RSAT matrix-clustering: dynamic exploration and redundancy reduction of transcription factor binding motif collections

Jaime Abraham Castro-Mondragon[1], Sébastien Jaeger[2], Denis Thieffry[3], Morgane Thomas-Chollier[3*] and Jacques van Helden[1*]

[1] Aix-Marseille Univ, Inserm, TAGC, Technological Advances for Genomics and Clinics, UMR_S 1090, Marseille, France.

[2] Aix Marseille Univ, CNRS, INSERM, CIML, Marseille, France

[3] Computational Systems Biology, Institut de Biologie de l'Ecole Normale Supérieure (IBENS), CNRS, Inserm, Ecole Normale Supérieure, PSL Research University, F-75005 Paris, France.

* To whom correspondence should be addressed:

Jacques van Helden
Tel: +33 4 91 82 87 49
Email: Jacques.van-Helden@univ-amu.fr

Morgane Thomas-Chollier
Tel: +33 1 44 32 23 53
Email: mthomas@biologie.ens.fr

ABSTRACT

Transcription Factor Binding Motifs (TFBMs) databases contain many similar motifs, from which non-redundant collections are derived by manual curation. However, the numbers of motifs and collections are exploding. Meta-databases merging these collections do not offer non-redundant versions, because automatically regrouping similar motifs into clusters cannot be easily achieved with available tools. Motif discovery from genome-scale data sets (e.g. ChIP-seq peaks) also produces redundant motifs, hampering the interpretation of results. We present *matrix-clustering*, a versatile tool that clusters similar TFBMs into multiple trees, and automatically creates non-redundant collections of motifs. A feature unique to *matrix-clustering* is its dynamic visualisation of aligned TFBMs, facilitating and accelerating the analysis of motif collections. It can simultaneously cluster multiple collections from various sources. We demonstrate how *matrix-clustering* considerably simplifies the interpretation of combined results from multiple motif discovery tools. It facilitates the comparison of ChIP-seq datasets, and highlights biologically relevant variations of similar motifs. By clustering 12 entire databases (>5000 motifs), we show that *matrix-clustering* correctly groups motifs belonging to the same TF families, and can drastically reduce motif redundancy. It is integrated within the RSAT suite (http://rsat.eu/), accessible through a user-friendly web interface or command-line for its integration in pipelines.

INTRODUCTION

Transcription Factor Binding Motifs (TFBM) are commonly represented as position-specific scoring matrices (PSSMs) (1) and visualized as sequence logos. Although the adequacy of PSSMs has been questioned for some particular Transcription Factor (TF) classes (2–5), e.g. in cases of dependencies between adjacent nucleotides, they are still considered as the most suitable method to represent the binding specificity of a TF. Thousands of PSSMs are available in private or public TFBM databases, such as JASPAR (6), RegulonDB (7), TRANSFAC (8), CisBP (9), FootprintDB (10), Hocomoco (11), which constitute key resources to interpret functional genomics results. A well-known issue with these databases is motif redundancy (10, 12), caused by various reasons: (i) for a given TF, multiple PSSMs can be built from different collections of sites characterized with alternative methods (i.e. DNAse footprinting, SELEX, Protein-Binding arrays, ChIP-seq, etc); (ii) the binding specificity is often conserved between TFs of the same family; (iii) some databases contain PSSMs obtained from orthologous TFs in different organisms; (iv) some unrelated TFs recognize similar DNA motifs; (v) the annotations may contain some errors.

In addition to this intra-database redundancy, inter-database redundancy and the exponential growth of motif collections are becoming a major issue. Indeed, the development of high-throughput methods to characterize genome-wise TF binding locations (e.g. ChIP-seq, DNAseI, ATAC-seq) has led to an explosion of motifs, with a fast expansion of databases (e.g. JASPAR 2016 almost doubled in size since its 2014 version, from 590 to 1092 motifs) (6, 13). In parallel, recent studies targeting many TFs (3) (14) (15) resulted in collections with as many motifs as certain databases. This constant increase in the number of motifs and redundant collections represents a real challenge for the community. Which collection to use? How important is the overlap between the different collections? Efforts to collect and integrate numerous up-to-date collections into a single metadatabase like FootprintDB (10) are critical for the community. This metadatabase however does not deal yet with the redundancy issue, and keeps increasing in size (9.037 PSSMs as of July 2016). This now constitutes a major bottleneck, by drastically increasing the time needed to compare motifs to, or to scan sequences with a whole motif database.

Individual analysis of high-throughput datasets such ChIP-seq also results in sets of redundant motifs. It is common practice to simultaneously use multiple *de novo* motif discovery tools (16–21) in order to benefit from their complementarity. While some motifs will be discovered exclusively by a given tool, most will be found independently by different tools, hence producing redundant motifs with small variations in length and/or nucleotide frequencies at some positions. Such variations may be important biologically, but remain undetected when inspecting unordered collections of motif logos.

Motif redundancy can be automatically reduced by identifying sets of similar motifs and clustering them. Quantifying the similarity between motifs is nevertheless far from trivial. Many efforts have been done to develop statistical methods and to find adequate metrics to compare the motifs, each one with its own strengths and drawbacks (22–38). Despite this intensive research activity to refine motif similarity metrics, no general consensus has emerged about the best motif similarity metric. Currently, a handful of tools are available for motif comparison: *STAMP (26), TomTom (22), matlign (29)*, *macro-ape* (30)*, DMINDA (39), DbcorrDB (38) and* RSAT *compare-matrices (40, 41).* Other tools are specialized in motif clustering to automatically identify groups of similarity among a set of input motifs: *STAMP (26)*, m*2match (28)*, *MATLIGN* (29), *GMACS (31), DMINDA (39)* and *motIV (38)* (see Table 1 for a comparison of their functionalities). However, these tools present limitations in either restricting the analysis to a single metric, or in the number of input motifs, or in the visualisation interfaces.

We have developed *matrix-clustering* within the RSAT suite (40–43)*,* motivated by the crucial need for a tool to cluster similar motifs, align them to facilitate visual comparison, explore each cluster in a dynamic way, and reduce redundancy either automatically or in a supervised yet user-friendly way. We first show with two study cases that *matrix-clustering* simplifies the interpretation of motif discovery results, and that a dynamic view of aligned logos can reveal biologically relevant motif variants. Two more study cases using complete databases demonstrate that the program identifies groups of motifs belonging to the same TF families, and can be used to explore the complementarity between multiple motif collections. This paves the way towards creating systematic non-redundant motif collections.

MATERIAL AND METHODS

matrix-clustering overview

matrix-clustering clusters similar motifs using hierarchical clustering, followed by a partitioning step that generates individual clusters (Figure 1). The partitioning of the tree into a set of clusters relies on a combination of thresholds on one or several similarity metrics. Within a cluster, PSSMs are aligned to facilitate visual comparison. The program accepts as input different file formats for the PSSMs, organized in collections of motifs to trace the provenance of the motifs. The results are displayed on a dynamic user-friendly web report enabling to collapse or expand subtrees at will.

Input formats

matrix-clustering receives as input one, several or many collections of PSSMs. Collections of motifs are provided as separate files with an associated "collection name" (e.g. several motif collections obtained with different motif discovery tools, in order to identify which motifs are discovered by each tool). This program supports different file formats: TRANSFAC (default), MEME, HOMER, JASPAR, etc; and has no restriction on the number of input PSSMs, but users should be aware that the processing time increases quadratically with the number of motifs.

Motif comparison

Similarity between each pair of input motifs is computed with the tool RSAT *compare-matrices (40, 41)*, which calculates multiple (dis)similarity statistics in the same analysis: Pearson correlation (cor), Sum Of Squared Distances (SSD), Mutual Information, Logo Dot Product, Euclidian Distances (dEucl), Sandelin-Wasserman Similarity (SW), as well as width-normalized version of these metrics (Supplementary Notes) Each possible comparison is done for each pair of matrices in both orientations, and the program returns the best matching alignment for each matrix pair. The corresponding alignment score, relative orientation and offset (shift between the two compared PSSMs) are used for the subsequent clustering steps.

Hierarchical clustering

To build the *global hierarchical tree* encompassing all input motifs, the user must select one of the motif (dis)similarity metrics and one linkage methods (average, complete, median or single). Some metrics directly measure distances (Euclidean, SSD, SW); for the metrics measuring similarities (e.g. cor, with a range from -1 to +1) the values are first transformed into distances (Dcor = 2 – r, where r is the correlation coefficient), to create a distance matrix between each motif pair, which serves as input for the hierarchical clustering.

Identification of PSSM clusters by tree partitioning

This global tree is segmented into a motif forest, where each sub-tree represents a motif cluster. This partitioning takes into account one or several user-specified thresholds. As *compare-matrices* returns simultaneously several metrics, any combination of these can be selected to define thresholds for the partitioning, thereby allowing users to obtain groups of motifs with the desired level of stringency. The global tree is traversed in a bottom-up way and all the motifs below each node are evaluated with the multiple thresholds (e.g. Ncor >= 0.5 and cor >= 0.7 and alignment width >= 5 columns). For each metric selected as threshold, intermediate node values are computed from all descendent nodes according to the user-selected agglomeration rule (single, average, median or complete). Whenever an intermediate node fails to satisfy any of the threshold values, a new cluster is created by separating its two children branches. This means that, if the tree is built with the average agglomeration rule, the motifs within each cluster have a mean distance at least as low as the thresholds. It must be noticed that the tree topology can change according to (i) the agglomeration rule and (ii) the metric selected to create the hierarchical tree (Supplementary Notes).

Progressive alignment of the PSSMs

Once the *global tree* is partitioned into a motif forest, each subtree is used as guide to progressively align the PSSMs. First, the motifs are orientated (direct or reverse) and then shifted adding empty columns at their ends. Note that this algorithm does not add internal gaps, in contrast to other algorithms which support them. The result of this process is one multiple alignment for each internal node of each tree, ending with a root alignment including all the motifs of a cluster tree.

Branch-wise PSSMs, logos and consensus sequences

Once the motifs have been aligned, *matrix-clustering* calculates for each node of the tree a *branch-wise motif* by summing or averaging the frequencies of the descendent aligned motifs. These branch-wise motifs are then used to generate their corresponding consensus sequences and *logos*. Branch-wise motifs introduced here are a generalization of the familial binding profiles (FBP) (26, 34)

Dynamic visualisation of the clusters

The clusters are displayed as a motif forest, i.e. a collection of trees with a logo displayed at each leave. A unique feature of *matrix-clustering* is that the motif tree can be browsed dynamically: each branch can be collapsed by clicking, and the resulting sub-tree is replaced by the logo of the branch motif. A second click on the same node expands it again (Figure 1).

Cross-coverage of motif collections

When two or more motif collections are given as input, the cross-coverage indicates the percentage of motifs from each collection found in clusters also containing motifs from another collection. The coverage of a motif collection *A* by a motif collection *B* ($c_{A,B}$) is calculated as the the number of motifs from *A* co-clustered with motifs from *B ($|A_{with B}|$)*, divided by the the total number of motifs in *A* ($|A|$).

$$c_{A,B} = \frac{|A_{with B}|}{|A|}$$

Reciprocally, the coverage of collection *B* by collection *A* is computed as follows.

$$c_{B,A} = \frac{|B_{with A}|}{|B|}$$

It must be noted that the coverage is not symmetrical between two collections. This asymmetrical comparison provides a more realistic interpretation of the importance of the intersection relative to the respective sizes of collections of different sizes (e.g. a comparison between a very large database and a small motif set). The cross-coverage is displayed as a heatmap, and a Venn diagram is drawn for each pair of collections. The percentage of motifs exclusive to each collection is also provided.

Implementation

matrix-clustering is implemented in Perl and R. The Logo trees are implemented in HTML5 with the D3 JavaScript library for manipulating documents based on data (http://d3js.org/). The website dynamic elements are implemented using the JavaScript libraries Jquery (http://jquery.com/) and DataTables (http://www.datatables.net/).

Motif datasets of the study cases

To illustrate the clustering of redundant motifs we used 359 motifs discovered with the RSAT tool *peak-motifs (16, 17)* in 12 TF ChIP-seq peak-sets obtained from Chen *et al (44).* For the full database clustering, we analysed 22 taxon-specific collections from 18 motif databases: vertebrates (JASPAR (6), Hocomoco mouse and human (11), CisBP (9), Jolma 2013 “HumanTF” (3), Jolma 2015 “HumanTF_dimers” (14), Uniprobe (45), Fantom5 'novel' motifs (46), hPDI (47) and epigram (48)), plants (JASPAR, Athamap (49), CisBp, ArabidopsisPBM (50) and Cistrome (15)) and insects (OntheFly (51), JASPAR, dmmpmm, idmmpmm (52), CisBP, FlyFactorSurvey (53), DrosphilaTF (54)).

Programs used in study cases

For all the study cases, hierarchical clustering was based on average linkage agglomeration, and the distance matrix was derived from normalized correlation (Ncor) between all PSSM pairs. For study case 1, we used *MEME-ChIP (20, 55)* with the following parameters: (-order 3 -meme-mod zoops -meme-minw 6 -meme-maxw 20 -meme-nmotifs 3 -meme-minsites 4 -dreme-e 0.05 -dreme-m 10 -centrimo-score 7.0 -centrimo-ethresh 15.0), *Homer (56)* with these parameters: (-len 10,13,15 -strand both -mis 3 -S 15) and *peak-motifs* with these parameters (-top_peaks 2000 -max_seq_len 800 -min_markov -2 -max_markov -2 -disco oligos,positions,local_words -nmotifs 4 -minol 6 -maxol 8 -no_merge_lengths -2str -origin center). For the motif comparison in Figure 2C, we used *RSAT compare-matrices* with default parameters. The thresholds used in *matrix-clustering* were: (-lth Ncor 0.45 -lth cor 0.65 -lth w 5).

