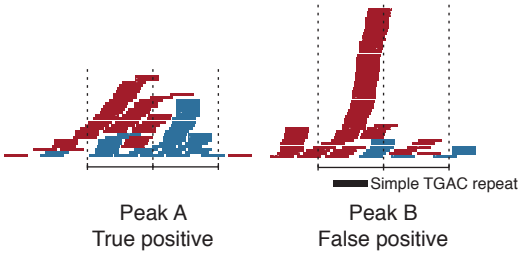
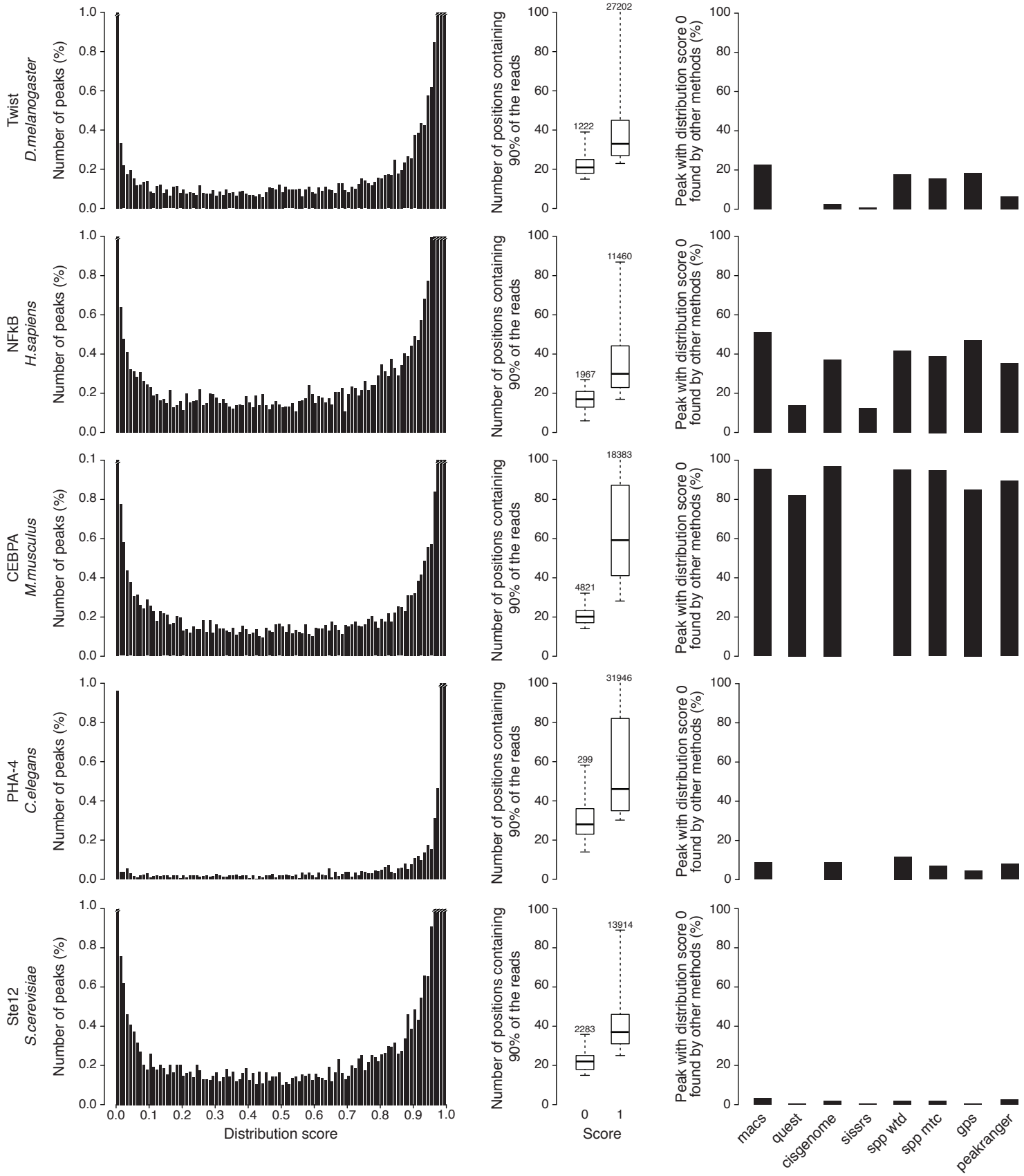
Supplementary Figure 1: Example of false positive peak

