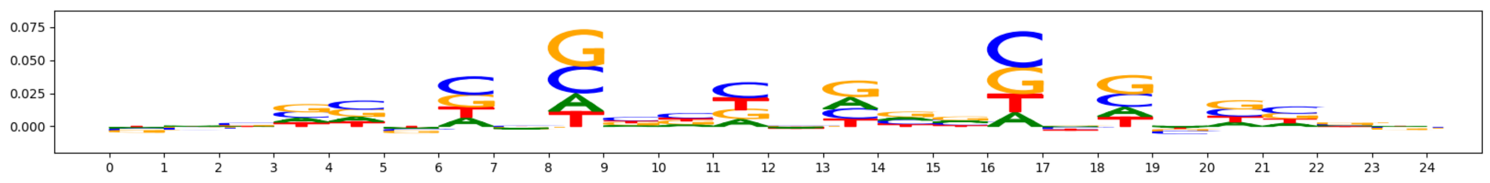
Bias factorized ChromBPNet training and quality check report

Preprocessing report

The image below should look closely like a Tn5 or DNase bias enzyme motif.



Bias model performance in peaks

Counts Metrics: The pearsonr in peaks should be greater than -0.3 (otherwise the bias model could potentially be capturing AT bias). MSE (Mean Squared Error) will be high in peaks.

Profile Metrics: Median JSD (Jensen Shannon Divergence between observed and predicted) lower the better. Median norm JSD is median of the min-max normalized JSD where min JSD is the worst case JSD i.e JSD of observed with uniform profile and max JSD is the best case JSD i.e 0. Median norm JSD is higher the better. Both JSD and median norm JSD are sensitive to read-depth. Higher read-depth results in better metrics.

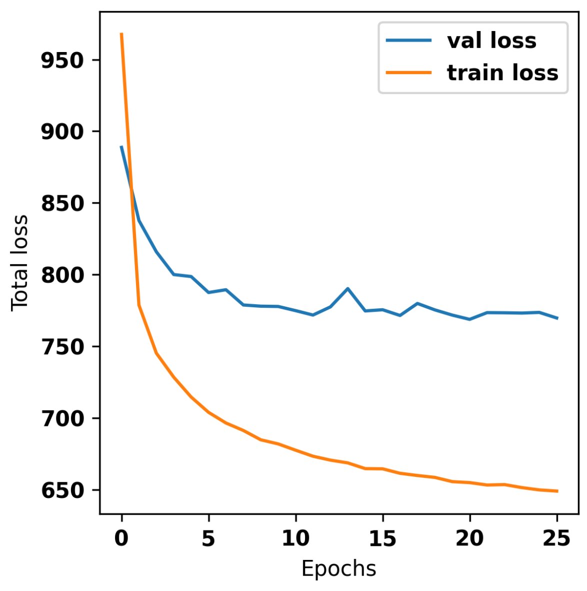
What to do if your pearsonr in peaks is less than -0.3? In the range of -0.3 to -0.5 please be wary of your chrombpnet_wo_bias.h5 TFModisco results showing lots of GC rich motifs (> 3 in the top-10). If this is not the case you can continue using the chrombpnet_wo_bias.h5. If you end up seeing a lot of GC rich motifs it is likely that bias model has learnt a different GC distribution than your GC-content in peaks. If you are transferring a bias model from a different sample you can consider using a different bias model or training a bias model for this sample. If you have trained a bias model for this sample and encounter this you might have to increase the bias_threshold_factor argument input to the *chrombpnet bias pipeline* or *chrombpnet bias train* command used in training the bias model and retrain a new bias model. For more intuition about this argument refer to the FAQ section in wiki. If the value is less than -0.5 the pipline will automatically throw an error.

	peaks.pearsonr	peaks.mse
counts_metrics	0.386012	12.086116

	peaks.median_jsd	peaks.median_norm_jsd
profile_metrics	0.636548	0.081944

Training report

The val loss (validation loss) will decrease and saturate after a few epochs.



ChromBPNet model performance in peaks

Counts Metrics: The pearsonr in peaks should be greater than 0.5 (higher the better). MSE (Mean Squared Error) will be low in peaks.

