Accounting for GC-content bias reduces systematic errors and batch effects in ChIP-Seq data

Mingxiang Teng and Rafael A. Irizarry

Correspondence: rafa@jimmy.harvard.edu

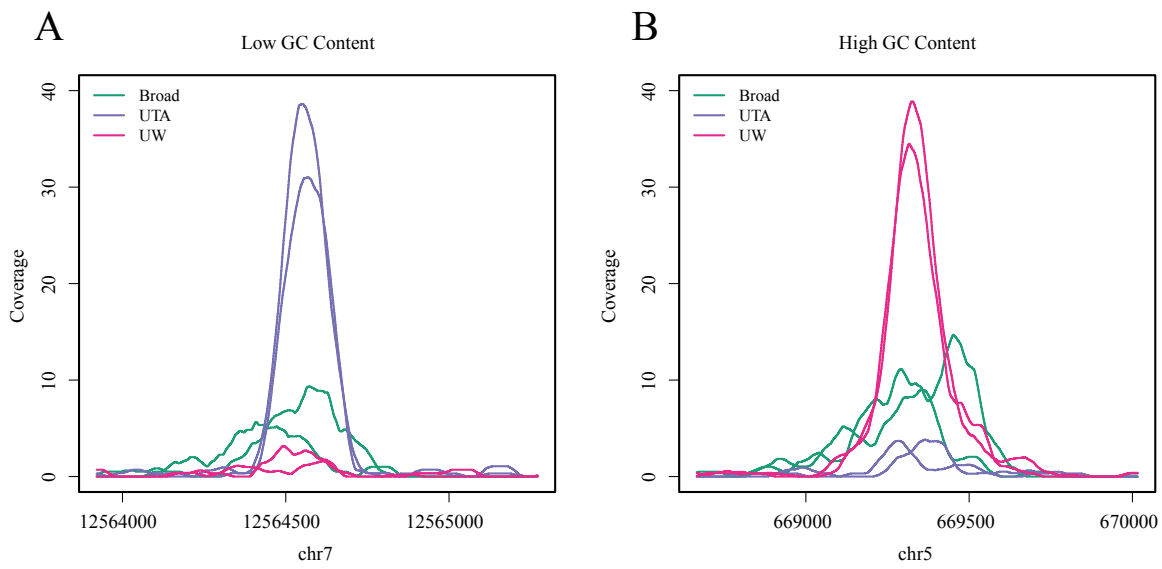

Figure S1 – Smoothed coverage plots for example regions shown in Figure 1G (A) and Figure 1H (B). Sliding window of 100bp is applied for smoothing.

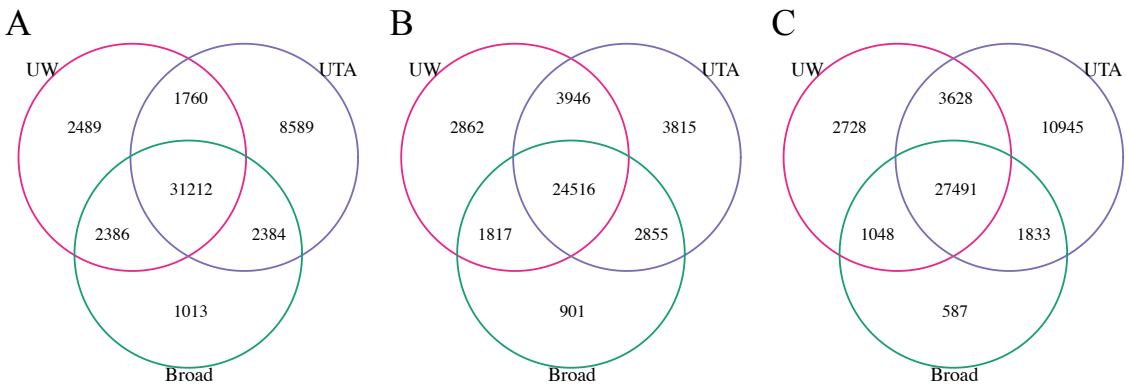

Figure S2 – Overlaps of peaks by different laboratories reported by ENCODE portal (A), by our algorithm with IDR (B) and by SPP with IDR (C). A uniform list of peaks was first generated by combining peaks from different labs and merging potential overlapped peaks. A Venn diagram was generated using overlaps between the uniform list and peaks from individual labs.

