

Tissue-specific prediction of directly regulated genes

Robert C. McLeay, Chris J. Leat and Timothy L. Bailey

June 5, 2011

Abstract

Direct binding by a transcription factor (TF) to the proximal promoter of a gene is strong evidence that the TF regulates the gene. Assaying the genome-wide binding of every TF in every cell type and condition is currently impractical. Histone modifications correlate with tissue/cell/condition-specific (“tissue-specific”) TF binding, so histone ChIP-seq data can be combined with traditional position-weight matrix (PWM) methods to make tissue-specific predictions of TF-promoter interactions.

Results:

We use supervised learning to train a naïve Bayes predictor of TF-promoter binding. The predictor’s features are the histone modification levels and a PWM-based score for the promoter. Training and testing uses sets of promoters labeled using TF ChIP-seq data, and we use cross-validation on 23 such datasets to measure accuracy. A PWM+histone naïve Bayes predictor using a single histone modification (H3K4me3) is substantially more accurate than a PWM score or a conservation-based score (phylogenetic motif model). The naïve Bayes predictor is more accurate (on average) at all sensitivity levels, and makes only half as many false positive predictions at sensitivity levels from 10% to 80%. On average, it correctly predicts 80% of bound promoters at a false positive rate of 20%. Accuracy does not diminish when we test the predictor in a different cell type (and species) from training. Accuracy is barely diminished even when we train the predictor *without* using TF ChIP-seq data.

Availability:

Our tissue-specific predictor of promoters bound by a TF is called DR GENE and is available at <http://bioinformatics.org.au/drgene>.

Contact:

t.bailey@imb.uq.edu.au

1 Supplementary Methods

1.1 Training the naïve Bayes classifier

We train our naïve Bayes classifier using a training set of labeled examples of the form (X, B) where B is the class (*bound* or *unbound*) and X is the vector (see main methods for full definition):

$$X = \langle M_f(p), H_{1,t}(p), \dots, H_{n,t}(p) \rangle,$$

where $M_f(p)$ is the PWM score (see Eqn. 1 in main paper), and $H_{i,t}(p)$ is the histone score (see Eqn. 2, main paper). We define each set of labeled examples in the results in the main paper.

We use the R package e1071 [2] to implement the naïve Bayes classifier. The classifier is trained using the *naiveBayes* function. To predict bound genes, we use the *predict* function (also available in the e1071 package) to calculate the probability that each test feature vector (X) is bound (see Eqn. 3, main paper).

1.2 Phylogenetic Motif Model Scores

We use two sequence conservation-based score functions based on the Monkey [8] algorithm, which we call “Monkey” and “Monkey+” scores, respectively. We selected Monkey as it was found to be most accurate in a recent comparison of phylogenetic motif model scanners [4]. To score a promoter, we run Monkey twice. First we run Monkey on a multiple alignment of mouse (mm8) and human (hg18) promoters twice, once with the settings, `monkey [motif] [multiple alignment] -tree [treefile] -m HB -freq [background] -cut 0`, using the same background described previously for PWM scanning, and once with the `-cut -1e6` parameter. The former `-cut` parameter produces species-specific predictions, only scoring bases with a strong match in the species of interest, but will not score all bases or promoters. We use the best score from species-specific Monkey results, if available, otherwise we use the best non-species-specific score from Monkey using the `-cut -1e6` parameter. We use the *multiz30way* alignment and tree from the UCSC Genome Browser [6], only using mouse (mm9) and human (hg18) from this alignment, as we find that this performs better than using a larger tree.

1.3 Monkey+ scores

The above approach, however, still assigns no score to many promoters and bases due to multiple alignment gaps. We therefore integrate Monkey’s species-specific scores with PWM scanning. We term this “Monkey+” (Monkey+PWM). We run Monkey with the `-cut 0` parameter (to produce a species-specific score). We then use the best score from both the Monkey and a PWM scan results for each promoter.