Profile Metrics: Median JSD (Jensen Shannon Divergence between observed and predicted) lower the better. Median norm JSD is median of the min-max normalized JSD where min JSD is the worst case JSD i.e JSD of observed with uniform profile and max JSD is the best case JSD i.e 0. Median norm JSD is higher the better. Both JSD and median norm JSD are sensitive to read-depth. Higher read-depth results in better metrics.

	peaks.pearsonr	peaks.mse
counts_metrics	0.661027	0.819197

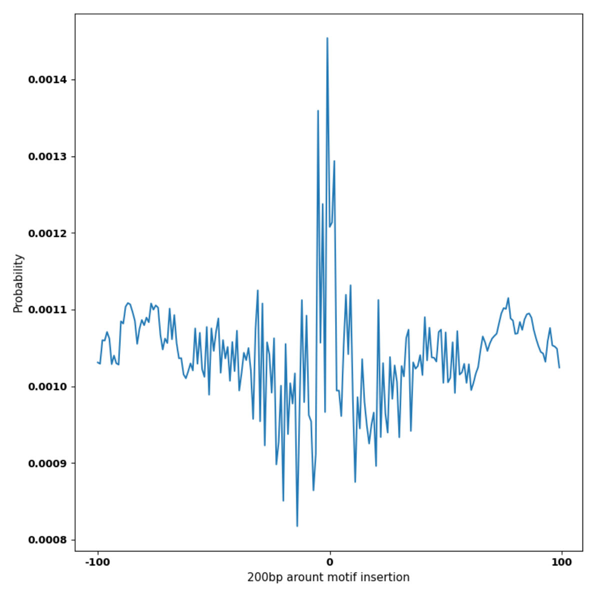
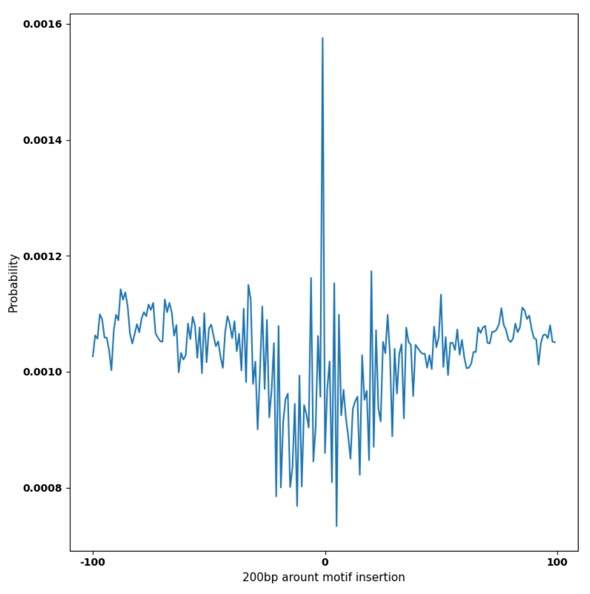
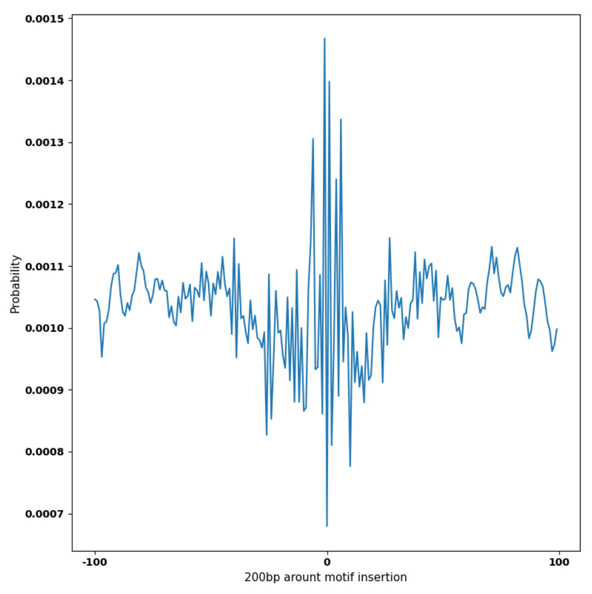
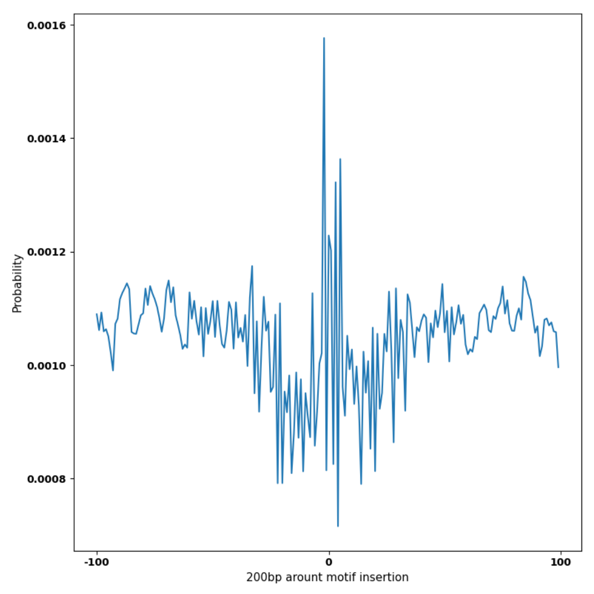
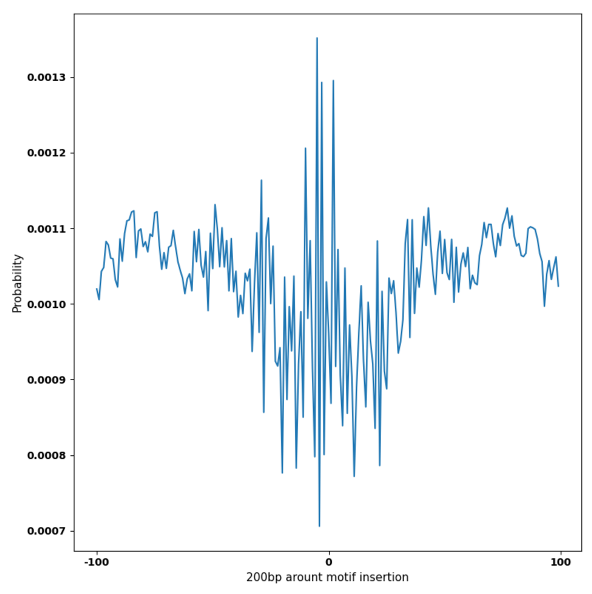
	peaks.median_jsd	peaks.median_norm_jsd
profile_metrics	0.517681	0.24542

