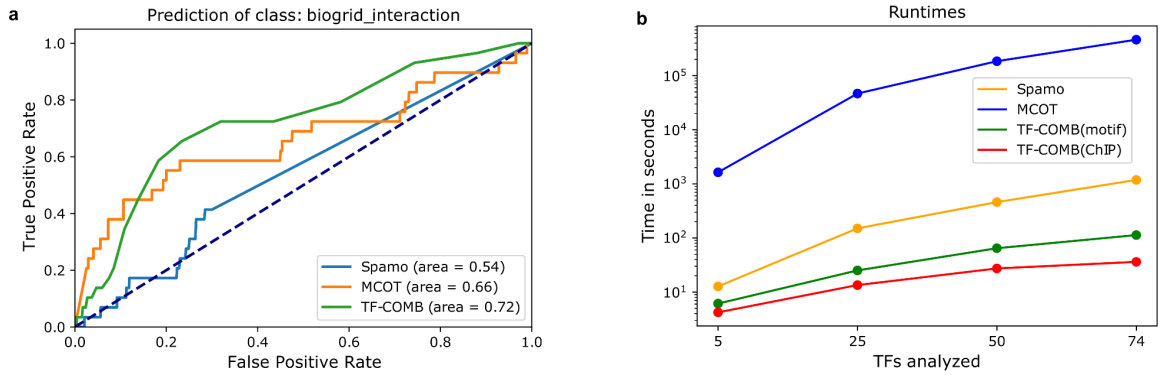
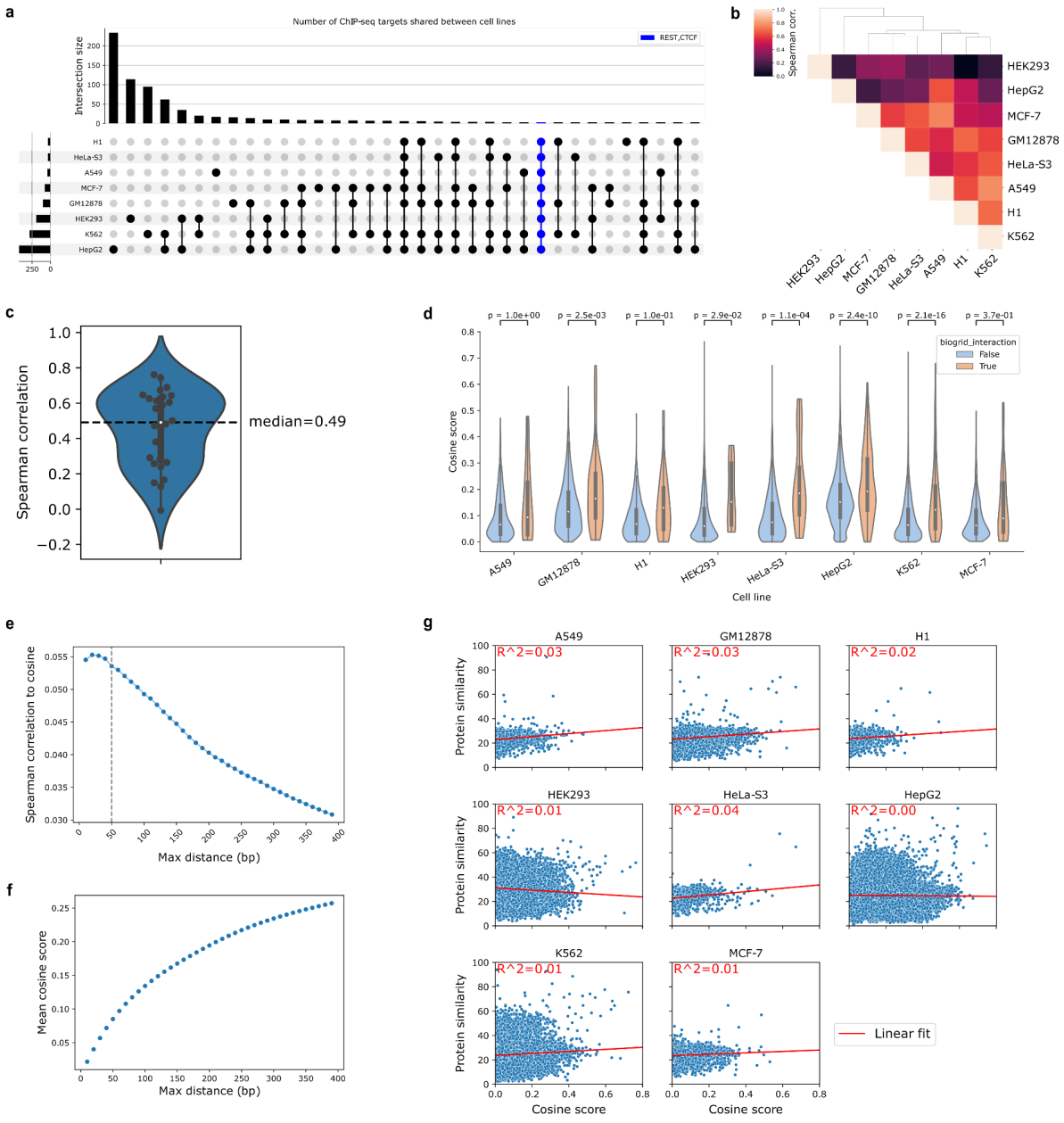
Supplementary Figures



