Figures

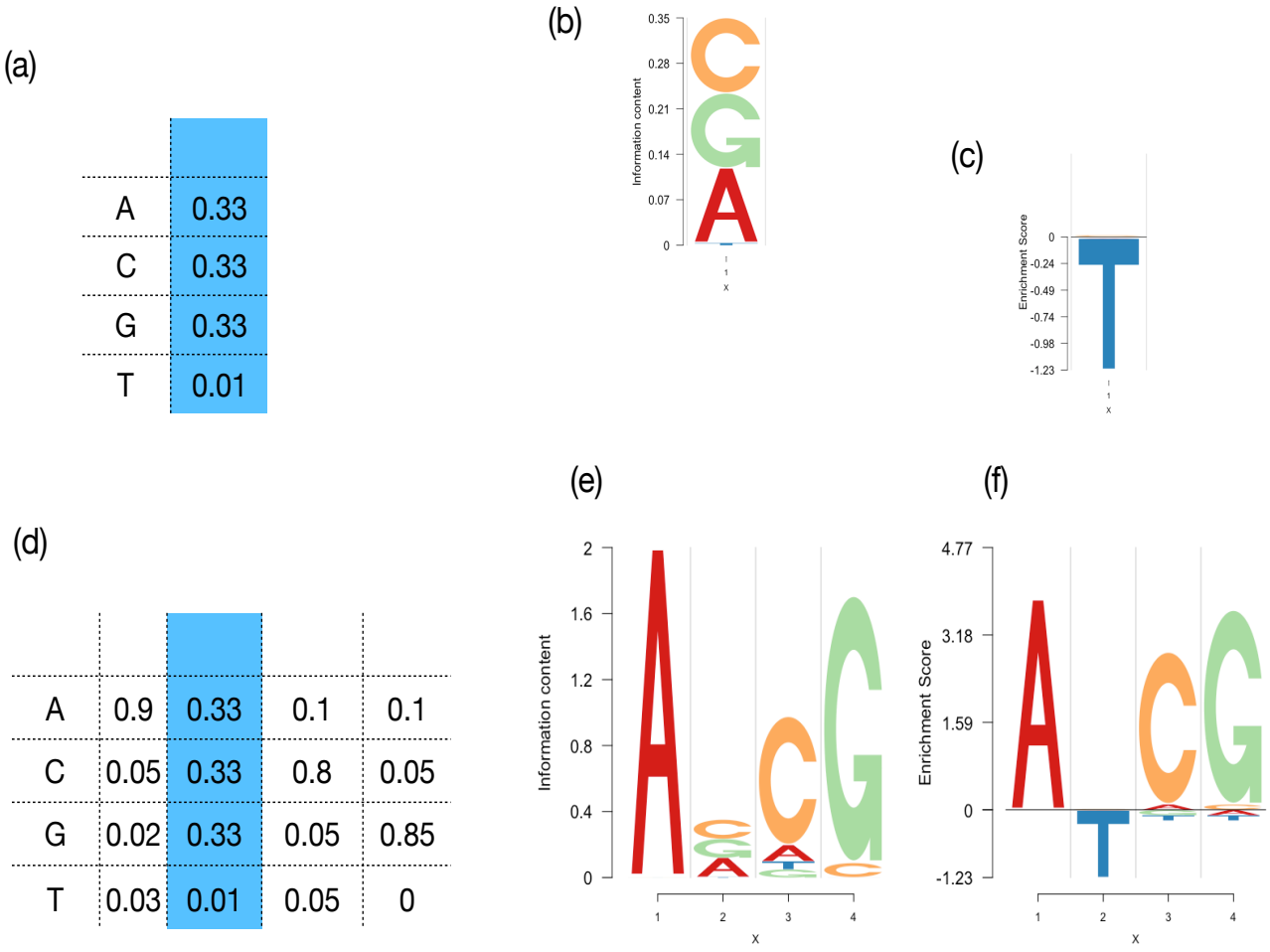
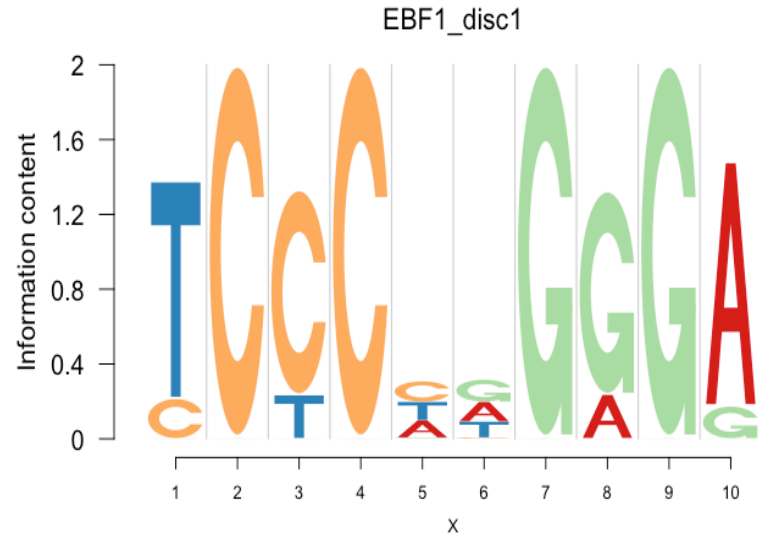