ChromBPNet marginal footprints on tn5 motifs

The marginal footprints are the response of the ChromBPNet no bias model to the hetergenous bias motifs. If the bias correction is complete the max of the profiles below should be below 0.003 on all the bias motifs.

For your convenience we calculate here the average of the max of the profiles: 0.024 And the model according to this is **uncorrected**

What to do if your model looks uncorrected (i.e max of profiles is greater than 0.003)?
Look at the motifs below captured by TFModisco and you should be able to see motifs that closely look like the bias motifs showing incomplete bias correction. This indicates that your bias model was not completely capturing the response of the bias. We recommend that you use a different pre-trained bias model. For more intuition on choosing the correct pre-trained model or retraining your bias model refer to FAQ section in wiki.

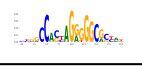
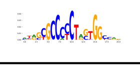
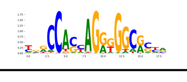
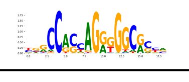
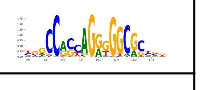
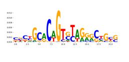
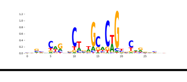
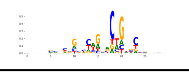
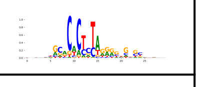
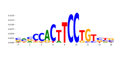
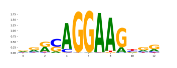
tn5 motif 1	tn5 motif 2	tn5 motif 3	tn5 motif 4	tn5 motif 5
				

