Supplementary Notes

1 GBR optimization

In this section we derive an efficient alternating optimization algorithm for the GBR objective (Methods). We first describe how to compute $\text{argmax}_q J'_{\text{GBR}}(\theta, q)$, then describe how this algorithm can be used in combination with an EM-like algorithm for learning θ.

GBR can be employed either as a regularizer for training the parameters or for inference directly. In the training case, an EM-like algorithm described is used to compute and output θ, which can then be used for inference either with or without GBR. In the inference case, q is computed and output as the posterior marginals. In our genomics experiments we trained our models without GBR and used GBR for inference only.

Optimizing q

The GBR regularizer $\mathcal{R}'_{\text{GBR}}(\theta, q)$ is convex in q; therefore, we could compute q using any convex optimization algorithm. However, general-purpose convex optimization algorithm do not scale to problems with millions or billions of variables such as those present in genomics. Therefore, we instead propose a novel alternating maximization strategy for performing this optimization more efficiently.

To enable efficient inference, we reformulate $J'_{\text{GBR}}(\theta, q)$ by introducing a new variable $r^M(X_H)$. Like q, r^M is a distribution over X_H, but we require that r^M be factorizable as a product of marginals—that is $r^M(x_H) = \prod_h r^M_h(x_h)$. We define the graph regularizer over r^M and add an additional term $\lambda_{\text{R1}} D(q(X_H) \| r^M(X_H))$, which encourages q and r^M to be similar. As we will show below, restricting r^M in this way means that the reformulated objective is a lower bound on the original rather than being equivalent. We will maximize this lower bound as an approximation to maximizing the original. The reformulated regularizer is

$$\mathcal{PR}'_{\text{GBR-R1}}(q, r^M) \triangleq -\lambda_{\text{R1}} D(q(X_H) \| r^M(X_H)) + f_{\text{R1}}(r^M) \tag{1}$$

$$f_{\text{R1}}(r^M) \triangleq -\lambda_G \sum_{(u,v)\in F_{GBR}} w(u,v) D(r^M(X_u) \| r^M(X_v)), \tag{2}$$

and $J'_{\text{GBR-R1}}(\theta, q, r^M)$ and $\mathcal{R}'_{\text{GBR-R1}}(\theta, q, r^M)$ are defined according to Equations (2) and (4) respectively using the corresponding regularizers. That is,

$$\text{maximize}_{\theta,q,r^M} \quad J'_{\text{GBR-R1}}(\theta, q, r^M) \triangleq \mathcal{L}(\theta) + \mathcal{R}'_{\text{GBR-R1}}(\theta, q, r^M), \tag{3}$$

$$\mathcal{R}'_{\text{GBR-R1}}(\theta, q, r^M) \triangleq -D(q(X_H) \| p_\theta(X_H|\bar{x}_O)) + \mathcal{PR}'_{\text{GBR-R1}}(q, r^M). \tag{4}$$

First, we show that $r^M \approx q$ for large values of λ_{R1}, so optimizing the reformulated regularizer is equivalent to optimizing a lower bound on the original.

Lemma 1. *For distributions $p \in \mathcal{P}$ and $q \in \mathcal{Q}$ where $\mathcal{P} \cap \mathcal{Q} \neq \emptyset$ and a continuous function $J(p, q)$, let $\tilde{J}(p, q; \lambda) = J(p, q) - \lambda D(p\|q)$, and $p^*_\lambda, q^*_\lambda \in \text{argmax}_{p\in\mathcal{P}, q\in\mathcal{Q}} \tilde{J}(p, q; \lambda)$. Then the following hold:*

$$\lim_{\lambda\to\infty} D(p^*_\lambda \| q^*_\lambda) = 0, \tag{5}$$

$$\lim_{\lambda\to\infty} \|p^*_\lambda - q^*_\lambda\|_\ell = 0 \quad \textit{for any } \ell\textit{, where } \|\cdot\|_\ell \textit{ is the } \ell\textit{-norm, and} \tag{6}$$

$$\lim_{\lambda\to\infty} \max_{p\in\mathcal{P}, q\in\mathcal{Q}} \tilde{J}(p, q; \lambda) \leq \max_{p\in\mathcal{P}} J(p, p). \tag{7}$$

Proof. Consider any $\epsilon > 0$ and any $p' \in \mathcal{P}, q' \in \mathcal{Q}$ such that $D(p' \| q') > \epsilon$. Let $\hat{p} \in \operatorname{argmax}_{p \in \mathcal{P} \cap \mathcal{Q}} J(p, p)$ and consider any $\lambda' \geq (1/\epsilon)(J(p', q') - J(\hat{p}, \hat{p}))$.

$$\begin{aligned} \tilde{J}(p', q'; \lambda') &= J(p', q') - \lambda' D(p' \| q') && (8)\\ &< J(p', q') - \lambda' \epsilon && (9)\\ &\leq J(p', q') - \cancel{\epsilon}\cancel{(1/\epsilon)}(J(p' \| q') - J(\hat{p} \| \hat{p})) && (10)\\ &= J(\hat{p}, \hat{p}) && (11) \end{aligned}$$

Therefore, $D(p^* \| q^*) \leq \epsilon$ when $\lambda \geq \lambda'$. This proves Proposition (5).

We have that

$$D(p \| q) \geq \frac{1}{2} \|p - q\|_1^2 \geq \frac{1}{2} \|p - q\|_\ell^2 \tag{12}$$

$$\tag{13}$$

for any ℓ-norm. The first inequality is Pinsker's inequality and the second follows from the relationship of ℓ-norms. Proposition (6) follows from this combined with Proposition (5).

Due to Proposition (6) and the continuity of $J(p, q)$,

$$\lim_{\lambda \to \infty} \max_{p \in \mathcal{P}, q \in \mathcal{Q}} \tilde{J}(p, q; \lambda) - \max_{p \in \mathcal{P} \cap \mathcal{Q}} J(p, p) = 0. \tag{14}$$

Proposition (7) follows from this and the fact that $\mathcal{P} \cap \mathcal{Q} \subseteq \mathcal{P}$. □

Therefore, for sufficiently large λ_{R1}, optimizing Equation (2) is equivalent to optimizing a lower bound on Equation (5) of the main text. This form allows us to compute q efficiently, which is shown as follows.

Theorem 2. *Define* $q^*(X_H) \triangleq \operatorname{argmax}_q J'_{\text{GBR-R1}}(\theta, q, r^M)$. *Then,*

$$q^*(x_H) = \frac{p_\theta(x_H, \bar{x}_O)^{1/(1+\lambda_{\mathrm{R1}})} \prod_{h \in H} r_h^M(x_h)^{\lambda_{\mathrm{R1}}/(1+\lambda_{\mathrm{R1}})}}{\sum_{x'_H} p_\theta(x'_H, \bar{x}_O)^{1/(1+\lambda_{\mathrm{R1}})} \prod_{h \in H} r_h^M(x'_h)^{\lambda_{\mathrm{R1}}/(1+\lambda_{\mathrm{R1}})}}. \tag{15}$$

Proof. For ease of notation, we group all terms that do not depend on q into one function $K_2(r)$. Since we must respect the sum-to-one property of q, we form the Lagrangian by adding the term $\lambda_2(1 - \sum_{x_H} q(x_H))$

$$L_2(q, \lambda_2) = -D(q(X_H) \| p_\theta(X_H | X_O)) - \lambda_{\mathrm{R1}} D(q(X_H) \| r^M(X_H)) - \lambda_2(1 - \sum_{x_H} q(x_H)) + K_2(r) \tag{16}$$

$$= \sum_{x_H} q(x_H) \log \frac{p_\theta(x_H | \bar{x}_O) r^M(x_H)^{\lambda_{\mathrm{R1}}}}{q(x_H)^{1+\lambda_{\mathrm{R1}}}} - \lambda_2(1 - \sum_{x_H} q(x_H)) + K_2(r) \tag{17}$$

$$= \sum_{x_H} q(x_H) \log \frac{p_\theta(x_H, \bar{x}_O) r^M(x_H)^{\lambda_{\mathrm{R1}}}}{q(x_H)^{1+\lambda_{\mathrm{R1}}}} - \log p_\theta(\bar{x}_O) - \lambda_2(1 - \sum_{x_H} q(x_H)) + K_2(r) \tag{18}$$

$$0 = \frac{\partial L}{\partial q(x_H)} = -\log p_\theta(x_H, \bar{x}_O) r^M(x_H)^{\lambda_{\mathrm{R1}}} + \log q(x_H)^{1+\lambda_{\mathrm{R1}}} + 1 + \lambda_{\mathrm{R1}} - \lambda_2 \tag{19}$$

$$\implies q(x_H) \propto p_\theta(x_H, \bar{x}_O)^{\frac{1}{1+\lambda_{\mathrm{R1}}}} r^M(x_H)^{\frac{\lambda_{\mathrm{R1}}}{1+\lambda_{\mathrm{R1}}}} \tag{20}$$

□

Critically, because r^M is factorizable such that each factor involves just one variable X_h, $q^*(X_H)$ obeys the same factorization properties as the unregularized model $p_\theta(X_H, \bar{x}_O)$. For example, if the original model was an HMM, q still factors as a chain. Therefore, the normalization constant can be computed using any algorithm for exact or approximate probabilistic inference on factorized models, such as belief propagation, with similar computational cost as the unregularized model.

