Are Computationally Predicted Footprints a Result of DNase I Cleavage Bias?

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) [3–7]. 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 occupancy score – FOS) [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 test these claims using more sophisticated digital genomic footprinting methods. Furthermore, we verify whether it is possible to improve computational methods by correcting DNase I cleavage bias.

DNase-seq Data

Crawford Lab (DU) [1]

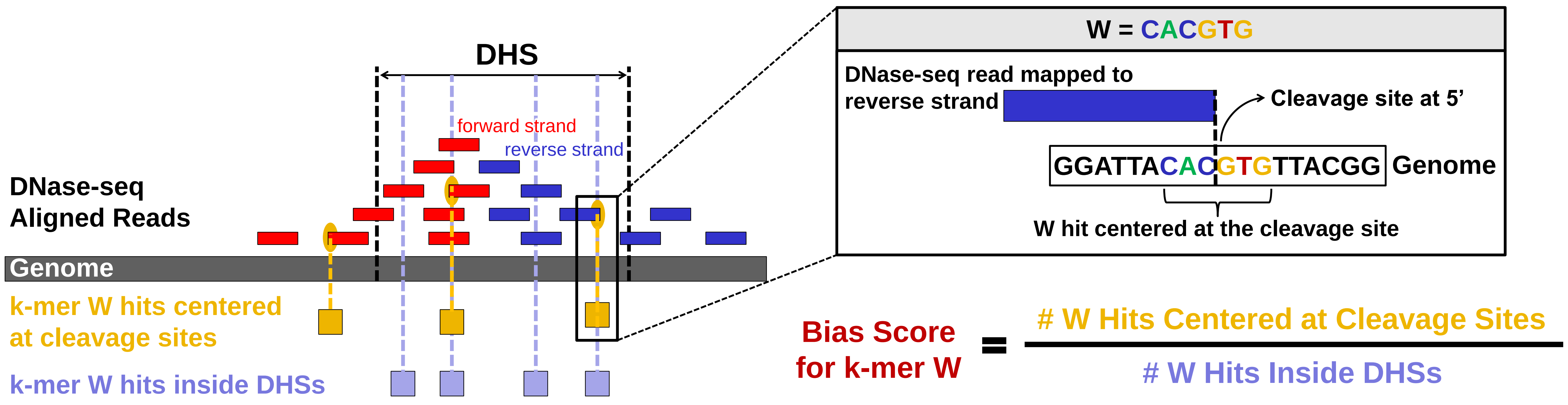
Cell Type	# Mapped Reads
H1-hESC	110303078
HeLa-S3	54267867
HepG2	50838536
Huvec	31848532
K562	365820647

Stamatoyannopoulous Lab (UW) [1]

Cell Type	# Mapped Reads
HepG2	168883956
Huvec	429088276
K562	179970820

Estimation of DNase I Cleavage Bias

Estimation of intrinsic DNase I cleavage bias was performed by calculating, for each k-mer W, the ratio of the number of observed cleavage sites centered at W and the number of times it occurred within DNase-seq hypersensitivity sites (DHSs) [2].



$$\text{Bias Score for k-mer W} = \frac{\text{\# W Hits Centered at Cleavage Sites}}{\text{\# W Hits Inside DHSs}}$$