Fig 1. Illustration of the sparse logo representation. We present an illustration of how sparse logo representation accounts for depletion signal and provides a more informative visualization of the sequence motif. In panel (a), we present a positional weight vector with a depletion of the base T while the other bases have similar position weights. In panel (b), we present the standard logo representation of this positional vector that highlights the abundance of A, C and G at this position. In panel (c), we present the sparse logo representation (ratio approach) that highlights the depletion of T instead. In panel (d), we present a position weight matrix with the position weight vector in panel (a) at the second position flanked by enrichments around it. In panel (e), we present the corresponding standard logo plot representation of the PWM matrix at panel (d). The signal at the second position gets swamped by the bias towards enrichment signals flanking it. In panel (f), we present the sparse logo representation of the PWM matrix in (d), where both the enrichment signals as well as the depletion signal at position 2 are clearly observed.

(a)



(b)

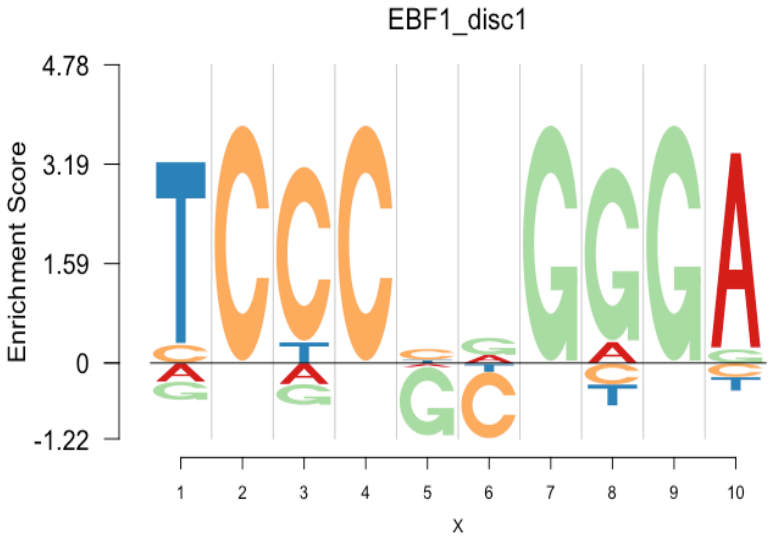


Fig 2. Application of sparse logo in detecting depletion patterns in Transcription Factor Binding sites (TFBS). We present the standard logo and the sparse logo representation of a transcription factor EBF1-disc1. The logo standard logo plot in panel (a) seems to indicate that the transcription factor binds in a dimerized form to its binding site. However, it fails to capture the depletion of G and C in the two positions in the middle of the dimer, which is apparently captured by the sparse logo representation in panel (b). The stack heights in the sparse logo representation in this plot has been determined by the *log* approach. The PWM data for EBF1-disc1, computed from the ENCODE TF Chip-seq datasets, is hosted on the webpage `http://compbio.mit.edu/encode-motifs/` [?]

