Supplementary Figures

Maria Osmala and Harri Lähdesmäki

List of Figures

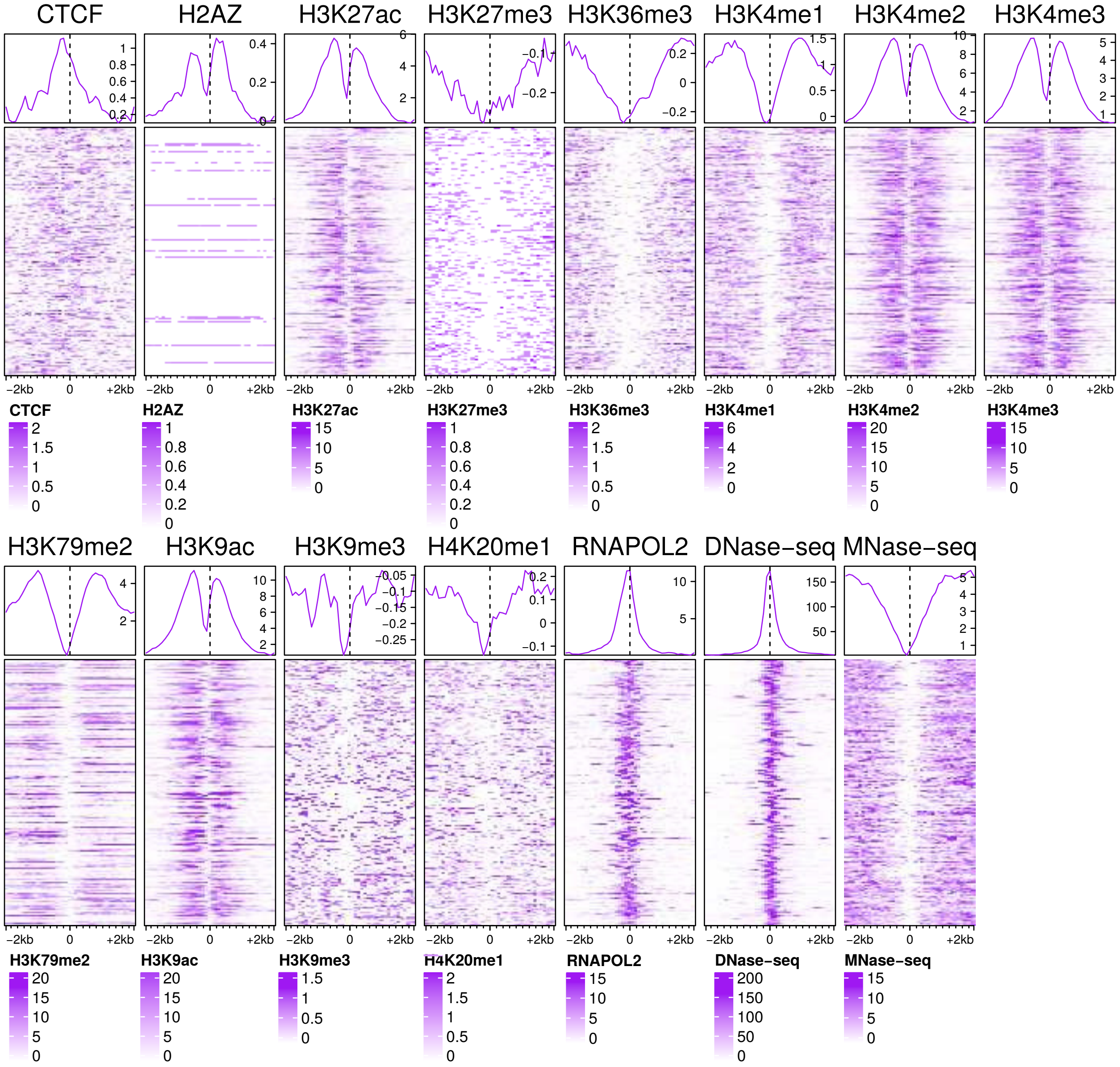

Figure S1: The coverage of different chromatin features at 1000 individual promoters, and their average profiles over promoters. The data originates from cell type K562, and features are presented in 4 kb window with bin size 100 bp. The promoters are unoriented, i.e. the direction of transcription from TSS is not utilized to direct the promoters.

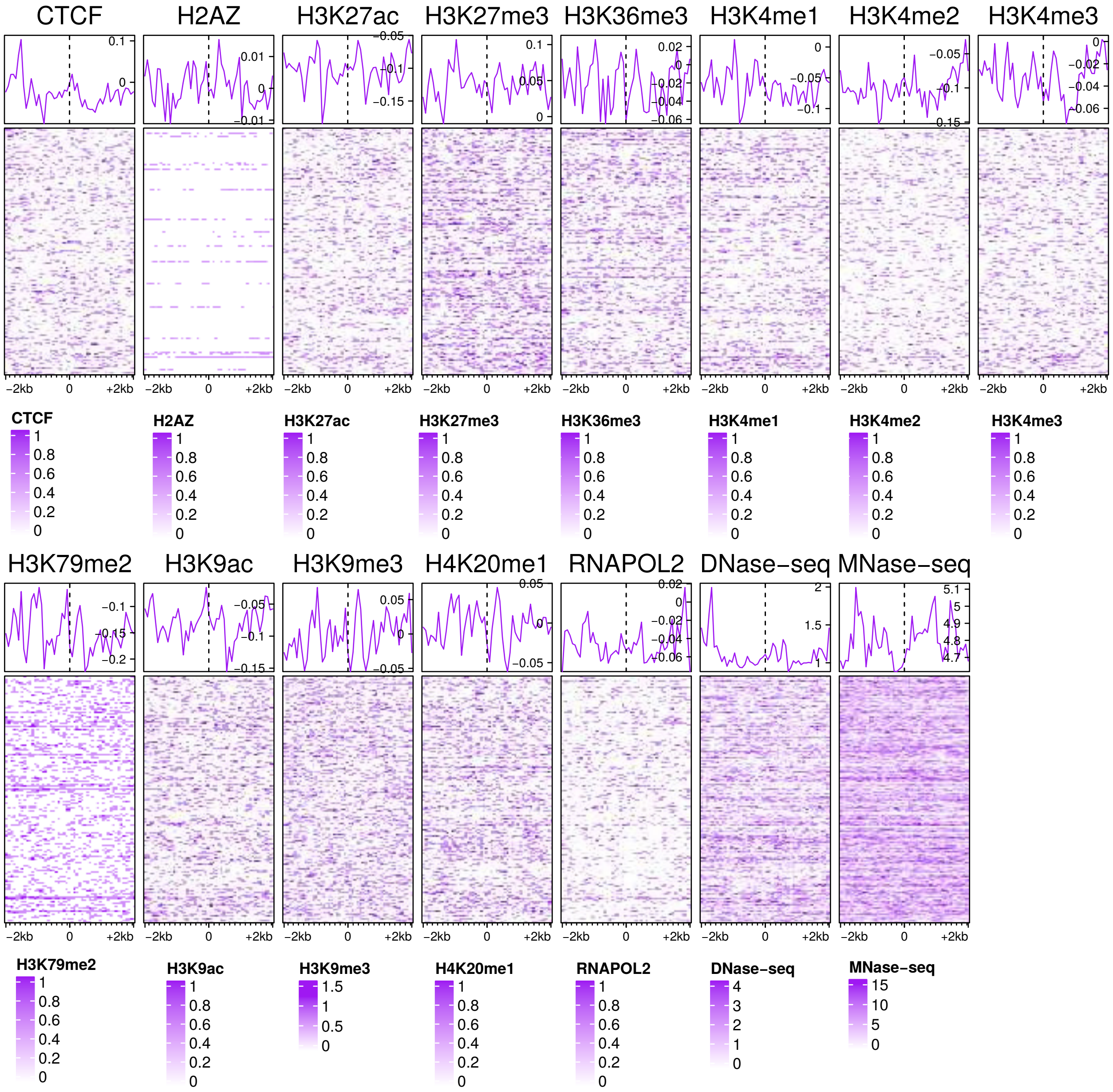

Figure S2: The coverage of different chromatin features at 1000 individual pure random genomic locations, and their average profile over the locations. The data originates from cell type K562, and features are presented in 4 kb window with bin size 100 bp.

