

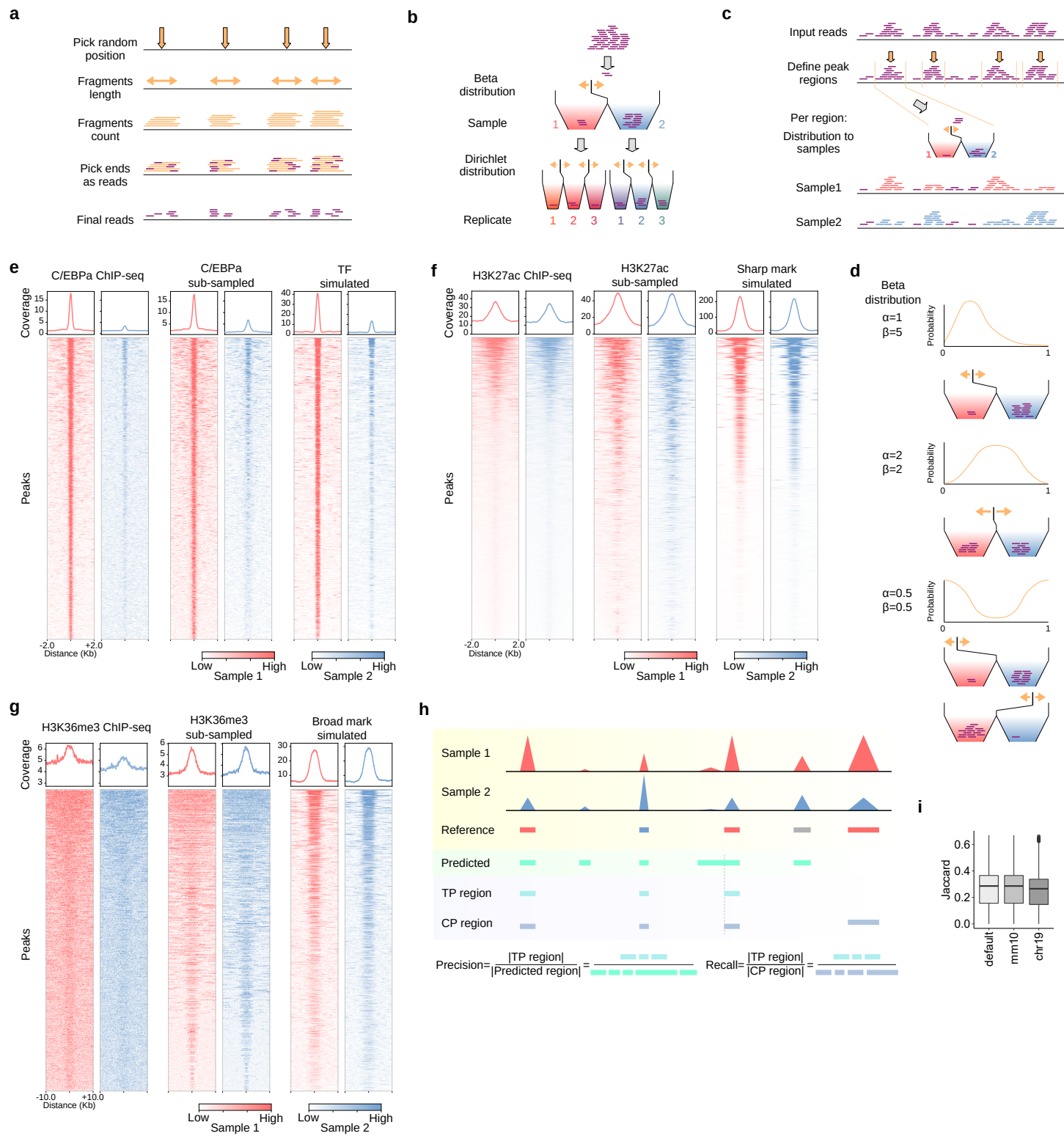


Figure S1: Simulation and sub-sampling details

a, Overview of DCSsim *in-silico* simulation. A defined number of random positions was picked within the provided genome sequence. The individual fragment length and the number of fragments per position were drawn from respective distributions (see methods). For the final reads, 3' or 5' ends of the fragments were randomly selected. **b**, Reads were distributed to samples and replicates via beta- and Dirichlet distributions, respectively. **c**, Scheme of DCSsub sub-sampling. Starting from aligned sequence reads, reads were distributed to samples and replicates as shown in b. This was done for all predefined peak regions indicated by orange arrows. The reads outside these regions, reduced by a user defined percentage, were randomly distributed to the two samples. **d**, Details of the gating strategy. Per region, the position of the gate was drawn from a beta distribution defined by the alpha and beta values. Alpha 1 and beta 5: high probability of sorting most reads into sample 2 (top). Alpha 2 and beta 2: most cases result in a nearly equal number of reads in both samples (middle). Alpha and beta of 0.5: the majority of reads sorted to either sample 1 or sample 2 (bottom). **e**, Heatmaps and profile plots for all peak regions of ChIP-seq experiments for the TF C/EBPa and after its knockdown (left), DCSsub sub-sampling of a 100:0 scenario from this data (middle) and the simulation with DCSsim for TF peak shape and 100:0 regulation (right). **f**, Heatmaps and profile plots from genuine ChIP-seq data (left), DCSsub sub-sampled (middle) and DCSsim simulated data (right) from an H3K27ac histone mark ChIP-seq experiment before and after perturbation of C/EBPa, with a 50:50 regulation scenario. **g**, Heatmaps and profile plots from genuine ChIP-seq data (left), DCSsub sub-sampled (middle) and DCSsim simulated data (right) from an H3K36me3 histone mark ChIP-seq experiment before and after perturbation of SETD2. **h**, Schematic representation of the evaluation procedure. Reference regions were compared to the predictions from the DCS tools. Colors of the reference regions indicate the origin of the higher signal (red sample 1 and blue sample 2). True positive (TP) and condition positive (CP) regions were used to calculate precision and recall. **j**, Influence of the MACS2 -g parameter. Boxplots resemble the Jaccard similarity between the resulting bed files and the simulated reference (default: 2.7e9, mm10: 1.87e9 and mm10 chr19: 5,8205,856).

