HINT-BC – HMM-based Identification of Transcription Factor Footprints on Bias-Corrected DNase-seq Data

Eduardo G. Gusmão*, Martin Zenke and Ivan G. Costa
Institute for Biomedical Engineering, RWTH Aachen University Medical School, Aachen, Germany
*eduardo.gusmao@rwth-aachen.de

Introduction

DNase I cleavage followed by massive sequencing (DNase-seq) has proven to be a powerful genome-wide technique for identifying active transcription factor (TF) binding sites [1–4]. Several computational approaches have been proposed to find nucleotide-resolution footprints (5-20 bp regions within two DNase-seq peaks) [3–11]. Recently, He et al. (2014) demonstrated that DNase-seq signals have biases towards the preference of DNase I to cleave particular sequences. Moreover, they show that the performance of a digital footprint method (footprint score – FS) [3] correlates with the cleavage bias of the underlying TF motif and that footprints are outperformed by simple DNase-seq tag count scoring (TC). Here, we propose the integration of a bias-correction strategy into our previous method HINT [4], which will be termed HINT-BC [12]. We investigate whether the bias-correction strategy has a significant impact on TF binding site prediction performance and perform a comprehensive evaluation including 13 footprinting methods.

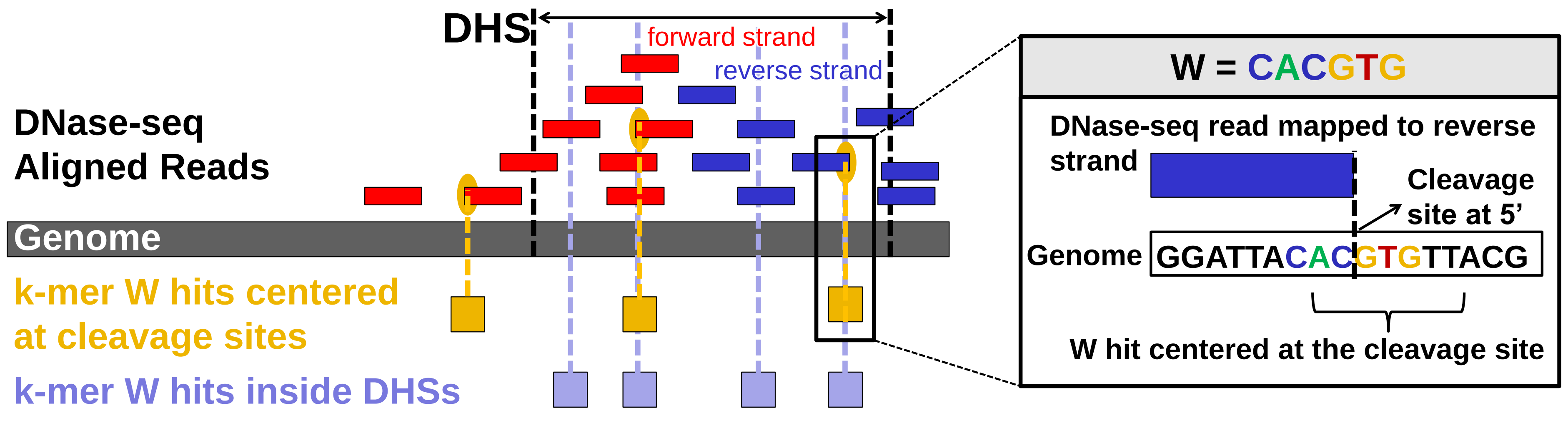
Data

Single-hit protocol (DU)	# Reads
H1-hESC	110303078
HeLa-S3	54267867
HepG2	50838536
Huvec	31848532
K562	365820647
LNCaP	163625945
MCF-7	89113893
K562-DP*	202001412
MCF-7-DP*	210715393

Double-hit protocol (UW)	# Reads
H7-hESC	302050785
HepG2	168883956
Huvec	429088276
IMR90-DP*	138604440
K562	179970820
m3134	127594903