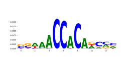
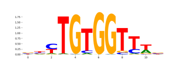
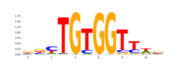
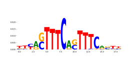
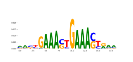
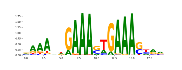
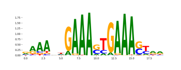
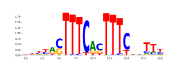
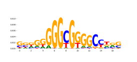
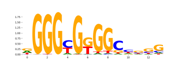
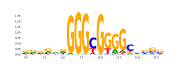
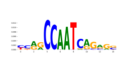
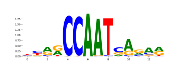
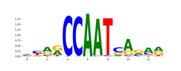
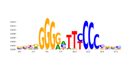
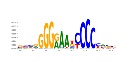
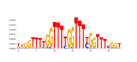
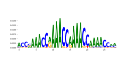
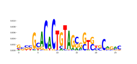
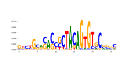
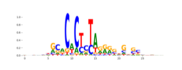
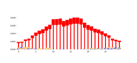
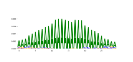
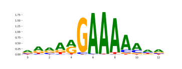
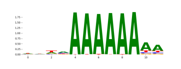
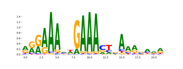
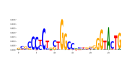
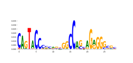
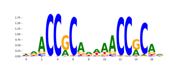
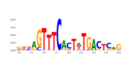
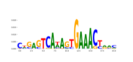
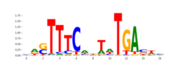
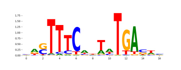
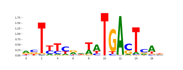
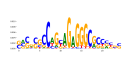
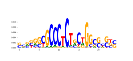
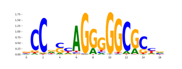
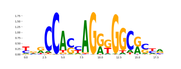
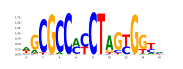
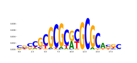
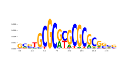
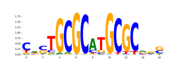
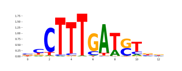
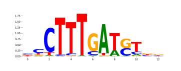
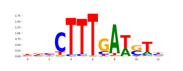
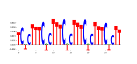
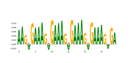
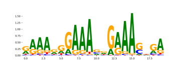
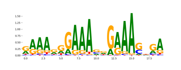
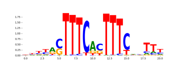
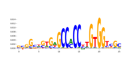
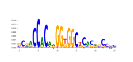
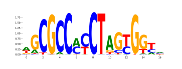
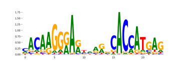
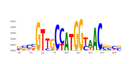
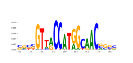
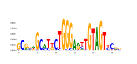
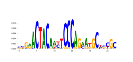
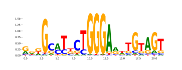
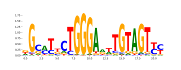
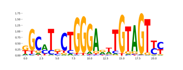
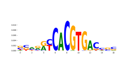
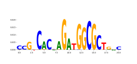
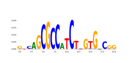
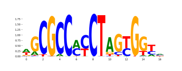
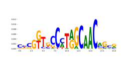
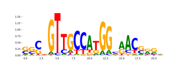
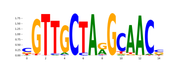
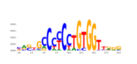
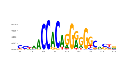
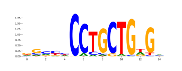
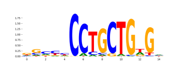
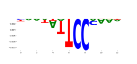
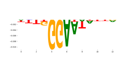
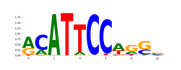
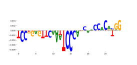
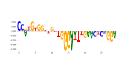
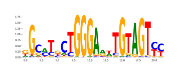
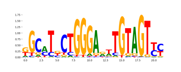
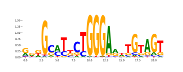
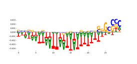
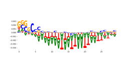
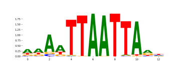
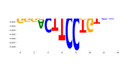
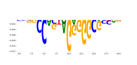
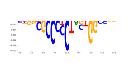
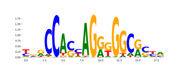
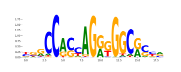
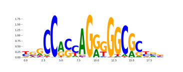
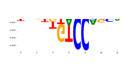
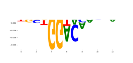
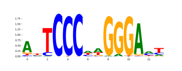
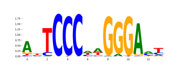
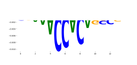
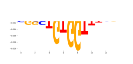
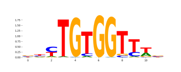
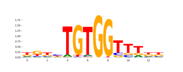
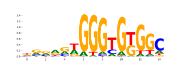
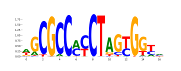
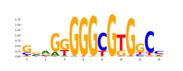
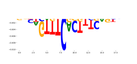
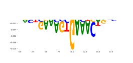
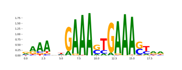
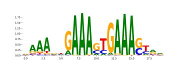
TFModisco motifs learnt from ChromBPNet after bias correction (chrombpnet_nobias.h5) model