Optimizing r^M

Despite the last reformulation, the objective still does not admit closed-form updates for r^M. Therefore, we again reformulate $\mathcal{PR}'_{\text{GBR-R1}}(\theta, q, r^M)$ by adding a new variable s^M, where s^M is also a distribution over X_H restricted to be factorizable as a product of marginals. As before, we add a term $\lambda_{\text{R2}} D(s^M(X_H) \| r^M(X_H))$, which encourages $s^M \approx r^M$. We define the graph regularizer KL divergence terms to have s^M on the left and r^M on the right—that is, in the form $D(s^M_u(X_u) \| r^M_v(X_v))$—which will enable efficient optimization for both variables.

$$\mathcal{PR}'_{\text{GBR-R2}}(q, r^M, s^M) \triangleq -\lambda_{\text{R1}} D(q(X_H) \| r^M(X_H)) + \max_{s^M} f_{\text{R2}}(r^M, s^M) \tag{21}$$

$$f_{\text{R2}}(r^M, s^M) \triangleq -\lambda_{\text{R2}} D(s^M(X_H) \| r^M(X_H)) - \lambda_G \sum_{(u,v) \in F_{GBR}} w(u,v) D(s^M_u(X_u) \| r^M_v(X_v)),$$

and $J'_{\text{GBR-R2}}(\theta, q, r^M, s^M)$ and $\mathcal{R}'_{\text{GBR-R2}}(\theta, q, r^M, s^M)$ are defined according to Equations (2) and (4) respectively using the corresponding regularizers. That is,

$$\text{maximize}_{\theta, q, r^M, s^M} \quad J'_{\text{GBR-R2}}(\theta, q, r^M, s^M) \triangleq \mathcal{L}(\theta) + \mathcal{R}'_{\text{GBR-R2}}(\theta, q, r^M, s^M), \tag{22}$$

$$\mathcal{R}'_{\text{GBR-R1}}(\theta, q, r^M, s^M) \triangleq -D(q(X_H) \| p_\theta(X_H | \bar{x}_O)) + \mathcal{PR}_{\text{GBR-R2}}(q, r^M, s^M). \tag{23}$$

By Lemma 1, optimizing $\mathcal{R}_{\text{GBR-R2}}(q)$ is equivalent to optimizing $\mathcal{R}_{\text{GBR-R1}}(q)$ for large values of λ_{R2}. This regularizer can be optimized in r^M and s^M using closed-form updates, shown as follows.

Theorem 3. *For notational simplicity, define a new regularization graph with self-edges of weight* $\lambda_{\text{R2}}/\lambda_G$*,* $E'_{\text{GBR}} \triangleq E_{\text{GBR}} \cup \{(h,h) \mid h \in H\}$*, and* $w'(u,v) \triangleq w(u,v) + \delta(u=v)\lambda_{\text{R2}}/\lambda_G$*. Let* $r^{M*}(X_H) \in \text{argmax}_{r^M} J'_{\text{GBR-R2}}(\theta, q, r^M, s^M)$ *and* $s^{M*}(X_H) \in \text{argmax}_{s^M} J'_{\text{GBR-R2}}(\theta, q, r^M, s^M)$*. Then,*

$$r^M_v(x_v) = \frac{\lambda_{\text{R1}} q^M_v(x_v) + \lambda_G \sum_{(u,v) \in E'_{\text{GBR}}} w'(u,v) s^M_u(x_v)}{\lambda_{\text{R1}} + \lambda_G \sum_{(u,v) \in E'_{\text{GBR}}} w'(u,v)}, \tag{24}$$

$$s^{M*}_u(x_u) = \frac{\exp \frac{\sum_{(u,v) \in E'_{\text{GBR}}} w'(u,v) \log r^M_v(x_u)}{\sum_{(u,v) \in E'_{\text{GBR}}} w'(u,v)}}{\sum_{x'_u} \exp \frac{\sum_{(u,v) \in E'_{\text{GBR}}} w'(u,v) \log r^M_v(x'_u)}{\sum_{(u,v) \in E'_{\text{GBR}}} w'(u,v)}}. \tag{25}$$

Proof. In its current form, $\mathcal{PR}_{\text{GBR-R2}}(q)$ involves a sum over all values of $q(X_H)$. However, the following lemma shows how the factorizability of r^M facilitates expressing the objective in a form that involves only sum over values of each variable X_h.

Lemma 4. *For distribution* $p(X_V)$ *and factorizable distribution* $q^M(X_V) = \prod_{v \in V} q^M_v(X_v)$*, define* $p^M_v(X_v) \triangleq \sum_{X_{V \setminus v}} p(X_V)$*.*

$$D(p \| q^M) = \sum_{v \in V} D(p(X_v) \| q(X_v)) - H(p) + \sum_{v \in V} H(q(X_v)). \tag{26}$$

Proof.

$$\begin{aligned}
D(p\|q^M) &= -H(p) - \sum_{x_V} p(x_V)\log\left(\prod_{v\in V} q_v^M(X_v)\right) && (27)\\
&= -H(p) - \sum_{x_V} p(x_V)\sum_{v\in V}\log q_v^M(X_v) && (28)\\
&= -H(p) - \sum_{v\in V}\sum_{x_v}\sum_{x_{V\neq v}} p(x_V)\log q_v^M(X_v) && (29)\\
&= -H(p) - \sum_{v\in V}\sum_{x_v}\left(\log q_v^M(X_v)\right)\sum_{x_{V\neq v}} p(x_V) && (30)\\
&= -H(p) - \sum_{v\in V}\sum_{x_v}\left(\log q_v^M(X_v)\right)p_v^M(x_v) && (31)\\
&= \sum_{v\in V} D(p(X_v)\|q(X_v)) - H(p) + \sum_{v\in V} H(q(X_v)) && (32)
\end{aligned}$$

□

Define q_h^M to be the marginal distribution of q over X_h, $q_h^M(X_h) \triangleq \sum_{X_{H\setminus h}} q(X_H)$. Using Lemma 4,

$$\mathcal{PR}_{\text{GBR-R2}}(q) = \max_{r^M}\left(-\lambda_{\text{R1}}\sum_{h\in H} D(q_h^M(X_h)\|r_h^M(X_h)) + H(q) - \sum_{h\in H} H(q_h^M(X_h)) + f_{\text{R2}}(r^M)\right). \tag{33}$$

We now proceed to derive the update steps. We first derive the update for r^M. The Lagrangian for the optimization of r_v^M is

$$\begin{aligned}
&L_{3\text{-}1}(r_v^M, \lambda_{3\text{-}1}) && (34)\\
&= \lambda_{\text{R1}} D(q_v^M(X_v)\|r_v^M(X_v)) + \lambda_G \sum_{(u,v)\in E'_{\text{GBR}}} w'(u,v) D(s_u^M(X_u)\|r_v^M(X_u)) && (35)\\
&\qquad + \lambda_{3\text{-}1}(1 - \sum_{x_v} r_v^M(X_v)) + K_{3\text{-}1}(q, s^M, r_{H\setminus v}^M) && (36)
\end{aligned}$$

$$0 = \frac{\partial L}{\partial r_v^M(x_v)} = -\left(\lambda_{\text{R1}} q_v^M(x_v) + \lambda_G \sum_{(u,v)\in E'_{\text{GBR}}} w'(u,v) s_u^M(x_v)\right)\frac{1}{r_v^M(x_v)} + \lambda_{3\text{-}1} \tag{37}$$

$$\implies r_v^M(x_v) \propto \lambda_{\text{R1}} q_v^M(x_v) + \lambda_G \sum_{(u,v)\in E'_{\text{GBR}}} w'(u,v) s_u^M(x_v) \tag{38}$$

$$\begin{aligned}
&\sum_{x_v}\left(\lambda_{\text{R1}} q_v^M(x_v) + \lambda_G \sum_{(u,v)\in E'_{\text{GBR}}} w'(u,v) s_u^M(x_v)\right) && (39)\\
&= \lambda_{\text{R1}}\cancelto{1}{\sum_{x_v} q_v^M(x_v)} + \lambda_G \sum_{(u,v)\in E'_{\text{GBR}}} w'(u,v)\cancelto{1}{\sum_{x_v} s_u^M(x_v)} && (40)
\end{aligned}$$

$$\implies r_v^M(x_v) = \frac{\lambda_{\text{R1}} q_v^M(x_v) + \lambda_G \sum_{(u,v)\in E'_{\text{GBR}}} w'(u,v) s_u^M(x_v)}{\lambda_{\text{R1}} + \lambda_G \sum_{(u,v)\in E'_{\text{GBR}}} w'(u,v)} \tag{41}$$

We next derive the update for s^M. The Lagrangian for the optimization of s_u^M is

$$L_{3\text{–}2}(s_u^M, \lambda_{3\text{–}2}) \tag{42}$$

$$= \lambda_G \sum_{(u,v)\in E'_{\text{GBR}}} w'(u,v) D(s_u^M(X_u) \| r_v^M(X_u)) + \lambda_{3\text{–}2}(1 - \sum_{x_u} s_u^M(X_u)) + K_{3\text{–}2}(q, r^M, s^M_{H\setminus u}) \tag{43}$$

$$= \lambda_G \sum_{x_u} s_u^M(x_u) \sum_{(u,v)\in E'_{\text{GBR}}} w'(u,v) \log \frac{s_u^M(x_u)}{r_v^M(x_u)} + \lambda_{3\text{–}2}(1 - \sum_{x_u} s_u^M(X_u)) + K_{3\text{–}2}(q, r^M, s^M_{H\setminus u}) \tag{44}$$

$$0 = \frac{\partial L}{\partial s_u^M(x_u)} = \lambda_G \sum_{(u,v)\in E'_{\text{GBR}}} w'(u,v) \log \frac{1}{r_v^M(x_u)} + \left(\lambda_G \sum_{(u,v)\in E'_{\text{GBR}}} w'(u,v)\right)(1 + \log s_u^M(x_u)) - \lambda_{3\text{–}2} \tag{45}$$

$$\implies s_u^M(x_u) \propto \exp \frac{\sum_{(u,v)\in E'_{\text{GBR}}} w'(u,v) \log r_v^M(x_u)}{\sum_{(u,v)\in E'_{\text{GBR}}} w'(u,v)} \tag{46}$$

□

Updating θ

The preceding section described an algorithm for computing $\text{argmax}_q\, \mathcal{R}_{\text{GBR-R2}}$. This algorithm can be combined with an EM-like algorithm in order to learn a θ that (locally) optimizes $J_{\text{GBR-R2}}$, as we describe in this section. We use an alternating EM-like algorithm to compute θ.

