Supplementary material for: TF-Prioritizer: a java pipeline to prioritize condition-specific transcription factors

Markus Hoffmann[1,2,*,**], Nico Trummer[1,*], Jakub Jankowski[3], Hye Kyung Lee[3], Lina-Liv Willruth[1], Olga Lazareva[4,5,6], Kevin Yuan[7], Nina Baumgarten[8,9,10], Florian Schmidt[11], Jan Baumbach[12,13], Marcel H. Schulz[8,9,10], David B. Blumenthal[14], Lothar Hennighausen[2,3,†], and Markus List[1,†]

[1]Big Data in BioMedicine Group, Chair of Experimental Bioinformatics, TUM School of Life Sciences, Technical University of Munich, Freising, Germany

[2]Institute for Advanced Study (Lichtenbergstrasse 2 a, D-85748 Garching, Germany), Technical University of Munich, Germany

[3]National Institute of Diabetes, Digestive, and Kidney Diseases, Bethesda, MD 20892, United States of America

[4]Division of Computational Genomics and Systems Genetics, German Cancer Research Center (DKFZ), Heidelberg, Germany

[5]Junior Clinical Cooperation Unit Multiparametric methods for early detection of prostate cancer, German Cancer Research Center (DKFZ), Heidelberg, Germany

[6]European Molecular Biology Laboratory, Genome Biology Unit, Heidelberg, Germany

[7]Big Data Institute, Nuffield Department of Population Health, University of Oxford, Oxford, United Kingdom

[8]Institute for Cardiovascular Regeneration, Goethe University, 60596 Frankfurt am Main, Germany

[9]German Center for Cardiovascular Research, Partner site Rhein-Main, 60590 Frankfurt am Main, Germany

[10]Cardio-Pulmonary Institute, Goethe University Hospital, 60596 Frankfurt am Main, Germany

[11]Laboratory of Systems Biology and Data Analytics, Genome Institute of Singapore, 60 Biopolis Street, Singapore, 138672, Singapore

[12]Chair of Computational Systems Biology, University of Hamburg, Hamburg, Germany

[13]Computational BioMedicine Lab, University of Southern Denmark, Odense, Denmark

[14]Department Artificial Intelligence in Biomedical Engineering (AIBE), Friedrich-Alexander University Erlangen-N̈uremberg (FAU), Erlangen, Germany

[*]The authors wish to be known that in their opinion the first two authors should be considered as shared first authors

[†] The authors wish to be known that in their opinion the last two authors should be considered as shared last authors

** *contact:* markus.daniel.hoffmann@tum.de; markus.list@tum.de

ABSTRACT

Background

Eukaryotic gene expression is controlled by cis-regulatory elements (CREs) including promoters and enhancers which are bound by transcription factors (TFs). Differential expression of TFs and their putative binding sites on CREs cause tissue and developmental-specific transcriptional activity. Consolidating genomic data sets can offer further insights into the accessibility of CREs, TF activity, and thus gene regulation. However, the integration and analysis of multi-modal data sets are hampered by considerable technical challenges. While methods for highlighting differential TF activity from combined ChIP-seq and RNA-seq data exist, they do not offer good usability, have limited support for large-scale data processing, and provide only minimal functionality for visual result interpretation.

