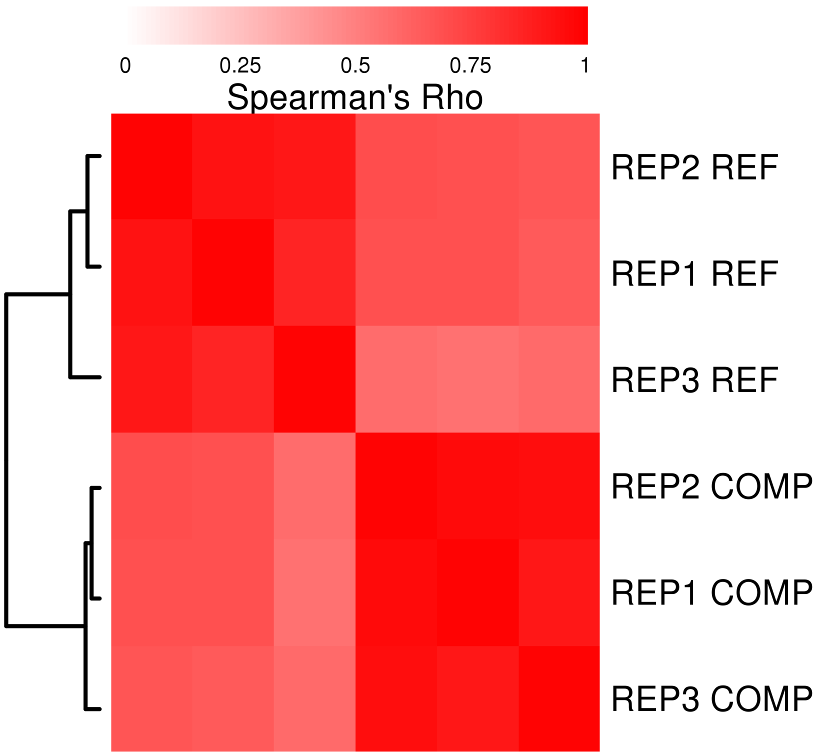
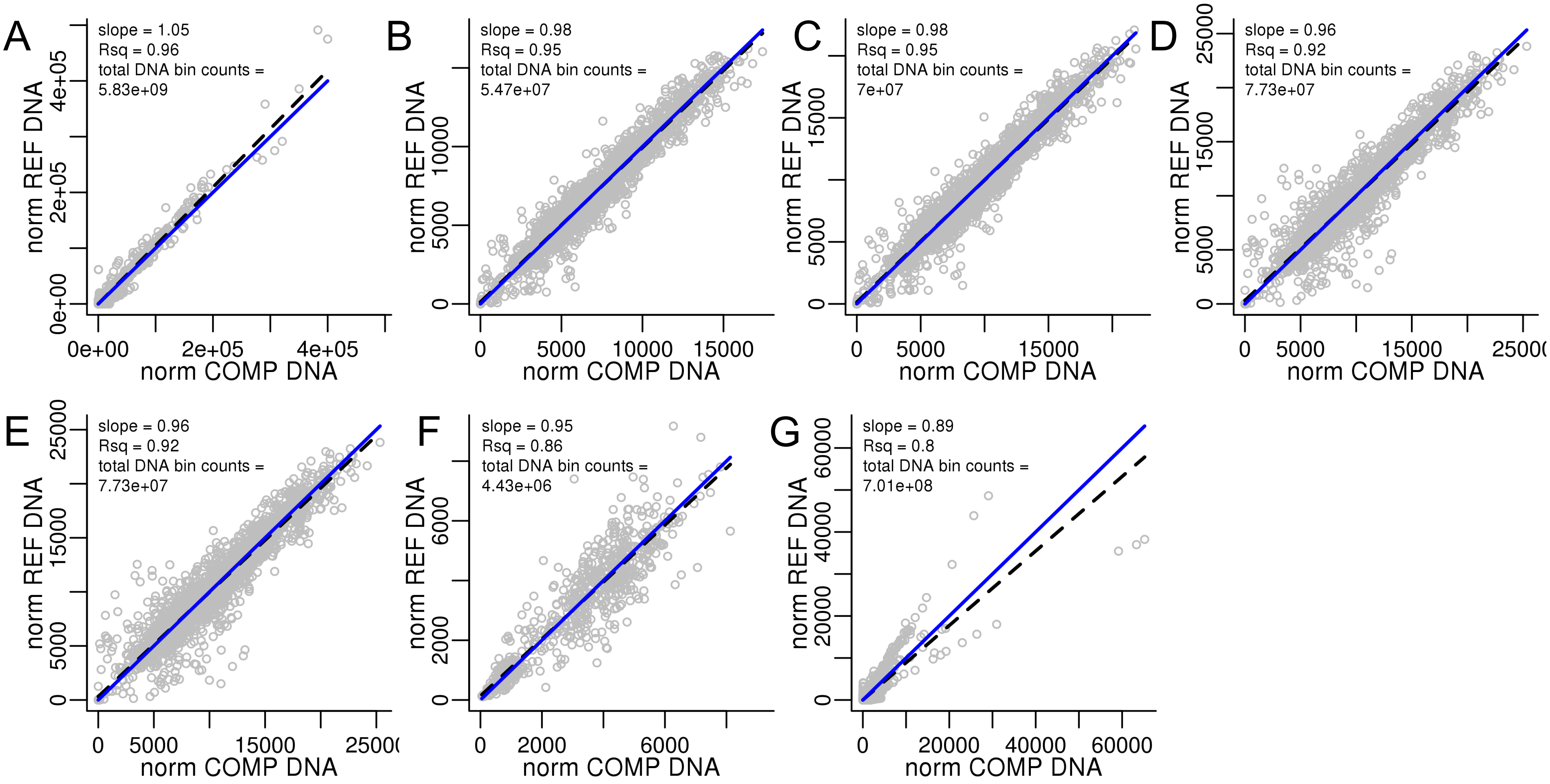
region	chr	start	end	BAC or fosmid ID
HTT locus	chr4	2709241	2881534	RP11-167G4
HTT locus	chr4	2838556	2981627	CTD-2314C7
HTT locus	chr4	2967192	3009931	WI2-2391F2
HTT locus	chr4	3001985	3201056	RP11-1069C14
HTT locus	chr4	3174998	3307130	CTD-2050N17
HTT locus	chr4	3303952	3476944	RP11-194F20
HTT locus	chr4	3360982	3536145	RP11-1142N16
HTT locus	chr4	3478122	3669683	RP11-1079H13
HTT locus	chr4	3590004	3780954	RP11-717M10
HTT locus	chr4	3778247	3814359	WI2-2977B8
HTT locus	chr4	3805281	3847050	WI2-1839N15
HTT locus	chr4	3840493	3884992	WI2-2269K14