For study case 2, re-ran *peak-motifs* analyses with the same parameters as in (16), as there has been some enhancements to the program since its publication (-top_peaks 0 -task purge,seqlen,composition,disco,merge_motifs,split_motifs,motifs_vs_motifs,timelog,archive,synthesis, small_summary -disco oligos,positions -2str -noov -nmotifs 5 -origin center -minol 6 -maxol 8 -min_markov -2 -max_markov -2 -max_seq_len 800 -scan_markov 1) + matrix-clustering (-lth Ncor 0.45 -lth cor 0.65 -lth w 5).

For study case 3, the thresholds used in *matrix-clustering* were: (-lth Ncor 0.55 -lth cor 0.75) to cluster the Jaspar vertebrates, HumanTF_dimers, Hocomoco human and mouse motifs. For Jaspar insects, we used the following thresholds: (-lth Ncor 0.45 -lth cor 0.65), based on empirical observation of the resulting heatmaps.

For study case 4, the thresholds used in *matrix-clustering* were: (-lth Ncor 0.65 -lth cor 0.8).

Availability

The tool *matrix-clustering* is freely available on the RSAT Web servers (http://www.rsat.eu/) (41). It can also be downloaded with the stand-alone RSAT distribution to be used on the Unix shell, allowing to include it in automated pipelines.

RESULTS

We have developed *matrix-clustering* to deal with the increasing number of motifs and reduce the inherent redundancy within collections. It takes as input one or more collections of motifs (PSSMs), measures the similarity between them using several motif comparison metrics, builds a motif similarity tree by hierarchical clustering, cuts this tree into a motif forest (one tree per cluster), computes branch-wise motifs at each branching point, and generates different graphical representations, including a dynamic visualization enabling fast manual curation (Figure 1).

As there is no general agreement about the best metric to measure PSSM similarity, we have systematically tested a variety of metrics to group the motifs (Supplementary Notes), and chosen as default the Normalized Pearson Correlation (Ncor), a corrected version of the Pearson Correlation (cor) where the normalization factor is the number of overlapping columns between two aligned matrices divided by the total columns of the alignment (28, 40). In the study cases below, we therefore used the Ncor metric to cluster the motifs, and then a partitioning rule based on a combination of thresholds on alignment width, correlation and normalized correlation (see Material & Methods and Discussion). The web reports for each study case are available on the supplementary website:

http://teaching.rsat.fr/data/published_data/Castro_2016_matrix-clustering/

Case study 1: identification of TF binding motifs variants within motifs discovered with multiple tools in ChIP-seq datasets

The first study case aims at comparing motifs detected in ChIP-seq peaks with various *ab initio* motif discovery tools: RSAT *peak-motifs (16, 17)*, Homer (56) and MEME-ChIP (20, 55). These motif discovery programs rely on different detection methods (over-representation, positional bias, enrichment, expectation maximization). Each tool produced a set of motifs, which were provided as a separate collection to matrix-clustering with appropriate parameters, to highlight the similarities and differences between the motifs found by the different tools.

To obtain the input motif collections, we re-analysed the ChIP-seq peaks for the TF Oct4 (also named Pou5f1) in mouse embryonic stem cells (ESC) from Chen et al (44). Oct4 binding consensus sequence is *5'-ATGCAAAT-3'*. In ESC, Oct4 often interacts with another TF, Sox2, that binds to the motif *5'-CATTGTA-3'*. The two TFs form an heterodimer that bind a composite motif called SOCT (5'-*CATTGTATGCAAAT-3'*) and co-regulates specific genes (57). Moreover, Oct4 can form homodimers that regulate different target genes (58).

In total, we obtained 66 motifs in three collections (22 motifs discovered by RSAT peak-motifs, 25 by MEME-ChIP and 19 by Homer), which *matrix-clustering* partitioned into 13 distinct clusters (Supplementary Figure 1, Supplementary website). Cluster_1 regroups the 37 motifs corresponding to Sox, Oct and other Oct-like motifs (Figure 2A). Since the name of the source collection (RSAT, MEME-ChIP, HOMER) is displayed besides each logo, we observe that very similar motifs have indeed been discovered by multiple tools, and that a given tool also returns several variants of a motif.

We used the dynamic visualisation to guide us in reducing the redundancy, by collapsing very similar motifs until finding the non-redundant motifs (Figure 2B). The logo alignments help pinpoint the local and global similarities of the clustered motifs. We obtained 6 non-redundant motifs and annotated them by searching for similar motifs in FootprintDB (vertebrate) (10). The tree includes 3 branches corresponding respectively to the canonical Oct4 (blue box), Sox2 (orange) and the composite SOCT motif (red) (Figure 2C). Interestingly, the remaining branches of this cluster highlight three motifs variants documented in the literature, but not stored in databases: an alternative configuration of Oct4 (black) (59), a palindromic motif bound by an Oct-Oct dimer known as MORE (More Palindromic-Oct-factor-Recognition-Element) (purple) motif (58), and an octamer-repeat motif known as Ocr (60). Of note, these last two motifs were only found by *peak-motifs* (Figure 2B).

The contributions of the respective motif discovery tool to the clusters are unbalanced (Figure 2D). While *peak-motifs* contributes to three clusters shared with MEME and HOMER, MEME-CHIP raised one single-motif cluster (singleton) and HOMER six (Figure 2D, Supplementary Figure 1). The cross-coverage between the tools (Figure 2E) confirms that altogether, *peak-motifs* and MEME show a pretty high overlap, whereas the HOMER motifs are quite dissimilar from those obtained with the other tools. Of note, many motifs found by HOMER only are actually of low-complexity (2-residue repeats) and are not likely to correspond to *bona fide* TFBMs.

Altogether, this study case highlights that *matrix-clustering* can guide and accelerate human-based reduction of a highly redundant collection of motifs, arising from separate motif discovery tools. The clustering moreover highlights the existence of TFBM variants and combinations (e.g. homodimer, heterodimer).

Case study 2: identification of exclusive or shared motifs between various ChIP-seq experiments

To demonstrate how motifs from different experiments can be clustered altogether to identify specific motifs (found exclusively in one dataset), or shared motifs (found in several or all datasets), we extended our previous analysis of Oct4 to the 12 TFs studied by Chen at al (44). We applied an *ab initio* motif discovery (*peak-motifs*) in each peak set separately, and obtained 359 PSSMs, from which *matrix-clustering* identified 40 non-redundant motif clusters (Supplementary Figure 2, supporting website). Some clusters contain set-specific motifs, e.g. motifs discovered exclusively in peak sets from Stat3 (cluster_14), Oct4 (cluster_16), Nanog (cluster_18), Ctcf (cluster_21) and Zfx (cluster_25), while other clusters are composed of motifs found in two or more peak sets (Figure 3A), such as the Sox motif (cluster_11) and Oct motif (cluster_1) respectively found in three (Oct4, Sox2, and Nanog) and five (Oct4, Sox2, Nanog, Stat3, Tcfcp2l1) TF peak sets. These TFs are known to cooperatively regulate common target genes, explaining why their binding motifs are found across multiple collections (16, 44). The cross-coverage heatmap (Figure 3B) provides a global view of the contribution of each collection (12 TF peak sets) to the clusters and the overall overlap of each collection. This representation confirms that Oct, Sox and Nanog collections contain highly similar motifs that cluster together. This is also the case for the c-Myc and n-Myc motifs collections, as well as for E2f1 and Zfx, which are functionally related as histone genes regulators (61). By contrast, the motifs discovered in CTCF peak sets are mostly specific to this collection, and the few other motifs shared with other peak sets are mostly low-complexity motifs (e.g. cluster_19, cluster_20, cluster_24), likely corresponding to artefactual motifs found in several TF peak sets or locally under-represented AT-rich motifs (62). Last, motifs discovered in Klf4 peaks were found in the twelve sets, consistent with its role as pioneer factor in pluripotency maintenance (63).

This second study case shows how *matrix-clustering* can be used to identify motifs specific to one collection (e.g. in a single TF peak set) or shared among several of them. By summarizing multiple motifs sets into a reasonable number of non-redundant motifs (in this case, reduction from 359 to 40 motifs), interpretation becomes less complex.

Case study 3: full-database analysis of relationships between motif clusters and TF families

TFs are classified in families according to their DNA-binding domain (DBD) (64, 65), which usually recognize similar binding sites. The binding specificity of these TFs is thus represented by similar TFBMs, which constitute a source of intra-database redundancy. We separately explored the redundancy within complete databases: taxon-specific motifs from JASPAR (vertebrates and insects, resp.), and species-specific motifs from Hocomoco (human and mouse, resp.). The clustering of JASPAR insects (133 motifs in total) reveals a large cluster of motifs (Figure 4A; Supplementary website) encompassing almost half of the database. This large cluster (64 motifs manually reduced to 9 visually distinct motifs, Figure 4B) corresponds to homeodomain-containing TFs, whose binding motif is characterized by the core consensus 5'-TAAT-3' (66). The numerous members of this TF family in the database reflects a bias in the Insect database, as most of these motifs result from a single analysis covering many homeodomain TFs (67).

By contrast in vertebrates, Hocomoco human is divided into 255 small clusters with much less redundancy (Figure 4C, similar results for JASPAR vertebrates and Hocomoco mouse in Supplementary Figures 3A and 3B, supporting website). As Hocomoco includes an annotation of TF families, we analysed the correspondence between motif clusters produced by *matrix-clustering* and TF families. The large majority of the clusters (204 out of 255) indeed regroup motifs bound by TFs of a single family (Figure 4D). Besides, most of the other clusters actually regroup TFs belonging to different families of the same class. The remaining clusters encompass TFs from different classes but nevertheless bound to similar motifs, and thus correctly grouped by *matrix-clustering*.

Reciprocally, for each TF family we counted the number of covered clusters (Supplementary Figure 4). Among the 78 families from Hocomoco, 29 are consistently packed in a single cluster, 10 in two clusters and 16 in three clusters. On the other extreme, some TF families are split into many clusters, e.g. the Zinc finger families (e.g. for the family “Factors with multiple dispersed zinc fingers”, each motif comes in a separate cluster). This dispersion is perfectly consistent with the well-known properties of these TF families: the Zinc finger domain is characterized by a wide variability of binding motifs, determined by the specific amino acids entering in contact with the DNA (68).

Of note, the original 641 human motifs in Hocomoco were automatically reduced to a set of 255 non-redundant motifs, almost one third of the database size (Supplementary website). For this reduction of intra-database redundancy, we used stringent threshold values (Ncor >= 0.55, cor >= 0.75) which produced clusters with motifs of similar size.

This third study case demonstrates how *matrix-clustering* can handle larger collections of motifs and automatically reduce the redundancy of motifs within a database, while correctly regrouping motifs belonging to the same TF Family.

Case study 4: inter-database redundancy: comparison and integration of multiple motif databases

The growing number of motif databases is becoming a major problem, since it results in partly redundant collections, and complicates the choice for users who need to use a database for their projects. To tackle this problem of inter-database redundancy, we performed a more challenging analysis by merging and clustering 18 full databases encompassing 22 motif collections separated by taxa, and evaluated the cross-coverage and specificity of these motif collections (see methods for the complete list of collections).

We first merged the public motif databases for insects (7 databases; 1895 motifs), plants (5 databases; 1590 motifs) and vertebrates (10 databases; 5384 motifs), and then applied *matrix-clustering*. We obtained respectively 354 (19%), 306 (19%) and 1757 (33%) clusters for insects, plants, and vertebrates (supporting website). In this case, two motifs were considered similar if they satisfied stringent thresholds (cor >= 0.8 and Ncor >= 0.65): the threshold on correlation ensures that the clustered motifs are highly similar and the additional threshold on normalized correlation selects the alignments covering most of the motif lengths, in order to separate composite motifs (e.g. bound by a TF dimer) from their elementary components (specific motifs for each interacting protein).

Figure 5 shows the cross-coverage between the different motif collections for each taxon. This representation allows to visualize and quantify the pairwise similarity of the different motif collections. For the insect databases, CisBP, OnTheFly, FlyFactorSurvey and Jaspar are the most similar in content, while DrosophilaTF content is drastically different (Figure 5A), likely because these motifs, discovered exclusively on Drosophila promoters are unknown (54).

Consistently for the plant databases, Jaspar and CisBP are most similar to each other (Figure 5B). The two other plant motif databases both have around 50% coverage with CisBP and Jaspar, but are very different from each other (17% coverage). By contrast the Cistrome database (15) covers Jaspar and CisBP but is only partly covered by other databases. This is consistent with Cistrome containing novel motifs obtained by DAP-seq, a new experimental *in vitro* method.

For the vertebrate motifs (Figure 5C), five databases have a similar content (Hocomoco human and mouse, Jaspar, CisBp, Jolma 2013 “HumanTF”), which is explained, as above, by the integration of Hocomoco and Jaspar in CisBP, as well as the similitude of the original datasets used to build the TFBMs (mostly public ChIP-seq and Selex-seq datasets), e.g. 87% of Jaspar motifs are found in clusters having CisBP (9) motifs, which is coherent with the fact that CisBP is a metadatabase that integrates Jaspar, among other databases (Figure 5D). The Uniprobe collection has a lower coverage with these databases, possibly because it relies solely on universal Protein Binding Microarray (PBM) to build the TFBMs, which indicates a possible bias in the results of this type of data as discussed by Zhao and Stormo (69). This is also the case for the hPDI (human Protein-DNA interactome) motifs, which are built from a restricted number of sites (17 per motif on average), which is very low compared to the hundreds of sites used to built motifs from high-throughput experiments (SELEX, ChIP-seq). The Fantom5 collection of “novel” motifs has a very low coverage with all other databases (<1.2%). This is in agreement with the particularity of this collection, which, by definition, is restricted to the motifs without any matches in reference databases (6). Similarly, the epigram motifs are not covered well by (and do not cover well) the other collections, since these motifs, constructed from 9-mers over-represented across several histone marks and cell types, do not match known motifs (15). Last, the heatmap shows that humanTF_dimers (14) differs considerably from the other databases, reflecting the distinct grouping of dimers and monomers by *matrix-clustering*. The content of this collection slightly covered by the other collections is explained by the few composite motifs corresponding to TF dimers (e.g. SOCT motifs) present in other motif databases.