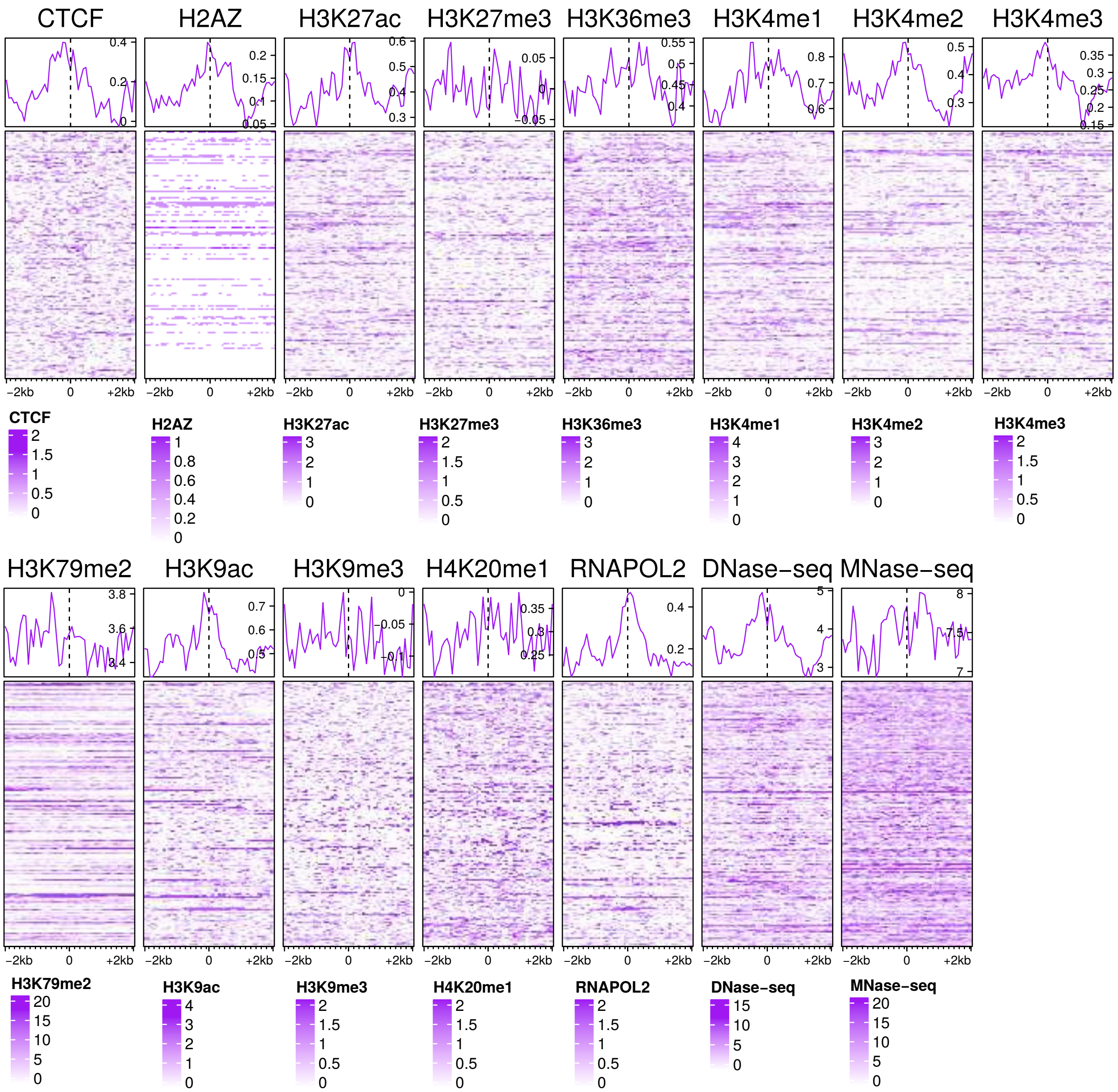

Figure S3: The coverage of different chromatin features at 1000 individual random genomic locations with signal, and their average profile over the locations. The data originates from cell type K562, and features are presented in 4 kb window with bin size 100 bp.

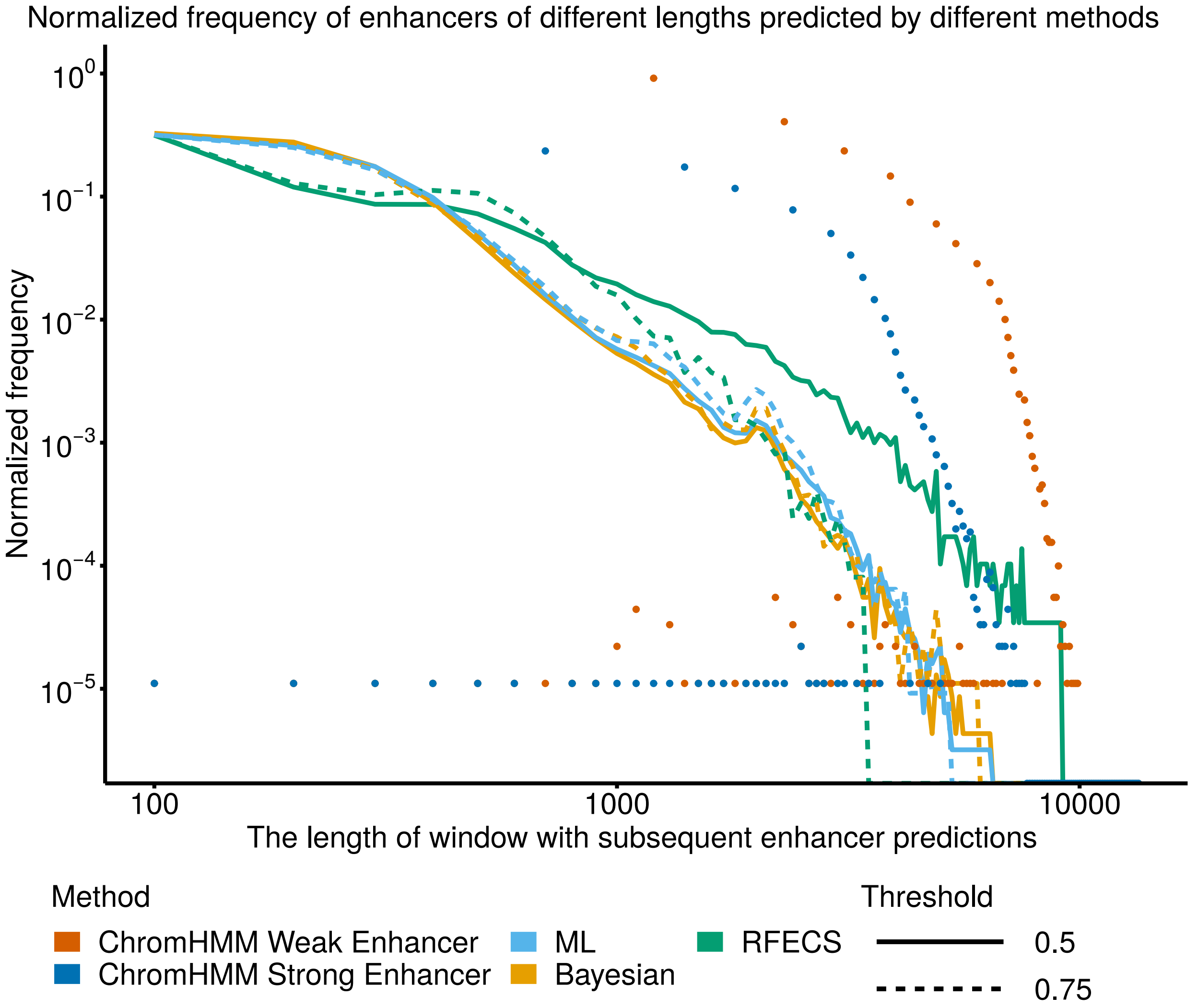

Figure S4: The normalized frequencies of varying lengths of enhancers predicted in cell line K562 by different methods using two prediction thresholds (0.5 and 0.75). PREPRINT and RFECS were trained on the random regions with signal. For each method and threshold, the frequencies were divided by the total number of regions predicted as enhancers by each method. The regions were formed by combining the subsequent enhancer predictions into one region.