HTT locus	chr4	3878490	3943231	CTD-2255O16
SORT1 locus	chr1	109267735	109416464	RP11-47M16
SORT1 locus	chr1	109197415	109400739	RP11-463O24

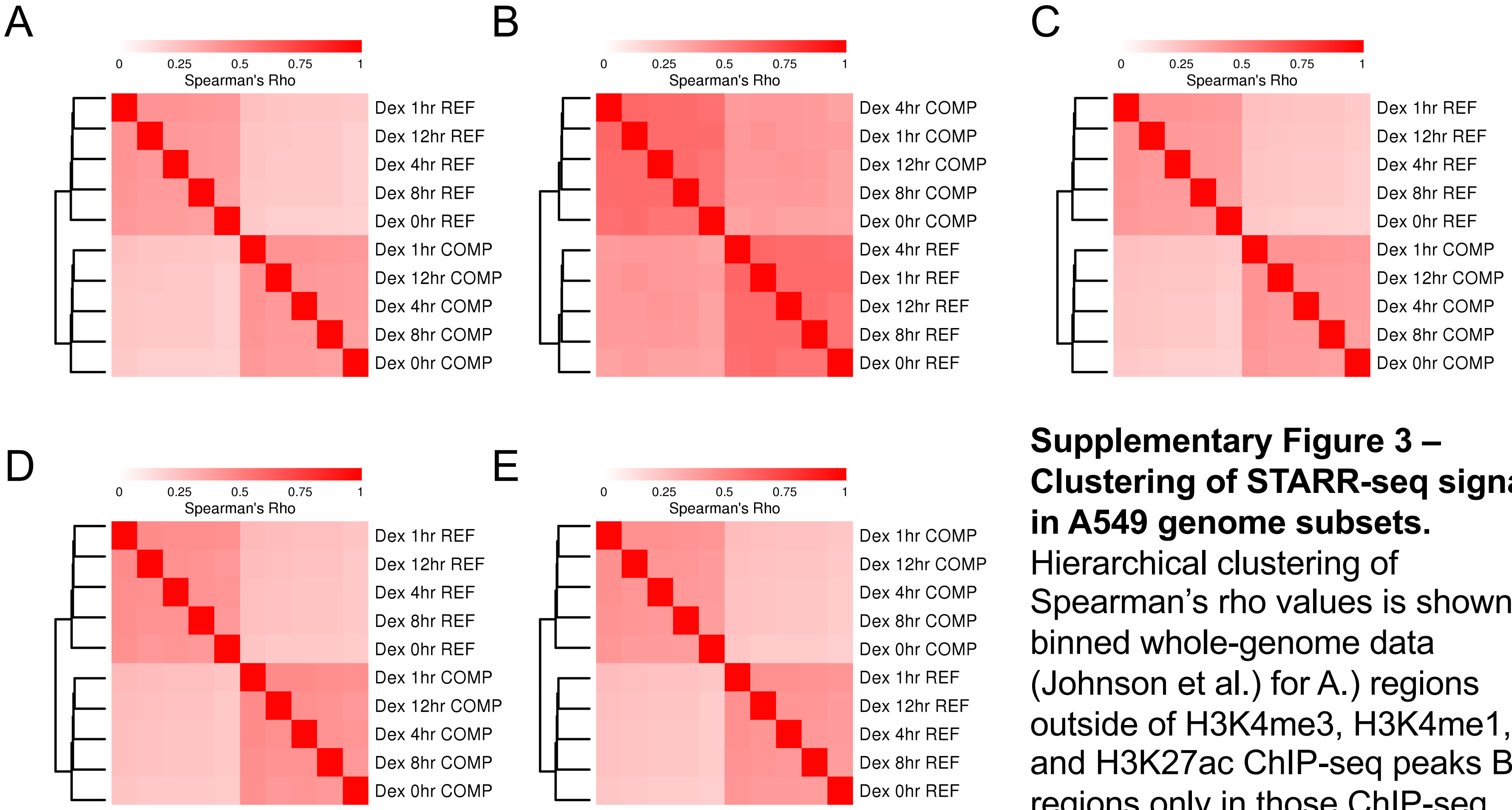
Supplementary Table 1 – BAC information.
Genomic position (hg38), BAC ID, and the associated locus are given for the BAC-derived STARR-seq libraries.



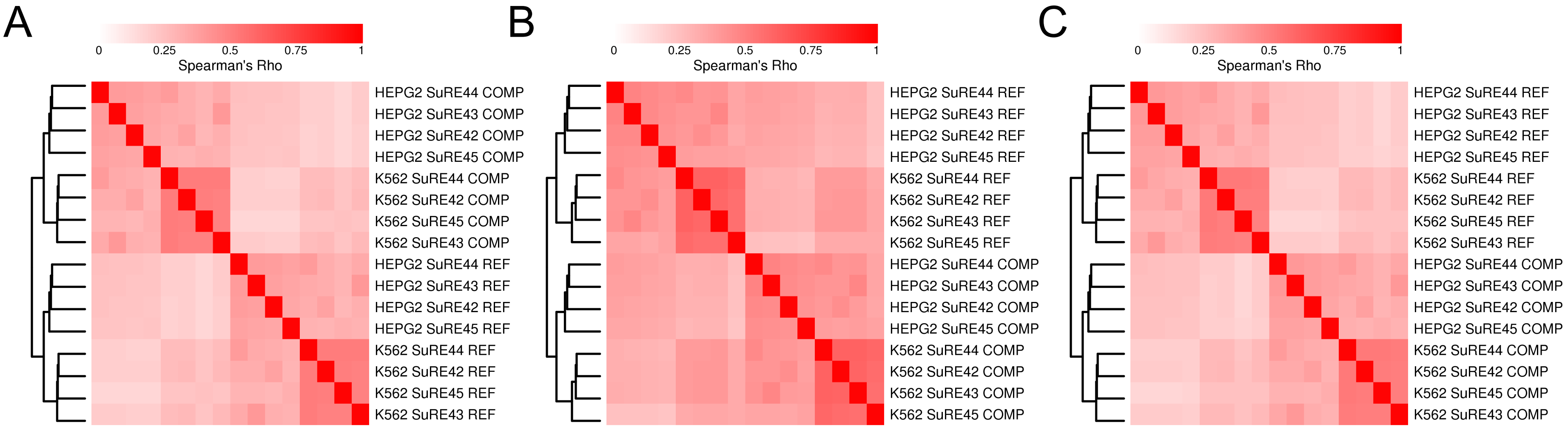
Supplementary Figure 1 – Clustering of STARR-seq signal in the *SORT1* locus. Hierarchical clustering of Spearman's rho values is shown in binned data from HepG2 cells derived from BACs spanning the *SORT1* gene locus.



