Supplementary materials for: Integrated analysis of motif activity and gene expression changes of transcription factors

Jesper Grud Skat Madsen[1,4], Alexander Rauch[1,4], Elvira Laila Van Hauwaert[1], Søren Fisker Schmidt[1,2], Marc Winnefeld[3] and Susanne Mandrup[1,5]

1. Department of Biochemistry and Molecular Biology, University of Southern Denmark, Odense, Denmark.

2. Present address: Institute for Diabetes and Cancer, Helmholtz Center Munich, German Research Center for Environmental Health, Neuherberg, Germany.

3. Research and Development, Beiersdorf AG, Hamburg, Germany.

4. These authors contributed equally.

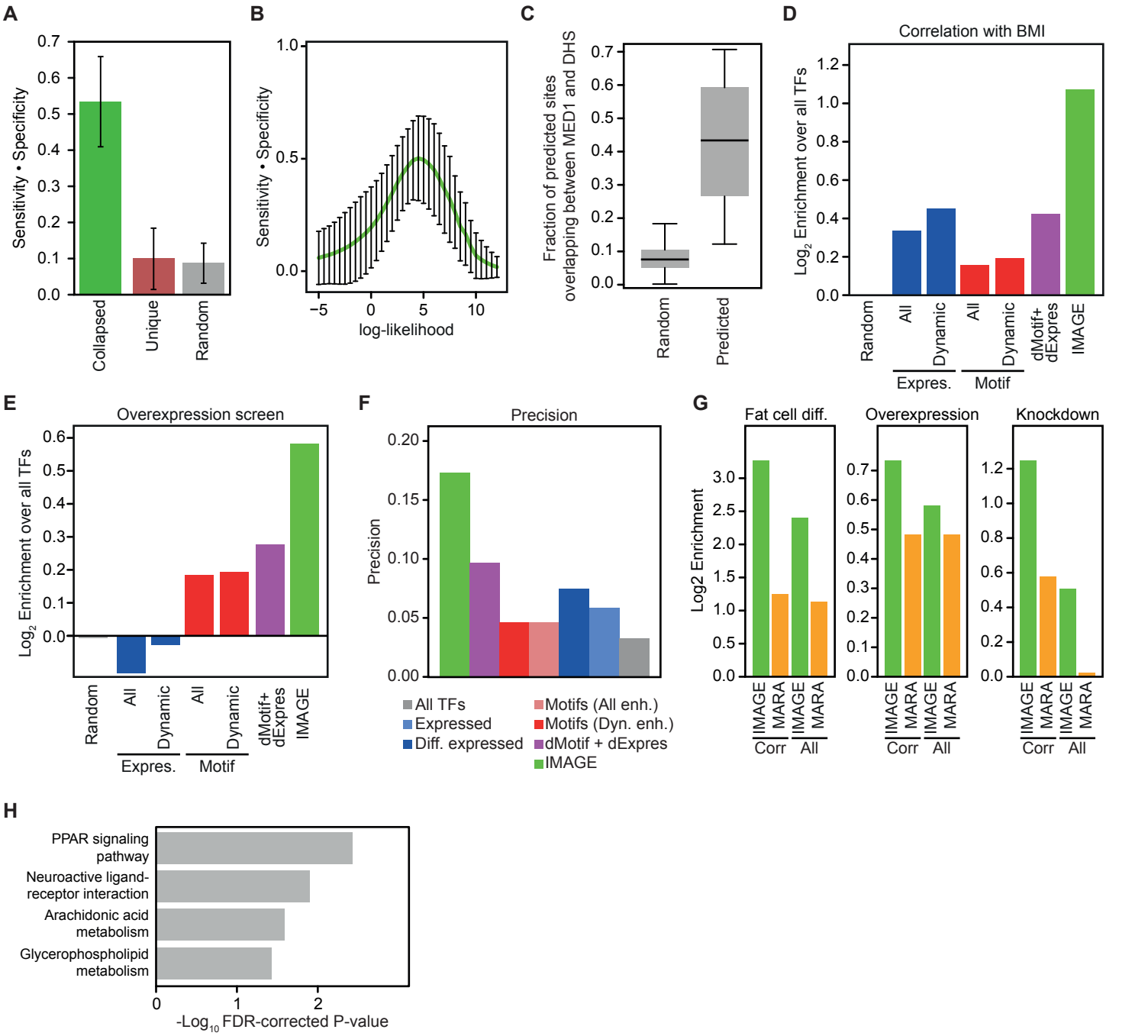
5. Corresponding author (s.mandrup@bmb.sdu.dk)

