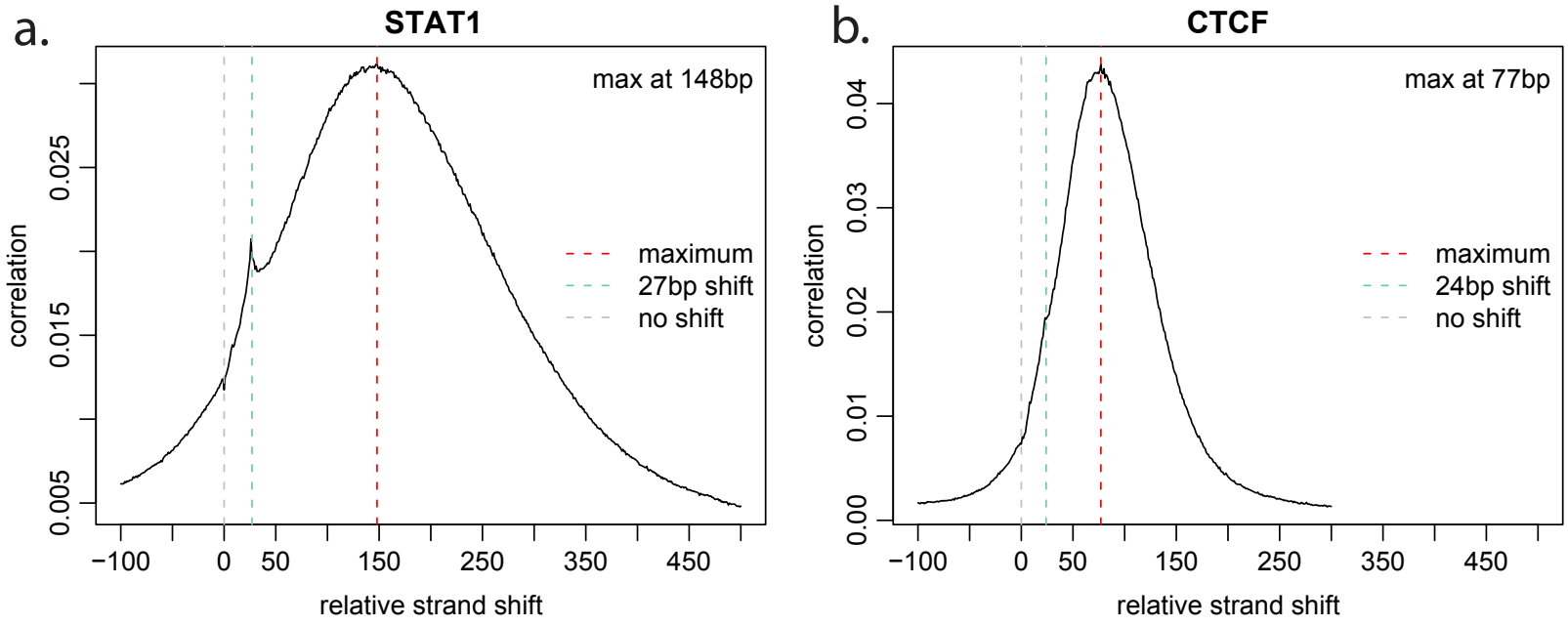
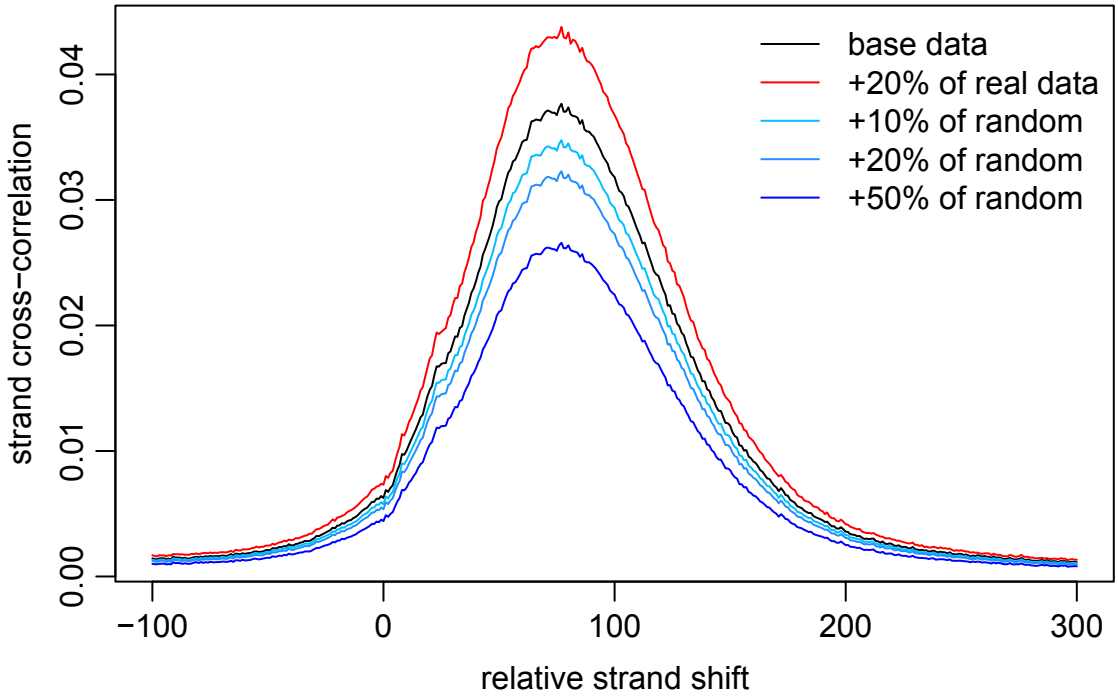
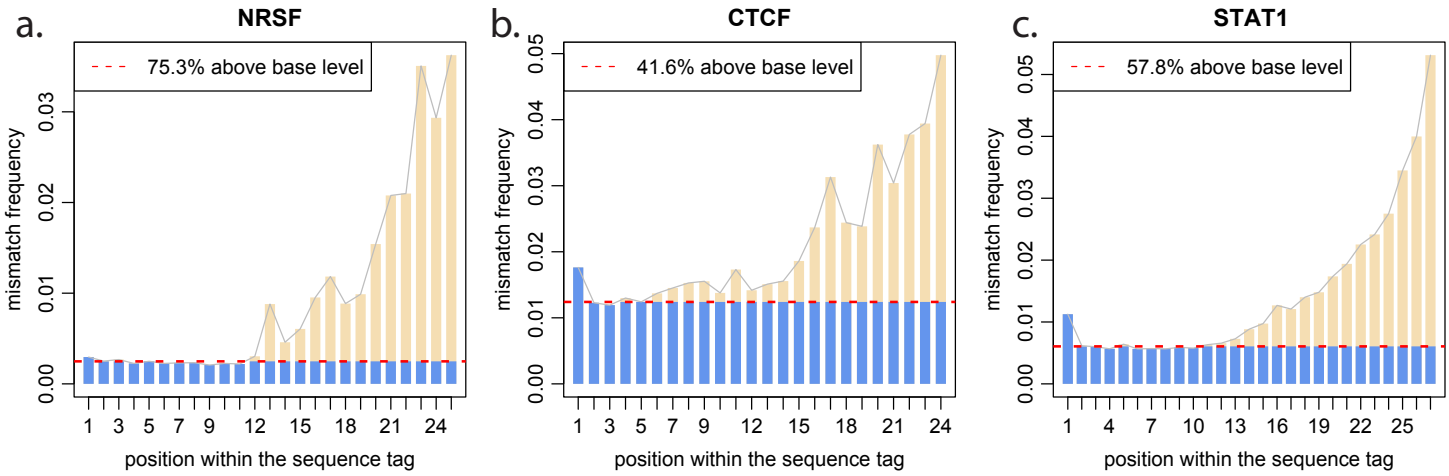