* Deproteinized DNA

Estimation of DNase I Cleavage Bias



Given:

- $G^s[i..j]$ DNA sequence from i to j for strand s ∈ {+,-}
- $\mathbf{x} = \langle x_1, ..., x_N \rangle$ DNase-seq signal vector
- $H = \{h_1, ..., h_L\}$ Set of DNase hypersensitivity regions
- $\mathbf{1}(\cdot)$ Indicator function

We are able to evaluate:

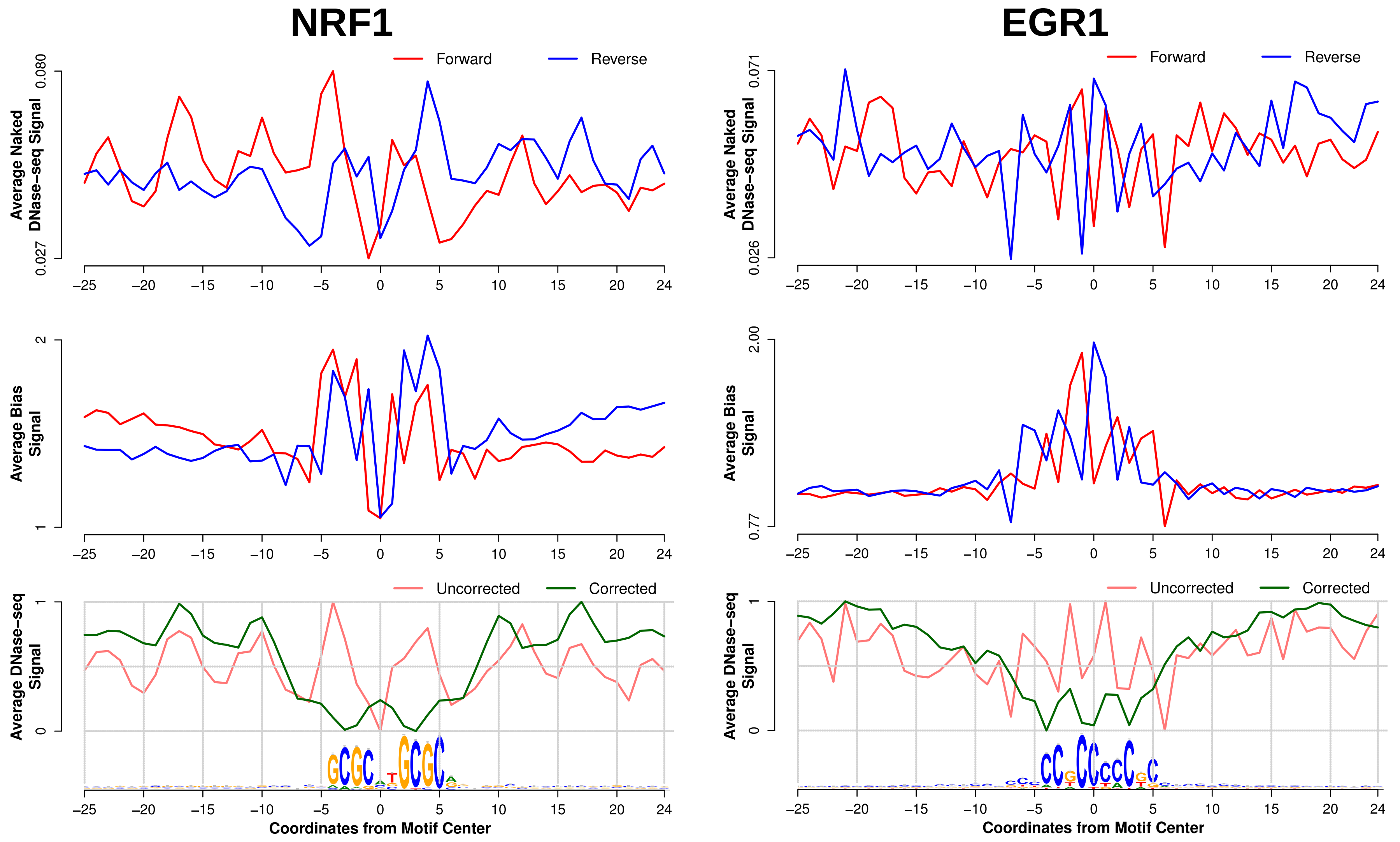
- Observed cleavage frequency for k-mer w $$o_w^s = 1 + \sum_{i=1}^{L} \sum_{j \in h_i} x_j^s \mathbf{1}\left(G^s[j - \frac{k}{2}..j + \frac{k}{2}] = w \right)$$
- Background cleavage frequency for k-mer w $$r_w^s = 1 + \sum_{i=1}^{L} \sum_{j \in l} \mathbf{1}\left(G^s[j - \frac{k}{2}..j + \frac{k}{2}] = w \right)$$
- Cleavage bias signal $b_i^s = o_w^s \cdot R / r_w^s \cdot O^s$ where: $O^s = \sum_{i=1}^{L} \sum_{j \in h_i} x_j^s$ and $R = \sum_{i=1}^{L} \sum_{j \in h_i}$

