Genome analysis

chromswitch: a flexible method to detect chromatin state switches

Selin Jessa and Claudia L. Kleinman*

Department of Human Genetics, McGill University, Montreal, QC, Canada and Lady Davis Research Institute, Jewish General Hospital, Montreal, QC, Canada

*To whom correspondence should be addressed.

Abstract

Summary: Chromatin state plays a major role in controlling gene expression, and comparative analysis of ChIP-seq data is key to understanding epigenetic regulation. We present chromswitch, an R/Bioconductor package to integrate epigenomic data in a defined window of interest to detect an overall switch in chromatin state. Chromswitch accurately classifies a benchmarking dataset, and when applied genome-wide, the tool successfully detects chromatin changes that result in brain-specific expression.

Availability and implementation: Chromswitch is implemented as an R package available from Bioconductor at https://bioconductor.org/packages/chromswitch. All data and code for reproducing the analysis presented in this paper are available at https://doi.org/10.5281/zenodo.1101260.

Contact: claudia.kleinman@mcgill.ca

Supplementary information: Supplementary data are available at Bioinformatics online.

1 Introduction

A routine question in epigenomic studies is how histone modifications and chromatin state differ among groups of samples. These can represent, for instance, distinct developmental, differentiation, or disease states. Several methods exist to process and interpret the raw signal from chromatin immunoprecipitation sequencing (ChIP-Seq), typically by identifying regions of significant enrichment (peaks) (Furey, 2012) or learning a genome-wide chromatin state segmentation (Ernst and Kellis, 2012; Hoffman et al., 2012; Mammana and Chung, 2015; Song and Chen, 2015). Combining the output of these methods across samples is not straightforward, and various tools have been developed for differential analysis at the peak level (Chen et al., 2015; Liang and Keleş, 2012; Ross-Innes et al., 2012), or for the joint analysis of samples to discover regions of change throughout the genome (Sohn et al., 2015; Yen and Kellis, 2015; Zeng et al., 2013). However, answering the common question of whether a large chromatin state change occurs in a specific, predefined region of interest usually requires ad hoc downstream analyses to interrogate inferred peaks or chromatin state assignments across the region.

Here, we present a flexible strategy to identify chromatin state changes in genomic windows. Given a query region and epigenomic features in two biological conditions as input, chromswitch uses hierarchical clustering and external validity measures to predict a chromatin state switch.

2 Materials and methods

2.1 Input and pre-processing

Chromswitch takes as input: (i) a query region specified by its genomic coordinates and (ii) BED files storing epigenetic features previously inferred for each sample, such as ChIP-seq peak calls or chromatin state segmentations. For simplicity, we refer to these features as peaks in what follows, but the algorithm is applicable to any feature represented by genomic coordinates and some associated metrics. For example, peak calling tools typically output fold change and significance values for each peak. Chromswitch can filter peaks using user-defined thresholds on these metrics. Next, these metrics are normalized genome-wide on a per-sample basis: for a vector of genome-wide values for one metric, we rescale the central 99% of values to the range [0, 1] and bound lower and upper outliers to 0 and 1 respectively. This step accounts for some of the technical variation between samples (e.g. in ChIP-seq efficiency) and allows for comparison of metrics with different ranges.

2.2 Feature matrix construction

Chromswitch then constructs a sample-by-feature matrix from the data to use as input for clustering, following one of two strategies (Fig. 1a, Supplementary Fig. S1). In the summary strategy, the feature vector for
each sample contains summary statistics compiled from all peaks present in the query region. These can include the mean, median and max of each metric, as well as the fraction of the region overlapped by peaks or their average number. In the binary strategy, in turn, a set of unique peaks is defined as the union of all peaks present in at least one sample, where peaks are collapsed if they have a reciprocal overlap above a user-specified minimum fraction. Peaks can also be merged if they are separated by less than a user-specified gap. The feature vector for each sample is, in this case, a binary vector encoding the presence or absence of each unique peak in that sample (Supplementary Fig. S1, see Supplementary Methods for details).

2.3 Chromatin state switch calls
Chromswitch then clusters samples using hierarchical clustering and selects the partition with the highest average Silhouette width, which measures cluster cohesion and separation (Rousseeuw, 1987). Inferred cluster assignments are then scored using a consensus (mean) of three external cluster validity indices designed to validate clustering solutions against ground-truth class labels: the Adjusted Rand Index, the Normalized Mutual Information and the V measure (Supplementary Table S1). The consensus score can then be used to threshold or rank putative chromatin state switches (see Supplementary Methods).