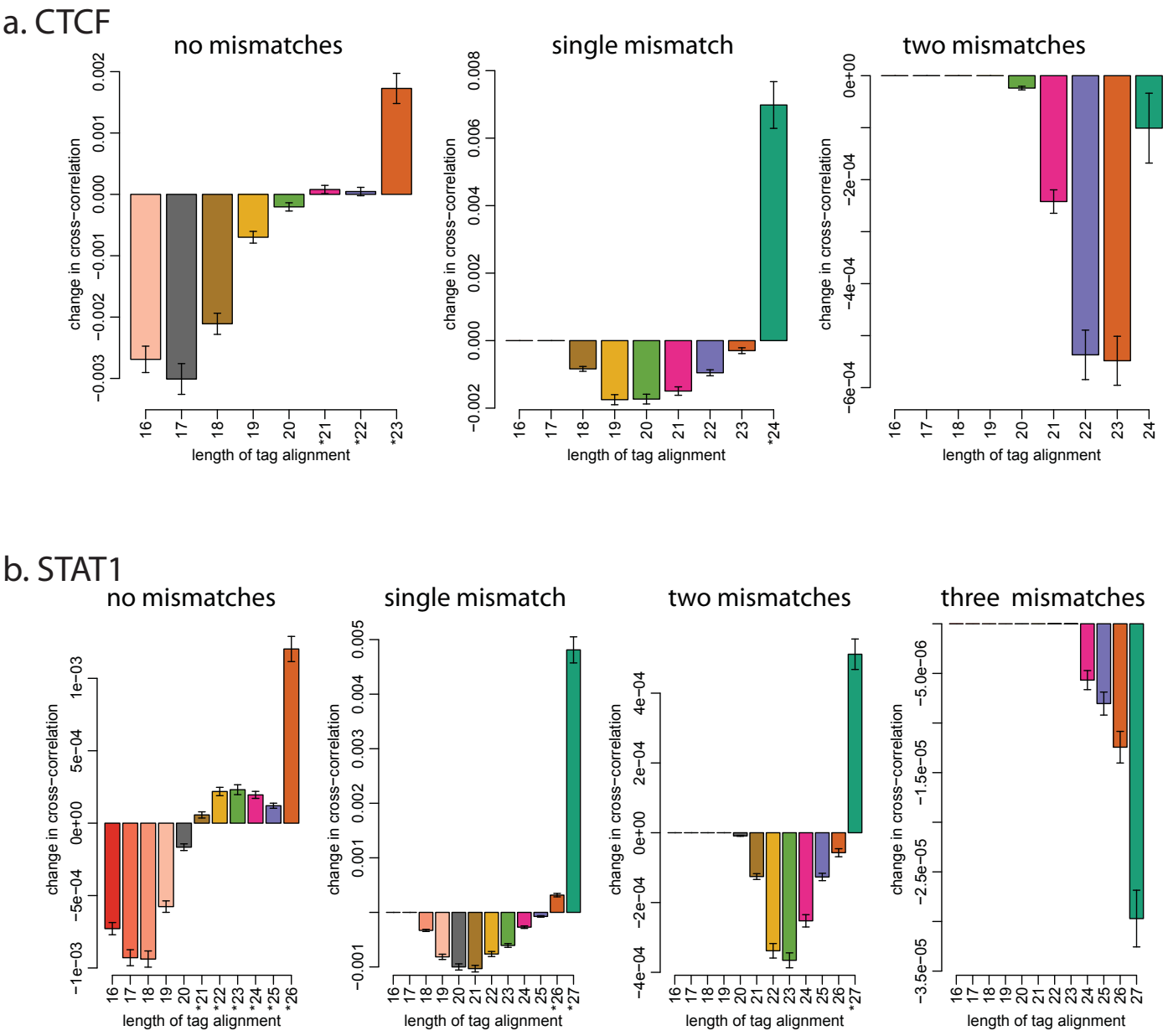
Supplementary Figure 1. Strand cross-correlation profiles are shown for STAT1 (**a**) and CTCF (**b**). The dashed red line marks the position of the maximum cross-correlation. The light-blue lines mark strand shift corresponding to the length of the Solexa tag reads. A jump of cross-correlation at such shift is present in some datasets, in particular STAT1. The gray line marks 0bp strand shift.



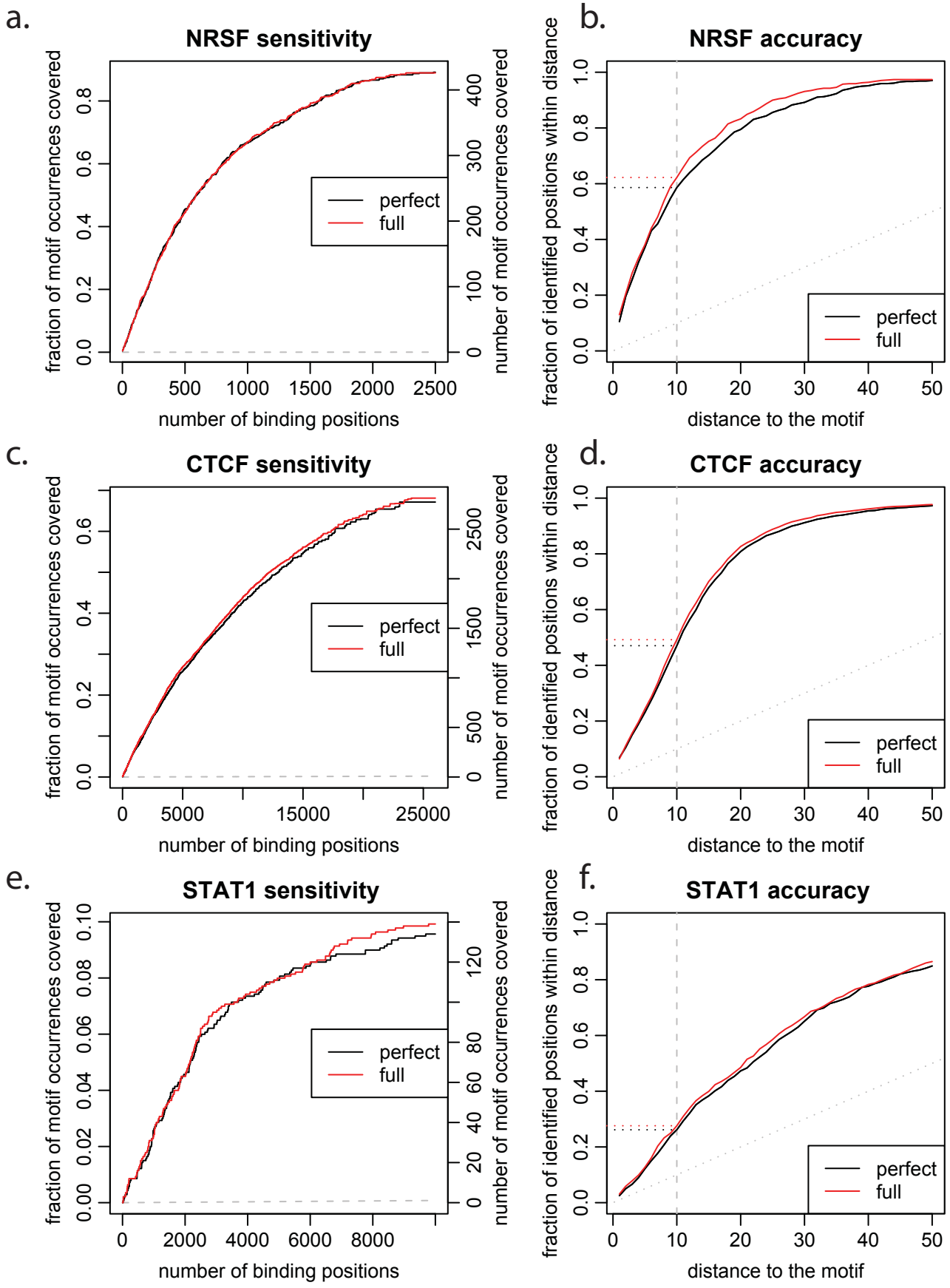
Supplementary Figure 2. Changes of cross-correlation profile with addition of different tag sets. The black curve shows cross-correlation profile for a fraction (83.3%) of the complete CTCF data. Addition of the remaining true data (increase of 20%) improves the cross-correlation (red curve), whereas addition of increasing amounts of completely random (Poisson process) data reduces it (blue curves).



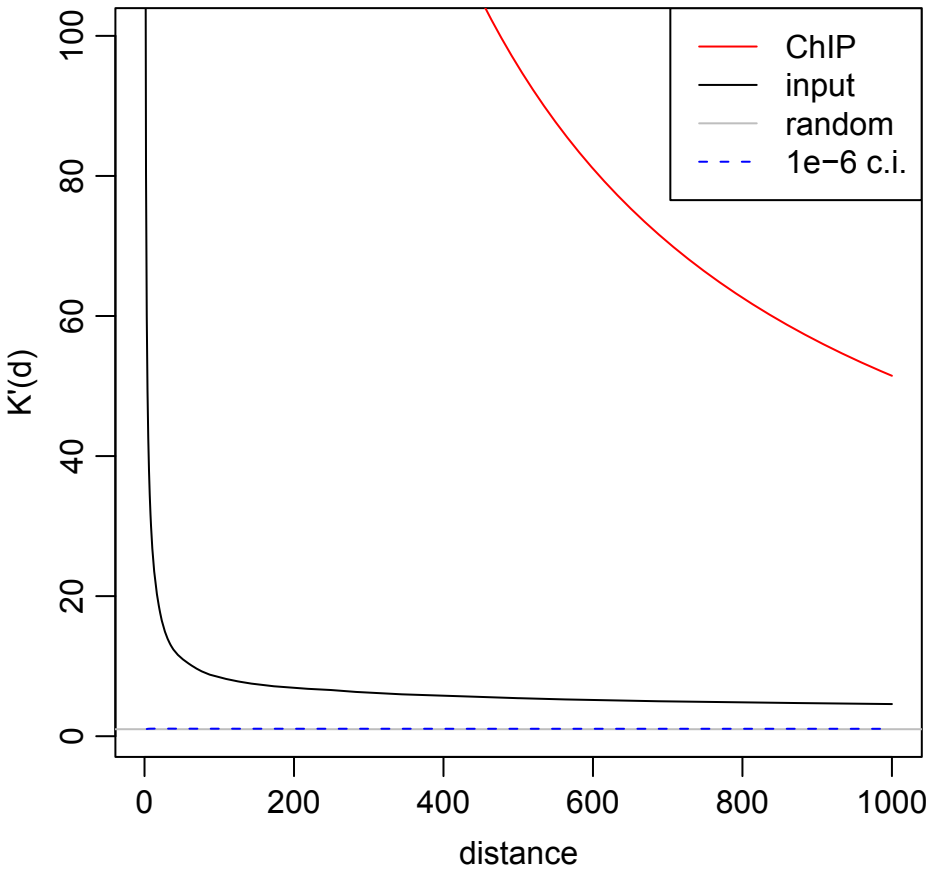
Supplementary Figure 3. Mismatch occurrence at different positions within the sequence tags. The plots show mismatch frequency for different positions within the tags of three datasets. In all cases, the mismatch frequency increases towards the 3' end of the tags. To estimate the fraction of mismatches that can be explained by this increase, we estimated the base level of mismatches based on the positions 2-5 (dashed red line). The overall fraction of mismatches stemming from the position-specific effects is marked by wheat color, whereas remaining fraction marked in blue. The fraction is given in the upper left corner. Since the first nucleotide in the tag sequence tends to exhibit increased error rate, and therefore does not follow the overall trend, it was excluded when counting position-driven mismatch fraction.