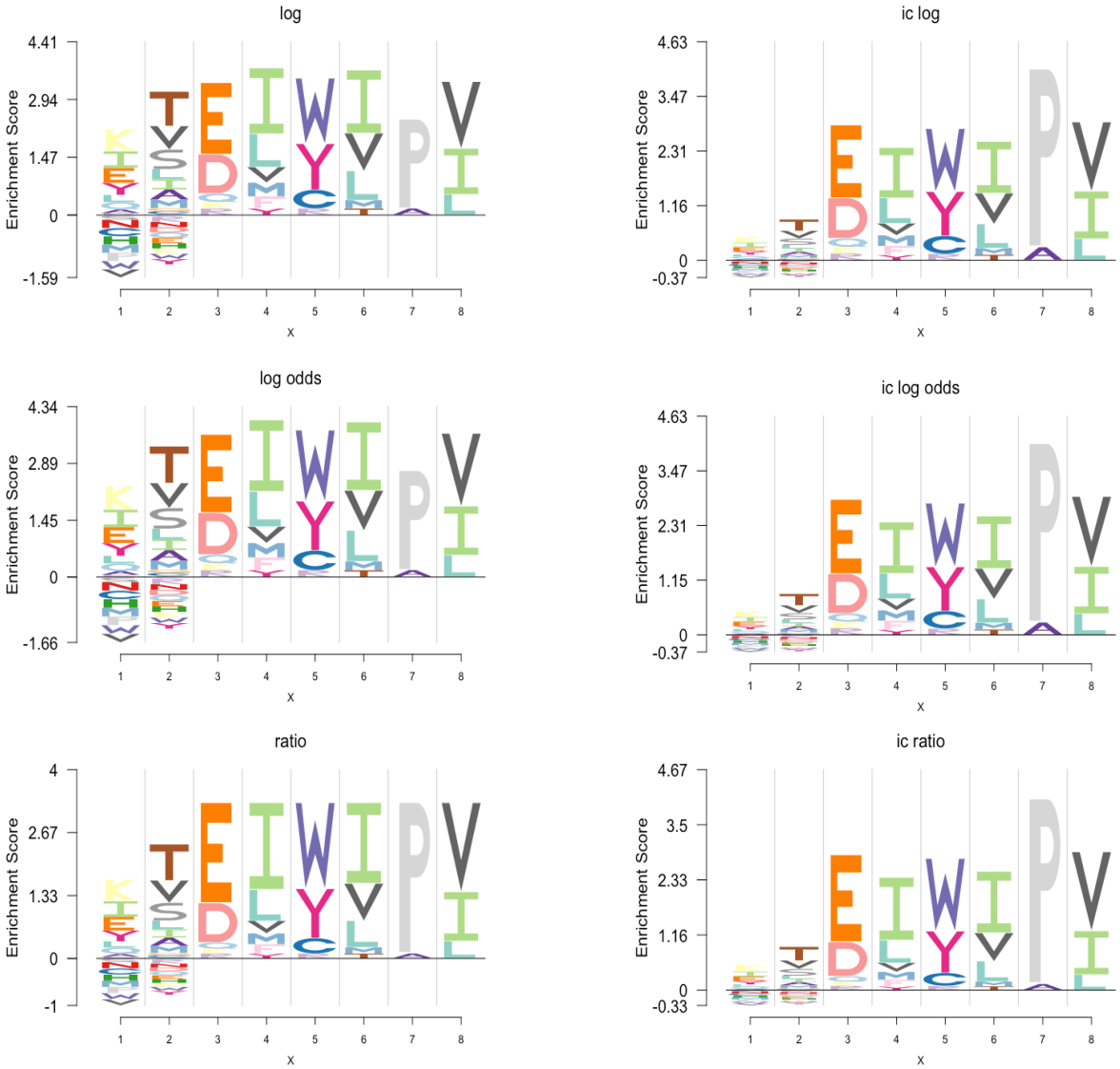
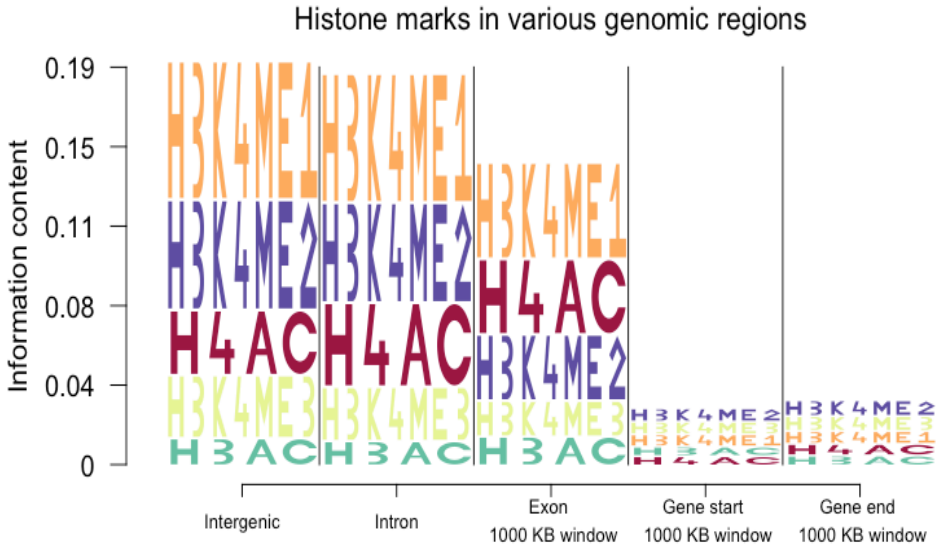