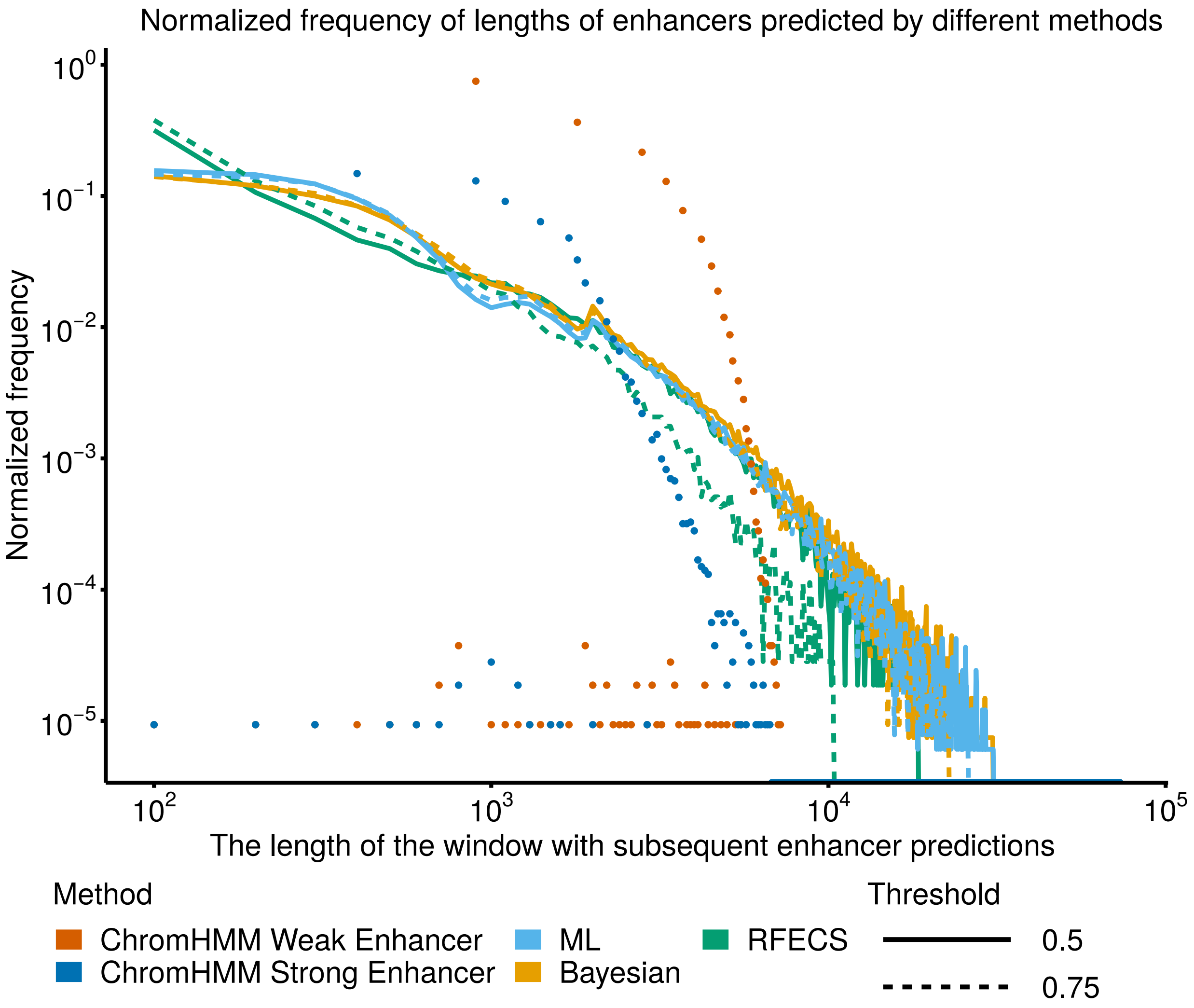

Figure S5: The normalized frequencies of varying lengths of enhancers predicted in cell line GM12878 by different methods using two prediction thresholds (0.5 and 0.75). PREPRINT and RFECS were trained on the pure random regions. For each method and threshold, the frequencies were divided by the total number of regions predicted as enhancers by each method. The regions were formed by combining the subsequent enhancer predictions into one region.

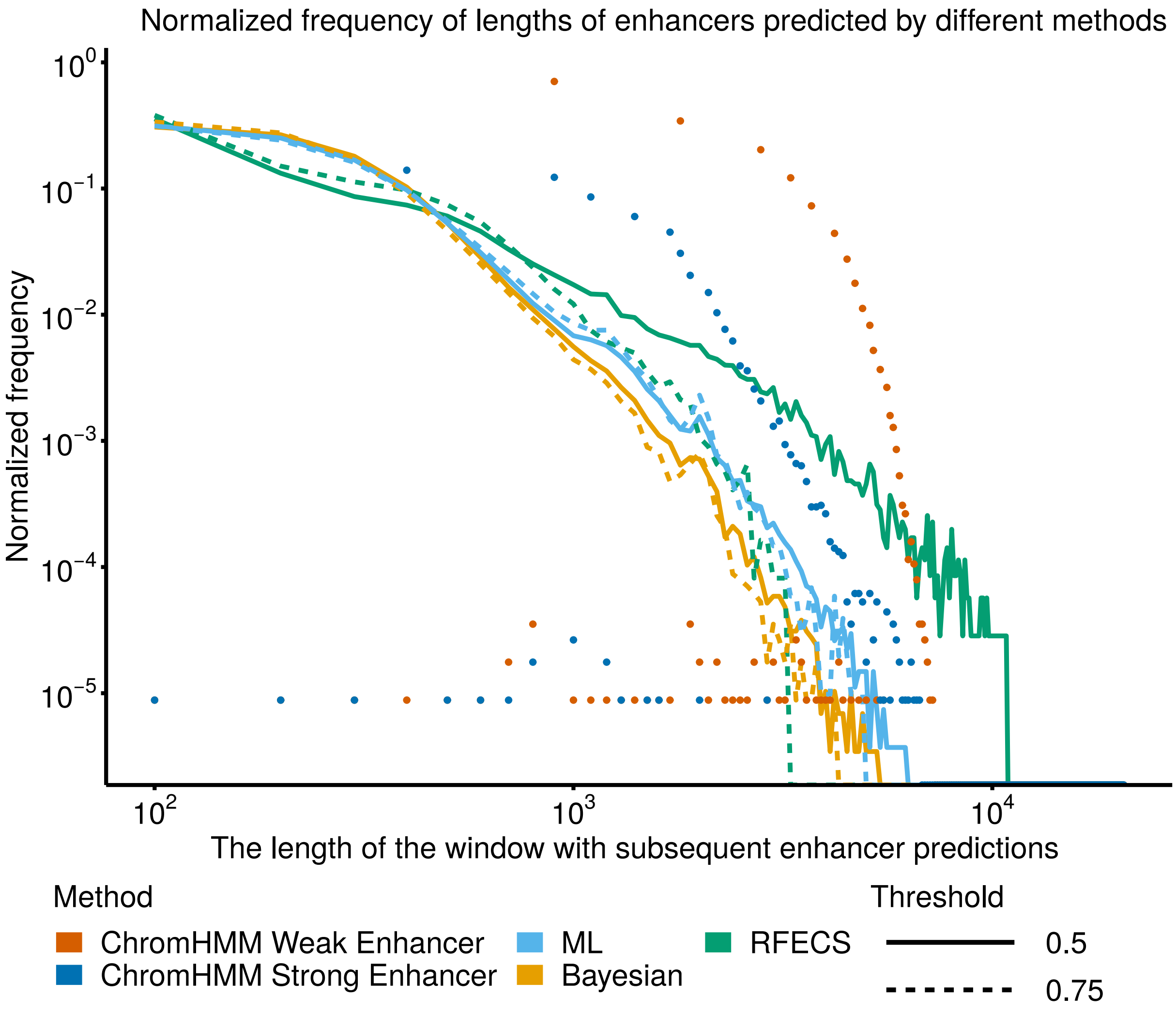

Figure S6: The normalized frequencies of varying lengths of enhancers predicted in cell line GM12878 by different methods using two prediction thresholds (0.5 and 0.75). PREPRINT and RFECS were trained on the random regions with signal. For each method and threshold, the frequencies were divided by the total number of regions predicted as enhancers by each method. The regions were formed by combining the subsequent enhancer predictions into one region.

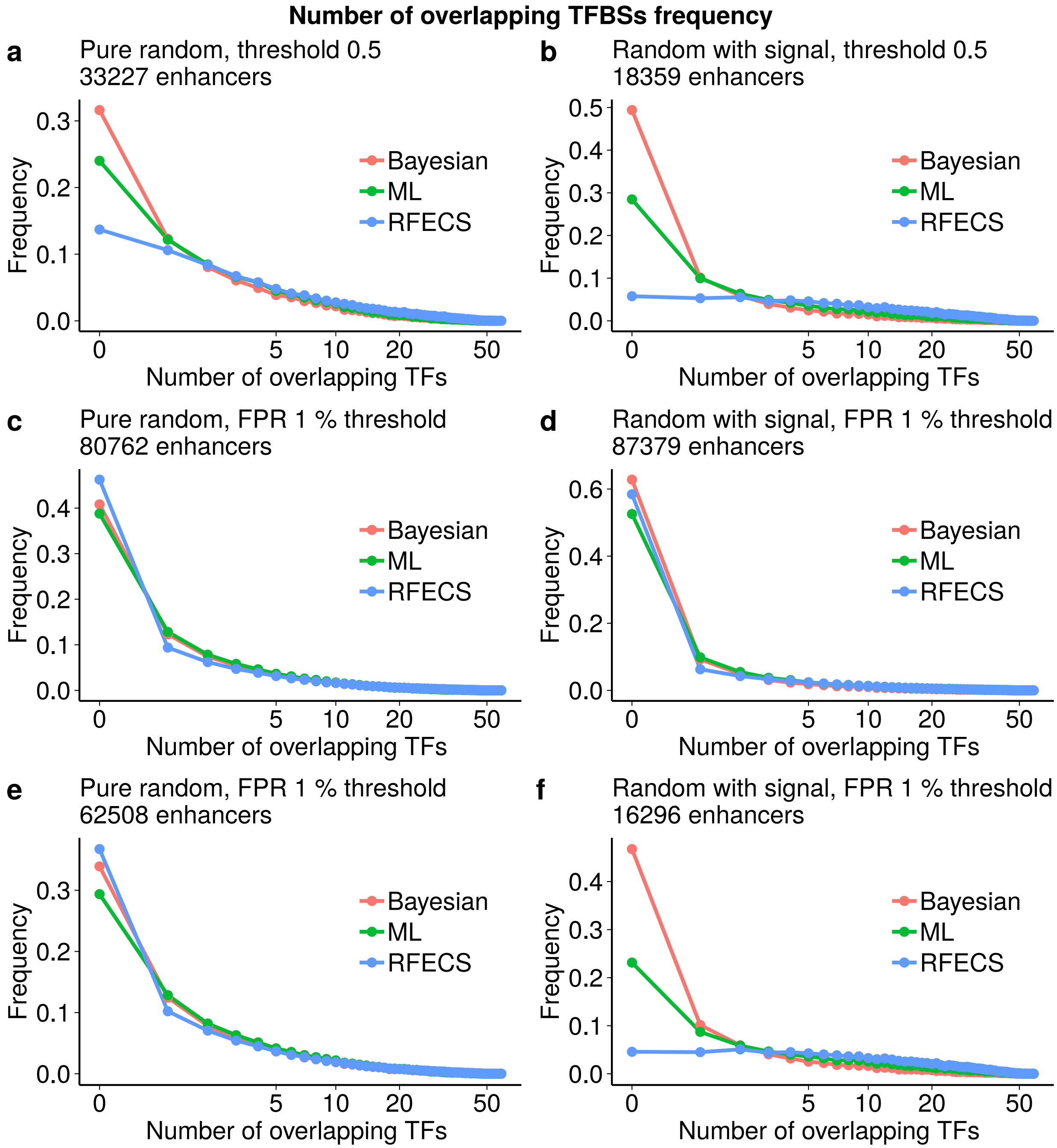

Figure S7: (Caption next page.)