Supplementary Figure 4. Selecting informative tags for CTCF and STAT1 datasets. Similar to Figure 2 of the main manuscript, the plots show cross-correlation changes resulting from consideration of different tag alignment quality classes together with the base set of perfectly aligned tags. The plots are given for (**a**) CTCF and (**b**) STAT1.

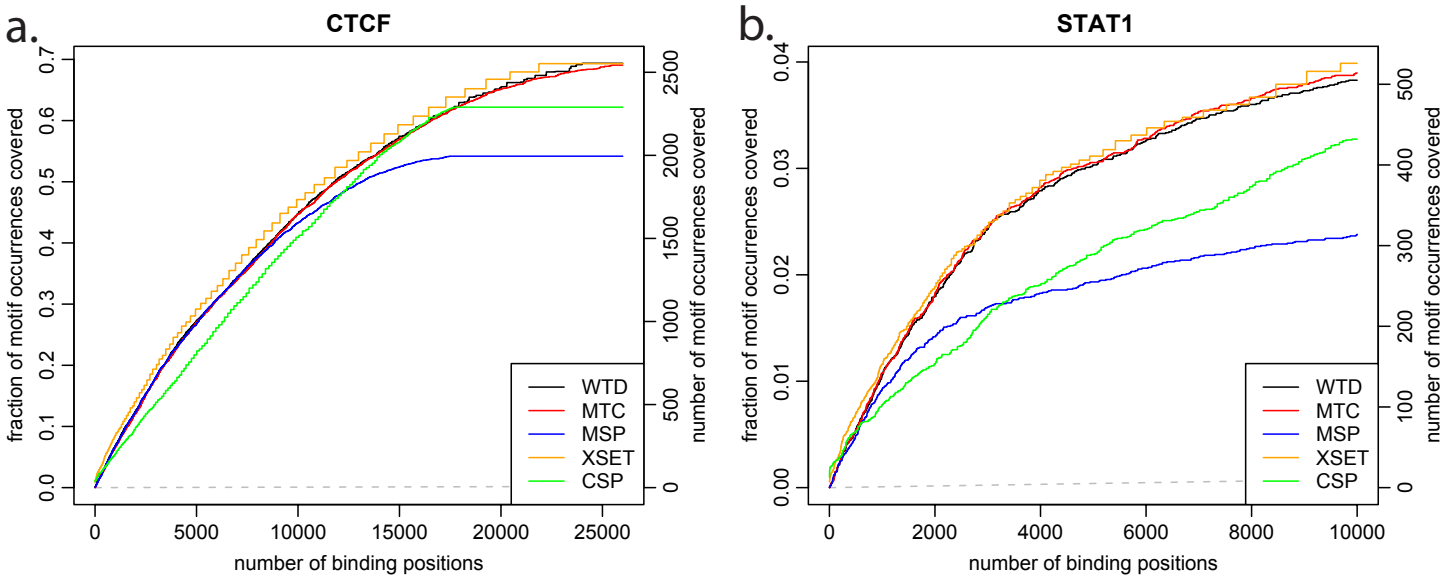


Supplementary Figure 5. The effect of tag increase on motif coverage and accuracy of identified binding positions. The plots show the difference in coverage of high-scoring motif positions (left column) and accuracy of identified binding positions (right column) for three datasets. In each plot, black lines show results based only on the set of tags showing perfect alignment, and red curves show the result on the set of tags determined by the acceptance procedure described in the manuscript. The later performs better in nearly all cases, except for the NRSF coverage where both sets of tags result in equivalent motif coverage.

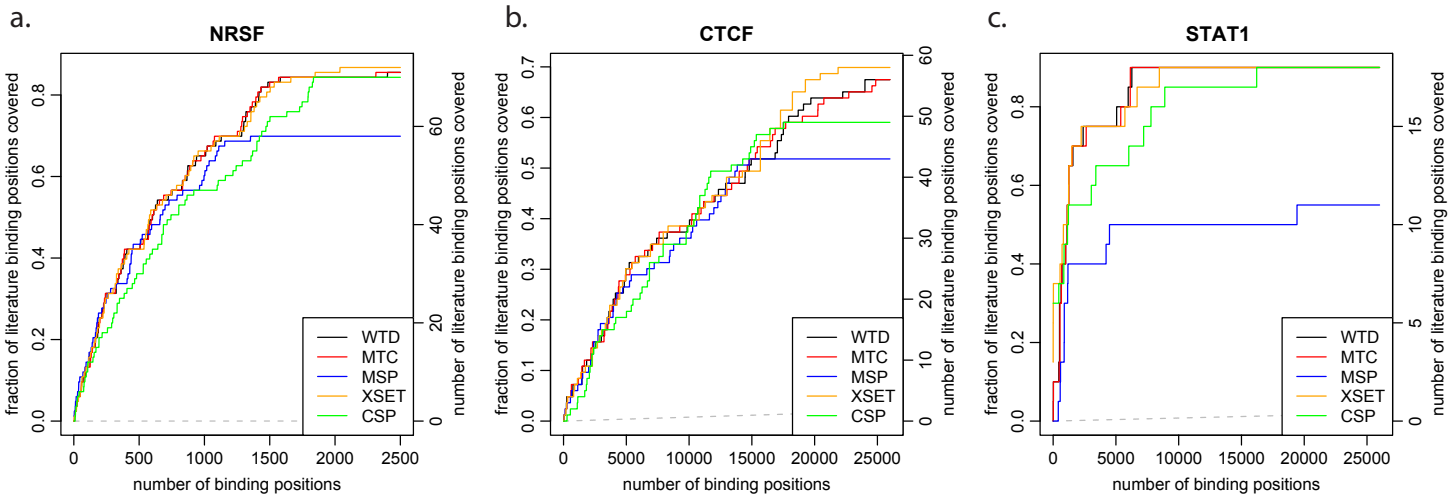


Supplementary Figure 6. Non-uniform properties of the background distribution. To compare background (input) tag distribution with a completely randomized model (Poisson spatial process) we use Rippley's K function, which reflects the degree of spatial clustering of a spatial point distribution[1] . For a Poisson spatial process in one dimension, $K(d) = 2d$, where d is a distance on the chromosome. The y-axis of the plot shows $K'(d) = \frac{K(d)}{2d}$ (so that Poisson process curve appears as constant). The $K'(d)$ observed for the background tag distribution (black) is significantly higher than that expected of a Poisson spatial process. The significance of the observed difference is demonstrated by the 10^{-6} confidence interval of the Poisson process $K'(d)$, which was estimated based on 2×10^6 simulations of a Poisson spatial process on the genome with the number of tags equivalent to that present in the experimental input tag dataset.

