A simple scaling normalization for comparing ChIP-Seq samples

Paul Manser, Mark Reimers

In ChIP-Seq and DNase-Seq experiments, the density of background reads can vary from sample to sample. Differences in background read densities between samples do not necessarily correspond to proportional changes of read densities in true ChIP-Seq peaks. Therefore, scaling by total library size as a means for normalizing called ChIP-Seq peaks across samples may be ineffective. We suggest a simple easily implemented alternative to scaling by total library size that scales only by the total number of reads mapped to called peaks. We then demonstrate the effectiveness of the modified scaling in K4me3 and K27ac ChIP-Seq data from the BrainSpan project as well as DNase-Seq data from the ENCODE project.

PeerJ PrePrints | https://peerj.com/preprints/175v1/ | v1 received: 27 Dec 2013, published: 27 Dec 2013, doi: 10.7287/peerj.preprints.175v1

A simple scaling normalization for comparing ChIP-Seq samples

PAUL MANSER[1,2], MARK REIMERS[1,2,3]

1. Department of Biostatistics, Virginia Commonwealth University, Richmond, VA 23284.
2. Virginia Institute for Psychiatric and Behavioral Genetics, Virginia Commonwealth University, Richmond, VA 23284.
3. Department of Psychiatry, Virginia Commonwealth University, Richmond, VA 23284.

1. Introduction

Although most ChIP-Seq experiments focus on finding 'peaks' of enrichment, a growing number of studies compare ChIP-Seq data across samples (Creyghton et al 2010). A natural step in normalizing ChIP-Seq data when comparing peaks between samples is to scale by library size as is commonly done for RNA-Seq data (Mortazavi et al. 2008). However different samples have different signal-to-noise ratios (SNRs) i.e. different levels of background reads. Therefore, peaks in different samples with the same heights can have different relative heights compared to their respective background levels. This issue was recognized by (Zhen et al. 2012), but their method allows one to compare only two samples at a time, and is thus unsuitable for group comparisons.

Figure 1 shows a one megabase region from chromosome 13 from two K27Ac biological replicates from the BrainSpan data (brainspan.org). The y-axes are scaled so that the peaks are visually comparable. We can see that the relationship between peak height and background level differs substantially between the two samples. Peaks called by MACS are indicated below the X-axis (Zhang et al. 2008). We can see in this case that an increase in library size does not imply proportional increases in both peaks and in background. Although background is typically low, it extends over the vast majority of the genome - typically less than 2% of the genome lies in peaks - and therefore a substantial fraction of reads (up to half) may count as background. Therefore, scaling by total library size for each sample will not necessarily make peak heights comparable across samples.

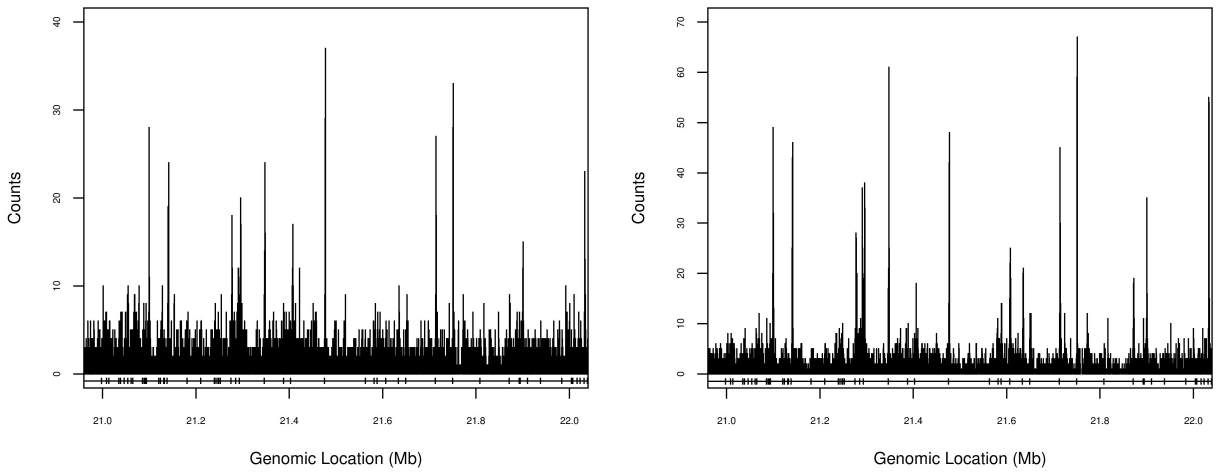


Figure 1: A one megabase region of chromosome 13 from two K27Ac cerebellum samples shows differing levels of background relative to peak heights. Called peaks using MACS are indicated at the bottom.

