# 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, 354340, Russian Federation
2. BIOSOFT.RU, LLC, Novosibirsk, 630090, Russian Federation
3. Federal Research Center for Information and Computational Technologies, Novosibirsk, 630090, Russian Federation
4. Novosibirsk State University, 630090, Russian Federation

*kolmykovsk@gmail.com

The Ministry of Science and Higher Education of the Russian Federation,
grant № 075-15-2021-1344

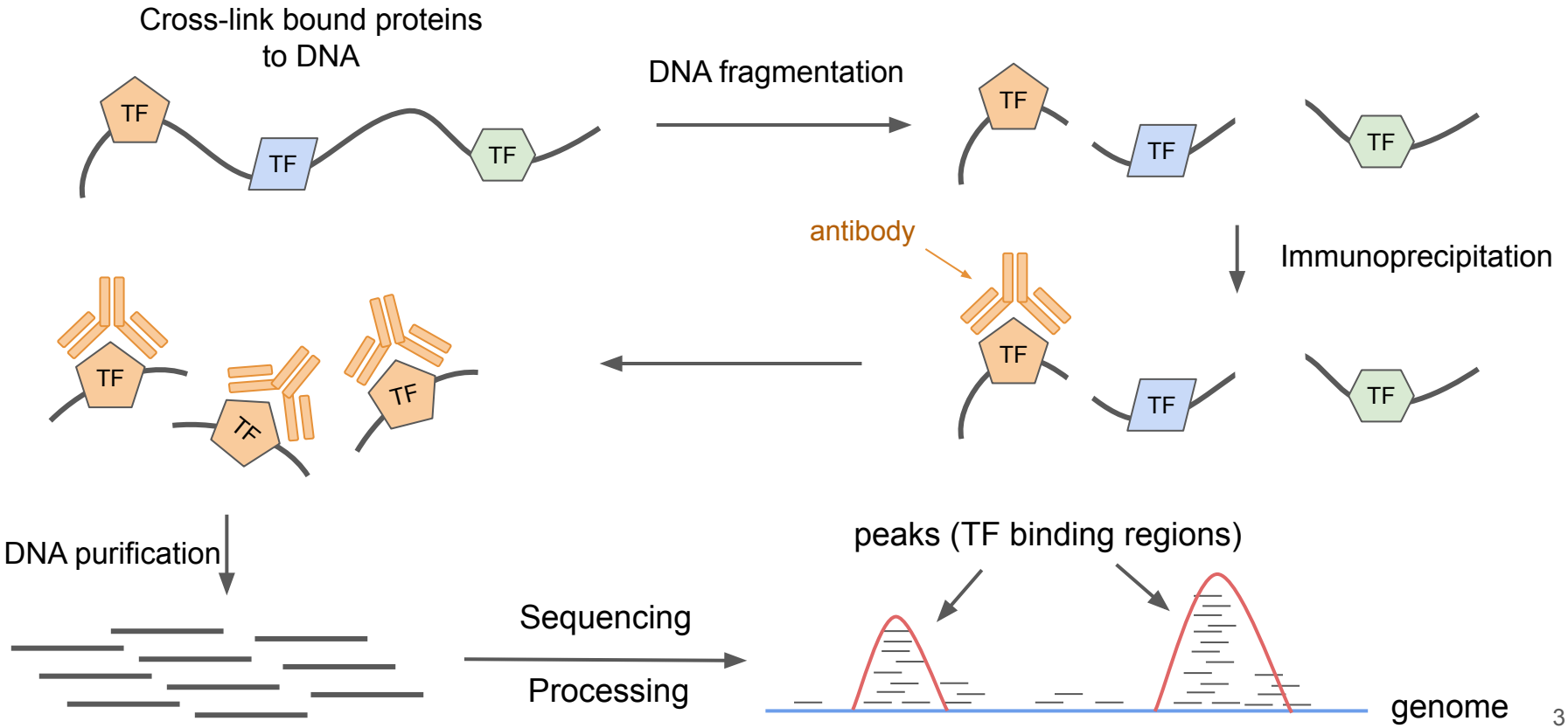
04-8 July 2022, Novosibirsk, Russia

## *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/).

# *Motivation and Aim: ChIP-seq*



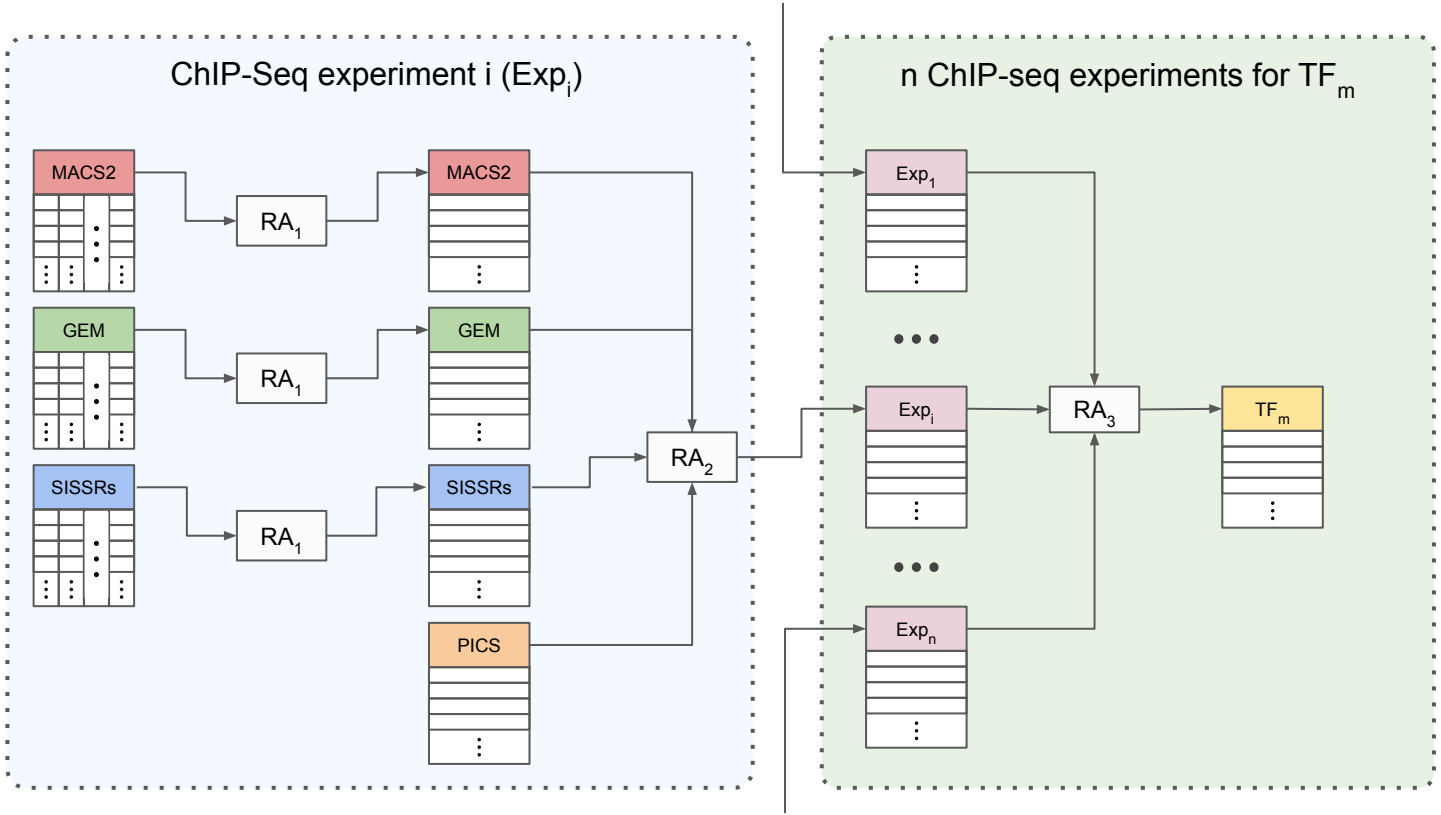
# *Motivation and Aim: Algorithm Development*

The development of a method for generalization of TF binding sites (meta-clusters) based on a large set of ChIP-seq experiments allows us to solve a number of problems:

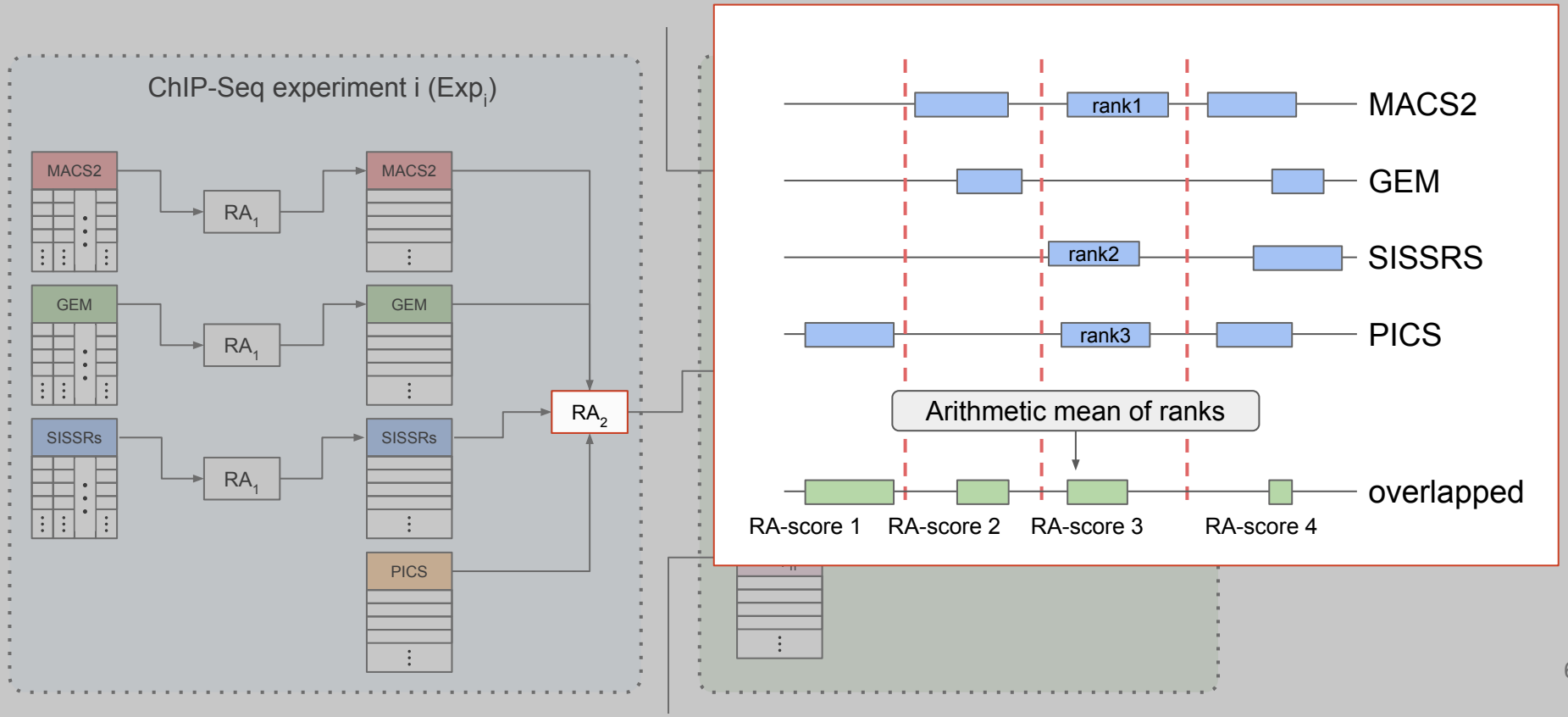
- Integration of large set of ChIP-seq experiments for a given TF;
- More convenient visualization and comparison of TF binding sites with other types of NGS data;
- Analysis of variability and tissue specificity of TF binding sites, as well as the binding motifs identification;
- Identification of stable TF binding sites and their boundaries.

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.

