Title
Differential Enriched Scan 2 (DEScan2): a fast pipeline for broad peak analysis

List of authors
Righelli Dario(1,2), Koberstein John(3), Nancy Zhang(4), Angelini Claudia(1), Peixoto Lucia(5), Risso Davide(6,7).

Affiliations:
1. CNR, Institute for Applied Mathematics “M. Picone”, Naples, Italy
2. DISA-MIS, Università di Salerno, Salerno, Italy
3. Oregon Health and Science University, Portland, Oregon, USA
4. Wharton University of Pennsylvania, Philadelphia, Pennsylvania, USA
5. Washington State University, Spokane, Washington, USA
6. Department of Statistical Sciences University of Padua, Padua, Italy
7. Weill Cornell Medical College, New York, New York, USA

Corresponding author:
Dario Righelli
Via Pietro Castellino 111, 80131, Napoli, Italy
Email address: d.righelli@na.iac.cnr.it

Abstract:

Nowadays, the analysis of RNA-seq and BS-Seq can be considered well established, whereas the analysis of broad peaks data as Sono-Seq/ATAC-Seq and histone modification (HM) ChIP-Seq is still challenging. To fill the gap in existing methods, we present DEScan2 a novel bioconductor package [2] for the analysis of broad peaks data.
The method consists of three main steps: 1) a peak caller, 2) peak filtering and alignment across replicates and 3) a method to efficiently compute a count matrix of the filtered peaks.
Using an already published ATAC-Seq dataset for chromatin accessibility our method shows interesting results, also by comparing it with other well-known tools for this kind of data analysis.

Introduction:

Next Generation Sequencing (NGS) techniques revolutionized biology enabling to examine biological processes by different points of view producing a vast amount of data. In this context, the most widely investigated aspects are the transcriptional level with RNA-Seq, the epigenetic state with ChIP-Seq and BS-seq and the chromatin accessibility with Sono-Seq/ATAC-Seq. Nowadays, the analysis of RNA-seq, BS-Seq can be considered well established, whereas the analysis of Sono-Seq/ATAC-Seq and histone modification (HM) ChIP-Seq is still challenging. Despite the lack of robust computational methods for their analysis, recent studies have demonstrated the relevance of Sono-Seq/ATAC-Seq to unveil the significant role of open state regions of chromatin linked to diseases like autism [Koberstein2018], among many others. To fill the gap in existing methods, we present DEScan2 a novel bioconductor package [descan2018] for the analysis of broad peaks data as Sono-Seq/Atac-Seq and Histone Modification ChIP-Seq.

Materials & Methods:

The method consists of three main steps: 1) a peak caller, 2) peak filtering and alignment across replicates and 3) a method to efficiently compute a count matrix of the filtered peaks.

The peak caller in step 1) is a moving scan window that compares the coverages within a sliding window to the coverages in a larger region outside the window, using a Poisson likelihood and providing a final score for each detected peak. However, the package can work with any external peak

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27357v1 | CC BY 4.0 Open Access | rec: 19 Nov 2018, publ: 19 Nov 2018

caller returning results in terms of bed files, indeed the package provides additional functionalities to load bed files of peaks and handle them as GenomicRanges structures [Lawrance2013].
The filtering and alignment step is aimed to determine if a peak is a "true peak" on the basis of its replicability in other samples. Basing on this idea, we developed this step on two user-given thresholds, one on the peaks's score and another on the minimum number of samples, in order to filter out those peaks not present in at least the given number of samples. In the light of this, the user can decide the minimum number of samples where each peak has to be detected. Moreover, a further threshold can be used over the peak score.
Finally, the third step produces a count matrix where each column is a sample and each row a filtered peak computed in the filtering step. The value of the matrix cell is the number of reads for the peak in the sample.
Furthermore, our package provides several functionalities for GenomicRanges data structure handling**.** One over the others gives the possibility to split a GenomicRanges over the chromosomes to speed-up the computations parallelizing them over the chromosomes.

Results:
We choose an already published dataset [Su2017] for chromatin accessibility for illustrating our method performances. The dataset describes in vivo adult mouse dentate granule neurons before and after synchronous neuronal activation using ATAC-Seq and RNA-Seq technologies, of which we selected the first 8 samples, 4 per each condition.

Using this dataset we show a possible pipeline (as illustrated in Figure 1, where DEScan2 steps are highlighted in yellow) for ATAC-Seq data analysis and their integration with RNA-Seq data, by comparing at each step our performances with other well-known software for the analysis of this type of data.

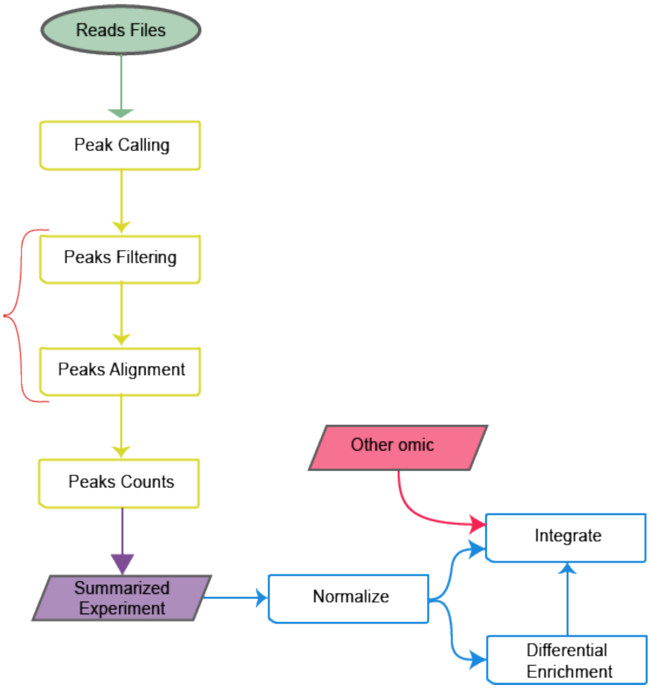

Figure 1: A possible pipeline for broad peaks data analysis and integration using DEScan2.

References:

1. KOBERSTEIN, J.N., et al. Learning-dependent chromatin remodeling highlights noncoding regulatory regions linked to autism. Sci. Signal., 2018, 11.513: eaan6500.
2. https://doi.org/doi:10.18129/B9.bioc.DEScan2

3. LAWRENCE, M, et al. Software for Computing and Annotating Genomic Ranges.. *PLoS Computational Biology*, 2013.
4. SU, Y., et al. Neuronal activity modifies the chromatin accessibility landscape in the adult brain. Nature neuroscience, 2017, 20.3: 476.