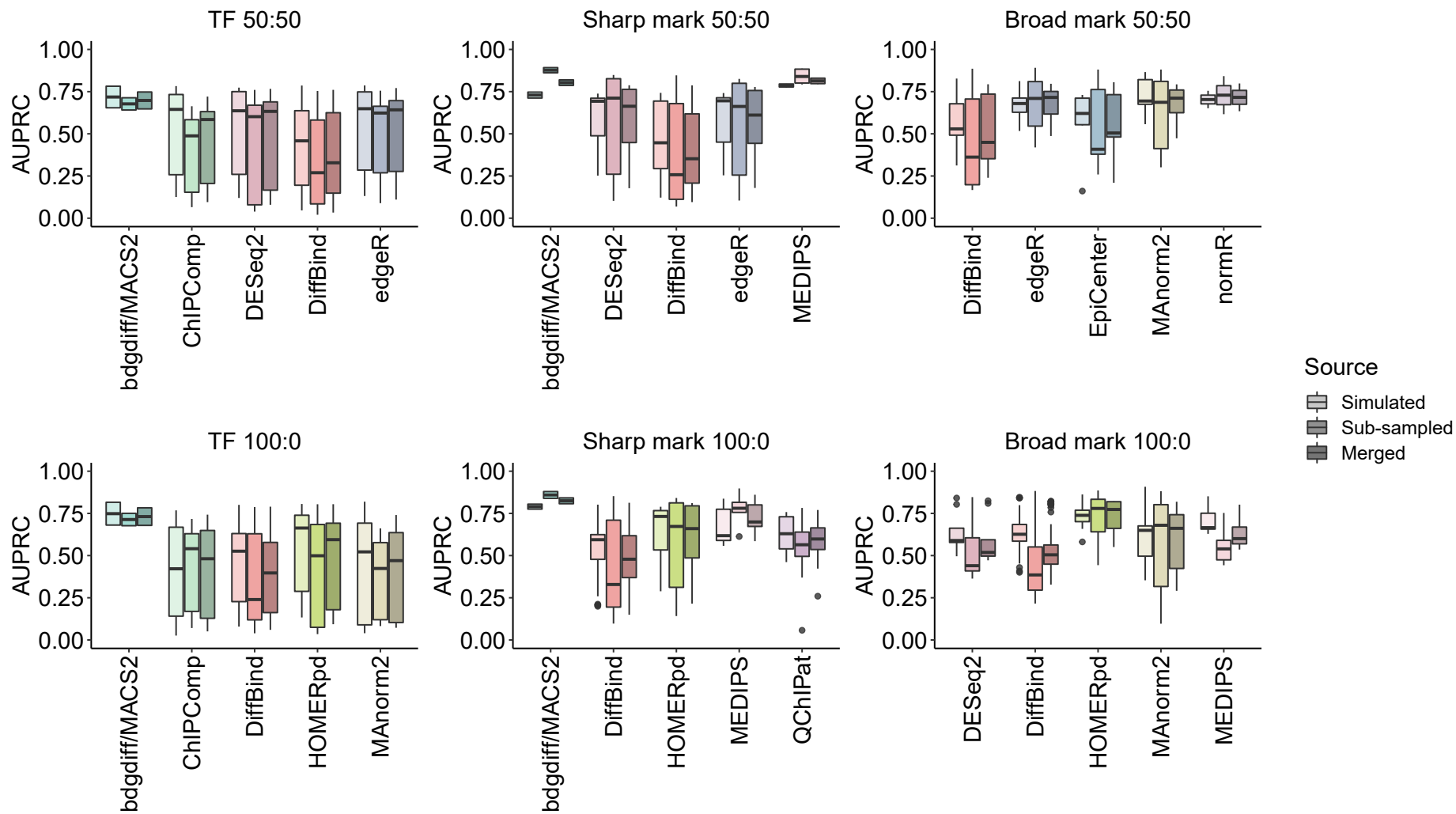
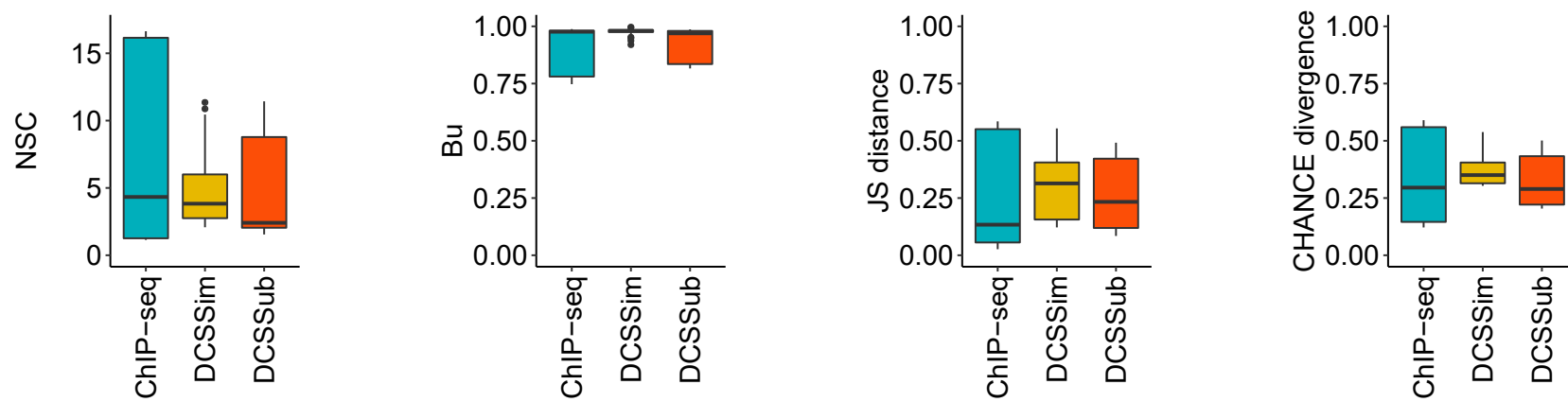
