

Cross-platform motif discovery and benchmarking to explore binding specificities of poorly studied human transcription factors

Supplementary Information

Ilya E. Vorontsov, Ivan Kozin, Sergey Abramov, Alexandr Boytsov, Arttu Jolma, Mihai Albu, Giovanna Ambrosini, Katerina Faltejskova, Antoni J. Gralak, Nikita Gryzunov, Sachi Inukai, Semyon Kolmykov, Pavel Kravchenko, Judith F. Kribelbauer-Swietek, Kaitlin U. Lavery, Vladimir Nozdrin, Zain M. Patel, Dmitry Penzar, Marie-Luise Plescher, Sara E. Pour, Rozita Razavi, Ally W.H. Yang, Ivan Yevshin, Arsenii Zinkevich, Matthew T. Weirauch, Philipp Bucher, Bart Deplancke, Oriol Fornes, Jan Grau, Ivo Grosse, Fedor A. Kolpakov, The Codebook/GRECO-BIT Consortium, Vsevolod J. Makeev, Timothy R. Hughes, Ivan V. Kulakovskiy

Contents (pdf): Supplementary Figures 1-12.

Supplementary Data (xlsx):

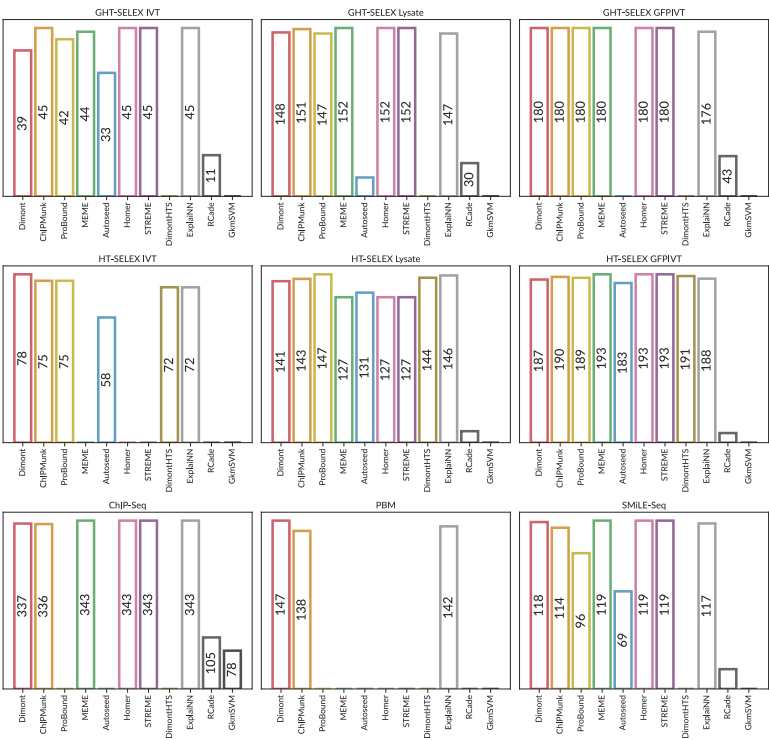
Supplementary Data 1. Overview of the Codebook experiments used in the study.

Supplementary Data 2. Analysis of the transferability of Archipelago models.

Supplementary Data 3. List of software tools used in the study.

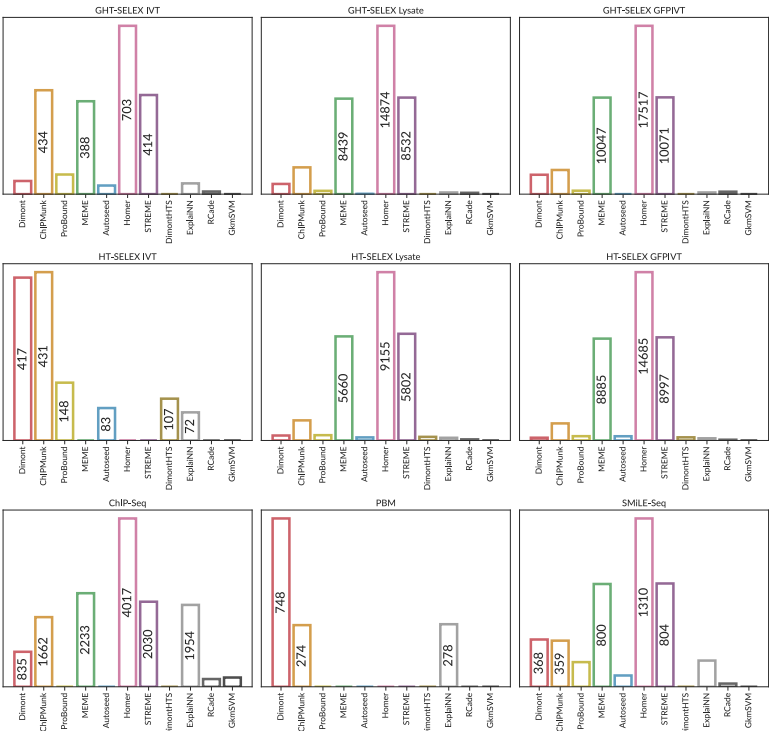
A

The total number of experiments of a particular type successfully processed by each motif discovery tool



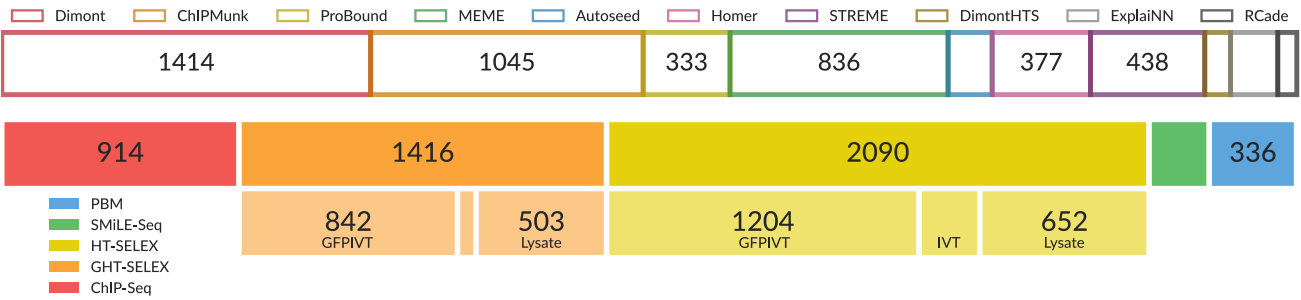
B

The number of motifs generated from each type of experiment by each motif discovery tool

