DOI 10.18699/SBB-2022-036

# Transcription factor binding sites: data integration, stable identifiers and incremental builds

Kolmykov S.[1, 2]*, Kondrakhin Y.[2, 3], Sharipov R.[1, 2, 4], Yevshin I.[1, 2], Ryabova A.[1, 2], Kolpakov F.[1, 2, 3]

[1] *Sirius University of Science and Technology, Sochi, Russia*

[2] *BIOSOFT.RU, LLC, Novosibirsk, Russia*

[3] *Federal Research Center for Information and Computational Technologies, Novosibirsk, Russia*

[4] *Novosibirsk State University, Novosibirsk, Russia*

** kolmykovsk@gmail.com*

**Key words:** transcription factor binding sites, data integration, stable identifiers, incremental builds

*Motivation and Aim*: Transcriptional regulation is a complex mechanism at different levels, and transcription factors (TF) play a key role in this process. Most TFs bind to DNA by recognizing their short regulatory elements called TF binding sites (TFBSs). These TFBSs are the key components of transcriptional regulation, as well. One of the most widely used methods for experimental identification of TF binding sites (TFBS) is chromatin immunoprecipitation followed by high throughput sequencing (ChIP-seq). Currently, TFBS datasets from various ChIP-seq experiments are being accumulated in databases such as GTRD (http://gtrd.biouml.org), ChIP-Atlas (https://chip-atlas.org/), ReMap (https://remap2022.univ-amu.fr/) and ENCODE (https://www.encodeproject.org/).

*Methods and Algorithms*: For meta-processing of individual ChIP-seq datasets for a given TF, the METARA method (METa Analysis of ChIP-seq datasets using a Rank Aggregation approach) was previously developed. It produced a single integrated meta-dataset (known as a metacluster set in GTRD) by processing datasets obtained from individual ChIP-seq experiments for the given TF. It is important to note that the special Rank Aggregation score (RA-score) was assigned to each meta-cluster. It evaluates the quality (or reliability) of each meta-cluster.

*Results*: We have developed a new method IMETARA (Incremental METARA). There are two main reasons for creating IMETARA: (a) to significantly reduce the run time of meta-cluster building; (b) introduce stable identifiers for each meta-cluster. Thus, IMETARA only recalculates the RA scores for existing meta-clusters and adds novel meta-clusters when several new ChIP-seq experiments are added to the ChIP-seq data already collected in GTRD.

It is important to note that there is a strong IMETARA continuity, namely that most (99.3–99.9 %) of the meta-clusters identified by METARA are also identified by IMETARA for many TFs. Transition from METARA to IMETARA makes it easier to control the completeness of the resulting set of meta-clusters. For this purpose, it is sufficient to consider the dependence of such a characteristic as "Percentage of new meta-clusters at this IMETARA step". Figure 1 demonstrates that the new steps result in only 5–10 % of new meta-clusters for ESR1 being included, while this percentage for FOXA1 is about 35–40 %.

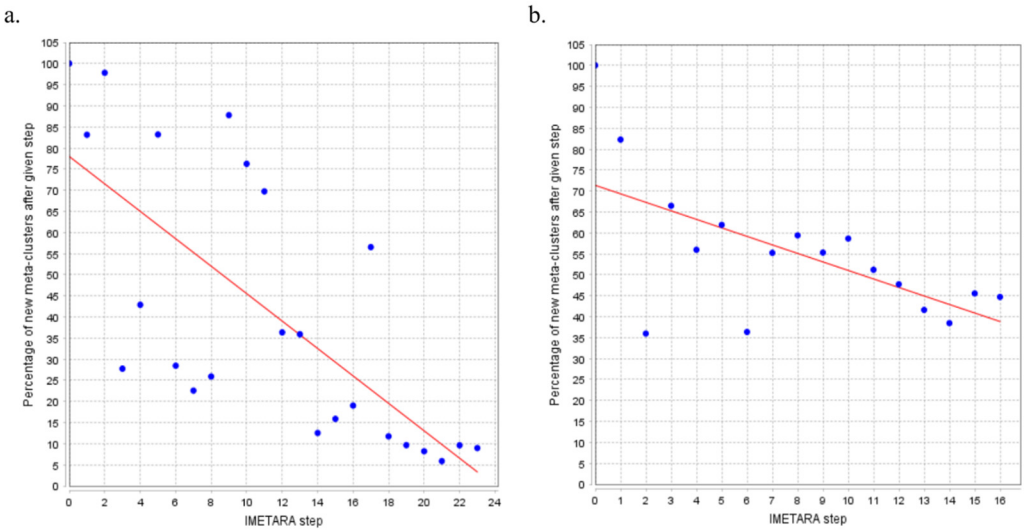


**Fig. 1.** Percentage of new meta-clusters at the given IMETARA step for (a) ESR1 and (b) FOXA1

The RA scores can be interpreted as reliability scores. As a result, the algorithm of decomposition of a sample of RA-scores into several components of normal mixture (Fig. 2) demonstrates that meta-clusters can be divided into 2–3 reliability groups.

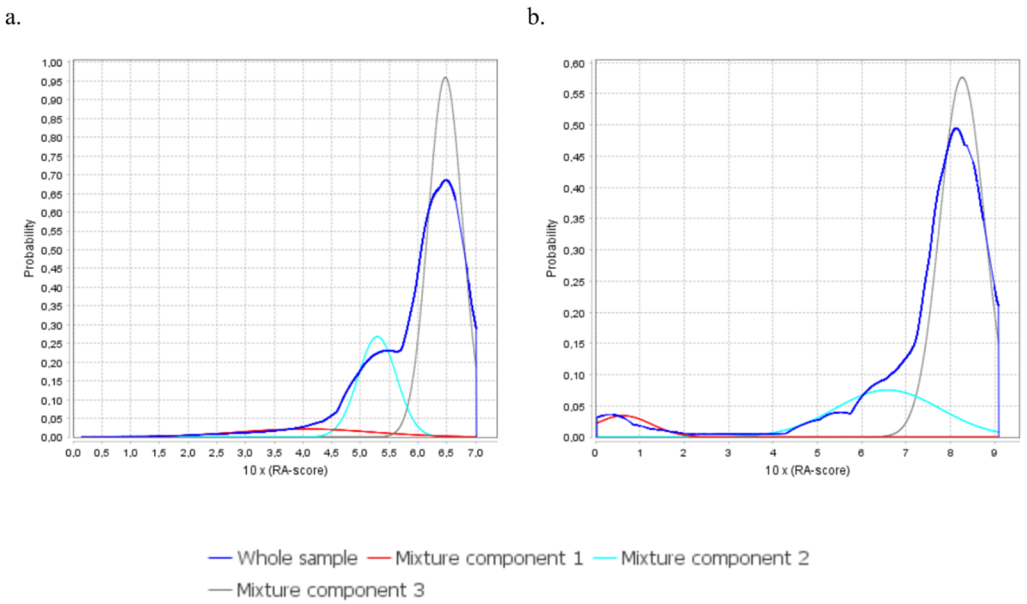


**Fig. 2.** Normal mixture of RA-scores for (a) JUND and (b) REST

*Acknowledgements*: The study was supported by the Ministry of Science and Higher Education of the Russian Federation, grant No. 075-15-2021-1344.