Fig 3. Various sparse logo representations of protein sequence motif. The sparse logo representation under various stack height and stack composition methods - *log*, *log-odds*, *ratio*, *ic-log*, *ic-log-odds* and *ic-ratio* for the Bacterial transcription activator, effector binding domain protein PF06445 (motif 4, Start=153 Length=8). The data is fetched from the 3PFDB website `http://caps.ncbs.res.in/cgi-bin/mini/databases/3pfdb/get_entry.cgi?id=PF06445`.

(a)



(b)

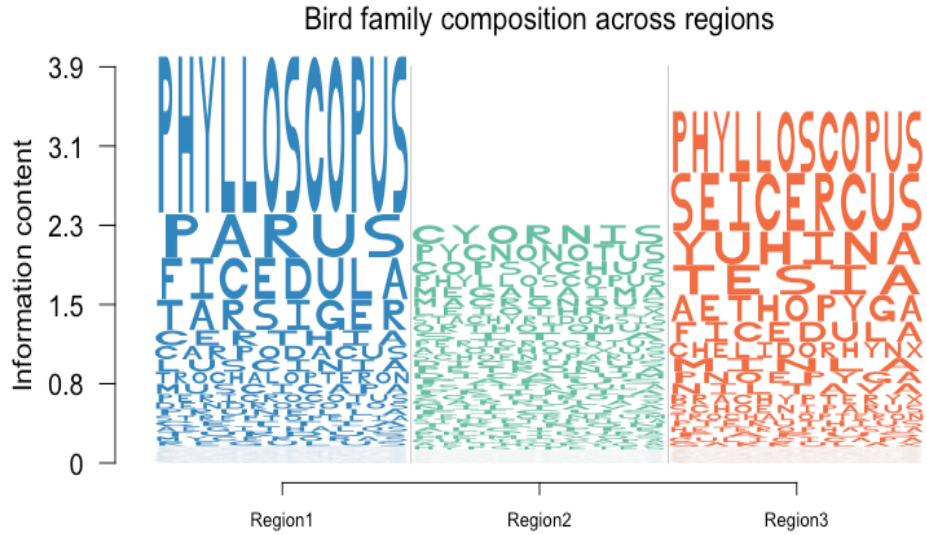


Fig 4. Example applications of string logo plots. We present two example biological applications of string logo plots. In panel (a), we present the logo plot representation of the composition of histone modification types that overlap with an intergenic region, intron, exon, gene start or gene end for the lymphoblastoid cell line GM06990 as reported in Koch et al 2007 [**?**]. In panel (b), we present the logo plot representation of the abundance compositions of bird species families in three different regions of the Himalayas - colored red, green and blue. The data has been taken from White et al 2017 [ref].