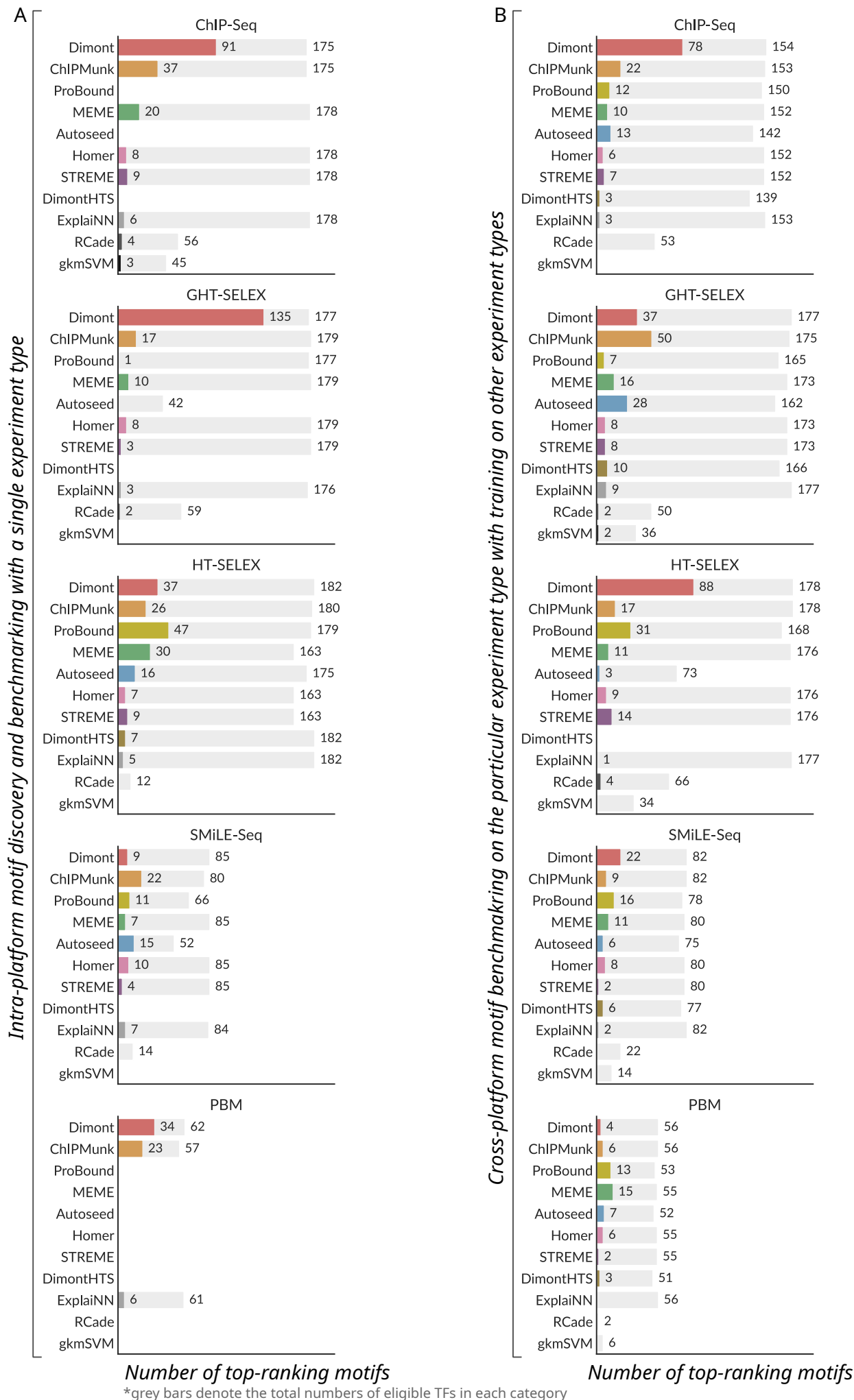


C

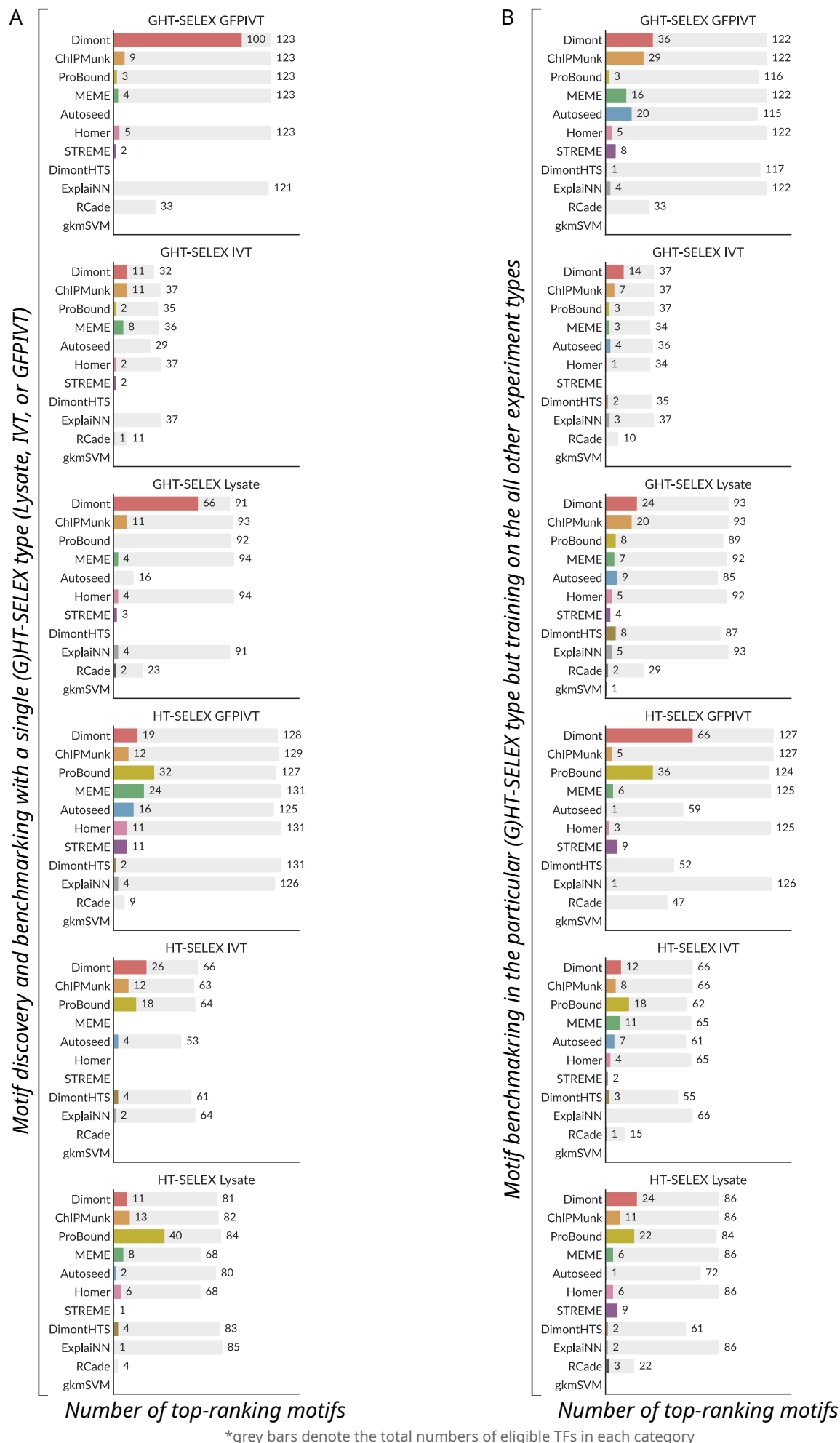
Top-20 motifs per TFs: number of motifs originating from particular motif discovery tools and experiment types