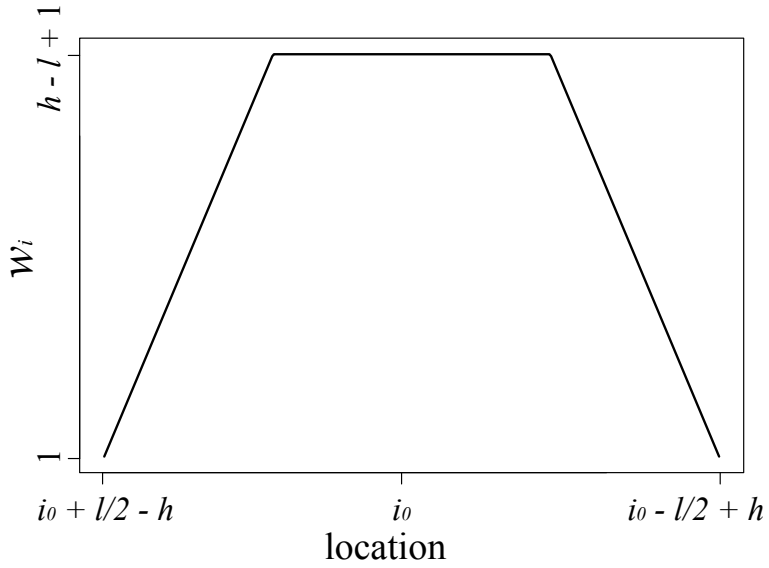

Figure S3 – Illustration of nucleotide weights when calculating effective GC content for a bin centered at location i_0.

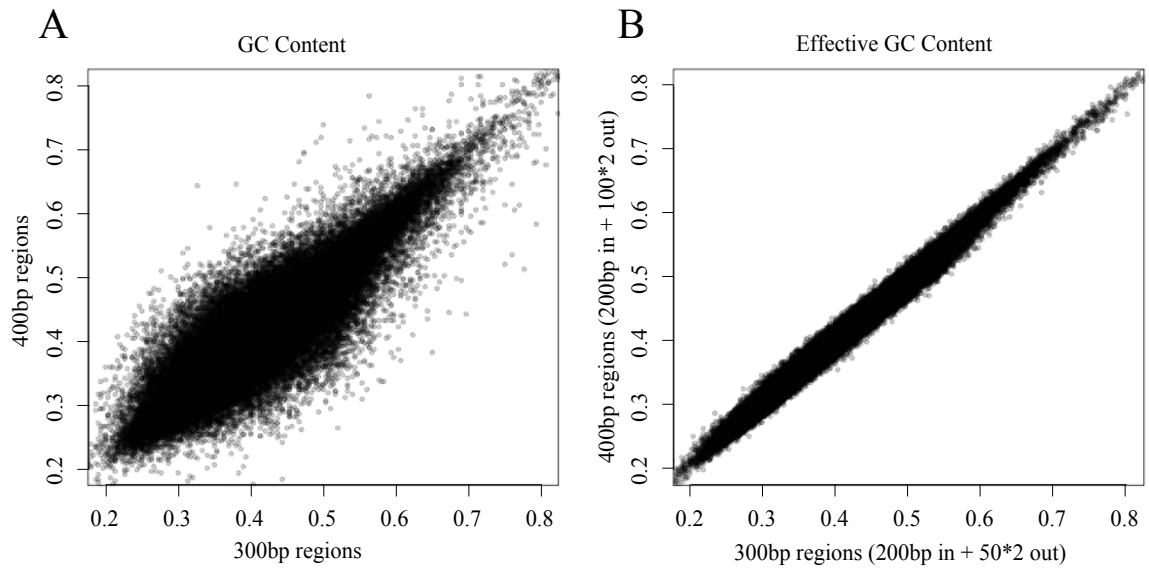

Figure S4 – Effective GC content is less sensitive to the bin size than GC-content. (A) Scatter plot between GC content of 300bp bins and corresponding 400bp bins. First, non-overlapping genome wide bins were generated using window sizes of 300bp and 400bp separately. Then, only 300bp bins located completely inside any 400bp bins were kept for GC content calculation. GC content of those corresponding 400bp bins were also calculated. (B) Same as (A) but using effective GC-content calculation. Here, l equals 200; h equals 300 and 250 for regions of sizes 400bp and 300bp, respectively.