E-step: Compute $\left(q^{(t+1)}, r^{M(t+1)}, s^{M(t+1)}\right) \in \text{argmax}_{q,r^M,s^M}\, J'_{\text{GBR-R2}}(\theta^{(t)}, q, r^M, s^M)$

M-step: Compute $\theta^{(t+1)} \in \text{argmax}_\theta\, J'_{\text{GBR}}(\theta, q^{(t+1)})$

The preceding section showed how to perform the E-step. To compute the M-step,

$$\underset{\theta}{\text{argmax}}\, J'_{\text{GBR}}(\theta, q^{(t+1)}) = \underset{\theta}{\text{argmax}}\, E_{q^{(t+1)}(X_H)}\left[\log p_\theta(X_H, \bar{x}_O))\right] \tag{47}$$

The M-step takes the same form as the EM algorithm presented in (Neal and Hinton, 1999). The update for θ depends on the particular factorization and parameterization properties of the model. Because the posterior distribution $q(X_H)$ obeys the same factorization properties as the unregularized model $p_\theta(X_H, X_O)$, the same closed-form updates for θ can be used.

Therefore, $J_{\text{GBR-R2}}$ can be optimized using a three-way alternating maximization algorithm, which proceeds by alternating updates to r and s to convergence, alternating this whole update of r/s with updates to q until convergence, then finally alternating updates to q and θ until convergence. A schematic of the algorithm and objective appear in Supplementary Figure 8, and the algorithm is shown in full in Algorithm 1.

Theorem 5. *The modified EM algorithm monotonically increases the GBR objective:*

$$J_{\text{GBR-R2}}(\theta^{(t)}) \leq J_{\text{GBR-R2}}(\theta^{(t+1)}). \tag{48}$$

Proof. Function $q^*(\cdot)$ of Algorithm 1 implements coordinate descent on q, r^M and s^M. $D(p\|q)$ and $D(p\|p)$ are jointly strictly convex in p and q and bounded below by 0. Thus, $J'_{\text{GBR-R2}}$ is bounded

Algorithm 1 Efficient and scalable algorithm to optimize $J'_{\text{GBR-R2}}$

```
function ${r^M}^*(q)$
    for $h \in H$ do
        $q^M_h(x_h) \leftarrow \sum_{x_{H \neq h}} q(x_H)$                                    (belief propagation)
    end for
    Initialize ${r^M}^{(0)}$, ${s^M}^{(0)}$ arbitrarily.
    $t_1 \leftarrow 1$
    while not converged do
        for $v \in H$ do
            ${r^M}_v^{(t_1)}(x_v) \leftarrow \frac{\lambda_{\text{R1}} q^M_v(x_v) + \lambda_G \sum_{(u,v) \in E'_{\text{GBR}}} w'(u,v) {s^M}_u^{(t_1-1)}(x_v)}{\lambda_{\text{R1}} + \lambda_G \sum_{(u,v) \in E'_{\text{GBR}}} w'(u,v)}$
        end for
        for $u \in H$ do
            ${s^M}_u^{(t_1)}(x_u) \leftarrow \frac{\exp \frac{\sum_{(u,v) \in E'_{\text{GBR}}} w'(u,v) \log {r^M}_v^{(t_1-1)}(x_u)}{\sum_{(u,v) \in E'_{\text{GBR}}} w'(u,v)}}{\sum_{x'_u} \exp \frac{\sum_{(u,v) \in E'_{\text{GBR}}} w'(u,v) \log {r^M}_v^{(t_1-1)}(x'_u)}{\sum_{(u,v) \in E'_{\text{GBR}}} w'(u,v)}}$
        end for
        $t_1 \leftarrow t_1 + 1$
    end while
    return ${r^M}^{(t_1)}$
end function

function $q^*(\theta)$
    $t_2 \leftarrow 1$
    Initialize ${r^M}^{(0)}$ arbitrarily.
    while not converged do
        $q^{(t_2)}(x_H) \leftarrow \frac{p_\theta(x_H, \bar{x}_O)^{1/(1+\lambda_{\text{R1}})} \prod_{h \in H} {r^M}_h^{(t_2-1)}(x_h)^{\lambda_{\text{R1}}/(1+\lambda_{\text{R1}})}}{\sum_{x'_H} p_\theta(x'_H, \bar{x}_O)^{1/(1+\lambda_{\text{R1}})} \prod_{h \in H} {r^M}_h^{(t_2-1)}(x'_h)^{\lambda_{\text{R1}}/(1+\lambda_{\text{R1}})}}$    (belief propagation)
        ${r^M}^{(t_2)} \leftarrow {r^M}^*(q^{(t_2)})$
        $t_2 \leftarrow t_2 + 1$
    end while
end function

Initialize $\theta^{(0)}$ arbitarily.
$t_3 \leftarrow 1$
while not converged do
    $q^{(t_3)} \leftarrow q^*(\theta^{(t_3-1)})$
    $\theta^{(t_3)} \leftarrow \text{argmax}_\theta \, E_{q^{(t_3)}(X_H)} [\log p_\theta(X_H, \bar{x}_O))]$      (EM update)
end while
Output $\theta^{(t_3)}$
```

below and jointly strictly convex in q, r^M and s^M. Convergence to the global optimum of $J'_{\text{GBR-R2}}$ in q, r^M and s^M follows from its strict convexity (Warga, 1963).

$$\begin{aligned} J_{\text{GBR-R2}}(\theta^{(t)}) &= J'_{\text{GBR-R2}}(\theta^{(t)}, q^{(t+1)}, {r^M}^{(t+1)}, {s^M}^{(t+1)}) \\ &\leq J'_{\text{GBR-R2}}(\theta^{(t+1)}, q^{(t+1)}, {r^M}^{(t+1)}, {s^M}^{(t+1)}) \\ &\leq J_{\text{GBR-R2}}(\theta^{(t+1)}) \end{aligned}$$

The first equality follows from the global optimality of $q^{(t+1)}$, ${r^M}^{(t+1)}$ and ${s^M}^{(t+1)}$. The second inequality follows from the fact that $\theta^{(t+1)}$ is chosen to maximize $J'_{\text{GBR-R2}}$. The third inequality follows from the fact that $J'_{\text{GBR-R2}}(\theta, q, r^M, s^M)$ is a lower bound on $J_{\text{GBR-R2}}(\theta)$. □

Computation

Probabilistic inference for computing q and θ was performed on the DBN model using the graphical models toolkit (GMTK) (Bilmes, 2010). GMTK computations were distributed over a cluster using Grid Engine. Alternating minimization for updating r^M and s^M were performed using the Measure Propagation package (Subramanya and Bilmes, 2011).

2 Graph-based regularization outperforms an alternative approach based on approximate inference

To evaluate the efficacy of GBR, we compared GBR to two related methods: 1) approximate inference on a graphical model with the same dependence structure, and 2) GBR using squared-error penalties, as described in He et al. (2013). We compared to the approximate inference method loopy belief propagation (LBP) because it is one of the most widely used approximate inference methods. While we would have preferred to perform this comparison using real data sets, it appeared that even our fastest implementations of these methods would take months to converge. Therefore, we instead performed this comparison using synthetic data. We generated a chain of length $n = 200$, with $(X_H, X_O) = (Z_{1:200}, Y_{1:200})$, where $Z_{1:200} \in \{0, 1\}^n$ and $Y_{1:200} \in \mathbb{R}^n$. We defined an HMM over this chain with transition probabilities $\Pr(Z_i = Z_{i+1}) = 0.9$ and emission probabilities $Y_i \sim N(Z_i, \sigma)$, where we vary σ to control the difficulty of the problem—higher σ results in more challenging inference. We generated a graph $W \in \mathbb{R}^{n \times n}$ over the vertices of the chain by setting $w_{ij} = 1$ with probability 0.4 if $Z_i = Z_j$, $w_{ij} = 1$ with probability 0.1 if $Z_i \neq Z_j$, and $w_{ij} = 0$ otherwise. This model is meant to simulate the task of labeling a chain (such as a genomic sequence) where we have noisy information about which pairs of positions have the same label.

