

# Supplementary materials for “Increased peak detection accuracy in over-dispersed ChIP-seq data with supervised segmentation models”

datasets	type of ChIP-Seq experiment	number of folds
H3K36me3_AM_immune	H3K36me3 (broad peaks)	10
H3K36me3_TDH_immune	H3K36me3 (broad peaks)	4
H3K36me3_TDH_other	H3K36me3 (broad peaks)	4
H3K4me3_PGP_immune	H3K4me3 (sharp peaks)	10
H3K4me3_TDH_immune	H3K4me3 (sharp peaks)	10
H3K4me3_TDH_other	H3K4me3 (sharp peaks)	10
H3K4me3_XJ_immune	H3K4me3 (sharp peaks)	10

Table 1: Summary of the number of folds in the cross-validation procedure by dataset. Two of the seven labeled histone mark datasets, i.e. H3K36me3\_TDH\_immune & H3K36me3\_TDH\_other, can be considered small datasets as they include biological replicates from four independent genomic chunks. In order to satisfy the assumption of independence between the training and test set in the cross-validation, we could not exceed 4-fold for both of them.

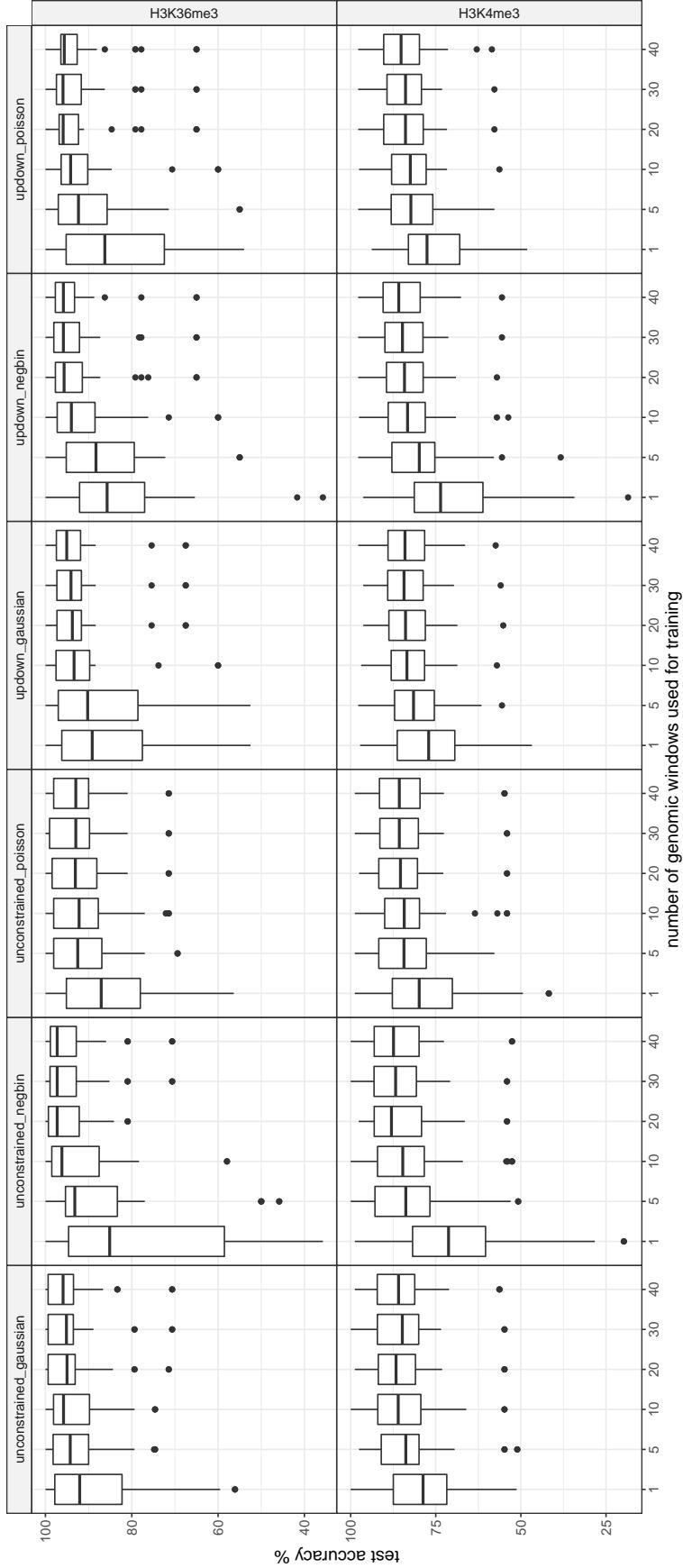


Figure 1: To determine how many labels are necessary to learn a model close to the maximum accuracy we used the cross-validation procedure described in the manuscript. Each test fold is assessed for a variable number of labeled genomic windows from the training folds (from 1 to 40). In all datasets, 10 genomic windows, with on average 5 labels per window, are enough to learn penalties and associated segmentation close to the model-specific maximum accuracy. Note that in addition to improving the average accuracy on test folds, a few genomic windows are enough to effectively reduce the variance of the results.