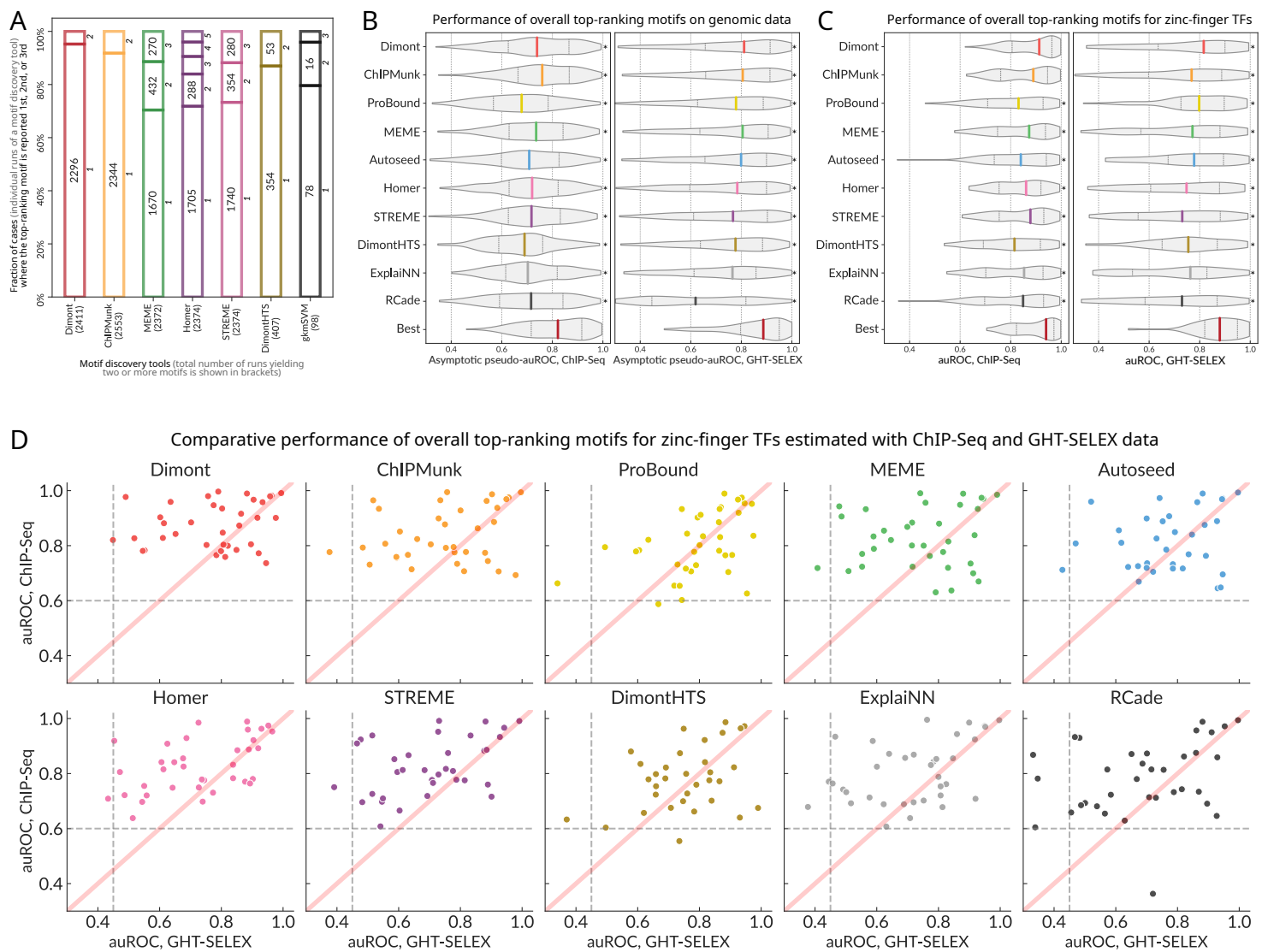
Supplementary Figure 1. A: The number of experiments successfully processed by each motif discovery tool. **B:** The number of motifs generated from each type of experiment by each motif discovery tool. **C:** The composition of the collection of overall top-20 motifs shown in the same way as the set of top-ranking motifs in Figure 1.



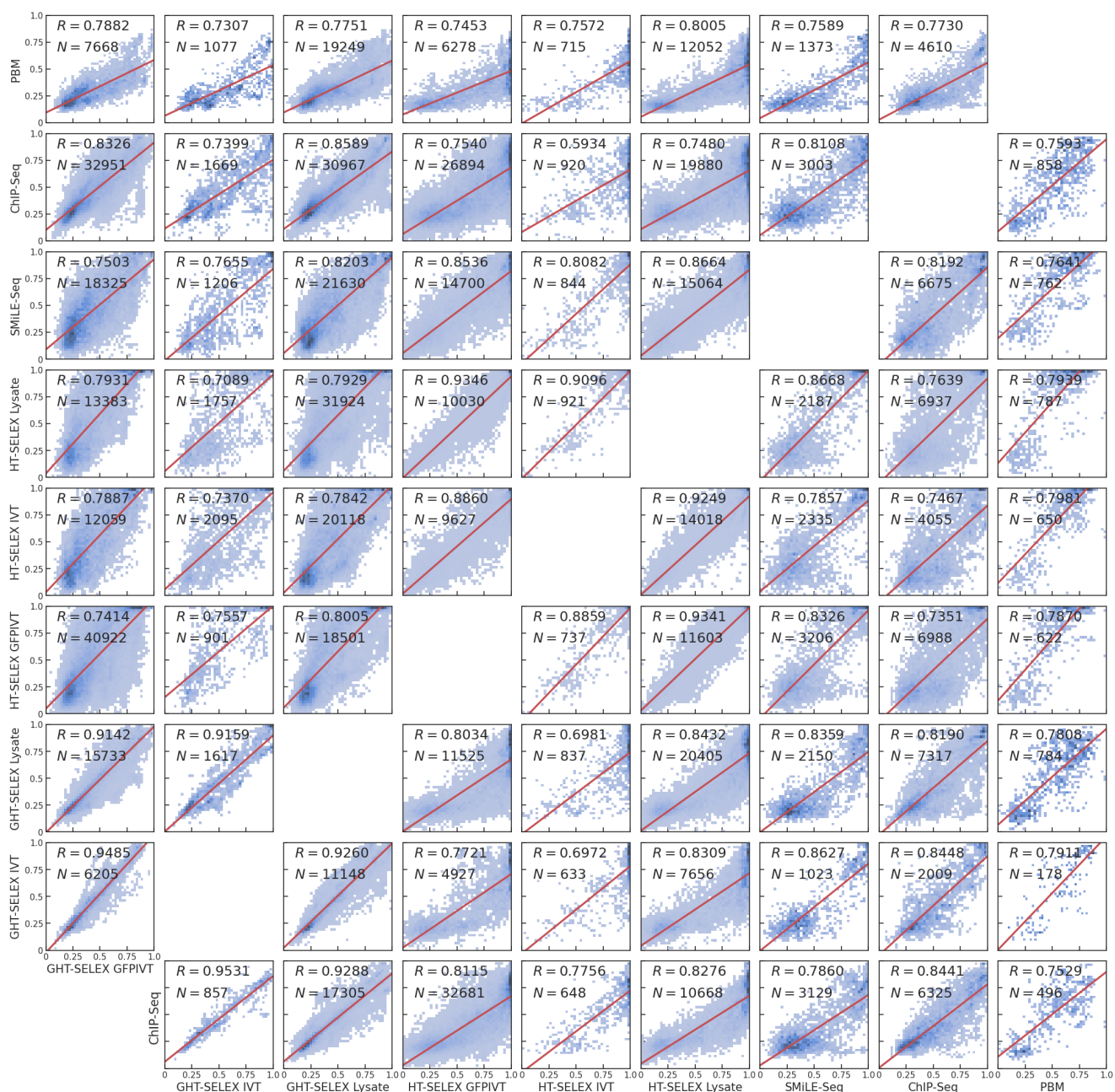
Supplementary Figure 2. Numbers of top-ranking motifs yielded by different motif discovery tools detected in intra- (single experiment type) and cross-platform benchmarking (between experiment types). **A:** The number of motifs (TFs) for which a particular tool produced the top-ranking motif if trained and tested on the same type of experiment. **B:** The number of motifs (TFs) for which a particular tool yielded the top-ranking motif when tested on a particular type of experiment but trained on any of the other types. The total number of TFs with at least one benchmarked motif for a particular combination of a tool and an experiment type is shown in gray.