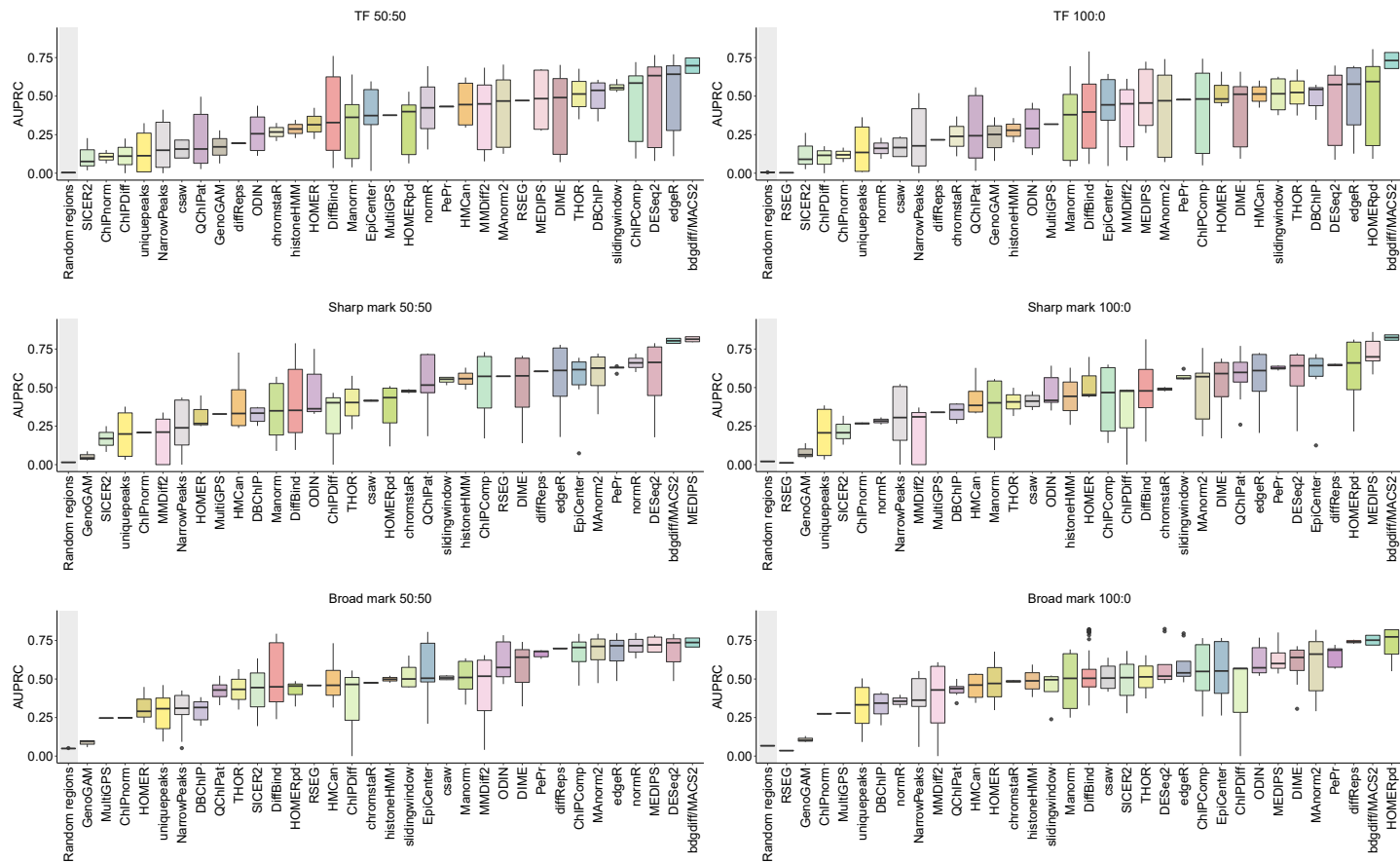
a**b**