In summary, this fourth study case highlights how *matrix-clustering* can be used to automatically reduce motif redundancy across multiple databases, even if the overall motif number is very large (several thousands of motifs). The concise representation provided by the cross-coverage heat map enables to intuitively grasp the overlap between each pair of individual collections.

Comparison with other motif clustering tools

Table 1 provides a list of features supported by existing motif clustering tools. For the sake of comparison, we submitted some of our study cases to alternative motif clustering tools (using default parameters): STAMP (26), m2match (28), Matlign (29) and Gmacs (31). A detailed report of the results is available in the Supplementary notes and the supporting website.

This analysis was restricted to study cases 1 and 3, since no other tool currently supports multiple collections as those of our study cases 2 and 4. In summary, none of the tested tools presents functionalities equivalent to *matrix-clustering*. In particular, *matrix-clustering* is the only tool enabling a dynamic browsing of motif trees with custom collapsing/expansion of branches, and the comparison of multiple motif collections. It provides multiple ways to inspect the results: logos forest; motif correlation heat map; searchable table of motifs and clusters; contributions of the respective collections to each cluster; cross-coverage heat map between collections. Of course, this flexibility has a certain cost in computing time (see Supplementary notes for a comparison of time efficiency between STAMP and *matrix-clustering*).

DISCUSSION

With the advent of large-scale experimental approaches to uncover TF binding specificity such as ChIP-seq, Selex-seq, and Protein Binding microarrays, the number of TFBMs has recently exploded, and motif redundancy has become a critical bottleneck for sequence analyses. Although other studies propose new metrics and software tools to measure motif similarity in the perspective of matching *de novo* motifs with reference motifs databases, only a few tools are truly specialized in motif clustering. We have performed a comprehensive survey of motif clustering tools and compared their functionalities (Table 1), which revealed many limitations that prompted us to develop *matrix-clustering.*

The key feature that distinguishes *matrix-clustering* from the other tools is its dynamic interface to browse hierarchies of clustered TFBMs. This feature substantially facilitates the visual exploration and reduces the time for human-driven analysis of the motif dataset. Notably, this visualization has enabled us to identify the Ocr motif in the Oct4 ChIP-seq dataset (Figure 2). This motif was already present in our previous analysis of this dataset (16), but we had not been able to detect this subtle motif variation among all other motifs, despite our experience in visual comparison of logos. We thus foresee that this dynamic visualisation of motif clusters will be beneficial to both experts and non-experts users, by providing support for human-based annotation of motif variants. Furthermore, the dynamic exploration of motif trees might serve to develop user-friendly interfaces to directly browse motif databases on their own websites. The other advantages of *matrix-clustering* compared to other clustering algorithms are: (i) partitioning of the input motif set into separated trees (forest), rather than forcing a single motif tree, (ii) generation of branch-wise motifs (also known as Familial Binding Profiles, FBPs) at each node of the trees, rather than just at the roots, (iii) specification of thresholds on custom combinations of similarity metrics to integrate multiple criteria for motif partitioning, (iv) support for multiple input motif sets (collections), (v) alternative representations of the clusters (hierarchical trees with logo alignments, searchable motif table, motif similarity heat maps and collection cross-coverage heat maps), (vi) automated production of non-redundant motif collections.

Our method relies on hierarchical clustering and on a bottom-up partitioning combining thresholds on multiple metrics (e.g. width >= 5, Ncor >= 0.5 and cor >= 0.7). The tree is thus segmented based on the similarity between all the descendent motifs of each branch, which strongly differs from the usual cut-off at an arbitrary height of the clustering tree. We also evaluated an alternative segmentation method called dynamic tree cut, which relies on tree topology to produce balanced clusters (70). However, for TFBMs, our multi-threshold heuristics produces more relevant results (not shown). Although hierarchical clustering is known to produce 'frozen' artefacts (i.e. a pair of nodes early grouped in the tree cannot be relocated in later steps) and motifs are not free to move across the tree (31). Note that this is not the case for the alignment-free methods (25, 31, 37), while this is issue is circumvented by iterative refinement in STAMP (26). In matrix-clustering, these limitations are compensated by the flexible partitioning step. Indeed, when such artefacts occur, setting a more stringent threshold or selecting another metric will properly separate the clusters.

The size and composition of the clusters are determined by the chosen agglomeration method in combination with the partitioning threshold (Supplementary Notes). This threshold should be tuned by the user to reach the desired granularity of clusters, as shown in the study cases. In the first study case, we used loose thresholds (Ncor >= 0.4 and cor >= 0.6) to group the motifs of the same TFs and observe motifs variants within a single cluster. To identify TFs from the same TF Family (study case 3), we took intermediate thresholds (Ncor >= 0.55 and cor >= 0.75). To obtain a non-redundant collection of motifs (study case 4), stringent thresholds can be used (Ncor >= 0.65 and cor >= 0.80), which ensures the motifs are highly similar in information content and width.

The running time grows quadratically because hierarchical clustering relies on a matrix indicating the distance between each pair of input motifs. For small-sized motif collections such as motif discovery results, the running time enables *matrix-clustering* usage via the website (e.g. 7 min for the first study case with 66 motifs). Very large datasets (e.g. full databases) have to be treated locally (e.g. 85 min on a single CPU to treat the 643 motifs of Hocomoco human).

Several motif databases like Jaspar and Hocomoco already provide non-redundant collections, obtained by a time-consuming manual curation, which will become complicated with the increasing number of available motifs. An advantage of *matrix-clustering* is to provide in a single command a full workflow for motif comparisons, clustering, partitioning and visualisation, whereas many other studies use distinct tools to run these tasks step by step (15, 24, 46, 48). Moreover, the dynamic visualization of aligned motif logos can significantly reduce the curation time. Of note, in most motif databases, the term *non-redundant* denotes the restriction to one motif per TF (6). However, distinct TFs may also bind very similar motifs (e.g. Oct4, Oct9, and Oct11). In *matrix-clustering* results, the term *non-redundant* refers to a single TFBM summarizing a set of highly similar motifs, independently of the binding TF. Groups of similar TFBMs are thus reduced to a single motif with a generic name (e.g. cluster_14 in study case 2), while all names of the original motifs are retained (e.g. the list of Oct motifs composing cluster_22 derived from the Jaspar vertebrate, supporting website). Still, there is no one-to-one correspondence between motif clusters and TF families (Supplementary Figure 3), as shown with Zinc fingers sharing similar DNA-binding domains, but not recognizing the same DNA sequences (68).

Reducing the size of motif collections is becoming indispensable to limit the processing time of tools relying on full motif databases (e.g. motif enrichment, motif comparisons, identification of regulatory variants). Additionally, the interpretability of the results is also limited by the multiplicity of partially redundant motifs. As a proof-of-concept, we have shown that *matrix-clustering* can be used to compare full collections, but also to drastically reduce the inter-database redundancy: in study case 4 above, motifs were merged from 22 collections of different sources, to build taxon-specific collections.

The selection of non-redundant motifs respectively reduced the insect, plant and vertebrate collections to 19%, 19% and 32% of their original sizes. We thus foresee that meta-databases, such as footprintDB (10), could benefit from *matrix-clustering* to offer a non-redundant motif collection.

The possibility to cluster several collections simultaneously makes *matrix-clustering* a versatile tool, as demonstrated with the chosen study cases (identification of motif variants, integration of motifs found by multiple motif discovery tools, comparison of motifs obtained from 12 datasets). It could also be used to compare motifs obtained in different experimental conditions. Given the compatibility with many PSSMs formats (Transfac, MEME, HOMER) and its Web access, this tool will be of interest to the broad community of biologists and bioinformaticians involved in the analysis of regulatory sequences.

ACKNOWLEDGEMENT

We wish to thank to Bruno Contreras-Moreira, Aïtor Gonzales, Carl Herrmann, Samuel Collombet, Roberto Tirado-Magallanes, Lambert Moyon and Coby Viner for the suggestions to improve the method, the visualization and useful feedback. We are also thankful to the Jaspar team for checking the validity of the clustering of JASPAR motif collections.

FUNDING

This work was supported by the French Agence Nationale pour la Recherche iBone [ANR-13-EPIG-0001-04]; EchiNodal [ANR-14-CE11-0006-02]; CONACyT-Mexico [Fellowship 391575]; a PhD grant from the Ecole Doctorale des Sciences de la Vie et de la Santé, Aix-Marseille Université [to J.A.C.M.]. Funding for open access charge: French Agence Nationale pour la Recherche .

REFERENCES

1. Stormo,G.D. (2000) DNA binding sites: representation and discovery. *Bioinformatics*, **16**, 16–23.

2. Weirauch,M.T., Cote,A., Norel,R., Annala,M., Zhao,Y., Riley,T.R., Saez-Rodriguez,J., Cokelaer,T., Vedenko,A., Talukder,S., *et al.* (2013) Evaluation of methods for modeling transcription factor sequence specificity. *Nat. Biotechnol.*, **31**, 126–34.

3. Jolma,A., Yan,J., Whitington,T., Toivonen,J., Nitta,K.R., Rastas,P., Morgunova,E., Enge,M., Taipale,M., Wei,G., *et al.* (2013) DNA-binding specificities of human transcription factors. *Cell*, **152**, 327–39.

4. Mathelier,A. and Wasserman,W.W. (2013) The next generation of transcription factor binding site prediction. *PLoS Comput. Biol.*, **9**, e1003214.

5. Keilwagen,J. and Grau,J. (2015) Varying levels of complexity in transcription factor binding motifs. *Nucleic Acids Res.*, 10.1093/nar/gkv577.

6. Mathelier,A., Fornes,O., Arenillas,D.J., Chen,C., Denay,G., Lee,J., Shi,W., Shyr,C., Tan,G., Worsley-Hunt,R., *et al.* (2015) JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, 10.1093/nar/gkv1176.

7. Gama-Castro,S., Salgado,H., Santos-Zavaleta,A., Ledezma-Tejeida,D., Muñiz-Rascado,L., García-Sotelo,J.S., Alquicira-Hernández,K., Martínez-Flores,I., Pannier,L., Castro-Mondragón,J.A., *et al.* (2015) RegulonDB version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond. *Nucleic Acids Res.* , 10.1093/nar/gkv1156.

8. Matys,V. (2003) TRANSFAC(R): transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.*, **31**, 374–378.

9. Weirauch,M.T., Yang,A., Albu,M., Cote,A.G., Montenegro-Montero,A., Drewe,P., Najafabadi,H.S., Lambert,S.A., Mann,I., Cook,K., *et al.* (2014) Determination and Inference of Eukaryotic Transcription Factor Sequence Specificity. *Cell*, **158**, 1431–1443.

10. Sebastian,A. and Contreras-Moreira,B. (2014) FootprintDB: A database of transcription factors with annotated cis elements and binding interfaces. *Bioinformatics*, **30**, 258–265.

11. Kulakovskiy,I. V, Vorontsov,I.E., Yevshin,I.S., Soboleva,A. V, Kasianov,A.S., Ashoor,H., Ba-alawi,W., Bajic,V.B., Medvedeva,A., Kolpakov,F.A., *et al.* (2015) HOCOMOCO : expansion and enhancement of the collection of transcription factor binding sites models. 10.1093/nar/gkv1249.

12. D'haeseleer,P. (2006) How does DNA sequence motif discovery work? *Nat. Biotechnol.*, **24**, 959–961.

13. Mathelier,A., Zhao,X., Zhang,A.W., Parcy,F., Worsley-Hunt,R., Arenillas,D.J., Buchman,S., Chen,C., Chou,A., Ienasescu,H., *et al.* (2014) JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, **42**, D142–7.

14. Jolma,A., Yin,Y., Nitta,K.R., Dave,K., Popov,A., Taipale,M., Enge,M., Kivioja,T., Morgunova,E. and Taipale,J. (2015) DNA-dependent formation of transcription factor pairs alters their binding specificity. *Nature*, **527**, 384–8.

15. O'Malley,R.C., Huang,S.C., Song,L., Lewsey,M.G., Bartlett,A., Nery,J.R., Galli,M., Gallavotti,A. and Ecker,J.R. (2016) Cistrome and Epicistrome Features Shape the Regulatory DNA Landscape. *Cell*, **165**, 1280–1292.

16. Thomas-Chollier,M., Herrmann,C., Defrance,M., Sand,O., Thieffry,D. and van Helden,J. (2012) RSAT peak-motifs: motif analysis in full-size ChIP-seq datasets. *Nucleic Acids Res.*, **40**, e31.

17. Thomas-Chollier,M., Darbo,E., Herrmann,C., Defrance,M., Thieffry,D. and van Helden,J. (2012) A complete workflow for the analysis of full-size ChIP-seq (and similar) data sets using peak-motifs. *Nat. Protoc.*, **7**, 1551–68.

18. Kulakovskiy,I. V, Boeva,V. a, Favorov, a V and Makeev,V.J. (2010) Deep and wide digging for binding motifs in ChIP-Seq data. *Bioinformatics*, **26**, 2622–3.

19. Bailey,T.L. (2011) DREME: motif discovery in transcription factor ChIP-seq data. *Bioinformatics*, **27**, 1653–9.

20. Ma,W., Noble,W.S. and Bailey,T.L. (2014) Motif-based analysis of large nucleotide data sets using MEME-ChIP. *Nat. Protoc.*, **9**, 1428–50.

21. Grau,J., Posch,S., Grosse,I. and Keilwagen,J. (2013) A general approach for discriminative de novo motif discovery from high-throughput data. *Nucleic Acids Res.*, **41**, e197.

