Supporting Information

Honkela et al. 10.1073/pnas.0914285107

SI Text

Derivation of the Gaussian Process Model. The linear system of ordinary differential equations underlying our model is

$$\frac{dp(t)}{dt}=f(t)-\delta p(t), \quad \text{[S1]}$$

$$\frac{dm_j(t)}{dt}=B_j+S_j p(t)-D_j m_j(t), \quad \text{[S2]}$$

where $f(t)$ denotes the TF mRNA, $p(t)$ the transcription factor (TF) protein and $m_j(t)$ the target gene mRNA concentration. Assuming steady-state initial conditions $p(0)=0$ and $m_j(0)=B_j/D_j$, the solution of the system is

$$p(t)=\exp(-\delta t)\int_0^t f(v)\exp(\delta v)dv, \quad \text{[S3]}$$

$$m_j(t)=\frac{B_j}{D_j}+S_j\exp(-D_j t)\int_0^t \exp(D_j u)\exp(-\delta u)\int_0^u f(v)\exp(\delta v)dv\,du. \quad \text{[S4]}$$

Both of these are linear operators on $f(t)$. Hence, placing a Gaussian process prior on $f(t)$ implies a joint Gaussian process model over all $[f(t),p(t),m_j(t)]$ (1, 2). This Gaussian process is completely characterized by its mean and covariance functions. Assuming $\mathrm{E}[f(t)]=0$, the above solutions (Eqs. **S3** and **S4**) imply $\mathrm{E}[p(t)]=0$, $\mathrm{E}[m_j(t)]=B_j/D_j$.

What remains is to determine the covariance functions. These can be evaluated as expectations

$$k_{xy}(t,t')=\mathrm{E}[\{x(t)-\mathrm{E}[x(t)]\}\{y(t')-\mathrm{E}[y(t')]\}], \quad \text{[S5]}$$

where $x,y\in\{f,p,m_j\}$. Assuming the squared exponential covariance* for $f(t)$,

$$k_{ff}(t,t')=a\exp\left(-\frac{(t-t')^2}{l^2}\right), \quad \text{[S6]}$$

all the required covariance functions can be derived in closed form by repeated application of the identity

$$\int_0^t \exp(Du)\mathrm{erf}(u/l+E)du=\frac{1}{D}\left\{\exp(Dt)\mathrm{erf}(E+t/l)-\mathrm{erf}(E)+\exp\left[\left(\frac{Dl}{2}\right)^2-EDl\right][\mathrm{erf}(E-Dl/2)-\mathrm{erf}(E-Dl/2+t/l)]\right\}. \quad \text{[S7]}$$

Covariance function k_{fp}. The covariance of the TF mRNA and TF protein k_{fp} is the same as the cross-covariance derived in ref. 2. In our model, this is needed only for inference of the protein concentration (such as in Figs. 1 and 2). The covariance is

$$k_{fp}(t,t')=\exp(-\delta t')\int_0^{t'}\exp(\delta u)k_{ff}(t,u)du=\frac{\sqrt{\pi}al}{2}\exp\left[\left(\frac{\delta l}{2}\right)^2+\delta(t-t')\right]\{\mathrm{erf}(\delta l/2+t/l)-\mathrm{erf}[\delta l/2+(t-t')/l]\}. \quad \text{[S8]}$$

Covariance function k_{fm_j}. The covariance of the TF mRNA and target mRNA k_{fm_j} is

$$\begin{aligned}k_{fm_j}(t,t')&=S_j\exp(-D_j t')\int_0^{t'}\exp[(D_j-\delta)u]\times\int_0^u\exp(\delta v)k_{ff}(t,v)dv\,du\\&=S_j\frac{\sqrt{\pi}al}{2(\delta-D_j)}\exp[-(D_j+\delta)t')]\left\{\exp\left[\left(\frac{D_j l}{2}\right)^2+D_j t+\delta t'\right]\times\{\mathrm{erf}(D_j l/2+t/l)-\mathrm{erf}[D_j l/2+(t-t')/l]\}-\exp\left[\left(\frac{\delta l}{2}\right)^2+\delta t+D_j t'\right]\times\{\mathrm{erf}(\delta l/2+t/l)-\mathrm{erf}[\delta l/2+(t-t')/l]\}\right\}.\end{aligned} \quad \text{[S9]}$$