Figure S2: AUPRCs of DCS tools based on simulated, sub-sampled or merged data

a, AUPRCs of the best performing tools per test scenario for simulated or sub-sampled data only and the merged AUPRCs. Rows showing TFs (left), sharp (middle) and broad marks (right) and the columns 50:50 (top) and 100:0 (bottom) regulation. Box plot limits, 25% and 75% quantiles; center line, median; whiskers, 1.5x interquartile range. **b**, Assessment of signal-to-noise metrics is shown by normalized strand coefficient (NSC), background uniformity (Bu), Jensen-Shannon (JS) distance and CHANCE divergence for real, DCSSim and DCSSub data.

a



b

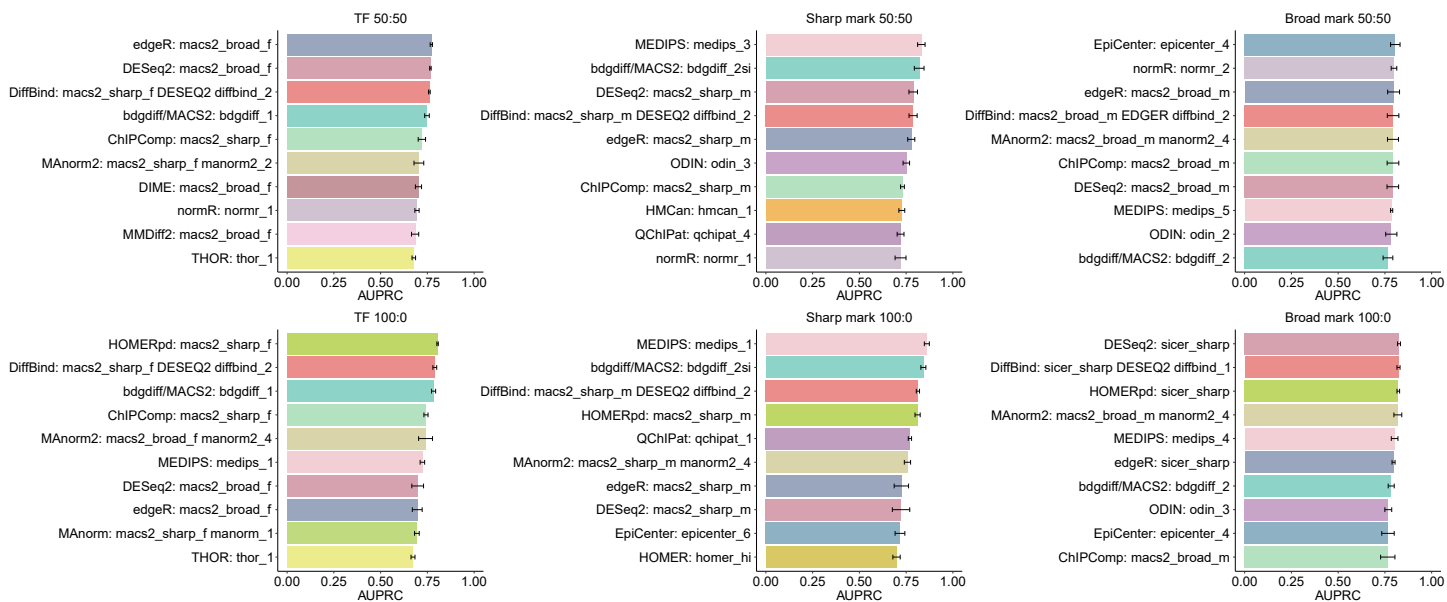


Figure S3: Assessment of DCS tools per peak shape and regulation scenario based on AUPRCs

a, AUPRC values of DCS tools for each of the six shape and regulation scenarios ranked by median AUPRC plus the respective random predictions on the left (gray background). Box plot limits, 25% and 75% quantiles; center line, median; whiskers, 1.5x interquartile range. **b**, AUPRCs of the respective best parameter setups for the top ten DCS tools per scenario. Labels indicate DCS tool and peak caller and/or parameter setup identifier (for details see Table S3) if applicable; whiskers, standard error of the mean.