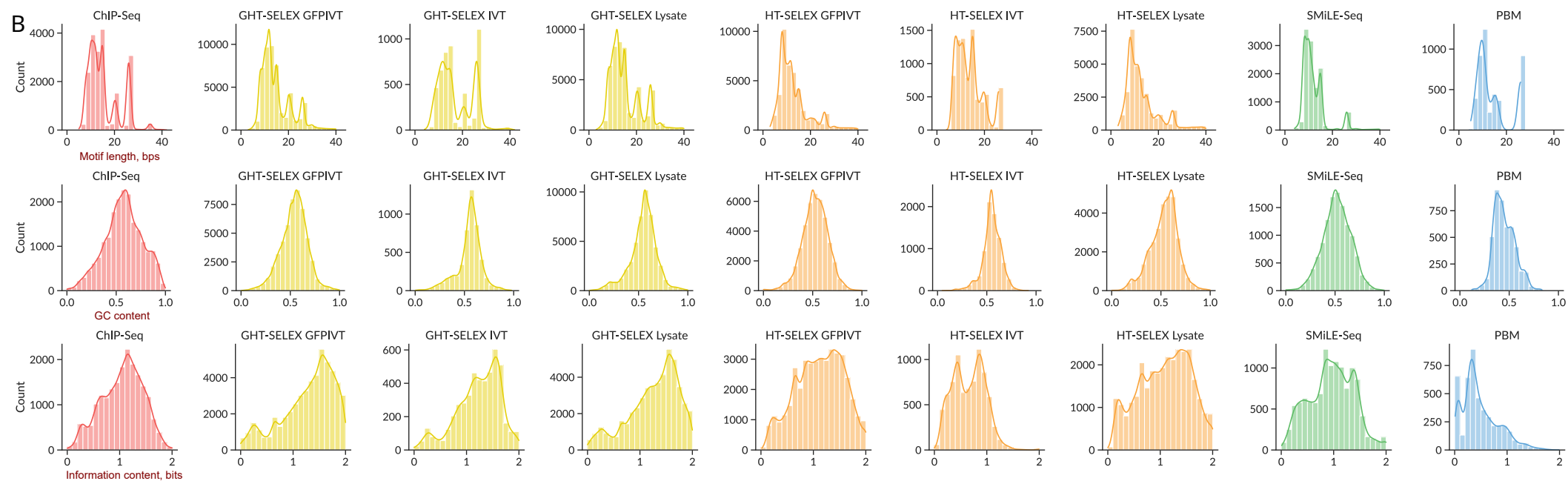
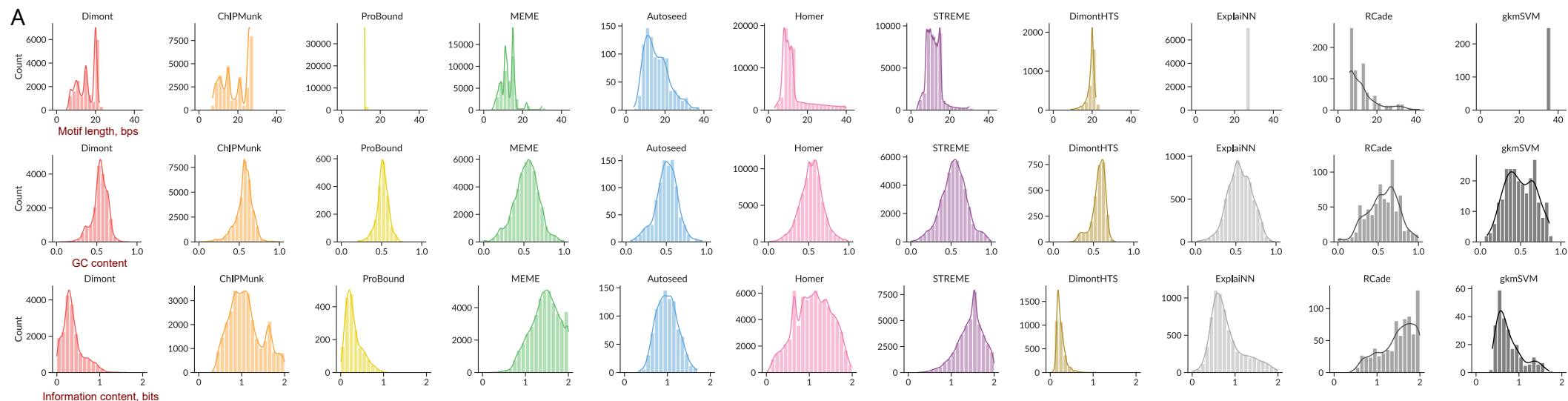
Supplementary Figure 3. Numbers of top-ranking motifs from different motif discovery tools detected in intra- and cross-platform benchmarking considering types of GHT-SELEX and HT-SELEX (IVT, Lysate, GFPIVT) independently. **A:** Motif discovery and testing performed with the same experiment type. **B:** Motif discovery and testing performed with different platforms, including training and testing between (G)HT-SELEX types.