Covariance function k_{pp}. Again following ref. 2, the covariance of the TF protein k_{pp} is

$$k_{pp}(t,t')=\exp[-\delta(t+t')]\int_0^t\exp(\delta u)\int_0^{t'}\exp(\delta u')k_{ff}(u,u')du'du=\frac{\sqrt{\pi}al}{4\delta}\exp\left[\left(\frac{\delta l}{2}\right)^2-\delta(t+t')\right][h(t',t)+h(t,t')],$$

where

$$h(t',t)=\exp[2\delta t]\left[\mathrm{erf}\left(\frac{\delta l}{2}+\frac{t}{l}\right)-\mathrm{erf}\left(\frac{\delta l}{2}+\frac{t-t'}{l}\right)\right]+\left[\mathrm{erf}\left(\frac{\delta l}{2}-\frac{t'}{l}\right)-\mathrm{erf}\left(\frac{\delta l}{2}\right)\right]. \quad \text{[S10]}$$

This is needed only for inference of the protein concentrations.

Covariance function k_{pm_j}. The covariance of the TF protein and target mRNA k_{pm_j} is

*The form used here differs from standard squared exponential by reparametrization $2l^2\to l^2$ that makes the arguments of erfs simpler.

$$
\begin{aligned}
&\frac{k_{pm_j}(t,t')}{S_j\exp(-\delta t-D_jt')}\\
&=\int_0^{t'}\exp[(D_j-\delta)u']\int_0^{t}\exp(\delta v)\int_0^{u'}\exp(\delta v')k_{ff}(v,v')dv'dvdu'\\
&=\frac{\sqrt{\pi}al}{4\delta}\exp\left[\left(\frac{\delta l}{2}\right)^2\right]\left(\frac{2\delta\exp(-D_jt'-\delta t)}{\delta^2-D_j^2}[\mathrm{erf}(\delta l/2-t/l)-\mathrm{erf}(\delta l/2)]\right.\\
&+\frac{\exp[-\delta(t+t')]}{\delta-D_j}[2\mathrm{erf}(\delta l/2)-\mathrm{erf}(\delta l/2-t'/l)-\mathrm{erf}(\delta l/2-t/l)]\\
&+\frac{\exp[\delta(t'-t)]}{\delta+D_j}\{\mathrm{erf}(\delta l/2+t'/l)-\mathrm{erf}[\delta l/2-(t-t')/l]\}\\
&\left.+\frac{\exp[\delta(t-t')]}{\delta-D_j}\{\mathrm{erf}[\delta l/2+(t-t')/l]-\mathrm{erf}(\delta l/2+t/l)\}\right)\\
&+\frac{\sqrt{\pi}l}{2(\delta^2-D_j^2)}\exp\left[\left(\frac{D_jl}{2}\right)^2-D_jt'-\delta t\right]\{\mathrm{erf}(D_jl/2-t'/l)-\mathrm{erf}(D_jl/2)\\
&\left.+\exp[(D_j+\delta)t][\mathrm{erf}(D_jl/2+t/l)-\mathrm{erf}[D_jl/2+(t-t')/l]\}\right). \qquad \textbf{[S11}
\end{aligned}
$$

This is needed only for inference of the protein concentrations.

Covariance function $k_{m_jm_k}$. The final covariance between target genes $k_{m_jm_k}$ is

$$
\begin{aligned}
k_{m_jm_k}(t,t')&=S_jS_k\exp(-D_jt-D_kt')\int_0^t\exp[(D_j-\delta)u]\\
&\times\int_0^{t'}\exp[(D_k-\delta)u']\int_0^{u}\exp(\delta v)\\
&\times\int_0^{u'}\exp(\delta v')k_{ff}(v,v')dv'dv\,du'du\\
&=\frac{\sqrt{\pi}alS_jS_k}{2}[h_{jk}(t,t',\delta)+h_{kj}(t',t,\delta)\\
&-h_{jk}(t,t',D_j)-h_{kj}(t',t,D_k)], \qquad \textbf{[S12}
\end{aligned}
$$

where

$$
\begin{aligned}
h_{jk}(t,t',D_x)&=\exp\left[\left(\frac{D_xl}{2}\right)^2\right]\frac{\exp(-D_xt-D_kt')}{(D_x+\delta)(D_j-\delta)}\\
&\times\left[\left(\frac{\exp[(D_k-\delta)t']-1}{D_k-\delta}+\frac{1}{D_k+D_x}\right)[\mathrm{erf}(D_xl/2-t/l)\right.\\
&-\mathrm{erf}(D_xl/2)]+\frac{\exp[(D_k+D_x)t']}{D_k+D_x}\{\mathrm{erf}(D_xl/2+t'/l)\\
&\left.-\mathrm{erf}[D_xl/2-(t-t')/l]\}\right]. \qquad \textbf{[S13}
\end{aligned}
$$

Gaussian Process Inference. Denoting all the observations of replicate r by $\boldsymbol{y}_r$ and a diagonal matrix with their measurement variance parameters by $\Sigma_r=\mathrm{diag}(\sigma_{1f}^2,\ldots,\sigma_{nf}^2,\{\sigma_{1jm}^2,\ldots,\sigma_{njm}^2\})$, the full kernel is $K_r=K+\Sigma_r$, where K can be evaluated using the above formulas.