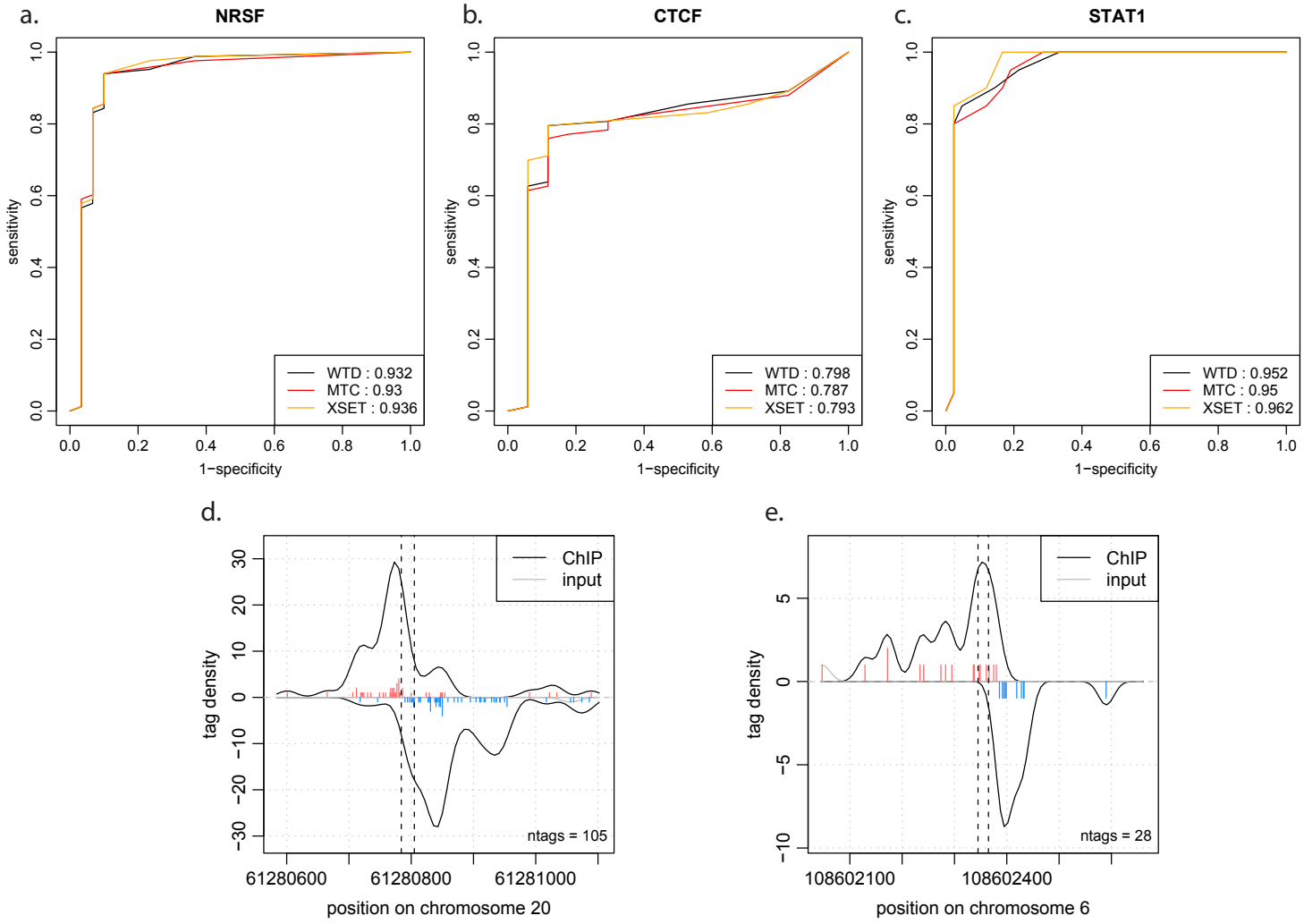
1. Diggle, P. Statistical analysis of spatial point patterns, Edn. 2nd. (Oxford University Press, London; 2003).



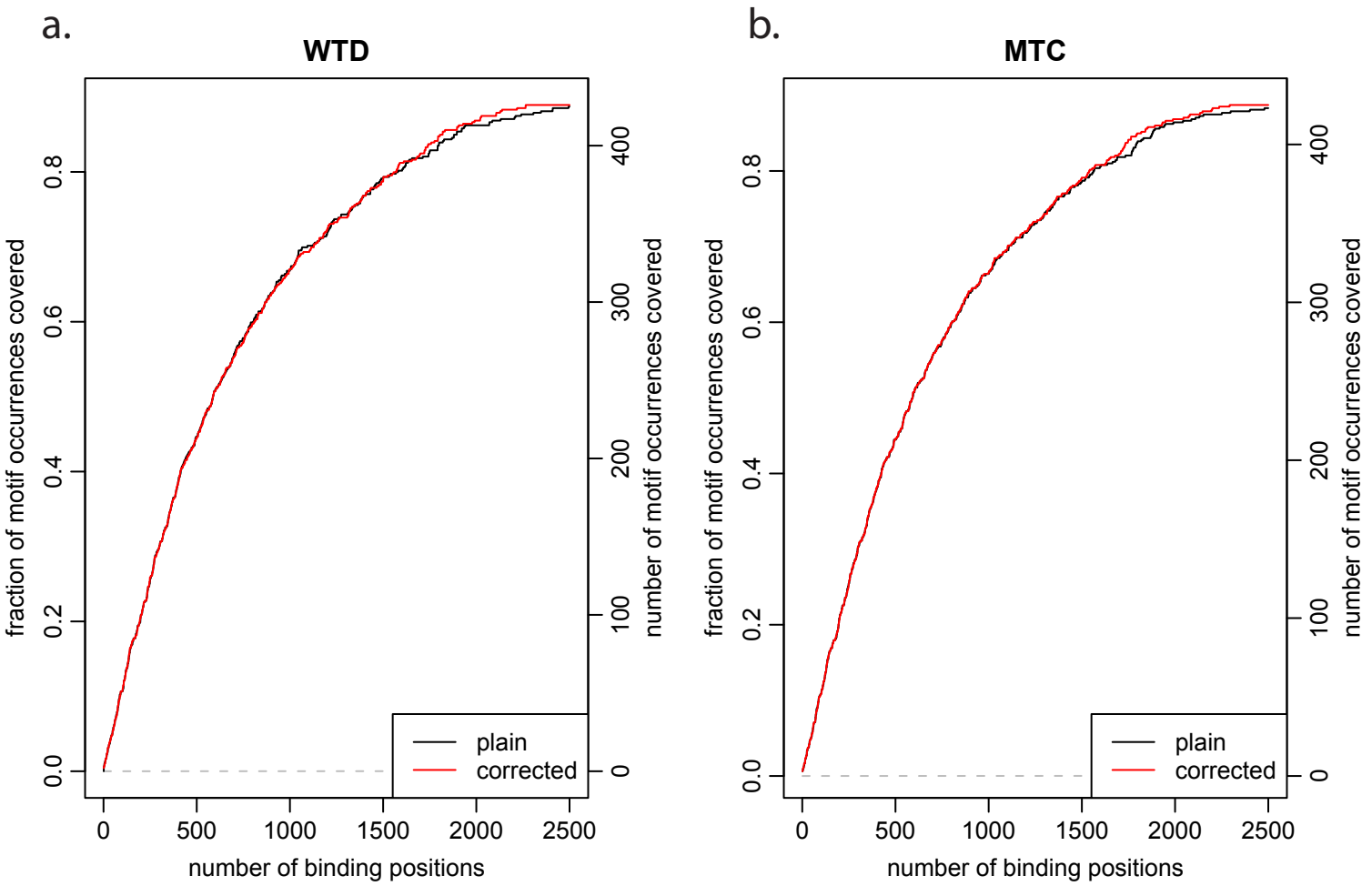
Supplementary Figure 7. (**a**) CTCF and (**b**) STAT1 motif coverage by different binding detection methods. Similar to Figure 4d of the main manuscript, the fraction of high-confidence motif positions coinciding (within 50bp) with the predicted binding position is shown for increasing number of top positions determined using different methods. Note that we were not able to generate required 26000 top peaks for the CTCF data using CSP method, resulting in abrupt, early saturation (**a**). A total number of high-confidence NRSF motifs was 764, 3682 for CTCF, and 13191 for STAT1.



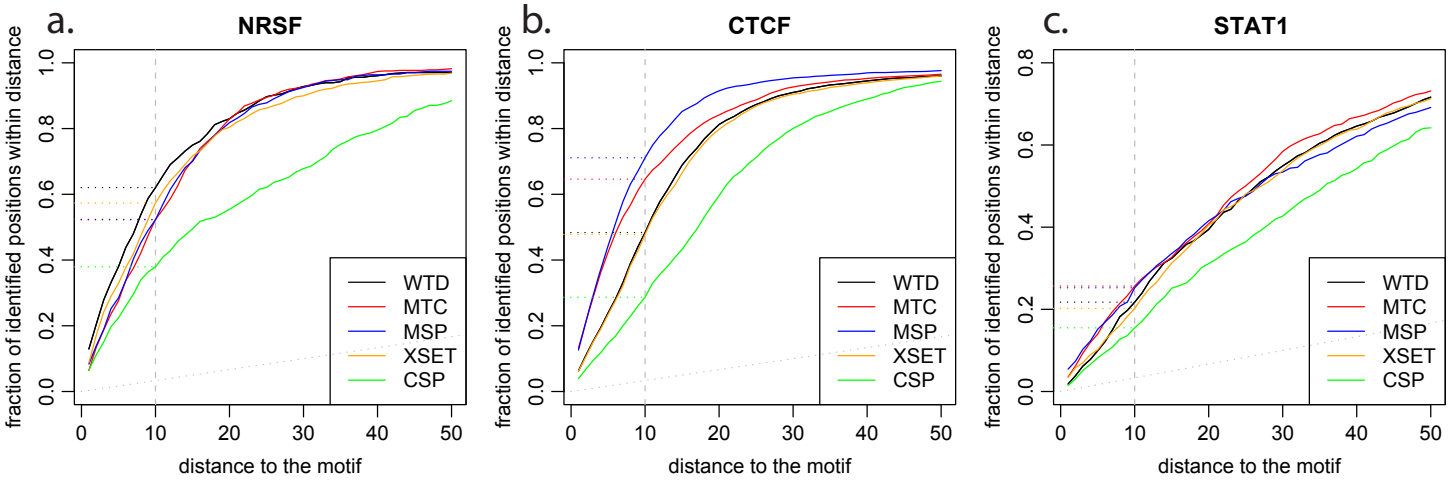
Supplementary Figure 8. In an analysis similar to Figure 4d of the main manuscript, the plots show fraction of the literature-validated binding sites covered (y-axis) with increasing number of top predicted binding positions (x-axis). The qPCR-validated literature sites for NRSF, CTCF and STAT1 were taken from [2], [16] and [11] respectively. There are a total of 83 such positions for NRSF, 83 for CTCF and 20 for STAT1.