DNase I Cleavage Bias Correction

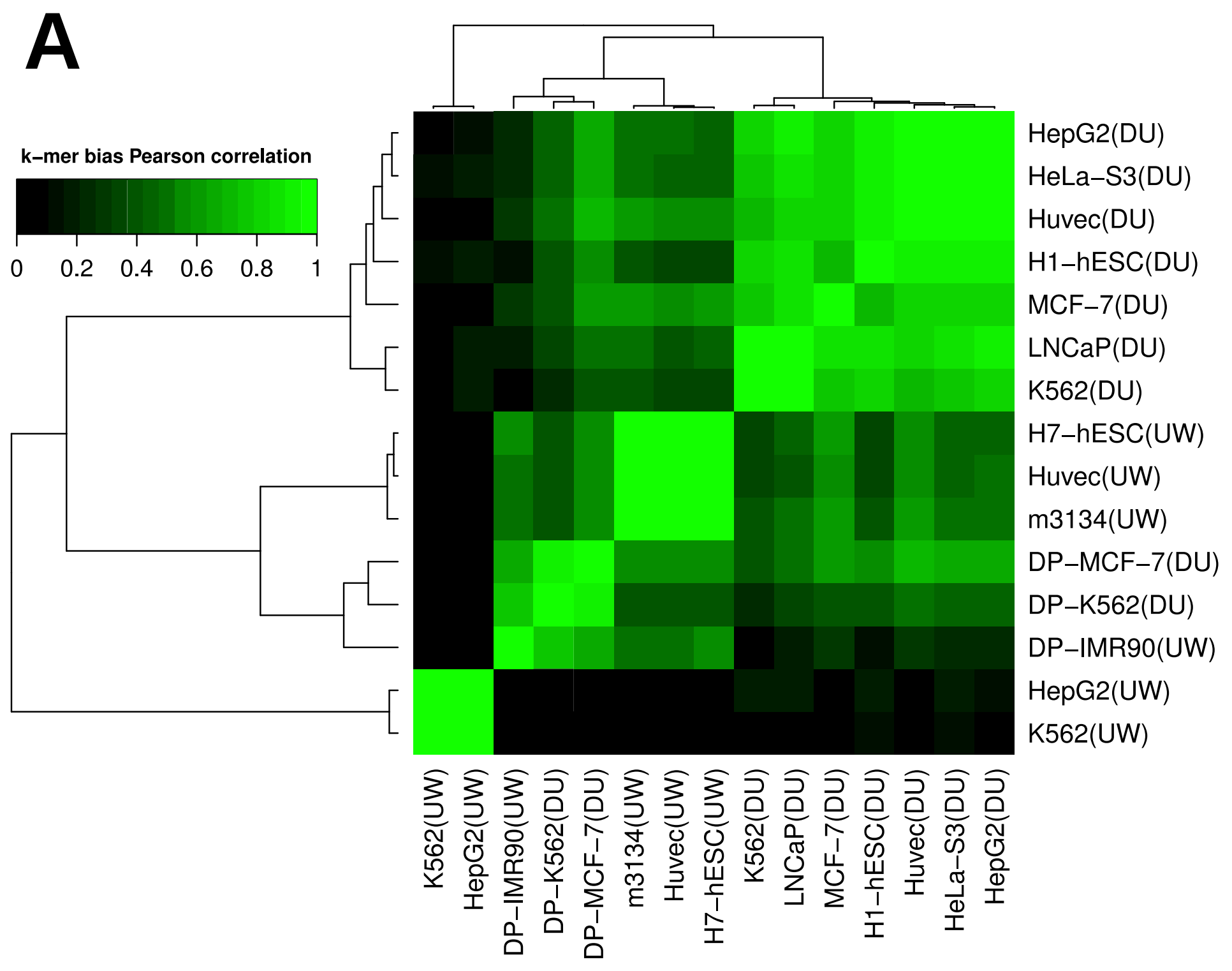
Corrected DNase-seq signal (y_i^s) is evaluated based on a correction signal (c_i^s), which is calculated based on smoothed versions of the DNase-seq signal ($\hat{x}_i^s$) and the bias signal ($\hat{b}_i^s$).