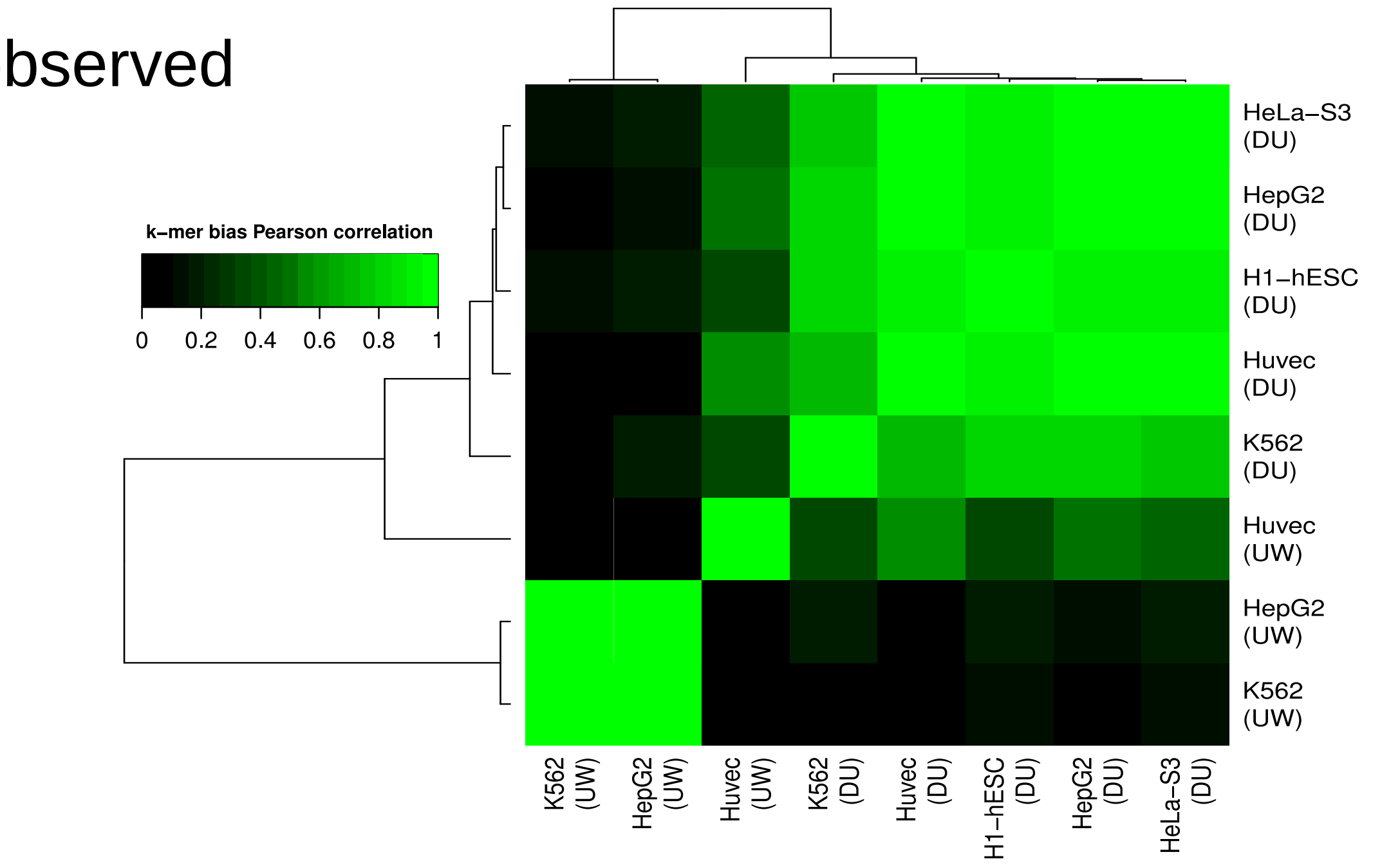


Fig. 1: Correlation of bias scores between different DNase-seq datasets given all possible DNA 6-mers.