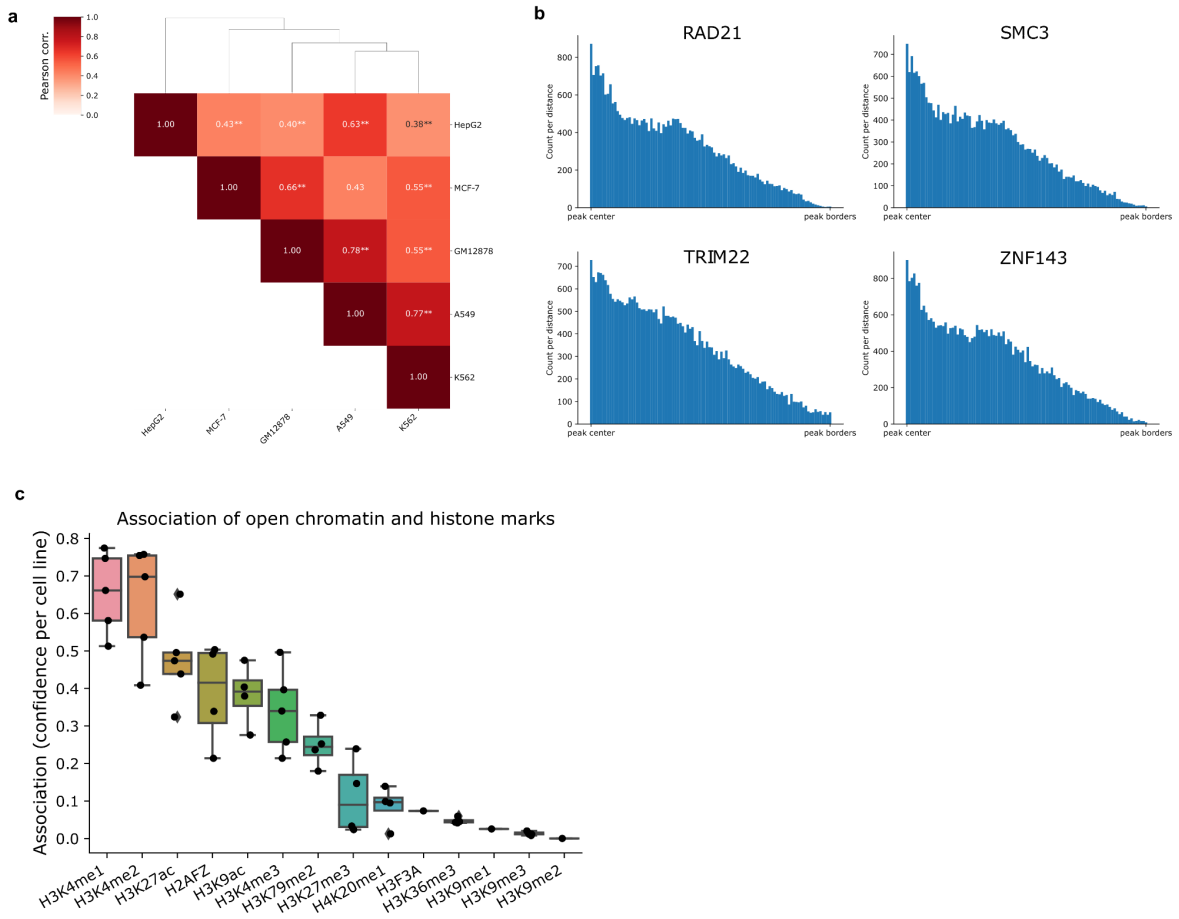
Supplementary Figure S1: Validation of TF-COMB and other tools