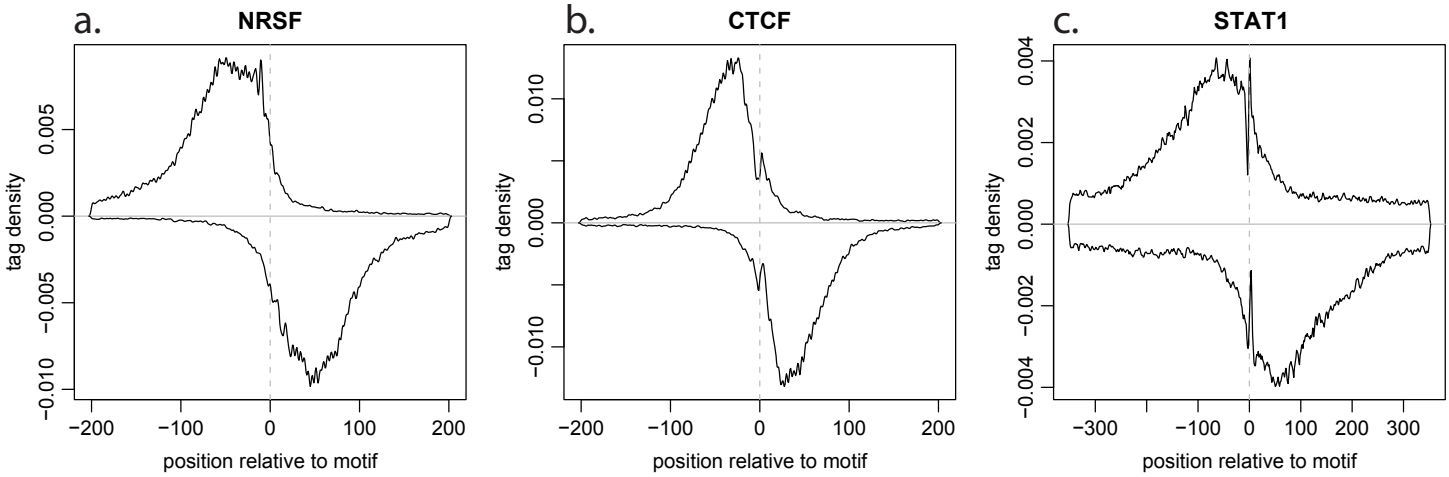
Supplementary Figure 9. a-c. The receiver-operating characteristic (ROC) curves show relationship between specificity and sensitivity for NRSF (**a**), CTCF (**b**) and STAT1 (**c**). The calculations utilize qPCR-validated literature positions described in Supplementary Figure 8. The curves are shown only for a subset of methods whose implementations allow calculation of binding scores at arbitrary genome positions. **d,e**. The differences observed between the ROC curves of different methods are explained predominantly by scoring of several "true negative" positions from the literature. The plots illustrate tag distribution around top two such positions from NRSF dataset. While qPCR tests of these two positions report enrichment ratios of 1 and 2.08, the tag profiles clearly show pronounced tag patterns typical of real binding, and little input tag density – suggesting that these positions are indeed bound by NRSF. Therefore the minor differences observed in the ROC curves may be due to presence of true positive positions within the set of true negatives.



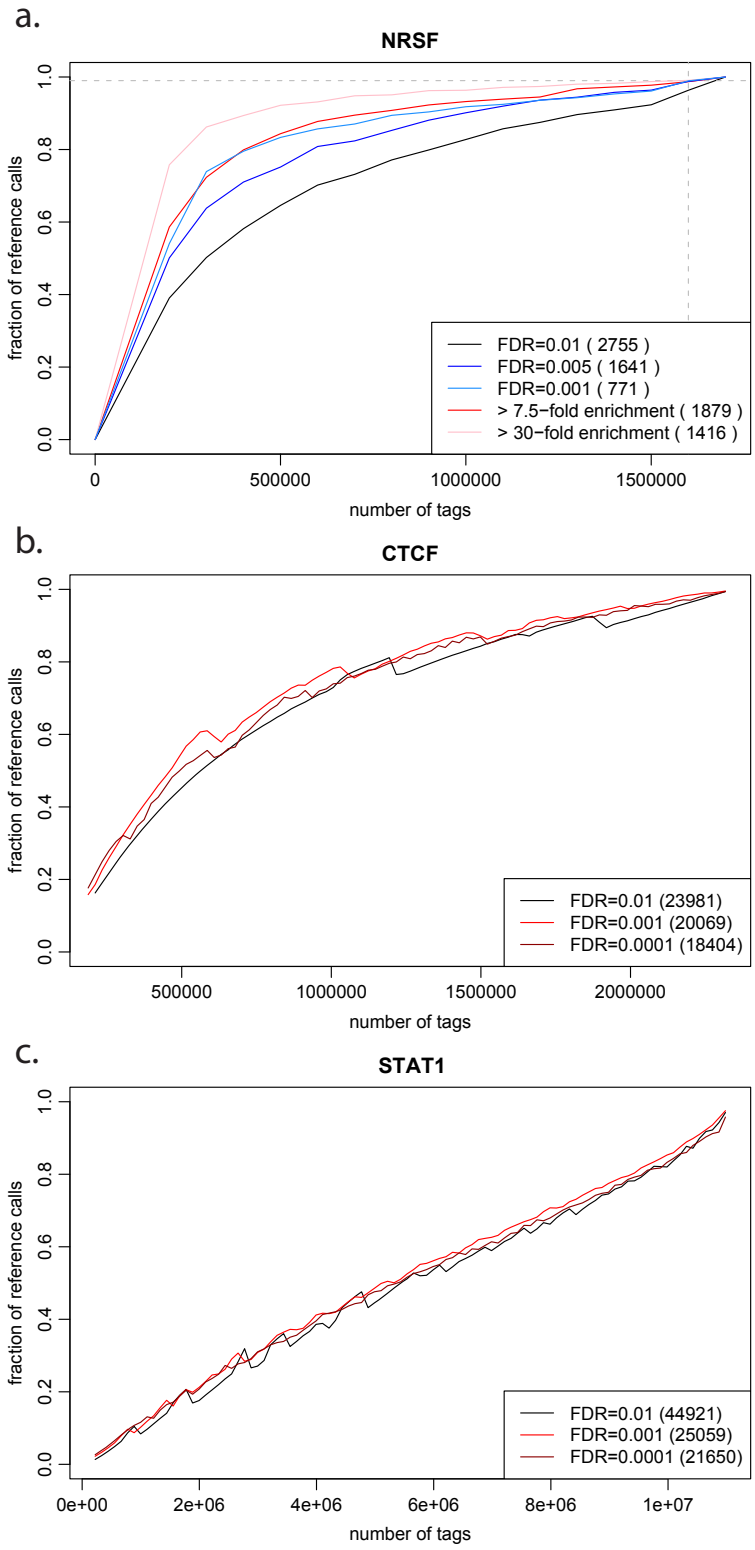
Supplementary Figure 10. Difference in coverage of high-scoring NRSF sequence motifs due to background corrections. The plots show fraction of motifs covered for increasing number of top predicted binding positions, with (red) and without (black) corrections for the background tag density. The results are shown for WTD method (**a**), and MTC method (**b**). While background subtraction has almost no effect within the top 1500 positions, background correction improves motif coverage at less prominent positions, allowing to achieve the same level of coverage with up to 11.3% fewer binding positions.