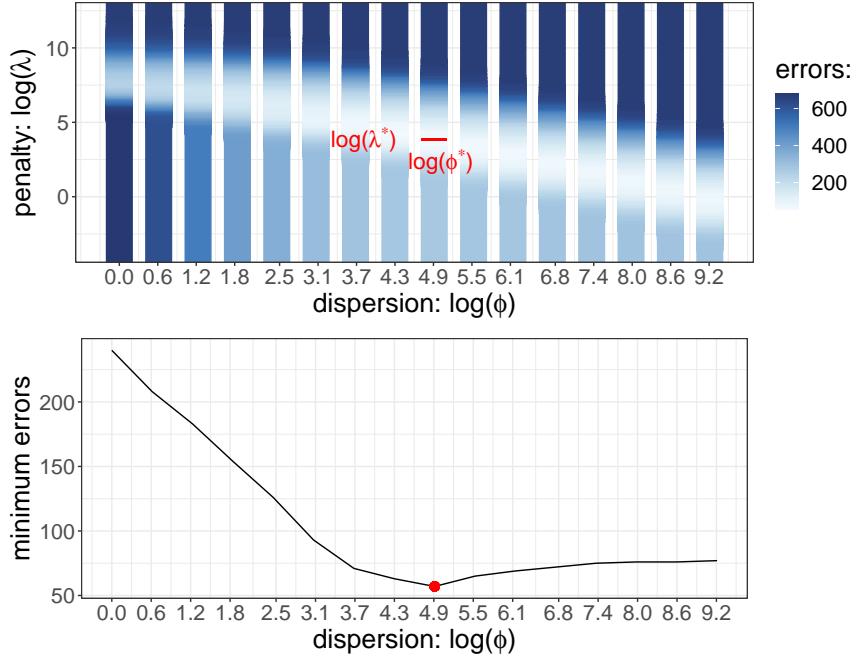


Figure 2: **(Top)** Visualization of  $\sum_{m \in \text{training set}} E_m(\phi \in \Phi, \lambda)$ . The global minimum error (57), shown in red ■, is reached for  $\lambda^* = 46.86$  and  $\phi^* = 135.94$ . **(Bottom)** For each  $\phi_i$ , i.e 16 values evenly placed on the log scale between 1 and 10000, the minimum error of  $E_m(\phi_i, \lambda)$  has been plotted. We can see the errors growing constantly at the left en right side of  $\phi^*$  which suggests that this range of  $\phi$  is appropriate for learning a suitable dispersion parameter value.