22. Gupta,S., Stamatoyannopoulos,J. a, Bailey,T.L. and Noble,W.S. (2007) Quantifying similarity between motifs. *Genome Biol.*, **8**, R24.

23. Tanaka,E., Bailey,T., Grant,C.E., Noble,W.S. and Keich,U. (2011) Improved similarity scores for comparing motifs. *Bioinformatics*, **27**, 1603–9.

24. Habib,N., Kaplan,T., Margalit,H. and Friedman,N. (2008) A novel Bayesian DNA motif comparison method for clustering and retrieval. *PLoS Comput. Biol.*, **4**, e1000010.

25. Pape,U.J., Rahmann,S. and Vingron,M. (2008) Natural similarity measures between position frequency matrices with an application to clustering. *Bioinformatics*, **24**, 350–7.

26. Mahony,S. and Benos,P. V (2007) STAMP: a web tool for exploring DNA-binding motif similarities. *Nucleic Acids Res.*, **35**, W253–8.

27. Mahony,S., Auron,P.E. and Benos,P. V (2007) Inferring protein-DNA dependencies using motif alignments and mutual information. *Bioinformatics*, **23**, i297–304.

28. Stegmaier,P., Kel,A., Wingender,E. and Borlak,J. (2013) A discriminative approach for unsupervised clustering of DNA sequence motifs. *PLoS Comput. Biol.*, **9**, e1002958.

29. Kankainen,M. and Löytynoja,A. (2007) MATLIGN: a motif clustering, comparison and matching tool. *BMC Bioinformatics*, **8**, 189.

30. Vorontsov,I.E., Kulakovskiy,I. V and Makeev,V.J. (2013) Jaccard index based similarity measure to compare transcription factor binding site models. *Algorithms Mol. Biol.*, **8**, 23.

31. Broin,P.Ó., Smith,T.J. and Golden,A.A. (2015) Alignment-free clustering of transcription factor binding motifs using a genetic-k-medoids approach. *BMC Bioinformatics*, **16**, 22.

32. Schones,D.E., Sumazin,P. and Zhang,M.Q. (2005) Similarity of position frequency matrices for transcription factor binding sites. *Bioinformatics*, **21**, 307–13.

33. Zhang,S., Zhou,X., Du,C. and Su,Z. (2013) SPIC: a novel similarity metric for comparing transcription factor binding site motifs based on information contents. *BMC Syst. Biol.*, **7 Suppl 2**, S14.

34. Sandelin,A. and Wasserman,W.W. (2004) Constrained Binding Site Diversity within Families of Transcription Factors Enhances Pattern Discovery Bioinformatics. *J. Mol. Biol.*, **338**, 207–215.

35. Kielbasa,S.M., Gonze,D. and Herzel,H. (2005) Measuring similarities between transcription factor binding sites. *BMC Bioinformatics*, **6**, 237.

36. Communication,S. (2006) MACO: A Gapped-Alignment Scoring Tool for Comparing Transcription Factor Binding Sites. **6**, 307–310.

37. Xu,M. and Su,Z. (2010) A novel alignment-free method for comparing transcription factor binding site motifs. *PLoS One*, **5**, e8797.

38. Grau,J., Grosse,I., Posch,S., Keilwagen,J. and Julius,K. (2015) Motif clustering with implications for transcription factor interactions. *PeerJ Prepr.*, 10.7287/peerj.preprints.1302v1.

39. Ma,Q., Zhang,H., Mao,X., Zhou,C., Liu,B., Chen,X. and Xu,Y. (2014) DMINDA: An integrated web server for DNA motif identification and analyses. *Nucleic Acids Res.*, **42**, 12–19.

40. Thomas-Chollier,M., Defrance,M., Medina-Rivera,A., Sand,O., Herrmann,C., Thieffry,D. and van Helden,J. (2011) RSAT 2011: regulatory sequence analysis tools. *Nucleic Acids Res.*, **39**, W86–91.

41. Medina-Rivera,A., Defrance,M., Sand,O., Herrmann,C., Castro-Mondragon,J. a., Delerce,J., Jaeger,S., Blanchet,C., Vincens,P., Caron,C., *et al.* (2015) RSAT 2015: Regulatory Sequence Analysis Tools. *Nucleic Acids Res.*, 10.1093/nar/gkv362.

42. Thomas-Chollier,M., Sand,O., Turatsinze,J.-V., Janky,R., Defrance,M., Vervisch,E., Brohée,S. and van Helden,J. (2008) RSAT: regulatory sequence analysis tools. *Nucleic Acids Res.*, **36**, W119–27.

43. van Helden,J. (2003) Regulatory Sequence Analysis Tools. *Nucleic Acids Res.*, **31**, 3593–3596.

44. Chen,X., Xu,H., Yuan,P., Fang,F., Huss,M., Vega,V.B., Wong,E., Orlov,Y.L., Zhang,W., Jiang,J., *et al.* (2008) Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell*, **133**, 1106–17.

45. Hume,M.A., Barrera,L.A., Gisselbrecht,S.S. and Bulyk,M.L. (2015) UniPROBE, update 2015: new tools and content for the online database of protein-binding microarray data on protein-DNA interactions. *Nucleic Acids Res.*, **43**, D117–22.

46. Forrest,A.R.R., Kawaji,H., Rehli,M., Baillie,J.K., de Hoon,M.J.L., Lassmann,T., Itoh,M., Summers,K.M., Suzuki,H., Daub,C.O., *et al.* (2014) A promoter-level mammalian expression atlas. *Nature*, **507**, 462–70.

47. Xie,Z., Hu,S., Blackshaw,S., Zhu,H. and Qian,J. (2010) hPDI: A database of experimental human protein-DNA interactions. *Bioinformatics*, **26**, 287–289.

48. Whitaker,J.W., Chen,Z. and Wang,W. (2015) Predicting the human epigenome from DNA motifs. *Nat Methods*, **12**, 265–272.

49. Bülow,L., Engelmann,S., Schindler,M. and Hehl,R. (2009) AthaMap, integrating transcriptional and post-transcriptional data. *Nucleic Acids Res.*, **37**, 983–986.

50. Franco-Zorrilla,J.M., López-Vidriero,I., Carrasco,J.L., Godoy,M., Vera,P. and Solano,R. (2014) DNA-binding specificities of plant transcription factors and their potential to define target genes. *Proc. Natl. Acad. Sci. U. S. A.*, **111**, 2367–72.

51. Shazman,S., Lee,H., Socol,Y., Mann,R.S. and Honig,B. (2014) OnTheFly: a database of Drosophila melanogaster transcription factors and their binding sites. *Nucleic Acids Res.*, **42**, D167–71.

52. Kulakovskiy,I. V, Favorov,A. V and Makeev,V.J. (2009) Motif discovery and motif finding from genome-mapped DNase footprint data. *Bioinformatics*, **25**, 128–131.

53. Zhu,L.J., Christensen,R.G., Kazemian,M., Hull,C.J., Enuameh,M.S., Basciotta,M.D., Brasefield,J.A., Zhu,C., Asriyan,Y., Lapointe,D.S., *et al.* (2011) FlyFactorSurvey: A database of Drosophila transcription factor binding specificities determined using the bacterial one-hybrid system. *Nucleic Acids Res.*, **39**, 111–117.

54. Down,T.A., Bergman,C.M., Su,J. and Hubbard,T.J.P. (2007) Large-scale discovery of promoter motifs in Drosophila melanogaster. *PLoS Comput. Biol.*, **3**, 0095–0109.

55. Machanick,P. and Bailey,T.L. (2011) MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics*, **27**, 1696–7.

56. Heinz,S., Benner,C., Spann,N., Bertolino,E., Lin,Y.C., Laslo,P., Cheng,J.X., Murre,C., Singh,H. and Glass,C.K. (2010) Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Mol. Cell*, **38**, 576–589.

57. Mistri,T.K., Devasia,A.G., Chu,L.T., Ng,W.P., Halbritter,F., Colby,D., Martynoga,B., Tomlinson,S.R., Chambers,I., Robson,P., *et al.* (2015) Selective influence of Sox2 on POU transcription factor binding in embryonic and neural stem cells. *EMBO Rep.*, **16**, 1177–91.

58. Tantin,D., Gemberling,M., Callister,C. and Fairbrother,W.G. (2008) High-throughput biochemical analysis of in vivo location data reveals novel distinct classes of POU5F1(Oct4)/DNA complexes. *Genome Res*, **18**, 631–639.

59. Mason,M.J., Plath,K. and Zhou,Q. (2010) Identification of context-dependent motifs by contrasting ChIP binding data. *Bioinformatics*, **26**, 2826–32.

60. Morgan,G.T. and Middleton,K.M. (1990) Short interspersed repeats from Xenopus that contain multiple octamer motifs are related to known transposable elements. *Nucleic Acids Res.*, **18**, 5781–6.

61. Gokhman,D., Livyatan,I., Sailaja,B.S., Melcer,S. and Meshorer,E. (2013) Multilayered chromatin analysis reveals E2f, Smad and Zfx as transcriptional regulators of histones. *Nat. Struct. Mol. Biol.*, **20**, 119–26.

62. Telorac,J., Prykhozhij,S. V., Schöne,S., Meierhofer,D., Sauer,S., Thomas-Chollier,M. and Meijsing,S.H. (2016) Identification and characterization of DNA sequences that prevent glucocorticoid receptor binding to nearby response elements. *Nucleic Acids Res.*, 10.1093/nar/gkw203.

63. Di Stefano,B., Collombet,S., Schou Jakobsen,J., Wierer,M., Sardina,J.L., Lackner,A., Stadhouders,R., Segura-Morales,C., Francesconi,M., Limone,F., *et al.* (2016) C/EBPα creates elite cells for iPSC reprogramming by upregulating Klf4 and increasing the levels of Lsd1 and Brd4. **18**.

64. Mahony,S., Golden,A., Smith,T.J. and Benos,P. V (2005) Improved detection of DNA motifs using a self-organized clustering of familial binding profiles. *Bioinformatics*, **21 Suppl 1**, i283–91.

65. Wingender,E., Schoeps,T. and Dönitz,J. (2013) TFClass: an expandable hierarchical classification of human transcription factors. *Nucleic Acids Res.*, **41**, D165–70.

66. Noyes,M.B., Christensen,R.G., Wakabayashi,A., Stormo,G.D., Brodsky,M.H. and Wolfe,S. a (2008) Analysis of homeodomain specificities allows the family-wide prediction of preferred recognition sites. *Cell*, **133**, 1277–89.

67. Portales-Casamar,E., Thongjuea,S., Kwon,A.T., Arenillas,D., Zhao,X., Valen,E., Yusuf,D., Lenhard,B., Wasserman,W.W. and Sandelin,A. (2010) JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, **38**, D105–D110.

68. Najafabadi,H.S., Mnaimneh,S., Schmitges,F.W., Garton,M., Lam,K.N., Yang,A., Albu,M., Weirauch,M.T., Radovani,E., Kim,P.M., *et al.* (2015) C2H2 zinc finger proteins greatly expand the human regulatory lexicon. *Nat Biotechnol*, **33**, 555–562.

69. Stormo,G.D. and Zhao,Y. (2010) Determining the specificity of protein-DNA interactions. *Nat. Rev. Genet.*, **11**, 751–60.

70. Langfelder,P., Zhang,B. and Horvath,S. (2008) Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics*, **24**, 719–20.

TABLE AND FIGURES LEGENDS

Table 1. Features of software tools available to perform clustering of PSSMs.

Figure 1. Schematic flow chart of the *matrix-clustering* algorithm. The program takes as input one (or several) collection(s) of PSSMs, and calculates the motif similarity using several metrics. One of these metrics is used to group the motifs with hierarchical clustering. A threshold consisting in a combination of metrics is used to cut the global tree in a set of subtrees (forest). Each resulting tree then serves as a guide to progressively align the motifs. The root motifs of each tree are exported as the non-redundant motifs. The trees can be collapsed or expanded at each node dynamically when displayed on the Web page.

Figure 2. Clustering of PSSMs discovered in the Oct4 ChIP-seq peaks using several motif discovery tools. The TF peaks of Oct4 identified by Chen et al (44) were submitted to three *de novo* motif discovery programs: RSAT peak-motifs, MEME-ChIP and HOMER. All discovered PSSMs were clustered simultaneously by *matrix-clustering*. **A.** Hierarchical tree corresponding to cluster_1 (37 motifs), where different Oct motif variants and Sox2 motifs can be observed (highlighted with different colored boxes). The leaves are annotated with the name of the submitted motif, and the name of its collection (one of the three programs). **B.** Reduced tree showing six non-redundant motifs, obtained after manual curation of the cluster_1, by collapsing the branches. **C.** Annotation of the 6 non-redundant variants ("branch motifs") by alignment to reference motifs (see main text). When available in databases (Jaspar or Hocomoco), the ID of the reference motif is indicated. Otherwise, it is replaced by the PMID of the publication where the motif is mentioned. **D**. Heatmap summarising the number of motifs from each collection found in each cluster. **E**. Heatmap of the cross-coverage between each collection.

Figure 3. Clustering of 12 sets of motifs discovered in mouse ESCs TF ChIP-seq peaks. A. Motif tree for one of the 40 clusters identified from 359 motifs discovered in ChIP-seq peak-sets for 12 different TFs involved in ESC pluripotency and proliferation. This cluster contains the Oct-binding motif. The bold black text displayed at each leaf indicates the name of the collection where the motif was discovered. **B.** Heatmap showing the cross-coverage between the 12 motif collections corresponding to the ESC peak-sets. Each row and column corresponds to one TF-specific peak-set. The color of each cell indicates the percentage of motifs from the column-associated collection found in clusters also containing motifs of the row-associated collection. Columns and rows were clustered independently in order to highlight the similarities between motif collections. The table is thus asymmetrical.