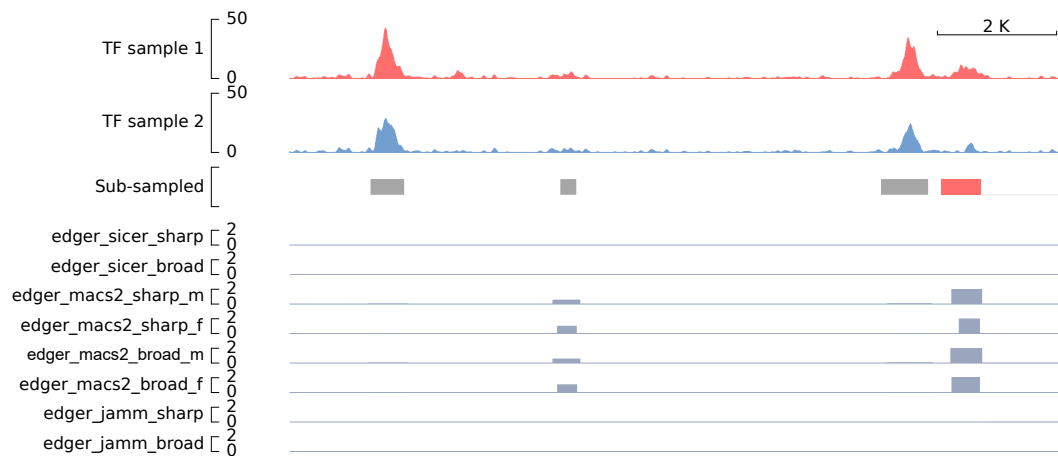
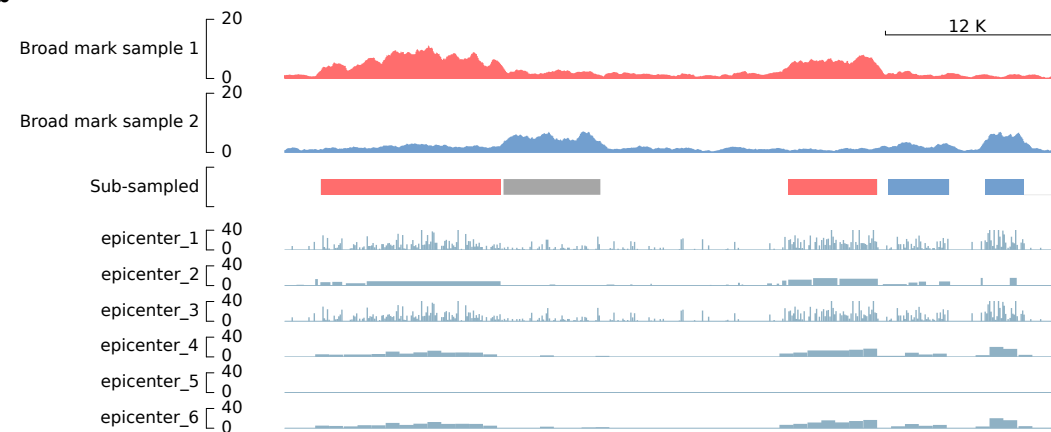
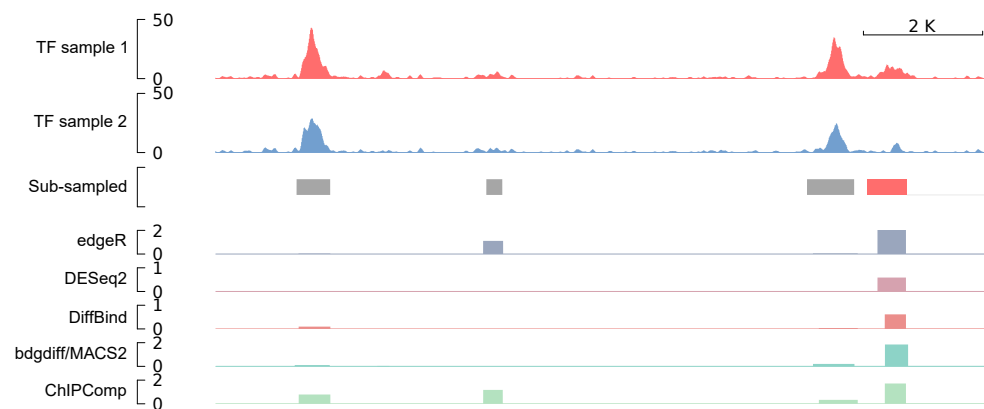
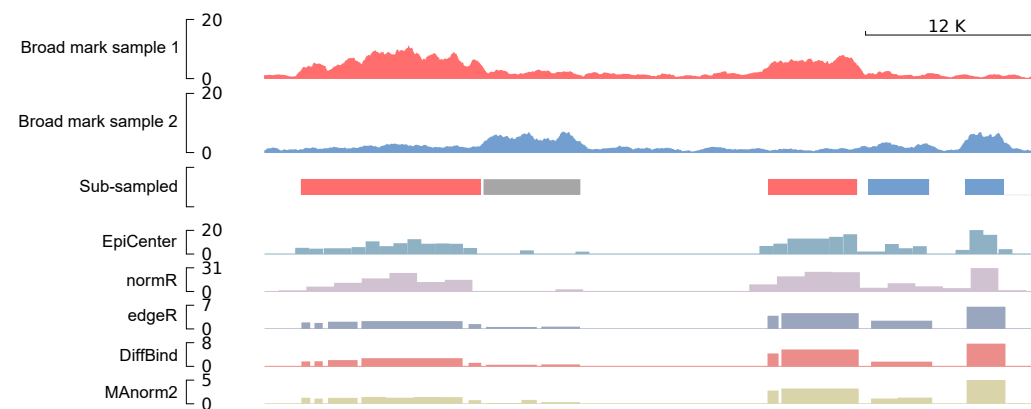
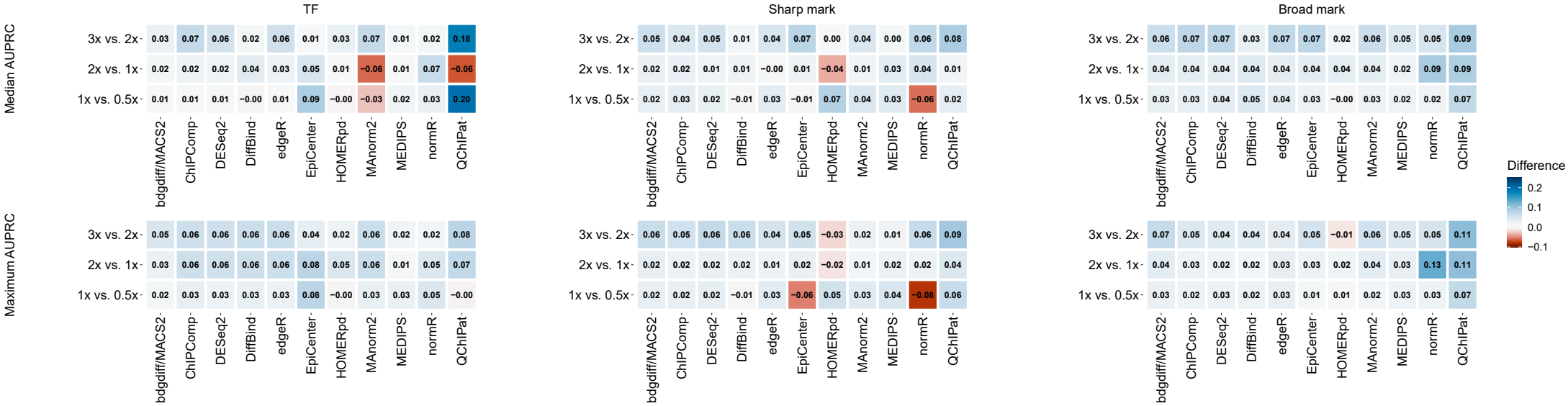
a**b****c****d**