Table of Contents

Supplemental Figures **2-3**

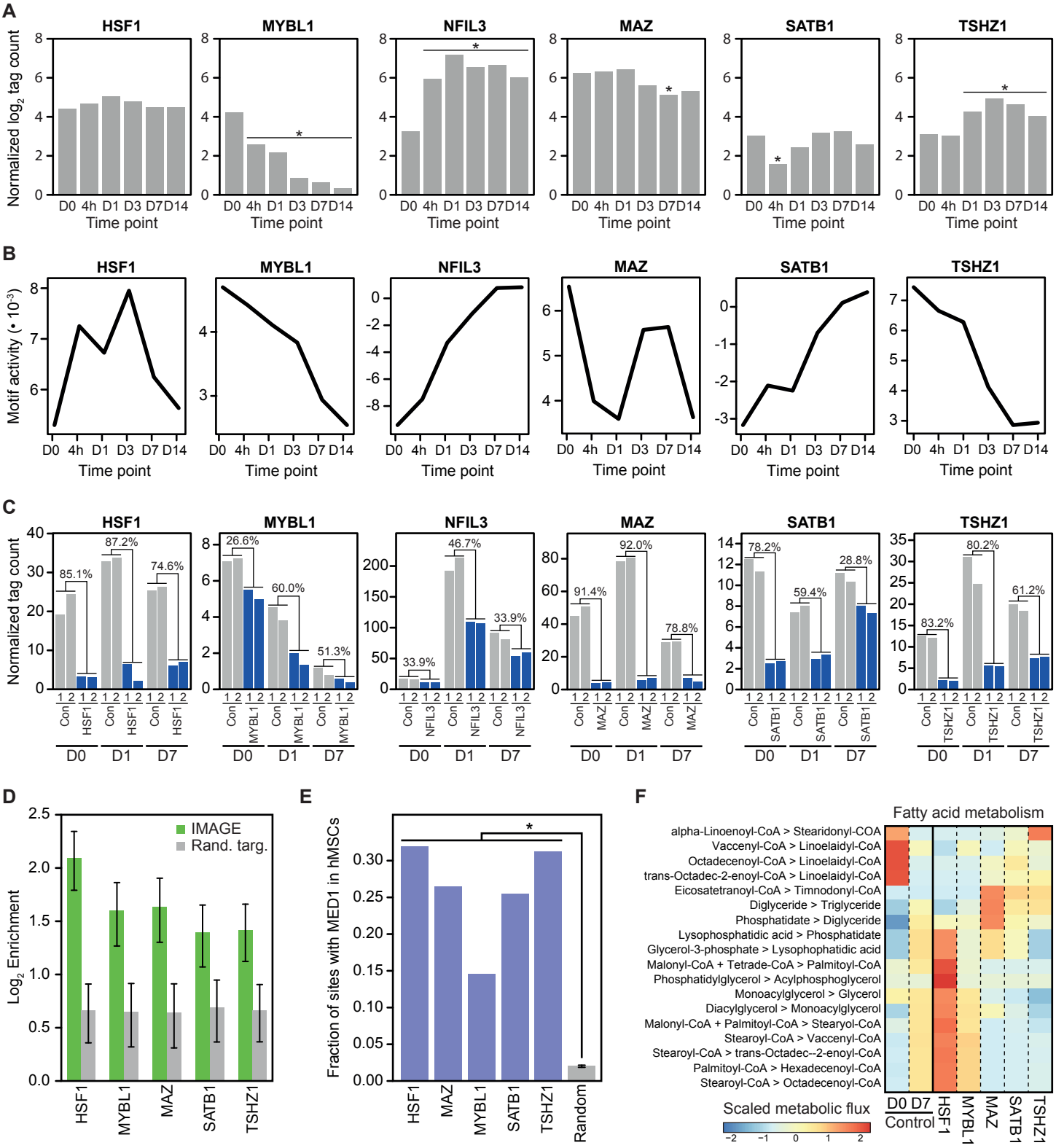
Supplemental Table Legend **4**

Supplemental Figures



Supplemental Fig S1: Related to Figure 2, 3 and 4

A) Validation that motifs can be collapsed, while retaining binding specificities. Binding sites among 5000 bound sites and 20,000 unbound sites were predicted using 104 motifs, as well as 14 collapsed motifs, assigned to a total of 14 transcription factors (as in Fig. 2A). Bars represent the mean of the sensitivity multiplied by the specificity from the predictions based on the collapsed motifs, from the predictions unique to each of the 104 motifs and by random selection of enhancers. Error bars represent the standard deviation. **B**) There is a local maximum in sensitivity multiplied by specificity using log-likelihood PWM scoring. The line shows averaged prediction statistics across all transcription factors for which ChIP-seq data was retrieved from ENCODE (as in Fig. 2A) at different log-likelihood score cut offs. Error bars indicate the standard deviation. **C**) IMAGE predicts similar sets of binding sites using either MED1 ChIP-seq or DNase-seq data in 3T3-L1 cells. Boxplot showing the fractional overlap of all predicted target enhancers for all motifs that are predicted with either MED1 ChIP-seq or DNase-seq data and the fractional overlap of randomized, size-matched groups of enhancers for all motifs. Target enhancers of all motifs were identified using IMAGE based on either MED1 ChIP-seq or DNase-seq. **D-E**) Comparison of the predictive power of IMAGE and various other methods for predicting transcriptional regulators of adipogenesis in 3T3-L1 preadipocytes. The bar plot shows the degree to which the indicated methods (same as in Fig. 4B and C) predicts causal transcription factors in 3T3-L1 adipocytes whose expression in white adipose tissue correlate with BMI ($P \leq 0.05$) (D), or with expression of genes that affect lipid accumulation upon overexpression in high density screens (Gubelmann et al. 2014) (E). The bar plot shows the predictive power of the indicated methods as determined by the enrichment of the predicted factors over all transcription factors. **F**) IMAGE has the highest precision for prediction of causal transcription factors in 3T3-L1 adipogenesis that belong to GO term 'Fat cell differentiation'. The bar plot indicates the precision of prediction using each group of candidates (predicted as in Fig. 4B). **G**) Comparison of predictive power of IMAGE and MARA. Gene expression data from 3T3-L1 used for IMAGE analysis were submitted to MARA analysis. Causal transcription factors were predicted as all transcription factors with a significant change in motif activity in MARA analysis ($Z \geq 2$) (All MARA), or all transcription factors with a significant change in motif activity in the MARA analysis ($Z \geq 2$) and strong positive correlation between motif activity and gene expression ($R \geq 0.8$) (Corr. MARA). Similarly, we defined either all candidates identified by IMAGE (All IMAGE) or those candidates with strong positive correlation between motif activity in IMAGE and gene expression ($R \geq 0.8$) (Corr. IMAGE). We calculated the enrichment of these four groups of transcription factors compared to all transcription factors within the 'fat cell differentiation' GO term and among genes that affect lipid accumulation upon overexpression (Gubelmann et al. 2014) or knockdown (Söhle et al. 2012). **H**) Predicted PPARG target genes are enriched for PPAR-related biological pathways. Bar plot showing the –log10 FDR-adjusted P-value of the three path-ways that are significantly enriched (Padj ≤ 0.05) for PPARG target genes predicted by IMAGE. Pathway analysis was performed using goseq (Young et al. 2010) and the KEGG database (Kanehisa et al. 2016).