Figure 4. Clustering of full Insect and Human motif databases. A. Heatmap representing the distance (Ncor) between all 133 motifs of JASPAR Insects. The grouping of these motifs into 43 clusters is indicated with a colored bar above the heatmap. The black square indicate the large cluster (almost half of the motifs) containing the very similar Homeodomain motifs. **B**. The 64 Homeodomain motifs were manually reduced via manually collapsing the tree branches into 9 motifs. The collapsed

tree is displayed along with the corresponding aligned branch motifs. **C**. Heatmap representing the distance (Ncor) between all 641 motifs of Hocomoco Human. The grouping of these motifs into 255 clusters is indicated with a colored bar above the heatmap. The black squares highlights the multiple small clusters that correspond to different TF Families (e.g. Fox, Gata, Sox, Gli, Stat, etc) . **D**. Repartition of the transcription factor (TF) families within clusters. The barplot indicates that most clusters are composed of a single TF family. The pie chart represents the reasons for observing multiple TF families in a single cluster. The text data used to generate these plots is available as Supplementary Supp_data_Fig_4D.tab

Figure 5. Cross-coverage of public motif databases. Several full public collections were merged and clustered, separately by taxa. The heatmaps of the cross-coverage between each collection is plotted for **A.** Seven insect collections, **B.** Five plant databases and **C.** Ten vertebrate databases. The heatmaps show the cross-coverage of the motifs between all the databases. Note that there are not symmetrical matrices because the number of motifs of each database differs (see methods). **D**. Venn diagrams showing the asymmetry on cross-coverage between two databases of different size.

Supplementary Figures

Supplementary Figure 1. Forest of motifs obtained with study case 1. See supporting website for a dynamically browsable version.

Supplementary Figure 2. Forest of motifs obtained with study case 2. See supporting website for a dynamically browsable version.

Supplementary Figure 3. Clustering of full motif databases. A. Heatmap representing the distance (Ncor) between all motifs of Hocomoco Mouse. The grouping of these motifs into clusters is indicated with a colored bar above the heatmap. **B**. Heatmap representing the distance (Ncor) between all motifs of Jaspar Vertebrates. The grouping of these motifs into clusters is indicated with a colored bar above the heatmap.

Supplementary Figure 4. Clustering of Hocomoco human motifs. A. The barplot shows the distribution of number of clusters per TF family. The left side of the histogram corresponds to consistently clustered families, and the right side to families dispersed across many clusters. **B**. Scatterplot comparing the number of members of each TF family as a function of the number of covered clusters. The name of the families with more than 20 members are shown.

Table 1

	Matlign	STAMP	m2match	DMINDA	motIV	GMACS	matrix-clustering
Year	2007	2007	2013	2014	2014	2015	2016
Clustering method							
Hierarchical	YES	YES	YES	no	YES	no	YES
SOTA	no	YES	no	no	YES	no	no
Genetic Algorithm	no	no	no	no	no	YES	no
minimum-spanning-tree	no	no	no	YES	no	no	no
Multiple alignment of logos	no	no	YES	no	no	no	YES
Alignment with internal gaps	no	YES	no	no	no	no	no
Logo tree	no	no	YES	no	no	no	YES
Familial binding profiles	YES	YES	YES	no	YES	no	YES
Partitioning of input motif set in distinct clusters	no	YES	YES	no	no	YES	YES
Consensus/Logo at tree root	no	YES	YES	no	YES	no	YES
Consensus/Logo at each branch	no	no	no	no	no	no	YES
Multiple collections as input	no	no	no	no	no	no	YES
Accessible via website	YES	YES	YES	YES	no	no	YES
Restriction on number of matrices	no	no	30 (trial account)	no	no	no	no
Publication (PMID or biorxiv DOI)	17559640	17478497	23555204	24753419	Bioconductor package	25627106	This article

Supplementary notes — RSAT matrix-clustering: dynamic exploration and redundancy reduction of transcription factor binding motif collections

Jaime Abraham Castro-Mondragon[1], Sébastien Jaeger[2], Denis Thieffry[3], Morgane Thomas-Chollier[3*] and Jacques van Helden[1*]

Table of Contents

Impact of the similarity metrics and linkage rules (hierarchical clustering step)

Since the last decade there are continuous efforts of the community to explore alternative ways to measure the similarity between motifs, but any metric proposed so far fails to capture some similarities considered as relevant from the inspection of logos, or for biological reasons. Hence, no ideal metric has been found yet. To face this issue, *matrix-clustering* supports all the metrics implemented in the companion program *compare-matrices* [1,2], including the most commonly used metrics (correlation, Euclidian distance, SSD, Sandelin-Wasserman) as well as some custom metrics, e.g. logo dot product, correlation of information content, width-normalized versions of the Pearson Correlation Coefficient (Ncor) and Euclidian Distance (NdEucl), where the values are normalized by the number of aligned columns divided by the total columns of the alignment.

For the hierarchical clustering step, we used distance matrices derived from all the similarity metrics supported in *compare-matrices*, and focused more particularly on the popular Pearson's correlation (*cor*) and Euclidian Distance (dEucl). A previously reported drawback of these metrics is that the motifs may be aligned on positions with low-information content, at the extremities of the motifs [3,4]. Figure Ia shows an example of such spurious alignment obtained with Pearson's correlation (*cor*) using as similarity cutoff cor >= 0.8.

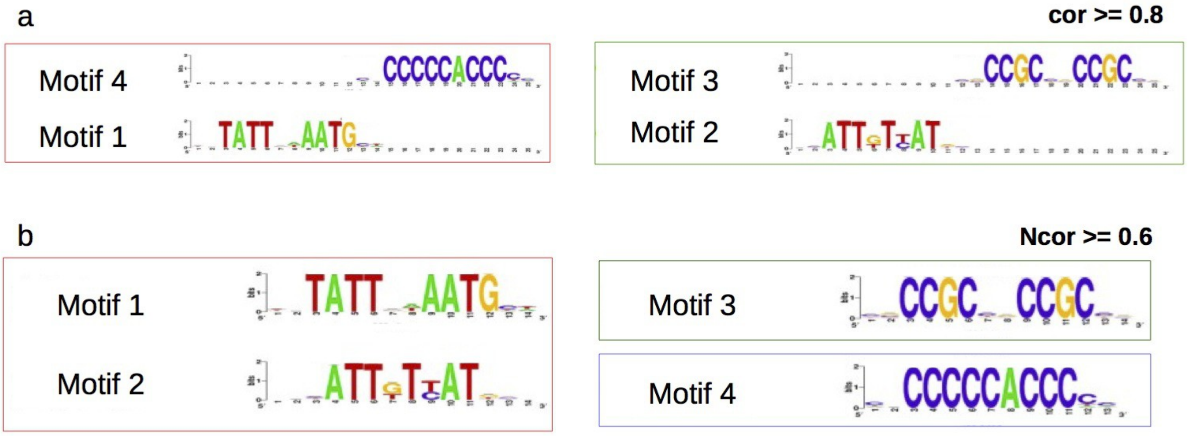


Figure I: Example of motif alignments obtained with RSAT compare-matrices. a. Spurious alignment obtained with Pearson's correlation (*cor*), where the motifs are aligned on one or two positions at the extremities of the motifs. **b.** Correct alignments obtained when using the normalized version of the Pearson's correlation (*Ncor*).

One approach to filter out such spurious alignment is to impose a threshold on the width of the alignment (e.g. at least 5 columns should be shared), but this would systematically discard small motifs from the alignments. A more suitable solution to this problem is to use width-normalized Pearson correlation (Ncor) [1,4]. When using these normalized metrics, the spurious alignments are indeed not observed anymore (Fig. Ib) [1,4].

In addition of the similarity metric, another parameter that strongly impacts on the number and composition of the clusters is the linkage rule used to build the global tree. This parameter affects not only the number of clusters, but also the structure of the trees (order of motif incorporation in the progressive alignment), and hence the motifs repartition among the clusters and the branch logos. To illustrate this, we analysed the study case 1 (Oct4 motifs discovered by HOMER, MEME and RSAT) by running *matrix-clustering* (Ncor as metric to build the tree, and Ncor >= 0.4, cor >= 0.6 as combined threshold for partitioning) with three alternative agglomeration rules: single, average, complete linkage (Fig. II). There are notable differences of tree topologies and motif clusters, depending on the linkage rule. The 66 motifs are regrouped respectively in 9 (single linkage), 13 (average), 19 (complete)clusters. The largest cluster (in red) obtained with single and average rules (Fig. IIa, IIb) is split into 4 smaller clusters (red, green, light green and orange) with the complete linkage (Fig. IIc). As the complete linkage is the most restrictive one, the number of clusters is generally higher than with the other methods. By contrast, the single linkage is the most permissive, it allows more motifs to be grouped together and hence defines a smaller number of clusters. Generally, we suggest to run *matrix-clustering* using the average linkage method.

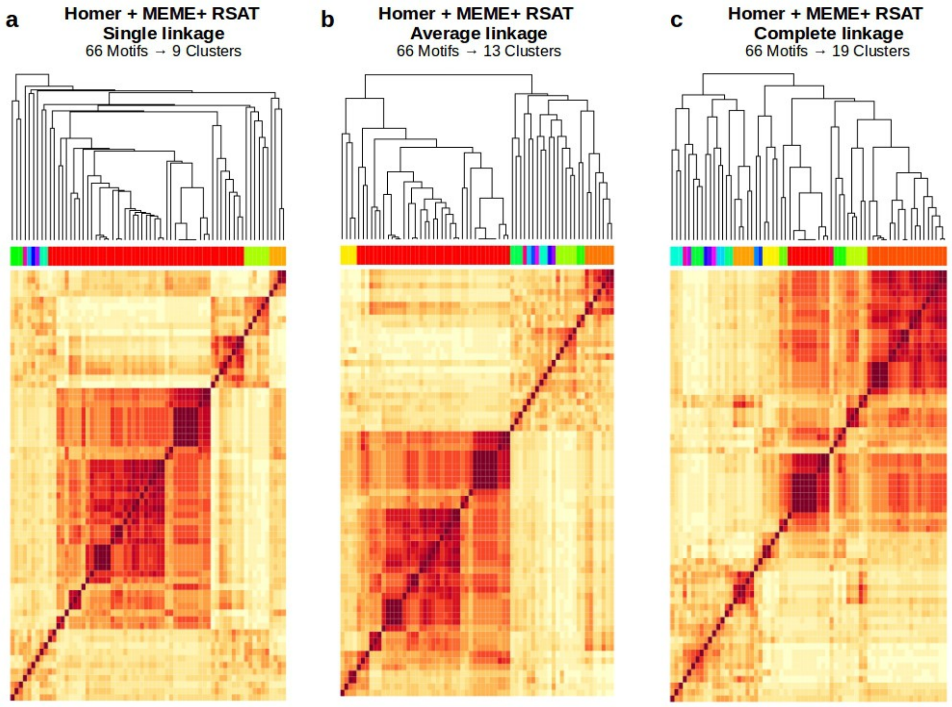


Figure II: Effect of the linkage rules on the final clusters. The data of study case 1 (Oct4 motifs obtained with Homer, MEME and RSAT) is clustered with the same parameters except for the agglomeration rule: single (**a**), average (**b**) or complete linkage (**c**). Heatmaps represent the all-versus-all motifs, with a color scale reflecting

the width-normalized correlation (Ncor). The hierarchical tree derived from the Ncor distance matrix is depicted above. Clusters are highlighted with a color bar below the trees.

Impact of the threshold (partitioning step)

We analysed the impact of different thresholds on the number of clusters with the data of study case 3 (JASPAR Insects). We systematically tested all combinations of thresholds on *cor* and *Ncor,* for values ranging from 0 to 1 with an increment of 0.1. For each result, we counted the number of clusters in order to observe the impact of the combination of thresholds (Fig. III). For very low *cor* and *Ncor* values, most motifs are grouped in a single cluster, whilst with higher values the number of clusters increases, to a point where all motifs are separated in singletons. In our experience, relevant results are obtained with thresholds ranging from Ncor >= 0.4-0.7 combined with cor >=0.6-0.8 (Fig. III, red square).

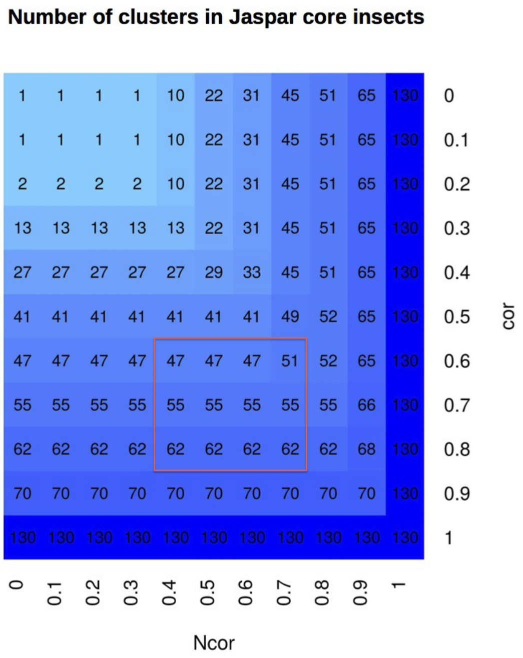


Figure III: Impact of threshold Combinations on the number of clusters. Heatmap showing the number of clusters obtained with threshold combinations of *Ncor* and *cor* ranging from 0 to 1. The red square highlights the combinations that usually result in coherent clusters.

In general, we made two observations: (i) a threshold consisting in a combination of multiple metrics improves the consistency of the clusters, (ii) given that the hierarchical tree is built on one metric, the normalized metrics group the motifs more coherently than those not normalized.