# *The METARA (METa Analysis of ChIP-seq datasets through the RA approach) method description*



# *The METARA (METa Analysis of ChIP-seq datasets through the RA approach) method description*



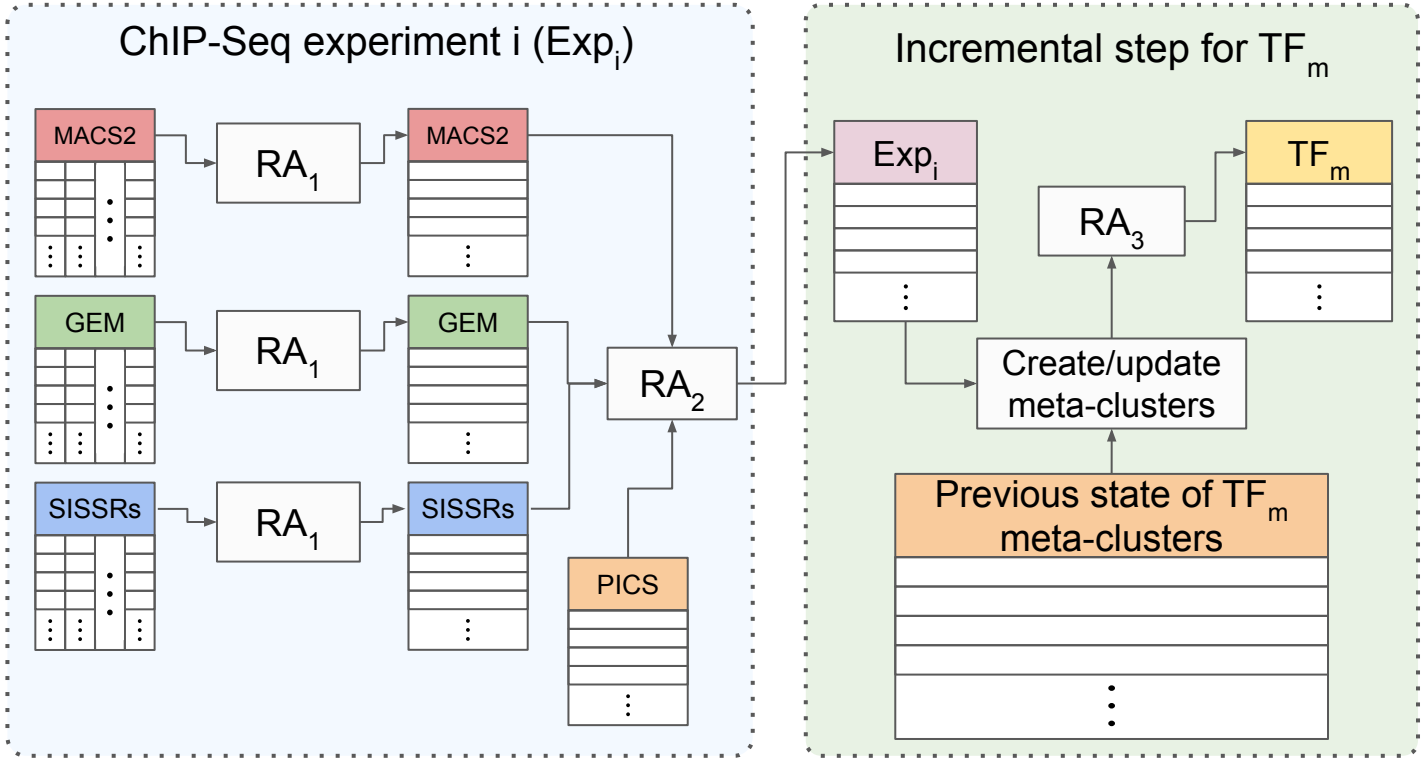
## *Motivation and Aim: Algorithm Development*

We have developed a new method **IMETARA** (Incremental METARA). There are two main reasons for creating IMETARA:

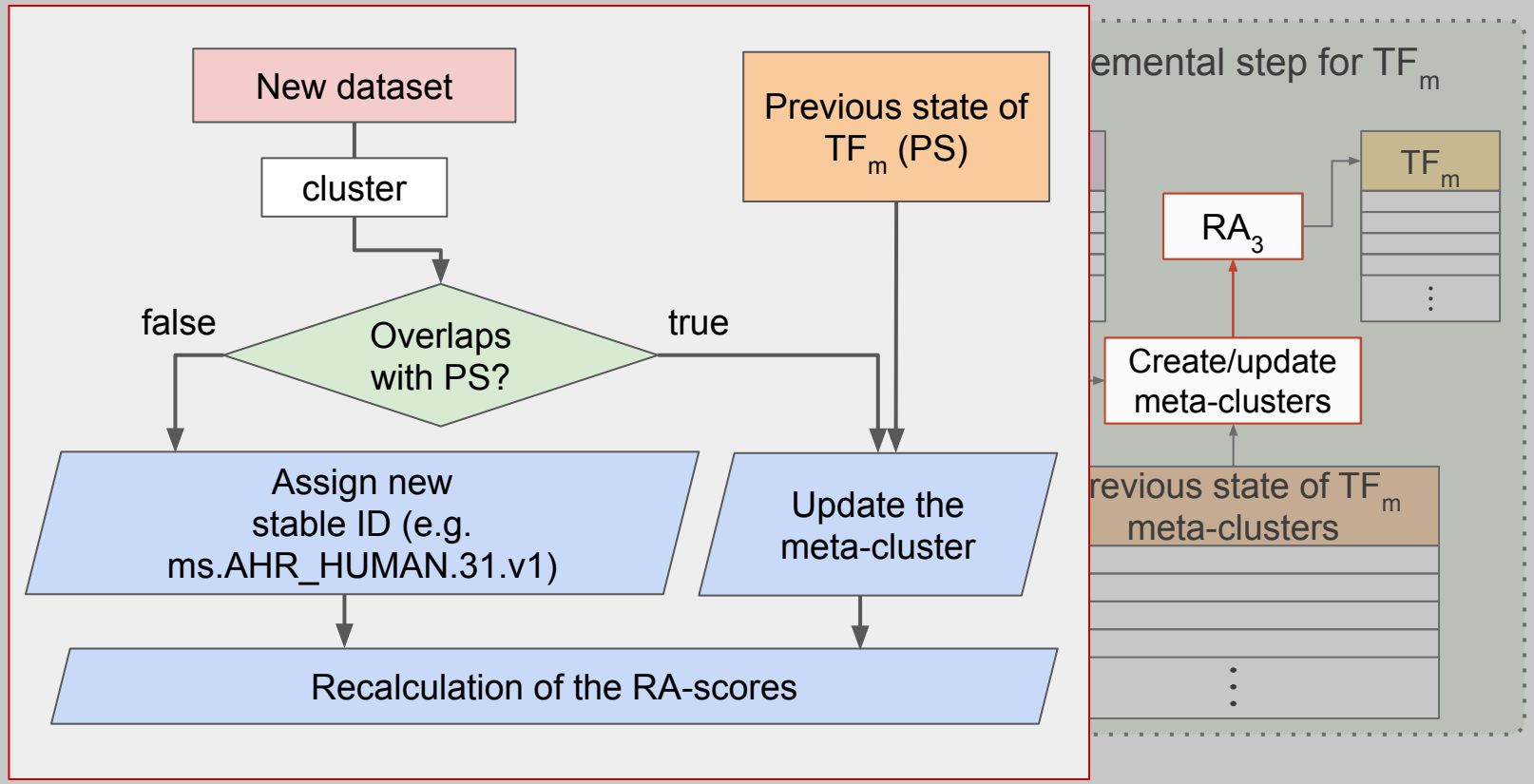
- to significantly reduce the run time of meta-cluster building;
- to introduce stable identifiers for each meta-cluster.

Thus, IMETARA only recalculates the RA scores for existing meta-clusters and adds novel meta-clusters with stable IDs, when new ChIP-seq experiments are added to the ChIP-seq data collection already stored in the GTRD database.

# *The IMETARA (Incremental META Analysis of ChIP-seq datasets through the RA approach) method description*



# IMETARA: Stable Identificators



# *Stable Identifiers*

In the developed algorithm, each meta-cluster (master site) is assigned a stable identifier, which will allow researchers to refer directly to the meta-clusters of TF binding sites, as they currently refer to SNP.

For example, the identifier of one of the TF binding sites has a specific pattern:

**ms.AHR_HUMAN.31.v1** - it consists of the **ms** prefix (GTRD master site), the name of the transcription factor (Uniprot DB protein name), the stable serial number of this site within this TF (31) and the version of this master site (v1).

As new data arrives in GTRD, the genomic boundaries of this site can be revised, in which case only the version will be changed, and the ID will be changed to ms.AHR_HUMAN.31.v2.

# Results

There is a strong IMETARA continuity, the 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.

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%.