3 Results
To evaluate chromswitch, we first assembled a benchmark dataset comprising data for 7 brain and 16 other tissues from the NIH Roadmap Epigenomics Project (Roadmap Epigenomics Consortium, 2015). We included ChlP-seq peaks for the H3K4me3 mark (a histone modification associated with active transcription), DNase I hypersensitive sites (associated with open chromatin), RNA-seq (measuring expression, the functional consequence of chromatin changes) and chromatin state assignments by a ChromHMM model trained on 60 Roadmap epigenomes. We identified a set of 60 5kbp...
regions surrounding transcription start sites (TSS) of genes where a clear chromatin state switch between brain and other tissues was evident in a genome browser for all data types, and 60 control regions (Supplementary Table S2). Area under the Receiver-Operating Characteristic curve (AUROC) calculation shows that chromswitch accurately classified these benchmark regions over a range of combinations of input data (Fig. 1b, Supplementary Fig. S2, and Supplementary Methods). Chromswitch preserved accuracy with small sample sizes (e.g. AUC = 0.95 for a dataset of 4 samples per condition, Supplementary Fig. S3) and high class imbalance (Supplementary Fig. S3). Furthermore, the method is robust to changes in tuning parameters when constructing the feature matrix using the binary strategy (gap and minimum reciprocal overlap, Supplementary Fig. S4). An exploration of the effect of variations in window size, which determines the signal to noise ratio, is presented in Supplementary Figure S5. Finally, chromswitch is applicable to the analysis of sparse marks covering broad domains, such as H3K27me3 (Supplementary Figs S6–S8).

We next evaluated chromswitch genome-wide, assessing whether it could identify chromatin switches that result in tissue-specific expression. Using as input either H3K4me3 peaks, DNase I hypersensitivity sites or ChromHMM assignments to the state ‘active transcription’, we applied chromswitch to 5 kbp windows surrounding all annotated TSS in the RefSeq annotation (see Supplementary Methods). Genes for which chromswitch detected an active state in brain samples and a silent state otherwise were validated using gene expression data in corresponding tissues from the Genotype-Tissue Expression Project (GTEx). We found that brain-specific chromatin state changes detected by chromswitch were recapitulated at the expression level (Fig. 1c, Supplementary Figs S9, S10). As expected, the median fold change of expression across candidate switches increased as a function of the threshold score used for predictions (Fig. 1d, Supplementary Fig. S11).

**4 Conclusion**

Chromswitch equips users to detect spatial, temporal, or tissue-specific chromatin state changes in specific query regions. The tool is free from data-intensive training steps, suitable for histone marks with diverse profiles and applicable downstream of existing tools for chromatin analysis. The method is robust to small sample sizes and high class imbalance, common scenarios in functional genomics projects. Chromswitch is implemented as an R package, designed for modularity and ease of use to facilitate investigation into epigenetic regulation and its consequences.

**Acknowledgements**

We thank Nicolas De Jay, Alexis Blanchet-Cohen, Steven Hébert and Karine Choquet for code review and testing. We thank Maud Marques and Vincenzo Forgetta for critical reading of the manuscript.

**Funding**

This work was supported by the Natural Sciences and Engineering Research Council of Canada (RGPIN-2016-04911). C.L.K. receives salary awards from the Fonds de Recherche du Québec-Santé. S.J. is a recipient of a Canadian Institutes of Health Research award. The computing and networking infrastructure were provided by Calcul Quebec and Compute Canada.

**Conflict of Interest**: none declared.

**References**

Chen, L. et al. (2015) A novel statistical method for quantitative comparison of multiple ChIP-seq datasets. *Bioinformatics*, 31, 1889–1896.

Ernst, J. and Kellis, M. (2012) ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods*, 9, 215–216.

Furey, T.S. (2012) ChIP-seq and Beyond: new and improved methodologies to detect and characterize protein-DNA interactions. *Nat. Rev. Genet.*, 13, 840–852.

Hoffman, M.M. et al. (2012) Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nat. Methods*, 9, 473–476.

Li, T. and Keles, S. (2012) Detecting differential binding of transcription factors with ChIP-seq. *Bioinformatics*, 28, 121–122.

Mannan, A. and Chung, H.R. (2015) Chromatin segmentation based on a probabilistic model for read counts explains a large portion of the epigenome. *Genome Biol.*, 16, 151.

Roadmap Epigenomics Consortium (2015) Integrative analysis of 111 reference human epigenomes. *Nature*, 518, 317–330.

Ross-Innes, C.S. et al. (2012) Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. *Nature*, 481, 389–393.

Rousseeuw, P.J. (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Computat. Appl. Math.*, 20, 53–65.

Sohn, K.-A. et al. (2015) hiHMM: bayesian non-parametric joint inference of chromatin state maps. *Bioinformatics*, 31, 2066–2074.

Song, J. and Chen, K.C. (2015) Spectacle: fast chromatin state annotation using spectral learning. *Genome Biol.*, 16, 33.

Yen, A. and Kellis, M. (2015) Systematic chromatin state comparison of epigenomes associated with diverse properties including sex and tissue type. *Nat. Commun.*, 6, 7973.

Zeng, X. et al. (2013) jMOSAICs: joint analysis of multiple ChIP-seq datasets. *Genome Biol.*, 14, R38.