Negative control with column-permuted motifs

In order to test the relevance of the clusters returned by *matrix-clustering*, the program can create negative controls by randomly permuting the input PSSM's columns using the tool RSAT *convert-matrix*. This approach has the advantage of maintaining the residue frequencies and the information content of each PSSM, but the consensus and the order of the relevant positions (the biological properties) are usually lost after permutation. The exact number of clusters in the negative control varies each time the set of matrices is permuted. To have an estimate of how many clusters can be found each time the motifs are permuted, we generated 100 sets of permutations of each collection.

Figure IV displays the clustering of the study case 3 collections (JASPAR insects and Hocomoco human), where all motifs are column-permuted. For the insects motifs (133 motifs), the number of clusters is quite high after permutations (Fig IVa, IVc), with a median at 85 clusters for 133 motifs, meaning that many clusters are singletons. However, we noticed in study case 3 that most insect motifs are grouped in a big cluster containing a high number of the Hox-like motifs. Permuting these short motifs with ATTA core produces similar motifs that can still be grouped together. This behavior is also observed in low-complexity motifs (e.g. A-rich motifs). For human motifs (641 motifs), after permutation, the number of cluster is drastically increased, with a median of 590 clusters over the 100 repetitions, thus with even more singletons than for the insect collection (Fig IVb, IVd).

This test shows the importance of the order of the conserved residues in the motifs as a factor which affects considerably the similarity. Indeed, these permuted matrices have the same information content as the real motif; the only difference is that the columns were sorted in a different way. This test also shows the adequacy of the agglomeration rule and selected thresholds to separate unrelated motifs.

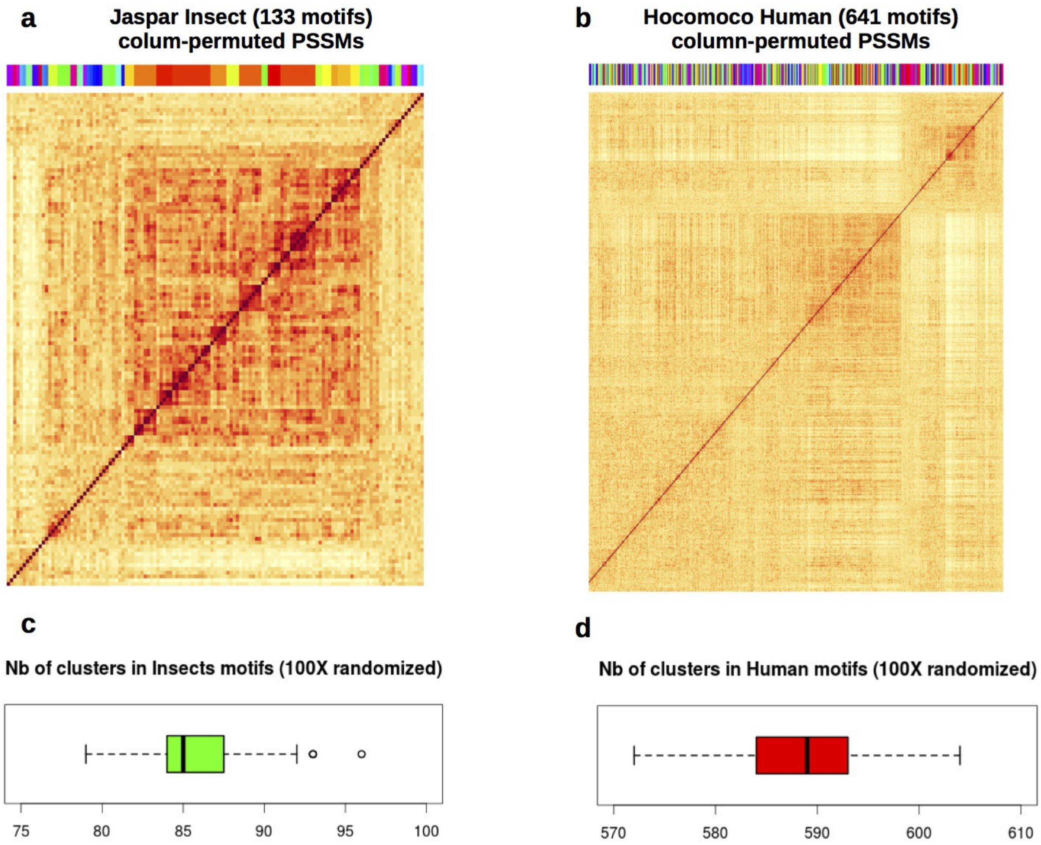


Figure IV: Clusters obtained after permutations of full collections. (a, c) Column-permuted motifs from Jaspar core insects**. (b,d)** Column-permuted motifs from Hocomoco human. Heatmap color scales indicate the width-normalized correlation (Ncor) of all-versus-all motifs. Clusters are highlighted with a color bar above the heatmap. The distribution of the number of clusters across 100 permutations is shown as a boxplot.

Comparison with other motif clustering tools

Table 1 provides a list of features supported by existing motif clustering tools.

For the sake of comparison, we submitted some of our study cases to alternative motif clustering tools. This analysis was restricted to study cases 1 and 3, since no other tool currently supports multiple collections as those of our study cases 2 and 4. For study case 1, we merged the 66 motifs from peak-motifs, MEME and HOMER in a single file in order to submit it to the alternative matrix clustering tools. Moreover, the public Web interface of m2match is restricted to 30 motifs, we thus restricted the analysis to the 22 motifs discovered by peak-motifs. Each tool was run on its web interface, with default parameters. In addition, we ran STAMP on the command line, because the Web interface does not enable to activate the tree partitioning option.

All results are available on the supporting Web site:
http://teaching.rsat.fr/data/published_data/Castro_2016_matrix-clustering/

STAMP

With our first study case, STAMP returns a tree whose main branches correctly separate the Oct, SOCT (composite Oct-Sox), and the GC-rich motifs. The familial binding profile summarizes the clustering tree as a whole, and only retains the 8 most conserved columns of the alignment between input motifs, which corresponds to the canonical Oct binding motif. When used on the command line, stamp supports an option to partition the tree in separate clusters[6]. STAMP identified 14 clusters among the 66 merged motifs discovered by MEME, Homer and Oct4, whereas matrix-clustering found 13 clusters.

With Insect JASPAR database, STAMP web interface produces a single tree regrouping all the 133 motifs, together with a multiple alignment of all their consensus strings, encompassing very different motifs, which would a priori seem non-alignable (e.g. GMCCCCCCGCNG and TATGCAAATNA). The global multiple alignment is summarized by the familial binding profile “ATTA” which corresponds to the core binding consensus of the Hox factors. This reflects the over-representation of Hox motifs in Jaspar insects, but is not representative of the diversity of motifs in the full database. Besides, STAMP presents all alignments in the form of consensus strings, whereas *matrix-clustering* represents them as logos. Its command-line interface however allows to activate a tree partitioning option, which identifies 33 clusters (43 clusters obtained with matrix-clustering), however it does not produce any graphical output (logos, tree views). The clustering of the full HOCOMOCO Human database gives similar results: all motifs are regrouped into a single tree (website) which is partitioned in 180 clusters (command-line version) in contrast with the 255 clusters produced by matrix-clustering.

In summary, STAMP Web interface and command-line produce complementary information, whereas matrix-clustering provides both the clusters and a rich and dynamic browsable visualization interface, irrespective of the submission mode (Web or command line).

STAMP is remarkably fast as compared to matrix-clustering, especially for motif collection of moderate size (Supp. Table 1). Note that for *matrix-clustering*, the relationship between database size and computing time is not monotonous. The processing time depends on the structure of the dataset (number of clusters, cluster sizes, singletons, ...). Also note that matrix-clustering produces branch motifs and logos for each branch of each tree, whereas STAMP only produces on family-binding profile (FBP) per cluster.

Supp. Table I: comparison of time required for matrix-scan and STAMP to cluster full motif databases.

Collection	Motif nb	Matrix-clustering (minutes)	STAMP (minutes)
Jaspar Insects	133	10	1
Hocomoco mouse	427	20	13
Jaspar vertebrate	519	98	23
Hocomoco human	641	55	42

Matalign

Matlign[7] splits the 66 Oct motifs into 22 clusters. The result is presented as a bitmap image where motif names are unreadable due to the vertical orientation of the tree, followed by a detailed list of merged matrices corresponding to the different nodes of the tree.

m2match

m2match is part of TRANSFAC tools[4], which are under license. However, the public interface allows to submit a restricted number of motifs (max 30). We tested the tool with the 22 motifs discovered by RSAT peak-motifs in the Oct4 ChIP-seq peaks. The tool produces a tree, which is partitioned into distinct clusters, whose matrices are aligned to produce branch motifs ("Familial Binding Profiles") and logos.

Gmacs

Gmacs[8] only runs as a command line tool. It is extremely easy to install and run, with a very few options. However, it only produces a text file indicating the cluster composition, without any other information (for example there is no motif alignments). The use of this version is thus rather limited.

References

1. Thomas-Chollier, M. *et al*. RSAT 2011: Regulatory sequence analysis tools. *Nucleic Acids Res*. **39,** (2011).

2. Medina-Rivera, A. *et al*. RSAT 2015: Regulatory Sequence Analysis Tools. *Nucleic Acids Research* 1–7 (2015). doi:10.1093/nar/gkv362

3. Habib, N., Kaplan, T., Margalit, H. & Friedman, N. A novel Bayesian DNA motif comparison method for clustering and retrieval. *PLoS Comput. Biol.* **4,** (2008).

4. Stegmaier, P., Kel, A., Wingender, E. & Borlak, J. A Discriminative Approach for Unsupervised Clustering of DNA Sequence Motifs. *PLoS Comput. Biol.* **9,** (2013).

5. Mathelier, A. *et al*. JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res*. **42,** D142–7 (2014).

6. Mahony, S., Auron, P. E. & Benos, P. V. DNA familial binding profiles made easy: Comparison of various motif alignment and clustering strategies. *PLoS Comput. Biol.* **3,** 0578–0591 (2007).

7. Kankainen, M. & Löytynoja, A. MATLIGN: a motif clustering, comparison and matching tool. *BMC Bioinformatics* **8,** 189 (2007).

8. Broin, P. Ó., Smith, T. J. & Golden, A. A. Alignment-free clustering of transcription factor binding motifs using a genetic-k-medoids approach. *BMC Bioinformatics* **16,** 22 (2015).

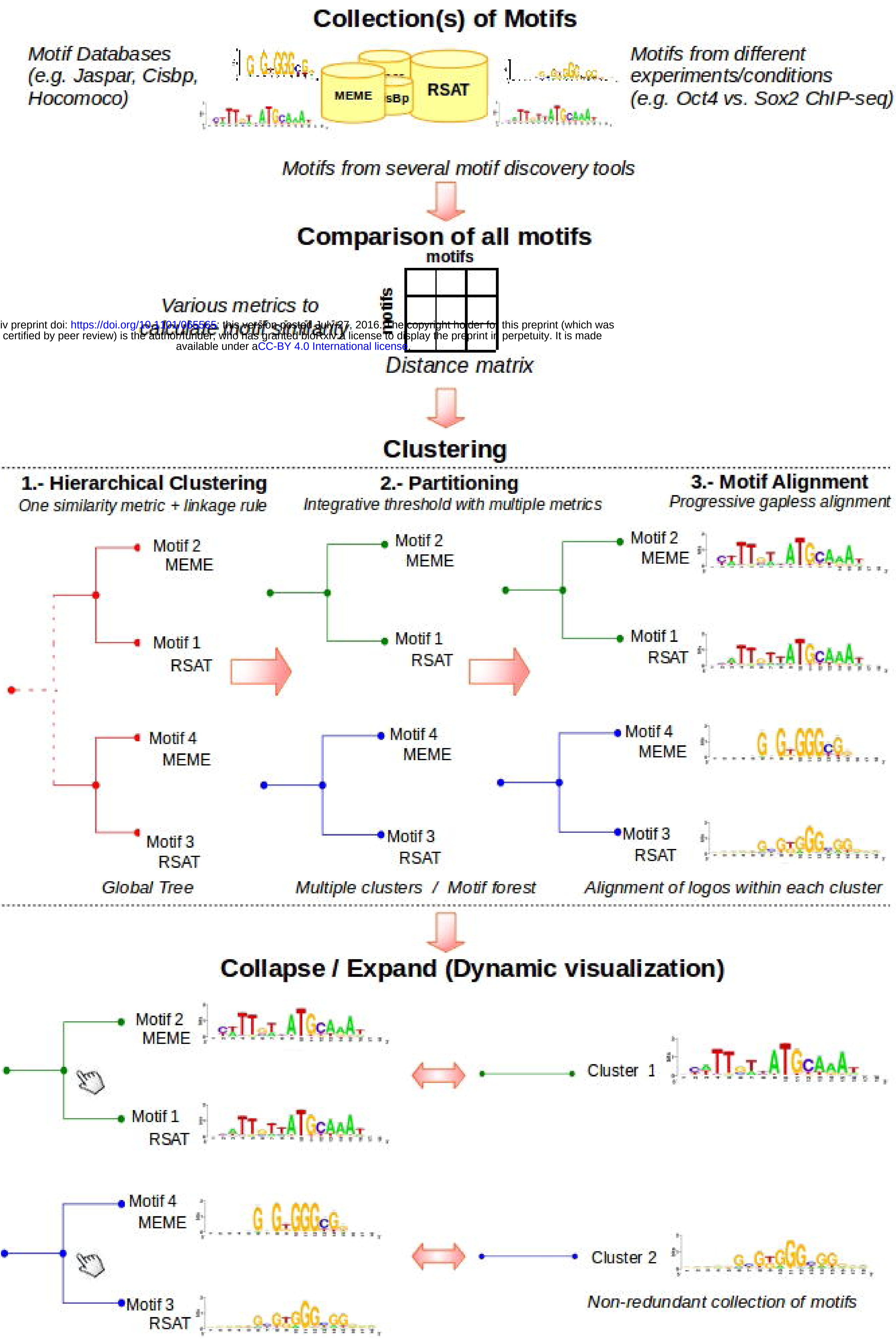

Collection(s) of Motifs
Motif Databases (e.g. Jaspar, Cisbp, Hocomoco)
MEME
sBp
RSAT
Motifs from different experiments/conditions (e.g. Oct4 vs. Sox2 ChIP-seq)
Motifs from several motif discovery tools
Comparison of all motifs
motifs
motifs
Various metrics to calculate motif similarity
Distance matrix
Clustering
1.- Hierarchical Clustering
One similarity metric + linkage rule
2.- Partitioning
Integrative threshold with multiple metrics
3.- Motif Alignment
Progressive gapless alignment
Motif 2 MEME
Motif 1 RSAT
Motif 4 MEME
Motif 3 RSAT
Global Tree
Multiple clusters / Motif forest
Alignment of logos within each cluster
Collapse / Expand (Dynamic visualization)
Cluster 1
Cluster 2
Non-redundant collection of motifs