Compared to peak A, peak B contains an unbalanced proportion of reads mapped to the positive (red) and negative (blue) strand due to the presence of a simple TGAC repeat that biases read mapping – even when only reads are considered that have a single best mapping position in the genome as done here. These kind of artifacts that are heavily penalized by the model used by peakzilla.



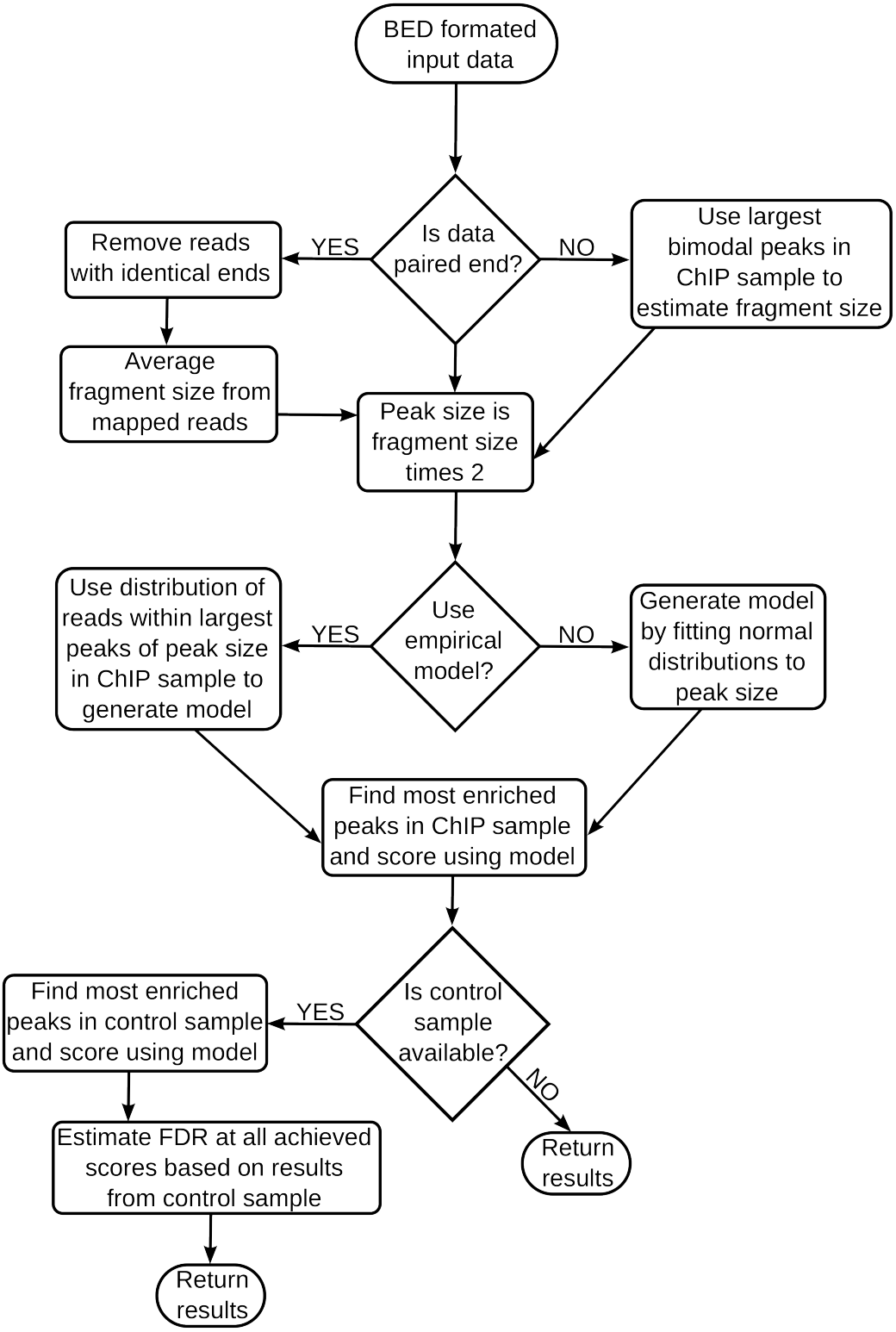
Supplementary Figure 2: Evaluation of peak scoring

