Supplementary material for
RoboCOP: Jointly computing chromatin occupancy profiles for numerous factors from chromatin accessibility data

May 12, 2020

<table>
<thead>
<tr>
<th>Abf1</th>
<th>Abf2</th>
<th>Ace2</th>
<th>Adr1</th>
<th>Aft1</th>
<th>Aft2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aro80</td>
<td>Asg1</td>
<td>Azf1</td>
<td>Bas1</td>
<td>Cad1</td>
<td>Cat8</td>
</tr>
<tr>
<td>Cbf1</td>
<td>Cep3</td>
<td>Cha4</td>
<td>Cin5</td>
<td>Crz1</td>
<td>Cst6</td>
</tr>
<tr>
<td>Cup9</td>
<td>Dal80</td>
<td>Dal82</td>
<td>Ecm22</td>
<td>Ecm23</td>
<td>Fhl1</td>
</tr>
<tr>
<td>Fkh1</td>
<td>Fkh2</td>
<td>FzT1</td>
<td>Gal4</td>
<td>Gat1</td>
<td>Gat3</td>
</tr>
<tr>
<td>Gat4</td>
<td>Gen4</td>
<td>Ger1</td>
<td>Gis1</td>
<td>Glh3</td>
<td>Gsm1</td>
</tr>
<tr>
<td>Gzf3</td>
<td>Hac1</td>
<td>Hal9</td>
<td>Hap1</td>
<td>Hcm1</td>
<td>Hmlalpha2</td>
</tr>
<tr>
<td>Hmra2</td>
<td>Hsf1</td>
<td>Leu3</td>
<td>Lys14</td>
<td>Matalpha2</td>
<td>Mbp1</td>
</tr>
<tr>
<td>Mcm1</td>
<td>Met31</td>
<td>Met32</td>
<td>Mga1</td>
<td>Mig1</td>
<td>Mig2</td>
</tr>
<tr>
<td>Mig3</td>
<td>Mot3</td>
<td>Msn1</td>
<td>Msn2</td>
<td>Msn4</td>
<td>Ndt80</td>
</tr>
<tr>
<td>Nhp10</td>
<td>Nhp6a</td>
<td>Nhp6b</td>
<td>Nrg1</td>
<td>Nrg2</td>
<td>Oaf1</td>
</tr>
<tr>
<td>Pbfl</td>
<td>Pbf2</td>
<td>Pdr1</td>
<td>Pdr3</td>
<td>Pld1</td>
<td>Pdr8</td>
</tr>
<tr>
<td>Pho2</td>
<td>Pho4</td>
<td>Put3</td>
<td>Rap1</td>
<td>Rdr1</td>
<td>Rds1</td>
</tr>
<tr>
<td>Rds2</td>
<td>Reb1</td>
<td>Re1</td>
<td>Rfx1</td>
<td>Rgm1</td>
<td>Rgt1</td>
</tr>
<tr>
<td>Rim101</td>
<td>Rox1</td>
<td>Rph1</td>
<td>Rpn4</td>
<td>Rsc3</td>
<td>Rsc30</td>
</tr>
<tr>
<td>Rtg3</td>
<td>Sf1</td>
<td>Sfp1</td>
<td>Sig1</td>
<td>Sip4</td>
<td>Skn7</td>
</tr>
<tr>
<td>Sko1</td>
<td>Smn1</td>
<td>Sok2</td>
<td>Spt15</td>
<td>Srd1</td>
<td>Stb3</td>
</tr>
<tr>
<td>Stb4</td>
<td>Stb5</td>
<td>Ste12</td>
<td>Stp1</td>
<td>Stp2</td>
<td>Stp3</td>
</tr>
<tr>
<td>Stp4</td>
<td>Sum1</td>
<td>Sut1</td>
<td>Sut2</td>
<td>Swi4</td>
<td>Swi5</td>
</tr>
<tr>
<td>Tbf1</td>
<td>Tbs1</td>
<td>Tca1</td>
<td>Tec1</td>
<td>Tos8</td>
<td>Tye7</td>
</tr>
<tr>
