High Resolution Modeling of Chromatin Interactions

Christopher Reeder and David Gifford

Massachusetts Institute of Technology
{reeder,gifford}@mit.edu

Abstract. SPROUT is a novel generative model for ChIA-PET data that characterizes physical chromatin interactions and points of contact at high spatial resolution. SPROUT improves upon other methods by learning empirical distributions for pairs of reads that reflect ligation events between genomic locations that are bound by a protein of interest. Using these learned empirical distributions Sprout is able to accurately position interaction anchors, infer whether read pairs were created by self-ligation or inter-ligation, and accurately assign read pairs to anchors which allows for the identification of high confidence interactions. When SPROUT is run on CTCF ChIA-PET data it identifies more interaction anchors that are supported by CTCF motif matches than other approaches with competitive positional accuracy. SPROUT rejects interaction events that are not supported by pairs of reads that fit the empirical model for inter-ligation read pairs, producing a set of interactions that are more consistent across CTCF biological replicates than established methods.

Keywords: Chromatin Interactions, ChIA-PET, CTCF.

1 Introduction

Chromatin interactions are a key component of gene regulation as looping induced interactions bring distal genomic regulatory sequences spatially proximal to their regulatory targets [8]. Identifying the connections between regulatory elements and the genes they regulate is required for understanding transcriptional regulation. Thus, the precise characterization of looping based interactions would help refine our understanding of how genes are controlled. Other forms of looping can implement other kinds of transcriptional regulation, such as isolating regions of the genome from transcriptional activity [4, 13–15, 17, 20].

Recently developed molecular approaches [19] identify chromatin interactions by producing single DNA molecules that combine pieces of DNA from both ends of an interaction event under appropriate ligation conditions. The base sequences at the ends of these DNA molecules are evidence in support of chromatin interactions at the genomic coordinates where the observed sequences originated. ChIA-PET is one such approach that measures chromatin interactions between genomic sites bound by a particular protein [5]. In ChIA-PET the dilute ligation step is preceded by fixation by formaldehyde, fragmentation by sonication, and

M. Deng et al. (Eds.): RECOMB 2013, LNBI 7821, pp. 186–198, 2013.

chromatin immunoprecipitation using an antibody designed to target the protein of interest. By using an antibody against a protein that is known to play a role in maintaining genome structure [16], subsequent analysis can focus specifically on chromatin contacts that involve that protein. However, ChIA-PET experimental data are polluted by pairs of reads whose ends do not correspond to binding events for the protein of interest. Such pairs of reads are much like the background reads observed in ChIP-Seq data [18]. This combined with the noisy positioning of reads around binding events presents two challenges for accurately analyzing ChIA-PET data. The first is to accurately identify the positions of the binding events that serve as potential anchors for interactions. The second is to accurately assign read pairs to chromatin interaction anchors or to a background noise model. Focusing on the set of chromatin interactions that are mediated by a specific regulatory protein or complex permits sequencing resources to be focused on the corresponding events. However, sophisticated computational methods are still required to accurately discover interactions from ChIA-PET data.

SPROUT is a novel computational approach for analyzing ChIA-PET data that integrates chromatin interaction discovery with the identification of interaction anchor points. SPROUT accomplishes this by modeling the empirical distribution of read positions around interaction anchors, allowing it to determine the positions of anchors and assign pairs of reads to anchors accurately. Previous approaches to analyzing ChIA-PET data [10] have separated anchor and interaction discovery eliminating the statistical strength that is gained from combining the two procedures. We note that SPROUT is theoretically applicable to datasets generated using related technologies such as Hi-C [11] when sufficient read coverage is available.

In the remainder of the paper we introduce the SPROUT model, discuss our results on CTCF ChIA-PET data, and conclude with observations about SPROUT's applicability.

2 Methods

SPROUT is a hierarchical generative model for ChIA-PET data that discovers interaction anchors, and a set of binary interactions between anchors. There are two types of pairs of reads that are present in ChIA-PET data. Self-ligation pairs arise from the ligation of a DNA molecule to itself. These pairs do not provide direct information about interactions between anchors and can be thought of as providing the same information as paired-end ChIP-Seq data. Inter-ligation pairs arise from the ligation of two distinct DNA molecules from the same chromosome or different chromosomes and thus provide information about a potential interaction.