Figure 1

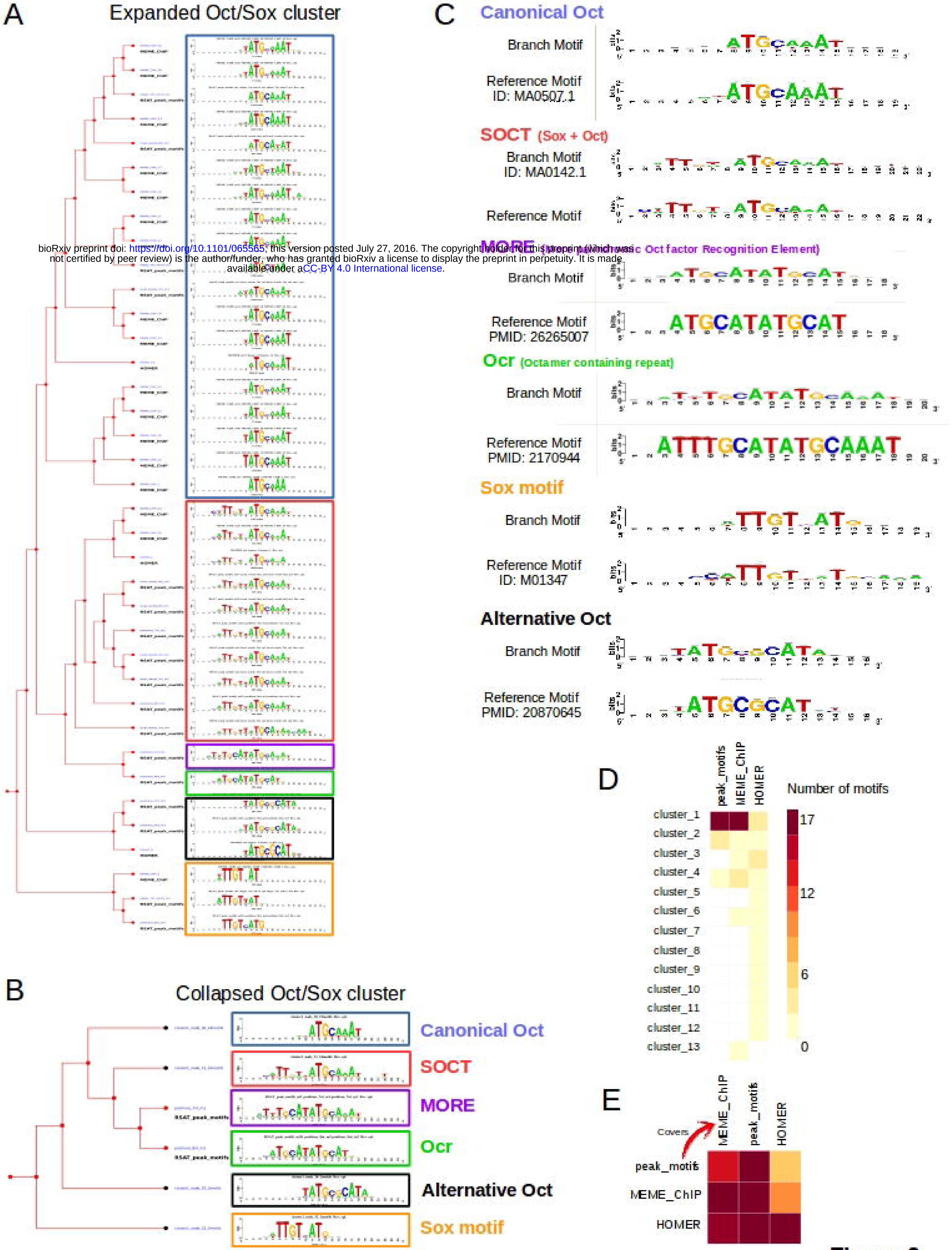

A
Expanded Oct/Sox cluster
B
Collapsed Oct/Sox cluster
Canonical Oct
SOCT
MORE
Ocr
Alternative Oct
Sox motif
C
Canonical Oct
Branch Motif
Reference Motif
ID: MA0507.1
SOCT (Sox + Oct)
Branch Motif
ID: MA0142.1
Reference Motif
MORE (More palindromic Oct factor Recognition Element)
Branch Motif
Reference Motif
PMID: 26265007
Ocr (Octamer containing repeat)
Branch Motif
Reference Motif
PMID: 2170944
Sox motif
Branch Motif
Reference Motif
ID: M01347
Alternative Oct
Branch Motif
Reference Motif
PMID: 20870645
bits
5'
3'
ATGCAAAT
ATGCATATGCAT
ATTTGCATATGCAAAT
TTGT AT
ATGCGCAT
D
peak_motifs
MEME_ChIP
HOMER
Number of motifs
cluster_1
cluster_2
cluster_3
cluster_4
cluster_5
cluster_6
cluster_7
cluster_8
cluster_9
cluster_10
cluster_11
cluster_12
cluster_13
17
12
6
0
E
Covers
MEME_ChIP
peak_motifs
HOMER
peak_motifs
MEME_ChIP
HOMER

Figure 2

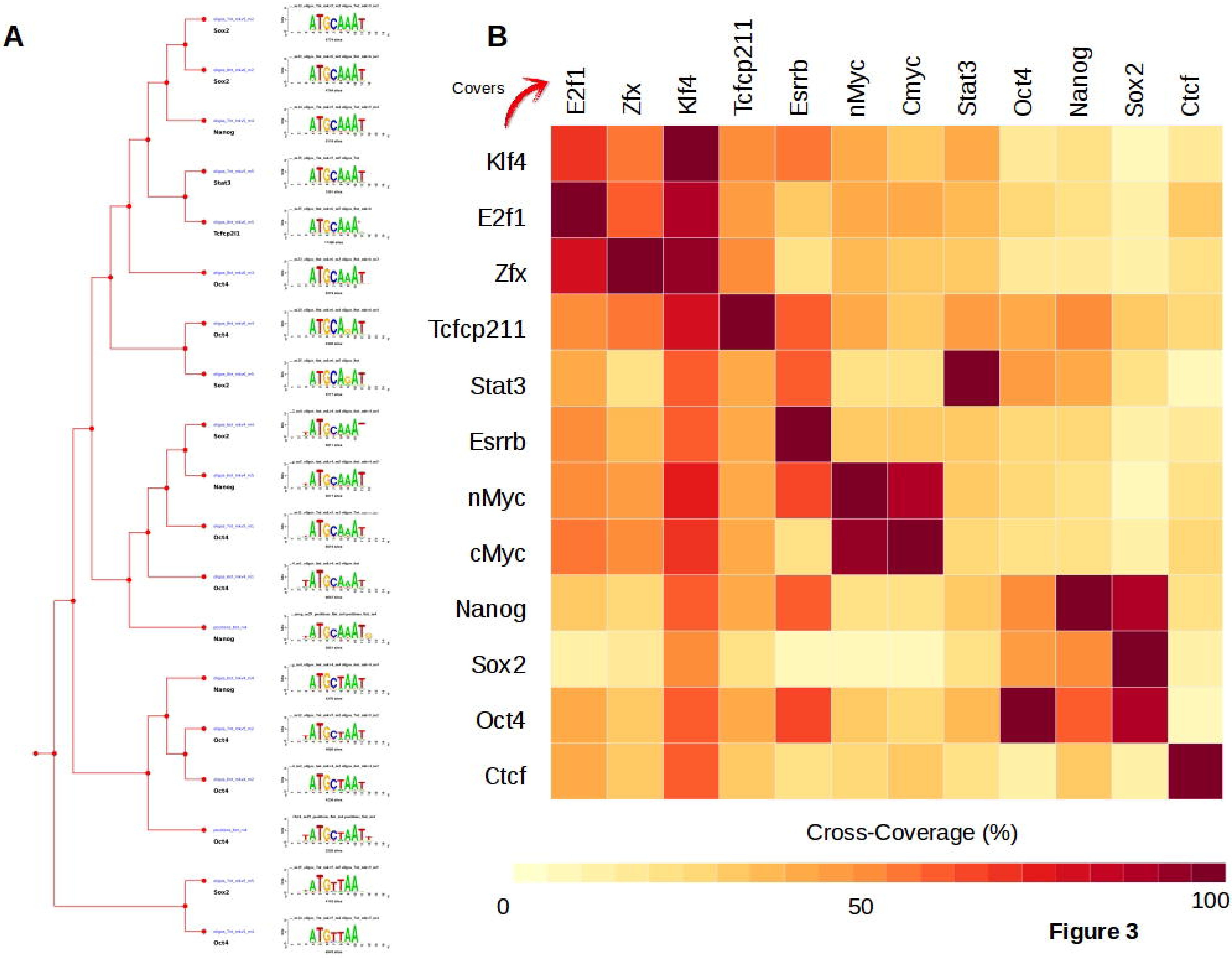

A
Sox2
Sox2
Nanog
Stat3
Tcfcp2l1
Oct4
Oct4
Sox2
Sox2
Nanog
Oct4
Oct4
Nanog
Nanog
Oct4
Oct4
Oct4
Sox2
Oct4
B
Covers
E2f1
Zfx
Klf4
Tcfcp211
Esrrb
nMyc
Cmyc
Stat3
Oct4
Nanog
Sox2
Ctcf
Klf4
E2f1
Zfx
Tcfcp211
Stat3
Esrrb
nMyc
cMyc
Nanog
Sox2
Oct4
Ctcf
Cross-Coverage (%)
0
50
100

Figure 3

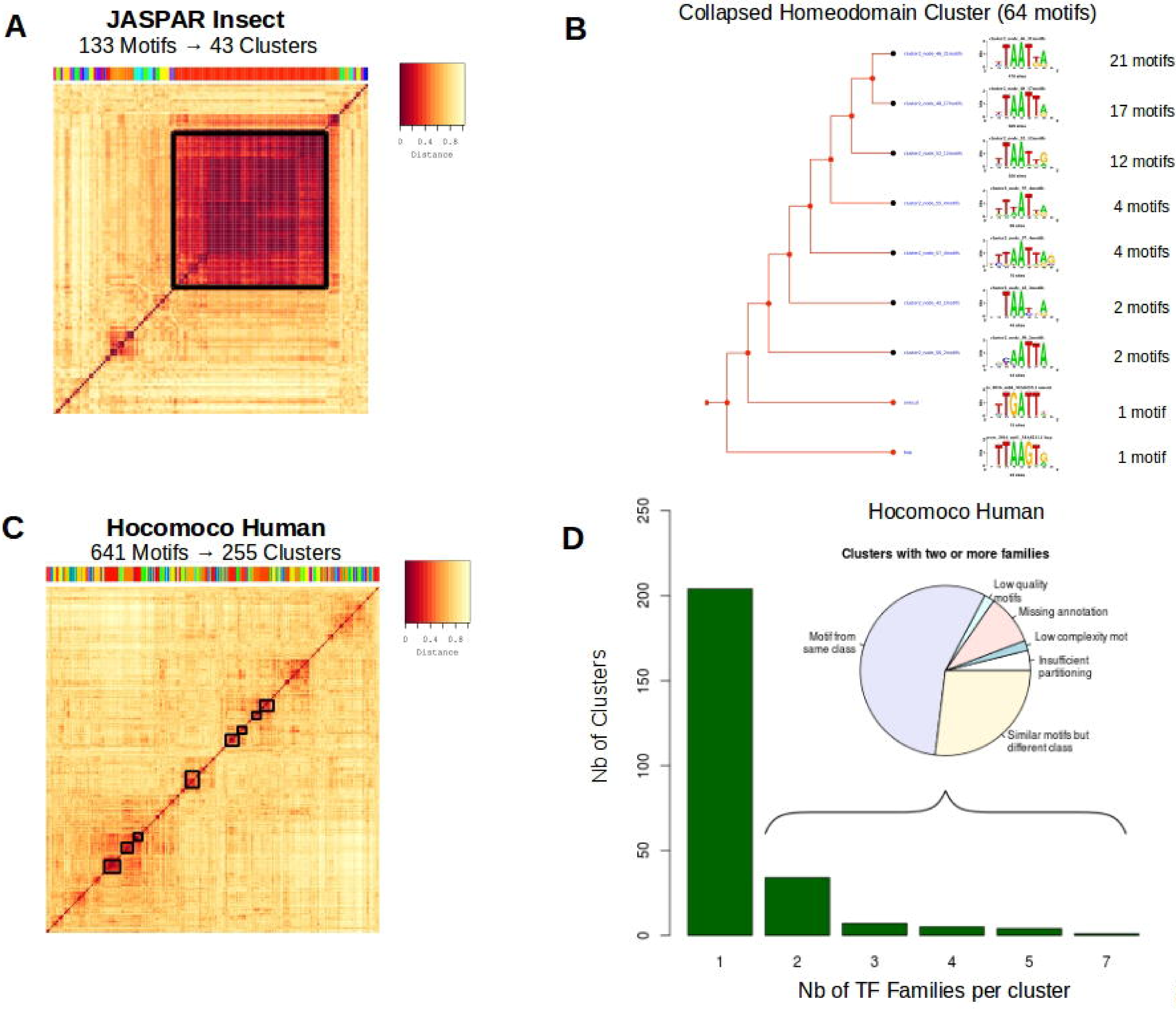

A
JASPAR Insect
133 Motifs → 43 Clusters
0 0.4 0.8
Distance
B
Collapsed Homeodomain Cluster (64 motifs)
21 motifs
17 motifs
12 motifs
4 motifs
4 motifs
2 motifs
2 motifs
1 motif
1 motif
C
Hocomoco Human
641 Motifs → 255 Clusters
0 0.4 0.8
Distance
D
Hocomoco Human
Clusters with two or more families
Low quality motifs
Missing annotation
Low complexity mot
Insufficient partitioning
Similar motifs but different class
Motif from same class
Nb of Clusters
0
50
100
150
200
250
1
2
3
4
5
7
Nb of TF Families per cluster