Figure S7: (Previous page.) The proportions of the genome-wide enhancer predictions having an overlap with the varying number of ChIP-seq peaks in cell line GM12878. The proportions are shown for the different random region definitions and for the different thresholds. In **a**, **c**, and **e**, the methods were trained on the pure random regions, and in **b**, **d**, and **f**, the methods were trained on the random regions with signal. The number of enhancers in each comparison are shown above the figure. In **a** and **b**, the number of enhancers was the minimum number of enhancers predicted by any of the methods with the threshold 0.5, in **c** and **d**, the number of enhancers was the minimum number of enhancers predicted by PREPRINT methods with their 1% FPR thresholds estimated on the K562 CV data, and in **e** and **f**, the number of enhancers was the minimum number of enhancers predicted by PREPRINT methods with their 1% FPR thresholds estimated on the GM12878 test data.

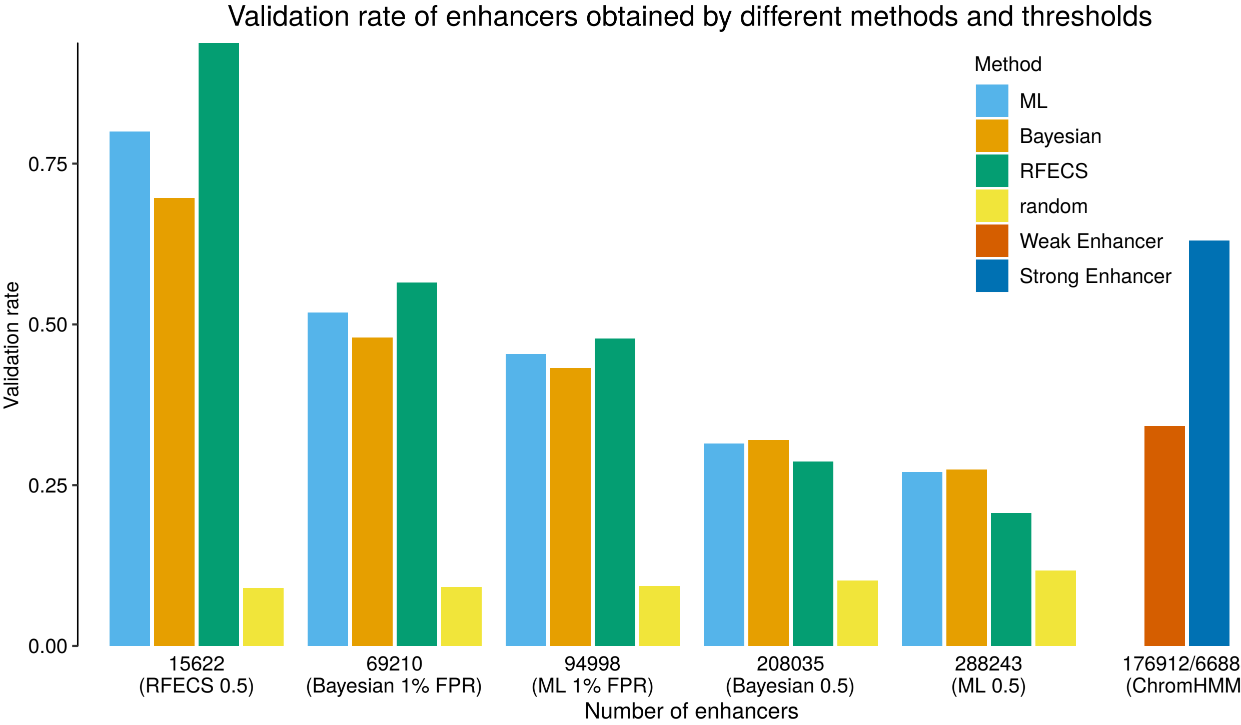

Figure S8: The validation rate of the genome-wide enhancer predictions obtained by the different methods and thresholds in cell line K562. The methods were trained on the random regions with signal. An enhancer prediction was validated if it overlapped at least 1 bp of at least one TF or co-regulatory ChIP-seq peak.

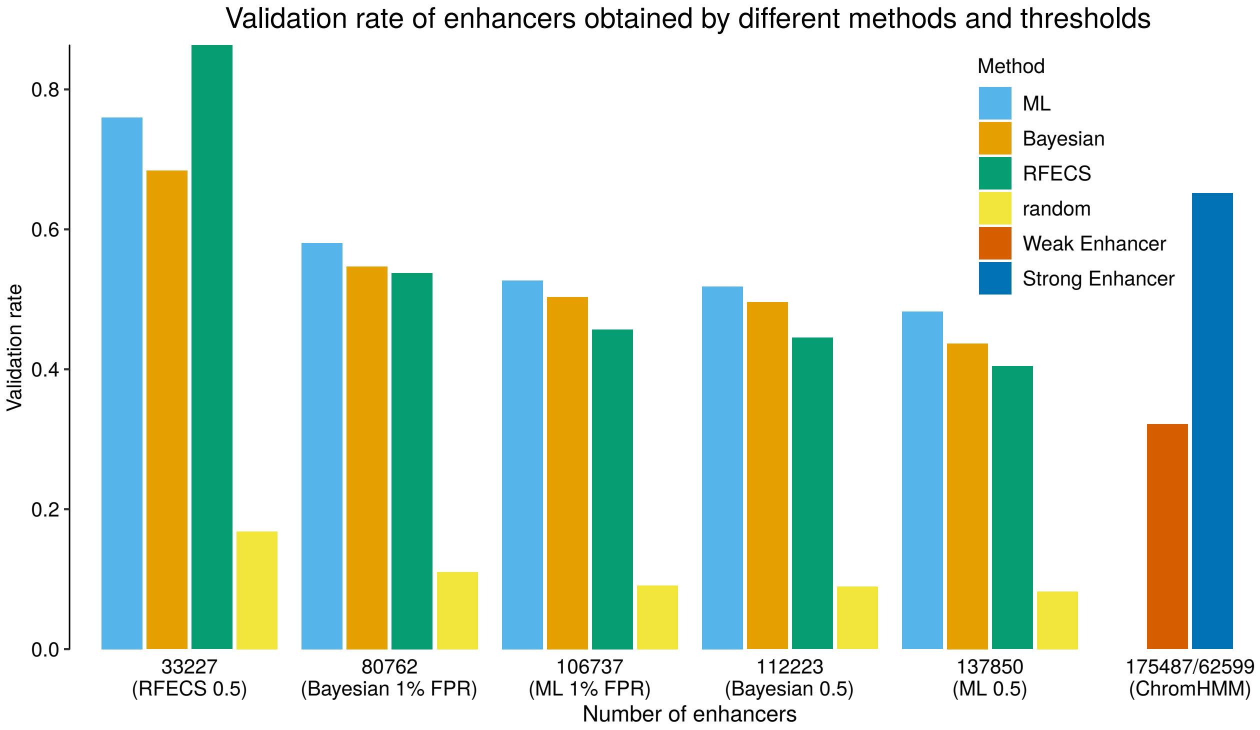

Figure S9: The validation rate of the genome-wide enhancer predictions obtained by the different methods and thresholds in cell line GM12878. The methods were trained on the pure random regions. An enhancer prediction was validated if it overlapped at least 1 bp of at least one TF or co-regulatory ChIP-seq peak.