<td>Uga3</td>
<td>Ume6</td>
<td>Upc2</td>
<td>Usv1</td>
<td>Vhr1</td>
<td>Xbp1</td>
</tr>
<tr>
<td>Yap1</td>
<td>Yap3</td>
<td>Yap6</td>
<td>YBR033W</td>
<td>YBR239C</td>
<td>YDR520C</td>
</tr>
<tr>
<td>YER064C</td>
<td>YER130C</td>
<td>YER184C</td>
<td>YGR067C</td>
<td>YKL222C</td>
<td>YLL054C</td>
</tr>
<tr>
<td>YER278C</td>
<td>YML081W</td>
<td>YNR063W</td>
<td>Yox1</td>
<td>YPR013C</td>
<td>YPR015C</td>
</tr>
<tr>
<td>YPR022C</td>
<td>YPR196W</td>
<td>Yrm1</td>
<td>Yrr1</td>
<td>Zap1</td>
<td>Zms1</td>
</tr>
</tbody>
</table>

Table S1: List of 150 TFs used by RoboCOP. TFs ones colored in gray do not have binding sites reported in (1). TFs colored in red have alternative validation sets from ChIP-exo (2) and ORGANIC (3) protocols. TFs colored in blue are known to be expressed at increased levels under specific experimental conditions, so may not be expressed in our own data (4).
<table>
<thead>
<tr>
<th>upmost 100</th>
<th>downmost 100</th>
<th>constant 100</th>
</tr>
</thead>
<tbody>
<tr>
<td>HSP12 (6.30)</td>
<td>RPL9A (−4.59)</td>
<td>AIF1 (0.00)</td>
</tr>
<tr>
<td>MET3 (5.87)</td>
<td>RPS22A (−4.51)</td>
<td>YMR272W-B (0.00)</td>
</tr>
<tr>
<td>PDC6 (5.69)</td>
<td>SSB1 (−4.43)</td>
<td>YBR013C (0.00)</td>
</tr>
<tr>
<td>YDL124W (5.59)</td>
<td>RPS31 (−4.42)</td>
<td>YNR070W (0.00)</td>
</tr>
<tr>
<td>HSP26 (5.41)</td>
<td>RPS3 (−4.36)</td>
<td>COS12 (0.00)</td>
</tr>
<tr>
<td>MET14 (5.41)</td>
<td>ASC1 (−4.34)</td>
<td>UBS1 (0.00)</td>
</tr>
<tr>
<td>DDR48 (5.15)</td>
<td>RPL3 (−4.34)</td>
<td>IPL1 (0.00)</td>
</tr>
<tr>
<td>AHP1 (4.99)</td>
<td>YEF3 (−4.31)</td>
<td>SEN54 (0.00)</td>
</tr>
<tr>
<td>MET10 (4.85)</td>
<td>RPS31 (−4.30)</td>
<td>OSW1 (0.00)</td>
</tr>
<tr>
<td>HSP26 (5.41)</td>
<td>RPS3 (−4.36)</td>
<td>COS12 (0.00)</td>
</tr>
<tr>
<td>MET14 (5.41)</td>
<td>ASC1 (−4.34)</td>
<td>UBS1 (0.00)</td>
</tr>
<tr>
<td>DDR48 (5.15)</td>
<td>RPL3 (−4.34)</td>
<td>IPL1 (0.00)</td>
</tr>
<tr>
<td>AHP1 (4.99)</td>
<td>YEF3 (−4.31)</td>
<td>SEN54 (0.00)</td>
</tr>
<tr>
<td>MET10 (4.85)</td>
<td>RPS31 (−4.30)</td>
<td>OSW1 (0.00)</td>
</tr>
<tr>
<td>upmost 100</td>
<td>downmost 100</td>