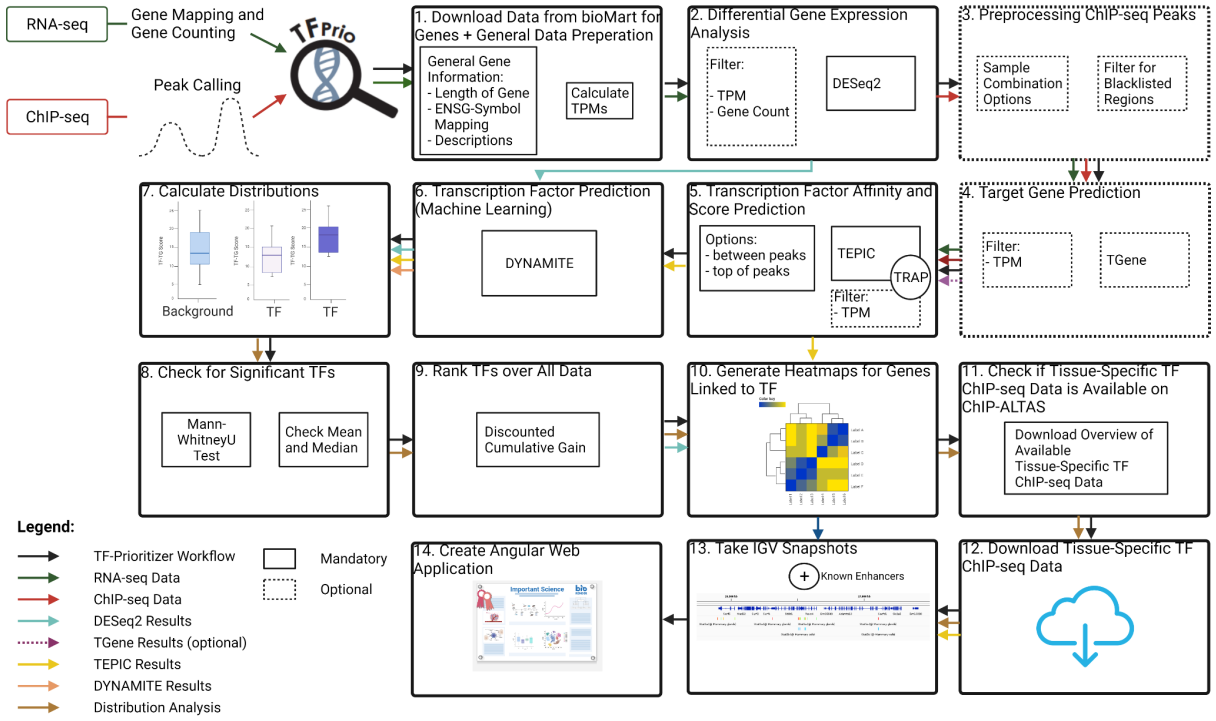
Results

We developed TF-Prioritizer, an automated java pipeline to prioritize condition-specific TFs derived from multi-modal data. TF-Prioritizer creates an interactive, feature-rich, and user-friendly web report of its results. To showcase the potential of TF-Prioritizer, we identified known active TFs (e.g., *Stat5, Elf5, Nfib, Esr1*), their target genes (e.g., milk proteins and cell-cycle genes), and newly classified lactating mammary gland TFs (e.g., *Creb1, Arnt*).