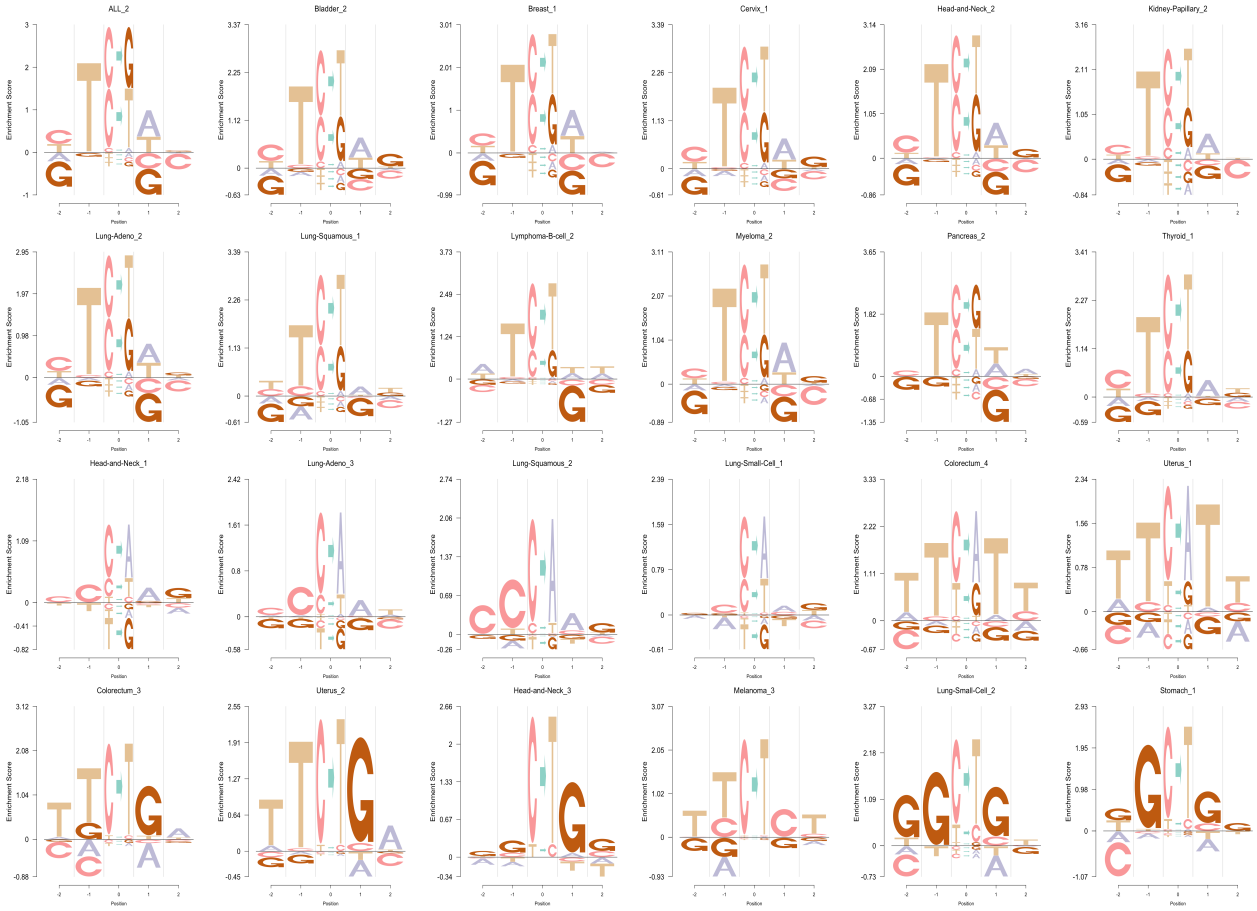


Fig 5. Logolas representation of cancer mutational signature profiles in alexandrov et al (2013). We present the sparse logo representations of the cancer mutational signature profiles across a number of tissues where the mutational signature data has been collected from 7042 cancers by Alexandrov et al (2013) [**?**]. Each mutational signature profile has the mutation type at the center and is flanked by two bases to the left and two bases to the right.