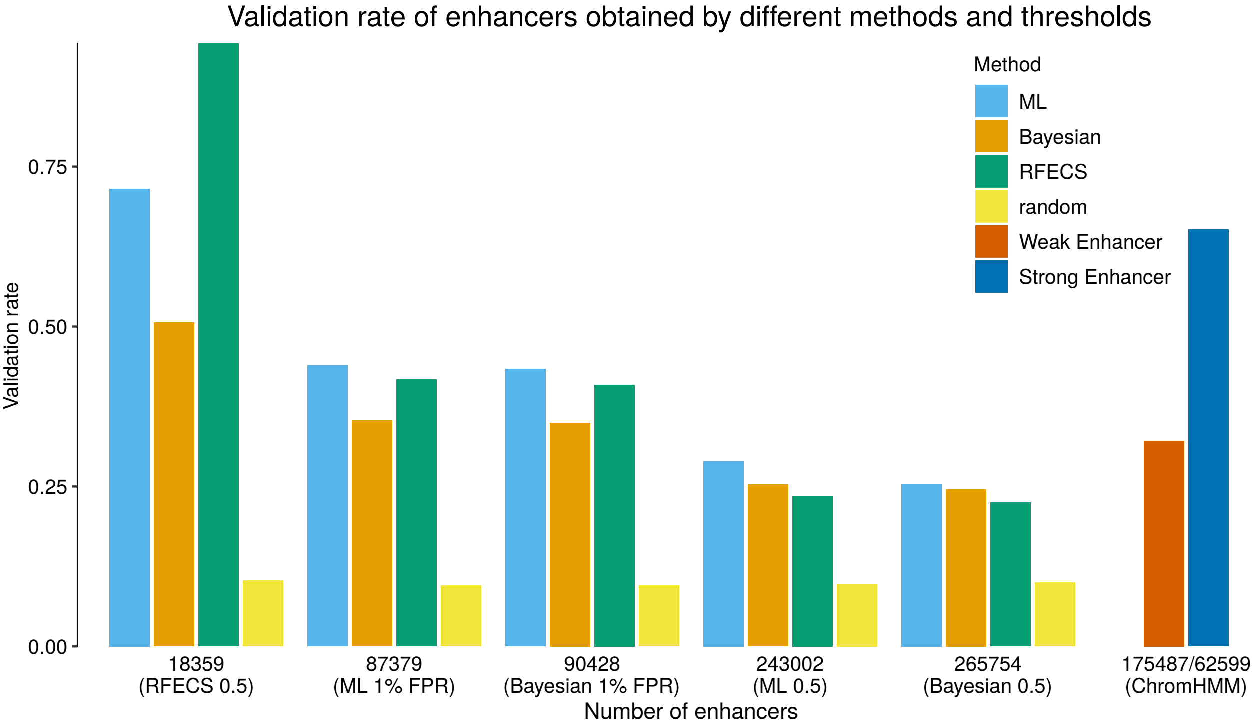

Figure S10: The validation rate of the genome-wide enhancer predictions obtained by the different methods and thresholds in cell line GM12878. The methods were trained on the random regions with signal. An enhancer prediction was validated if it overlapped at least 1 bp of at least one TF or co-regulatory ChIP-seq peak.

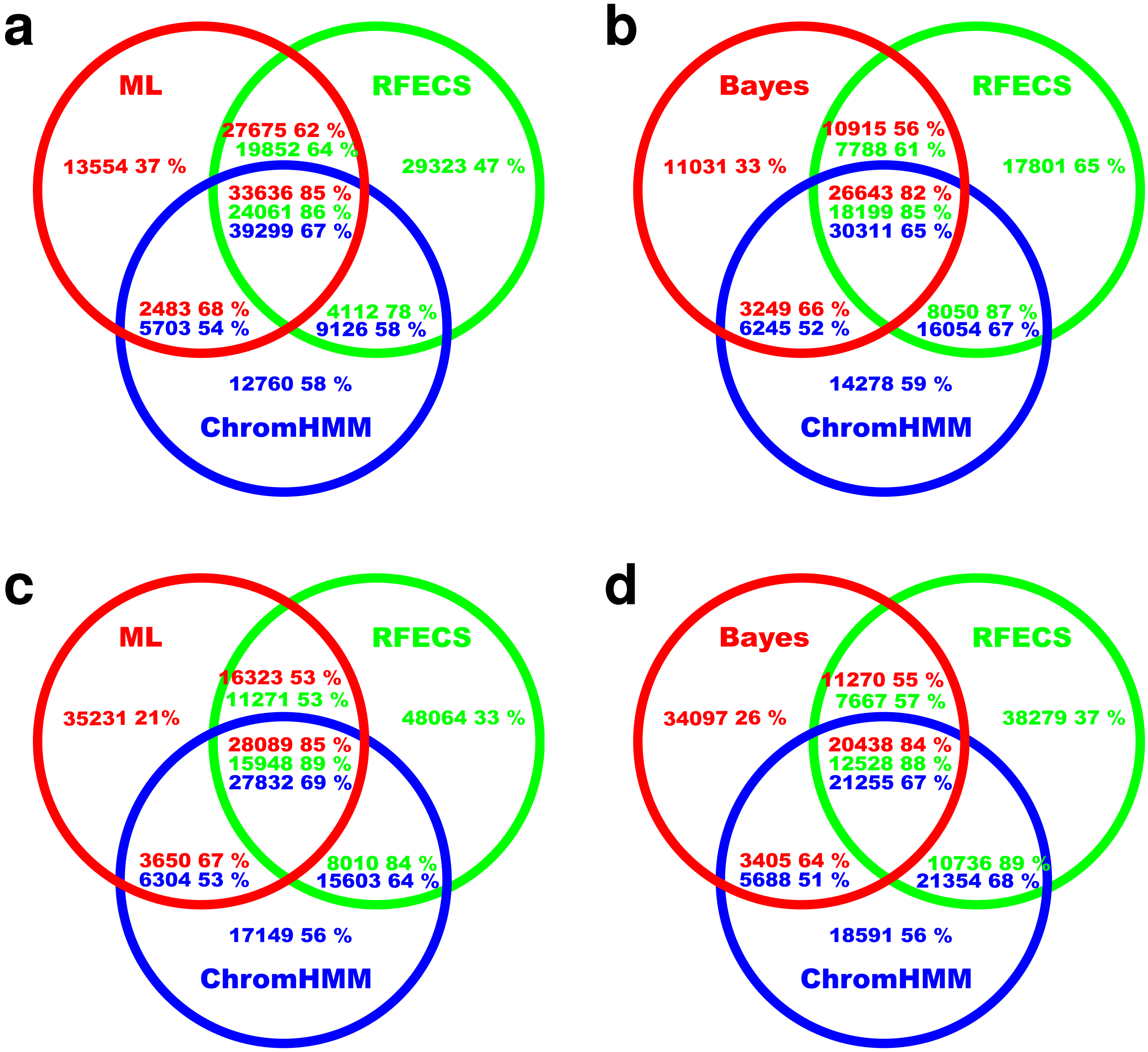

Figure S11: The unique and overlapping genome-wide enhancer predictions made by different methods in cell line K562 when using 1% FPR threshold for the PREPRINT methods. In **a** and **c**, predictions were obtained with the ML approach, and in **b** and **d**, the predictions were obtained with the Bayesian approach. The overlap between the PREPRINT, RFECS and ChromHMM predictions were quantified as the number of enhancers. In figures **a** and **b**, PREPRINT and RFECS were trained on the pure random regions, and in figures **c** and **d**, PREPRINT and RFECS were trained on the random regions with signal. In each figure, the number of enhancers predicted by PREPRINT and RFECS was the same. The number of enhancers was chosen to be the minimum number of enhancers predicted by either PREPRINT with the 1% FPR threshold or by RFECS with the lower threshold. The numbers were: **a** 51838, **b** 51838, **c** 69210, and **d** 69210. Inside every area, the number of enhancers belonging to the set is shown together with the proportion of validated enhancers in the set. The overlapping areas are not proportional to the number of overlapping regions due to the asymmetry of the overlaps.