2 Supplementary Results

2.1 Sensitivity at 1% False positive rate

Scoring Method	Tissue Type			Mean Sensitivity For All Tissues
	GM12878 (5 TFs)	K562 (10 TFs)	mES (8 TFs)	
Naïve Bayes PWM+H3k4me2	0.19 (0.03)	0.28 (0.06)	0.13 (0.03)	0.20 (0.05)
Naïve Bayes PWM+H3k4me2, H3k4me3	0.22 (0.05)	0.27 (0.05)	0.12 (0.02)	0.20 (0.04)
Naïve Bayes PWM+H3k4me3	0.21 (0.04)	0.28 (0.06)	0.13 (0.02)	0.21 (0.04)
Naïve Bayes PWM+H3k9ac	0.22 (0.05)	0.28 (0.06)	-	0.25 (0.03)
Naïve Bayes PWM+H3k27ac	0.22 (0.05)	0.28 (0.06)	-	0.25 (0.03)
Naïve Bayes PWM+H3k9ac, H3k27ac	0.22 (0.06)	0.26 (0.05)	-	0.24 (0.02)
Naïve Bayes PWM+H3k4me2, H3k4me3, H3k9ac, H3k27ac	0.22 (0.06)	0.25 (0.05)	-	0.23 (0.02)
Naïve Bayes Monkey++H3k4me3	-	-	0.15 (0.03)	0.15
PWM	0.09 (0.04)	0.17 (0.05)	0.14 (0.04)	0.13 (0.02)
H3K4me3	0.01 (0.00)	0.03 (0.00)	0.05 (0.01)	0.03 (0.01)
Monkey+	-	-	0.13 (0.04)	0.13
BBS	-	-	0.11 (0.04)	0.11

Table 1: **Sensitivity at FPR=0.01 of TF-promoter binding predictions.** The table shows the sensitivity at 1% false positive rate of different methods for scoring promoters. Results are shown for predictions in three different tissues. The number of TF ChIP-seq binding datasets used in each tissue is indicated. Results for naïve Bayes scores are the average sensitivity in a “hold-one-TF-out” experiment. Results for the other scoring methods are the mean of the sensitivity values for each of the TFs used in the given tissue. All results are rounded to two decimal places. Standard errors are given in parentheses. Highest accuracies for a given tissue and overall are in bold font. Missing data for mES is due to lack of availability of histone acetylation data. Monkey+ and BBS were only tested in mES cells.

2.2 Detailed accuracy results for predicting bound promoters

TF	Tissue	SAME _{TISSUE}	DIFF _{TISSUE}	SAME _{TF}	SLNB	PWM	H3K4me3	Monkey+	BBLs
Gata1	K562	0.882	0.876		0.875	0.686	0.814		
Gata2	K562	0.888	0.882		0.878	0.699	0.813		
Yy1	K562	0.889	0.887		0.828	0.646	0.862		
c-Myc	K562	0.875	0.875	0.882	0.860	0.702	0.874		
	mES	0.875	0.858	0.858	0.888	0.810	0.903	0.801	0.802
CTCF	GM12878	0.830	0.827	0.847	0.817	0.767	0.737		
	K562	0.830	0.819	0.865	0.799	0.808	0.713		
	mES	0.831	0.856	0.896	0.763	0.896	0.721	0.873	0.880
Egr1	GM12878	0.825	0.835	0.824	0.822	0.718	0.798		
	K562	0.918	0.919	0.877	0.898	0.893	0.769		
Esrrb	mES	0.856	0.858		0.780	0.719	0.801	0.685	0.708
Jund	GM12878	0.887	0.866	0.861	0.755	0.771	0.694		
	K562	0.950	0.947	0.930	0.954	0.889	0.843		
Klf4	mES	0.901	0.892		0.892	0.835	0.871	0.817	0.843
n-Myc	mES	0.888	0.876		0.896	0.832	0.899	0.825	0.810
Nfyb	K562	0.930	0.936		0.916	0.903	0.786		
Srf	GM12878	0.787	0.810	0.838	0.794	0.522	0.800		
	K562	0.855	0.841	0.844	0.865	0.597	0.830		
STAT3	mES	0.896	0.888		0.898	0.765	0.832	0.721	0.700
Tcfcp2l1	mES	0.881	0.879		0.834	0.823	0.819	0.782	0.766
Usf1	GM12878	0.958	0.949	0.955	0.921	0.909	0.780		
	K562	0.929	0.930	0.930	0.921	0.879	0.790		
Zfx	mES	0.880	0.870		0.884	0.839	0.883	0.833	0.835
Mean:		0.880	0.877	0.877	0.858	0.779	0.810	0.792	0.793
#Wins <i>cf.</i> PWM		22/23 [‡]	22/23 [‡]	12.5/13[†]	20/23 [‡]	-	0/8	0/8	1/8
#Wins <i>cf.</i> all non-NB		18/23 [†]	19/23 [†]	11.5/13[†]	14/23	10/23	13/23	0/8	0/8
#Wins <i>cf.</i> SAME _{TISSUE}		-	7.5/23	6/13	7/23	0.5/23	4/23	0/8	0/8
#Wins <i>cf.</i> all NB		7/23	5/23	6/13	6/23	0.5/23	2/23	0/8	0/8