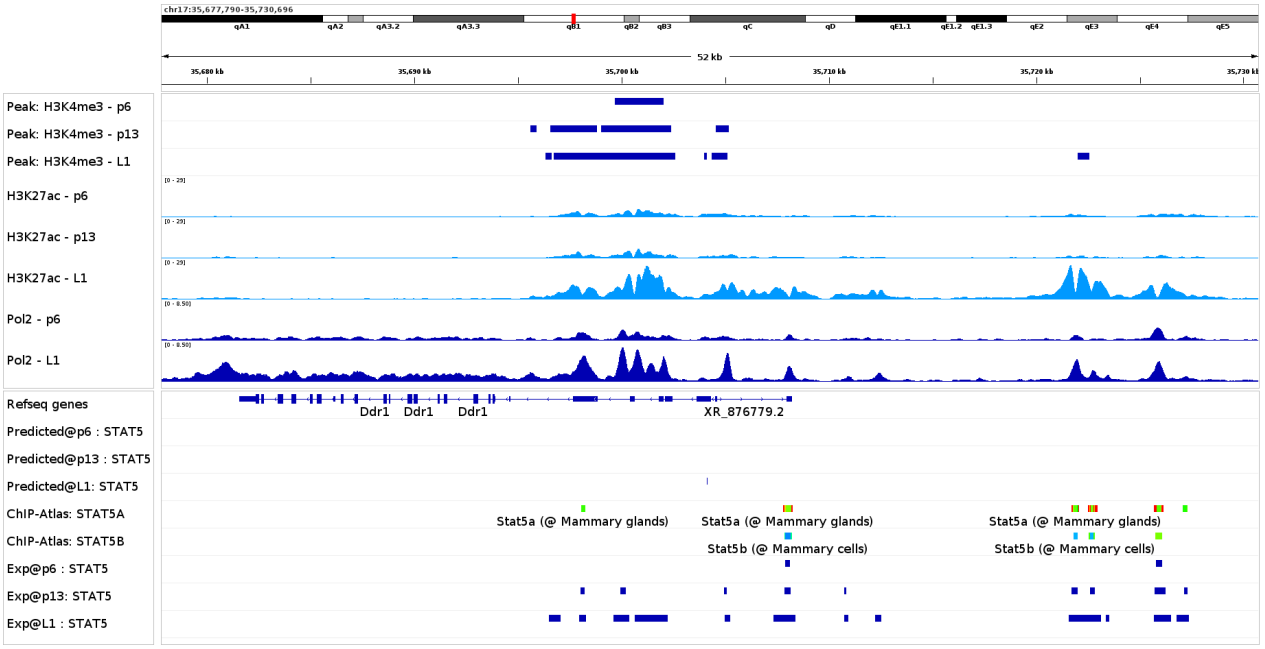
Conclusion

TF-Prioritizer accepts ChIP-seq and RNA-seq data, as input and suggests TFs with differential activity, thus offering an understanding of genome-wide gene regulation, potential pathogenesis, and therapeutic targets in biomedical research.

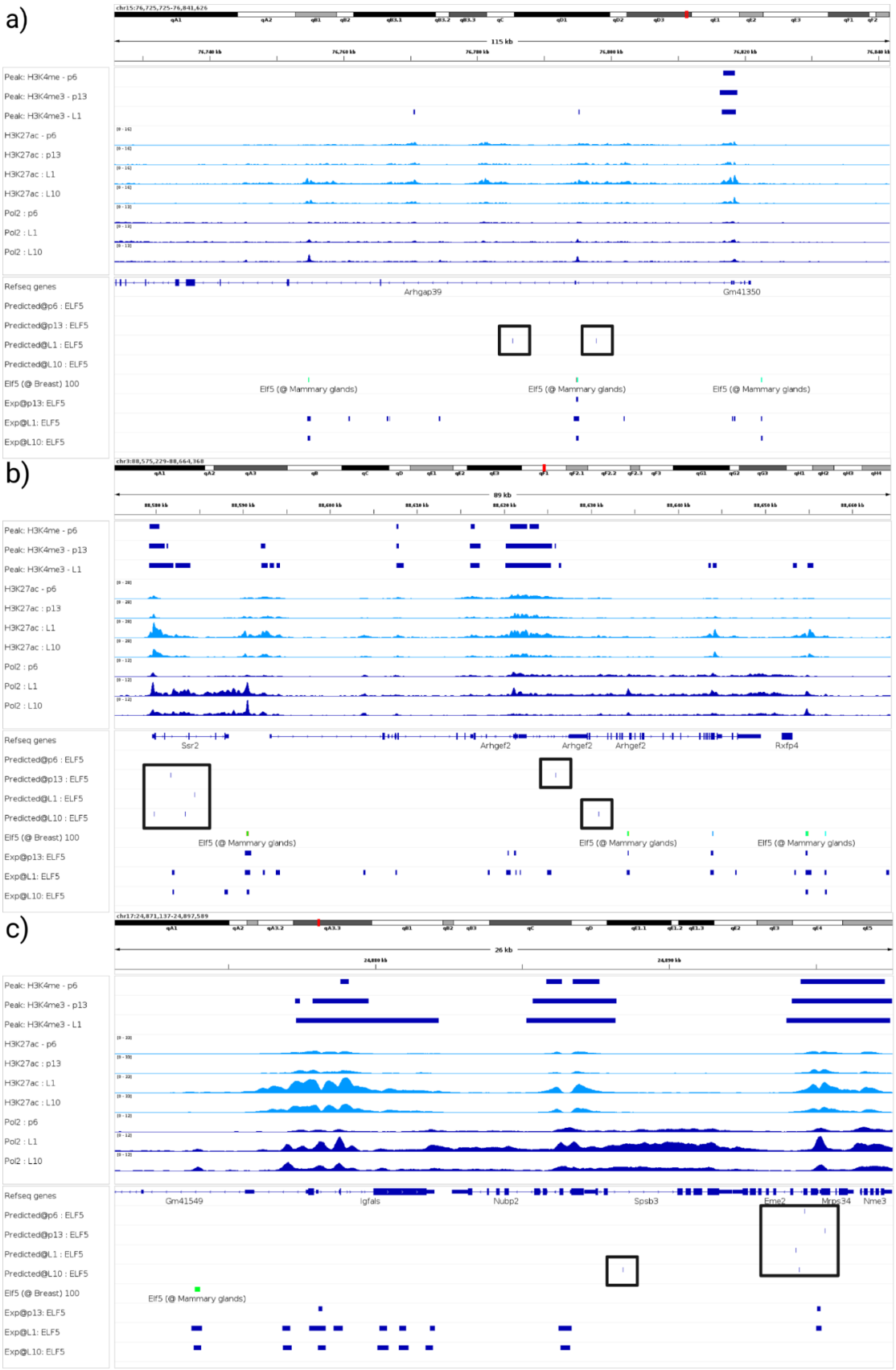
Supplementary Figure 1: Technical Workflow