Figure S4: Coverage and top tool predictions of example DCS regions

Coverage of one replicate of sub-sampled reads for sample 1 (row 1) and sample 2 (row 2) with the respective reference regions (row 3) in two example regions. The color of row 3 indicates up-regulation in sample 1 (red), up regulation in sample 2 (blue) and no difference (gray). The height of predicted regions represents the $-\log_{10}$ of p-value, adjusted p-value or FDR or the score derived from the respective DCS tool. Therefore, higher bars mean the respective DCS tool has more confidence in that region to be differential between the samples. **a**, Rows 4 to 11 show the predicted regions from all applied parameter setups of edgeR within this replicate of TF sub-sampled peaks (peak callers SICER2 and JAMM did not return peak-regions for this example region). **b**, Rows 4 to 9 show the predicted regions from all tested parameter setups of EpiCenter within this replicate of broad mark sub-sampled peaks (setup 5 did not result in predictions for the region shown, details of parameter setups: Table S3). **c**, Rows 4 to 8 show predicted regions from the best parameter setups of the top 5 DCS tools for TF and **d**, broad mark data and 50:50 regulation. Representative regions for sub-sampled (**a**, **c**) C/EBPa (TF) and (**b**, **d**) H3K36me3 (broad mark) in 50:50 regulation scenarios are shown.

a



b

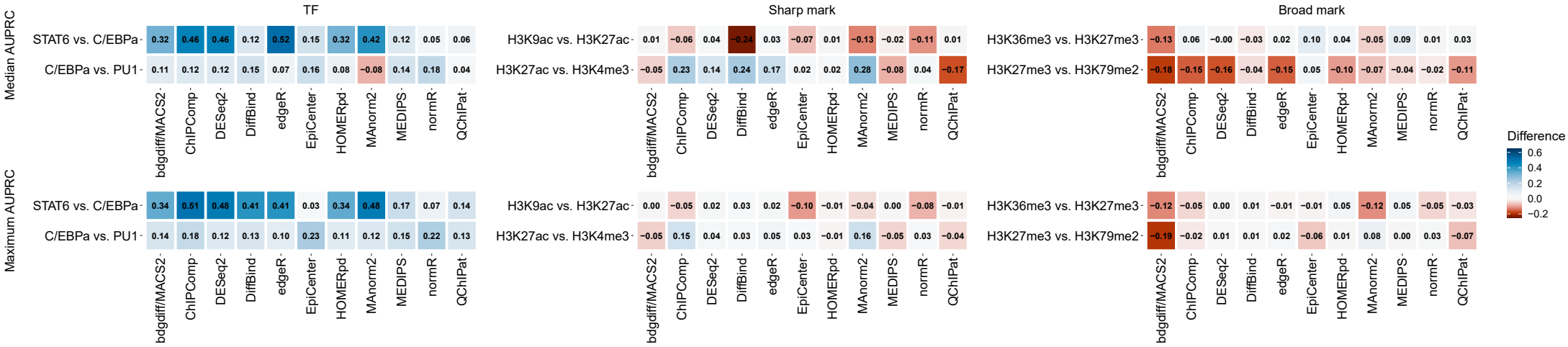


Figure S5: **Signal-to-noise ratio affects AUPRC**

Heatmaps showing the difference of AUPRC between lower and higher FRiP and higher and lower background signal for the top 11 DCS tools. **a**, The difference in AUPRC for simulated data with decreasing background, 3x versus 2x background, 2x versus 1x background and 1x versus 0.5x background. The upper row shows the differences in the median AUPRC and the lower row the differences in the maximum AUPRC per DCS tool. The left column shows the differences for TFs, middle column for sharp marks and the right column for broad marks. Colors indicate higher AUPRC (blue) for the sample with lower background, lower AUPRC (red) for the sample with lower background and now difference (white). Rounded values are shown inside tiles.

b, Heatmaps show differences in AUPRC for the top 11 DCS tools. The analyzed data is derived from sub-sampled ChIP-seq experiments from additional TFs and histone marks with different FRiP. Colors according to the differences in AUPRC between the samples with lowest vs. intermediate and intermediate vs. highest FRiP (lower AUPRC in lower FRiP sample: red, higher AUPRC in the higher FRiP sample: blue, no difference: white). The left column shows the differences for TFs, middle column for sharp marks and the right column for broad marks. The upper row shows the differences in the median AUPRC and the lower row the differences in the maximum AUPRC per DCS tool.