a) ROC curve of predictive ability of the assessed tools. Dashed line represents auROC=0.5.
b) Benchmark of runtime for assessed tools with increasing number of TFs analyzed.



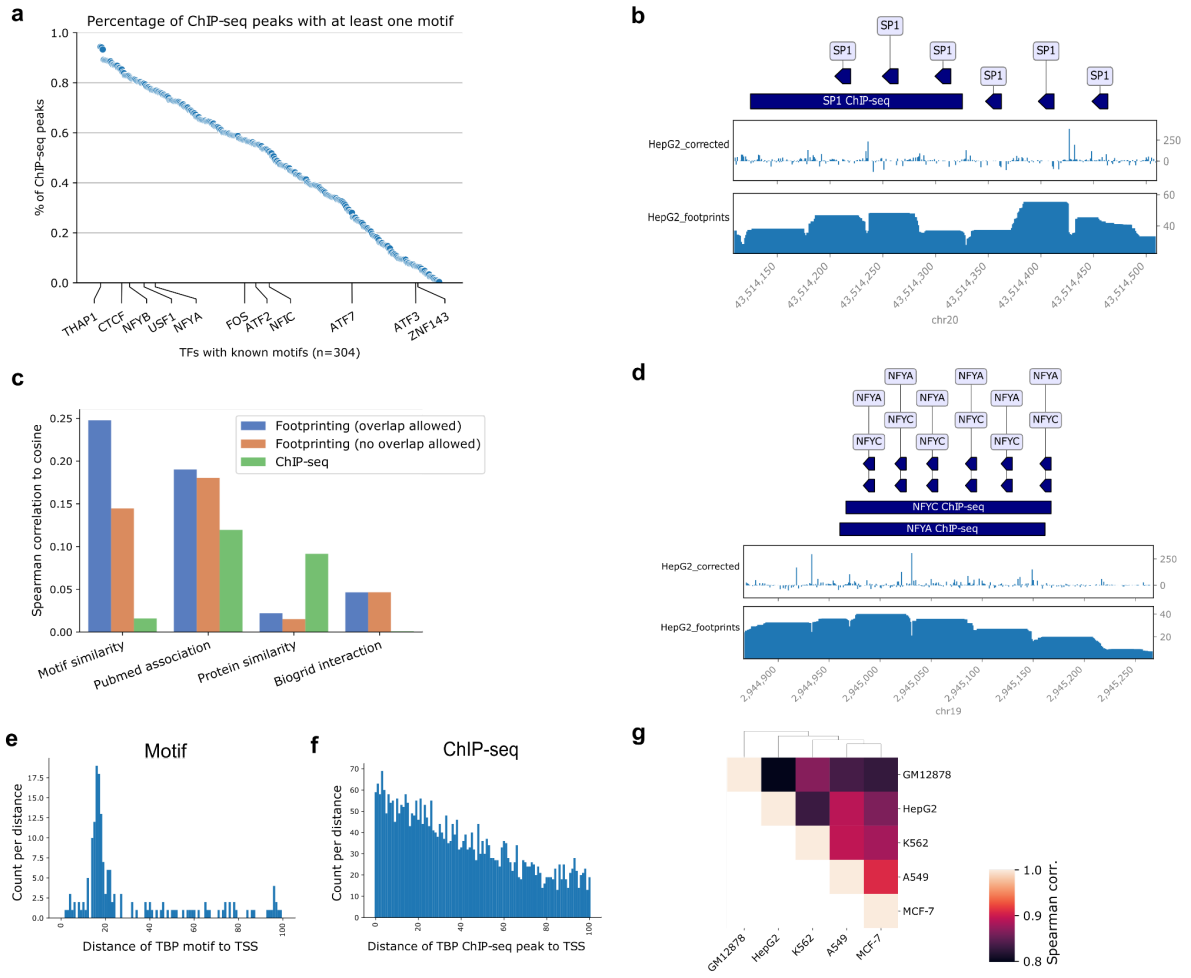
Supplementary Figure S2: Co-occurrence of ChIP-seq peaks across cell lines