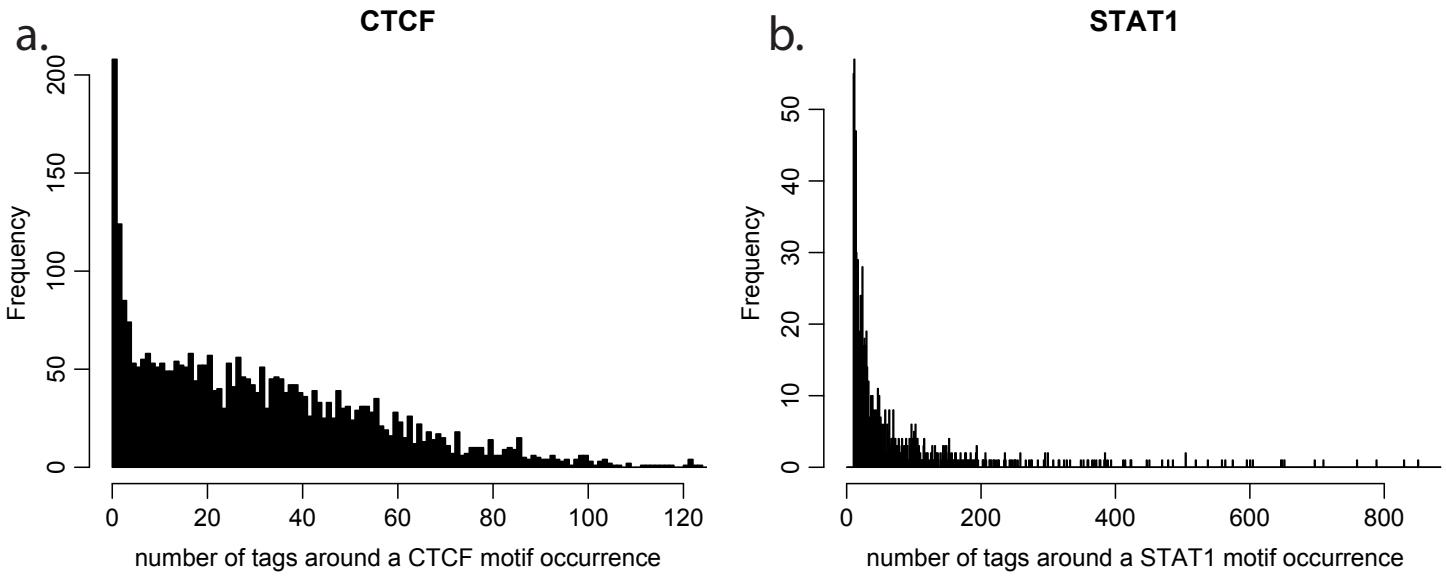
Supplementary Figure 11. Accuracy of binding positions determined by different methods. The plots show a fraction (y-axis) of binding positions determined by different methods that falls within a certain distance (x-axis) of a high-scoring motif occurrence. 10bp distance intersect is highlighted as an example. Only binding positions within 300bp of a high-scoring motif are considered in the analysis. The dotted gray line shows random expectation.



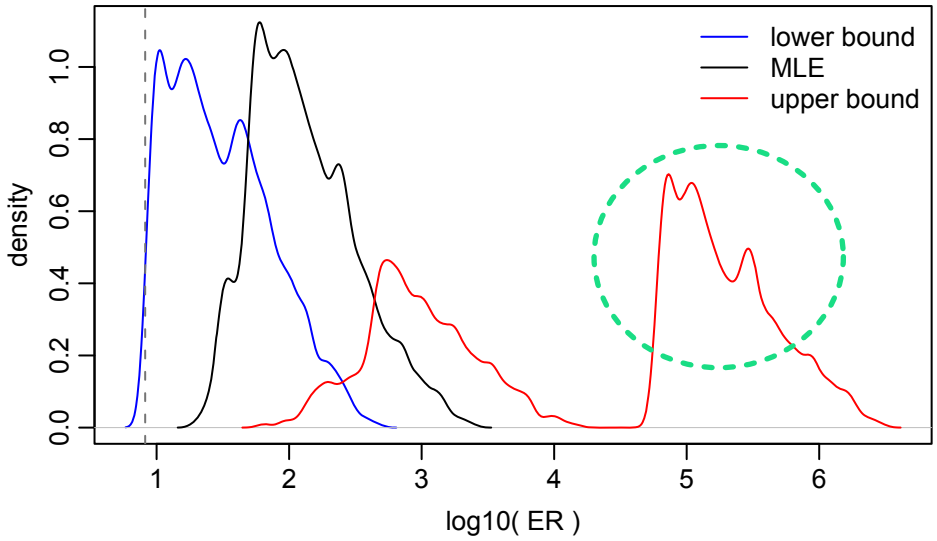
Supplementary Figure 12. Tag density profiles around motif positions. The plots frequency of tags mapping to positive (positive y-axis) and negative (negative y-axis) strands around high-scoring motif positions for **a.** NRSF, **b.** CTCF and **c.** STAT1.



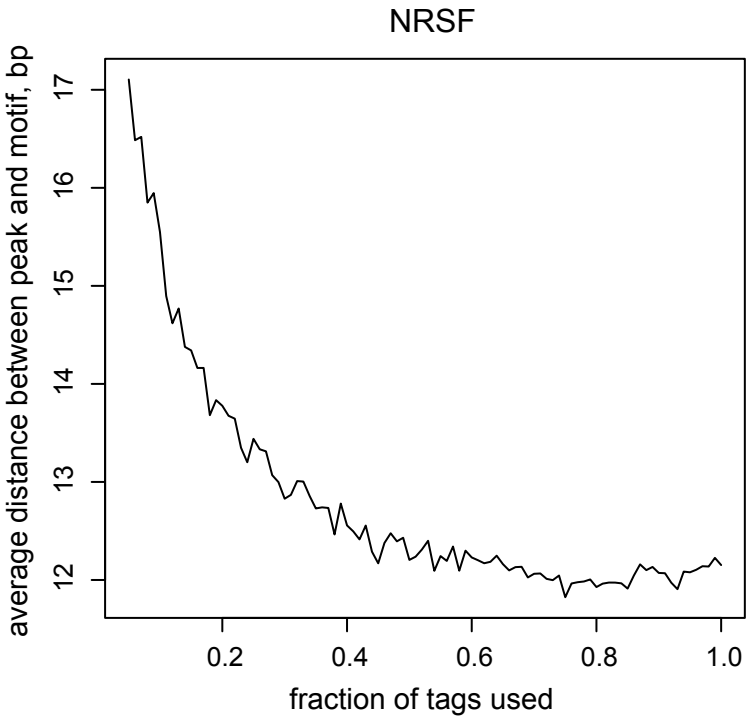
Supplementary Figure 13. Analysis of sequencing depth. The plots show fraction of reference binding positions (y-axis) that can be determined using a smaller random subset of the overall dataset. The x-axis shows the fraction of the overall tags that is being sampled for prediction. The plots are shown for NRSF (**a**), CTCF (**b**) and STAT1 (**c**). Predictions using different stringency thresholds (FDR) are shown in various red colors. The number of predicted binding positions is given in parenthesis next to each label. The NRSF plot (**a**) shows saturation curves where only binding positions confidently exceeding a certain fold enrichment ratio are considered (blue colors). All of the predictions are generated using WTD method.



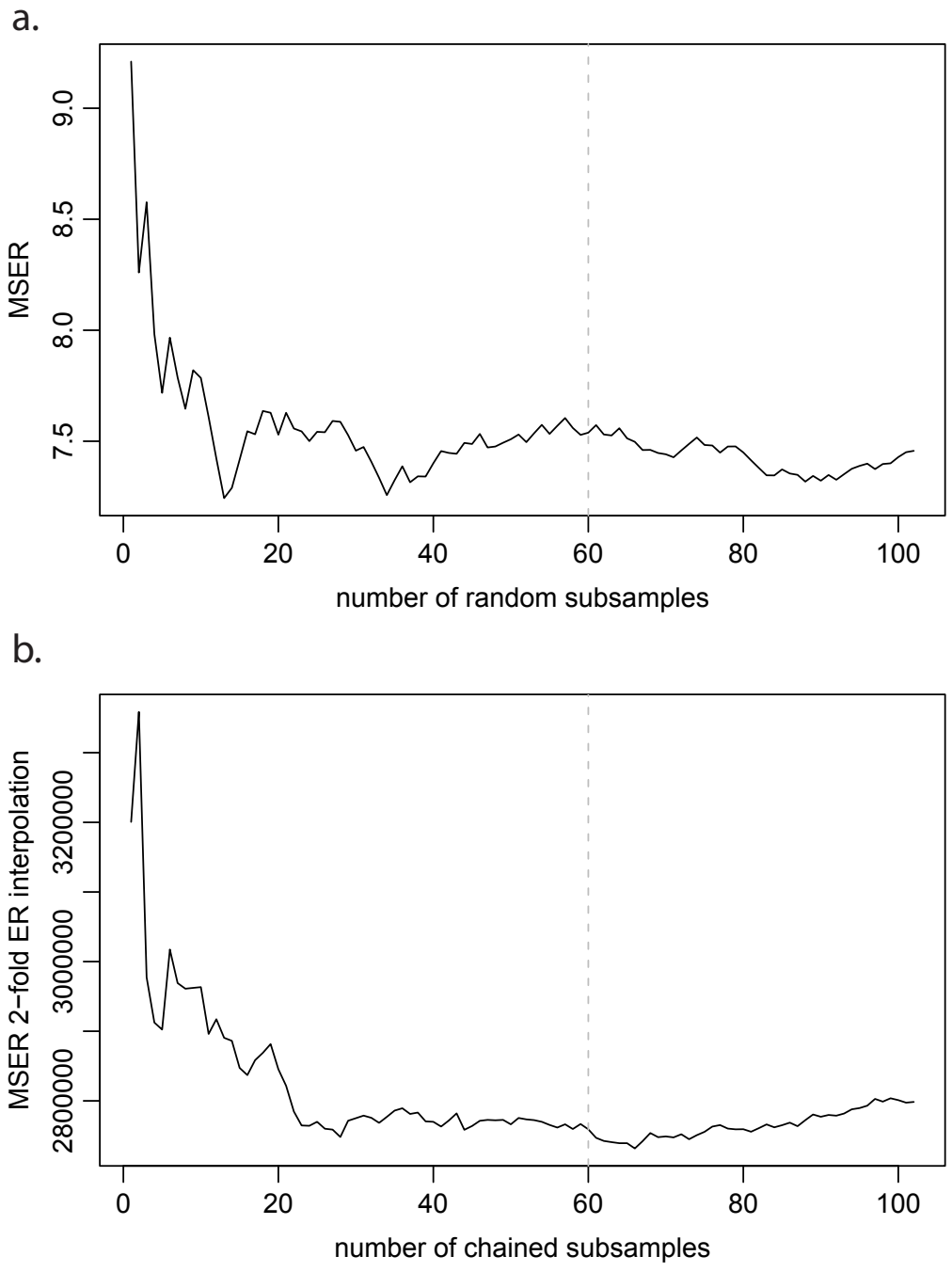
Supplementary Figure 14. Tag counts within 100bp around high-scoring CTCF (**a**) and STAT1 (**b**) motifs occurrences. The motifs with zero tags were excluded for visualization purposes.