Supplementary Figure 2: IGV screenshot of STAT5 target gene DDR1



Supplementary Figure 3: IGV screenshot of ELF5 target genes ARHGAP39, ARHGEF2, and IGFALS



Supplementary Figure 4: Confusion matrices and statistical measures

a) STAT5A..STAT5B

		Experimental	
		True	False
Predicted	True	54,805	15,319
	False	39,866	30,258

Sensitivity	57.89%	Specificity	66.39%
Precision	78.15%	Accuracy	60.65%
F1-Score	66.51%		

b) ELF5

		Experimental	
		True	False
Predicted	True	29,864	6,737
	False	8,636	27,965

Sensitivity	77.57%	Specificity	80.59%
Precision	81.59%	Accuracy	79.00%
F1-Score	79.53%		

c) ESR1

		Experimental	
		True	False
Predicted	True	9,214	1,003
	False	6,459	3,758

Sensitivity	58.79%	Specificity	78.93%
Precision	90.18%	Accuracy	63.48%
F1-Score	71.18%		

d) NFIB

		Experimental	
		True	False
Predicted	True	104,007	13,330
	False	30,954	86,383

Sensitivity	77.06%	Specificity	86.63%
Precision	88.64%	Accuracy	81.13%
F1-Score	82.45%		

e) CREB1

		Experimental	
		True	False
Predicted	True	45,263	3,699
	False	9,895	39,067

Sensitivity	82.06%	Specificity	91.35%
Precision	92.45%	Accuracy	86.12%
F1-Score	86.94%		

f) ARNT

		Experimental	
		True	False
Predicted	True	17,083	24,403
	False	3,811	37,675

Sensitivity	81.76%	Specificity	60.69%
Precision	41.18%	Accuracy	66.00%
F1-Score	54.77%		

e) ARNT..HIF1A

		Experimental	
		True	False
Predicted	True	59,919	83,115
	False	45,076	97,958

Sensitivity	57.07%	Specificity	54.10%
Precision	41.89%	Accuracy	55.19%
F1-Score	48.32%		

g) AHR..ARNT

		Experimental	
		True	False
Predicted	True	65,503	94,477
	False	38,243	121,737

Sensitivity	63.14%	Specificity	56.30%
Precision	40.94%	Accuracy	58.52%
F1-Score	49.68%		

Supplementary Figure 5: Co-occurrence analysis

Co-Occurrence Analysis

	ARNT	CREB1	ELF5	ESR1	KLF4	MYC	NFIB	SNAI2	STAT1	STAT5A	STAT5B	TFAP2C
ARNT	-	0.08	0.03	0.04	0.01	0.02	0.05	0.02	-	0.03	0.03	0.05
CREB1	0.08	-	0.22	0.07	-	0.01	0.29	0.01	-	0.21	0.11	0.13
ELF5	0.03	0.22	-	0.03	-	-	0.23	0.01	-	0.24	0.14	0.03
ESR1	0.04	0.07	0.03	-	-	-	0.05	0.01	-	0.03	0.01	0.06
KLF4	0.01	-	-	-	-	0.10	-	0.01	0.09	-	0.01	-
MYC	0.02	0.01	-	-	0.10	-	0.01	0.02	0.08	-	0.01	0.01
NFIB	0.05	0.29	0.23	0.05	-	0.01	-	0.01	-	0.21	0.10	0.07
SNAI2	0.02	0.01	0.01	0.01	0.01	0.02	0.01	-	0.01	0.01	0.03	0.02
STAT1	-	-	-	-	0.09	0.08	-	0.01	-	-	0.01	-
STAT5A	0.03	0.21	0.24	0.03	-	-	0.21	0.01	-	-	0.19	0.03
STAT5B	0.03	0.11	0.14	0.01	0.01	0.01	0.10	0.03	0.01	0.19	-	0.01
TFAP2C	0.05	0.13	0.03	0.06	-	0.01	0.07	0.02	-	0.03	0.01	-