DNase I Cleavage Bias Correction

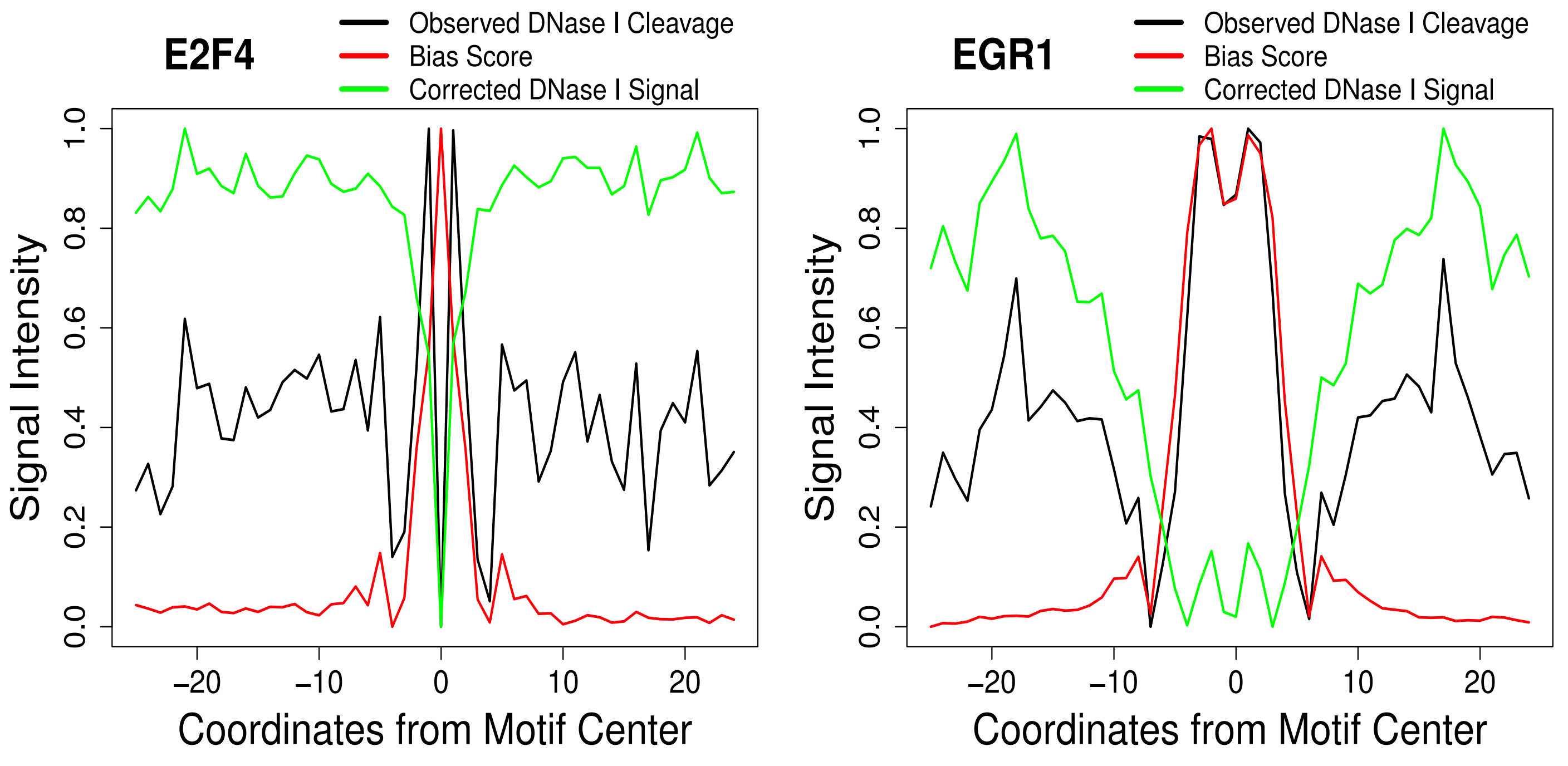


Fig. 2: Observed cleavage, bias score and corrected signal for TFs E2F4 and EGR1. Signals were standardized to be in [0,1].

DNase I Cleavage bias correction was based on smoothed versions of both DNase I and bias score signals [2].