<td>constant 100</td>
</tr>
<tr>
<td>-------------------</td>
<td>----------------------</td>
<td>---------------------------</td>
</tr>
<tr>
<td>PRX1 (3.07)</td>
<td>RPS4B (−3.46)</td>
<td>CCZ1 (0.00)</td>
</tr>
<tr>
<td>YMR315W (3.04)</td>
<td>RPS20 (−3.44)</td>
<td>SPT8 (0.00)</td>
</tr>
<tr>
<td>TPS1 (3.03)</td>
<td>RPL42B (−3.44)</td>
<td>MDM1 (0.00)</td>
</tr>
<tr>
<td>TFS1 (3.00)</td>
<td>RPL27A (−3.42)</td>
<td>HMI1 (0.00)</td>
</tr>
<tr>
<td>MXR1 (2.99)</td>
<td>RPS18B (−3.42)</td>
<td>VID24 (0.00)</td>
</tr>
<tr>
<td>PDI1 (2.99)</td>
<td>RPS0A (−3.39)</td>
<td>STB6 (0.00)</td>
</tr>
<tr>
<td>SAM2 (2.99)</td>
<td>RPL12B (−3.37)</td>
<td>CDC9 (0.00)</td>
</tr>
<tr>
<td>TRX2 (2.97)</td>
<td>RPS10A (−3.37)</td>
<td>YOR387C (0.00)</td>
</tr>
<tr>
<td>YBR085C-A (2.97)</td>
<td>RPL13B (−3.36)</td>
<td>SPC110 (0.00)</td>
</tr>
<tr>
<td>MRP8 (2.97)</td>
<td>RPL6B (−3.36)</td>
<td>STE5 (0.00)</td>
</tr>
<tr>
<td>FMO1 (2.95)</td>
<td>RPL8A (−3.35)</td>
<td>RTC4 (0.00)</td>
</tr>
<tr>
<td>MET2 (2.95)</td>
<td>RPL24A (−3.34)</td>
<td>TOP1 (0.00)</td>
</tr>
<tr>
<td>YOR285W (2.94)</td>
<td>NSR1 (−3.33)</td>
<td>YCR090C (0.00)</td>
</tr>
<tr>
<td>MUP3 (2.93)</td>
<td>RPS24B (−3.30)</td>
<td>YPR027C (0.00)</td>
</tr>
<tr>
<td>GSF2 (2.92)</td>
<td>RPS18A (−3.30)</td>
<td>PSR2 (0.00)</td>
</tr>
<tr>
<td>ALD4 (2.92)</td>
<td>RPS6B (−3.30)</td>
<td>BUB1 (0.01)</td>
</tr>
<tr>
<td>GRX1 (2.90)</td>
<td>RPL11B (−3.29)</td>
<td>YOR072W-B (0.01)</td>
</tr>
<tr>
<td>CMK2 (2.88)</td>
<td>RPS6A (−3.28)</td>
<td>LDB16 (0.01)</td>
</tr>
<tr>
<td>EN01 (2.88)</td>
<td>RPL6A (−3.26)</td>
<td>YDR042C (0.01)</td>
</tr>
<tr>
<td>YLR194C (2.86)</td>
<td>RPL20A (−3.23)</td>
<td>ATS1 (0.01)</td>
</tr>
<tr>
<td>SER33 (2.86)</td>
<td>RPS24A (−3.22)</td>
<td>LOH1 (0.01)</td>
</tr>
<tr>
<td>ARA1 (2.85)</td>
<td>RPL43A (−3.22)</td>
<td>TAF13 (0.01)</td>
</tr>
<tr>
<td>GRE2 (2.84)</td>
<td>RPS8B (−3.21)</td>
<td>YER188C-A (0.01)</td>
</tr>
<tr>
<td>GPD1 (2.83)</td>
<td>RPL33A (−3.21)</td>
<td>MSL1 (0.01)</td>
</tr>
<tr>
<td>MUP1 (2.82)</td>
<td>NOP1 (−3.20)</td>
<td>YDR194W-A (0.01)</td>
</tr>
<tr>
<td>HOR7 (2.82)</td>
