monaLisa: an R/Bioconductor package for identifying regulatory motifs

Dania Machlab[1,2,3], Lukas Burger[1,2], Charlotte Soneson[1,2], Filippo M. Rijli[1,3], Dirk Schübeler[1,3], and Michael B. Stadler[1,2,3, *]

[1]*Friedrich Miescher Institute for Biomedical Research, Basel, Switzerland*
[2]*SIB Swiss Institute of Bioinformatics, Basel, Switzerland*
[3]*University of Basel, Switzerland*
[*]*to whom correspondence should be addressed: michael.stadler@fmi.ch*

November 30, 2021

Abstract

Proteins binding to specific nucleotide sequences, such as transcription factors, play key roles in the regulation of gene expression. Their binding can be indirectly observed via associated changes in transcription, chromatin accessibility, DNA methylation and histone modifications. Identifying candidate factors that are responsible for these observed experimental changes is critical to understand the underlying biological processes. Here we present *monaLisa*, an R/Bioconductor package that implements approaches to identify relevant transcription factors from experimental data. The package can be easily integrated with other Bioconductor packages and enables seamless motif analyses without any software dependencies outside of R.
Availability: *monaLisa* is implemented in R and available on Bioconductor at `https://bioconductor.org/packages/monaLisa` with the development version hosted on GitHub at `https://github.com/fmicompbio/monaLisa`.
Contact: michael.stadler@fmi.ch

Introduction

Binding proteins that interact with specific nucleotide sequences, such as transcription factors (TFs), play key roles in the regulation of cellular functions and organismal development (Spitz & Furlong, 2012). Identifying candidate proteins that could play regulatory roles in development or act as drivers for an observed biological response is thus a crucial step in the interpretation of genomics data, such as absolute values or changes of DNA methylation, chromatin modifications, accessibility or transcription. There are various existing tools and methods for regulatory protein identification via their binding motifs. Most of these are command line tools or web servers that cannot be easily integrated with other Bioconductor (Huber *et al.*, 2015) packages for a seamless analysis in R, or they require the installation of additional software outside of R. Conceptually, many of these methods can be roughly divided into two types: enrichment-based methods that compare motif occurrences between sets of sequences, and model-based methods that estimate motif importance from their ability to explain experimental observations. Here, we present *monaLisa*, short for "motif analysis with Lisa", an R/Bioconductor package that implements both of these approaches and enables seamless motif identification analyses in R.

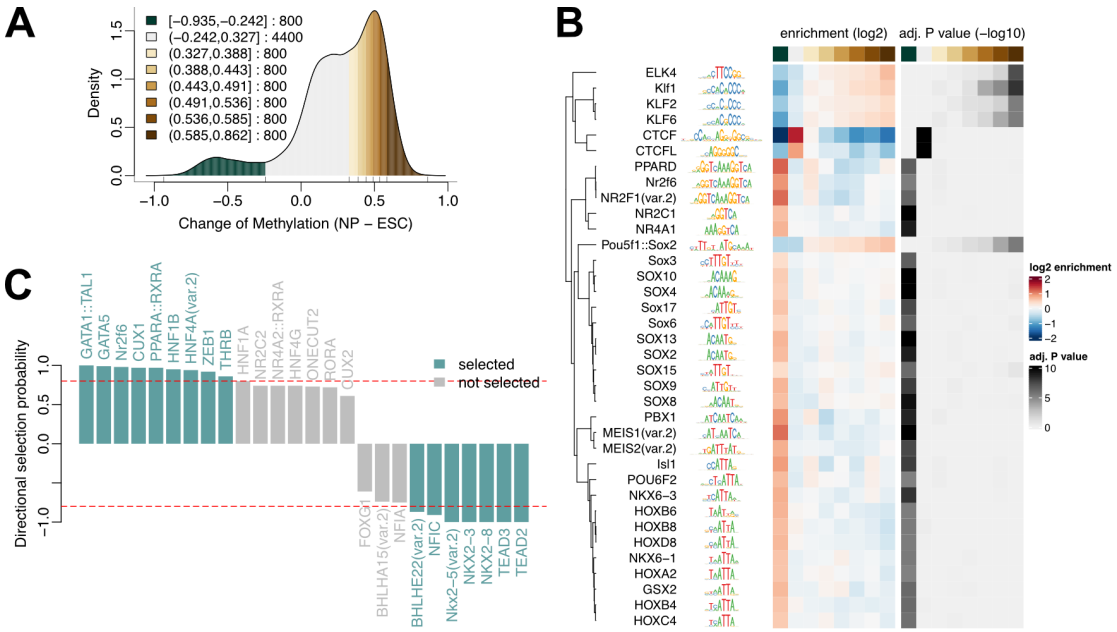


Figure 1: **A,B** Analysis of methylation changes between neuronal progenitors (NP) and embryonic stem cells (ESC): Binned density of methylation levels (A), with bin boundaries and sizes given in the legend, and enrichment and significance heatmaps (B) of motifs (rows) across bins (columns). **C** Analysis of accessibility changes between liver and lung: Directional selection probabilities for motifs identified using stability selection.