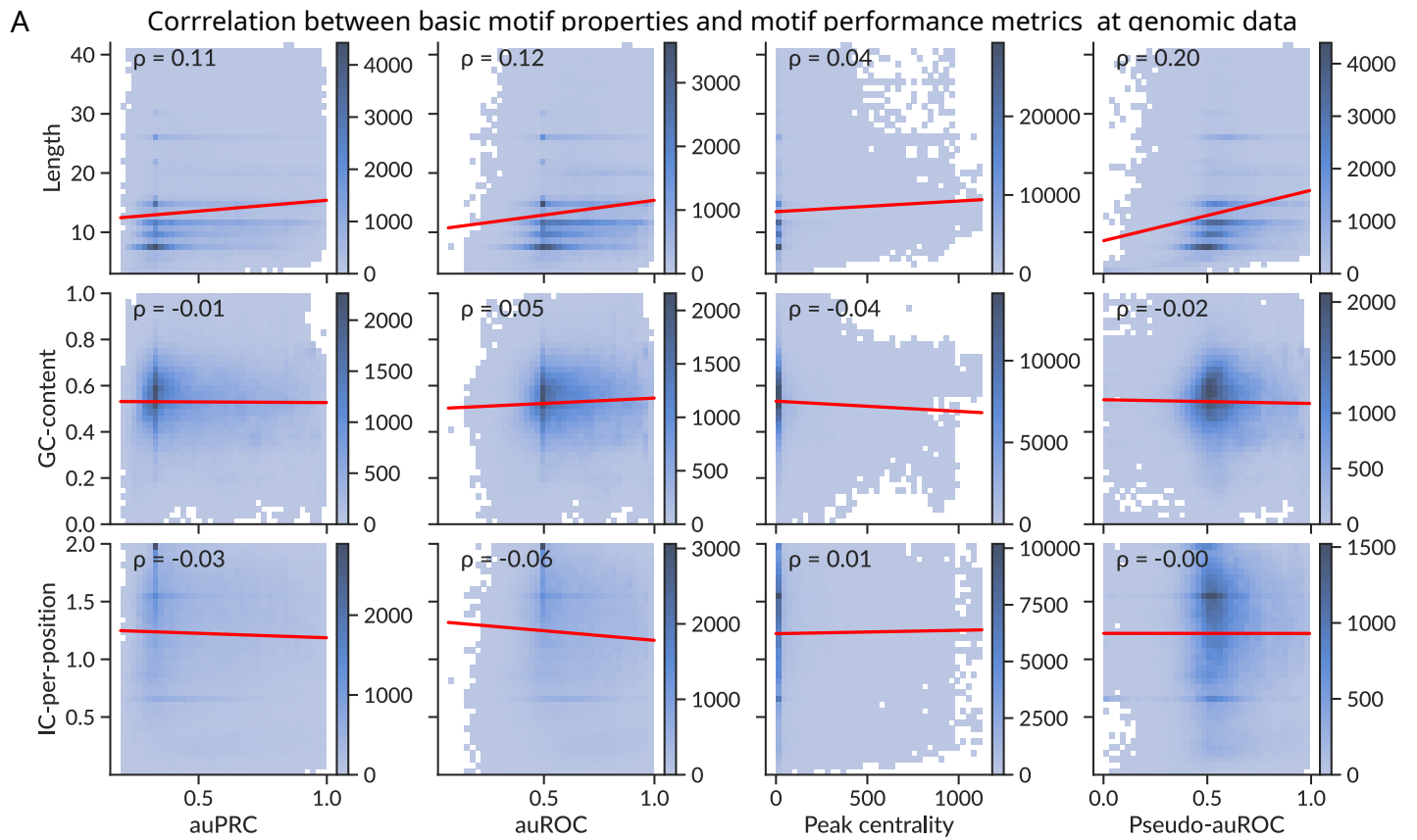
Supplementary Figure 4. A. The fraction of cases where the 1st, the 2nd, or the 3rd motif reported by a single run of a motif discovery tool are ranked the highest of these three in the overall benchmarking. **B:** Performance achieved by the overall top-ranking motifs produced by individual tools across TFs: distributions of pseudo-auROC values for ChIP-Seq and GHT-SELEX. **C:** Distributions of auROC values for ChIP-Seq and GHT-SELEX for zinc-finger TFs only. **D:** Scatterplots of auROC achieved at GHT-SELEX and ChIP-Seq for zinc-finger TFs only. * $p < 0.05$, paired Wilcoxon test against the best of motifs from all listed tools.



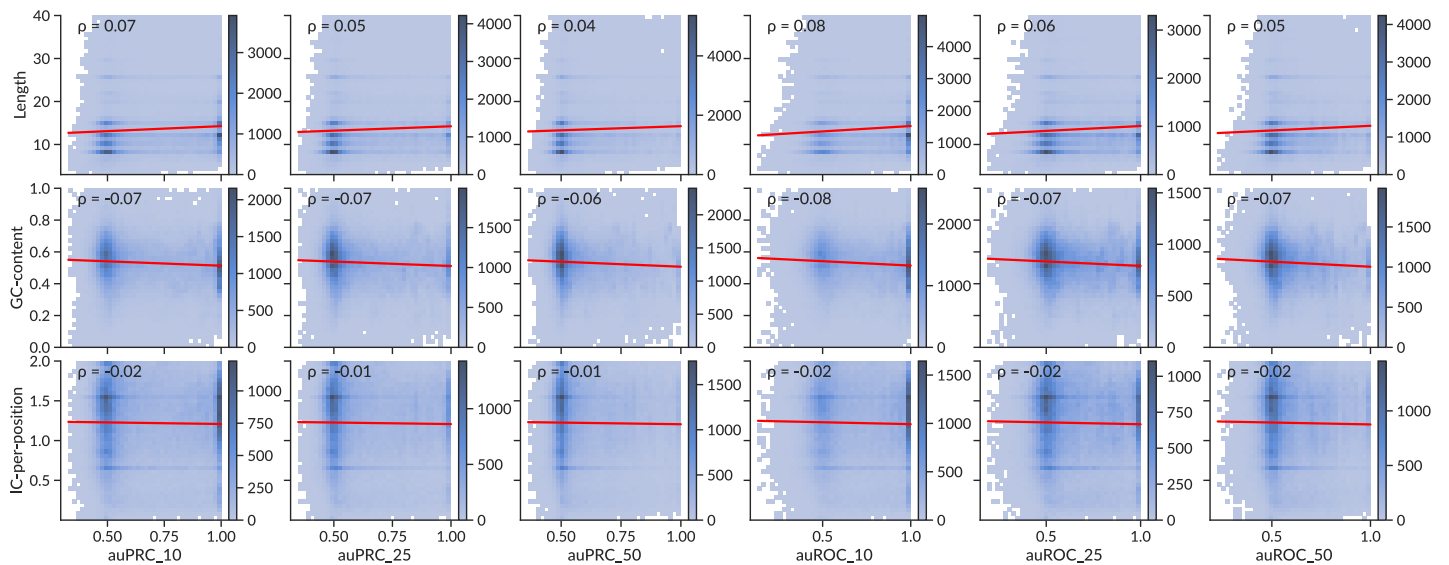
Supplementary Figure 5. Correlation between overall performance of motifs at the data of the same type as used for motif derivation (intra-platform) and at the test data from other experiment types (cross-platform).



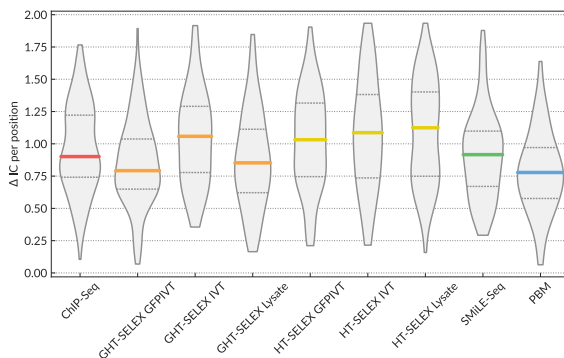
Supplementary Figure 6. Distributions of basic motif properties (top to bottom: length, GC%, information content).
A: Different motif discovery tools. **B:** Different experiment types.