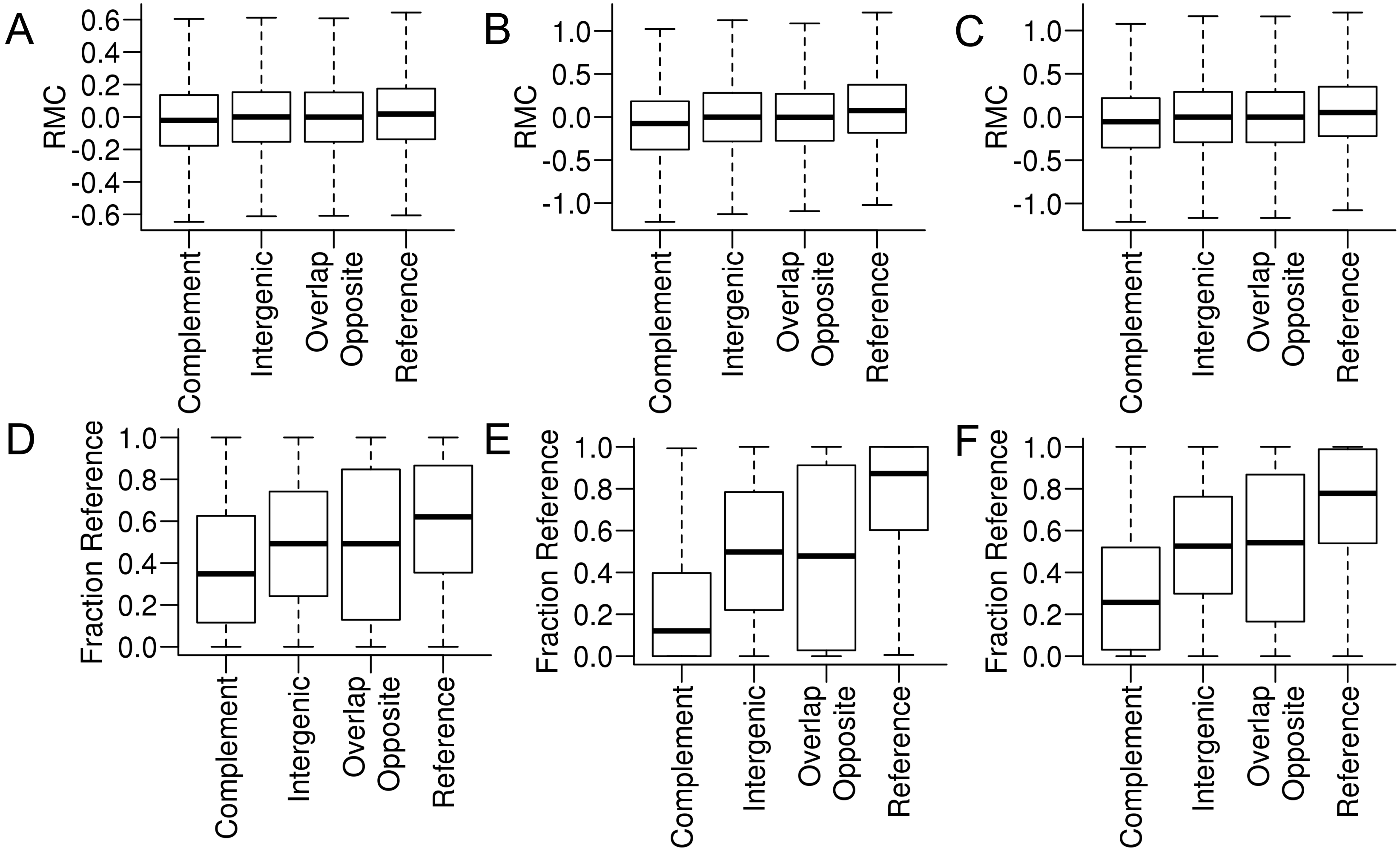
Supplementary Figure 2 – Agreement between stranded normalized reporter DNA levels. Normalized binned DNA counts from the Reference (REF) strand are plotted versus the Complement (COMP) strand for A.) A549 whole genome STARR-seq data from Johnson et al., B-E.) A549, BE(2)-C, HepG2, and K562 STARR-seq data derived from BACs spanning the *HTT* gene locus, F.) HepG2 STARR-seq data derived from BACs spanning the *SORT1* gene locus, and G.) LNCaP STARR-seq data from Liu et al. The solid blue line is y=x and the dashed black line is the linear model fit.