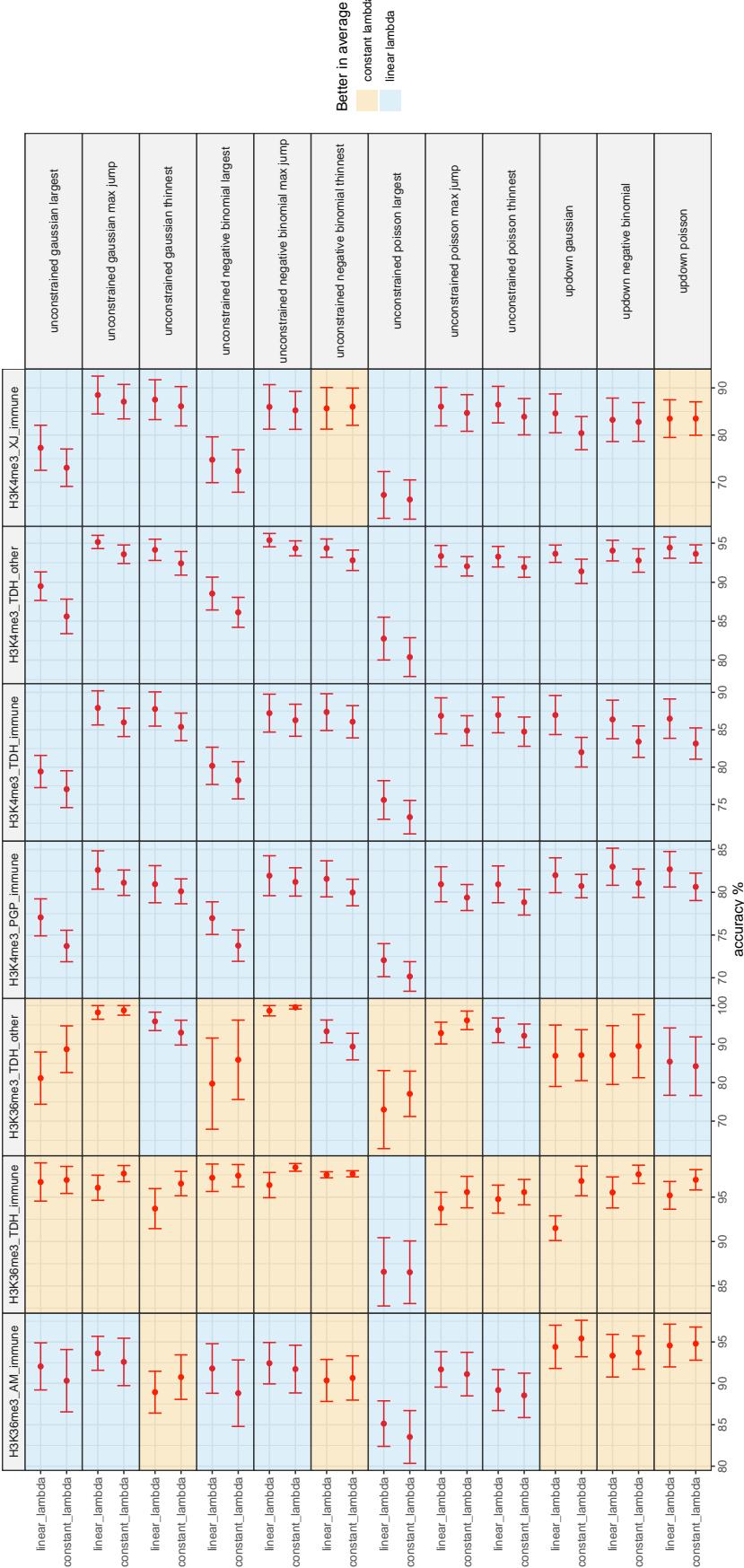


Figure 3: The *linear  $\lambda$*  learning method improves the accuracy of the segmentation models upon the *constant  $\lambda$*  learning method on H3K4me3 datasets. The mean accuracy and its 95% CI computed on the test folds is shown in red ■. In 46 of the 48 comparisons on the H3K4me3 datasets, the *linear  $\lambda$*  learning method was better in average than the *constant  $\lambda$*  learning method. After pooling the folds by type of experiment, we performed a paired t-test on each comparison. After correcting the p-values with the Benjamini & Hochberg method, 6 (/12) differences in mean accuracy were still significant (adjusted p-value  $< 0.05$ ). The concerned models are: unconstrained gaussian max jump; unconstrained gaussian largest; unconstrained negative binomial largest; unconstrained poisson thinnest; updown gaussian; updown negative binomial. In H3K36me3 datasets, there is no clear trend on which learning method is the best. 24 of the 36 comparisons are in favour of the *constant  $\lambda$*  learning method. After pooling the folds by type of experiment and performed a paired t-test following by a correction of the p-values, none of the 12 differences in mean were significant.

H3K4me3\_XJ\_immune/6/McGill0026  
MACS, parameter=2.7  
Up-Down poisson, lambda=1960.335  
unconstrained negative binomial, lambda=0.071, phi=10000  
unconstrained gaussian, lambda=328.597