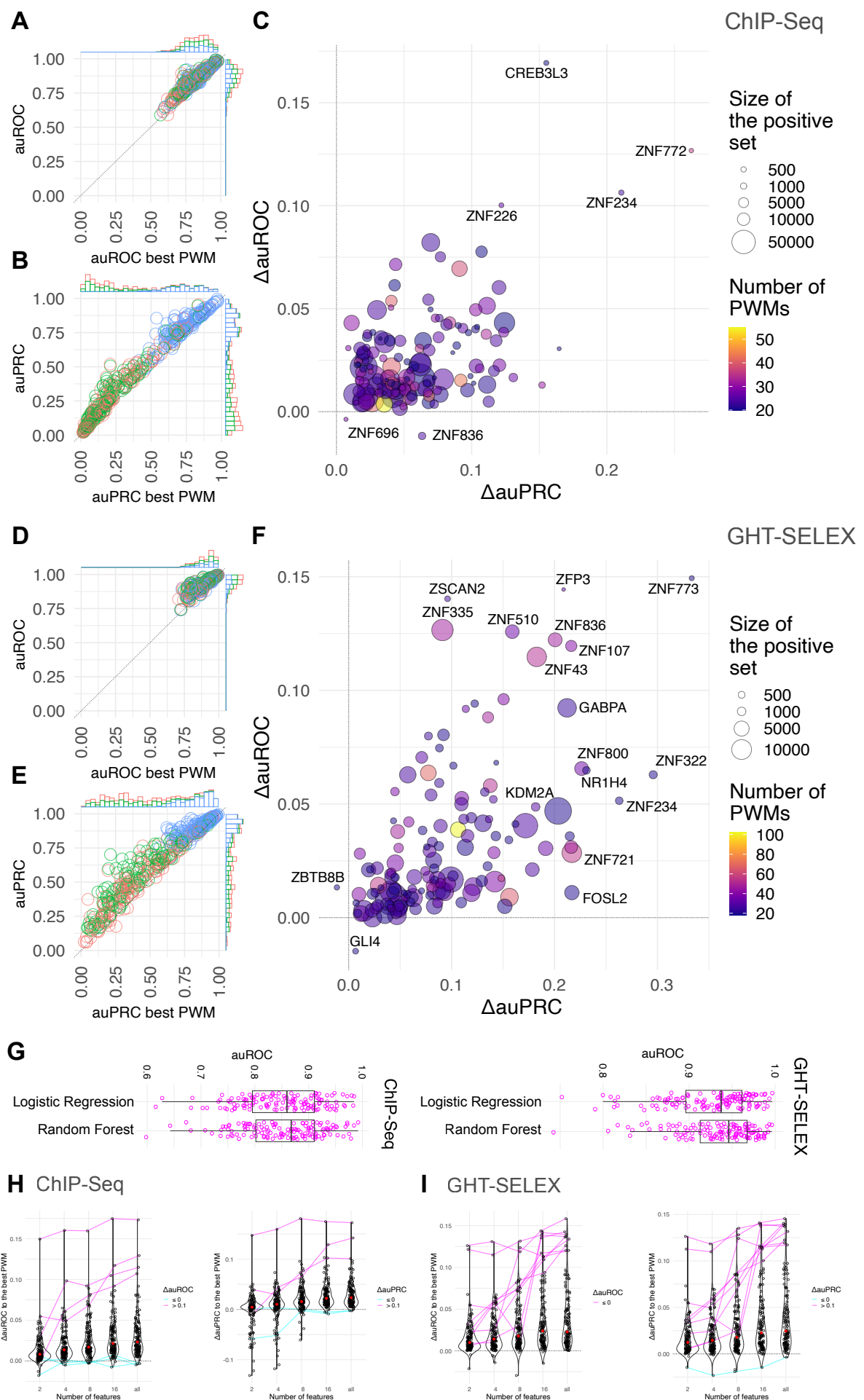
B Correlation between basic motif properties and motif performance metrics at HT-SELEX and SMiLE-Seq data



C Maximal difference in IC found for top 10 motifs per TF



Supplementary Figure 7. A: Correlations between individual performance metrics and basic motif properties (length, information content, GC%) for ChIP-Seq and GHT-SELEX test data. The color scale of the binned density plots denotes the numbers of motifs. **B:** Same as (A) for benchmarks with artificial sequences from HT-SELEX and SMiLE-Seq. **C:** A violin plot across TFs illustrating the maximal difference in the information content for the top 10 motifs for each TF. Individual experiment types are shown separately.



Supplementary Figure 8. Performance of the Archipelago model when trained and tested on the same type of experimental data. **A-C:** Absolute performance and relative gain of the Archipelago model compared to the best PWMs, ChIP-Seq. **D-F:** Same as **A-C** but for GHT-SELEX. Estimates obtained with alternative negative sets (random, aliens, and shades) are depicted by colored circles (green, red, and blue, respectively). **G:** Performance of the Random Forest and Logistic regression achieved when trained and tested on ChIP-Seq or GHT-SELEX data. **H, I:** Performance gain (purple) or loss (blue) depending on the number of PWMs in the model, ChIP-Seq and GHT-SELEX data.

CREB3L3

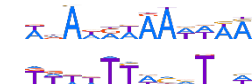
ChIP-Seq		GHT-SELEX	
1		1	
2		2	
3		3	
4		4	
5		5	

FOSL2

ChIP-Seq		GHT-SELEX	
1		1	
2		2	
3		3	
4		4	
5		5	

Supplementary Figure 9. Top 5 motifs of the Archipelago models sorted by feature importance. Each row lists the motif in both possible orientations.

ZNF772

ChIP-Seq		GHT-SELEX	
1		1	
2		2	
3		3	
4		4	
5		5	

ZNF773

ChIP-Seq		GHT-SELEX	
1		1	
2		2	
3		3	
4		4	
5		5	

Supplementary Figure 9, cont.