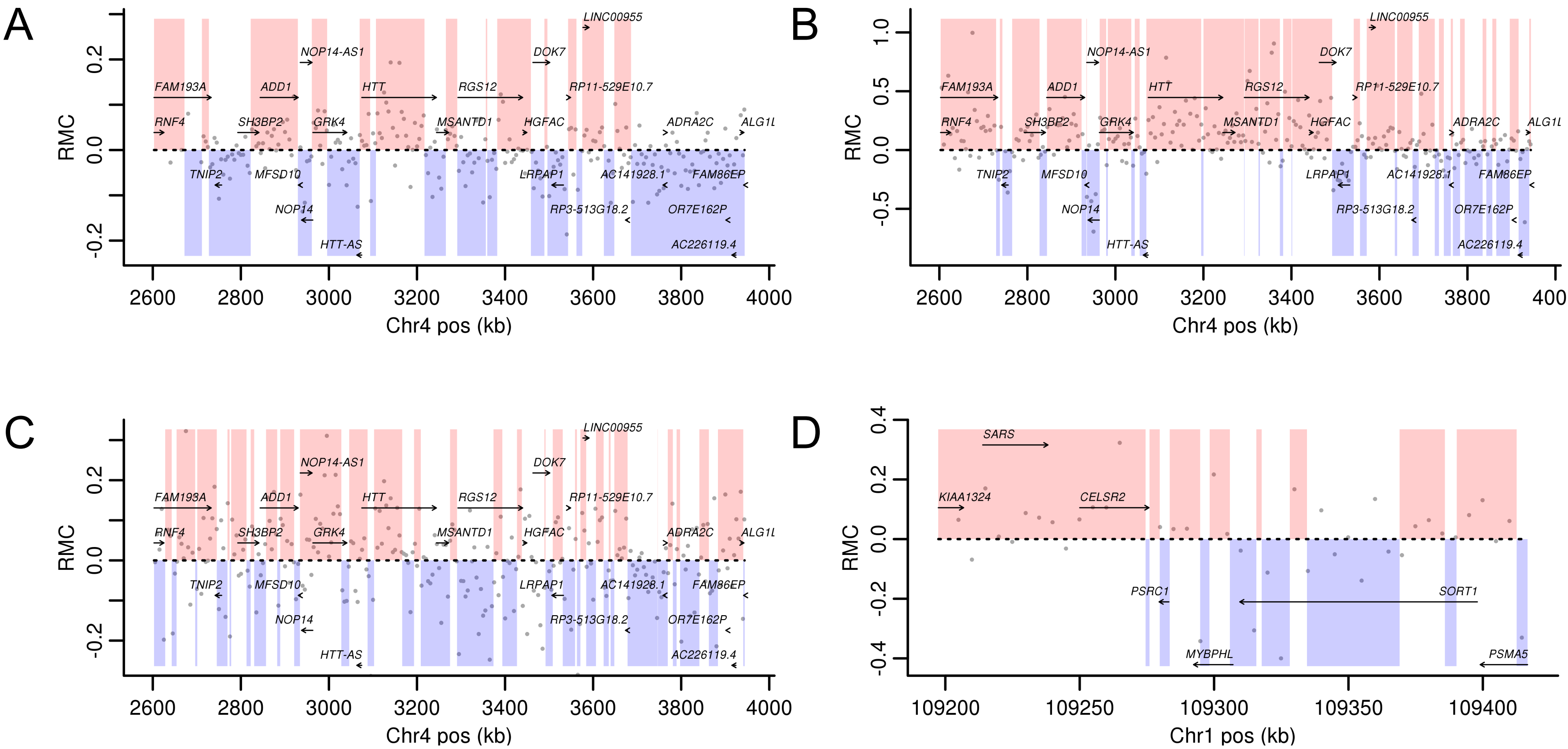
Supplementary Figure 3 – Clustering of STARR-seq signal in A549 genome subsets. Hierarchical clustering of Spearman’s rho values is shown in binned whole-genome data (Johnson et al.) for A.) regions outside of H3K4me3, H3K4me1, and H3K27ac ChIP-seq peaks B.) regions only in those ChIP-seq peaks C.) regions outside of gene bodies D.) regions within gene bodies E.) 1 million randomly sampled bins.



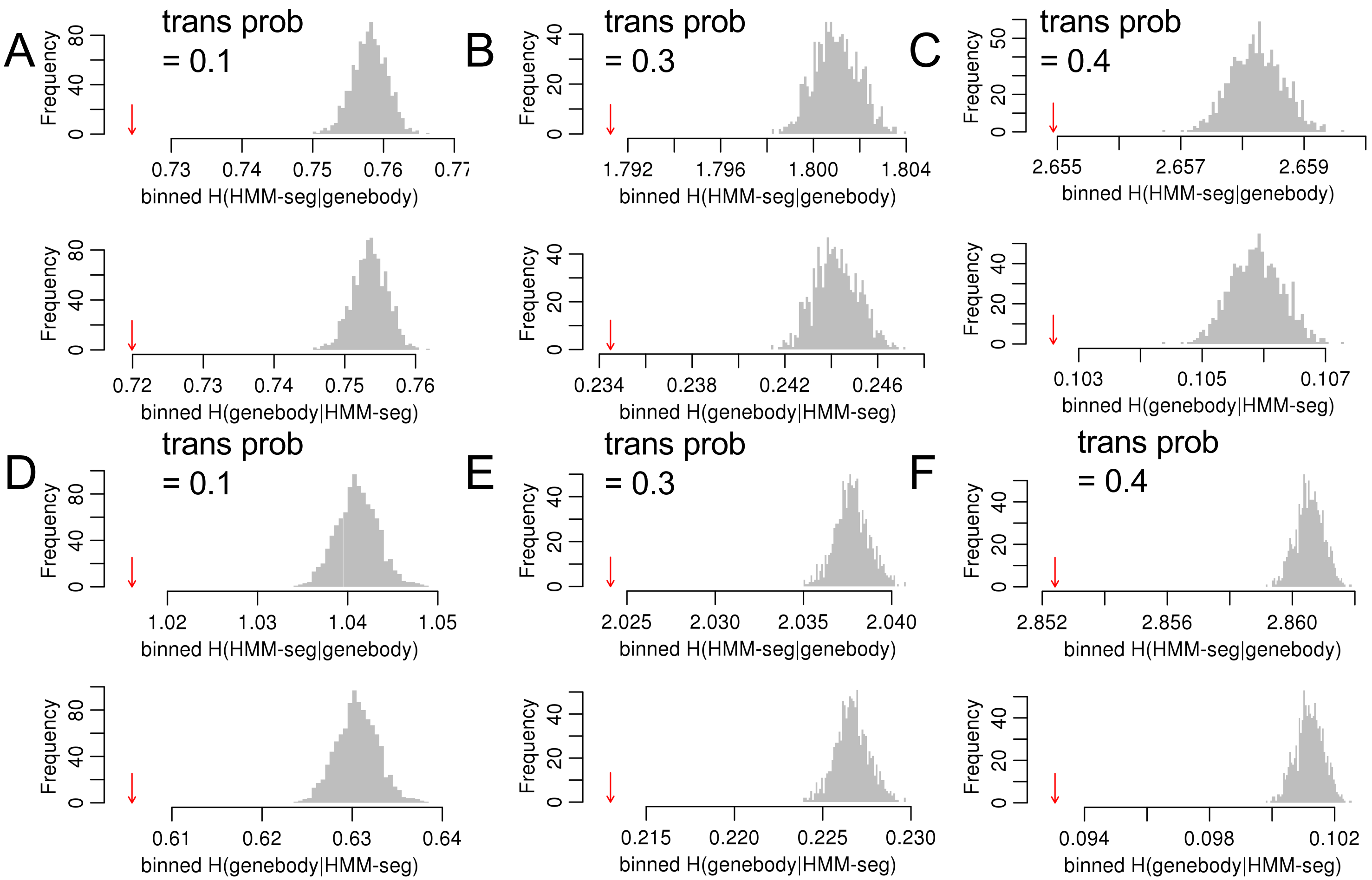
Supplementary Figure 4 – Clustering of promoter-less MPRA signal. Hierarchical clustering of Spearman’s rho values is shown in binned whole-genome in two cell types from data from Van Arensbergen et al. for A.) regions excluding promoters, B.) only promoter regions, and C.) 1 million randomly sampled bins. Promoter regions were defined as 2 kb upstream and 500 bp downstream of annotated TSSs (GTex v8).