Figure 4

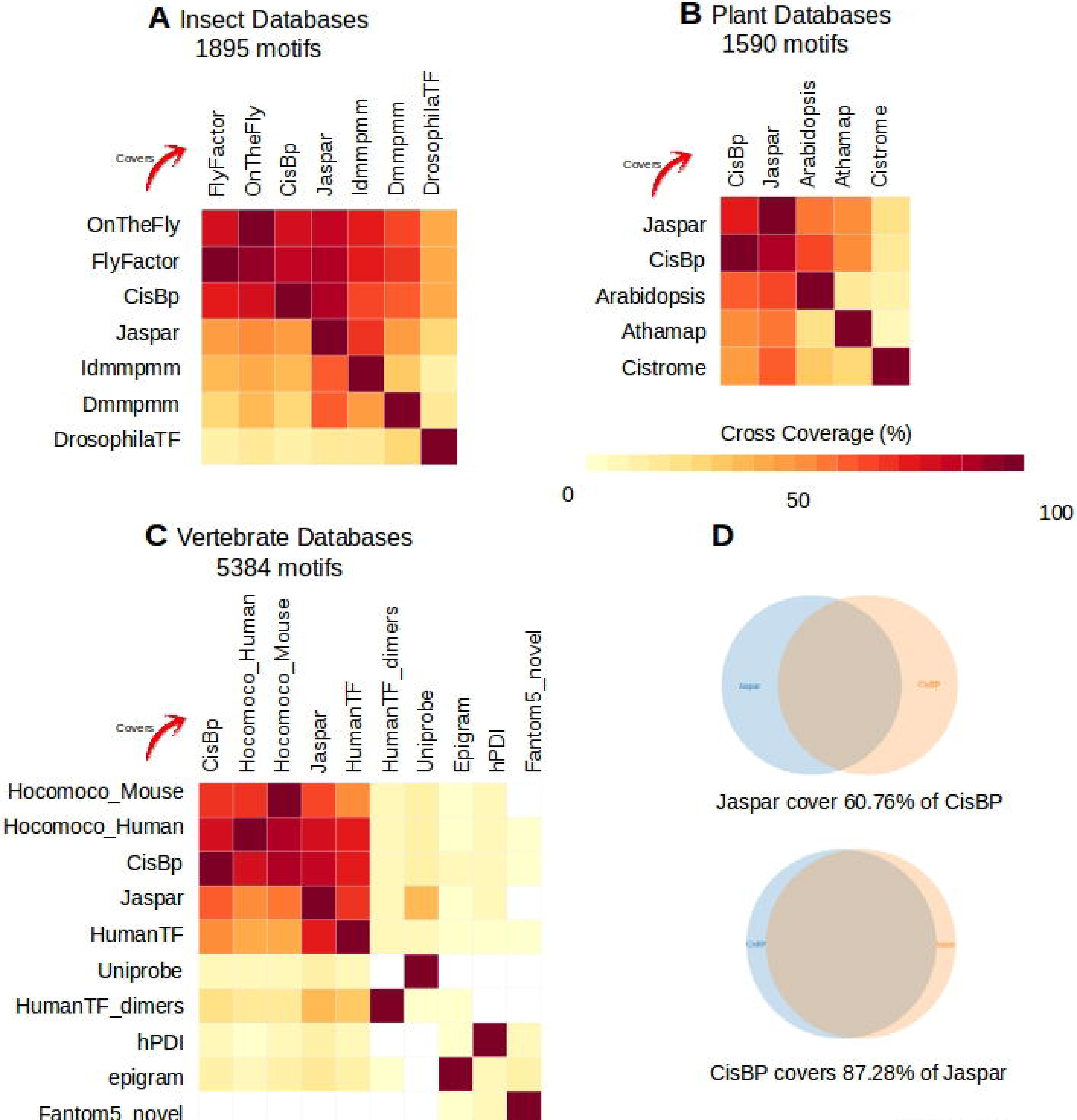
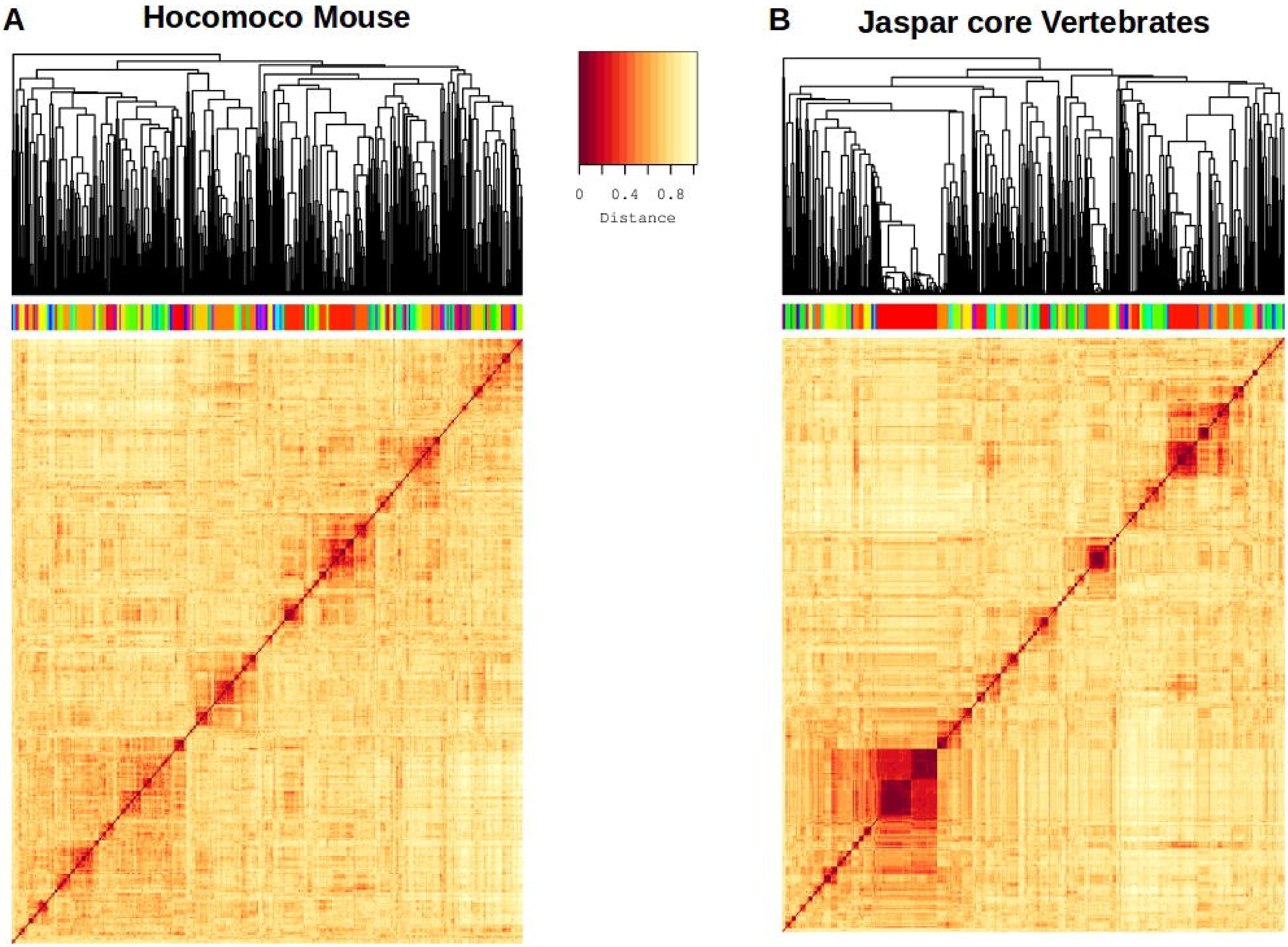
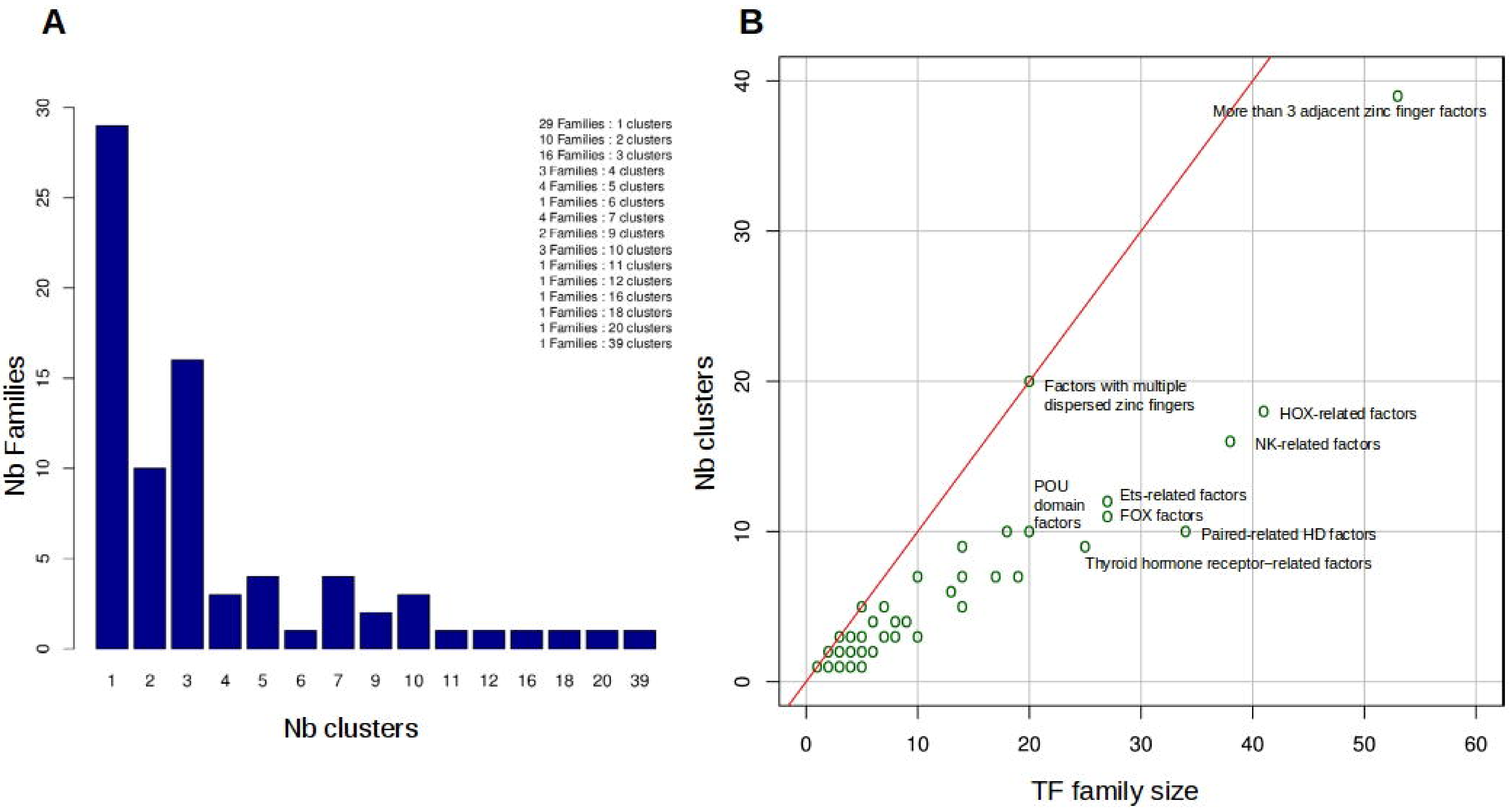

A Insect Databases
1895 motifs
Covers
FlyFactor
OnTheFly
CisBp
Jaspar
Idmmpmm
Dmmpmm
DrosophilaTF
OnTheFly
FlyFactor
CisBp
Jaspar
Idmmpmm
Dmmpmm
DrosophilaTF
B Plant Databases
1590 motifs
Covers
CisBp
Jaspar
Arabidopsis
Athamap
Cistrome
Jaspar
CisBp
Arabidopsis
Athamap
Cistrome
Cross Coverage (%)
0
50
100
C Vertebrate Databases
5384 motifs
Covers
CisBp
Hocomoco_Human
Hocomoco_Mouse
Jaspar
HumanTF
HumanTF_dimers
Uniprobe
Epigram
hPDI
Fantom5_novel
Hocomoco_Mouse
Hocomoco_Human
CisBp
Jaspar
HumanTF
Uniprobe
HumanTF_dimers
hPDI
epigram
Fantom5_novel
D
Jaspar
CisBP
Jaspar cover 60.76% of CisBP
CisBP
Jaspar
CisBP covers 87.28% of Jaspar

Figure 5



Supp Figure 3



Supp Figure 4