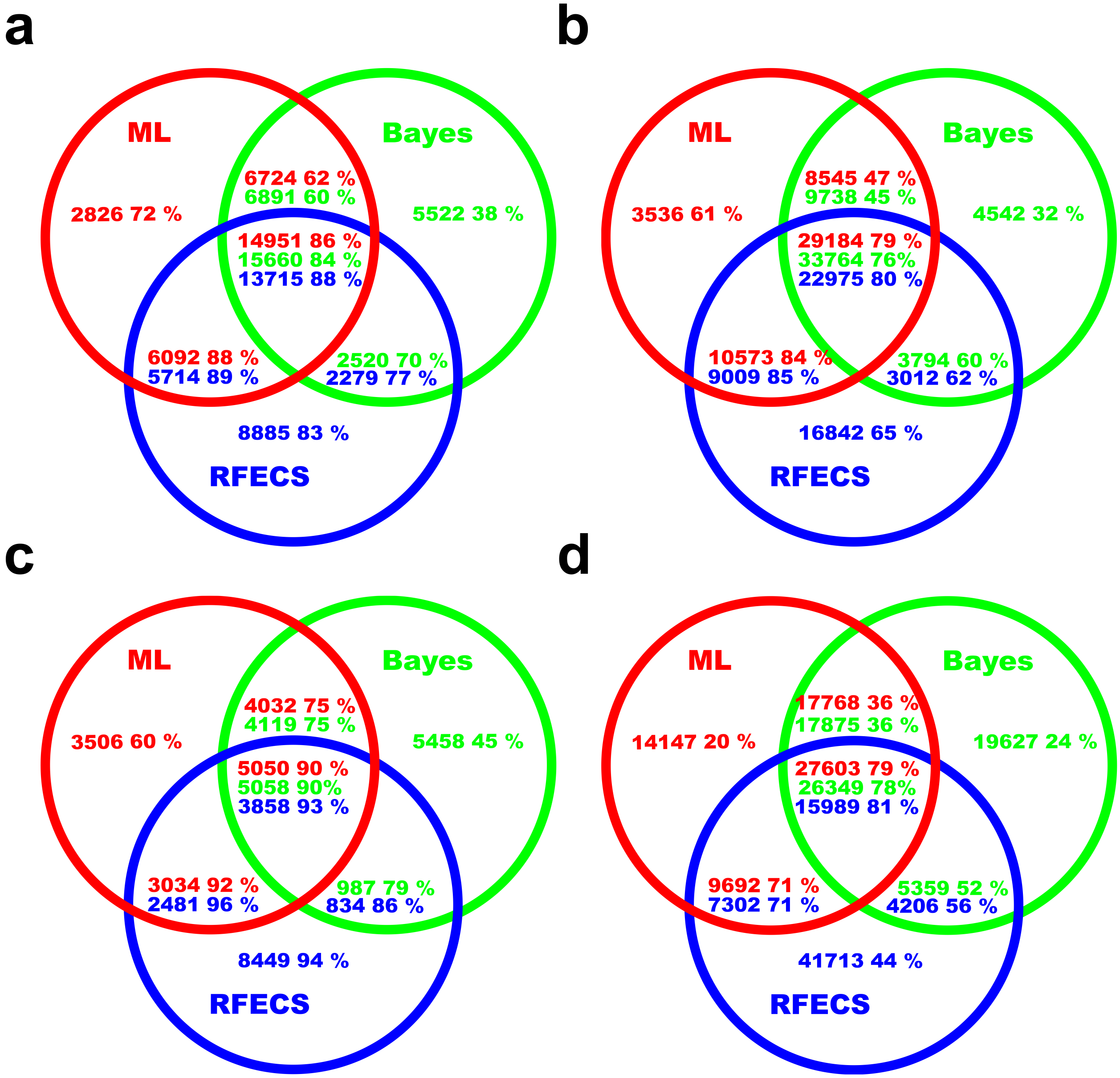

Figure S12: The unique and overlapping genome-wide enhancer predictions made by PREPRINT and RFECS in cell line K562. In **a** and **c**, the predictions were obtained using the 0.5 threshold, and in **b** and **d**, the PREPRINT predictions were obtained using the 1% FPR threshold. In figures **a** and **b**, the methods were trained on the pure random regions, and in **c** and **d**, the methods were trained on the random regions with signal. In **a** and **c**, an equal number of top enhancers was selected for each method, the number was the minimum across all methods with threshold 0.5. In **b** and **d**, the number of enhancers was the minimum number predicted across PREPRINT methods with their 1% FPR threshold. The numbers were: **a** 30593, **b** 51838, **c** 15622, and **d** 69210. Inside every area, the number of enhancers belonging to the set is shown together with the proportion of validated enhancers in the set. The overlapping areas are not proportional to the number of overlapping regions due to the asymmetry of the overlaps.

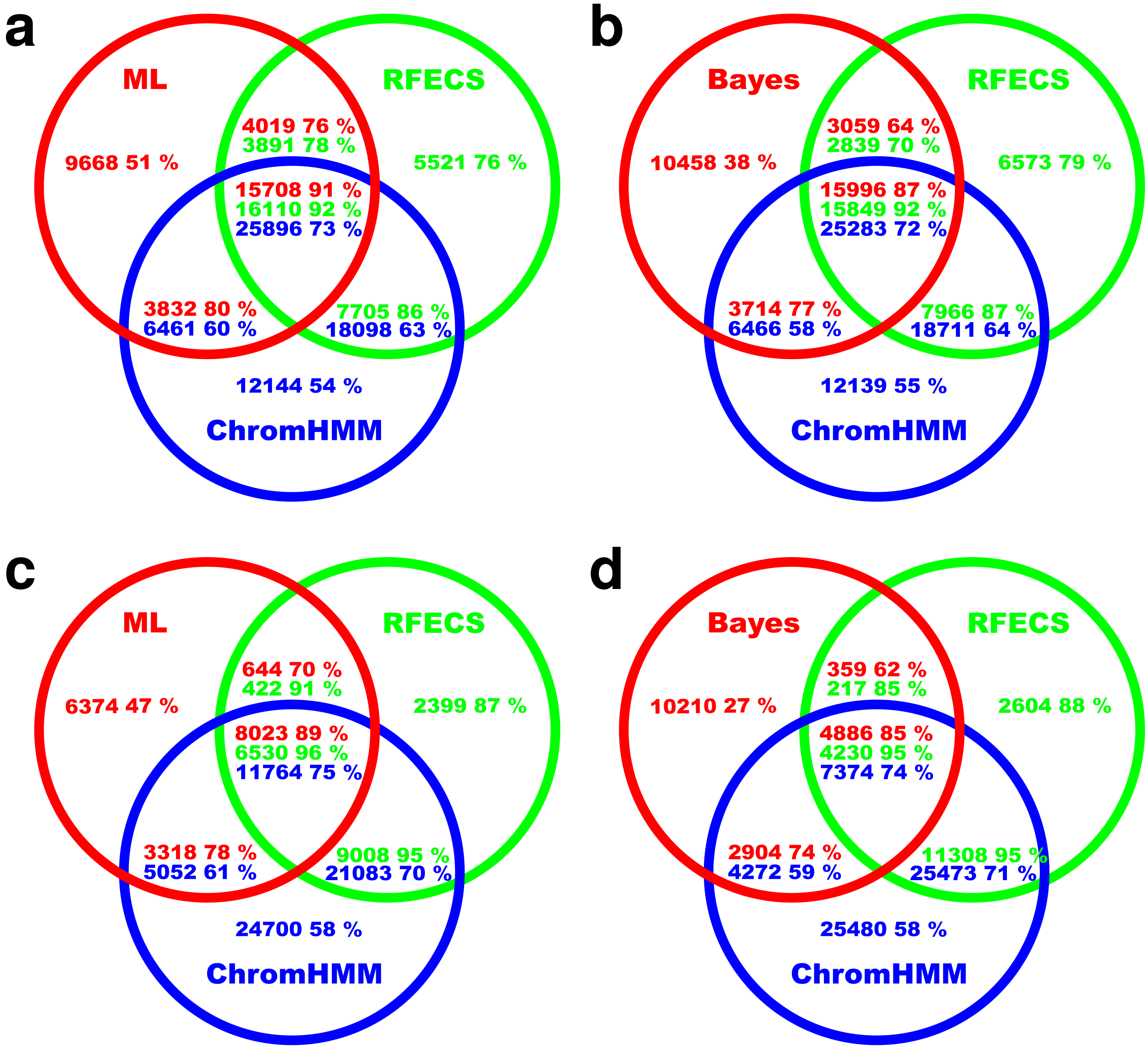

Figure S13: The unique and overlapping genome-wide enhancer predictions made by different methods in cell line GM12878. In **a** and **c**, the predictions were obtained by the ML approach, and in **b** and **d**, the predictions were obtained by the Bayesian approach. The overlap between the PREPRINT, RFECS and ChromHMM predictions were quantified as the number of enhancers. In figures **a** and **b**, PREPRINT and RFECS were trained on the pure random regions, and in **c** and **d**, PREPRINT and RFECS were trained on the random regions with signal. In each figure, the number of enhancers predicted by PREPRINT and RFECS was the same. The number of enhancers was chosen to be the minimum number of enhancers predicted by either PREPRINT or RFECS with the 0.5 threshold. The numbers were: **a** 33227, **b** 33227, **c** 18359, and **d** 18359. Inside every area, the number of enhancers belonging to the set are shown together with the proportion of validated enhancers in the set. The overlapping areas are not proportional to the number of overlapping regions due to the asymmetry of the overlaps.