Supplemental Fig S2: Related to Figure 5

IMAGE was used to determine motif activity based on MED1 ChIP-seq (n=2) and RNA-seq (n=2) during adipocyte differentiation of hMSC-TERT4 cells. **A**) HSF1 is not differentially expressed during adipocyte differentiation of hMSCs. The bar plots show the log2 transformed gene expression for 6 transcription factors (TSHZ1, MYBL1, MAZ, HSF1, NFIL3 and SATB1) predicted by IMAGE to be causal regulators of hMSC-TERT4 adipocyte differentiation. * denotes significantly (Padj ≤ 0.05) different expression level compared to day 0. **B**) The candidate factors all have changes in motif activity, but different patterns, during adipocyte differentiation in hMSCs. The graphs show the predicted motif activity for the six transcription factors identified by IMAGE. **C**) Validation at the mRNA level of the siRNA-mediated knockdown of the six transcription factors identified using IMAGE. Bar plots showing the expression of each factor is shown at day 0, day 1 and day 7 after induction of adipocyte differentiation. The percentage knockdown relative to control is indicated. **D**) IMAGE predicts target genes of different transcription factors with high accuracy in hMSCs. The bars show the enrichment of IMAGE predicted target genes of HSF1, MYBL1, MAZ, SATB1 and TSHZ1 among genes affected by knockdown of each transcription factor. Knockdown-dependent genes were defined as the top 1000 most differentially expressed genes between the control and the knockdown in hMSC-TERT4 cells stimulated to differentiate along the adipocyte lineage for 7 days. The enrichment is calculated by comparing the fraction of predicted target genes that are knockdown-dependent (the precision) to a randomized control fraction using size-matched randomized groups of target genes and dependent genes. Background enrichment was estimated by 1000 permutation of randomizing the predicted target genes and calculating the same enrichment. The error bars indicate the standard deviation across 1000 permutations. **E**) Ovelap between MED1-bound enhancers and the indicated transcription factors. * Denote significant (P ≤ 0.05) enrichment compared to randomized regions. **F**) The change in metabolic flux through most reactions assigned to triacylglycerol synthesis or fatty acid synthesis correlate with the change lipid accumulation upon knockdown of the candidate factors. The heatmap shows the scaled and centered metabolic flux of all metabolic reactions assigned to triacylglycerol synthesis or fatty acid synthesis, which reached a flux of at least 0.1 umol per gDW per hr. Metabolic fluxes were inferred by the SPOT method (Kim et al. 2016) from gene expression data at day 7 in either of the knockdowns or in the control, as well as day 0 in the control, using a model of human metabolism (Recon2) (Thiele et al. 2013).

Supplementary Table Legend

Supplemental Table S1 – Related to Figure 5

Pathway analysis using goseq and the Reactome database of differentially expressed genes in hMSC-TERT4 cells transfected with the indicated siRNAs. For the control cells, differentially expressed genes were defined as genes that change significantly in expression between day 0 and day 1 of adipocyte differentiation (induced or repressed), whereas for the five transcription factor knockdowns, differentially expressed genes were defined as genes that change significantly in expression between control cells and knockdown cells at day 1 of adipocyte differentiation (higher or lower expressed).