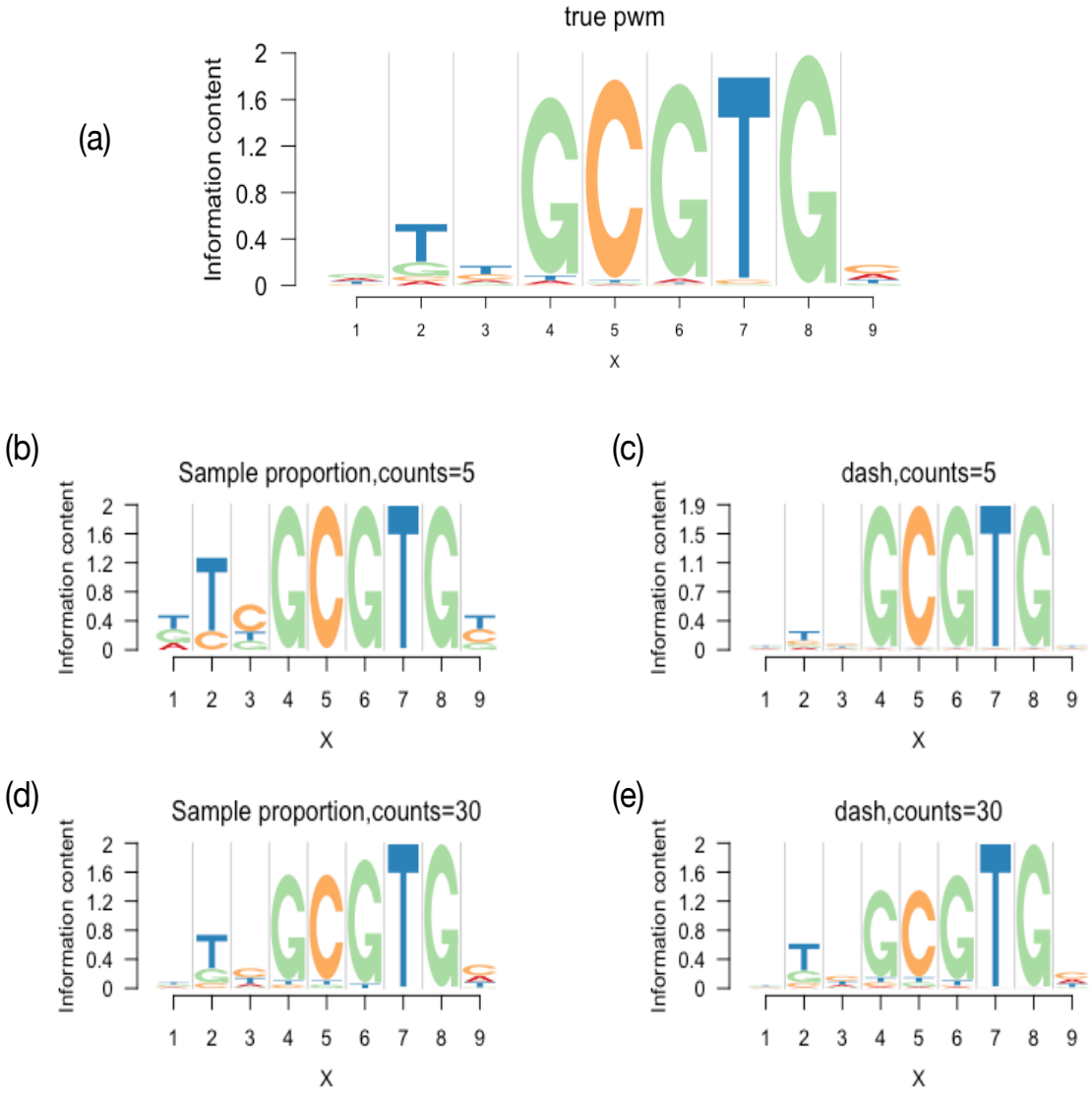


Fig 6. Subsampling experiment to validate the performance of Dirichlet Adaptive Shrinkage (dash). In panel (a), we present the standard logo plot representation of the position frequency matrix (PFM) of the Aryl hydrocardon receptor. This PFM matrix is then used to define the position weights and two random subsamples, A of size 5 and B of size 30 are generated from this position weight matrix. Panels (b) and (c) demonstrate the logo plot representation estimated position weight matrix from the 5 symbols in the subsample A before applying dash and post dash scaling. Panels (d) and (e) show the same results for subsample B, with size 30. We notice that for subsample A case, the dash scaled PWM in panel (c) is a closer approximation of the original PWM in panel (a) compared to the pre-dash version in panel (b). However, the effect of dash is comparatively lower for subsample B, since the subsample size (30) is much larger and both the pre-dash and post-dash PWMs of subsampled symbols in panels (d) and (e) are good approximations to the original PWM.