We compared five methods of inference: 1) inference on each position independently, with no chain model; 2) inference on the chain alone, without using W; 3) LBP on the chain plus extra factors of $\Pr(X_i = X_j) = \text{sigmoid}(\lambda w_{ij})$, where λ controls the strength of these factors; 4) a variant of GBR using the regularization graph W and a squared-error penalties as described in He et al. (2013); and 5) GBR using the regularization graph W. We chose hyperparameters for each model (λ_G, λ_{R1} and λ_{R2} for GBR and λ for LBP) using a training set of 200 simulations.

GBR significantly outperforms all other models for all experiments, providing nearly as much improvement in accuracy as the chain model does over the independent model (Supplementary Figure 9). The pattern of accuracy is instructive in understanding the properties of each model. LBP performs very well when there is little noise, but becomes easily stuck in local optima on harder problems. The variant of GBR with squared error provides a modest improvement over the chain model, but has poor performance relative to KL penalties, consistent with previous work on semi-supervised methods.

3 Segway model

We used graph-based regularization to augment the Segway semi-automated genome annotation method (Hoffman et al., 2012). Segway uses a dynamic Bayesian network model to perform genome annotation. The model is presented in detail in (Hoffman et al., 2012), but we describe it briefly here.

- We define a latent label variable $Y_i \in \{1..K\}$ for each position $i \in \{1..N\}$ in the genome, where K is the user-specified number of labels and N is the number of positions.

- We define observed signal data variables $X_{i,j}$ representing the value of signal data set $j \in \{1..M\}$ at genomic position i, where M is the number of signal data sets. We downsample the genome into bins of size R and average the signal data in each bin (after applying the inverse hyperbolic sine transform), so $N \approx 3 \times 10^9/R$. Because the sequencing depth of existing Hi-C data sets is too low to achieve single base pair resolution, we used $R = 10000$ for experiments using GBR to integrate Hi-C data. We used $R = 1$ for experiments using GBR to transfer information between cell types.

- The observed data variable $X_{i,j}$ depends only on the label at position i, Y_i. We model the variable $X_{i,j}$ as a Gaussian distribution with data set- and label-specific mean parameter $\mu_{i,j}$ and data set-specific variance parameter σ_j. In the case that some data values are missing due to mappability, we weight the observation of $X_{i,j}$ by the proportion of mappable positions $\mathring{X}_{i,j}$ in bin i in data set j.

- The label variable Y_i depends only on the label at the previous position Y_{i-1}. We model the label transition from label a to label b using a transition parameter $Q_{a,b}$.

- We model segment length, determined by self-transitions $Q_{a,a}$, separately from label transitions, determined by $Q_{a,b}$ for $a \neq b$. The self-transition is weighted by a hyperparameter $\lambda_{\text{transition}}$, which weighs the importance of segment length relative to signal data.

- The parameters $\mu_{1:K,1:M}$, $\sigma_{1:M}$ and $Q_{1:K,1:K}$ are learned through EM.

The overall log-likelihood of the Segway model is defined as:

$$\begin{aligned}
\log \Pr(X, Y \mid \mu, \sigma, Q) &= \sum_{i=1}^{N} \sum_{j=1}^{M} \mathring{X}_{i,j} \log N(X_{i,j} \mid \mu_{Y_i,j}, \sigma_j) \\
&+ \lambda_{\text{transition}} \sum_{i=1}^{N-1} \mathbf{1}(Y_i == Y_{i+1}) \log Q_{Y_i,Y_i} \\
&+ \sum_{i=1}^{N-1} \mathbf{1}(Y_i \neq Y_{i+1})(\lambda_{\text{transition}} \log(1 - Q_{Y_i,Y_i}) + \log Q_{Y_i,Y_{i+1}})
\end{aligned} \tag{49}$$

where $\mu_{\ell,j}$ is the mean associated with signal data set j and label ℓ; σ_j is the variance associated with signal data set j (shared between all labels); $\mathring{X}_{i,j}$ is the proportion of mappable positions in bin i for data set j; $Q_{a,b}$ is the transition probability parameter from label a to label b; and $\lambda_{\text{transition}}$ is a weight on the transitions relative to the emissions of the model.

4 Review of existing SAGA methods for using data from multiple cell types

Existing methods for semi-automated genome annotation work well on data from a single cell type, but annotating multiple cell types remains an active area of research. There are three simple strategies for performing annotation of multiple cell types. First, the simplest strategy is to apply the same model to both genomes (sometimes called "concatenated" annotation) (Sheffield et al., 2013), but this requires that all cell types have the same set of available data, which is not generally true. Moreover, in practice, experimental artifacts lead to poor performance for models which model multiple data from multiple experiments with the same parameters, exhibiting effects such as assigning separate sets of labels to each cell type in the model. Second, one could perform annotation separately on each cell type and find a mapping between the labels (for example, by using the Hungarian algorithm (Kuhn, 1955)). However, since different cell types generally have different types of activity and different sets of signal data sets, such a mapping is generally very poor. Third, one could use all data from all cell types in one model (sometimes called "stacked" annotation), but this strategy must either give the same label to each position for every cell type or use a separate label for each pattern of labels across cell types, which requires an exponentially-large number of labels.

Two additional methods have been proposed to annotate multiple cell types. The first, called hiHMM ("hierarchically-linked infinite HMM") maintains a separate model for each cell type and uses a regularization penalty to encourage the models to have similar parameters (Ho et al., 2014). This addresses the problem of requiring the same set of data across cell types, but does not share any position-specific information between cell types. The second method for performing multi-cell-line annotation, called TreeHMM, is given a tree over cell types and models the transition between labels between neighboring positions and also between neighboring developmental states (Biesinger et al., 2013). This model can integrate position-specific information between cell types, but requires that each cell type has the same available data and is sensitive to any cross-experiment artifacts. Moreover, the complexity of this model forced the authors to resort to approximate methods for inference, which likely decreases the quality of the resulting annotations.

These two problems—requiring a common set of data and failure to integrate evidence—are especially important because, although there are virtually limitless cell types and cell states that one would like to understand, very limited numbers of experiments have been performed in most of these cell types due to the cost of genomic experiments. For example, ENCODE has performed 335 experiments in its most-studied cell type, but has performed just 2-10 experiments in more than 100 cell types.

Transferring information with GBR removes the requirement for a common set of data across cell types and does integrate position-specific evidence across cell types. Therefore, GBR provides a method leveraging all available data in order to produce high-quality annotations of each cell type.

5 Related optimization methods

Clearly, the most straightforward way to express pairwise interactions in a graphical model is to encode them in the underlying graph and to use approximate methods (reviewed in (Wainwright and Jordan, 2008)) to enable inference. This form of interaction is quite general, in that when one adds a factor $\phi(y_i, y_j)$ between two random variables Y_i and Y_j, these random variables may have any type of interaction, expressed by $\phi(y_i, y_j)$. GBR, on the other hand, asks only for similarity between the marginals, meaning that $p(y_i|\cdot)$ and $p(y_j|\cdot)$ should be similar. Alternatively, a factor could encode such similarity, for example if $\phi(y_i, y_j) = \lambda \mathbf{1}(y_i = y_j)$. Such factors added to an HMM or CRF would result in a high treewidth model that can be dealt with using approximate inference. Doing so, however, loses any guarantee of optimality (which we preserve with GBR).

The posterior regularization framework of Ganchev et al. (Ganchev et al., 2010) takes an approach similar to ours, augmenting a simple model in a way that maintains tractable inference. This method adds a regularization term to an EM objective in order to require the posterior probabilities to satisfy logical constraints in expectation. Ganchev et al. show how to optimize this combined objective efficiently when the regularization term is linear in the posterior distribution of the model. Unfortunately, pairwise similarity relationships cannot be expressed with such a linear regularization term.

The most similar work to ours are the following three methods for graph regularization. First, Altun el al. (Altun et al., 2005) describe a graph regularization applied to a max-margin model applied to pitch-accent prediction and optical character recognition. However, this method involves a matrix inversion step, and thus cannot scale to large models. Second, Subramanya et al. (Subramanya et al., 2010) combine a temporal CRF with a regularizer that expresses pairwise squared-error penalties derived from unlabeled data. They apply this method to the part-of-speech tagging task (Subramanya et al., 2010) and later to related problems in natural language (Das and Petrov, 2011; Das and Smith, 2011). That work, however, resorts to a purely heuristic update step, and lacks any optimality guarantees. Third, He et al. (He et al., 2013) present an approach based on an exponentiated gradient descent algorithm. Like our approach, He's approach exhibits monotone convergence. Although He's work has many similarities with our approach, our methods were developed independently, and He's work differs from ours in three important ways. First, He et al. (He et al., 2013) use an exponentiated gradient descent strategy, while we use alternating minimization. Second, He's method uses a squared-error penalty, which is inappropriate for probability distributions, unlike our use of the Kullback-Leibler divergence (Bishop, 1995, p. 226). Third, the exponentiated gradient descent method is applied to handwriting recognition and part-of-speech tagging, while we apply GBR to genome annotation.