Usage and examples

Enrichment-based tools like *HOMER* (Heinz *et al.,* 2010) and *MEME* (Bailey *et al.,* 2015) identify novel or known motifs enriched in a given set of sequences compared to a suitable background. In *monaLisa*, this is done by first binning sequences, for example gene promoters or enhancers, according to their associated values. In the example here we use changes of DNA methylation between mouse embryonic stem cells and derived neuronal progenitors (Stadler *et al.,* 2011; Burger *et al.,* 2013) (Fig. 1A). A collection of known motifs, for example from the *JASPAR2020* package (Fornes *et al.,* 2020), are then evaluated for enrichment in each bin compared to the background, using *HOMER*'s normalization method to adjust for differences in sequence composition. This can also be diagnosed using available visualization functions (Suppl. Fig. 1A and B). Several ways to define background sequences are available, and the results can be visualized as a heatmap (Fig. 1B), that shows the enrichment of each motif in each bin, compared to all other bins. Additional confidence can often be gained by focusing on motifs for which the enrichment scales with the numerical value under consideration. *monaLisa* also offers to search for enriched k-mers (oligonucleotides of length k), which is particularly useful to complement the motif enrichment analysis and identify potential gaps in the database of known motifs (Suppl. Fig. 1C).

In the bin-based approach, motifs are analyzed independently of each other. In contrast, methods such as *REDUCE* (Roven & Bussemaker, 2003) or *ISMARA* (Balwierz *et al.,* 2014) use linear regression approaches to identify regulatory motifs that are most likely to explain the observed numerical responses. A similar model-based approach is also available in *monaLisa*, but uses a different regression framework: randomized lasso stability selec-

tion, introduced by Meinshausen & Bühlmann, 2010, with the improved error bounds proposed by Shah & Samworth, 2013. Regression is performed on several random subsets of the data to calculate motif selection probabilities. This type of regularization has advantages in selecting variables consistently, demonstrating better error control and not depending strongly on the initial regularization chosen (Meinshausen & Bühlmann, 2010). For illustration, we have used *monaLisa*'s regression with stability selection to identify transcription factor motifs that could explain the observed changes of accessibility between mouse liver and lung (data from The ENCODE Project Consortium, 2012) which represents the response variable. The predictor matrix consists of predicted binding sites for each TF, and additional variables, such as G+C composition, can also be included. The results can be visualized as stability paths (Suppl. Fig. 2), that show the selection probability for each motif as a function of the regularization steps, or as the final selection probabilities (Fig. 1C) combined with a sign to indicate if a motif correlates positively or negatively with changes in accessibility.

The illustrating examples and datasets above are included and described in detail in the package vignette. In addition to enrichment- and regression-based motif identification methods, *monaLisa* further provides helpful functions for motif analyses, including functions to predict motif matches and calculate similarity between motifs.

Summary

monaLisa is an R/Bioconductor package for motif analyses applicable to sequences with associated numerical data. Regulatory motifs explaining the observations can be identified using two complementary approaches. *monaLisa* requires no additional software tools and can be easily integrated with other Bioconductor packages for seamless analyses in R.

Acknowledgements

We would like to thank the members of the Rijli, Schübeler and Stadler groups, Luca Giorgetti, Florian Geier and our colleagues from the Novartis Institutes for Biomedical Research for suggestions on the software. Research in the groups of the authors is supported by the Novartis Research Foundation. DM was supported by the Swiss National Science Foundation (grant 31003A_175776 to FMR). DS furthermore acknowledges support from the Swiss National Science Foundation (310030B_176394) and the European Research Council under the European Union's (EU) Horizon 2020 research and innovation program grant agreements (ReadMe-667951 and DNAaccess-884664).

