
**Combining transcription factor binding affinities with an open chromatin prior for accurate gene expression prediction**

Florian Schmidt[1,2], Nina Gasparoni3, Gilles Gasparoni[3], Kathrin Gianmoena[4], Cristina Cadenas[4], Julia K. Polansky[5], Peter Ebert[2,6], Karl Nordström[3], Matthias Barann[7], Anupam Sinha[7], Sebastian Fröhler[8], Jieyi Xiong[8], Azim Dehghani Amirabad[1,2,6], Fatemeh Behjati Ardakani[1,2], Barbara Hutter[9], Gideon Zipprich[10], Bärbel Felder[10], Jürgen Eils[10], Benedikt Brors[9], Wei Chen[8], Jan G. Hengstler[4], Alf Hamann[6], Thomas Lengauer[2], Philip Rosenstiel[7], Jörn Walter[3], and Marcel H. Schulz[1,2]

*1: Cluster of Excellence for Multimodal Computing and Interaction, Saarland University, Saarbrücken, 66123, Germany.*
*2: Computational Biology & Applied Algorithmics, Max Planck Institute for Informatics, Saarbrücken, 66123, Germany.*
*3: Department of Genetics, University of Saarland, Saarbrücken, Germany.*
*4: Leibniz Research Centre for Working Environment and Human Factors IfADo, Ardeystrasse 67, Dortmund, 44139, Germany.*
*5: Experimental Rheumatology, German Rheumatism Research Centre, Berlin, Germany*
*6: International Max Planck Research School for Computer Science, Saarbrücken, 66123, Germany.*
*7: Institute of Clinical Molecular Biology, Christian-Albrechts-University, Kiel, Germany.*
*8: Berlin Institute for Medical Systems Biology, Max-Delbrück Center for Molecular Medicine, Berlin, Germany.*
*9: Applied Bioinformatics, Deutsches Krebsforschungszentrum, Heidelberg, Germany*
*10: Data Management and Genomics IT, Deutsches Krebsforschungszentrum, Heidelberg, Germany*

**Abstract**
The binding and contribution of Transcription Factors (TFs) to cell specific gene expression is often deduced from open- chromatin measurements to avoid cost and labour intensive TF ChIP-seq assays. It is important to develop reliable and fast computational methods for accurate TF binding prediction in Open chromatin regions (ORCs). Here, we report a novel segmentation- based method, TEPIC, to predict TF binding by combining sets of OCRs with position weight matrices. TEPIC can be applied to various open chromatin data, e.g. DNaseI-Seq and NOMe-Seq, using either peaks or footprints as input data. TEPIC computes TF affinities as a quantitative measure (parameter) of TF binding strength and we show that low affinity binding sites predicted in this way improve performance over a simple presence/absence classification. Further, we show that while footprints called from OCRs capture most essential TF binding events, OCR peaks deliver the best prediction performance. Using machine learning techniques, we assessed the importance of individual TFs for gene expression and found that TEPIC scores nearly reach the quality of TF ChIP-seq data. Finally we show that TEPIC predicts all major known key transcriptional regulators in primary human hepatocytes and CD4+ T-cells emphasizing the reliability and applicability of our method.