Supplementary Figures/Tables

Data type	URL
CAGE	http://www.gencodegenes.org/releases/7.html
ENCODE (ChIP-seq, DNase, Replication timing)	http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/
ChIP-seq (Roadmap)	http://www.roadmapepigenomics.org/data
Hi-C	http://yuelab.org/hi-c/download.html
Topological domains	http://www.cs.cmu.edu/∼ckingsf/software/armatus/

Supplementary Table 1: Data sources

	IMR90 domain annotation	Eight-cell type domain annotation	GM12878 reduced annotation
Input data sets	DNase H2aK5ac H2aK9ac H2a.Z H2bK120ac H2bK12ac H2bK15ac H2bK20ac H2bK5ac H3K14ac H3K18ac H3K23ac H3K27ac H3K27me3 H3K36me3 H3K4ac H3K4me1 H3K4me2 H3K4me3 H3K56ac H3K79me1 H3K79me2 H3K9ac H3K9me1 H3K9me3 H4K20me1 H4K5ac H4K8ac H4K91ac Repli-seq	DNase H2a.Z H3K27ac H3K27me3 H3K36me3 H3K4me1 H3K4me2 H3K4me3 H3K79me2 H3K9ac H3K9me3 H4K20me1	H3K4me1 H3K4me2 H3K3me3 H3K9ac H3K27ac H3K27me3 H3K36me3 H4K20me1
Number of labels	8	8	25
Transition weight ($\lambda_{\text{transition}}$)	48	12	1
Number of random EM initializations	10	10	10
GBR graph scale (λ_G)	1	1	1
GBR optimization hyperparameter (λ_{R1})	1	1	10

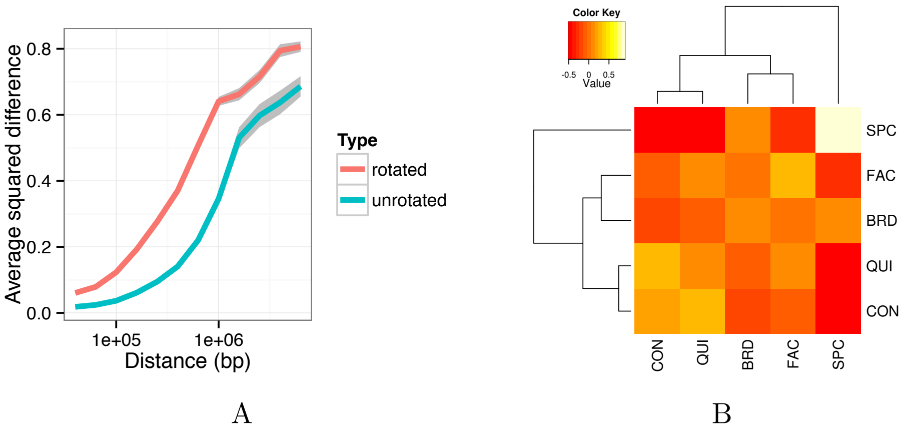
Supplementary Table 2: Parameters of all genome annotations

GOID	Bonferroni-corrected p-value	Name
GO:0008150	5.44809636039661e-49	biological process
GO:0070647	5.36690477736568e-14	protein modification by small protein conjugation or removal
GO:0016567	2.53980288482968e-13	protein ubiquitination
GO:0032446	3.48958938873894e-13	protein modification by small protein conjugation
GO:0071840	1.2816846277191e-07	cellular component organization or biogenesis
GO:0048522	3.15663523300463e-07	positive regulation of cellular process
GO:0016043	1.55013794212494e-06	cellular component organization
GO:0050953	2.07559208909527e-06	sensory perception of light stimulus
GO:0007601	2.79978686922173e-06	visual perception
GO:0006996	2.18396515274592e-05	organelle organization
GO:0016071	7.40708736771575e-05	mRNA metabolic process
GO:0048519	0.000123587368672724	negative regulation of biological process
GO:0023056	0.000191430490530164	positive regulation of signaling
GO:0048518	0.000239550248137195	positive regulation of biological process
GO:0032270	0.000250487526217915	positive regulation of cellular protein metabolic process
GO:0018146	0.000336577942863192	keratan sulfate biosynthetic process
GO:0007005	0.000402299808664309	mitochondrion organization
GO:0010647	0.000414643467676553	positive regulation of cell communication
GO:0032268	0.000530771233104358	regulation of cellular protein metabolic process
GO:0043928	0.000623789428174407	exonucleolytic nuclear-transcribed mRNA catabolic process involved in deadenylation-dependent decay
GO:0000288	0.000670978259119087	nuclear-transcribed mRNA catabolic process, deadenylation-dependent decay
GO:1903320	0.00117566239446111	regulation of protein modification by small protein conjugation or removal
GO:0009967	0.00131477493050438	positive regulation of signal transduction
GO:1902533	0.00136644459633516	positive regulation of intracellular signal transduction
GO:0048523	0.00148158842395948	negative regulation of cellular process
GO:0051340	0.00150521495507962	regulation of ligase activity
GO:0000291	0.00193533449146964	nuclear-transcribed mRNA catabolic process, exonucleolytic
GO:0031401	0.00194625332376889	positive regulation of protein modification process
GO:1903047	0.00215993757988771	mitotic cell cycle process
GO:0042531	0.00237481842965934	positive regulation of tyrosine phosphorylation of STAT protein
GO:0007267	0.0039286364798486	cell-cell signaling
GO:0006401	0.00399526165008678	RNA catabolic process
GO:0031396	0.00443702013802013	regulation of protein ubiquitination
GO:0000278	0.00470584872871359	mitotic cell cycle
GO:0051351	0.00495154943903681	positive regulation of ligase activity
GO:0051438	0.00524041402868326	regulation of ubiquitin-protein transferase activity
GO:0042339	0.0053487952592411	keratan sulfate metabolic process
GO:0051247	0.0056999635780996	positive regulation of protein metabolic process
GO:0016265	0.00629144608895318	death
GO:0046427	0.00671456368410498	positive regulation of JAK-STAT cascade
GO:0008219	0.0067422659494943	cell death
GO:0000956	0.00698342075165014	nuclear-transcribed mRNA catabolic process
GO:0031399	0.00786100228743956	regulation of protein modification process
GO:0044770	0.00795426575038111	cell cycle phase transition
GO:0051246	0.00937868333257509	regulation of protein metabolic process

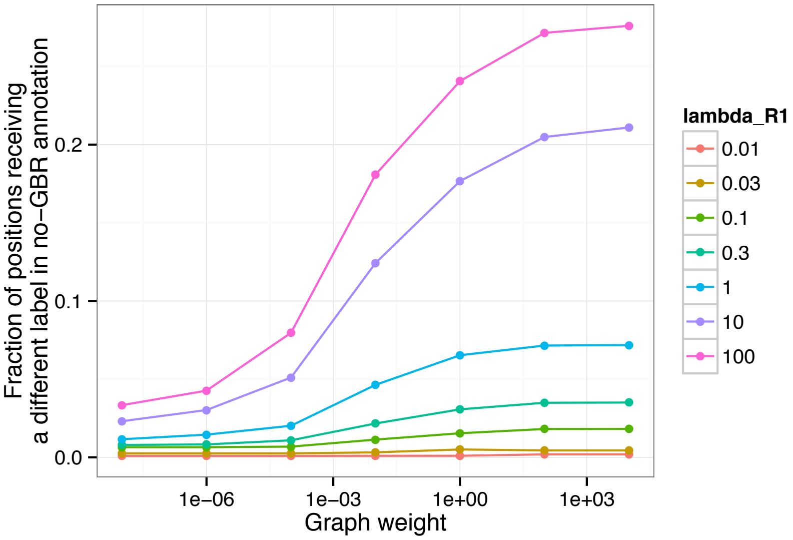
Supplementary Table 3: GO terms enriched for genes in BRD domains