References

1. Bailey, T. L., Johnson, J., Grant, C. E. & Noble, W. S. The MEME Suite. *Nucleic Acids Research* **43,** W39–W49. ISSN: 0305-1048 (W1 July 1, 2015).
2. Balwierz, P. J. *et al.* ISMARA: automated modeling of genomic signals as a democracy of regulatory motifs. *Genome Research* **24,** 869–884. ISSN: 1088-9051, 1549-5469 (May 1, 2014).
3. Burger, L., Gaidatzis, D., Schübeler, D. & Stadler, M. B. Identification of active regulatory regions from DNA methylation data. *Nucleic Acids Research* **41,** e155–e155. ISSN: 0305-1048 (Sept. 1, 2013).
4. Fornes, O. *et al.* JASPAR 2020: update of the open-access database of transcription factor binding profiles. *Nucleic Acids Research* **48,** D87–D92. ISSN: 0305-1048 (D1 Jan. 8, 2020).
5. Heinz, S. *et al.* Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Molecular Cell* **38,** 576–589. ISSN: 1097-2765 (May 28, 2010).

6. Huber, W. *et al.* Orchestrating high-throughput genomic analysis with Bioconductor. *Nature Methods* **12,** 115–121. ISSN: 1548-7105 (Feb. 2015).

7. Meinshausen, N. & Bühlmann, P. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **72,** 417–473. ISSN: 1467-9868 (2010).

8. Roven, C. & Bussemaker, H. J. REDUCE: an online tool for inferring cis-regulatory elements and transcriptional module activities from microarray data. *Nucleic Acids Research* **31,** 3487–3490. ISSN: 0305-1048 (July 1, 2003).

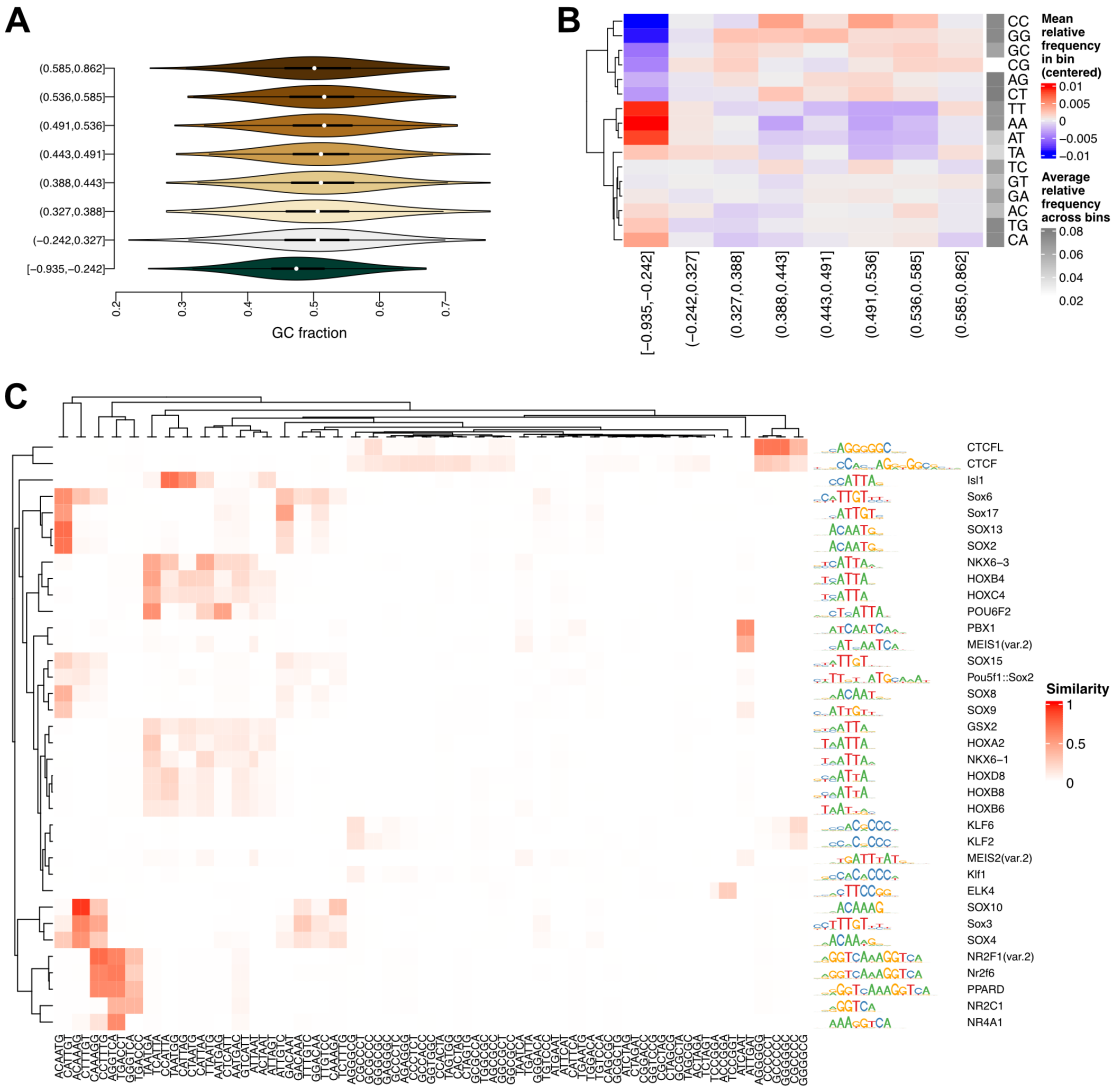
9. Shah, R. D. & Samworth, R. J. Variable selection with error control: another look at stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **75,** 55–80. ISSN: 1467-9868 (2013).