2. Methods

We suggest a modified scaling factor that scales only by the total number of reads mapped into called peaks rather than by whole library size. The set of called peaks for a set of samples is taken to be the union of the set of called peak intervals for each sample. This is typically only 1-2% of the genome. By effectively ignoring the differing levels of background, our method implicitly accounts for the different

PeerJ PrePrints | https://peerj.com/preprints/175v1/ | v1 received: 27 Dec 2013, published: 27 Dec 2013, doi: 10.7287/peerj.preprints.175v1

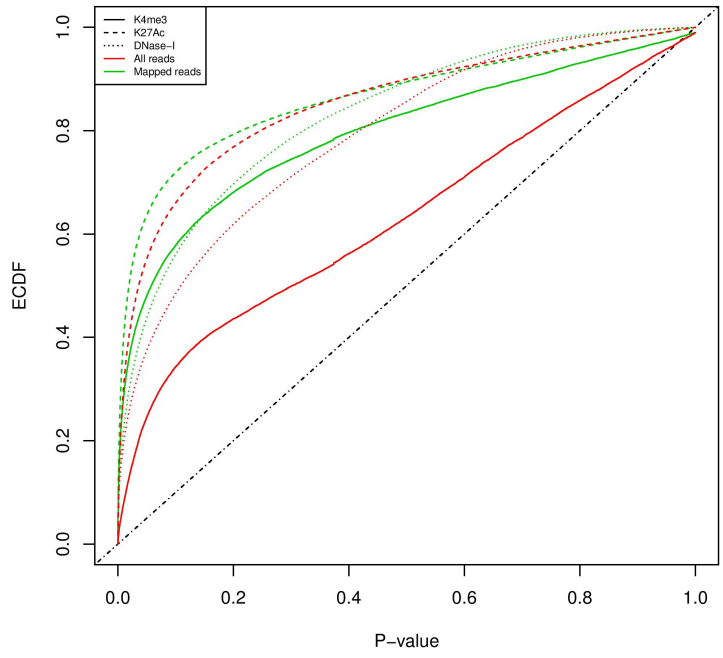
SNRs across samples. Since our method is implemented after peak calling, control samples used for peak calling are not required for normalization for purposes of comparing samples. Additionally, our method allows for implementation of standard downstream statistical analyses such as sample clustering and linear model fitting, as distinct from MAnorm, another ChIP-Seq normalization method, which allows only for pairwise comparison of peaks between two samples after normalization (Zhen et al. 2012). If we find N called peaks, we compute the scaled peak height for sample i and peak j as the original peak height X_{ij} scaled by the sum of all peak heights for that sample:

$$Z_{ij} = \frac{X_{ij}}{\sum_{j=1}^{N} X_{ij}} \quad (1)$$

3. Results

We demonstrate the effectiveness of our modified approach on K4me3 and K27ac data sets from BrainSpan (to appear on brainspan.org) and ENCODE DNase-Seq data (genome.ucsc.edu/ENCODE). The BrainSpan data consists of samples from 3 post-mortem brains, each sampled at cerebellum and prefrontal cortex. Our goal is to detect differences between these two brain regions. The ENCODE data consists of pairs of technical replicates of HeLa, GM12878, and two different astrocyte samples (NH-A and HAc) for which we want to again find peaks with different heights (Thurman et al 2012). Total read depths for DNase-Seq were estimated from a random sample of 1 kilobase intervals from chromosome 21 as the actual total read depths were not provided by ENCODE.

Scaling only by reads mapped to peaks decreases within group variability and increases power to detect differences between groups. For each peak in the master list, a standard two sample t-test was performed to detect differences between regions of cerebellum and prefrontal cortex in the BrainSpan data. Similarly, a one-way ANOVA was performed to test for differences in peaks between sample types for the ENCODE data. Figure 2 gives empirical cumulative density functions (ECDFs) of p-value distributions for each data set showing that there are more small p-values using the modified scaling method and therefore that our method increases power to detect differences for all three data types. Choosing a p-value level on the X-axis, the corresponding curve indicates the proportion of peaks with p-values less than the specified X-axis value. Therefore, a relative increase in power to detect differences is indicated by a steeper curve on the left side of the plot. The dashed black line indicates the theoretical ECDF corresponding to a completely flat p-value distribution which we would expect under the null hypothesis of no differences between groups.



PeerJ PrePrints | https://peerj.com/preprints/175v1/ | v1 received: 27 Dec 2013, published: 27 Dec 2013, doi: 10.7287/peerj.preprints.175v1

Figure 2: Empirical CDFs of p-value distributions testing for group differences show improved power when scaling by reads mapped to peaks compared to scaling using total read depth. Note that a larger fraction of differences between distinct groups, relative to differences between replicates, appear statistically significant.

4. Discussion

We have shown that differing signal to noise ratios occur in several widely-used data types used to assess chromatin modification using DNA sequencing. Our proposed modified scaling is a simple and effective method for accounting for read depths in a way that is robust to differing signal to noise ratios across samples. Furthermore it is simpler to compute than normal scaling in cases where the true read depths may not be known, and must be estimated from a subset of the data; this situation is common when using public data. Our approach reduces within-group variability and increases power to detect differences across groups.

References

Creyghton MP, Cheng AW, Welstead GG, Kooistra T, Carey BW, Steine EJ, Hanna J, Lodato MA, Frampton GM, Sharp PA, Boyer LA, Young RA, Jaenisch R: Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc Natl Acad Sci USA* 2010, 107:21931-21936.

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B: Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods* 2008, 5, 621-628.

Thurman RE, Rynes E, Humbert R, et al. The accessible chromatin landscape of the human genome. *Nature* 2012, 489:75-82.

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, Liu XS: Model-based analysis of ChIP-Seq (MACS). *Genome Biol* 2008, 9:R137.

Zhen S, Zhang Y, Yuan G, Orkin SH, Waxman DJ. MAnorm: a robust model for quantitative comparison of ChIP-Seq data sets. *Genome Biology* 2012, 13:R16.