$$\hat{x}_i^s = \sum_{j=i-25}^{i+24} x_j^s \qquad \hat{b}_i^s = \frac{b_i^s}{\sum_{j=i-25}^{i+24} b_j^s} \qquad c_i^s = \hat{x}_i^s \hat{b}_i^s$$

$$y_i^s = \log(x_i^s + 1) - \log(c_i^s + 1)$$

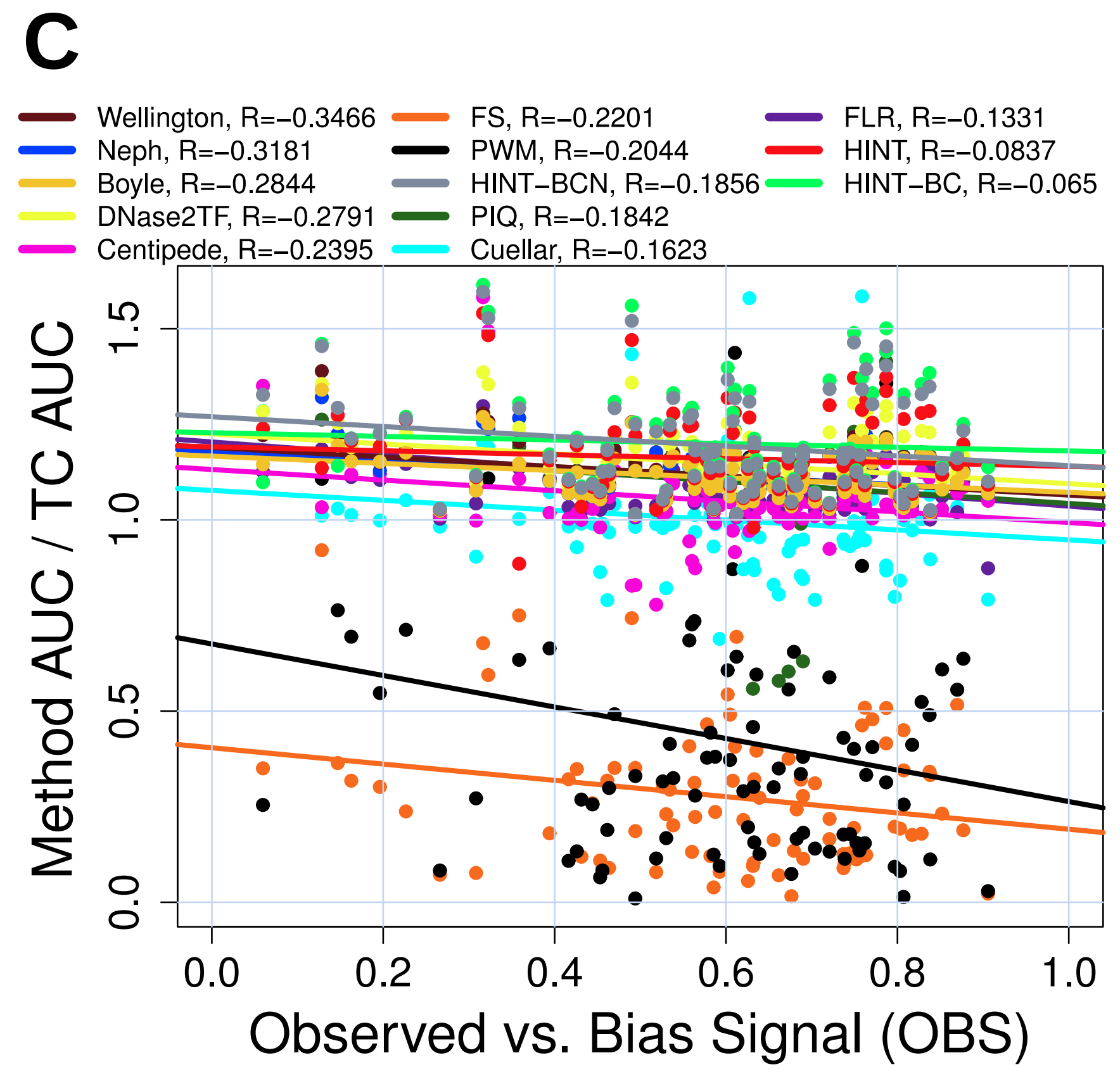


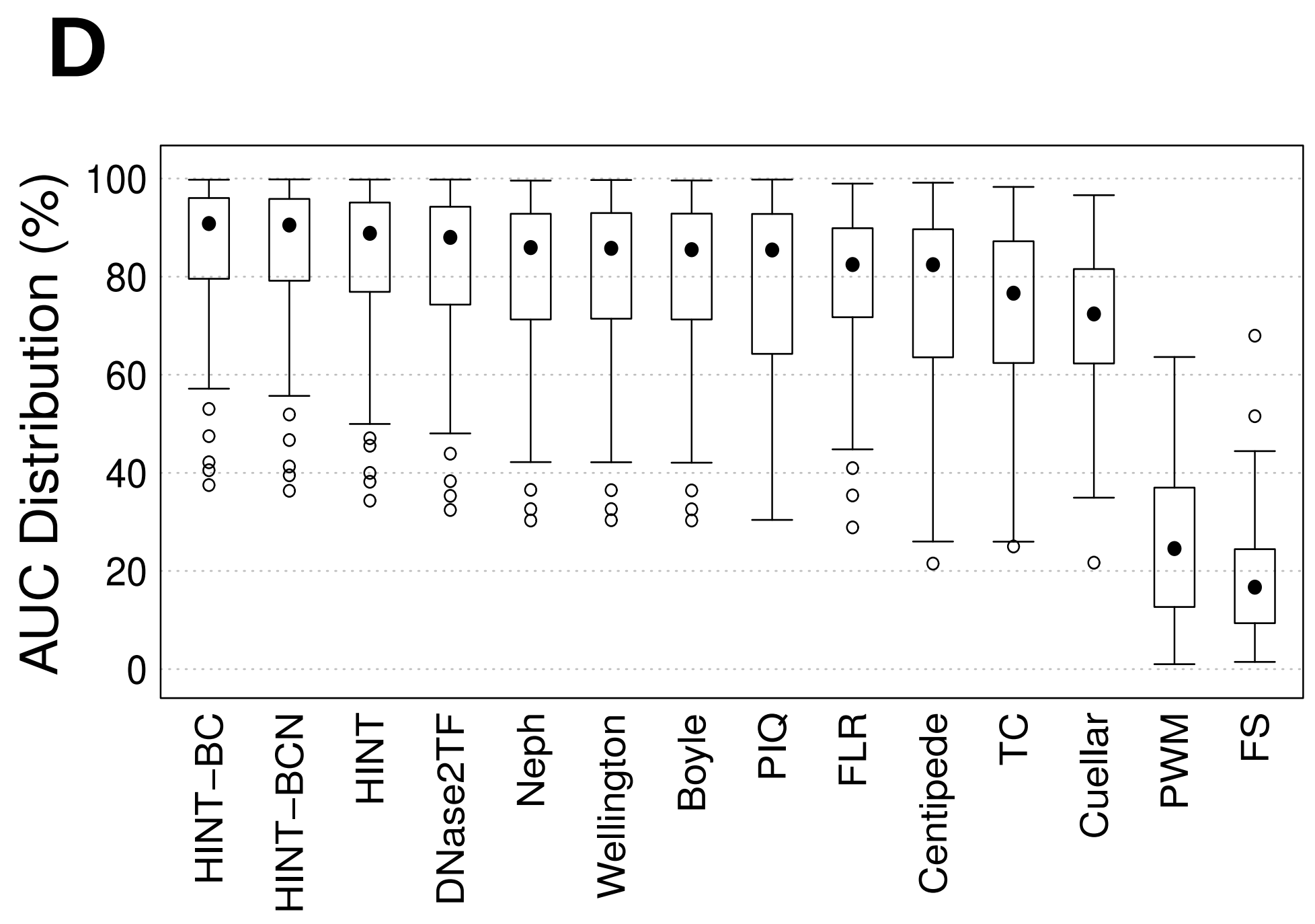
Results



B

		HINT-BC	HINT-BCN	HINT	DNase2TF	PIQ	Wellington	Neph	Boyle	FLR	Centipede	Cuellar	TC	PWM	FS
1.3409	HINT-BC														
2.0795	HINT-BCN														
3.6023	HINT	■													
4.0682	DNase2TF	■	■												
5.6932	PIQ	■	■	■											
6.3068	Wellington	■	■	■	■										
6.9091	Neph	■	■	■	■										
7.6477	Boyle	■	■	■	■										
9.2614	FLR	■	■	■	■	■	■	■							
9.8409	Centipede	■	■	■	■	■	■	■	■						
