Shape matters: Differential peak detection for Chip-seq data sets

Gabriele Schweikert, Guido Sanguinetti ✉
University of Edinburgh, Edinburgh, United Kingdom

Motivation and Objectives

ChIP-Seq has rapidly become the dominant experimental technique in functional genomic and epigenomic research. Statistical analysis of ChIP-Seq data sets however remains challenging, due to the highly structured nature of the data and the paucity of replicates. Current approaches to detect differentially bound or modified genomic regions are mainly borrowed from RNA-Seq analysis tools, e.g. (Ross-Innes, 2012, Anders and Huber, 2010). With these methods a given peak is represented by a single number: the total number of reads mapping to the peak region. Any information encoded in the structure of the peak is ignored. However, the shape of an enrichment peak at a given genomic location tends to be highly reproducible across biological replicates and increasing evidence hints towards a functional role of these profile structures (The ENCODE Project Consortium, 2012). To complement count-based methods, we present MMDiff, a new non-parametric statistical testing methodology to identify significant shape differences in profile patterns of enrichment peaks between different conditions.

Methods

The underlying idea is to treat each peak as a *distribution* over a finite space given by the starting positions of all reads. The problem of testing for differential binding is then reduced to testing whether two samples are generated by the same probability distribution. As there is a large variability for observed peak profiles at different genomic locations we cannot make any assumption about the type of distribution. We therefore take advantage of recent kernel-based tests developed in the machine learning community (Gretton *et al.*, 2012). Here, the non-linearity of the original data is captured with a positive definite Kernel and the data is mapped into a high-dimensional reproducing Kernel Hilbert space (RKHS). In the RKHS the mean element of a distribution contains the information of all higher-order moments. Furthermore, the distance between the mean elements of two distributions, the maximum mean discrepancy (MMD), can be used as test statistic. Intuitively, the greater the distance, the more different the distributions are. Here we use the 5' position of the mapped reads as features and an RBF Kernel, where the width of the Kernel is heuristically determined. To obtain empirical p-values we compute the probability of observing MMD values between biological replicates which are at least as extreme as those observed between conditions. To correct for multiple testing we compute false discovery rates (FDRs) according to (Benjamini and Hochberg, 1995).

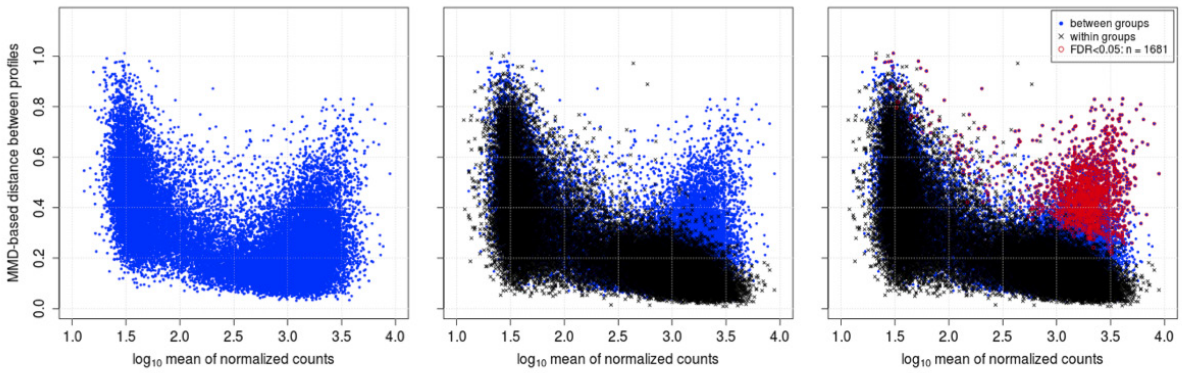

Figure 1. Scatter plots showing MMD as a function of averaged normalized total counts, where each dot represent one examined promoter. Left: MMD determined between WT and Null. Middle: Overlayed are the MMDs for biological replicates (black) Right: Additionally, profiles that are significantly different in WT vs Null ($p < 0.05$) are shown in red.

Results and Discussion

One of the best studied epigenomic marks is trimethylation of Lysine 4 at histone 3 (H3K4me3), which is associated with active gene promoters. However, the mechanisms and dynamics by which this mark is established at its target locations are not well understood. We apply our method to a recently published data set by (Clouaire *et al.*, 2012), which examines the role of one key player, the DNA binding protein Cfp1. To this end, H3K4me3 profiles in wt ES cells and mutant cell lines lacking Cfp1 are compared.

Our empirical analysis shows that MMDiff is complementary to count based methods, that it provides highly reproducible results and that it is able to detect biologically relevant changes in histone modifications.

To make the method available to a wider range of users we have developed an R package, called MMDiff which is released with the la-test Bioconductor version (2.12).

Acknowledgements

We would like to thank Arthur Gretton, Rory Stark and Gunnar Raetsch for helpful discussions. Shaun Webb is thanked for computing support. Thomas Clouaire and Adrian Bird provided the data and helped with discussions.

G.Schw. acknowledges support from EC through the FP7- Marie Curie project "Epigenome Informatics". G.S. acknowledges support from the European Research Council through grant MLCS306999.

References

Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biol* **11**(10), R106.

Benjamini and Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing; Journal of the royal statistical society. series a, Statistics in Society. JRSS. Series A, Statistics in society, **57**(1), 289.

Clouaire, T., Webb, S., Skene, P., Illingworth, R., Kerr, A., et al. (2012). Cfp1 integrates both CpG content and gene activity for accurate H3K4me3 deposition in embryonic stem cells. *Genes Dev* **26**(15), 1714– 28.

The ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schoelkopf, B., and Smola, A. J. (2012). A kernel two-sample test. *JMLR* **13**, 723–773.

Ross-Innes, C. S., Stark, R., Teschendorff, A. E., Holmes, K. A., Ali, H. R., et al. (2012). Differential Oestrogen receptor binding is associated with clinical outcome in breast cancer. *Nature* **481**(7381), 389–93.