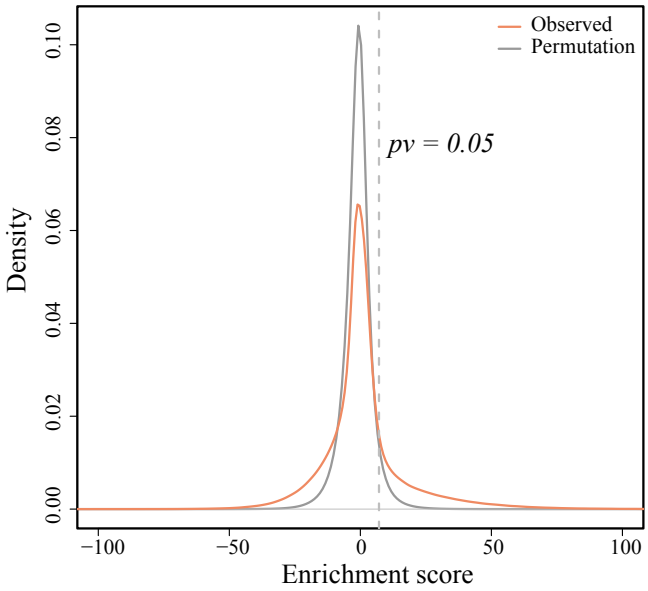

Figure S5 – Densities of observed enrichment scores and for enrichment scores obtained from the permutation procedure.

Table S1 – Datasets downloaded from ENCODE portal for analysis in this paper.

Accession ID	Laboratory	Replicate	Cell Line
ENCFF000BRG	Broad	rep1	HUVEC
ENCFF000BRD	Broad	rep2	HUVEC
ENCFF000RVE	UTA	rep1	HUVEC
ENCFF000RVI	UTA	rep2	HUVEC
ENCFF001HSL	UW	rep1	HUVEC
ENCFF001HSN	UW	rep2	HUVEC
ENCFF000ARV	Broad	rep1	GM12878
ENCFF000ARP	Broad	rep2	GM12878
ENCFF000ROU	UTA	rep1	GM12878
ENCFF000ROZ	UTA	rep2	GM12878
ENCFF000ROX	UTA	rep3	GM12878
ENCFF001HHX	UW	rep1	GM12878
ENCFF001HIA	UW	rep2	GM12878
ENCFF000BAS	Broad	rep1	HeLa-S3
ENCFF000BAT	Broad	rep2	HeLa-S3
ENCFF000RTA	UTA	rep1	HeLa-S3
ENCFF000RTC	UTA	rep2	HeLa-S3
ENCFF001HNI	UW	rep1	HeLa-S3
ENCFF001HNP	UW	rep2	HeLa-S3
ENCFF000BED	Broad	rep1	HepG2
ENCFF000BEI	Broad	rep2	HepG2
ENCFF000RUI	UTA	rep1	HepG2
ENCFF000RUJ	UTA	rep2	HepG2
ENCFF001HNT	UW	rep1	HepG2
ENCFF001HNU	UW	rep2	HepG2
ENCFF000BWM	Broad	rep1	K562
ENCFF000BWR	Broad	rep2	K562
ENCFF000RWH	UTA	rep1	K562
ENCFF000RWK	UTA	rep2	K562
ENCFF000RWO	UTA	rep3	K562
ENCFF001HTO	UW	rep1	K562
ENCFF001HTP	UW	rep2	K562
ENCFF000CMM	Broad	rep1	NHEK
ENCFF000CMF	Broad	rep2	NHEK
ENCFF000SCQ	UTA	rep1	NHEK
ENCFF000SCV	UTA	rep2	NHEK
ENCFF001HVN	UW	rep1	NHEK
ENCFF001HVQ	UW	rep2	NHEK