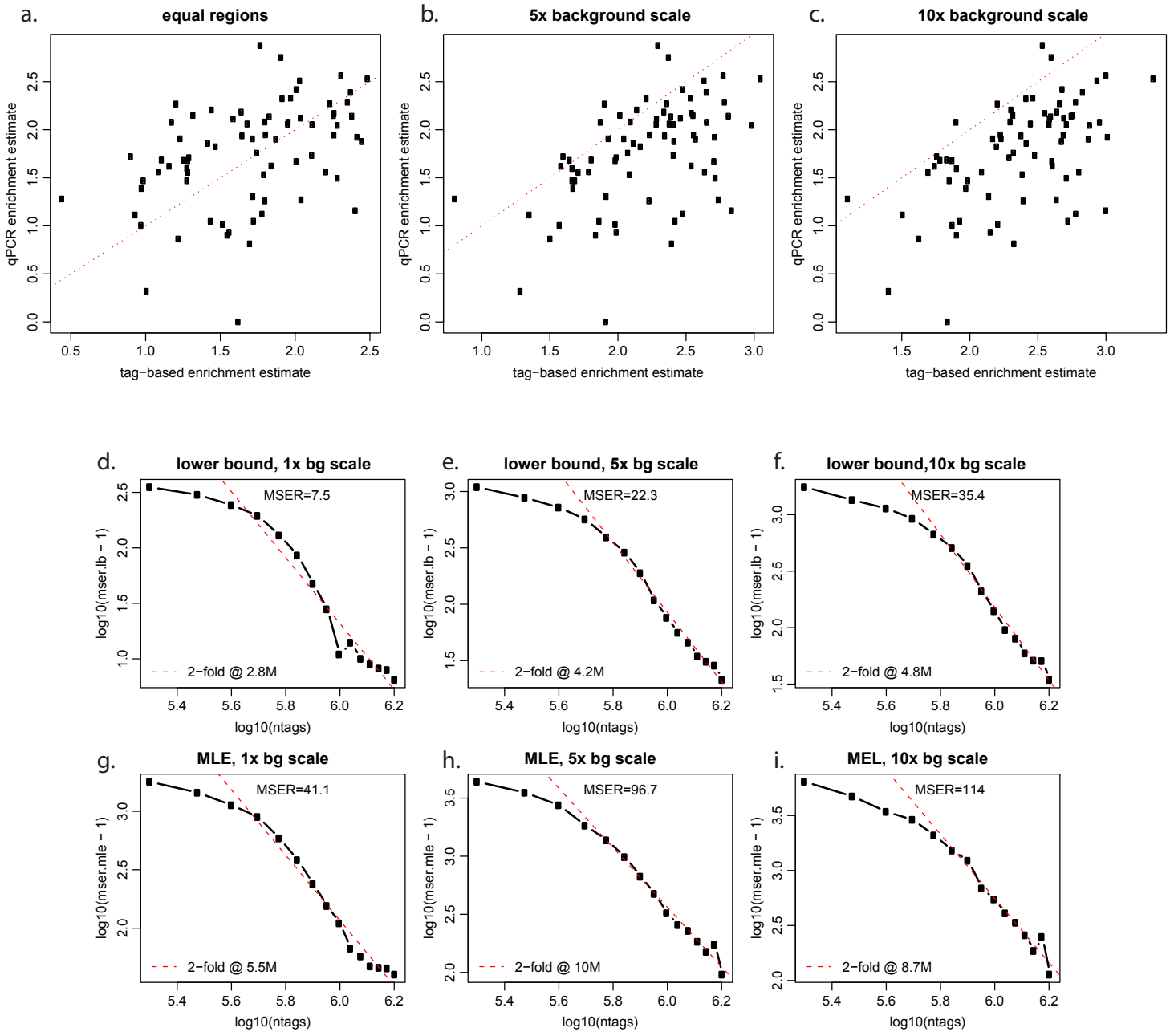
Supplementary Figure 15. Potential range of true enrichment ratios. While the fold-enrichment ratio of a particular binding position cannot be calculated precisely, we estimate a 95% confidence interval for each predicted position. The plot shows distribution of lower bounds (blue), maximum likelihood estimates (MLE, black curve), and upper bounds (red) for the entire set of the NRSF binding positions predicted using WTD method with FDR 0.01, whose enrichment is significantly higher than 7.5 (MSER value, marked by a vertical dashed line). The green circle points out a population of peaks for which the upper bound cannot be properly estimated due to lack of input tags in that region.



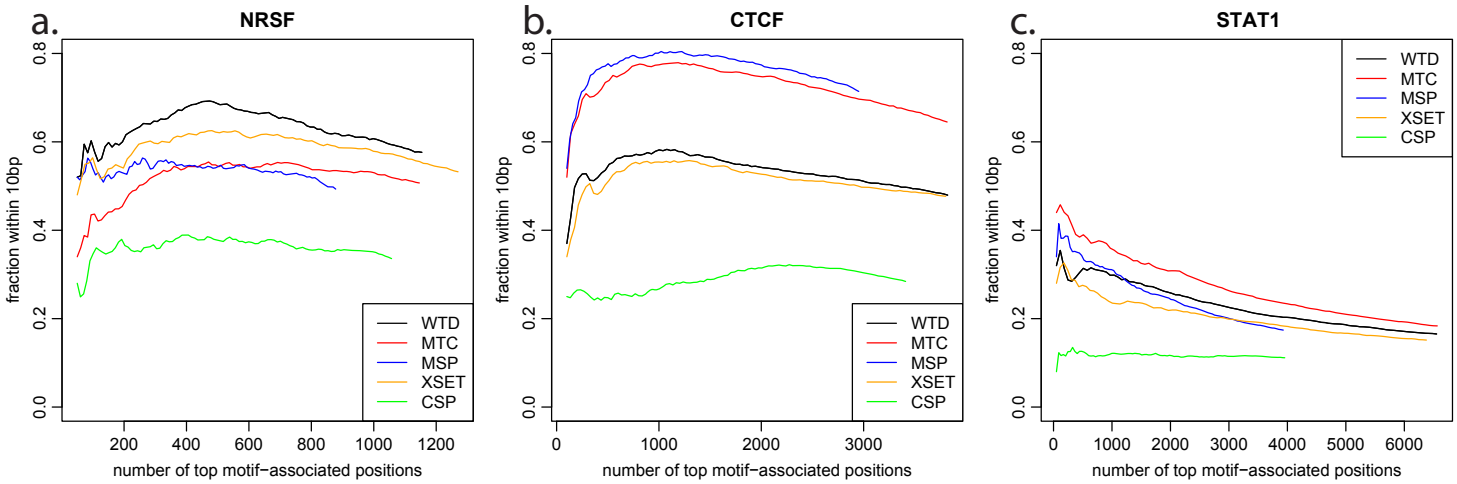
Supplementary Figure 16. The mean distance between the detected NRSF binding positions and the centers of the high-confidence sequence motifs matches (y-axis) are shown as a function of the fraction of tags used for binding detection (x-axis). The binding positions were determined using FDR of 0.01 using WTD method. Only the peaks that had a motif within 100bp from them were considered.