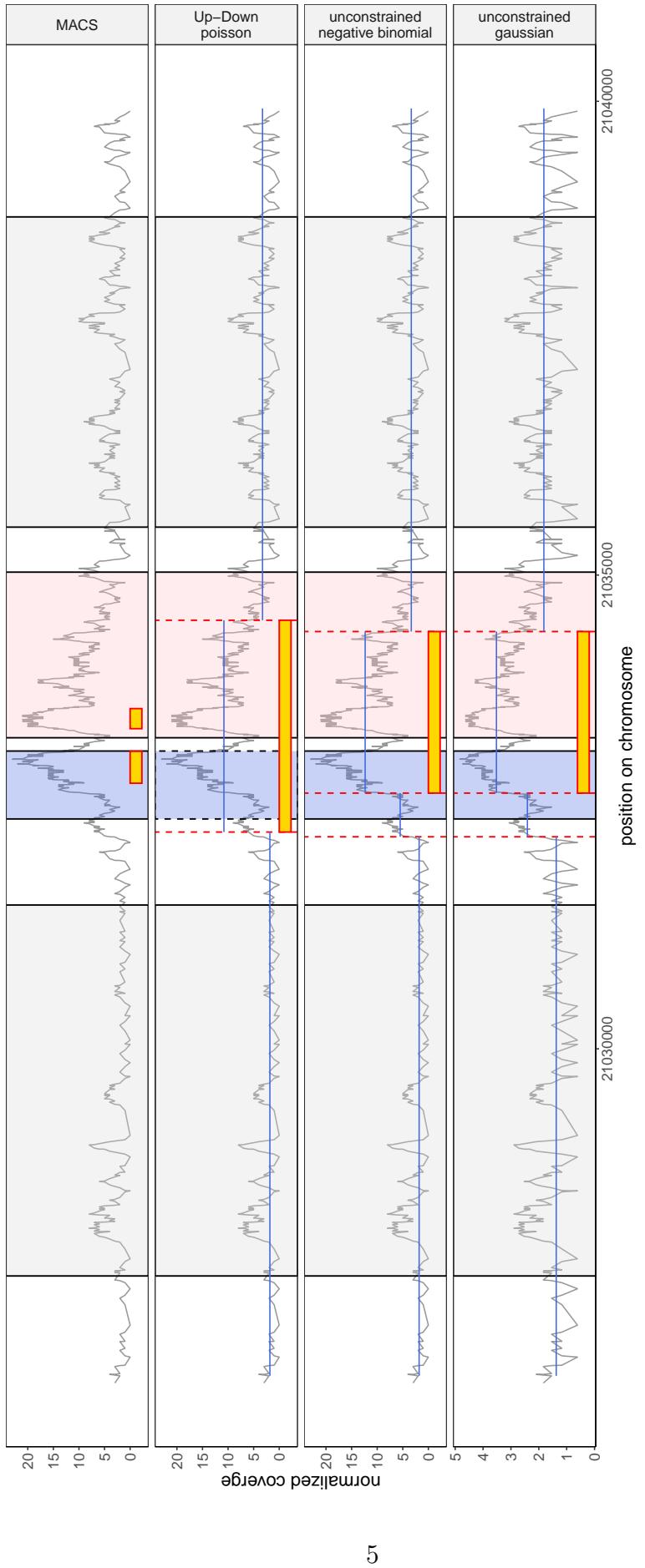


Figure 4: Visualization of the trained MACS, trained Up-Down Poisson, trained unconstrained negative binomial and trained unconstrained Gaussian models, on the normalized coverage profile (in the Gaussian case the data are additionally transformed as explained in the manuscript) from one of the biological samples (McGill0026) of the ChIP-Seq experiment directed against H3K4me3 histone marks. (**Top**) Summary of the tuning parameter values learned on the training set for each model. The parameter learned for MACS is the q-value threshold parameter.

H3K4me3\_TDH\_immune/3/McGill0059  
MACS, parameter=1.5  
Up-Down poisson, lambda=3691.67  
unconstrained negative binomial, lambda=14.199, phi=251.189  
unconstrained gaussian, lambda=1321.224

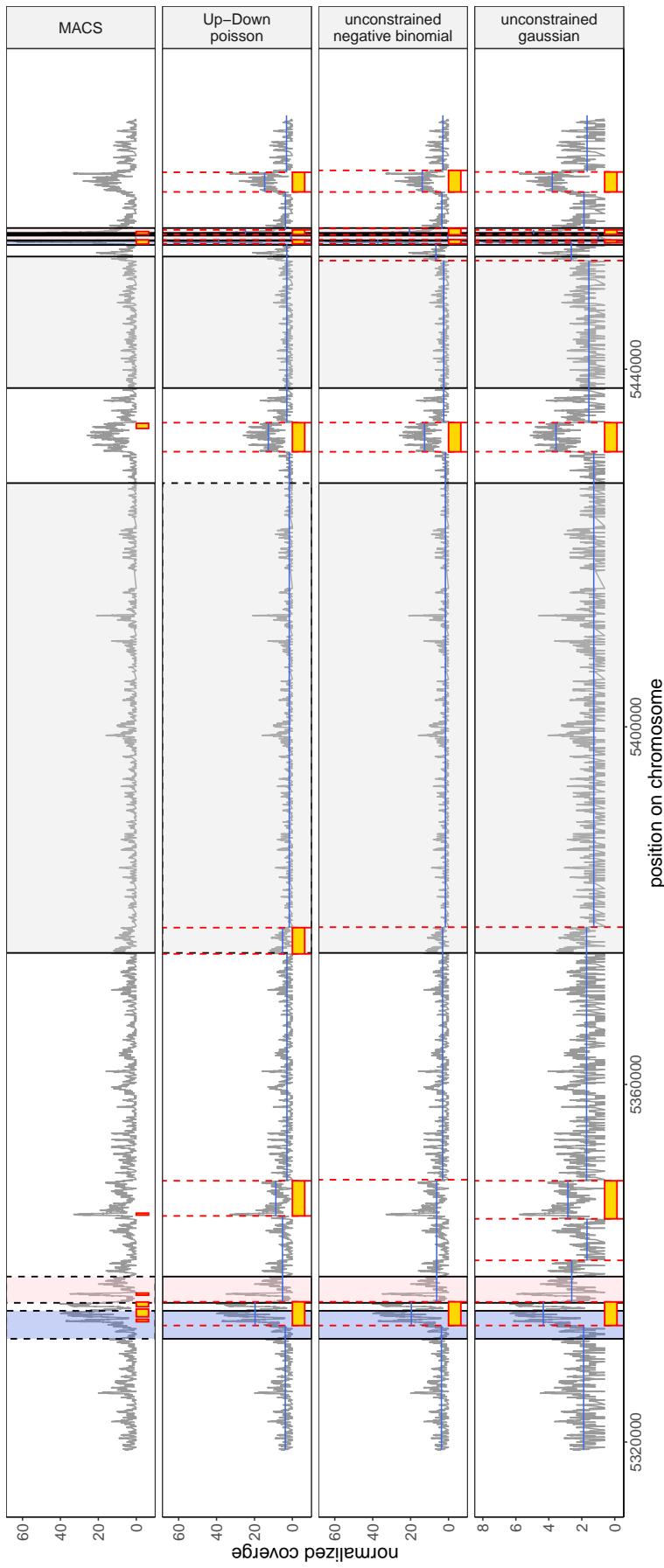


Figure 5: Visualization of the trained MACS, trained Up-Down Poisson, trained unconstrained negative binomial and trained unconstrained Gaussian models, on the normalized coverage profile (in the Gaussian case the data are additionally transformed as explained in the manuscript) from one of the biological samples (McGill0059) of the ChIP-Seq experiment directed against H3K4me3 histone marks. (**Top**) Summary of the tuning parameter values learned on the training set for each model. The parameter learned for MACS is the q-value threshold parameter.