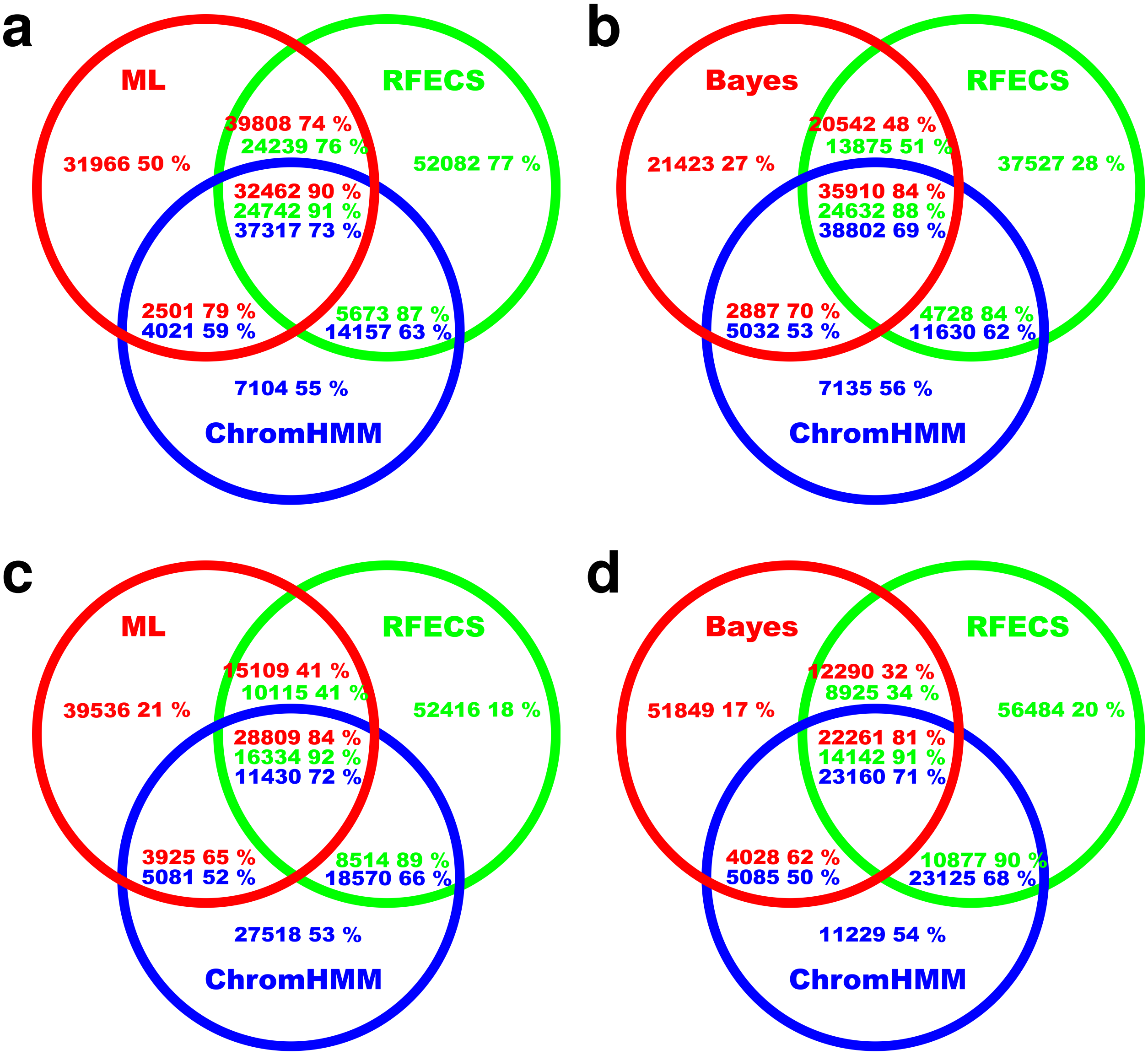

Figure S14: The unique and overlapping genome-wide enhancer predictions made by different methods in cell line GM12878. The PREPRINT enhancers were obtained using the 1% FPR thresholds. The 1% FPR thresholds were estimated from the K562 CV data. In **a** and **c**, the PREPRINT predictions were obtained with the ML approach, and in **b** and **d**, the PREPRINT predictions were obtained with the Bayesian approach. The overlap between the PREPRINT, RFECS and ChromHMM predictions were quantified as the number of enhancers. In figures **a** and **b**, PREPRINT and RFECS were trained on the pure random regions, and in **c** and **d**, PREPRINT and RFECS were trained on the random regions with signal. In each figure, the number of enhancers predicted by PREPRINT and RFECS was the same. The number of enhancers was chosen to be the minimum number of enhancers predicted by either PREPRINT with the 1% FPR threshold or by RFECS with the lower threshold. The numbers were: **a** 80762, **b** 80762, **c** 87379, and **d** 87379. Inside every area, the number of enhancers belonging to the set are shown together with the proportion of validated enhancers in the set. The overlapping areas are not proportional to the number of overlapping regions due to the asymmetry of the overlaps.

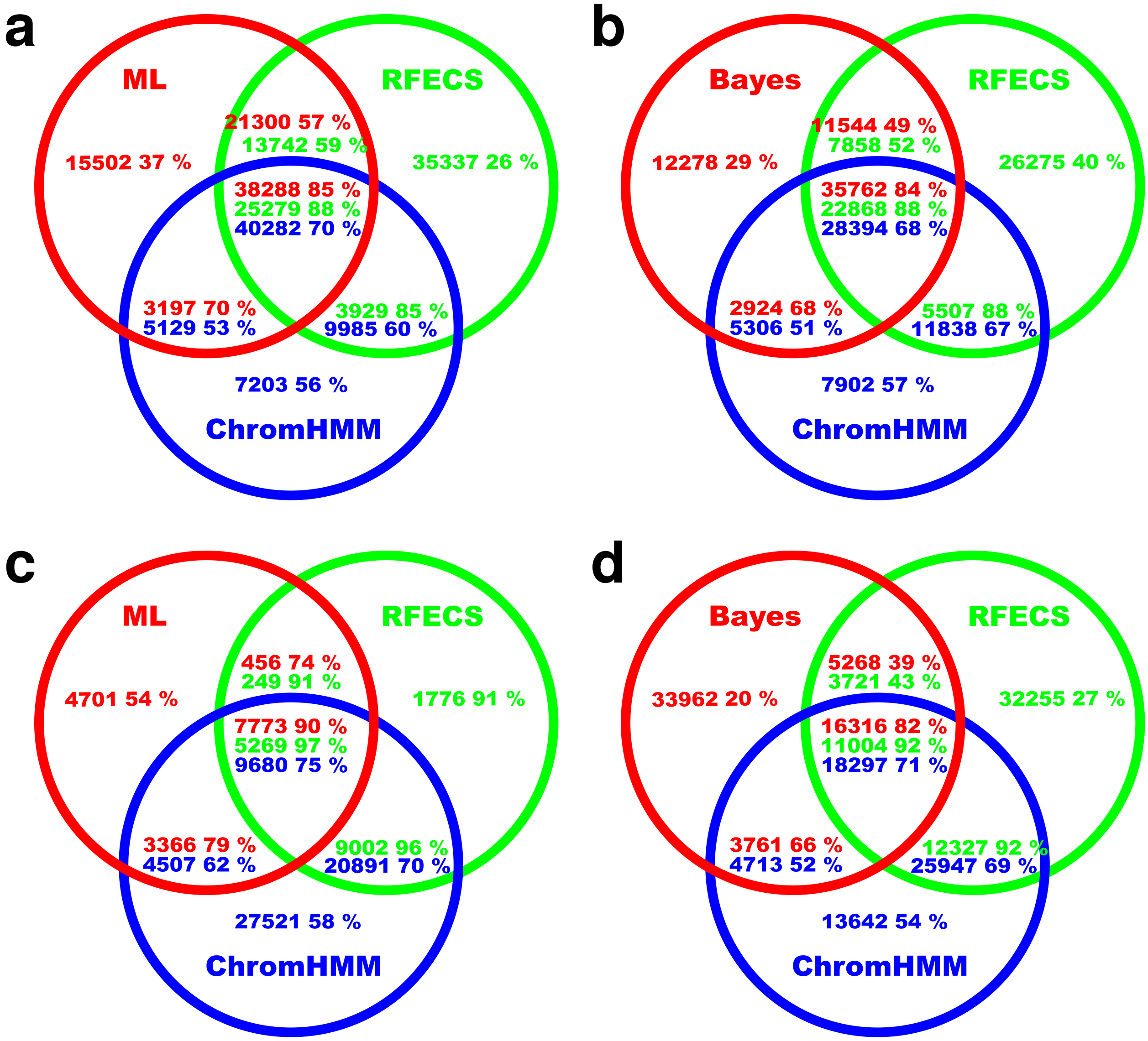

Figure S15: The unique and overlapping genome-wide enhancer predictions made by different methods in cell line GM12878. The PREPRINT enhancers were obtained using the 1% FPR threshold estimated from the GM12878 test data. In **a** and **c**, the predictions were obtained by the ML approach, and in **b** and **d**, the predictions were obtained by the Bayesian approach. The overlap between the PREPRINT, RFECS and ChromHMM predictions were quantified as the number of enhancers. In figures **a** and **b**, PREPRINT and RFECS were trained on the pure random regions, and in **c** and **d**, PREPRINT and RFECS were trained on the random regions with signal. In each figure, the number of enhancers predicted by PREPRINT and RFECS was the same. The number of enhancers was chosen to be the minimum number of enhancers predicted by either PREPRINT with the 1% FPR threshold or by RFECS with the lower threshold. The numbers enhancers were: **a** 62508, **b** 62508, **c** 59307, and **d** 59307. Inside every area, the number of enhancers belonging to the set are shown together with the proportion of validated enhancers in the set. The overlapping areas are not proportional to the number of overlapping regions due to the asymmetry of the overlaps.