10. Spitz, F. & Furlong, E. E. M. Transcription factors: from enhancer binding to developmental control. *Nature Reviews Genetics* **13,** 613–626. ISSN: 1471-0064 (Sept. 2012).

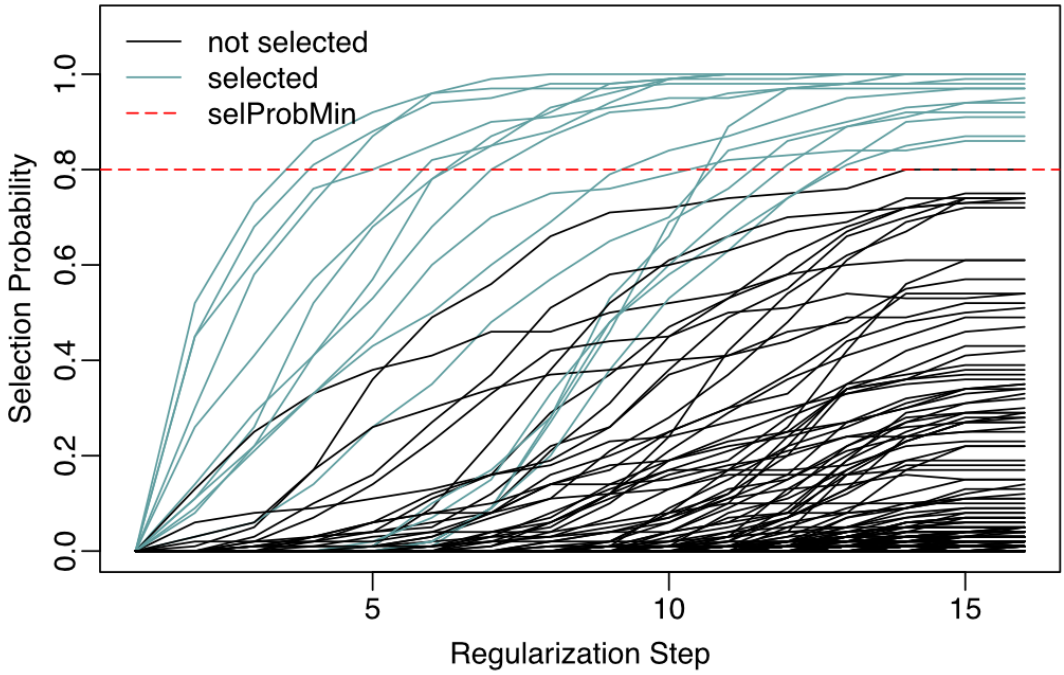
11. Stadler, M. B. *et al.* DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature* **480,** 490–495. ISSN: 1476-4687 (Dec. 2011).

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

Supplementary Figures



Supplementary Figure 1: Binned enrichment analysis of methylation changes between embryonic stem cells and neuronal progenitors. **A** Distributions of the fraction of G+C bases for sequences in each bin. **B** Differences of dinucleotide frequencies for sequences in each bin, relative to the mean over all bins. **C** Similarities between enriched motifs (rows) and k-mers (columns). Potential gaps in the motif analysis can be identified as k-mers without similar motifs.



Supplementary Figure 2: Regression-based analysis of accessibility changes between liver and lung. Stability paths showing the selection probability for each motif as a function of the regularization step, with the legend indicating the colors of selected motifs and the selection probability cutoff.