H3K4me3\_PGP\_immune/20/McGill0001  
MACS, parameter=1.6  
Up–Down poisson, lambda=1438.822  
unconstrained negative binomial, lambda=24.749, phi=73.564  
unconstrained gaussian, lambda=1129.816

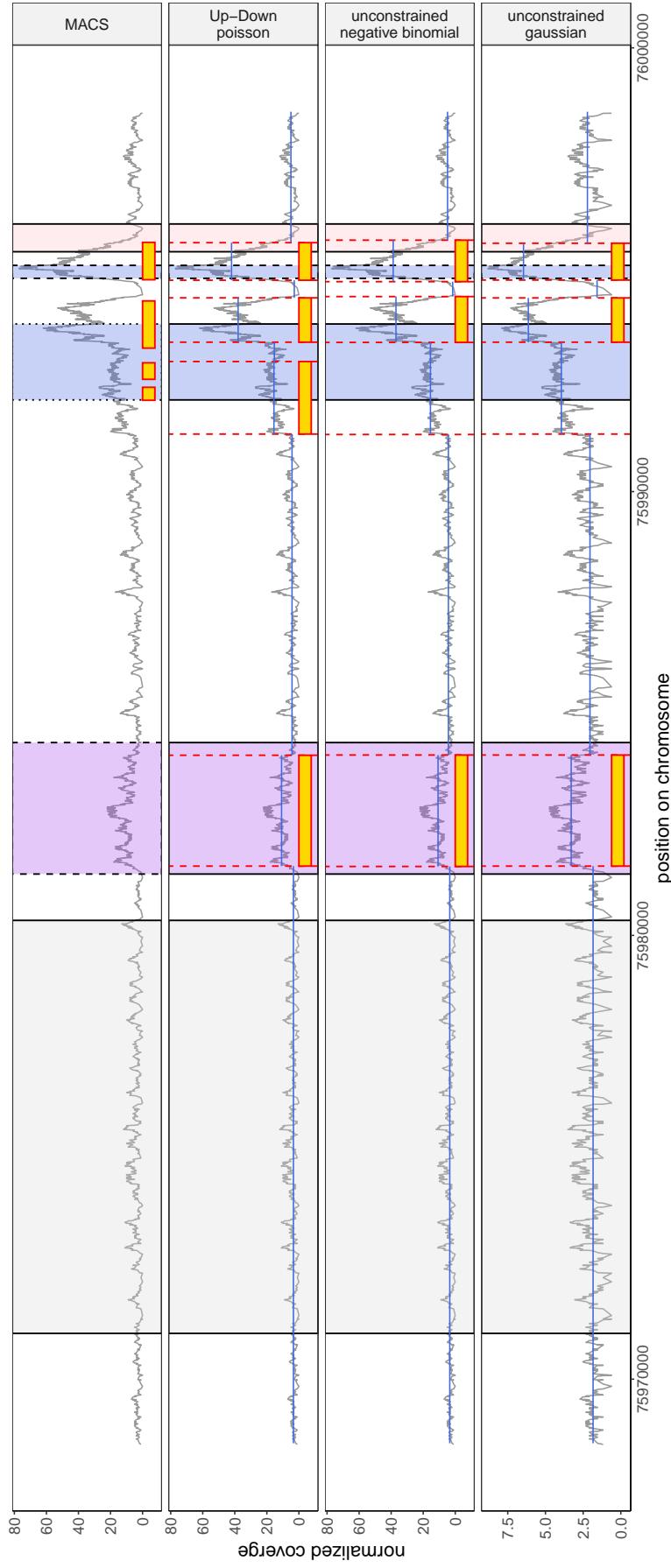


Figure 6: Visualization of the trained MACS, trained Up-Down Poisson, trained unconstrained negative binomial and trained unconstrained Gaussian models, on the normalized coverage profile (in the Gaussian case the data are additionally transformed as explained in the manuscript) from one of the biological samples (McGill0001) of the ChIP-Seq experiment directed against H3K4me3 histone marks. (**Top**) Summary of the tuning parameter values learned on the training set for each model. The parameter learned for MACS is the q-value threshold parameter.

H3K4me3\_TDH\_other/19/McGill0022  
MACS, parameter=5.5  
Up–Down poisson, lambda=11276.402  
unconstrained negative binomial, lambda=69.353, phi=135.936  
unconstrained gaussian, lambda=4528.46