GOID	Bonferroni-corrected p-value	Name
GO:0001944	2.9592753850599e-11	vasculature development
GO:0001568	1.34928597594565e-09	blood vessel development
GO:0072358	1.87785996344277e-09	cardiovascular system development
GO:0072359	1.94865728031881e-09	circulatory system development
GO:0048598	2.33761516882996e-09	embryonic morphogenesis
GO:0032501	4.29504418805107e-09	multicellular organismal process
GO:0044707	6.45108090271442e-09	single-multicellular organism process
GO:0007275	9.54873387980616e-09	multicellular organismal development
GO:0032502	9.87403900226897e-09	developmental process
GO:0009653	1.13900576119261e-08	anatomical structure morphogenesis
GO:0048646	1.27560082632202e-08	anatomical structure formation involved in morphogenesis
GO:0048856	3.44027824649974e-08	anatomical structure development
GO:0048514	3.64379056813118e-08	blood vessel morphogenesis
GO:0044767	4.96188838214423e-08	single-organism developmental process
GO:0008150	1.09920334986419e-07	biological process
GO:0009888	1.15062070196739e-07	tissue development
GO:0048731	1.34005756992352e-07	system development
GO:0030198	2.54744657740361e-07	extracellular matrix organization
GO:0043062	2.70008497648116e-07	extracellular structure organization
GO:0040011	4.09681657551741e-07	locomotion
GO:0048523	6.00924738320409e-07	negative regulation of cellular process
GO:0048513	1.27555684817328e-06	organ development
GO:0009887	1.67228271838845e-06	organ morphogenesis
GO:0060429	1.75248481849761e-06	epithelium development
GO:0009605	2.12388368291444e-06	response to external stimulus
GO:0048519	2.70812717394184e-06	negative regulation of biological process
GO:0048869	3.99639769208689e-06	cellular developmental process
GO:0009790	4.8142081410924e-06	embryo development
GO:0030154	8.17799427787657e-06	cell differentiation
GO:0048522	8.79037004023573e-06	positive regulation of cellular process
GO:0048870	9.70748540194382e-06	cell motility
GO:0051674	9.70748540194382e-06	localization of cell
GO:0048518	9.8762526690946e-06	positive regulation of biological process
GO:1902533	1.75047286259699e-05	positive regulation of intracellular signal transduction
GO:0071840	2.37299292646679e-05	cellular component organization or biogenesis
GO:0030334	2.38844411436045e-05	regulation of cell migration
GO:0007389	3.41057645381657e-05	pattern specification process
GO:0051270	4.15793772024666e-05	regulation of cellular component movement
GO:2000145	7.20700858370221e-05	regulation of cell motility
GO:0040012	7.2398136159996e-05	regulation of locomotion
GO:0016043	7.57778182419685e-05	cellular component organization
GO:0001501	9.16531463777006e-05	skeletal system development
GO:0008219	0.000103725100284498	cell death
GO:0001525	0.0001112117619674	angiogenesis
GO:0009966	0.000112414145595174	regulation of signal transduction
GO:0016265	0.000115956206339428	death
GO:0009967	0.000119265054523972	positive regulation of signal transduction
GO:0016477	0.000141195508435494	cell migration
GO:0048568	0.000148183903115669	embryonic organ development
GO:0001503	0.000148474000627697	ossification
GO:0006915	0.000174710834263385	apoptotic process
GO:0048729	0.000185804853094102	tissue morphogenesis
GO:0012501	0.000193299675551627	programmed cell death
GO:0010647	0.000264669391358318	positive regulation of cell communication
GO:0003007	0.000308615021011459	heart morphogenesis
GO:0010628	0.000312006014076936	positive regulation of gene expression
GO:0023056	0.000458831197316147	positive regulation of signaling
GO:0051239	0.000694243253157008	regulation of multicellular organismal process
GO:0031325	0.000823302329284296	positive regulation of cellular metabolic process
GO:0002009	0.000843967242091738	morphogenesis of an epithelium
GO:0048863	0.000980998073137653	stem cell differentiation
GO:0042325	0.00125849445147922	regulation of phosphorylation
GO:1902531	0.00144207328720399	regulation of intracellular signal transduction
GO:0044236	0.00151180971326667	multicellular organismal metabolic process
GO:0009893	0.00161958399151636	positive regulation of metabolic process
GO:0022603	0.00169126335423007	regulation of anatomical structure morphogenesis
GO:0035295	0.00223230684414793	tube development
GO:0006928	0.00230989986294592	cellular component movement
GO:0010604	0.00236857824547227	positive regulation of macromolecule metabolic process
GO:0010646	0.00252308440283661	regulation of cell communication
GO:0050793	0.0026847817189637	regulation of developmental process
GO:0048562	0.00285877724714378	embryonic organ morphogenesis
GO:0032879	0.00295955595643228	regulation of localization
GO:0080134	0.00335551717091664	regulation of response to stress
GO:0048705	0.00370439844279299	skeletal system morphogenesis
GO:0048864	0.00389984096973906	stem cell development
GO:0023051	0.0040158774598731	regulation of signaling
GO:0006334	0.00406223628626311	nucleosome assembly
GO:0045893	0.00414091776655817	positive regulation of transcription, DNA-templated
GO:0007167	0.00519580828357705	enzyme linked receptor protein signaling pathway
GO:0003002	0.00540223999981726	regionalization
GO:0051254	0.00549153180374605	positive regulation of RNA metabolic process
GO:1902680	0.00598393700399004	positive regulation of RNA biosynthetic process
GO:0023014	0.00676138544251521	signal transduction by phosphorylation
GO:0008283	0.0070310418956878	cell proliferation
GO:0048583	0.0072096008049822	regulation of response to stimulus
GO:0006333	0.00732272231577271	chromatin assembly or disassembly
GO:0007507	0.00736076227709712	heart development
GO:0042127	0.00771299785801664	regulation of cell proliferation
GO:2000026	0.008432344878453	regulation of multicellular organismal development
GO:0010941	0.00916678907171083	regulation of cell death
GO:0043408	0.00935020093242137	regulation of MAPK cascade
GO:0044259	0.00998773914798186	multicellular organismal macromolecule metabolic process

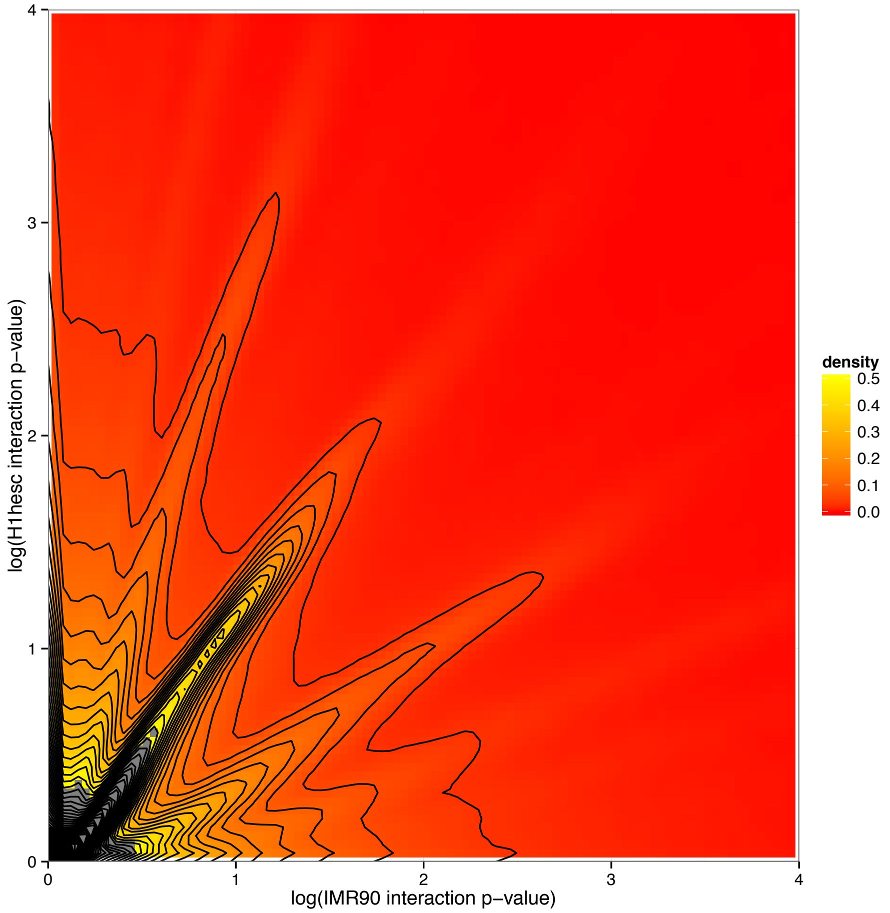
Supplementary Table 4: GO terms enriched for genes in SPC domains.



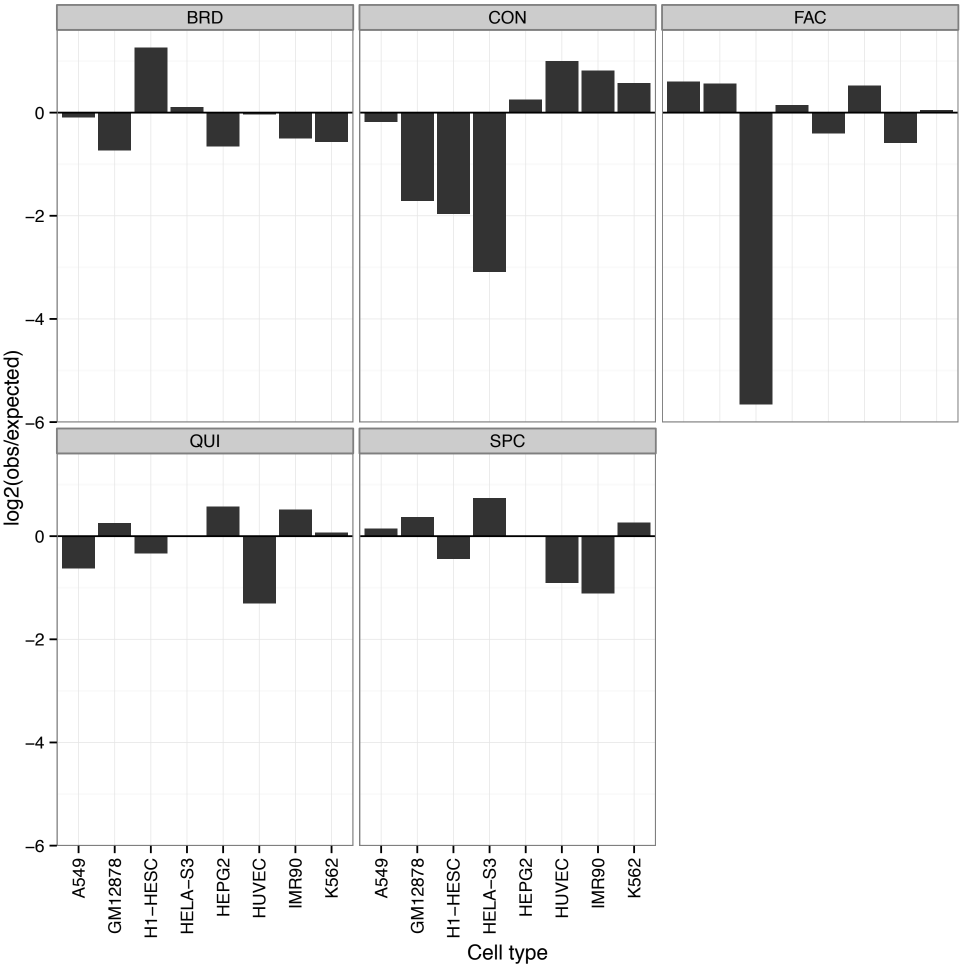
Supplementary Figure 1: (A) Average squared difference between replication timing values at left and right sides of significant contacts, as a function of genomic distance, relative to a permutation control (t-test 95% confidence interval grey error regions). (B) Confusion matrix of Segway annotation labels at left and right sides of significant contacts (without GBR). Color depicts $\log_2$(obs/expected) relative to a permutation control (Methods). Pairs of annotation labels at significantly interacting positions match more often than expected by chance (binomial test $p < 10^{-16}$).