Based on standard Gaussian process regression (1), the posterior distribution of a vector $\boldsymbol{x}$ consisting of values of f, p and m_j, not necessarily at times of observations, is Gaussian with

$$
\boldsymbol{x}|\boldsymbol{y}_r\sim\mathcal{N}(\boldsymbol{\mu}_x+K_{xy}K_r^{-1}(\boldsymbol{y}_r-\boldsymbol{\mu}_y),K_{xx}-K_{xy}K_r^{-1}K_{yx}), \qquad \textbf{[S14}
$$

where $\boldsymbol{\mu}_x$ and $\boldsymbol{\mu}_y$ are the means of $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively. K_{xy} and K_{xx} can again be evaluated using the above formulas.

1. Rasmussen CE, Williams, CKI (2006) *Gaussian Processes for Machine Learning* (MIT Press, Cambridge, MA).

2. Lawrence ND, Sanguinetti G, Rattray M (2007) *Advances in Neural Information Processing Systems*, eds Schölkopf B, Platt JC, Hofmann T (MIT Press, Cambridge, MA), Vol 19, pp 785–792.

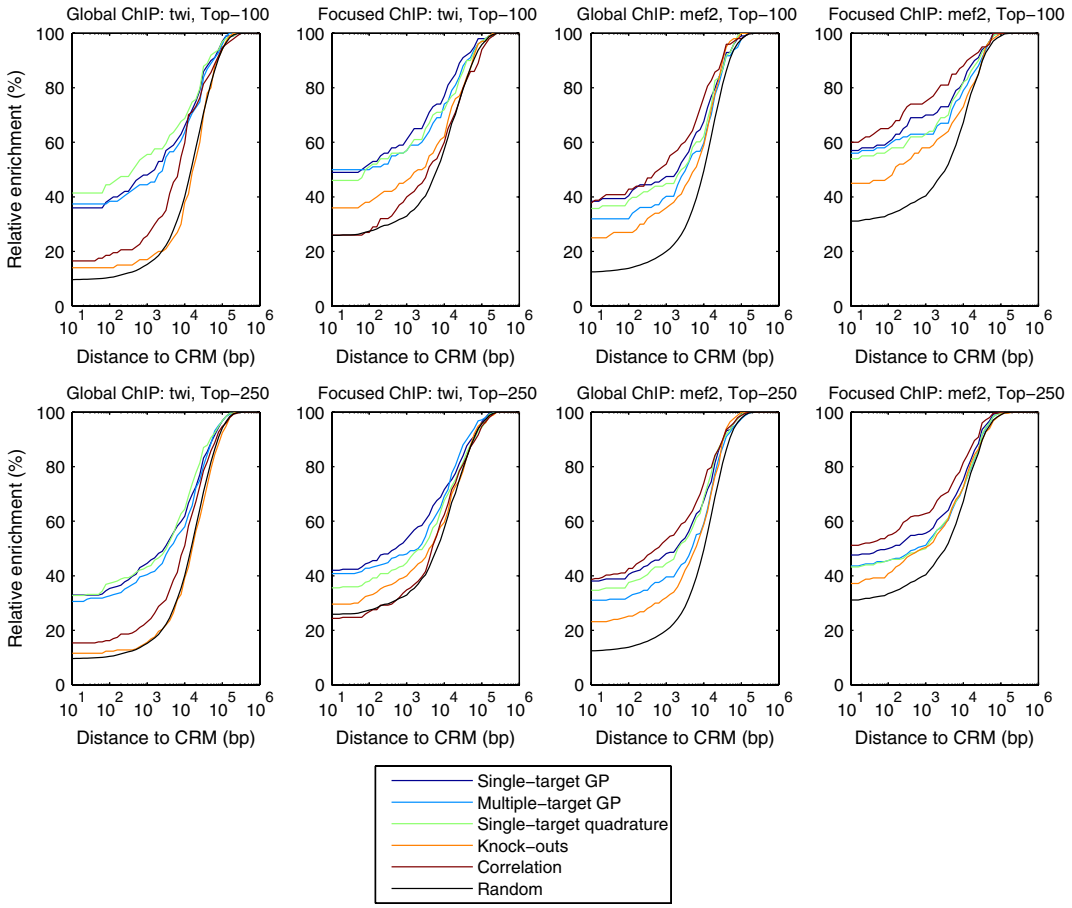

Fig. S1. The fraction of top 100 and top 250 predicted targets having a ChIP-chip binding site (*cis*-regulatory module, CRM) within *x* base pairs as a function of *x* in the different ChIP-chip evaluations used in the paper. Binding sites within an intron of a gene are considered to have a distance 0.