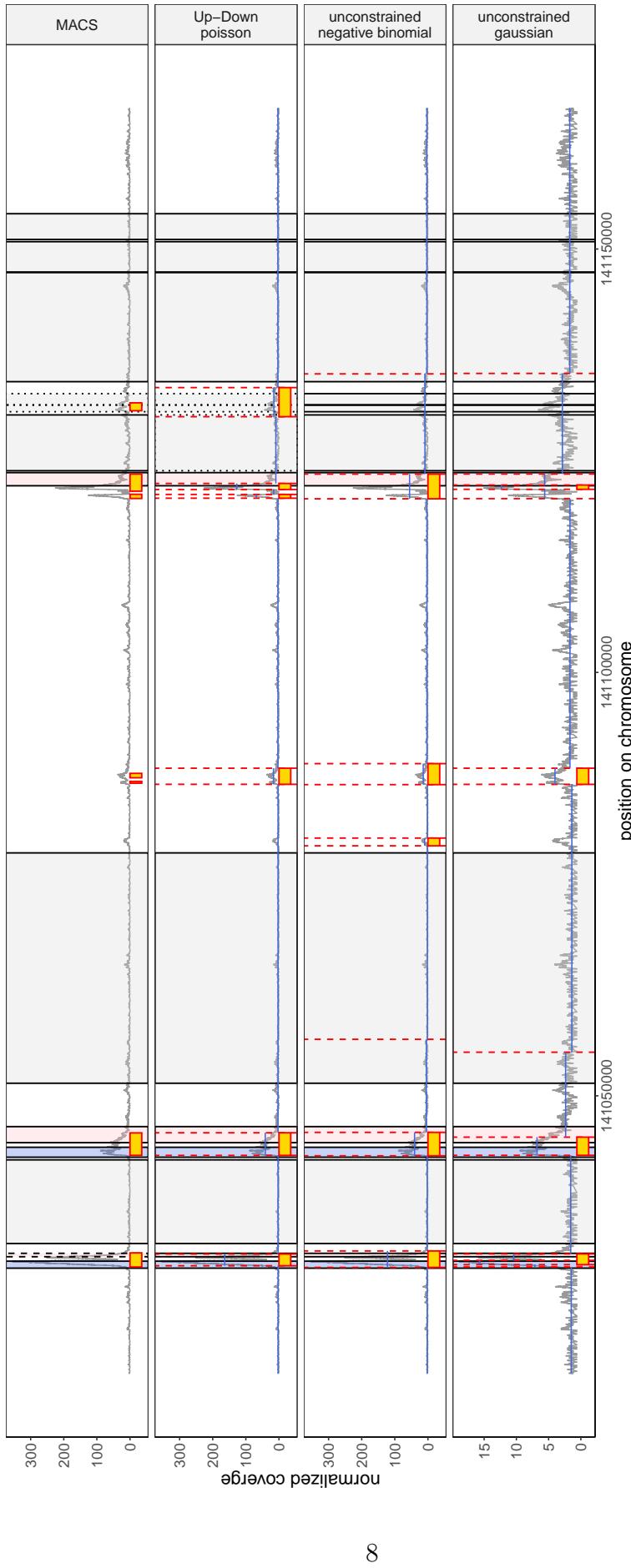


Figure 7: Visualization of the trained MACS, trained Up-Down Poisson, trained unconstrained negative binomial and trained unconstrained Gaussian models, on the normalized coverage profile (in the Gaussian case the data are additionally transformed as explained in the manuscript) from one of the biological samples (McGill0022) of the ChIP-Seq experiment directed against H3K4me3 histone marks. (**Top**) Summary of the tuning parameter values learned on the training set for each model. The parameter learned for MACS is the q-value threshold parameter.

H3K36me3\_TDH\_immune/3/McGill0024  
 HMCan, parameter=4.7  
 Up-Down poisson, lambda=62399.771  
 unconstrained negative binomial, lambda=6344.735, phi=6.31  
 unconstrained gaussian, lambda=37346.512