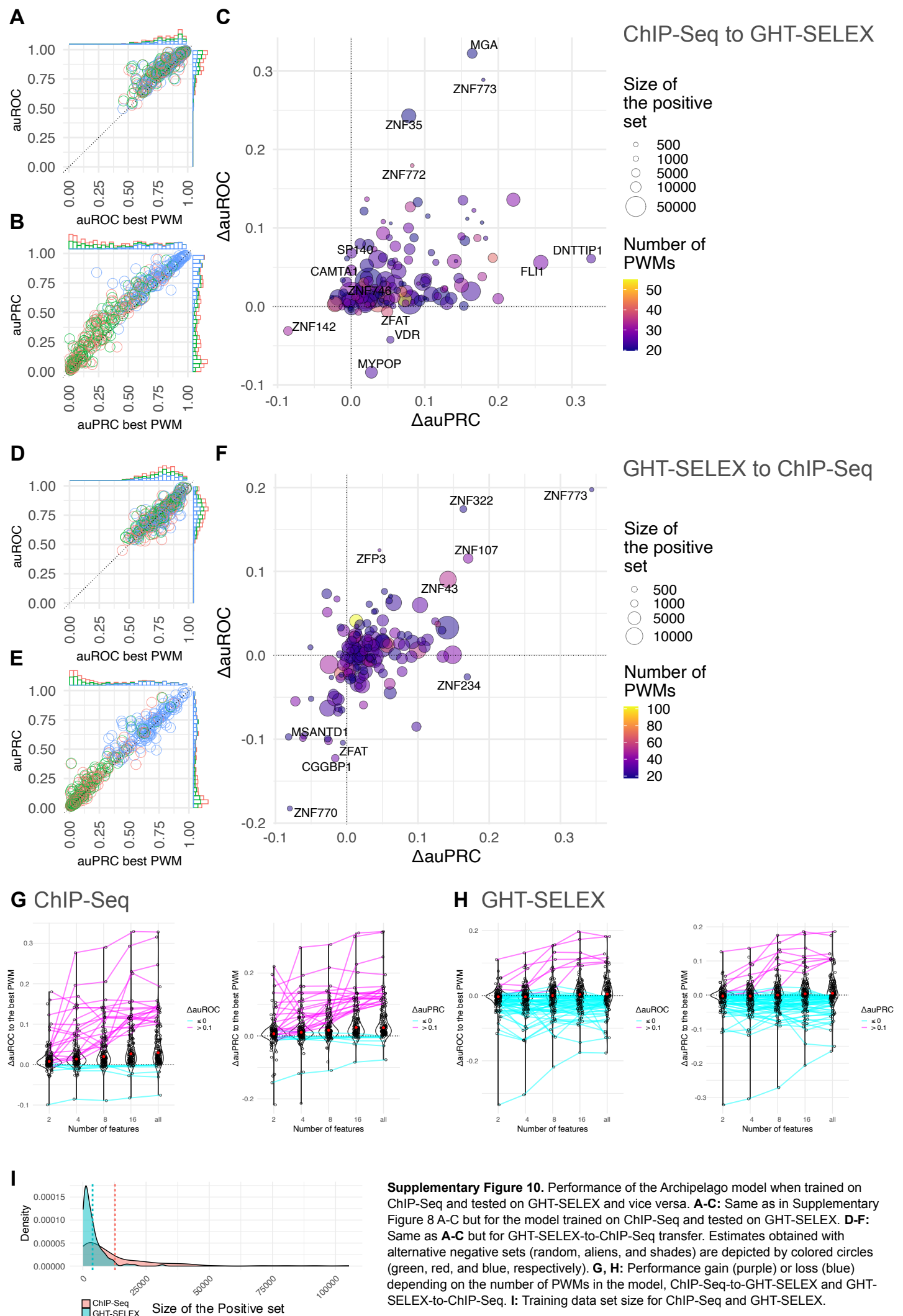
ZNF770

ChIP-Seq		GHT-SELEX	
1		1	
2		2	
3		3	
4		4	
5		5	

ZNF43

ChIP-Seq		GHT-SELEX	
1		1	
2		2	
3		3	
4		4	
5		5	

Supplementary Figure 9, cont.



Codebook/GRECO-BIT Motif Explorer Website Overview

 GRECO-BIT / Codebook Motif Explorer Approved data Help & Downloads

Codebook / GRECO-BIT: Motif Explorer

The Codebook Motif Explorer (MEX) is a detailed online catalog of DNA motifs built from a huge collection of experimental data produced by the Codebook Consortium to annotate the binding specificities of human transcription factors. Motif discovery was performed with multiple motif discovery tools, and the resulting motifs were rigorously benchmarked against held-out test data. On this website, you can find an interactive table browser listing the motifs, their performance metrics and ranks, as well as the metadata covering the underlying experimental datasets. Please check the accompanying manuscript and [Help & Downloads](#) for further details.

TF name

DNA-binding domain

Last page: AC092835 | motifs

Show recent pages

TF name	Total motifs	Total exp-s	DNA-binding domain	Best motif logo	ChIP-Seq exp-s	PBM exp-s	GHT-SELEX IVT exp-s	GHT-SELEX GFPVIT exp-s	GHT-SELEX Lysate exp-s	HT-SELEX IVT exp-s	HT-SELEX GFPVIT exp-s	HT-SELEX Lysate exp-s	SMILE-Seq exp-s
AC092835	288	2	C2H2 ZF		2								
AHCTF1	130	2	AT hook			2							
ATMIN	421	2	C2H2 ZF		2								
BATF2	669	4	bZIP		2				1			1	
BHLHA9	236	2	bHLH		2								
CAMTA1	1499	8	CG-1		1			1	2		1	2	1
CAMTA2	604	4	CG-1		1				1			1	1
CASZ1	1612	10	C2H2 ZF				1	1	2	1	1	2	2
CGGBP1	919	7	unknown		1	2		1	1			1	1
CPXCR1	254	2	C2H2 ZF		2								

Items per page: 10 1 - 10 of 236 |< < > >|

Approved data

Complete data

Help & Downloads

Motif sets

Metadata

MEX datasets

Benchmarking performance metrics

AC092835 | motifs

AC092835 | exp-s

MYF6 | motifs

RORB | motifs

LEUTX | motifs

C2H2 ZF

AT hook

bZIP

bHLH

CG-1

- 1 Switch between **Approved** and **Complete** sets of motifs and experiments
- 2 Link to the brief **Help** and **Downloads** page
- 3 Filter the list of TFs and representative motifs by **TF name** or **DNA-binding domain**
- 5 Quick **Page navigation** menu
- 6 Sort the table by clicking on **column headers** (such as the number of experiments a particular type)
- 7 Link to the page listing the **top-ranked motifs** of a particular **TF**
- 8 Link to the page with the **experiment-level metadata**
- 9 Table **pagination** settings

Supplementary Figure 11.

An overview of features of the Codebook Motif Explorer website (<https://mex.autosome.org>).

CREB3L3 (bZIP) motifs Columns Combined Rank (+6 of...) Reset columns

Prev: CPXCR1 (C2H2 ZF) Next: CTCF (C2H2 ZF)

Ranks Metrics

		ChIP-Seq		HT-SELEX IVT		SMILE-Seq	
		THC_0565		YWC_B_GA40NGGTCAT		UT380-042	
Combined Rank	ID	Name	Train experiment ID	Rank	Rank	Rank	
1	CREB3L3.FL@AFS.Lys@beady-lemon-olm@Halle.Dimont@Motif_1_imw20_asttrained	Motif_1_imw20_asttrained	YWO_B_AffSeq_F2_CREB3L3	1	8	94	
2	CREB3L3.FL@AFS.Lys@ugly-saffron-wallaby+breezy-taupe-rhino+ceros+beady-lemon-olm@Halle.Dimont@Motif_1_imw20_asttrained	Motif_1_imw20_asttrained	YWO_B_AffSeq_F2_CREB3L3	11	3	86	
3	CREB3L3.FL@AFS.Lys@breezy-taupe-rhinoceros@Halle.Dimont@Motif_1_imw20_asttrained	Motif_1_imw20_asttrained	YWO_B_AffSeq_F2_CREB3L3	10	16	97	
4	CREB3L3.FL@AFS.Lys@gimpy-dandelion-robin@autosome-ru.ChIPMunk@426seq_7to15_m0	426seq_7to15_m0	YWN_B_AffSeq_F2_CREB3L3	17	13	7	
5	CREB3L3.FL@AFS.Lys@gimpy-dandelion-robin@autosome-ru.ChIPMunk@425seq_15to7_m0	425seq_15to7_m0	YWN_B_AffSeq_F2_CREB3L3	14	17	6	

Items per page: 5 1 - 5 of 500

Heatmap for CREB3L3, Kendall's τ_b between motif ranks

Show in absolute scale

Heatmap visualization showing Kendall's τ_b correlation between motif ranks for CREB3L3 across various experiments.