a) Intersection count of available TFs ChIP-seq experiments for each cell line combination. Amount of TFs present in all cell lines highlighted in blue. Number of distinct TFs per cell line are on the left.
b) Correlation of TF-pair cosine values between cell lines.
c) Distribution of correlation values shown in b).
d) Distribution of cosine association scores for TF-pairs with and without PPIs across cell lines.
e) Correlation of cosine scores to PPIs with increasing allowed distance between TFs. Max distance of 50bp is marked with a dashed line.
f) Mean cosine association score with increasing allowed distance between TFs.
g) Cosine association compared to protein similarity of TF-pairs across cell lines. A linear fit is added in red and the R-squared is added at the left corner for each cell line.



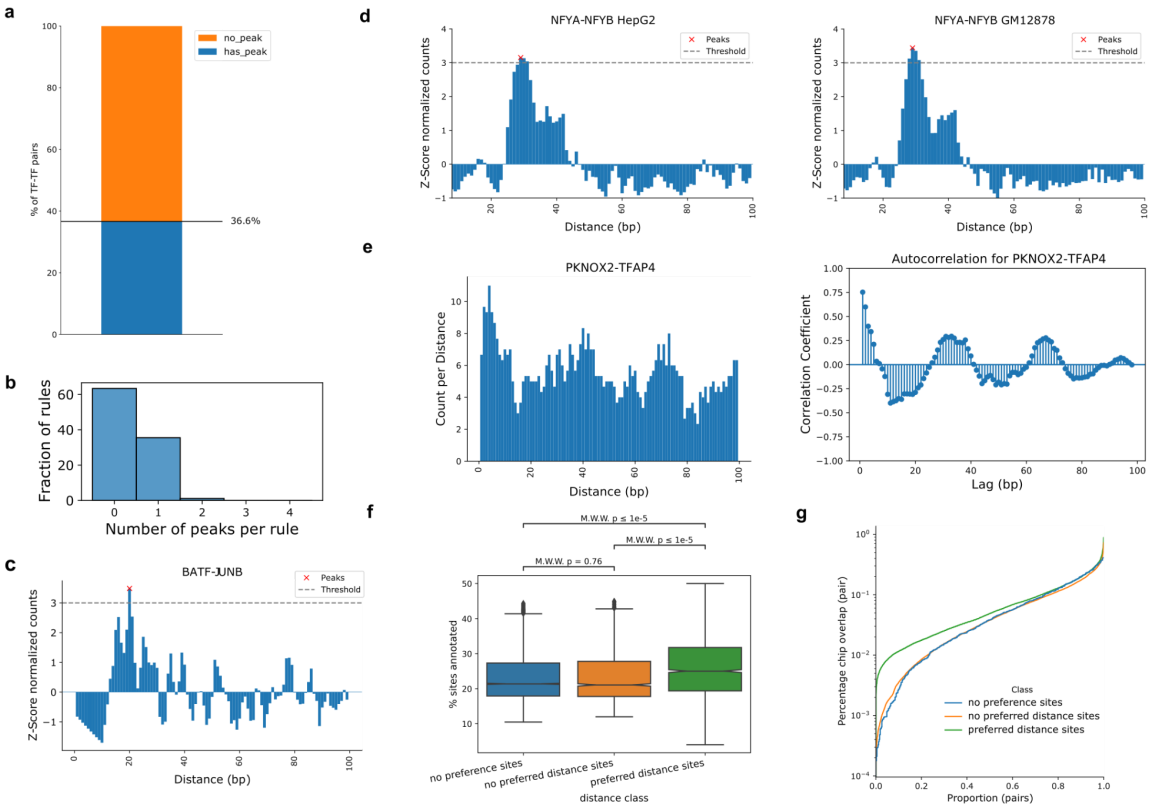
Supplementary Figure S3: Integration of epigenetic marks for co-occurrence