Histogram of peak distribution scores for different datasets (left). Peaks with distribution scores (multiplicative factors [0…1]) below 1 are penalized in their final score, to reflect that their read distributions do not fit to the specific double distribution of the model. Penalized peaks with a distribution score of 0 contain substantially less diverse sequence reads (fewer genomic positions containing 90% of read counts) than non-penalized peaks with a distribution score of 1 and are thus more likely to be false positives as described in supplementary figure 1 (middle). Number of estimated false positive peaks by peakzilla found by other methods (right).



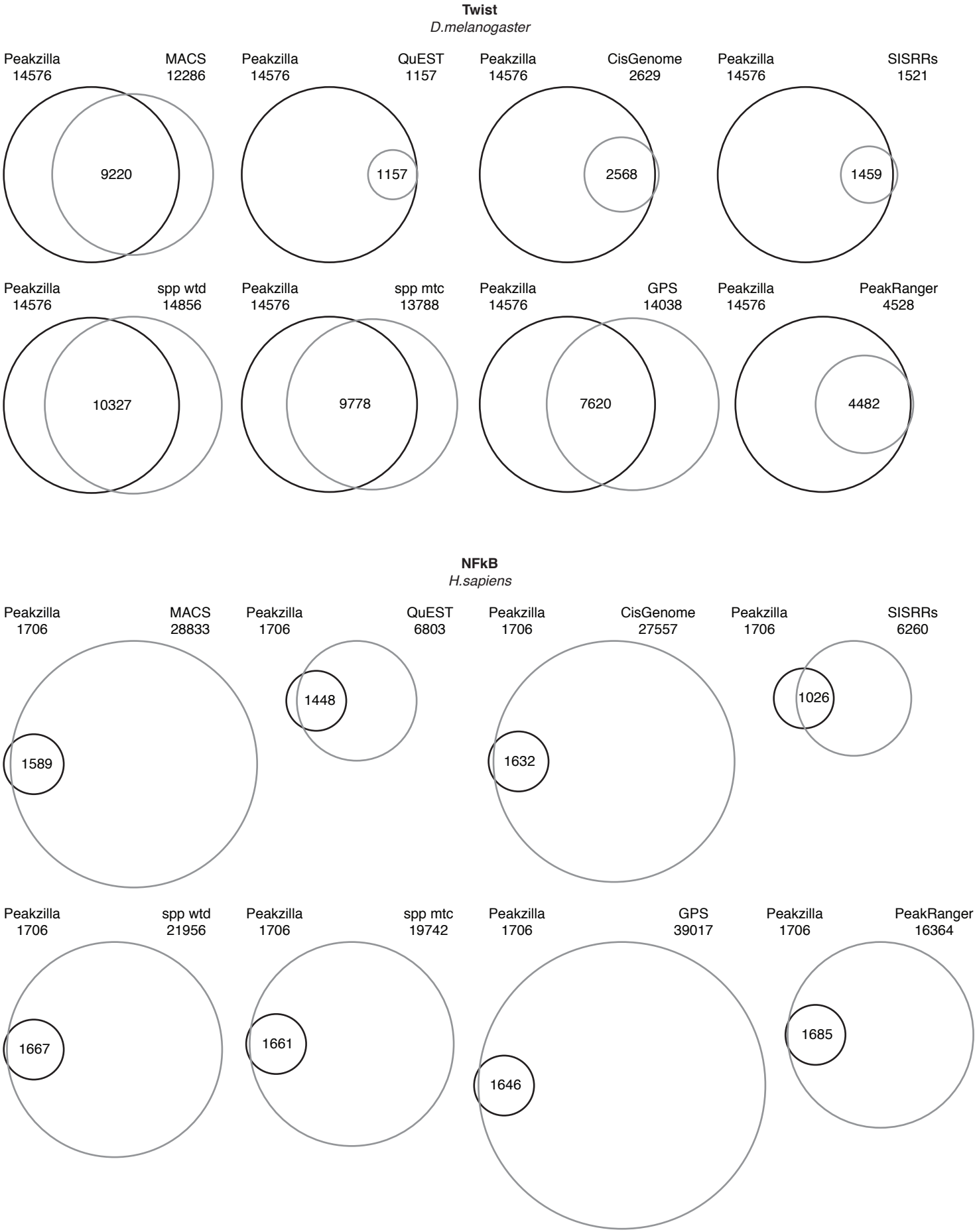
Supplementary Figure 3: Flowchart of the algorithm