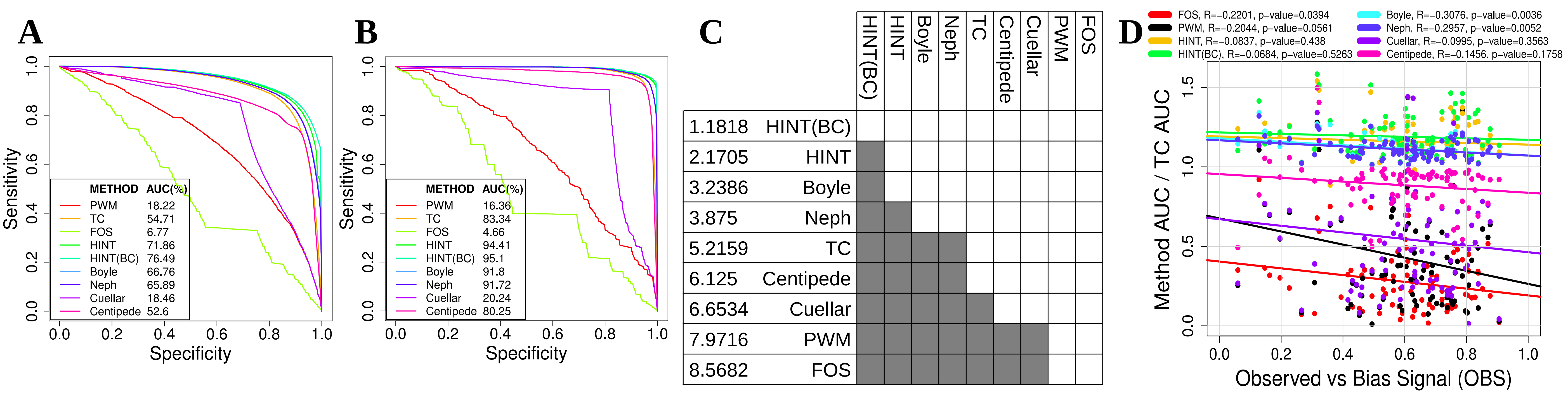
Smoothed DNase I Cleavage Signal = **Summation of the observed DNase I cleavage signal around a 50bp window.**

Smoothed Bias Score = **Summation of bias score signal around a 50bp window.**

$$\text{Corrected Smoothed Signal} = \frac{\text{Smoothed DNase I Cleavage Signal}}{\text{Smoothed Bias Score}}$$

$$\text{Corrected DNase I Signal} = \log\left(\frac{\text{Observed DNase I}}{\text{Cleavage Signal}}+1\right) - \log\left(\frac{\text{Corrected}}{\text{Smoothed Signal}}+1\right)$$

Results



		HINT(BC)	HINT	Boyle	Neph	TC	Centipede	Cuellar	PWM	FOS
1.1818	HINT(BC)									
2.1705	HINT	▨								
3.2386	Boyle	▨								
3.875	Neph	▨	▨							
5.2159	TC	▨	▨	▨	▨					
6.125	Centipede	▨	▨	▨	▨					
6.6534	Cuellar	▨	▨	▨	▨	▨				
7.9716	PWM	▨	▨	▨	▨	▨	▨	▨		
8.5682	FOS	▨	▨	▨	▨	▨	▨	▨		

Fig. 3: (A,B) Performance of methods as ROC curves for TFs E2F4 and EGR1. In the legend it is shown the AUC at 10% FPR. (C) Friedman-Nemenyi hypothesis test. Each row starts with the Friedman ranking for each method. A shadowed cell means that the method in the column outperformed the method in the row (95% confidence level). (D) Correlation between the performance of each method (in relation to the DNase I TC) and the OBS (correlation between observed and bias signal).

We applied the digital footprinting method HINT [4] to the DNase-seq signal and the bias corrected (BC) signal. We observed that bias corrected version of HINT – HINT (BC) – outperformed all other methods: the site-centric tag count (TC), footprint occupancy score (FOS), position weight matrix (PWM) bitscore, Boyle [5], Neph [3], Cuellar [6] and Centipede [7] (Fig.3C). Interestingly, the Friedman-Nemenyi test also showed that the bias corrected (BC) version of HINT significantly outperforms the original version. We also evaluated the correlation between the performance of each method (represented by their AUC relative to TC's AUC) and the observed vs bias signal (OBS) (Fig.3D) [2]. The latter corresponds to the correlation between the observed DNase I cleavage and bias score (Fig.2). Significant negative correlations were observed for FOS, Boyle and Neph. Again, since HINT (BC) portrayed a smaller correlation than HINT, the bias correction demonstrated to mitigate prediction biases.

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. Bioinformatics. btu519+ (2014).
5. Boyle AP et al. Genome Res. 21(3), 456-464 (2011).
6. Cuellar-Partida G et al. Bioinformatics. 28(1), 56-62 (2012).
7. Pique-Regi R et al. Genome Res. 21(3), 447-455 (2011).