<td>RPL4A (−3.19)</td>
<td>RUD3 (0.01)</td>
</tr>
<tr>
<td>DDR2 (2.82)</td>
<td>RPL14A (−3.17)</td>
<td>YHR100C (0.01)</td>
</tr>
<tr>
<td>GPM1 (2.80)</td>
<td>RPS19B (−3.17)</td>
<td>YOR293C-A (0.01)</td>
</tr>
<tr>
<td>MET22 (2.79)</td>
<td>TMA19 (−3.17)</td>
<td>SIN3 (0.01)</td>
</tr>
<tr>
<td>WTM1 (2.69)</td>
<td>RPL16A (−3.17)</td>
<td>FYV12 (0.01)</td>
</tr>
<tr>
<td>ARN2 (2.68)</td>
<td>RPL37A (−3.16)</td>
<td>TRP3 (0.01)</td>
</tr>
<tr>
<td>RTS3 (2.66)</td>
<td>RPL19A (−3.16)</td>
<td>SMK1 (0.01)</td>
</tr>
<tr>
<td>CPR1 (2.63)</td>
<td>RPL7A (−3.16)</td>
<td>PFD1 (0.01)</td>
</tr>
<tr>
<td>YER067W (2.62)</td>
<td>RPS17A (−3.15)</td>
<td>YGR146C-A (0.01)</td>
</tr>
<tr>
<td>GSH1 (2.61)</td>
<td>RPL17B (−3.14)</td>
<td>KAR9 (0.01)</td>
</tr>
<tr>
<td>PGM2 (2.61)</td>
<td>RPS23B (−3.12)</td>
<td>FIG2 (0.01)</td>
</tr>
<tr>
<td>GSC2 (2.61)</td>
<td>RPS27B (−3.11)</td>
<td>PGA1 (0.01)</td>
</tr>
<tr>
<td>PRC1 (2.60)</td>
<td>RPL20B (−3.11)</td>
<td>YKR051W (0.01)</td>
</tr>
<tr>
<td>MET8 (2.56)</td>
<td>RPS11B (−3.10)</td>
<td>SMC2 (0.01)</td>
</tr>
<tr>
<td>HOM6 (2.56)</td>
<td>RPL23A (−3.10)</td>
<td>SNF6 (0.01)</td>
</tr>
<tr>
<td>STF1 (2.55)</td>
<td>RPS28A (−3.08)</td>
<td>YLR412C-A (0.01)</td>
</tr>
<tr>
<td>HOR1 (2.55)</td>
<td>STM1 (−3.07)</td>
<td>FUS1 (0.01)</td>
</tr>
<tr>
<td>HNT1 (2.54)</td>
<td>RPL40B (−3.05)</td>
<td>MTW1 (0.01)</td>
</tr>
<tr>
<td>YET1 (2.52)</td>
<td>RPL34B (−3.05)</td>
<td>CNA1 (0.01)</td>
</tr>
<tr>
<td>YHR138C (2.52)</td>
<td>RPP1B (−3.04)</td>
<td>CDC12 (0.01)</td>
</tr>
<tr>
<td>MMP1 (2.51)</td>
<td>RPL39 (−3.03)</td>
<td>MscC7 (0.01)</td>
</tr>
<tr>
<td>HSP78 (2.51)</td>
<td>ILV3 (−3.02)</td>
<td>SMC6 (0.01)</td>
</tr>
<tr>
<td>HXK1 (2.51)</td>
<td>RPL23B (−2.99)</td>
<td>YMCl (0.01)</td>
</tr>
</tbody>
</table>

Table S2: Lists of 100 most up-regulated, 100 most down-regulated, and 100 most constant genes after 60 minutes of cadmium treatment. Upmost and downmost are sorted with greatest change first (log-fold expression change in parentheses). Most constant are sorted with least absolute change first (absolute log-fold expression change in parentheses).