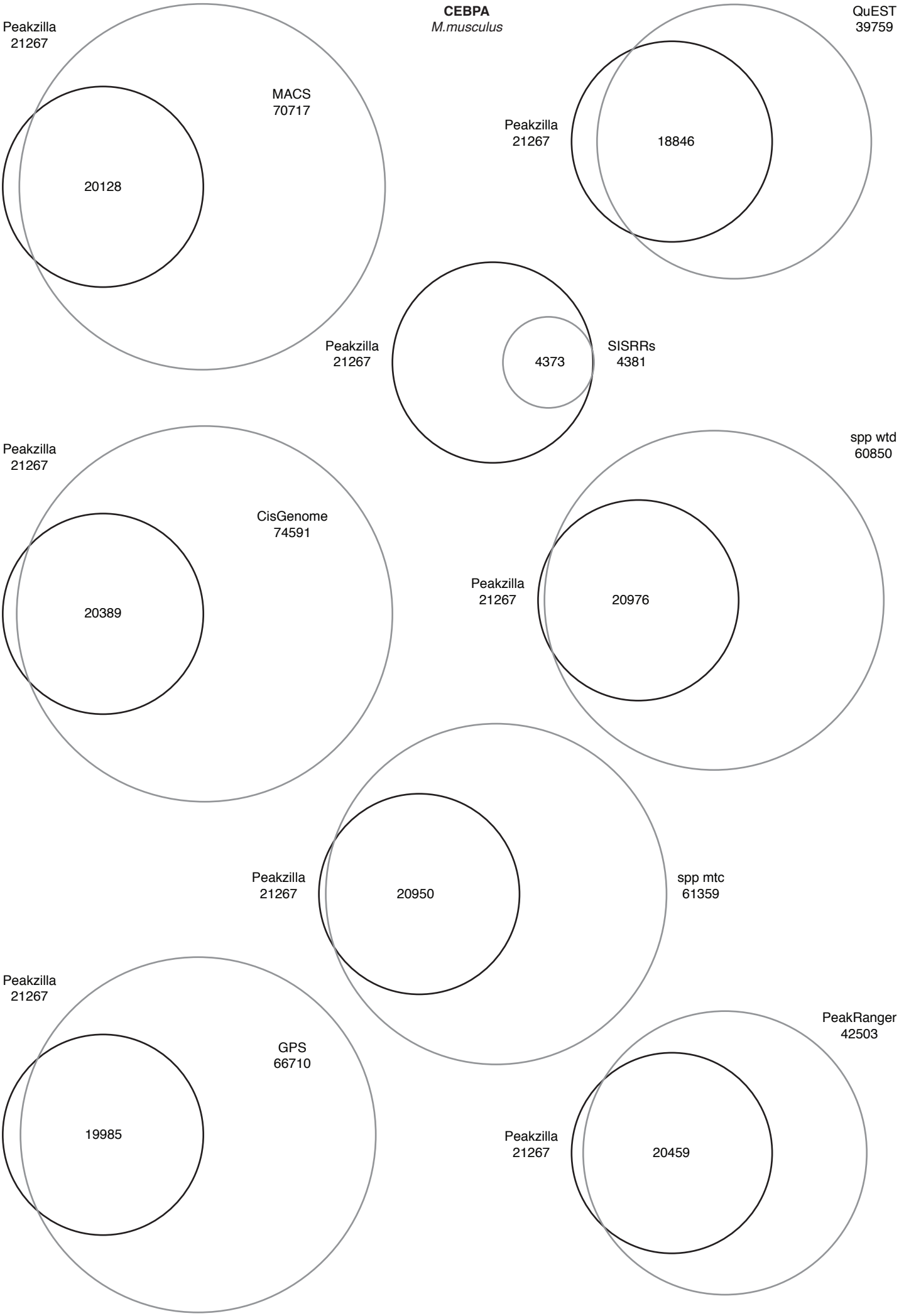
The flowchart provides a high-level description of every data processing step from input to output. Operations are described in the squares with rounded corners while within the diamonds experiment dependent choices are described.