Supplementary Material 1: Confusion matrices and the calculation of sensitivity, specificity, precision, accuracy, and F1-score

a) Confusion Matrix

		Experimental	
		True	False
Predicted	True	TP	FP
	False	FN	TN

TP = True Positives
FP = False Positives
TN = True Negatives
FN = False Negatives

b) Metrics

Sensitivity = $\frac{TP}{TP+FN}$

Specificity = $\frac{TN}{TN+FP}$

Precision = $\frac{TP}{TP+FP}$

Accuracy = $\frac{TP+TN}{TP+FP+TN+FN}$

F1-Score = $\frac{2 \times \text{Sensitivity} \times \text{Precision}}{\text{Sensitivity} + \text{Precision}}$

Supplementary Table 1: Feature comparison between TF-Prioritizer and diffTF

Feature	TF-Prioritizer	diffTF
Filter for chromosomes (e.g., sex, mitochondrial chromosomes)	✗	✓ mandatory, cannot be skipped
Classifies TFs into activators and repressors	✗	✓
Sample combination option to reduce runtime	✓	✗
Filter for blacklisted genomic regions	✓	✓
TPM filter for TFs	✓	✗
TPM filter for target genes	✓	✗
Biophysical Model for TF binding sites	✓ TRAP	✗ Use of all binding sites
Probability scores for TF target gene links	✓ TEPIC	✗
Prioritize TFs based on binding site energies,TF expression and target genes expression	✓ Linear Regression (DYNAMITE)	✗
Additional filtering layer for prioritized TFs	✓ Background distribution model	✓ Background distribution model
Searches automatically for outside experimental data	✓ ChIP-Atlas	✗
Creates genome browser screenshot on loci of interest using experimental and predicted data	✓ IGV browser	✗
Provide insight into predictions (e.g., show peaks in IGV)	✓	✗
Creates heatmaps for predicted target genes that have the highest binding affinity for each TF	✓	✗
Creates a feature-rich interactive web application	✓	✗

Supplementary Table 2: Technical comparison between TF-Prioritizer and diffTF

Technical Features	TF-Prioritizer	diffTF
Provides a docker or docker-like version of the pipeline to prevent technical dependency issues	✔ dockerized	— singularity, with dependency problems
Pipeline can resume if unexpected shutdown happened (e.g., server crash, process was forced to quit)	✔	✘
Simulation of the whole process to check for validity of parameters and files	✔	✔
Multi-Threading available	✔	✔
pipeline checks available RAM and uses less threads if not enough RAM is available	✔	✘

Supplementary Table 3: Comparison of prioritized transcription factors between TF-Prioritizer and diffTF

TF	diffTF	TF-Prioritizer
AHR	(x)	(x)
AIRE	(x)	
ANDR	(x)	
AP2B	(x)	
AP2C	(x)	
ARI3A.D	(x)	
ARI3A.S	(x)	
ARI5B	(x)	
ARID3A		(x)
ARID5A		(x)
ARNT	**(x)**	**(x)**
ARNT2	(x)	
ATF1	(x)	
ATF2	(x)	
ATF3	(x)	
BACH1	(x)	
BARX2	(x)	
BATF	(x)	
BCL6	(x)	
BHE40	(x)	
BHLHE40		(x)
BMAL1	(x)	
BRCA1	(x)	
CBFB		(x)
CDC5L	(x)	
CEBPA	(x)	(x)
CEBPB	(x)	

CEBPD	(x)	(x)
CEBPG	(x)	
CEBPZ	(x)	
CLOCK		(x)
COE1	(x)	
COT1.B	(x)	
COT1.S	(x)	
COT2.B	(x)	
COT2.S	(x)	
CREB1	**(x)**	**(x)**
CREM	(x)	
CTCF	(x)	
CUX1	(x)	
CXXC1	(x)	
DBP	(x)	(x)
DLX3	(x)	
E2F1	(x)	
E2F2	(x)	(x)
E2F3	(x)	
E2F4		(x)
E2F5	(x)	
E2F6	(x)	(x)
E2F7	(x)	(x)
E4F1	(x)	
EGR1	(x)	(x)
EGR2	(x)	(x)
EGR3	(x)	
ELF1	(x)	
ELF3	(x)	(x)
ELF5	**(x)**	**(x)**