a) Correlation of TF binding distances(ChIP-seq) in relation to open chromatin (ATAC-seq) between cell lines. Significance of the correlation is marked with stars, where a p-value of less than 0.01 and 0.001 are marked with “*” and “**”, respectively.
b) Examples of relative TF binding locations within peaks. The x-axis represents the binding location from center (left) to border (right) of open chromatin peaks.
c) Association of open chromatin to locations of histone modifications and variants. Each box shows the association for all cell lines to a specific histone modification.



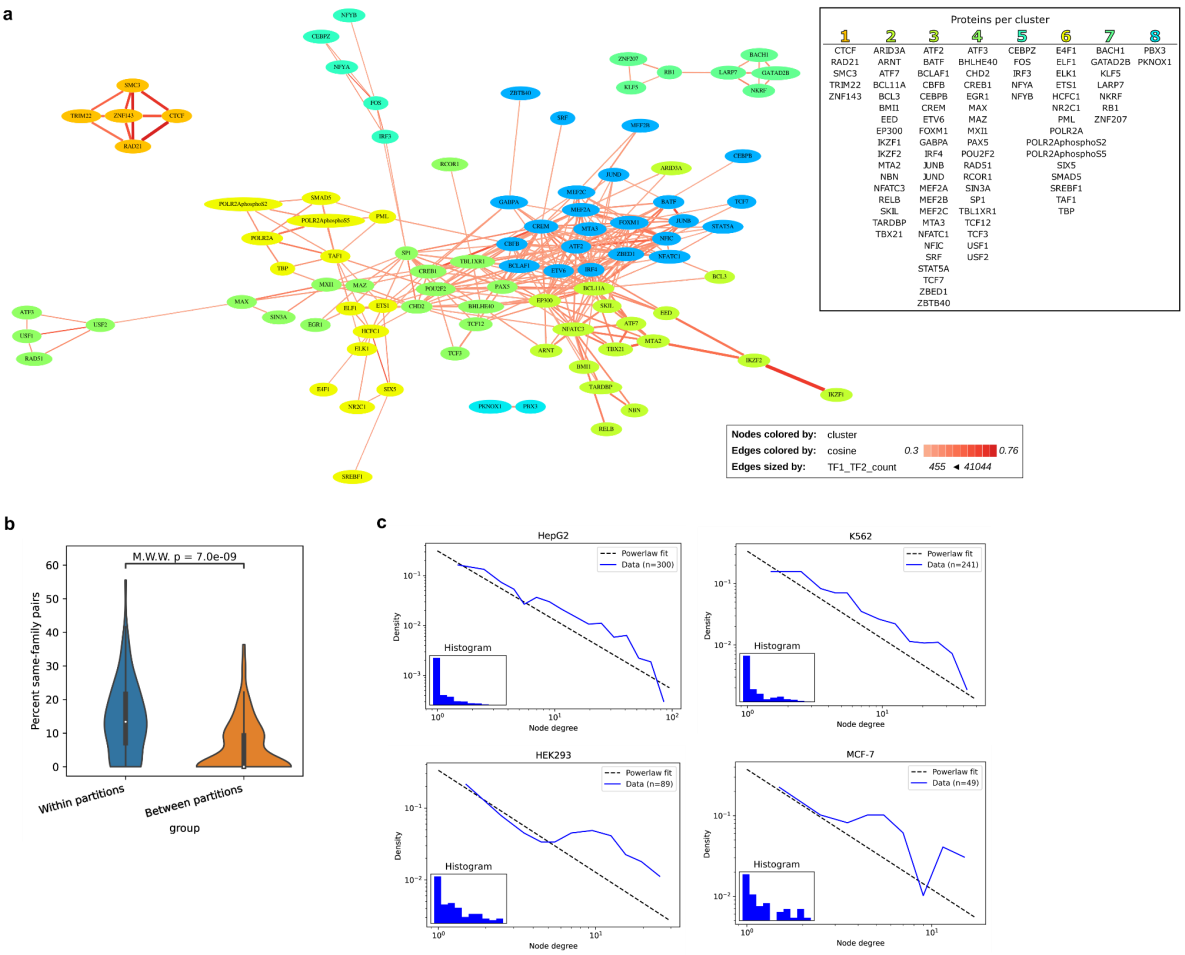
Supplementary Figure S4: Additional aspects of co-occurrence of footprinting