Supplementary Figure 4: Overlap of Peakzilla peaks with peaks from different methods

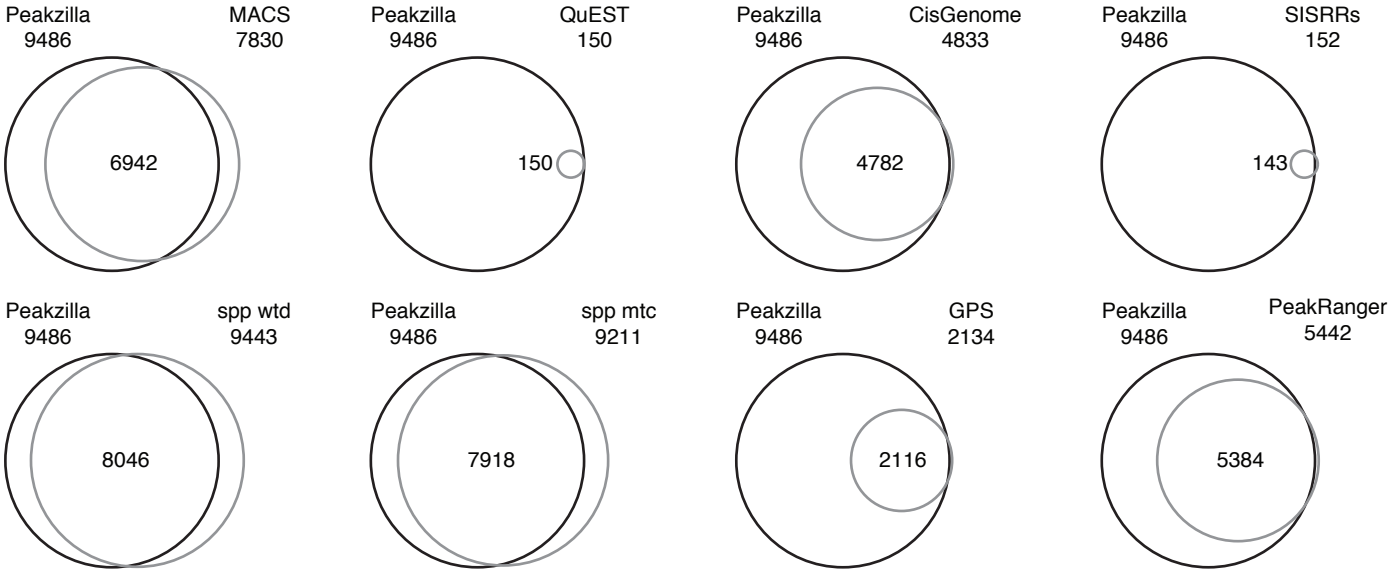
Venn diagrams showing for Venn diagrams for each dataset showing the number of peaks found by each method and their overlap with the peakzilla peaks. Note that while the overall number of peaks identified varies widely due to different thresholds the overlap remains large.