10.642	Cuellar	■	■	■	■	■	■	■	■						
11.1648	TC	■	■	■	■	■	■	■	■						
12.7386	PWM	■	■	■	■	■	■	■	■	■	■	■			
13.7045	FS	■	■	■	■	■	■	■	■	■	■	■	■		





(A) Correlation of bias scores between different DNase-seq datasets given all possible DNA 6-mers. Deproteinized experiments are marked with "DP". **(B)** Friedman-Nemenyi hypothesis test. The rows are sorted by the Friedman ranking. A shadowed cell means that the method in the column outperformed the method in the row (95% confidence). **(C)** Correlation between the performance of each method (in relation to the DNase-seq tag count; TC) and the OBS (correlation between observed and bias signal). **(D)** Distribution of the area under the ROC curve (AUC) at 10% specificity for all footprinting methods using a validation set with 88 ChIP-seq experiments.

Bibliography

1. ENCODE Project. Nature. 489(7414):57-74 (2012).
2. He HH et al. Nat Meth. 11(1):73-78 (2014).
3. Neph S et al. Nature. 489(7414):83-90 (2012).
4. Gusmao EG et al. Bioinf. 30(22):3143-51 (2014).
5. Boyle AP et al. Gen. Res. 21(3):456-64 (2011).
6. Cuellar-Partida G et al. Bioinf. 28(1):56-62 (2012).
7. Pique-Regi R et al. Gen. Res. 21(3):447-55 (2011).
8. Sung MH et al. Mol Cell. 56(2):275-85 (2014).
9. Yardimci GG et al. NAR. 42(19):11865-78 (2014).
10. Sherwood RI et al. Nat. Biotech. 32(3):171-8 (2014).
11. Piper J et al. NAR. 41(21):e201 (2013).
12. Gusmao EG et al. Nat. Meth. (in revision).