ELK1	(x)	
ELK3	(x)	
ELK4	(x)	(x)
ENOA	(x)	
EPAS1	(x)	(x)
ERG	(x)	
ERR2	(x)	
ESR1	**(x)**	**(x)**
ETS1	**(x)**	**(x)**
ETS2	(x)	(x)
ETV4	(x)	(x)
ETV5	(x)	
ETV6		(x)
EVI1	(x)	
FLI1	(x)	
FOS	(x)	
FOSB	(x)	
FOSL2	(x)	
FOXA1	(x)	
FOXA3	(x)	
FOXC1	(x)	
FOXI1	(x)	
FOXJ2	(x)	
FOXJ3.A	(x)	
FOXJ3.S	(x)	
FOXM1	(x)	
FOXO1	(x)	
FOXO3	(x)	(x)
FOXO4	(x)	(x)
FOXP2	(x)	

FOXP3	(x)	
FOXQ1	(x)	
FUBP1	(x)	
GABP1	(x)	
GABPA	(x)	(x)
GATA2	(x)	(x)
GATA3	(x)	
GATA6	(x)	
GCR.C	(x)	
GCR.S	(x)	
GFI1	(x)	
GLI1	(x)	
GLI2	(x)	
GLI3	(x)	
GLIS3	(x)	
GMEB1		(x)
HAND1		(x)
HBP1	(x)	
HES1	(x)	(x)
HEY2	(x)	
HIC1	(x)	(x)
HIF1A	(x)	(x)
HINFP	(x)	
HLF	(x)	(x)
HLTF	(x)	
HMGA1	(x)	
HNF1A	(x)	
HOXA5		(x)
HOXB4		(x)
HOXD9		(x)

HSF1	(x)	
HSF2	(x)	
HTF4	(x)	
HXA1	(x)	
HXA10	(x)	
HXA5	(x)	
HXA7	(x)	
HXB6	(x)	
HXB7	(x)	
HXB8	(x)	
HXC6	(x)	
HXC8	(x)	
HXD10	(x)	
HXD4	(x)	
HXD9	(x)	
IKZF1	(x)	
IRF1	(x)	
IRF2	(x)	
IRF3	(x)	
IRF7	(x)	(x)
IRF9	(x)	(x)
ITF2	(x)	
JUN	(x)	
JUNB	(x)	(x)
JUND	(x)	(x)
KAISO	(x)	
KLF1	(x)	
KLF15	(x)	(x)
KLF3		(x)
KLF4	(x)	(x)

KLF6	(x)	(x)
KLF8	(x)	
LEF1	(x)	
LYL1		(x)
MAF	(x)	
MAFB	(x)	(x)
MAFG	(x)	
MAFK.A	(x)	
MAFK.S	(x)	
MAZ	(x)	(x)
MBD2	(x)	
MCR	(x)	
MECP2	(x)	(x)
MEF2A	(x)	
MEF2C	(x)	
MEF2D	(x)	
MEIS1	(x)	
MEIS2	(x)	
MITF	(x)	
MSX2	(x)	
MTF1	(x)	
MXI1	(x)	(x)
MYB	(x)	(x)
MYBB	(x)	
MYC	(x)	(x)
MYCN	(x)	
NANOG.A	(x)	
NANOG.S	(x)	
NF2L1	(x)	
NF2L2	(x)	

NFAC1.A	(x)	
NFAC1.S	(x)	
NFAC3	(x)	
NFAC4	(x)	
NFAT5	(x)	
NFATC3		(x)
NFATC4		(x)
NFE2	(x)	
NFE2L1		(x)
NFIA.C	(x)	
NFIA.S	(x)	
NFIB		**(x)**
NFIL3	(x)	(x)
NFKB1	(x)	
NFKB2	(x)	(x)
NFYA.D	(x)	
NFYA.S	(x)	
NFYB	(x)	
NFYC	(x)	
NKX31	(x)	
NR1D1	(x)	(x)
NR1D2		(x)
NR1H2	(x)	(x)
NR1H3		(x)
NR1H4	(x)	
NR2C1	(x)	
NR2C2	(x)	
NR2E3	(x)	
NR2F2		(x)
NR2F6	(x)	