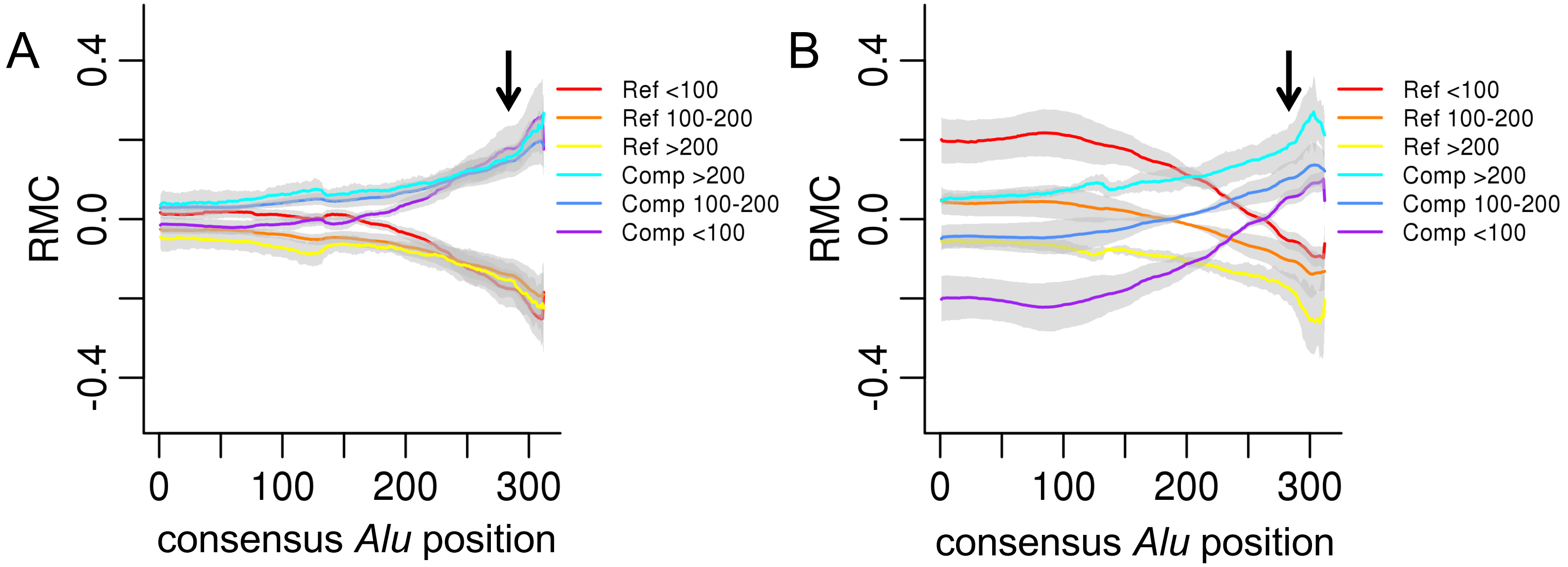
Supplementary Figure 5 – MPRA strand asymmetry effects in gene region types. Boxplots for RMC in A.) A549 cells from Johnson et al., B.) HepG2 cells or C.) K562 cells from Van Arensbergen et al. For D-F, the plots show the fraction of the gene region segmented as Reference by HMMSeg applied to the same data as in A-C. The center line represents the median, the boxes define the interquartile range, and the whiskers mark the most extreme point no greater than 1.5 times the interquartile range. Only regions greater 10 kb are plotted. The transition probability used in HMMSeg-derived plots is 0.3. Similar results were obtained for transition probabilities from 0.05 to 0.4.