Table 2: **Classifier Performance (AUC) by Method, TF, and Tissue.** The AUC for each method on each TF-tissue combination. Bold AUCs are the best achieved for each TF-tissue combination. Blank cells in the “NB: SAME_{TF}” column are due to training data being unavailable; for conservation-based methods (Monkey and Monkey+), only mES TFs were used. When counting wins, a draw between methods is counted as one-half each. “#Wins *cf.* all...” do not include the current column when comparing to other methods. Methods annotated with ‡ have p -value < 0.001 for the binomial test with the hypothesis that the method has higher performance. For †, p -value is < 0.05 . Note that Jund in GM12878 cells requires the use of a different PWM to that used in K562 cells. We discuss this in Sec. 2.3.

2.3 JunD displays divergent binding preferences across tissues

Unlike the other TFs included in our reference sets, the motif used for JunD is that of the AP1 transcription initiation complex, of which JunD is a member. Fig. 1(a) & 1(b) show that although the AP1/JunD PWM performs well in GM12878 cells, it is not correlated with TF binding in K562 cells. Given the surprising difference in PWM performance for JunD between these two tissues, we first investigate the JunD binding sites contained within the defined promoters in each tissue.

We apply MEME-ChIP [7] to GM12878 peaks. Of 96 peak regions, 46 contain the **tgaCtga** AP1 motif previously used for PWM scanning (see supp. Tab. 2). We find a second motif in 14 of 96 sequences with a consensus sequence of **tgaCGtca**. The PWM used for scanning, however, will not score perfect matches of this latter motif highly. Both the **tgaCtca** (TRE) and **tgaCGtga** (CRE) binding site motifs have been reported as binding sites for JunD’s bZip domain [3, 5]. We then apply MEME-ChIP [7] to the K562 promoter peak set. Unlike the GM12878 peaks, we find only one AP1 motif, (CRE – **tgaCGtga**) in 318 of 464 sequences. It cannot be readily detected by the PWM used (see supp. Tab. 2), explaining the random performance (AUC= 0.45) of the PWM method in Figure 1(b).

We rescan promoters using the CRE PWM, and use this as test data to our hold-one-TF out predictor, . Figure 1(c) shows the resulting ROC curve. The naïve Bayes predictor’s accuracy improves from an AUC of 0.82 to 0.95, while the PWM increases from an AUC of 0.45 to 0.89. Clearly, the key problem in predicting JunD-bound promoters is not our naïve Bayes predictor, but rather the correct choice of PWM.