TFModisco motifs generated from profile contribution scores of the ChromBPNet after bias correction model. cwm_fwd, cwm_rev are the forward and reverse complemented consolidated motifs from contribution scores in subset of random peaks. These CWM motifs should be free from any bias motifs and should contain only Transcription Factor (TF) motifs. For each of these motifs, we use TOMTOM to find the top-3 closest matches (match_0, match_1, match_2) from a database consisting of both MEME TF motifs and heterogenous enzyme bias motifs that we have repeatedly seen in our datasets. The qvals (qval0,qval1,qval2) should be low (< 0.0001) for most of the closest TF motif hits (i.e indicating that the closest match is the correct match) - this is also generally verifiable by eye as the closest match will look closely like the CWMs (atleast part of it in case of heterodimers). All the motifs in the list should look nothing like the enzyme bias motif.

What to do if you find an obvious bias motif in the list?
This indicates that your bias model was not completely capturing the response of the bias. We recommend that you use a different pre-trained bias model. For more intuition on choosing the correct pre-trained model or retraining your bias model refer to FAQ section in wiki.

What to do if you find an obvious bias motif in the list?

pattern	NumSeqs	cwm_fwd	cwm_rev	match0	qval0	match0_logo	match1	qval1	match1_logo	match2	qval2	match2_logo
pos__0	6676			CTCF_MA0139.1	2.365200e-08		CTCF_HUMAN.H11MO.0.A	2.135810e-07		CTCF_MOUSE.H11MO.0.A	2.252190e-07	
pos__1	4094			TN5_2	2.990870e-05		TN5_1	1.539880e-04		TN5_3	3.666160e-03	
pos__2	3641			ETS1_HUMAN.H11MO.0.A	1.706730e-03		ETV4_MOUSE.H11MO.0.B	1.988920e-03		ERG_HUMAN.H11MO.0.A	1.988920e-03	