Table S3: F1-score of nucleosome positions predicted by RoboCOP, COMPETE, and randomly selected nucleosome positions. COMPETE predicted nucleosomes are only slightly better than random prediction of nucleosome positions.

<table>
<thead>
<tr>
<th></th>
<th>RoboCOP</th>
<th>COMPETE</th>
<th>random</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.77</td>
<td>0.54</td>
<td>0.52</td>
</tr>
</tbody>
</table>

Figure S1: Transition diagram among the different DBFs. The DBFs are nucleosome, $K$ number of TFs and unbound state. The model transitions to any DBF (pink), say $\pi_k$, with probability $\alpha_k$ and then comes back to the silent state (green) with a probability of 1. From the silent state it can then transition to another DBF. The transitions among the DBFs are therefore independent of one another.

Figure S2: (a) State transition diagram of a TF. From the first hidden silent state (pink) of the TF, the model can transition to the forward motif (top) or the reverse motif (bottom) with equal probability of 0.5. It then transitions through the entire length, $L_k$, of the motif. After transitioning through the entire motif, the model transitions back to the central silent state (green) with probability of 1. (b) State transition diagram of the nucleosome model, $\pi_{K+1}$. Nucleosomes are modeled to be 147 bases long. After transitioning through the 147 nucleosome positions the model transitions back to the central silent state.
Estimated quantiles
Empirical quantiles at 91
0 5 10 15 20
0 10 20 30 40 50
Estimated quantiles
Empirical quantiles at 92
0 5 10 15 20
0 10 20 30 40 50
Estimated quantiles
Empirical quantiles at 93
0 5 10 15 20
0 10 20 30 40 50
Estimated quantiles
Empirical quantiles at 94
0 5 10 15 20
0 10 20 30 40 50
Estimated quantiles
Empirical quantiles at 95
0 5 10 15 20
0 10 20 30 40 50
Estimated quantiles
Empirical quantiles at 96
0 5 10 15 20
0 10 20 30 40 50
Estimated quantiles
Empirical quantiles at 97
0 5 10 15 20
0 10 20 30 40 50
Estimated quantiles
Empirical quantiles at 98
0 5 10 15 20
0 10 20 30 40 50
Estimated quantiles
Empirical quantiles at 99
0 5 10 15 20
0 10 20 30 40 50
Estimated quantiles
Empirical quantiles at 100
0 5 10 15 20
0 10 20 30 40 50
Estimated quantiles
Empirical quantiles at 101
0 5 10 15 20
0 10 20 30 40 50
Estimated quantiles
Empirical quantiles at 102
0 5 10 15 20
0 10 20 30 40 50
Estimated quantiles
Empirical quantiles at 103
0 5 10 15 20
0 10 20 30 40 50
Estimated quantiles
Empirical quantiles at 104
0 5 10 15 20
0 10 20 30 40 50
Estimated quantiles
Empirical quantiles at 105
0 5 10 15 20
0 10 20 30 40 50
Estimated quantiles
Empirical quantiles at 106
0 5 10 15 20
0 10 20 30 40 50
Estimated quantiles
Empirical quantiles at 107
0 5 10 15 20
0 10 20 30 40 50
Estimated quantiles
Empirical quantiles at 108
0 5 10 15 20
0 10 20 30 40 50
Estimated quantiles
Empirical quantiles at 109
0 5 10 15 20
0 10 20 30 40 50
Estimated quantiles
Empirical quantiles at 110
0 5 10 15 20
0 10 20 30 40 50
Estimated quantiles
Empirical quantiles at 111
0 5 10 15 20
0 10 20 30 40 50
Estimated quantiles
Empirical quantiles at 112
0 5 10 15 20
0 10 20 30 40 50
Estimated quantiles
Empirical quantiles at 113
0 5 10 15 20
0 10 20 30 40 50
Estimated quantiles
Empirical quantiles at 114
0 5 10 15 20
0 10 20 30 40 50
Estimated quantiles
Empirical quantiles at 115
0 5 10 15 20
0 10 20 30 40 50
Estimated quantiles
Empirical quantiles at 116
0 5 10 15 20
0 10 20 30 40 50
Estimated quantiles
Empirical quantiles at 117
0 5 10 15 20
0 10 20 30 40 50
Estimated quantiles
Empirical quantiles at 118
0 5 10 15 20
0 10 20 30 40 50
Estimated quantiles
Empirical quantiles at 119
0 5 10 15 20
0 10 20 30 40 50
Estimated quantiles
Empirical quantiles at 120
Figure S3: Quantile-quantile plots of negative binomial distributions for nucFrags midpoint counts at each of the 147 positions of a nucleosome. The shaded gray region corresponds to the 95% confidence interval of the plots. The plots show the estimated negative binomial distributions are good fits to the midpoint counts of nucFrags.