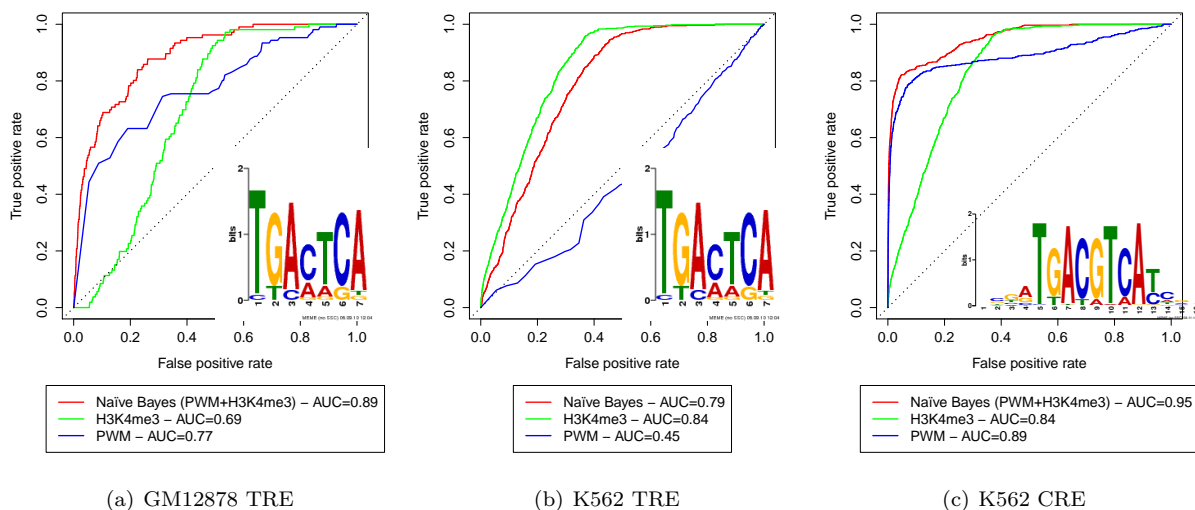


Figure 1: **ROC of predicting JunD-bound promoters using the AP1 TRE or CRE PWM.** Each panel shows ROC curves for predicting JunD binding using the naïve Bayes predictor (trained on other TFs from the same tissue), sorting by H3K4me3 tag count, and PWM scanning. In panels (a) and (b), we use the TRE PWM for the naïve Bayes predictor and PWM scanning in GM12878 and K562 cells respectively. Panel (c) shows the effectiveness of using the CRE PWM in K562 cells.

2.4 Classifier Parameters for each TF-tissue combination

<i>TF</i>	<i>Tissue</i>	$P(\text{Bound} \text{TF})$	<i>H3k4me3</i>				<i>PWM</i>			
			<i>Bound</i>		<i>Unbound</i>		<i>Bound</i>		<i>Unbound</i>	
			$\hat{\mu}$	$\hat{\sigma}$	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\mu}$	$\hat{\sigma}$
CTCF	mES	0.034	0.694	0.597	-0.024	1.003	1.750	1.035	-0.061	0.942
Esrrb	mES	0.027	0.912	0.385	-0.025	1.000	0.963	1.322	-0.026	0.977
Klf4	mES	0.067	1.040	0.280	-0.074	0.991	1.078	0.636	-0.077	0.977
c-Myc	mES	0.034	1.128	0.273	-0.040	0.993	1.026	0.637	-0.036	0.991
n-Myc	mES	0.074	1.093	0.236	-0.087	0.986	1.053	0.577	-0.084	0.979
STAT3	mES	0.006	1.011	0.248	-0.006	1.000	1.174	1.235	-0.007	0.994
Tcfcp2l1	mES	0.048	0.935	0.421	-0.047	0.998	1.099	0.759	-0.055	0.979
Zfx	mES	0.061	1.070	0.275	-0.070	0.990	1.065	0.521	-0.070	0.984
CTCF	K564	0.090	0.673	0.539	-0.066	1.010	1.037	1.201	-0.102	0.916
Egr1	K564	0.009	0.882	0.202	-0.008	1.001	0.763	0.822	-0.007	0.999
Gata1	K564	0.012	0.969	0.373	-0.012	0.999	0.680	1.047	-0.008	0.997
Gata2	K564	0.016	0.956	0.376	-0.015	0.999	0.705	1.009	-0.011	0.996
Jund	K564	0.002	0.653	0.354	-0.002	1.001	1.067	0.969	-0.003	0.999
c-Myc	K564	0.217	0.913	0.366	-0.253	0.972	0.544	0.665	-0.151	1.025
Nfyb	K562	0.094	0.816	0.454	-0.085	1.003	1.492	0.827	-0.155	0.882
Srf	K564	0.006	0.886	0.218	-0.006	1.001	0.276	1.471	-0.002	0.996
Usf1	K564	0.036	0.830	0.265	-0.031	1.004	1.494	0.609	-0.056	0.968
Yy1	K564	0.075	1.028	0.287	-0.083	0.991	0.337	0.654	-0.027	1.018
CTCF	GM12878	0.074	0.633	0.611	-0.051	1.008	1.232	1.177	-0.099	0.915
Egr1	GM12878	0.035	0.832	0.327	-0.030	1.003	1.381	0.621	-0.049	0.975
Jund	GM12878	0.014	1.045	0.348	-0.015	0.998	2.305	1.370	-0.033	0.954
Srf	GM12878	0.005	1.004	0.297	-0.005	1.000	0.652	1.590	-0.003	0.995
Usf1	GM12878	0.050	0.863	0.408	-0.045	1.001	1.353	0.750	-0.071	0.960

Table 3: **Naïve Bayes scoring parameters for each TF-tissue combination.** We calculate the sample mean ($\hat{\mu}$) and standard deviation ($\hat{\sigma}$) for the PWM and H3K4me3 features for both bound and unbound promoter for each TF-tissue combination. We also calculate the prior probability that a promoter is bound by a given TF in each tissue.