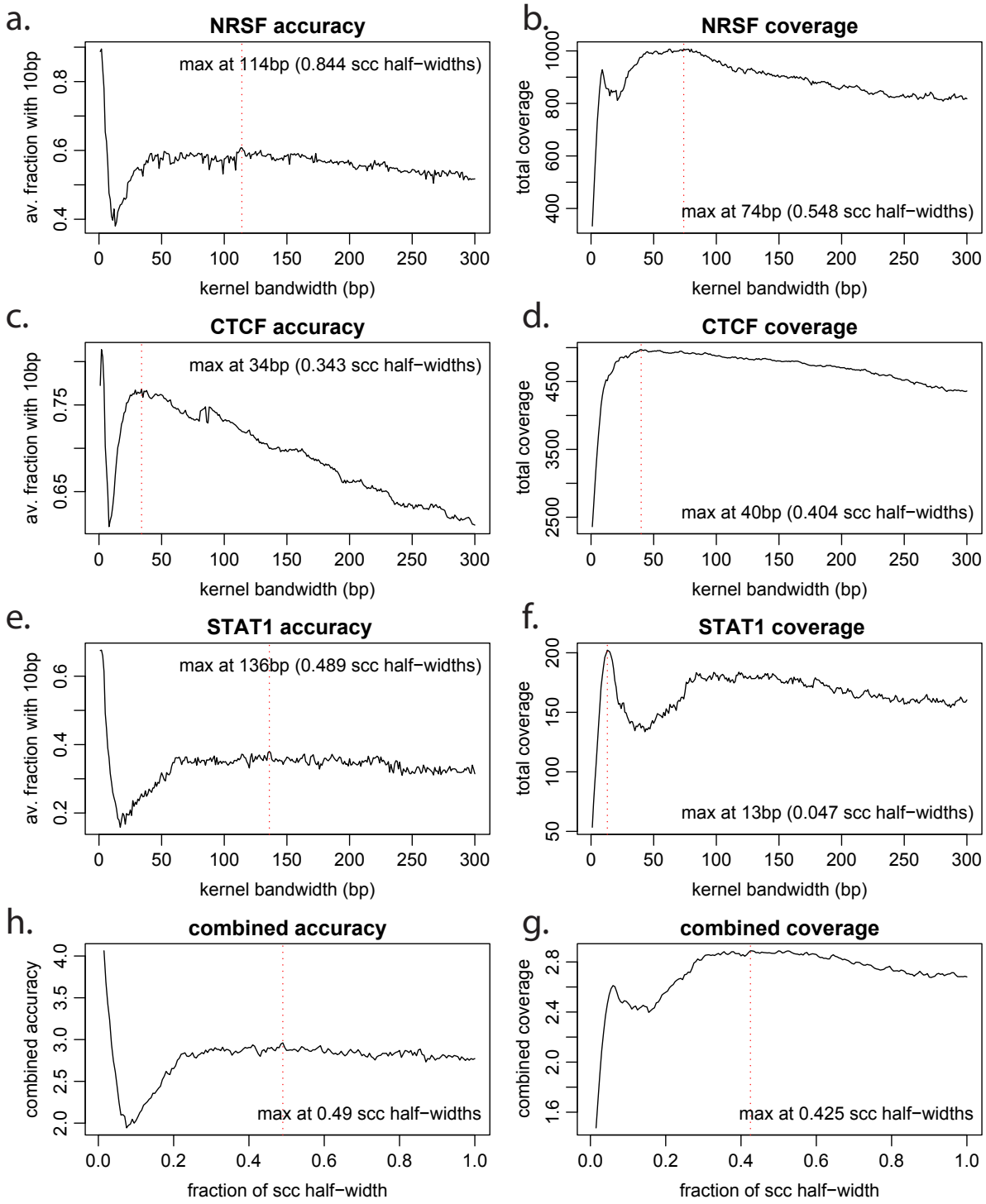
Supplementary Figure 17. Convergence of MSER and interpolated values with increasing number of random dataset subsamples. **a.** The NRSF dataset minimal saturated enrichment ratio (MSER) estimate (y-axis) is shown for an increasing number of random subsamples (x-axis). The standard deviation of the MSER estimate between 60 and 100 subsamples is 0.069, and the mean value is 7.49. **b.** The predicted sequencing depth required to reach 2-fold saturation for NRSF dataset (y-axis) is shown as a function of the number of chained random subsamples (x-axis). The standard deviation of the predicted depth between 60 and 100 subsamples is 2.0×10^4, and the mean is 2.8×10^6.



Supplementary Figure 18. Different methods for estimating enrichment ratio of predicted binding positions. **a-c.** Correlation between NRSF qPCR values and enrichment estimate lower bounds calculated using different background window scales. The enrichment ratio confidence interval for a particular binding position is assessed based on the ChIP and background (input) tag counts within 100bp window around the binding position (see Methods). Because the input tag density tends to be low, counting background tags in larger windows (and normalizing for the size ratio) should provide tighter confidence intervals. The plots show relationship between log10 qPCR enrichment estimate (y-axis) and log10 of the tag-based enrichment estimate lower bound (x-axis), using (**a**) 100bp, (**b**) 500bp, and (**c**) 1000bp background tag-counting window. Even though the enrichment ratios are normalized for the window size difference, using larger background tag windows results in tag-based enrichment estimates larger than qPCR values. **d-f.** MSER interpolation plots using enrichment estimates based on different background window sizes. **g-i.** An analogous interpolation using maximum likelihood estimate (MLE) values. Infinite MLE values were excluded from calculation.



Supplementary Figure 19. Spatial precision of the binding positions predicted by different methods using motif instances with lower confidence thresholds. The plots are analogous to Figure 5 **b-d.** of the main manuscript, but calculated using larger set of sequence motif instances derived using lower confidence thresholds (P-value < 10^{-7}, 10^{-7} and 10^{-5} for NRSF, CTCF and STAT1 respectively, resulting in a total of 2323 NRSF motifs, 5921 CTCF motifs, and 145067 STAT1 motifs)



Supplementary Figure 20. Selecting optimal kernel bandwidth for the MSP method. The performance of the MSP method depends on the kernel bandwidth used to calculate tag density. To determine optimal kernel bandwidth we have evaluated coverage and spatial accuracy of the MSP method for a range of bandwidth values from 1 to 300bp. The coverage was calculated as the area under the coverage curve (*e.g.* Figure 4d of the main text). The spatial accuracy as the mean fraction of binding positions within 10bp of the motif position (*e.g.* Figure 5b-d). The plots show accuracy and coverage dependencies on the kernel bandwidth for NRSF (**a, b**), CTCF (**c, d**) and STAT1 (**e, f**). The plots **h.** and **g.** show combined dependence on the kernel bandwidth, expressed as a fraction of the strand cross-correlation (SCC) half-width (Figure 1d of the main text, Supplementary Figure 1). The coverage and the accuracy values shown on the y-axis in figures h. and g. correspond to a sum of the values from each of the three examined proteins, scaled by the maximum of each individual profile. The plots show that the optimal accuracy is achieved with bandwidth corresponding to 0.49 SCC half-widths, and the optimal coverage is achieved at 0.425 SCC half-widths.

	mean distance	standard deviation	correlation	P-value
NRSF	95.15	14.89	0.10	0.0013
CTCF	74.57	12.82	0.04	0.0125
STAT1	144.25	16.36	-0.04	0.0950

Supplementary Table 1. Characteristics of distances between positive- and negative-strand tag peaks. For each of the three examined datasets, the table shows mean peak separation distance, standard deviation of such distances, the value of Pearson linear correlation coefficient between peak separation distance and the number of tags forming the peaks, and corresponding correlation test P-value. Only positions with more than 10 tags were included in calculations. A small correlation between peak separation and magnitude are observed for NRSF and CTCF; no statistically significant correlation is observed for STAT1.

a. STAT1

number of mismatches \ length of tag alignment	16	17	18	19	20	21	22	23	24	25	26	27
0	330259	432465	459594	334198	197693	126251	113759	98457	81141	56089	446832	8111030
1			141174	354617	442871	473218	374272	318560	175200	81041	181964	1616762
2					3761	51832	142288	155485	109057	58974	61065	258519
3							2	4	2412	3358	5276	12177

b. CTCF

number of mismatches \ length of tag alignment	16	17	18	19	20	21	22	23	24
0	201227	232959	173964	98321	56719	33689	20961	96088	1788287
1			54807	119864	119986	106697	70878	96605	401707
2					1538	15726	35031	39720	67564

Supplementary Table 2. Number of tags within different alignment quality classes for STAT1 and CTCF datasets. Similarly to Table 1 in the main manuscript, the table gives the number of tags whose best alignment falls within each class. The number of mismatches includes the number of nucleotides covered by gaps.

	control type								
FDR	**i**	**p**	**1**	**2**	**3**	**4**	**5**	**7**	**ss**
0.01	2755	9206 (0.078)	2985 (0.012)	2306 (0.0082)	2190 (0.0078)	2031 (0.0061)	2053 (0.0065)	1887 (0.0050)	1044 (<4e-4)
0.005	1879	6656 (0.043)	2660 (0.0096)	2111 (0.0074)	1846 (0.0046)	1843 (0.0045)	1850 (0.0046)	1681 (0.0038)	922 (<4e-4)
0.001	1416	5227 (0.023)	2129 (0.0076)	1679 (0.0038)	1521 (0.0023)	1488 (0.0022)	1375 (<4e-4)	1156 (<4e-4)	718 (<4e-4)

Control type legend:

i	input-based
p	Poisson random model
1 .. 7	randomization with bin size 1 .. 7 bp
ss	randomization of tag strand assignment without altering positions

Supplementary Table 3. Number of statistically significant binding positions under various background models. The table shows the number of NRSF binding positions predicted by the WTD method using different background tag distribution models (columns), and different FDR thresholds (rows). To provide further comparison of randomization-based method with the input-based background model, the table shows in parentheses the input-based FDR level corresponding to the number of peaks returned by the randomization model. The red colors corresponding to over-estimation of the number of positions under a randomization model, blue under-estimation. The Poisson-based randomization model consistently returns larger number of predicted positions than empirical input-based model for the same FDR threshold. Binned randomization models (where clusters of tags within a certain distance are maintained together) can return a number of positions comparable to that determined by input-based model; however, different bin size needs to be used for different FDR thresholds. Finally, a randomization model where only tag strand assignment is altered (positions remain the same), provides more conservative estimates of the number of significant binding positions.