CREB3L3 experiments

- GHT-SELEX GFPVIT: 1 (1 approved) sample
- GHT-SELEX Lysate: 2 (2 approved) samples
 - YWO_B_AffSeq_F2_CREB3L3
 - YWN_B_AffSeq_F2_CREB3L3
- ChIP-Seq: 1 (1 approved) sample
 - THC_0565
- PBM: 4 (2 approved) samples
 - HT-SELEX GFPVIT: 1 (1 approved) sample
 - HT-SELEX IVT: 2 (1 approved) samples
 - HT-SELEX Lysate: 2 (2 approved) samples
 - SMILE-Seq: 1 (1 approved) sample

Annotations:

- Heatmap of the pairwise Kendall rank correlation between experiments computed from motif ranks
- Show and hide individual columns
- Switch between motif ranks and raw performance metrics
- Open pop-up window with the experiment-level metadata

CREB3L3 (bZIP) motifs Columns Train data type (+14 of...) Reset columns

Prev: CPXCR1 (C2H2 ZF) Next: CTCF (C2H2 ZF)

Ranks Metrics

		ChIP-Seq		HT-SELEX IVT		SMILE-Seq								
		THC_0565		YWC_B_GA40NGGTCAT		UT380-042								
Algorithm	Combined Rank	HOCOMOCO similarity	Name	Train data type	Motif	auPRC ₁	auROC ₁	Rank	auROC ₂₅	auPRC ₁₀	auROC ₅₀	auROC ₁₀	Rank	auROC ₂₅
Dimont	1	0.11	ATF3	Motif_1_imw20_asttrained	GHT-SELEX Lysate	0.702	0.785	8	0.692	0.745	0.627	0.782	94	0.598
Dimont	2	0.13	ATF6A	Motif_1_imw20_asttrained	GHT-SELEX Lysate	0.696	0.779	3	0.693	0.744	0.627	0.782	86	0.599
Dimont	3	0.11	ATF6A	Motif_1_imw20_asttrained	GHT-SELEX Lysate	0.690	0.771	16	0.688	0.743	0.623	0.780	97	0.597
ChIPMunk	4	0.14	ATF6A	426seq_7to15_m0	GHT-SELEX Lysate	0.692	0.753	13	0.682	0.748	0.618	0.781	7	0.610
ChIPMunk	5	0.14	ATF6A	425seq_15to7_m0	GHT-SELEX Lysate	0.692	0.754	17	0.682	0.748	0.618	0.781	6	0.611

Items per page: 5 1 - 5 of 500

YWC_B_GA40NGGTCAT

Metadata of the experiment

Gene name: CREB3L3
Well on plate: E2
Insert ID: CREB3L3.DBD
flank 5: ACGACGCTCTTCCGATCTGA

DNA library ID: GA40NGGTCAT_v1
Plasmid ID: pTH13945
cycle: 3
flank 3: GGTCATAGATCGGAAGAGCA

Test datasets
crappy-cream-setter+hazy-cyan-snail+zippy-gold-uakari

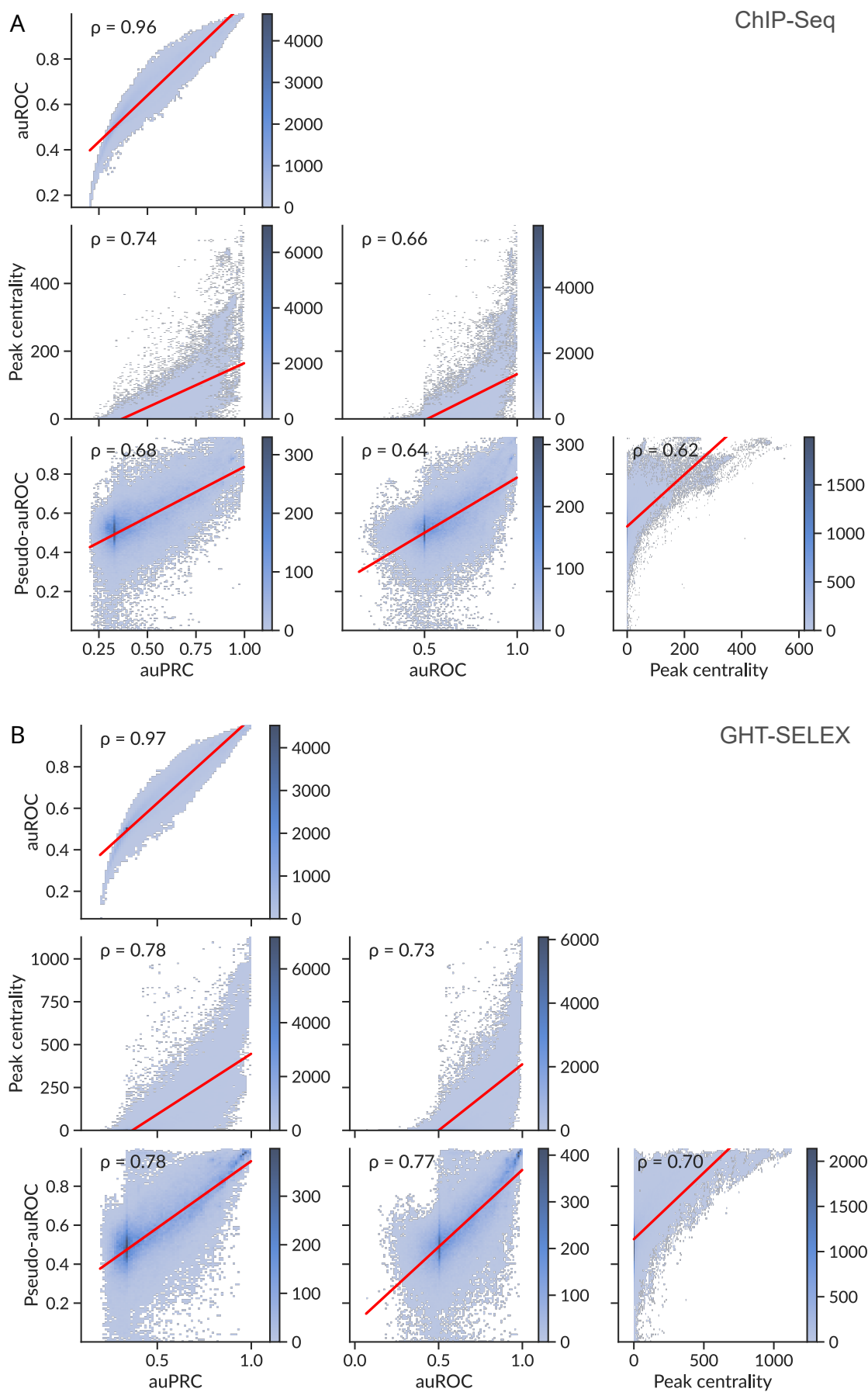
Construction type: DBD
Processing info: Reads

Annotations:

- Download individual motifs
- Show and hide individual columns
- Experiment metadata

Supplementary Figure 11, cont.

An overview of features of the Codebook Motif Explorer website (<https://mex.autosome.org>).



Supplementary Figure 12. Correlations of different benchmarking metrics considering all pairs of motifs and datasets. Linear regression trend (red line) and Pearson's rho are shown on the plots. **A:** ChIP-Seq. **B:** GHT-SELEX. The color scale of the binned density plots denotes the number of motifs.