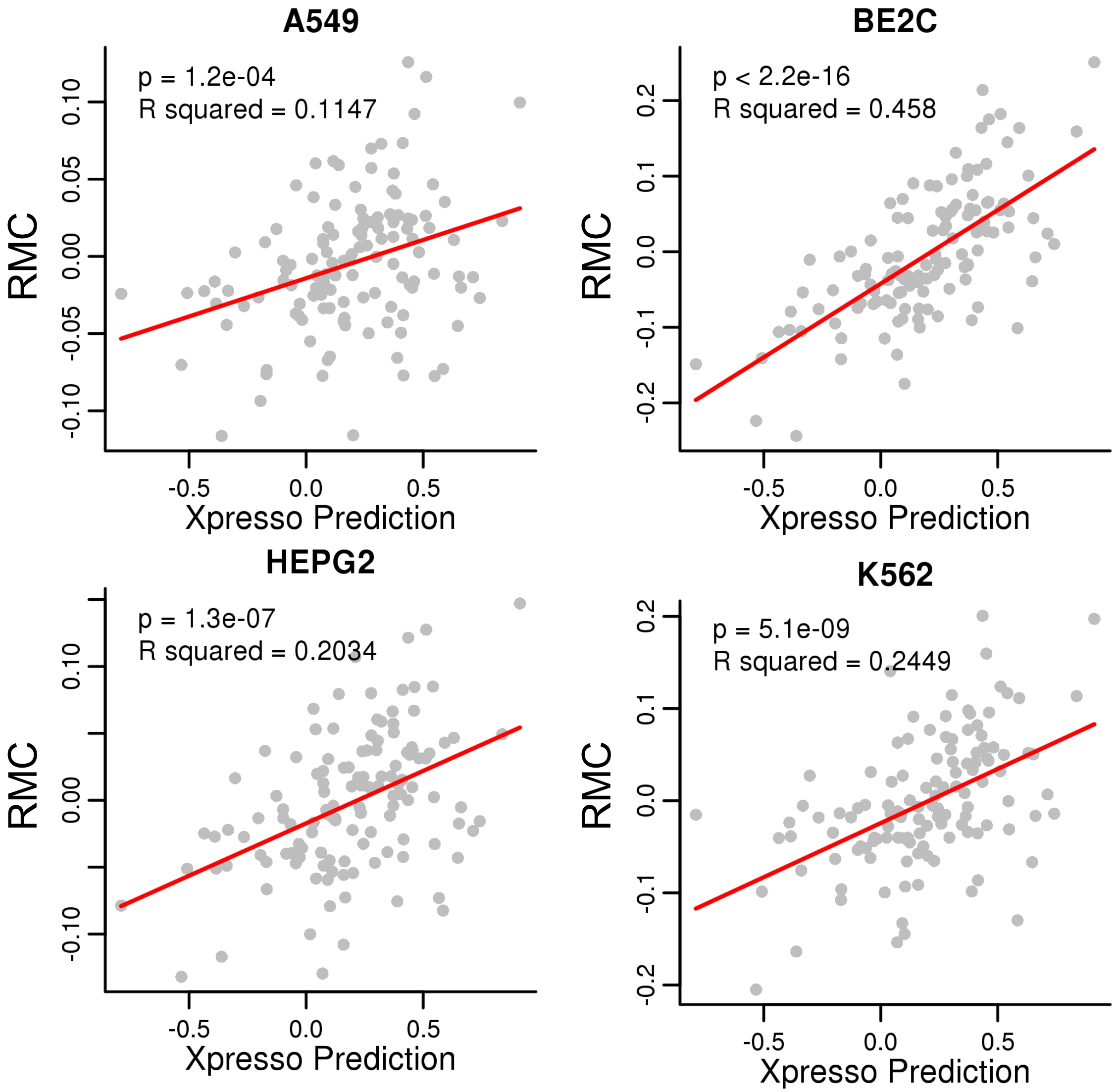
Supplementary Figure 6 – Correlation of MPRA strand asymmetry in multiple cell and assay types. RMC values are shown for A.) STARR-seq data from BACs in the *HTT* locus in A549 cells, B.) Arensbergen et al. whole genome data in the *HTT* locus for HepG2 cells, C.) STARR-seq whole genome data from Johnson et al. in the *HTT* locus in A549 cells, or D.) BAC-generated STARR-seq in the SORT1 locus in HepG2 cells. Gray dots are average RMC values in 5 kb windows. Pink and blue blocks were assigned to Reference and Complement, respectively by HMMSeg. For HMMSeg calculation, a transition probability of 0.3 was used for all data sets.



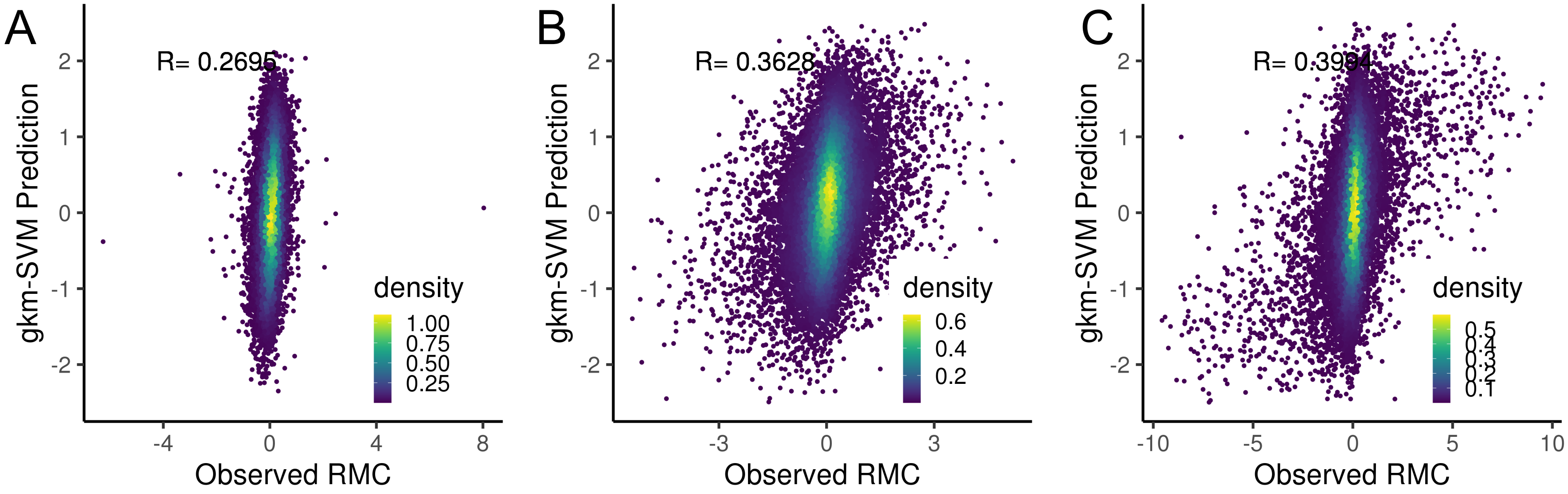
Supplementary Figure 7 – HMM-seg segmentation of the autosome by MPRA strand asymmetry is significantly similar to gene bodies at multiple transition probabilities. The gray histograms show the values of the conditional entropy (H) of the HMM-seg generated segmentation given Reference gene body segmentation (top) or the converse (bottom) from 1000 random shuffles of the HMM-seg segmentation (see Methods). The transition probability used in HMM-seg is listed in each plot pair. The H of the actual HMM-seg is shown as the red arrow. A-C are for segmentations generated from A549 cells in the Johnson et al. dataset. D-F are for segmentations generated from K562 in the Van Arensbergen et al. dataset. Similar plots were seen for Complement genes and in HepG2 cells in the Van Arensbergen et al. dataset.