a) Fraction of ChIP-seq peaks containing sites of associated TF motifs.
b) Illustration of ChIP-seq peaks (upper blue boxes) in comparison to Tn5 signal (middle track) and TOBIAS derived footprinting score (lower track) for SP1. Individual TF motif locations are shown as triangles (upper track).
c) Correlation of cosine score with motif similarity, literature association score, protein similarity and protein interaction (BioGrid) score respectively per TF pair. Shown for ChIP-seq and footprinting with overlap allowed and excluded respectively.
d) Same plot as described in b) for NFYA/NFYC.
e-f) Distribution and gain of resolution shown for TBP binding site distances to TSS sites based on footprinting data analysis e) and ChIP-seq analysis f).
g) Correlation of TF pair cosine scores between cell lines from footprinting derived data.



Supplementary Figure S5: Additional aspects of binding grammar

a) Percent of TF-pairs that exhibit at least one preferred binding distance (peak) as displayed in c-d).
b) Distribution of TF-pairs (rules) on their number of predicted preferred binding distances.
c) Z-score normalized TF-pair binding counts sorted by distance. Peaks above threshold are considered preferred binding distance.
d) Difference in binding distance distribution for NFYA-NFYB in HepG2 (left) and GM12878 (right).
e) Periodic binding distance preference for PKNOX2-TFAP4. Left plot shows the distribution of binding distances for all co-occurring sites. Right plot shows the calculated autocorrelation, i.e. lag of binding distances, for the pair.
f) Percentage of sites annotated to genes (using UROPA) per distance class.
g) Proportion of ChIP-seq overlap per distance class.



Supplementary Figure S6: Network analysis of co-occurrence

a) The GM12878 co-occurrence network. Nodes are colored on the basis of Louvain community clustering. On the right a list of TFs per cluster is shown.
b) Distribution of same-family pairs randomly picked from within or between partitions (clusters of a)). Percent same-family is estimated per subset for 20 randomly selected pairs across n=1000 iterations.
c) Node degree (count) for all TFs of different cell line networks. Node degrees follow power-law distribution.