CEBPA
M.musculus
Peakzilla
21267
MACS
70717
20128
QuEST
39759
Peakzilla
21267
18846
Peakzilla
21267
4373
SISRRs
4381
Peakzilla
21267
CisGenome
74591
20389
spp wtd
60850
Peakzilla
21267
20976
Peakzilla
21267
20950
spp mtc
61359
Peakzilla
21267
GPS
66710
19985
PeakRanger
42503
Peakzilla
21267
20459

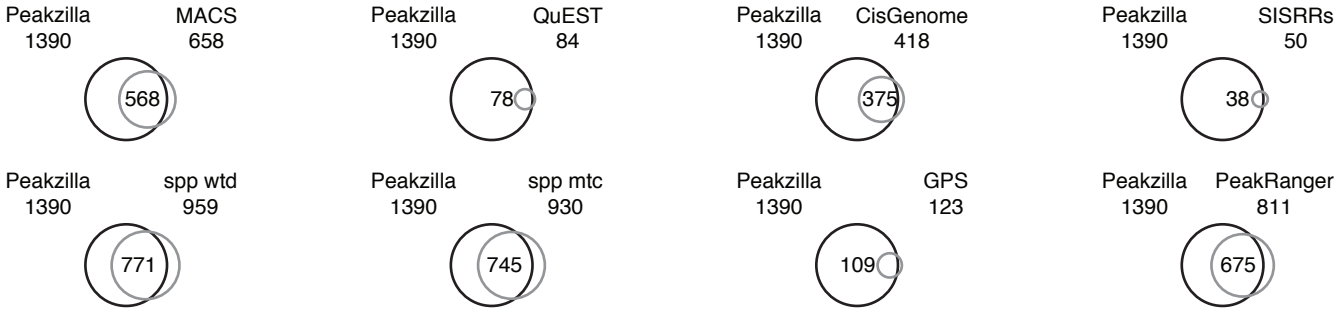
PHA-4

C.elegans



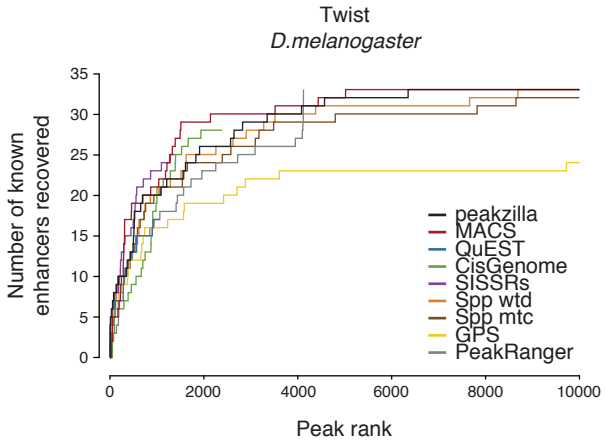
Ste12

S.cerevisiae



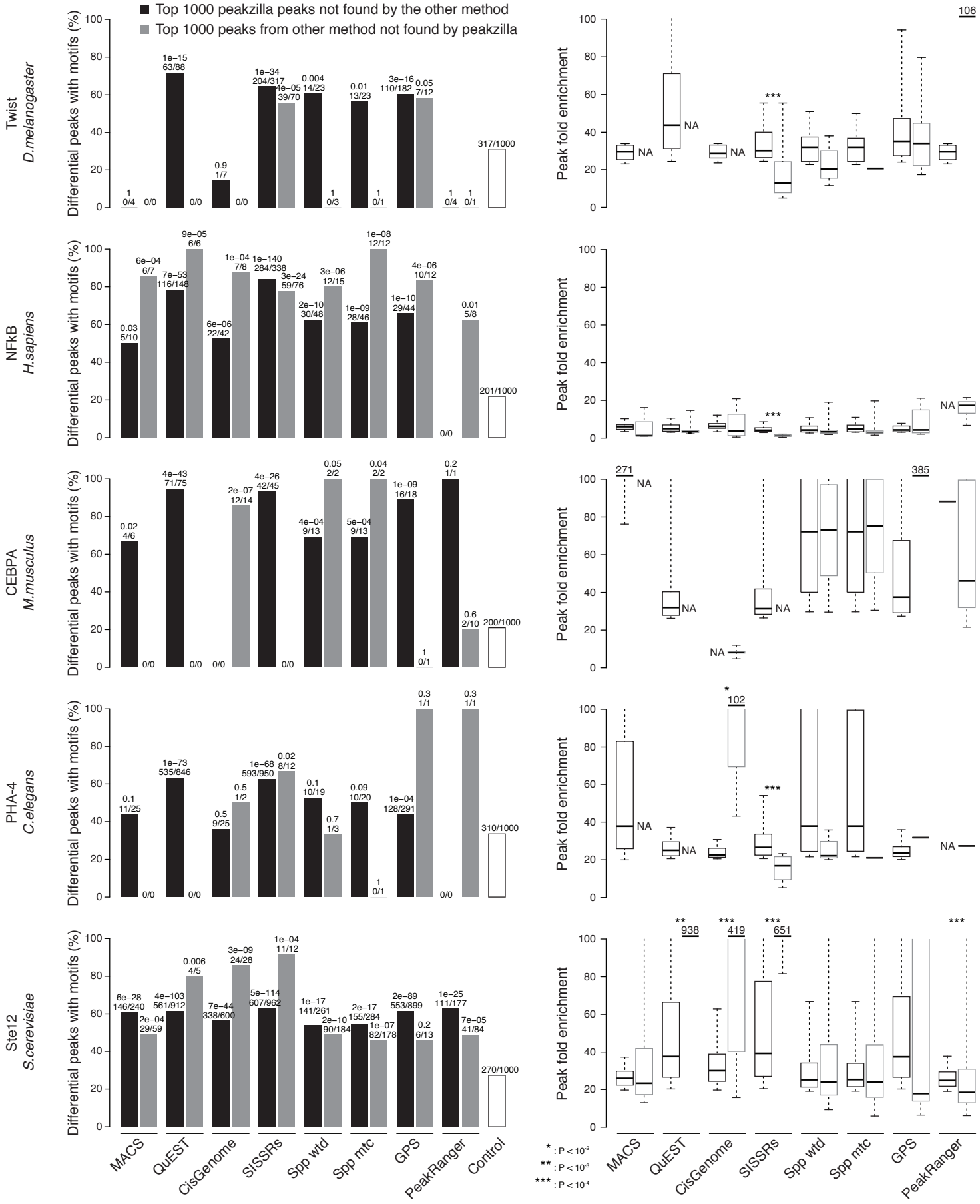
Supplementary Figure 5: Recovery of known enhancers

Cumulative number of known Twist enhancers overlapping with the peaks called by each method and ranked by score. Note that each method has a different total number of peaks called. All known Twist enhancers are identified by all methods, except for four and three enhancers that are not found by QuEST and GPS. Note that each method has a different total number of peaks called.