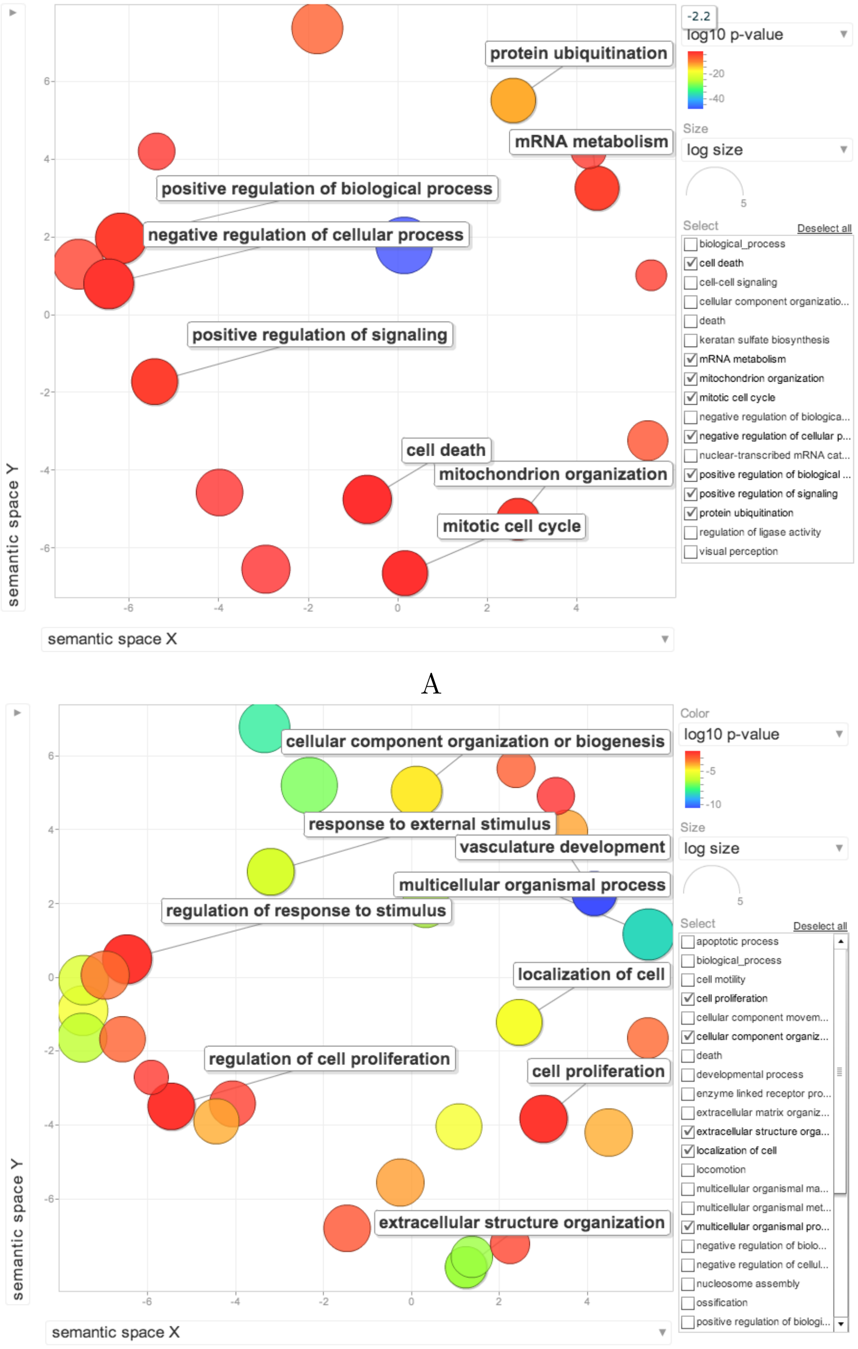
Supplementary Figure 2: Fraction of annotation changed by GBR. Y axis depicts the fraction of positions that receive a different label between an annotation without GBR and an annotation with GBR using a certain set of hyperparameters. X axis and color depict the λ_G and λ_{R1} hyperparameters respectively.



Supplementary Figure 3: Correlation of Hi-C contact strength between IMR90 and H1-hESC. X and Y axes are log *p*-values of association of a given pair of positions. Color indicates density of points. Black lines indicate density contours in 0.1% bins. Spearman $r = 0.57$.



Supplementary Figure 4: Distribution of domain labels across eight cell types. Y axis indicates log2(bases covered by label ℓ in celltype A / (bases covered by label ℓ / 8)).



log10 p-value ▾

Size

log size ▾

Select Deselect all

□ biological_process
☑ cell death
□ cell-cell signaling
□ cellular component organizatio...
□ death
□ keratan sulfate biosynthesis
☑ mRNA metabolism
☑ mitochondrion organization
☑ mitotic cell cycle
□ negative regulation of biologica...
☑ negative regulation of cellular p...
□ nuclear-transcribed mRNA cat...
☑ positive regulation of biological ...
☑ positive regulation of signaling
☑ protein ubiquitination
□ regulation of ligase activity
□ visual perception

A

cellular component organization or biogenesis
response to external stimulus
vasculature development
multicellular organismal process
regulation of response to stimulus
localization of cell
regulation of cell proliferation
cell proliferation
extracellular structure organization
semantic space Y
semantic space X
6
4
2
0
-2
-4
-6
-6
-4
-2
0
2
4
-5
-10
5

Color

log10 p-value ▾

Size

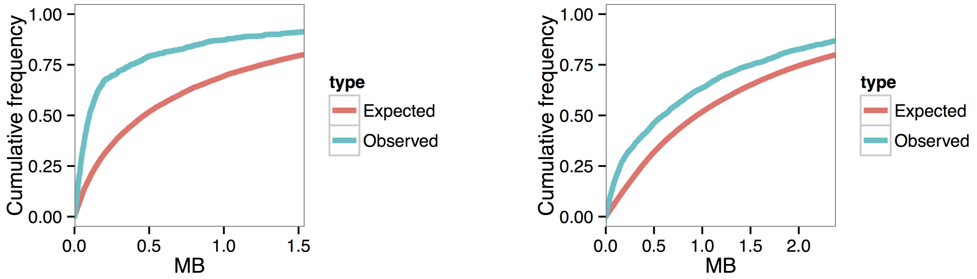
log size ▾

Select Deselect all

□ apoptotic process
□ biological_process
□ cell motility
☑ cell proliferation
□ cellular component movem...
☑ cellular component organiz...
□ death
□ developmental process
□ enzyme linked receptor pro...
□ extracellular matrix organiz...
☑ extracellular structure orga...
☑ localization of cell
□ locomotion
□ multicellular organismal ma...
□ multicellular organismal met...
☑ multicellular organismal pro...
□ negative regulation of biolo...
□ negative regulation of cellul...
□ nucleosome assembly
□ ossification
□ positive regulation of biologi...

B

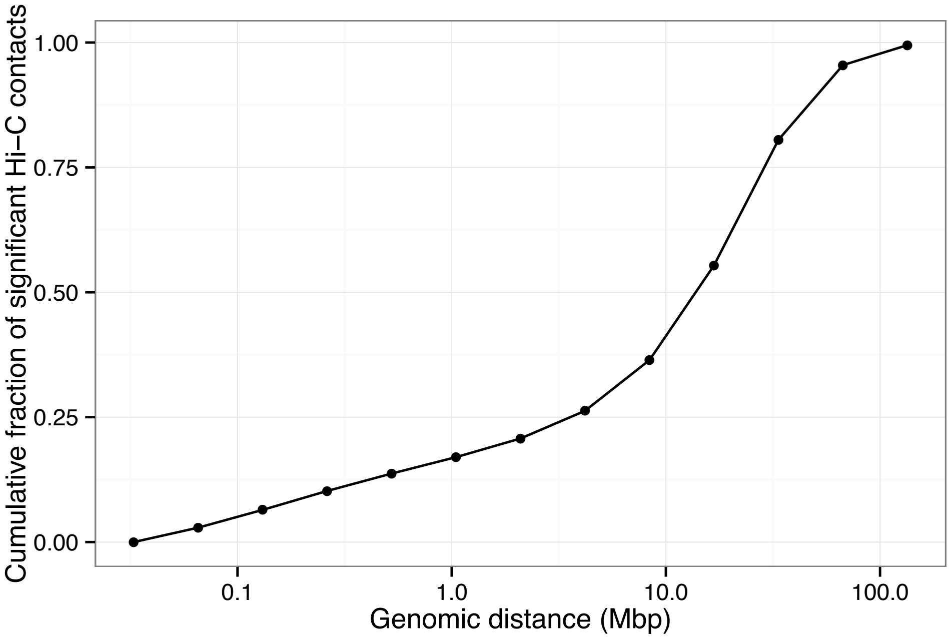
Supplementary Figure 5: Visualization of GO term enrichment for genes in IMR90 (A) BRD domains and (B) SPC domains using REVIGO (Supek et al., 2011). Each bubble represents a cluster of related enriched GO terms. X and Y axes are projected semantic axes defined using multidimensional scaling on the semantic similarity of each pair of terms.