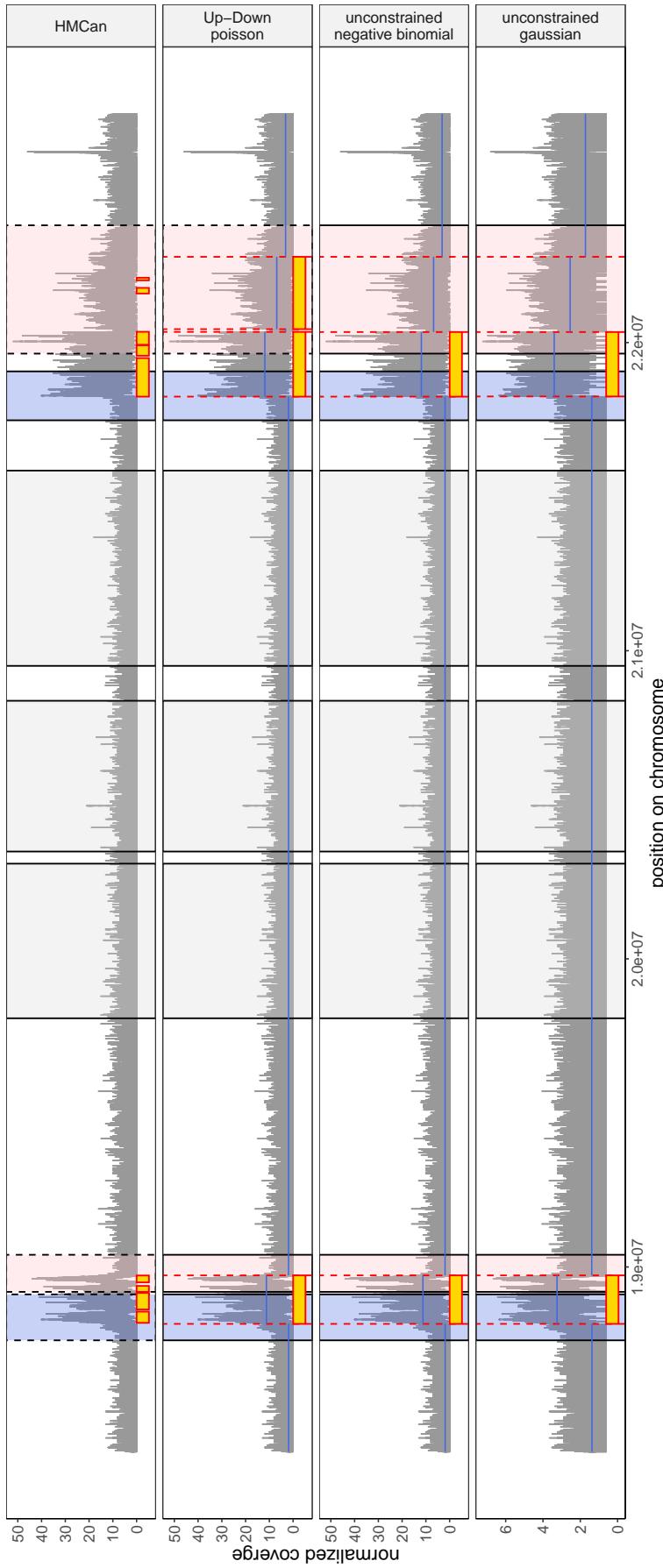


Figure 8: Visualization of the trained HMCan, trained Up-Down Poisson, trained unconstrained negative binomial and trained unconstrained Gaussian models, on the normalized coverage profile (in the Gaussian case the data are additionally transformed as explained in the manuscript) from one of the biological samples (McGill0024) of the ChIP-Seq experiment directed against H3K36me3 histone marks. (Top) Summary of the tuning parameter values learned on the training set for each model. The parameter learned for HMCan is the finalThreshold parameter.

H3K36me3\_TDH\_other/1/McGill0012  
 HMCan, parameter=4.4  
 Up-Down poisson, lambda=37287.707  
 unconstrained negative binomial, lambda=838.423, phi=11.659  
 unconstrained gaussian, lambda=2941.584