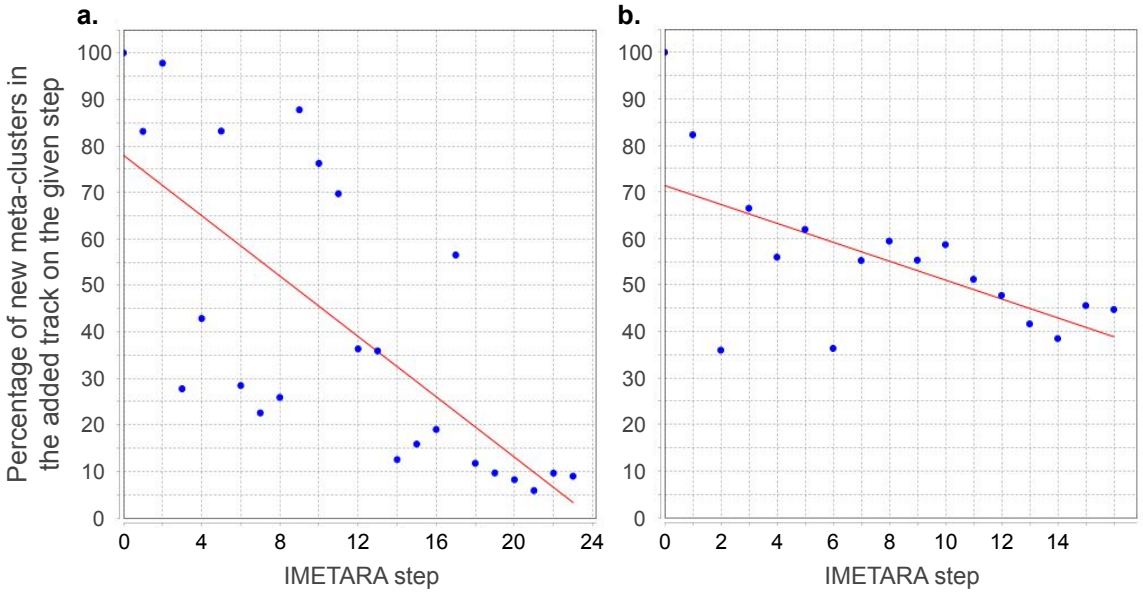



**Figure 1.** Percentage of new meta-clusters in the added track on the given step for (a) ESR1 and (b) FOXA1.

# Results

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 (see Figure 2) demonstrates that meta-clusters can be divided into 2-3 reliability groups.

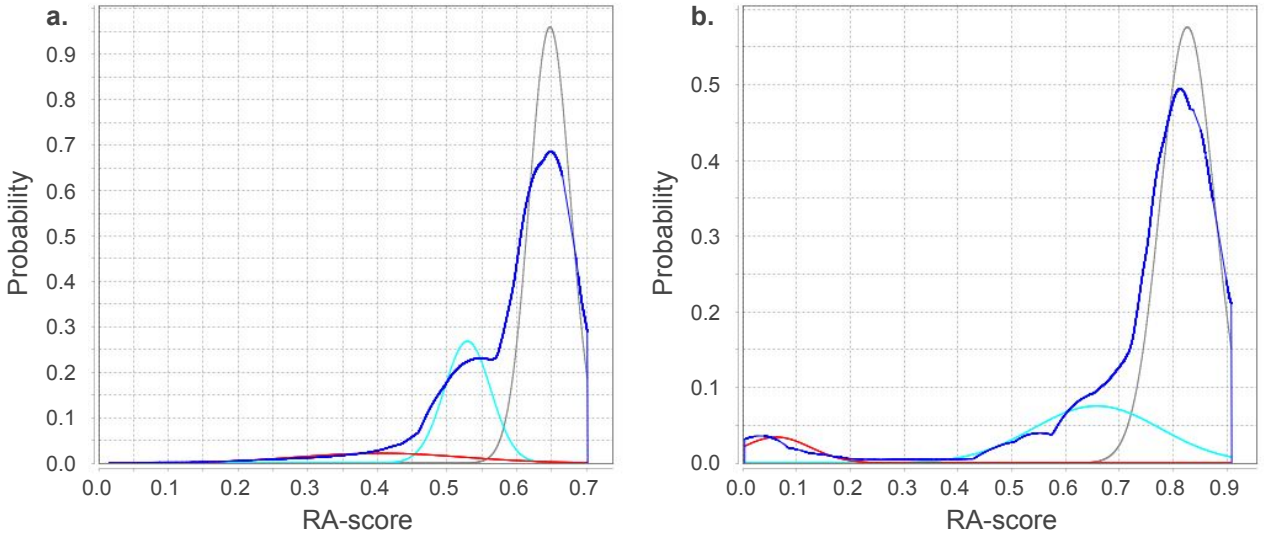



**Figure 2.** Normal mixture of RA-scores for (a) JUND and (b) REST.