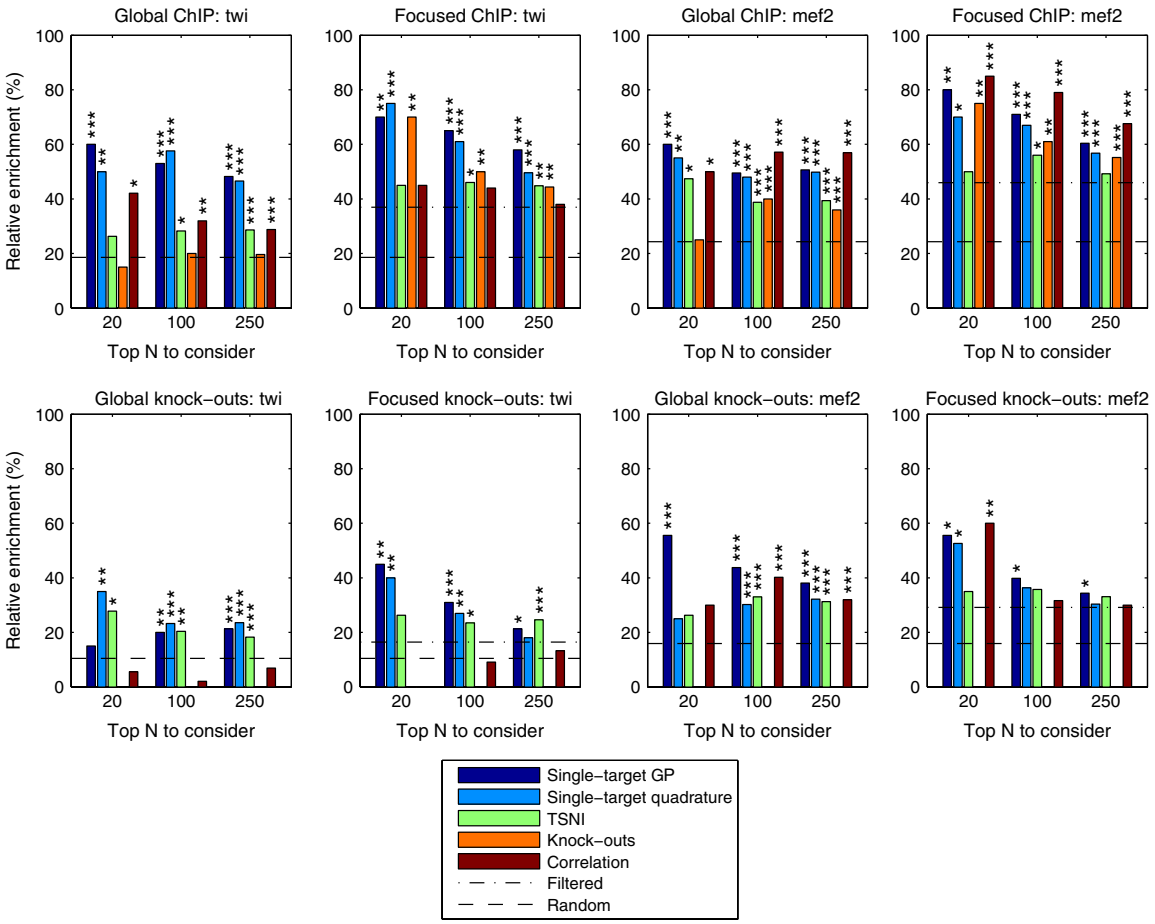

Fig. S2. Evaluation results of different rankings including a variant of the TSNI (time series network identification) method of Della Gatta et al. (1) using TF expression levels as a proxy for TF protein measurements. The plots show the relative frequency of positive predictions among *N* top-ranking targets, similar to Fig. 3 in the main text. The dashed line denotes the frequency in the full population and the dash-dotted line within the population considered in the focused evaluation. The first row shows the frequency of targets with ChIP-chip binding within 2,000 base pairs of the gene, whereas the second row shows the frequency of predicted targets with significant differential expression in TF knockouts. *p* values of results statistically significantly different from random are denoted by ***: $p < 0.001$, **: $p < 0.01$, *: $p < 0.05$. Comparison to knockout ranking is obviously omitted for knockout validation.