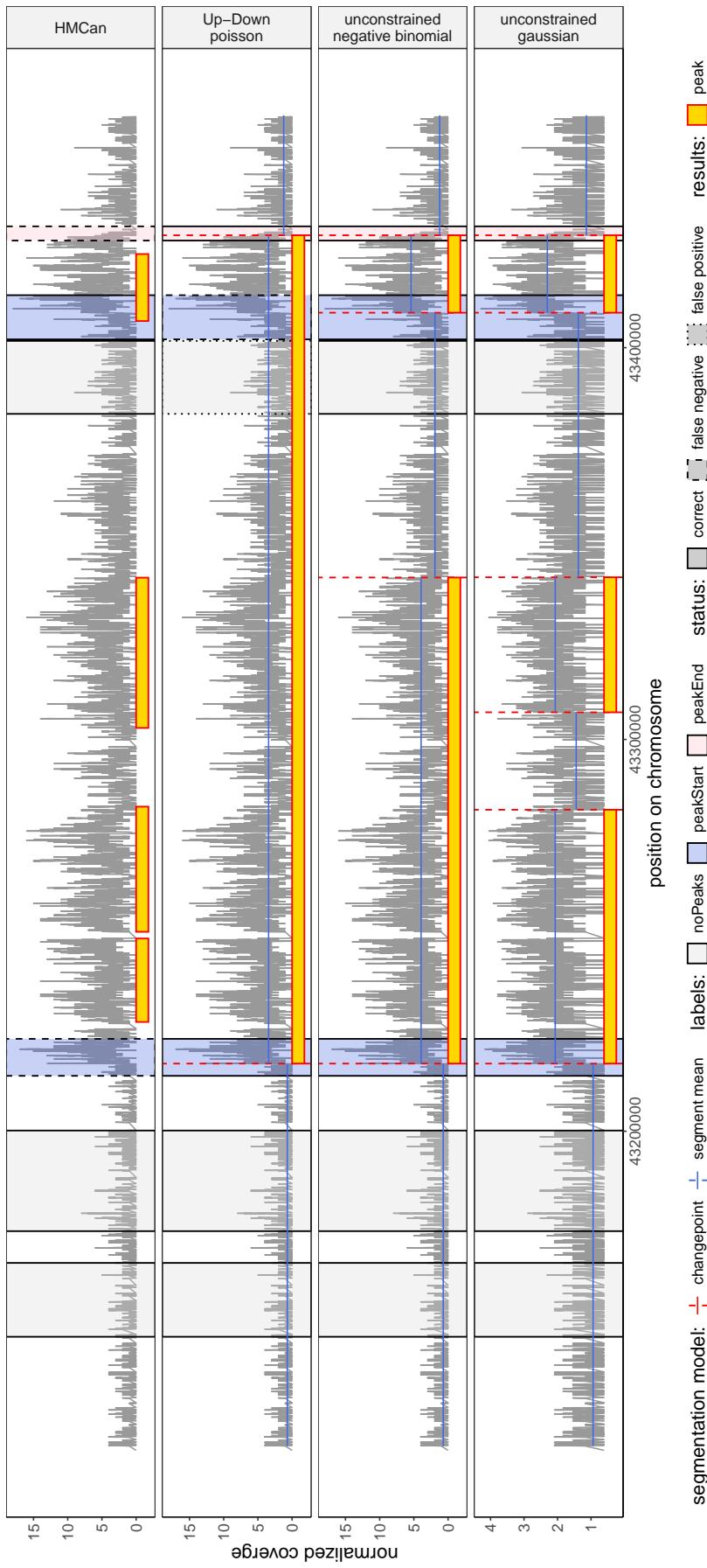


Figure 9: Visualization of the trained HMCan, trained Up-Down Poisson, trained unconstrained negative binomial and trained unconstrained Gaussian models, on the normalized coverage profile (in the Gaussian case the data are additionally transformed as explained in the manuscript) from one of the biological samples (McGill0012) of the ChIP-Seq experiment directed against H3K36me3 histone marks. (Top) Summary of the tuning parameter values learned on the training set for each model. The parameter learned for HMCan is the finalThreshold parameter.

H3K36me3\_AM\_immune/17/McGill0028  
 HMCan, parameter=4.4  
 Up–Down poisson, lambda=9281.377  
 unconstrained negative binomial, lambda=495.85, phi=11.659  
 unconstrained gaussian, lambda=4969.542

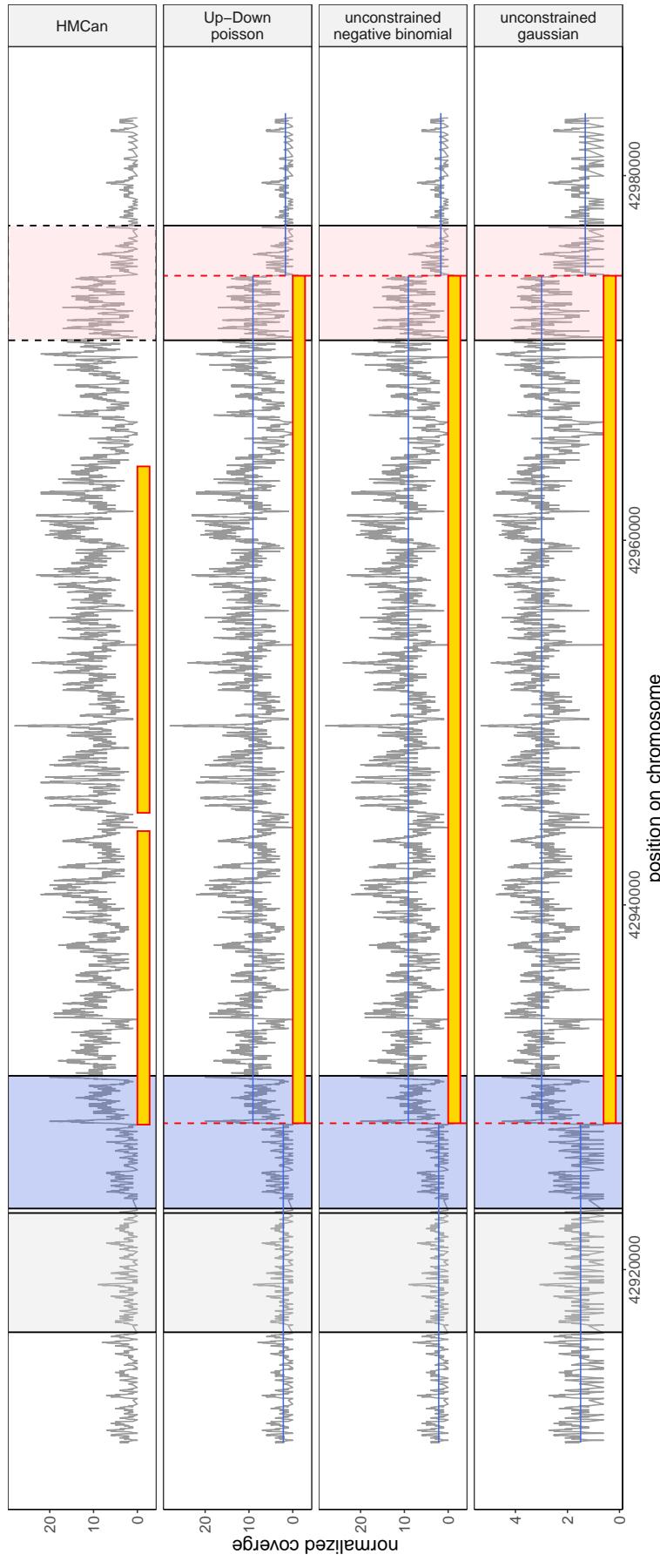


Figure 10: Visualization of the trained HMCan, trained Up-Down Poisson, trained unconstrained negative binomial and trained unconstrained Gaussian models, on the normalized coverage profile (in the Gaussian case the data are additionally transformed as explained in the manuscript) from one of the biological samples (McGill0028) of the ChIP-Seq experiment directed against H3K36me3 histone marks. (**Top**) Summary of the tuning parameter values learned on the training set for each model. The parameter learned for HMCan is the finalThreshold parameter.