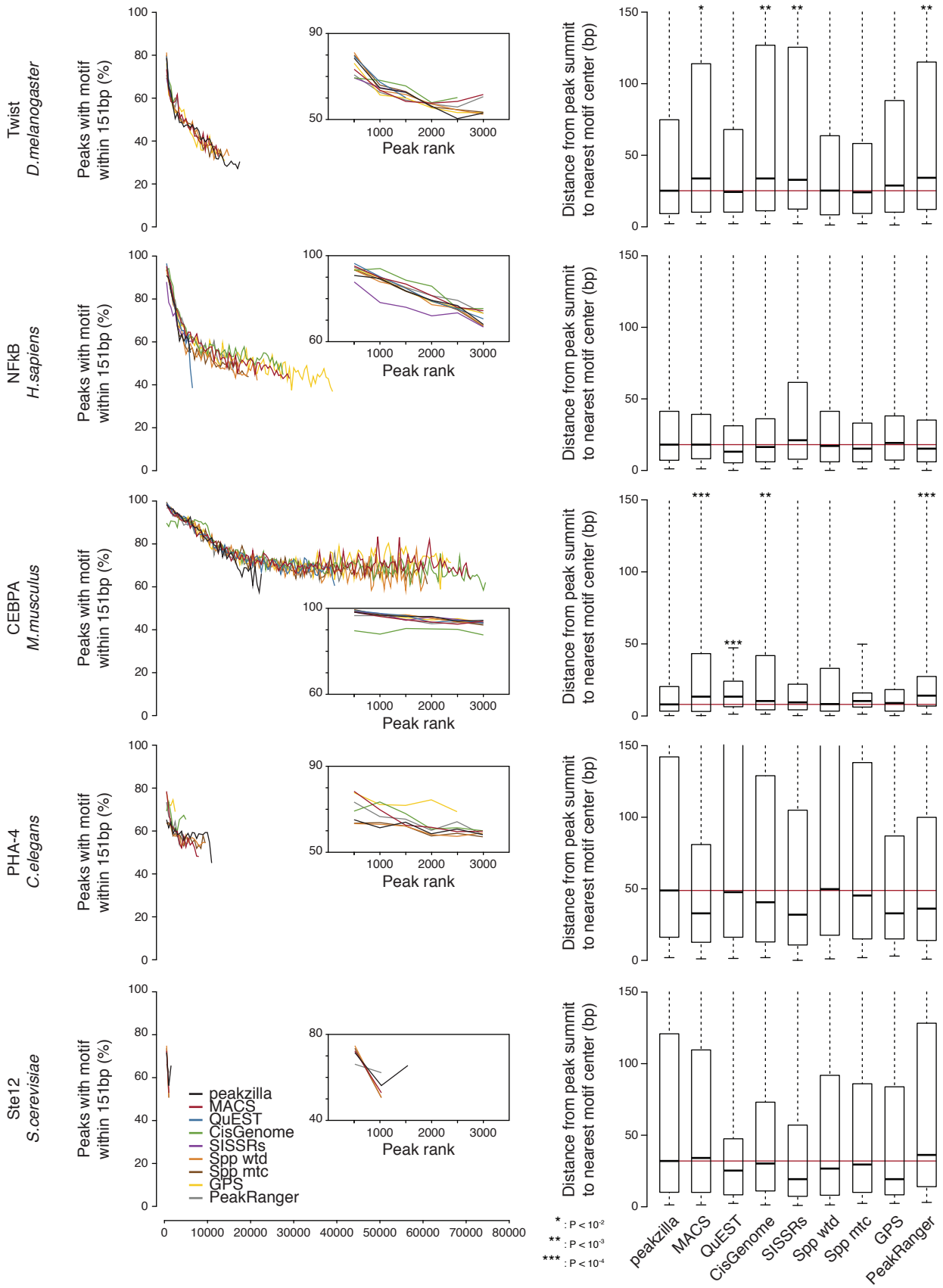
Supplementary Figure 6: Evaluation of differential peaks between Peakzilla and other methods

Enrichment of motifs in differential peaks (left). Binomial p-values of enrichment over control and number of differential peaks that contain a motif is shown on top of the bars. Fold enrichment values of differential peaks and associated Wilcoxon p-values (NA: no peak) (right; the Twist data [top row] are repeated from main Figure 2A & B to allow direct comparisons). For PHA-4 and Ste12, although in 4 cases, the median fold enrichment of peaks missed by peakzilla is high, it comes from only very few peaks (<28).



Supplementary Figure 7: Precision of peaks

The precision of locating peak summits (= TFBS) is estimated as the number of peaks with the corresponding TF motif within 151 bp around the summit (barplots) and the distance of the summits to the nearest motif (boxplots) of the best 500 peaks. Note that the overall fraction of peaks that contain a motif is comparable only for a subset of peaks in the PHA-4 dataset with lower rank GPS appears to outperform other peakfinders.



Supplementary Figure 8: Number of peaks split using different methods

Percentage of peaks from one method (rows) overlapping the peaks from another method (columns). The percent of peaks that are split is indicated by the shading. Overall these results show that the transcription factor binding regions identified by MACS QuEST, CisGenome and PeakRanger are frequently identified as multiple distinct binding sites by other algorithms.