaks.
- SICER and ZINBRA are most consistent in terms of peaks distribution for ZINBRA $fdr = 10^{-3}$.
- MACS2 is not capable to call broad peaks even with –*broad* command line option.
- Q: *How to distinguish narrow and broad ChIP-Seq peaks?*
  A: Peak can be considered as *broad* if length is at least 3kbp.