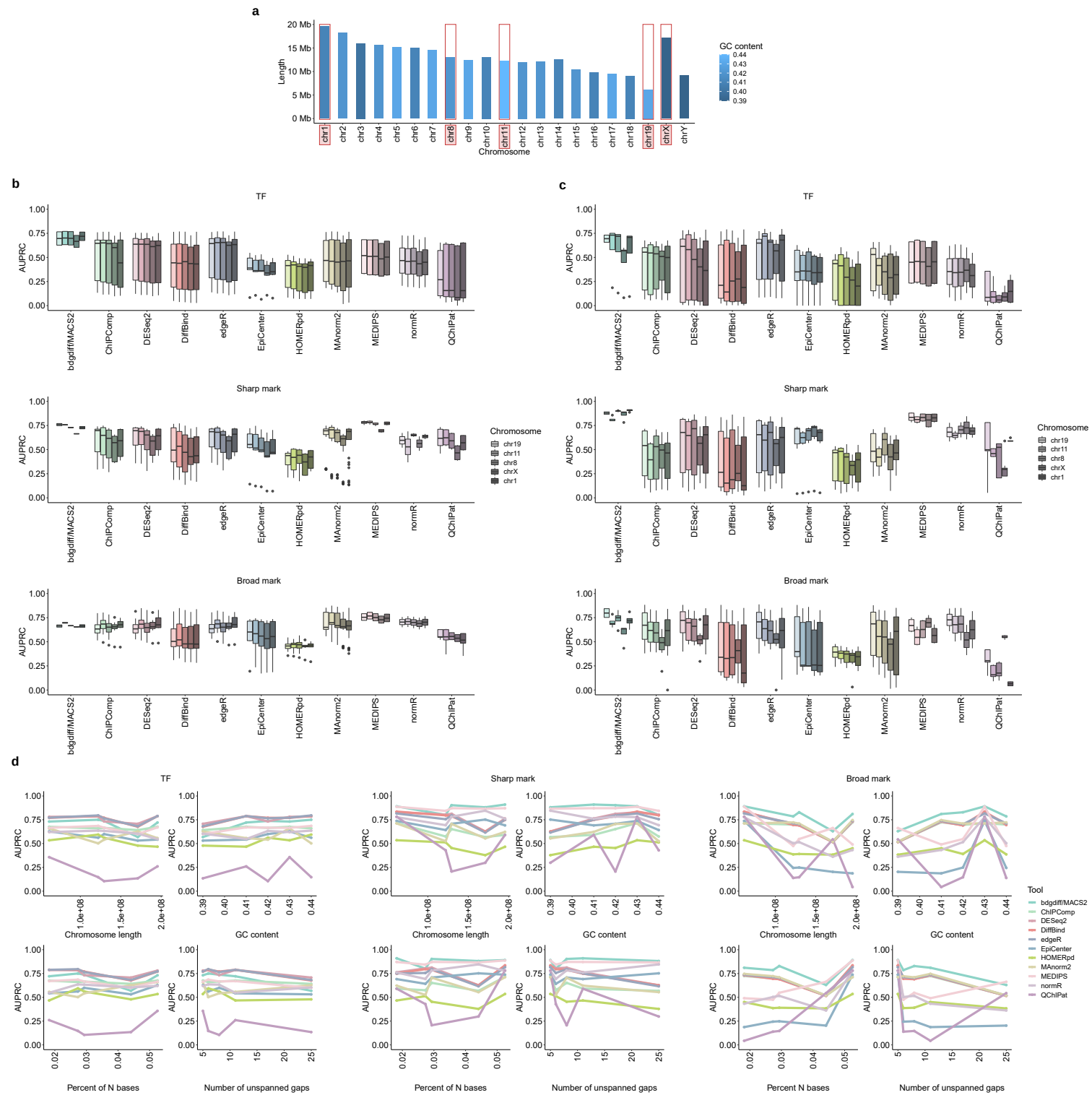


Figure S6: Effects of chromosome characteristics and ChIP-seq signal distribution on AUPRC

a, GC content and chromosome length in the mouse genome assembly (mm10). The investigated chromosomes are highlighted in red. **b**, AUPRCs from simulated data of the top 11 DCS tools (based on AUPRC for the initial six shape and regulation scenarios) for five chromosomes of mm10 (chr1, chr8, chr11, chr19 and chrX) for TFs, sharp and broad marks. **c**, AUPRCs from sub-sampled data of the top 11 DCS tools for C/EBPa (TF), H3K27ac (sharp mark) and H3K36me3 (broad mark) data. Chromosomes per DCS tool are ordered by length, from short to long. Box plot limits, 25% and 75% quantiles; center line, median; whiskers, 1.5x interquartile range. **d**, Line graphs show the AUPRC for the best performing parameter setup per DCS tool (for the top 11 tools from the initial analysis based on chr19) for chromosome length, GC content, percent of unknown (N) bases and the number of unspanned gaps of the five chromosomes from the mm10 assembly.