Figure S4: Quantile-quantile plot of negative binomial distribution for \textit{shortFrags} midpoint counts at annotated Abf1 and Reb1 binding sites (1). The plot shows the estimated negative binomial distribution is a good fit to the midpoint counts of \textit{shortFrags}.

Figure S5: Likelihood in the 10 iterations of EM of \textit{RoboCOP}. The likelihood converges in 10 iterations.
Figure S6: Transition weights of DBFs in every iteration of EM when initializing using multiples of $K_D$.

Figure S7: Choosing threshold (dotted red line) for transition probabilities of TFs. Threshold value is calculated to be 2 standard deviations more than the mean of all initial TF transition probabilities. Only three TFs (Nhp6b, Rox1, and Sig1) have an initial transition probability greater than the threshold.
Figure S8: (a) Aggregate of MNase-seq nucFrag and shortFrag from positions 60,000 to 65,000 of yeast chromosome I. Above the plot are genomic annotations for this region, with Watson strand genes depicted as green arrows and Crick strand genes as orange. Below the gene annotations, known TF binding sites (1) are indicated using triangles. This region contains two annotated binding sites for Abf1 (pink). (b) Running RoboCOP without a threshold and without an ‘unknown’ factor, (c) with a threshold but without an ‘unknown’ factor, (d) without a threshold but with an ‘unknown’ factor, and (e) with a threshold and with an ‘unknown’ factor (i.e., the final model). In (b,c), we observe large TF clusters that result in increased false positive predictions. In (d), without the threshold, we see that most TFs are predicted as being an ‘unknown’ factor. In (e), by adding an ‘unknown’ factor and a threshold on how large a TF transition probability can grow, we achieve a happy medium.
Figure S9: (a–c) Aggregate nucleosome dyad probability, as computed by COMPETE, around annotated +1 nucleosomes (5) of (a) the 100 most up-regulated genes (purple), (b) the 100 genes least changed in transcription (gray), and (c) the 100 most down-regulated genes (teal), before and 60 minutes after treating cells with cadmium. We find that COMPETE does not identify any change in these nucleosome occupancy profiles after cadmium treatment. (d) Two-dimensional plot of MNase-seq fragments near the HSP26 promoter (positions 380,700 to 382,350 of yeast chromosome II are shown) before treatment with cadmium (nucFrags in red; shortFrags in blue), along with the nucFrags and shortFrags signals that result from aggregating those midpoint counts. Gene annotations depicted with arrows at the top (Watson strand in green; Crick strand in orange). (e) COMPETE-predicted occupancy profile of this region before treatment with cadmium. (f,g) The same as (d,e), respectively, but 60 minutes after cadmium treatment. Even though the MNase-seq data have changed significantly after treatment (d,f), COMPETE does not make use of this data, so fails to capture the changes to the chromatin occupancy profile of the genome.

References