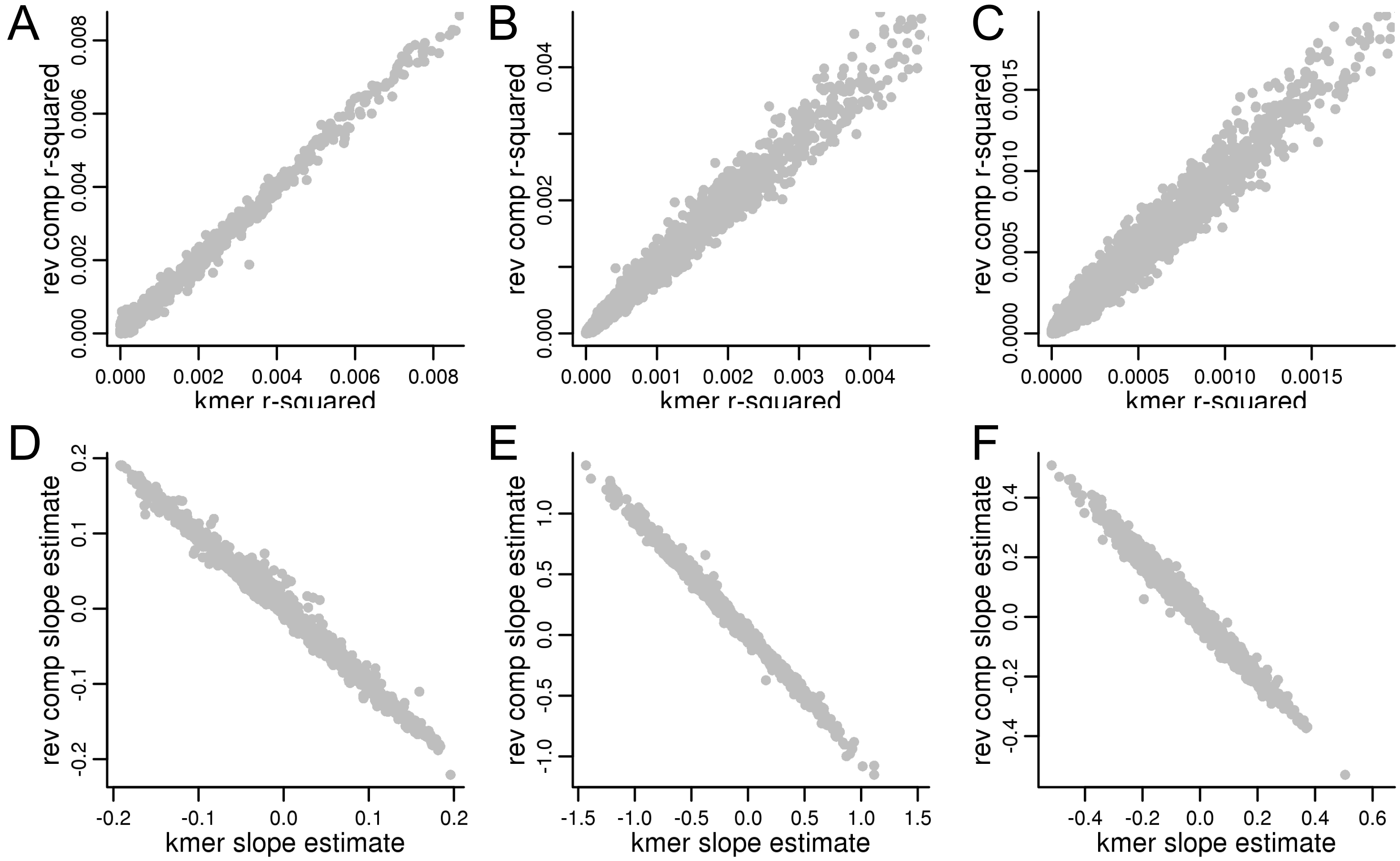
Supplementary Figure 8. Effects at *Alu* consensus positions in a MPRA with an upstream test element. From Van Arensbergen et al. data, the genome-wide median RMC (y-axis) for each annotated *Alu* consensus position (x-axis) is plotted for Reference- (Ref) or Complement- (Comp) oriented *Alu* insertions, grouped by levels of divergence (indicated in respective colors) measured by milliDiv units (e.g., < 100 corresponds to <10% divergence from the ancestral consensus, see Methods). Data is shown from A.) K562 cells, and B.) HepG2 cells. The gray bands represent two standard deviations from the median. The black arrow indicates the start of the A-tail sequence.



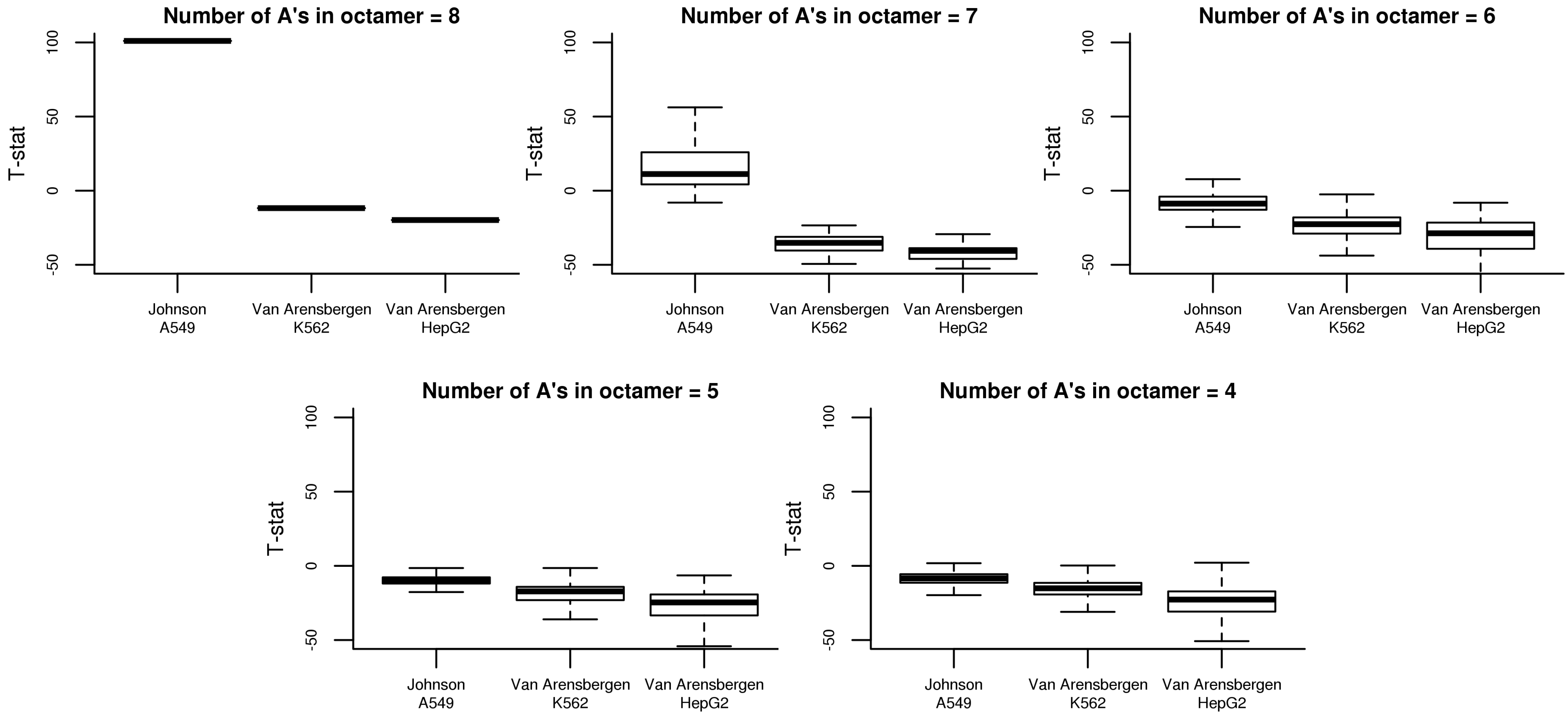
Supplementary Figure 9 – Xpresso transcriptional activity predictions correlate with RMC. Plotted are Xpresso predictions versus RMC from 4 cell types in the *HTT* locus. Both sets of values were mapped to 10 kb bins, corresponding to the sequence input size to Xpresso. A linear regression p-value and R squared is given for each cell type.