1. Della Gatta G, et al. (2008) Direct targets of the TRP63 transcription factor revealed by a combination of gene expression profiling and reverse engineering. *Genome Res* 18:939–948.

Table S1. The results of 100,000-fold bootstrap resampling of the dataset of observed genes to assess statistical significance of differences in ranking method performance

Twist global ChIP-chip

Top 20	ST	MT	QR	CO	KO
ST		*	.	+	***
MT					**
QR				.	**
CO					*
KO					

Top 100	ST	MT	QR	CO	KO
ST		*		***	***
MT				***	***
QR		+		***	***
CO					**
KO					

Top 250	ST	MT	QR	CO	KO
ST		**	.	***	***
MT				***	***
QR		+		***	***
CO					***
KO					

Twist global knockouts

Top 20	ST	MT	QR	CO	KO
ST		+		*	–
MT				.	–
QR	**	**		***	–
CO					–
KO	–	–	–	–	–

Top 100	ST	MT	QR	CO	KO
ST		.		***	–
MT				***	–
QR	.	+		***	–
CO					–
KO	–	–	–	–	–

Top 250	ST	MT	QR	CO	KO
ST		*		***	–
MT				***	–
QR	.	*		***	–
CO					–
KO	–	–	–	–	–

Twist focused ChIP-chip

Top 20	ST	MT	QR	CO	KO
ST		.		*	
MT				+	
QR	.	.		***	.
CO					
KO				+	

Top 100	ST	MT	QR	CO	KO
ST		+	.	***	**
MT				***	*
QR				***	**
CO					
KO				+	

Top 250	ST	MT	QR	CO	KO
ST		***	***	***	***
MT				***	*
QR				***	*
CO					
KO				**	

Twist focused knockouts

Top 20	ST	MT	QR	CO	KO
ST		.		***	–
MT				***	–
QR				***	–
CO					–
KO	–	–	–	–	–

Top 100	ST	MT	QR	CO	KO
ST		*	+	***	–
MT				***	–
QR		.		***	–
CO					–
KO	–	–	–	–	–

Top 250	ST	MT	QR	CO	KO
ST		*	**	***	–
MT				+	–
QR				**	–
CO					–
KO	–	–	–	–	–

Mef2 global ChIP-chip

Top 20	ST	MT	QR	CO	KO
ST		.	.	.	**
MT					**
QR					*
CO					*
KO					

Top 100	ST	MT	QR	CO	KO
ST		+			+
MT					.
QR					+
CO	*	**	*		***
KO					

Top 250	ST	MT	QR	CO	KO
ST		***			***
MT					*
QR		***			***
CO	**	***	**		***
KO					

Mef2 global knockouts

Top 20	ST	MT	QR	CO	KO
ST			*	*	–
MT			*	*	–
QR					–
CO					–
KO	–	–	–	–	–

Top 100	ST	MT	QR	CO	KO
ST		+	***		–
MT			**		–
QR					–
CO			*		–
KO	–	–	–	–	–

Top 250	ST	MT	QR	CO	KO
ST			**	**	–
MT			**	*	–
QR					–
CO					–
KO	–	–	–	–	–

Mef2 focused ChIP-chip

Top 20	ST	MT	QR	CO	KO
ST			+		
MT			*		.
QR					
CO			.		.
KO					

Top 100	ST	MT	QR	CO	KO
ST		*	*		**
MT					.
QR					.
CO	*	**	***		***
KO					

Top 250	ST	MT	QR	CO	KO
ST		**	**		*
MT					
QR					
CO	***	***	***		***
KO					

Mef2 focused knockouts

Top 20	ST	MT	QR	CO	KO
ST		+	.		–
MT					–
QR					–
CO					–
KO	–	–	–	–	–

Top 100	ST	MT	QR	CO	KO
ST			.	**	–
MT	*		***	***	–
QR				+	–
CO					–
KO	–	–	–	–	–

Top 250	ST	MT	QR	CO	KO
ST			**	*	–
MT			***	*	–
QR					–
CO					–
KO	–	–	–	–	–

The methods studied are ST, single-target GP method; MT, multiple-target GP method; QR, single-target quadrature method; CO, ranking by correlation; KO, ranking by *q* value in knockouts. For each pair of methods, the marks in the tables show how often the method on the corresponding row dominated the one on the corresponding column. The marks are interpreted as follows: .: >70% dominance, +: >80% dominance, *: >90% dominance, **: >95% dominance, ***: >99% dominance, –: comparison not applicable.