2.5 Maximum PWM Score and Normality

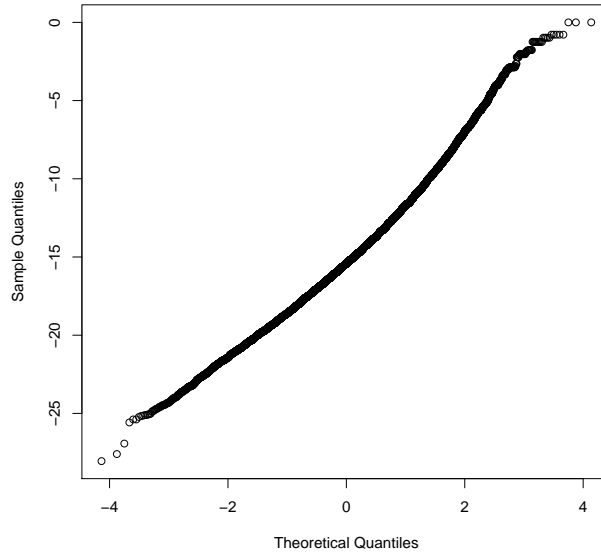


Figure 2: **Q-Q Normal Plot of Esrrb PWM Scores.** We demonstrate the normality of the distribution of values of log-transformed PWM scores for Esrrb. The data plotted is $\log(M_f(p))$ from Eqn. 1, for all mES promoters.

3 Supplementary Data

Table 4: **Transcription factor-tissue combinations used as our reference sets for evaluating the predictor.** The TF ChIP-seq data listed is used to define whether a given promoter is bound or unbound for each TF. An “X” present in a column means that TF ChIP-seq data is available for that TF in the tissue described in the column heading. If another member of the family was used to scan in place of the TF (e.g. GATA1 was used for GATA2), this is noted in parentheses in the PWM Type column.

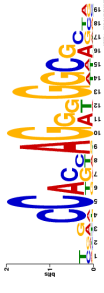
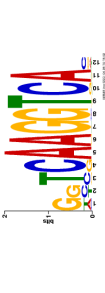
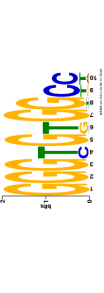
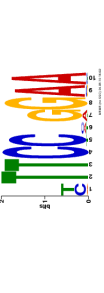

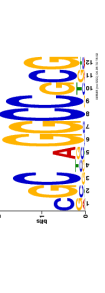
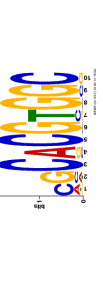

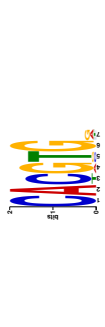


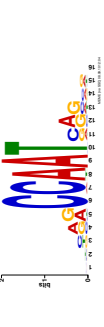
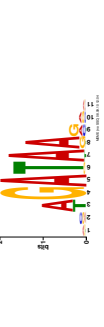
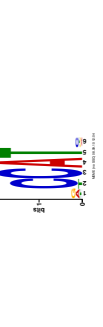
TF	mES	GM12878	K562	ChIP-seq Source	PWM	PWM Type	PWM Source
CTCF	X	X	X	Chen <i>et al.</i> [1], Thomas <i>et al.</i> [10]		ChIP-seq	Chen <i>et al.</i> [1]
Esrrb	X			Chen <i>et al.</i> [1]		ChIP-seq	Chen <i>et al.</i> [1]
Klf4	X			Chen <i>et al.</i> [1]		ChIP-seq	Chen <i>et al.</i> [1]
STAT3	X			Chen <i>et al.</i> [1]		ChIP-seq	Chen <i>et al.</i> [1]
Tcfcp2l1	X			Chen <i>et al.</i> [1]		ChIP-seq	Chen <i>et al.</i> [1]
Zfx	X			Chen <i>et al.</i> [1]		ChIP-seq	Chen <i>et al.</i> [1]
cMyc	X		X	Chen <i>et al.</i> [1], Thomas <i>et al.</i> [10]		ChIP-seq	Chen <i>et al.</i> [1]
nMyc	X			Chen <i>et al.</i> [1]		ChIP-seq	Chen <i>et al.</i> [1]

Table 4: Transcription factor-tissue combinations used as our reference sets (continued).