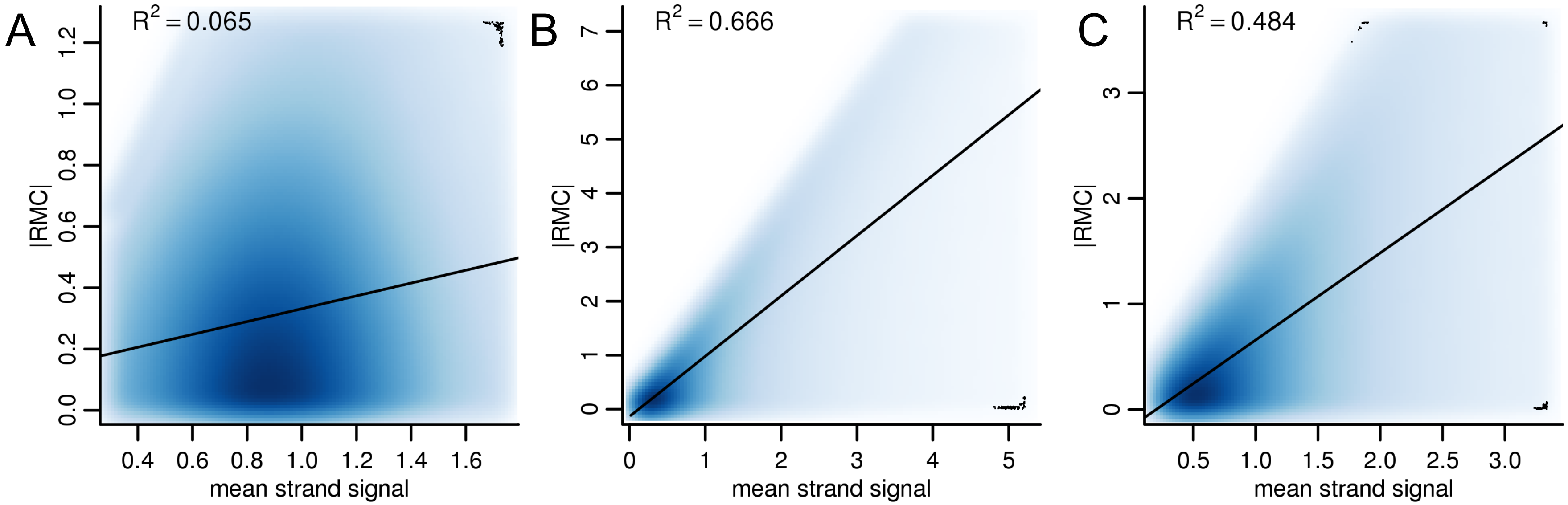
Supplementary Figure 10 – RMC prediction by gkm-SVM. Plotted are observed RMC values versus gkm-SVM predicted values for A.) Johnson et al. A549, B.) Van Arensbergen et al. K562 and C.) HepG2. To train the model, we selected 10000 regions with positive RMC and 10000 regions with negative RMC. The predictions were then made on an independent, randomly sampled set of 20000 regions. The Pearson R value is given on each plot.



Supplementary Figure 11 – Correlation of individual k-mers with RMC is equal and opposite in their reverse complement sequence. The r-squared value of the linear regression of a given k-mer with chromosome 1 RMC is plotted versus the r-squared of its reverse complement for A.) Johnson et al. data, B.) Van Arensbergen et al. HepG2 data, and C.) Van Arensbergen et al. K562 data. The slope estimate from the same linear regressions for a given k-mer is plotted versus its reverse complement for the datasets in D-F in the same order as A-C. For A-C, the axes maximums equal the 99th percentile of the data.



Supplementary Figure 12. Very high A content drives opposing effects in upstream and downstream MPRA data. Boxplots of the T-statistic from a linear regression of a given octamer's frequency versus chromosome 1 RMC values are shown for each of the indicated datasets (x-axis). The boxplots are for octamers containing 4-8 A's as indicated by each plot title. All plots have the same y-axis scale.



Supplementary Figure 13. Correlation of RMC magnitude with reporter signal varies by assay type. In each plot, the mean strand signal, calculated as the mean of the Reference and Complement normalized RNA to DNA counts ratios, is plotted versus the absolute value of RMC for A.) Johnson et al. data, B.) Van Arensbergen et al. HepG2 data, and C.) Van Arensbergen et al. K562 data. The fit line and r-squared value from a linear regression is shown for each dataset. Darker blue regions indicate a higher density of data points. Axes upper limits restricted to the 99th percentile of data.