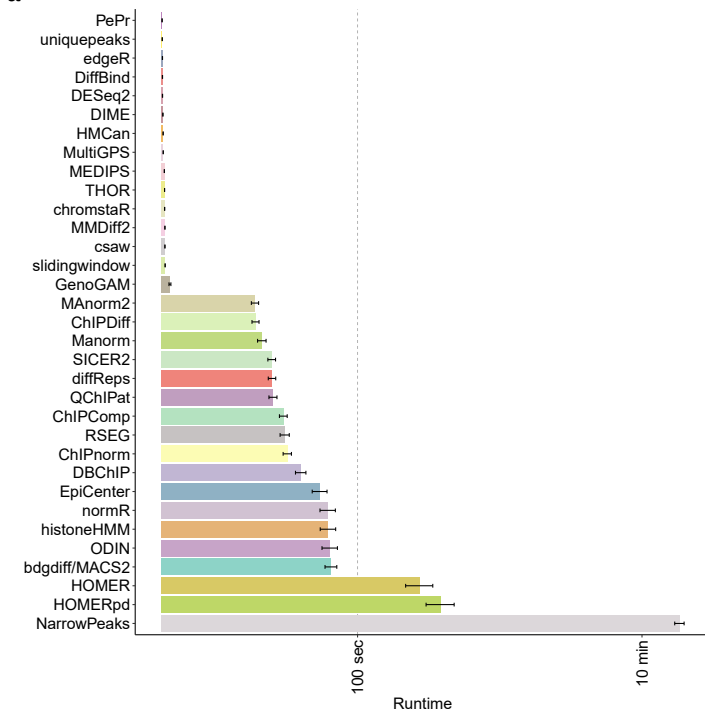
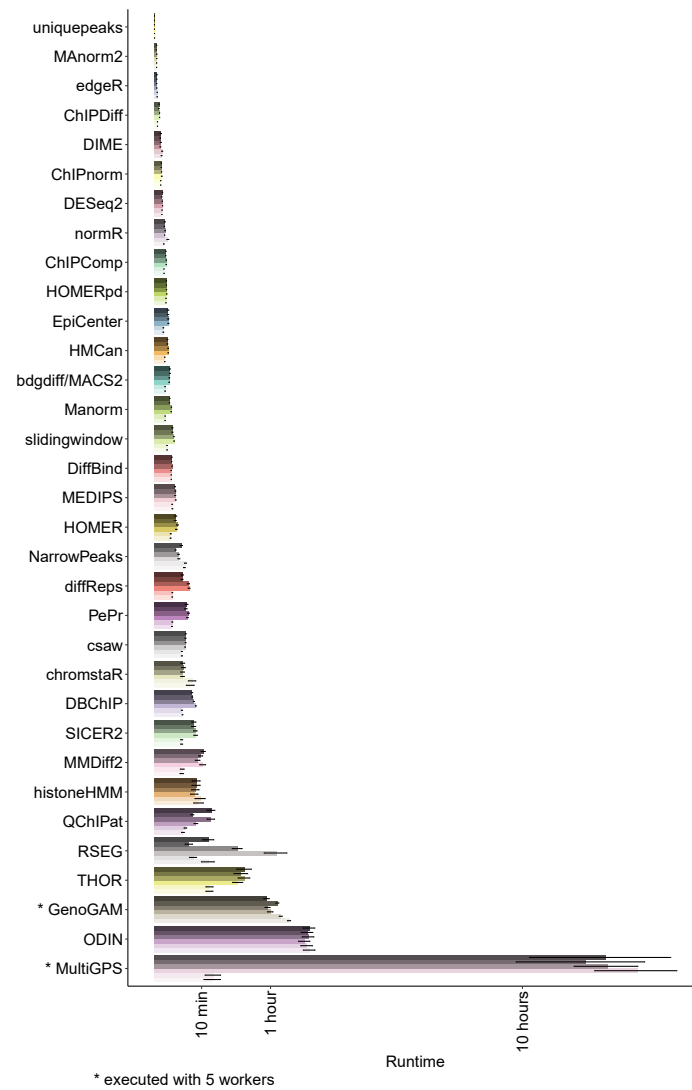
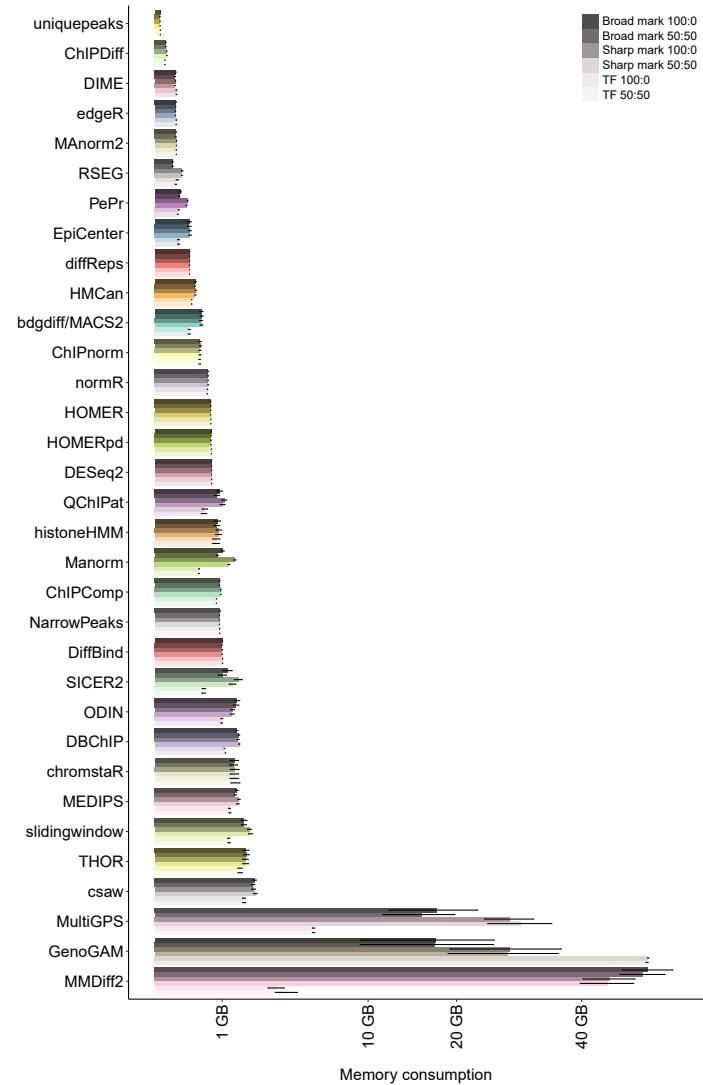
a**b****c**

Figure S7: **Preparation time and runtime plus memory requirements based on individual scenarios**

a, Average preparation time to reformat and process input files depending on the respective DCS tool over the six test scenarios. **b**, Runtime and **c**, memory consumption of all tested DCS tools per peak and regulation scenario. Due to their extensive runtimes GenoGAM and MultiGPS were executed with 5 workers. Whiskers indicate standard error of the mean.

Figure S8: **Performance measures and ranking of DCS tools per scenario**

a, Heatmap of DCS tool performance. The first two columns represent the average AUPRCs of all parameter setups per DCS tool and the top AUPRCs the best performing parameter setup per DCS tool for the six tested peak shape and regulation scenarios and their combinations. All other metrics are shown as average of all parameter setups per DCS tool over all test scenarios. In columns 3 and 4 the accuracy profiles and the stability in form of standard deviations between AUPRCs of the simulated and sub-sampled replicates and the number of NA results summarizing all failed and faulty execution runs or runs with empty outputs are shown. The runtime columns consist of the preparation time, time required for the DCS prediction on average and the runtime per peak-shape as the number of reads to process differ. The last columns show memory consumption per DCS tool over all test sets and for TFs, sharp and broad marks only. Tools were ordered by their mean DCS score over all test sets. **b**, The best performing parameter setups per DCS tool were ranked by DCS score for the respective scenario and scenario combinations (for details see Tables S2, S3, S4 and S5). Colored boxes indicate the applied peak caller if applicable and if default, default with custom windows or adjusted parameters were used.