NR4A1		(x)
NR4A2		(x)
NR5A2	(x)	
NR6A1	(x)	
NRF1	(x)	
ONEC2	(x)	
OVOL1	(x)	(x)
P53	(x)	
P63	(x)	
P73	(x)	
PAX5.D	(x)	
PBX1	(x)	
PBX2	(x)	(x)
PBX3	(x)	
PEBB	(x)	
PIT1	(x)	
PKNX1	(x)	
PLAG1.D	(x)	
PLAG1.S	(x)	
PO2F1	(x)	
PO2F2	(x)	
PO6F1	(x)	
PPARA.C	(x)	
PPARD	(x)	
PPARG		(x)
PPARG.A	(x)	
PPARG.S	(x)	
PRDM1	(x)	
PRGR.C	(x)	
PRGR.S	(x)	

PRRX1	(x)	
PRRX2	(x)	(x)
PURA	(x)	
RARB	(x)	
RARG.C	(x)	
RARG.S	(x)	
RBPJ		(x)
REL	(x)	
RELB	(x)	(x)
REST	(x)	
RFX1	(x)	
RFX2	(x)	
RFX3	(x)	
RORA	(x)	(x)
RORG	(x)	
RREB1	(x)	
RUNX1	(x)	(x)
RUNX2	(x)	(x)
RUNX3		(x)
RXRA	(x)	
RXRB		(x)
RXRG	(x)	
SHOX2		(x)
SMAD1	(x)	
SMAD2	(x)	(x)
SMAD3		(x)
SMAD4	(x)	
SMRC1	(x)	
SNAI1	(x)	
SNAI2	(x)	(x)

SOX10	(x)	(x)
SOX13	(x)	
SOX17	(x)	
SOX18	(x)	
SOX4	(x)	(x)
SOX5	(x)	
SOX6		(x)
SOX9	(x)	
SP1.A	(x)	
SP1.S	(x)	
SP2	(x)	
SP3	(x)	(x)
SP4	(x)	
SPIB	(x)	
SRBP1	(x)	
SRBP2	(x)	
SREBF2		(x)
STAT5A	**(x)**	**(x)**
STA5TB	**(x)**	**(x)**
STAT1	(x)	(x)
STAT3	(x)	
STAT4	(x)	
STAT6	(x)	
SUH	(x)	
TAL1.A	(x)	
TAL1.S	(x)	
TBP	(x)	
TBX2	(x)	
TBX3	(x)	(x)
TCF3		(x)

TCF4		(x)
TCF7	(x)	
TEAD1	(x)	
TEAD3	(x)	
TEAD4	(x)	
TEF	(x)	
TF2L1	(x)	
TF7L2	(x)	
TFAP2A		(x)
TFAP2B		(x)
TFAP2C		(x)
TFCP2	(x)	
TFDP1	(x)	
TFE2.S	(x)	
TFE3	(x)	
TFEB	(x)	
TGIF1		(x)
TGIF1.S	(x)	
THA.C	(x)	
THA.S	(x)	
THB.C	(x)	
THB.S	(x)	
TWST1	(x)	
TYY1	(x)	
USF1	(x)	
USF2	(x)	
VDR		(x)
VDR.C	(x)	
VDR.S	(x)	
XBP1	(x)	

YBOX1	(x)	
ZBT18	(x)	
ZBT7A	(x)	
ZBTB17		(x)
ZBTB6	(x)	
ZBTB7A		(x)
ZEB1	(x)	(x)
ZEP1	(x)	
ZEP2	(x)	
ZFHX3	(x)	
ZFP335		(x)
ZFX	(x)	(x)
ZN148	(x)	
ZN423	(x)	

Supplementary Table 4: Comparison of prioritized transcription factors before and after the filtering of the background distribution

	Number of TFs before filter	Number of TFs after filter
TEPIC	203	92
LOG2FC and TEPIC	203	95
TEPIC and DYNAMITE	203	97
DYNAMITE	203	102
LOG2FC and DYNAMITE	203	104
TF-TG score	203	104
LOG2FC	203	119