TF	mES	GM12878	K562	ChIP-seq Source	PWM	PWM Type	PWM Source
Usf1		X	X	Thomas <i>et al.</i> [10]		SELEX	Portales-Casamar <i>et al.</i> [9]
Jund		X	X	Thomas <i>et al.</i> [10]		Compiled (AP1)	Portales-Casamar <i>et al.</i> [9]
Egr1		X	X	Thomas <i>et al.</i> [10]		Bacterial 1-hybrid	Portales-Casamar <i>et al.</i> [9]
Nfyb			X	Thomas <i>et al.</i> [10]		Compiled (NYFA)	Portales-Casamar <i>et al.</i> [9]
Gata1			X	Thomas <i>et al.</i> [10]		ChIP-seq	Portales-Casamar <i>et al.</i> [9]
Gata2			X	Thomas <i>et al.</i> [10]		ChIP-seq (Gata1)	Portales-Casamar <i>et al.</i> [9]
Yy1			X	Thomas <i>et al.</i> [10]		Compiled	Portales-Casamar <i>et al.</i> [9]
Srf		X	X	Thomas <i>et al.</i> [10]		SELEX	Portales-Casamar <i>et al.</i> [9]

4 References

- [1] Chen, X., Xu, H., Yuan, P., Fang, F., Huss, M., Vega, V. B., Wong, E., Orlov, Y. L., Zhang, W., Jiang, J., Loh, Y.-H., Yeo, H. C., Yeo, Z. X., Narang, V., Govindarajan, K. R., Leong, B., Shahab, A., Ruan, Y., Bourque, G., Sung, W.-K., Clarke, N. D., Wei, C.-L., and Ng, H.-H. (2008). Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell*, **133**(6), 1106–1117.
- [2] Dimitriadou, E., Hornik, K., Leisch, F., Meyer, D., and Weingessel, A. (2005). Misc Functions of the Department of Statistics (e1071), TU Wien.
- [3] Hai, T. and Curran, T. (1991). Cross-family dimerization of transcription factors fos/jun and atf/creb alters dna binding specificity. *Proc Natl Acad Sci U S A*, **88**(9), 3720–3724.
- [4] Hawkins, J., Grant, C., Noble, W. S., and Bailey, T. L. (2009). Assessing phylogenetic motif models for predicting transcription factor binding sites. *Bioinformatics*, **25**(12), i339–i347.
- [5] John, M., Leppik, R., Busch, S. J., Granger-Schnarr, M., and Schnarr, M. (1996). Dna binding of jun and fos bzip domains: homodimers and heterodimers induce a dna conformational change in solution. *Nucleic Acids Res*, **24**(22), 4487–4494.
- [6] Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., and Haussler, D. (2002). The human genome browser at UCSC. *Genome Res*, **12**(6), 996–1006.
- [7] Machanick, P. and Bailey, T. (2011). MEME-ChIP: Tba. *Bioinformatics (to be submitted)*.
- [8] Moses, A. M., Chiang, D. Y., Pollard, D. A., Iyer, V. N., and Eisen, M. B. (2004). MONKEY: identifying conserved transcription-factor binding sites in multiple alignments using a binding site-specific evolutionary model. *Genome Biol*, **5**(12), R98.
- [9] Portales-Casamar, E., Thongjuea, S., Kwon, A. T., Arenillas, D., Zhao, X., Valen, E., Yusuf, D., Lenhard, B., Wasserman, W. W., and Sandelin, A. (2010). Jaspas 2010: the greatly expanded open-access database of transcription factor binding profiles. *Nucleic Acids Res*, **38**(Database issue), D105–D110.
- [10] Thomas, D. J., Rosenbloom, K. R., Clawson, H., Hinrichs, A. S., Trumbower, H., Raney, B. J., Karolchik, D., Barber, G. P., Harte, R. A., Hillman-Jackson, J., Kuhn, R. M., Rhead, B. L., Smith, K. E., Thakkapallayil, A., Zweig, A. S., Consortium, E. N. C. O. D. E. P., Haussler, D., and Kent, W. J. (2007). The ENCODE Project at UC Santa Cruz. *Nucleic Acids Res*, **35**(Database issue), D663–D667.