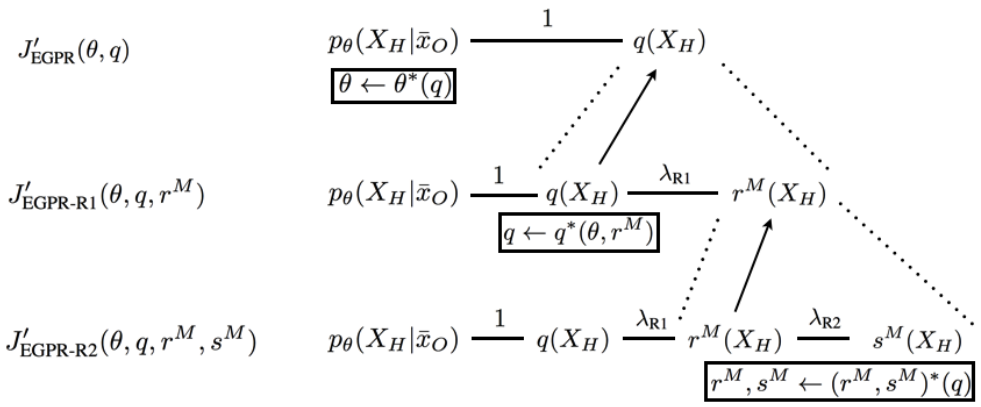
A

B

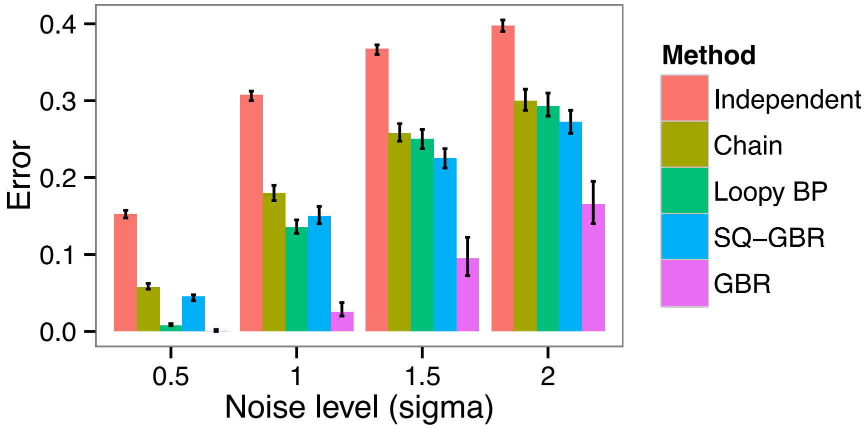
Supplementary Figure 6: Enrichment of consistent Segway boundaries for consistent replication domain boundaries. (A) Fraction of consistent replication domain boundaries overlapping consistent Segway domain boundaries as a function of the overlap distance. (B) Same as (A), but fraction of Segway domain boundaries. We used replication domain boundaries called by Pope et al. (2014). We defined replication boundaries occurring in more than 10 out of 18 cell types as consistent.



Supplementary Figure 7: Genomic distance distribution of significant IMR90 Hi-C contacts ($q < 0.05$).


$J'_{\text{EGPR}}(\theta, q)$
$p_\theta(X_H|\bar{x}_O)$
1
$q(X_H)$
$\theta \leftarrow \theta^*(q)$
$J'_{\text{EGPR-R1}}(\theta, q, r^M)$
$p_\theta(X_H|\bar{x}_O)$
1
$q(X_H)$
λ_{R1}
$r^M(X_H)$
$q \leftarrow q^*(\theta, r^M)$
$J'_{\text{EGPR-R2}}(\theta, q, r^M, s^M)$
$p_\theta(X_H|\bar{x}_O)$
1
$q(X_H)$
λ_{R1}
$r^M(X_H)$
λ_{R2}
$s^M(X_H)$
$r^M, s^M \leftarrow (r^M, s^M)^*(q)$

Supplementary Figure 8: Schematic of the three formulations of the objective and the alternating maximization strategy. Edges in this figure indicate KL terms, labeled according to their weight in the objective. Boxed formulae are update steps. We perform two reformulations, first splitting q into q and r^M linked by a KL term of weight λ_{R1}, then splitting r^M into r^M and s^M, linked by a KL term of weight λ_{R2}.



Supplementary Figure 9: Comparison between inference methods on synthetic data. The X axis shows σ, a hyperparameter controlling the difficulty of inference. The Y axis shows the average accuracy over 200 simulations of MAP inference on the model in question (95% Wilcoxon test confidence intervals). Bars correspond to five different inference methods: 1) inference on each position independently, with no chain model (Independent); 2) inference on the chain alone, without using W (Chain); 3) loopy belief propagation on the chain plus extra factors of $\Pr(X_i = X_j) = \text{sigmoid}(\lambda w_{ij})$, where λ controls the strength of these factors; (Loopy BP) 4) a variant of GBR using the regularization graph W and a squared-error penalties as described in (He et al., 2013) (SQ-GBR); and 5) GBR using the regularization graph W (our method, GBR).

References

Altun Y, Belkin M, and Mcallester DA. 2005. Maximum margin semi-supervised learning for structured variables. In *Advances in Neural Information Processing Systems*, pp. 33–40.

Biesinger J, Wang Y, and Xie X. 2013. Discovering and mapping chromatin states using a tree hidden Markov model. *Bioinformatics* **14**: S4.

Bilmes J. 2010. Dynamic graphical models. *IEEE Signal Processing Magazine* **27**: 29–42.

Bishop C. 1995. *Neural Networks for Pattern Recognition.* Oxford UP, Oxford, UK.

Das D and Petrov S. 2011. Unsupervised part-of-speech tagging with bilingual graph-based projections. In *NAACL*, pp. 600–609.

Das D and Smith N. 2011. Semi-supervised framesemantic parsing for unknown predicates. In *Association for Computational Linguistics.*

Ganchev K, Graça J, Gillenwater J, and Taskar B. 2010. Posterior regularization for structured latent variable models. *Journal of Machine Learning Research* **11**: 2001–2049.

He L, Gillenwater J, and Taskar B. 2013. Graph-based posterior regularization for semi-supervised structured prediction. In *Seventeenth Conference on Computational Natural Language Learning.*

Ho JW, Jung YL, Liu T, Alver BH, Lee S, Ikegami K, Sohn KA, Minoda A, Tolstorukov MY, Appert A, et al.. 2014. Comparative analysis of metazoan chromatin organization. *Nature* **512**: 449–452.

Hoffman MM, Buske OJ, Wang J, Weng Z, Bilmes JA, and Noble WS. 2012. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nature Methods* **9**: 473–476.

Kuhn HW. 1955. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly* **2**: 83–97.

Neal R and Hinton G. 1999. A view of the EM algorithm that justifies incremental, sparse, and other variants. In *Learning in graphical models*, pp. 355–368. MIT Press.

Pope BD, Ryba T, Dileep V, Yue F, Wu W, Denas O, Vera DL, Wang Y, Hansen RS, Canfield TK, et al.. 2014. Topologically associating domains are stable units of replication-timing regulation. *Nature* **515**: 402–405.

Sheffield NC, Thurman RE, Song L, Safi A, Stamatoyannopoulos JA, Lenhard B, Crawford GE, and Furey TS. 2013. Patterns of regulatory activity across diverse human cell types predict tissue identity, transcription factor binding, and long-range interactions. *Genome Research* **23**: 777–788.

Subramanya A and Bilmes J. 2011. Semi-supervised learning with measure propagation. *Journal of Machine Learning Research* **12**: 3311–3370.

Subramanya A, Petrov S, and Pereira F. 2010. Efficient graph-based semi-supervised learning of structured tagging models. In *Proc. of EMLNP 2010*, pp. 167–176. Association for Computational Linguistics.

Supek F, Bošnjak M, Škunca N, and Šmuc T. 2011. REVIGO summarizes and visualizes long lists of gene ontology terms. *PloS ONE* **6**: e21800.

Wainwright M and Jordan M. 2008. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning* **1**: 1–305.

Warga J. 1963. Minimizing certain convex functions. *Journal of the Society for Industrial and Applied Mathematics* **11**: 588–593.