pattern	NumSeqs	cwm_fwd	cwm_rev	match0	qval0	match0_logo	match1	qval1	match1_logo	match2	qval2	match2_logo
pos__3	2569			RUNX1_HUMAN.H11MO.0.A	6.755080e-02		RUNX1_MOUSE.H11MO.0.A	6.755080e-02		RUNX3_MOUSE.H11MO.0.A	6.755080e-02	
pos__4	1754			FOS+JUN_MA0099.3	1.820350e-04		JUN_HUMAN.H11MO.0.A	1.820350e-04		FOSL2+JUN_MA1130.1	1.820350e-04	
pos__5	1509			IRF1_HUMAN.H11MO.0.A	7.290600e-08		IRF1_MOUSE.H11MO.0.A	7.290600e-08		IRF1_MA0050.2	1.643320e-06	
pos__6	984			KLF5_MA0599.1	1.425480e-07		KLF1_HUMAN.H11MO.0.A	3.563700e-07		KLF3_HUMAN.H11MO.0.B	1.604560e-06	
pos__7	470			NFYA_HUMAN.H11MO.0.A	1.301930e-05		NFYA_MOUSE.H11MO.0.A	1.301930e-05		NFYC_HUMAN.H11MO.0.A	1.132290e-04	
pos__8	395			NFKB1_HUMAN.H11MO.1.B	3.672830e-06		NFKB1_MOUSE.H11MO.0.A	3.672830e-06		TF65_MOUSE.H11MO.0.A	1.312580e-05	
pos__9	292			IRF4_IRF_1	2.057930e-02		IRF4_MA1419.1	2.057930e-02		IRF5_IRF_1	2.618920e-02	
pos__10	246			TN5_3	4.519600e-01		TFE2_HUMAN.H11MO.0.A	4.519600e-01		TFE2_MOUSE.H11MO.0.A	4.519600e-01	
pos__11	184			PRDM6_HUMAN.H11MO.0.C	9.796820e-02		ZNF384_MA1125.1	9.796820e-02		STAT1_MOUSE.H11MO.0.A	1.591670e-01	
pos__12	156			ETS1_MOUSE.H11MO.0.A	6.099970e-01		ELK4_HUMAN.H11MO.0.A	6.099970e-01		RUNX3_RUNX_3	6.099970e-01	
pos__13	127			BATF3_HUMAN.H11MO.0.B	5.256100e-04		BATF3_MOUSE.H11MO.0.A	5.256100e-04		BATF_HUMAN.H11MO.0.A	5.847510e-02	
pos__14	126			CTCFL_HUMAN.H11MO.0.A	4.496240e-01		CTCF_MA0139.1	5.214380e-01		CTCF_C2H2_1	5.214380e-01	
pos__15	121			ATF2_HUMAN.H11MO.0.B	1.014090e-03		ATF2_MOUSE.H11MO.0.A	1.014090e-03		FOSB+JUNB_MA1136.1	1.014090e-03	
pos__16	117			NRF1_MA0506.1	1.985530e-03		NRF1_MOUSE.H11MO.0.A	1.985530e-03		NRF1_NRF_1	1.985530e-03	
pos__17	117			TF7L2_HUMAN.H11MO.0.A	1.072820e-05		TF7L2_MOUSE.H11MO.0.A	1.072820e-05		TF7L1_HUMAN.H11MO.0.B	8.191080e-05	
pos__18	105			IRF3_HUMAN.H11MO.0.B	6.471380e-05		IRF3_MOUSE.H11MO.0.A	6.471380e-05		IRF1_MA0050.2	2.691260e-04	
pos__19	103			CTCF_C2H2_1	7.974850e-01		ZNF41_HUMAN.H11MO.0.C	7.974850e-01		JUNB_HUMAN.H11MO.0.A	7.974850e-01	
pos__20	85			RFX5_RFX_2	1.500120e-07		RFX5_RFX_3	1.500120e-07		Rfx3.mouse_RFX_1	8.839410e-07	
pos__21	68			ZNF76_HUMAN.H11MO.0.C	2.988000e-16		ZN143_HUMAN.H11MO.0.A	5.750800e-10		THA11_MOUSE.H11MO.0.B	1.539750e-09	
pos__22	66			TFEB_MA0692.1	8.477910e-05		TFEB_bHLH_1	8.477910e-05		USF1_bHLH_1	8.477910e-05	
pos__23	49			CTCF_C2H2_1	4.840660e-02		TYY1_HUMAN.H11MO.0.A	4.840660e-02		TYY1_MOUSE.H11MO.0.A	5.248840e-02	
pos__24	31			Rfx1_MA0509.1	5.737180e-04		RFX2_MOUSE.H11MO.0.A	6.494520e-04		RFX3_RFX_2	6.494520e-04	
pos__25	30			ZIC3_HUMAN.H11MO.0.B	8.316770e-01		ZIC3_MOUSE.H11MO.0.A	8.316770e-01		ETV6_MOUSE.H11MO.0.C	8.316770e-01	
neg__0	144			RELB_MA1117.1	6.988540e-01		TEAD1_MOUSE.H11MO.0.A	6.988540e-01		TEAD3_MA0808.1	6.988540e-01	
neg__1	95			ZN143_MOUSE.H11MO.0.A	2.070950e-01		THA11_MOUSE.H11MO.0.B	2.070950e-01		ZNF76_HUMAN.H11MO.0.C	2.070950e-01	
neg__2	94			LHX3_HUMAN.H11MO.0.C	1.835040e-02		Arid3b_MA0601.1	1.835040e-02		ONECUT3_CUT_1	2.767970e-02	
neg__3	37			ETV2_HUMAN.H11MO.0.B	4.088310e-04		ETV2_MOUSE.H11MO.0.A	4.088310e-04		ETV4_MOUSE.H11MO.0.B	9.290560e-04	
neg__4	35			CTCF_MA0139.1	1.583940e-03		CTCF_HUMAN.H11MO.0.A	1.583940e-03		CTCF_MOUSE.H11MO.0.A	1.583940e-03	
neg__5	33			EBF1_EBF_1	1.000000e+00		EBF1_MA0154.3	1.000000e+00		E2F4_MOUSE.H11MO.0.A	1.000000e+00	
neg__6	30			RUNX1_HUMAN.H11MO.0.A	9.107180e-02		RUNX1_MOUSE.H11MO.0.A	9.107180e-02		RUNX2_HUMAN.H11MO.0.A	9.107180e-02	
neg__7	28			KLF4_MOUSE.H11MO.0.A	2.764430e-01		CTCF_C2H2_1	2.764430e-01		KLF9_HUMAN.H11MO.0.C	2.764430e-01	
neg__8	22			IRF1_MOUSE.H11MO.0.A	4.642490e-04		IRF2_HUMAN.H11MO.0.A	4.642490e-04		IRF2_MOUSE.H11MO.0.B	4.642490e-04