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.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.175v2 | CC-BY 3.0 Open Access | received: 10 Apr 2014, published: 10 Apr 2014

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.

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 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:

$$Z_{ij} = \frac{X_{ij}}{\sum_{j=1}^{N} X_{ij}} \tag{1}$$

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.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.175v2 | CC-BY 3.0 Open Access | received: 10 Apr 2014, published: 10 Apr 2014