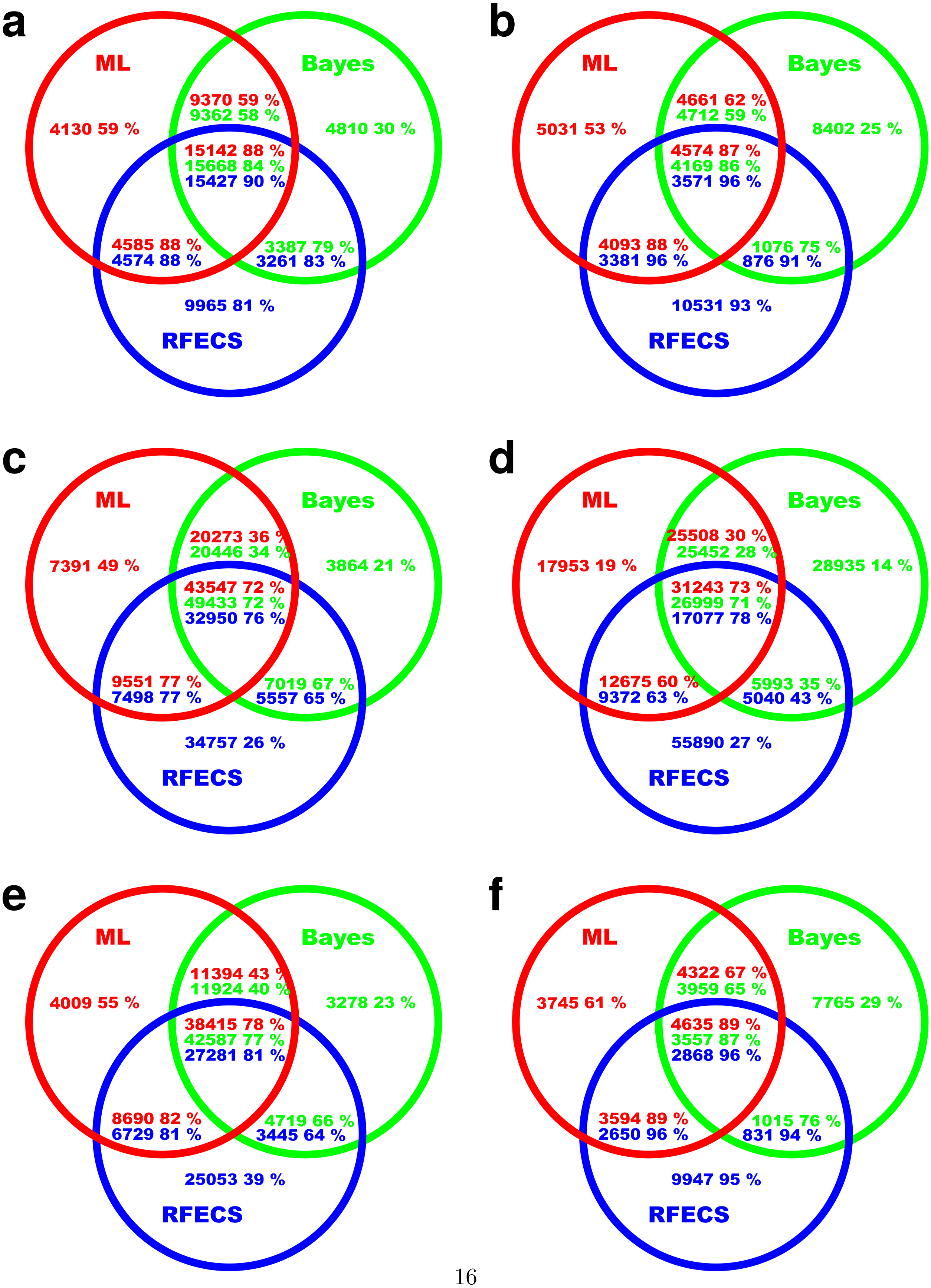

Figure S16: (Caption next page.)

Figure S16: (Previous page.) The unique and overlapping genome-wide enhancer predictions made by PREPRINT and RFECS in cell line GM12878. In figures **a**,**c**, and **e**, the methods were trained on the pure random regions, and in **b**, **d**, and **f**, the methods were trained on the random regions with signal. In each figure, the number of enhancers predicted by PREPRINT and RFECS was the same. In **a** and **b**, the number of enhancers was chosen to be the minimum number of enhancers predicted with threshold 0.5. In **c** and **d**, the number of enhancers was chosen to be the minimum number of enhancers predicted by either PREPRINT with the 1% FPR threshold or by RFECS with the lower threshold. The 1% FPR threshold estimated from the K562 CV data was used. In **e** and **f**, the number of enhancers was chosen to be the minimum number of enhancers predicted by either PREPRINT with the 1% FPR threshold or by RFECS with the lower threshold. The 1% FPR threshold estimated from the GM12878 test data was used. The numbers were **a** 33227, **b** 18359, **c** 80762, **d** 87379, **e** 62508, and **f** 59307. Inside every area, the number of enhancers belonging to the set are shown together with the proportion of validated enhancers in the set. The overlapping areas are not proportional to the number of overlapping regions due to the asymmetry of the overlaps.

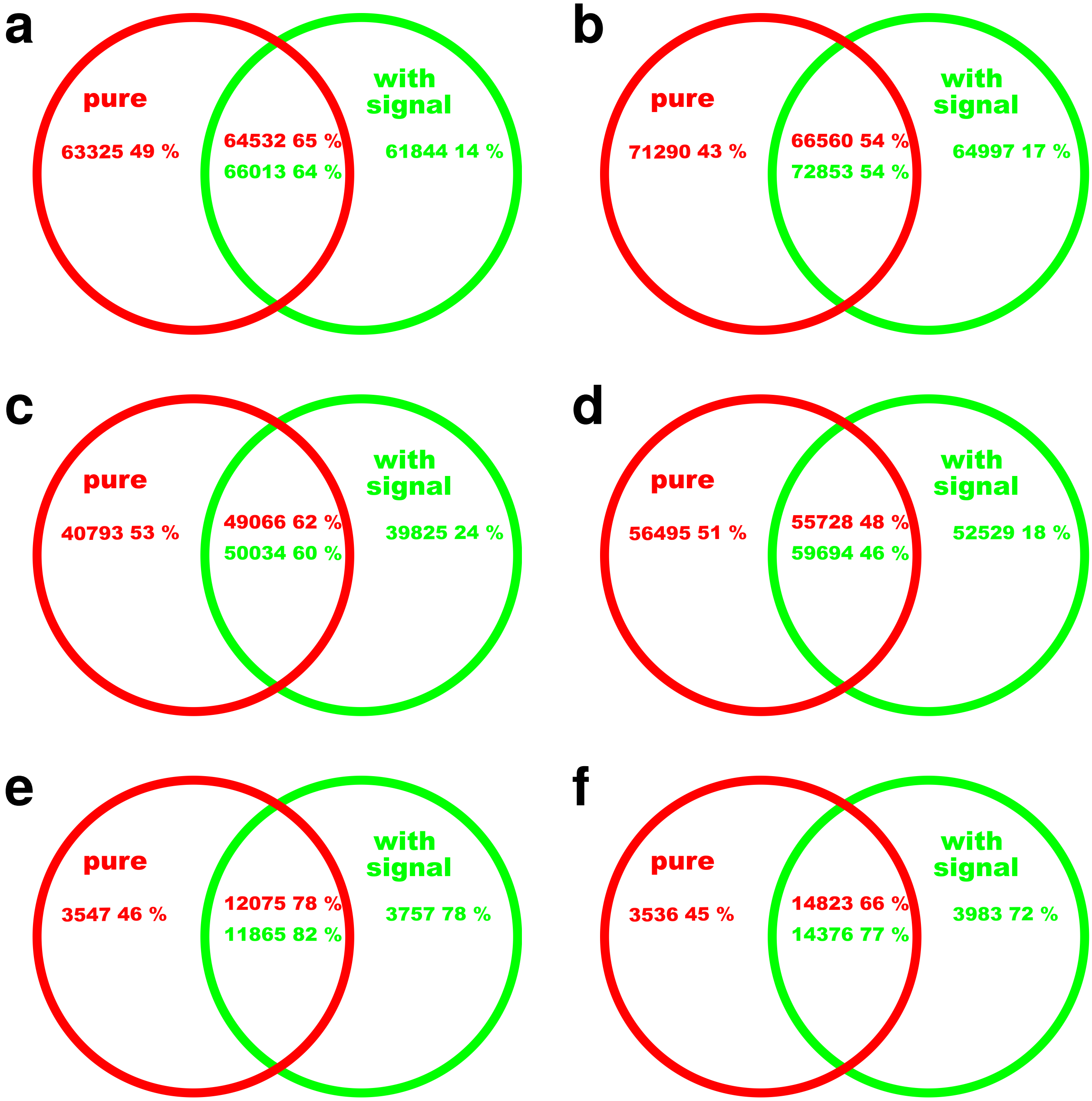

Figure S17: The Venn diagrams between the predictions obtained by the same method trained on the different random definitions. In each figure, the number of enhancers predicted in the two settings was the same. The minimum number of enhancers predicted with the 0.5 threshold were used. The comparisons were: **a**: PREPRINT with the ML approach in cell line K562, **b** PREPRINT with the ML approach in cell line GM12878, **c** PREPRINT with the Bayesian approach in cell line K562, **d**PREPRINT with the Bayesian approach in cell line GM12878, **e** RFECS in cell line K562, and **f** RFECS in cell line GM12878. Inside every area, the number of enhancers belonging to the set are shown together with the proportion of validated enhancers in the set. The overlapping areas are not proportional to the number of overlapping regions due to the asymmetry of the overlaps.