SPROUT models read-pair data with a mixture over distributions describing the generation of self-ligation pairs and inter-ligation pairs. The components of the model describing these two types of read pairs are themselves mixtures of distributions corresponding to the way pairs of reads are expected to be distributed around anchors. We assume that the paired-end sequencing data generated by

a ChIA-PET experiment have been processed appropriately resulting in a set $\mathbf{R} = \{r_1, \ldots, r_N\}$ such that each $r_i = \langle r_i^{(1)}, r_i^{(2)} \rangle$ is a pair of genomic coordinates corresponding to the aligned positions of a pair of reads. Such processing includes removing linker tags from the reads, filtering out pairs that are identified as chimeric because of their heterogeneous linker tags, and aligning the reads to the genome. The following is the likelihood of $\mathbf{R}$

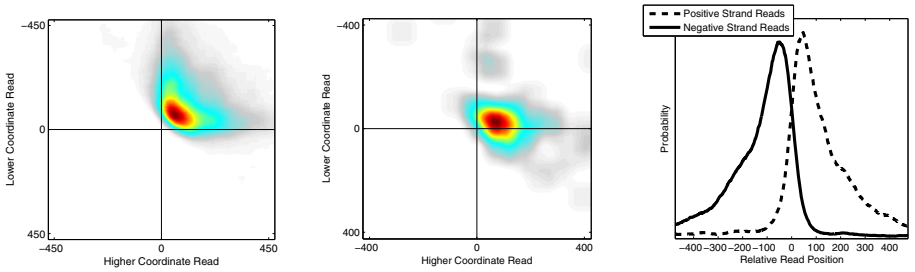
$$\Pr(\mathbf{R}, \pi, \psi, \rho, l) = \prod_{i=1}^{N} \left[\rho \left[\sum_{j=1}^{M} \pi_j \Pr(r_i|l_j) \right] + (1-\rho) \left[\sum_{j=1}^{M} \sum_{k=1}^{M} \psi_{j,k} \Pr(r_i|l_j, l_k) \right] \right] \quad (1)$$

Where $0 \leq \rho \leq 1$, $\sum_{i=1}^{N} \pi_i = 1$, $\sum_{i=1}^{N} \sum_{j=1}^{N} \psi_{i,j} = 1$

SPROUT identifies a set $l = \{l_1, \ldots, l_M\}$ that specifies the locations of sites that are bound by the protein of interest and are potential anchors for interactions. ρ is the probability that a pair of reads was generated by self-ligation. Self-ligation pairs reflect the ligation of a DNA fragment to itself to form a circular fragment. Such pairs are associated with one anchor and the self-ligation component of the model is a mixture of distributions each taking a single parameter to specify the location of the anchor position. These distributions take the form $\Pr(r_i|l_j)$ (Fig. 1a). A relative weight π_j is associated with each anchor j. These distributions describe the length and arrangement of fragments around an anchor which are induced by the fragmentation step of the ChIA-PET protocol.

Inter-ligation pairs can be associated with either the same anchor or two different anchors that were in close proximity in the nucleus. The inter-ligation component of the model is a mixture of distributions each taking two parameters that specify the locations of the anchor(s) that the fragments were associated with. A relative weight $\psi_{j,k}$ is associated with each pair of anchors j and k. The distributions $\Pr(r_i|l_j, l_k)$ take different forms because if $j = k$ (Fig. 1b) then there are constraints on the ends of the fragments involved in the ligation. For example, the fragments cannot have been overlapping since they were part of the same chromosome prior to fragmentation. If $j \neq k$ (Fig. 1c) it is assumed that the ends were generated independently by two one-dimensional distributions centered around the two anchors $\Pr(r_i|l_j, l_k) = \Pr(r_i^{(1)}|l_j) \Pr(r_i^{(2)}|l_k)$. We also assume that r_i implicitly carries information about the strandedness of the reads because in both the case where $j = k$ and $j \neq k$ the distributions depend on the strandedness of the reads.

ChIA-PET data are noisy, and we observe reads that do not correspond to anchors. To account for these reads, we introduce a noise component with dummy variable l_B ($B \notin \{1, \ldots, M\}$). In this work we consider uniform $\Pr(r_i|l_B)$, however knowledge about the propensity for genomic regions to generate background noise could be incorporated into a more refined noise distribution. We assume that $\Pr(r_i|l_j, l_k)$ where $j = B$ or $k = B$ is defined in the same way as the case in which j and k specify two different anchors: $\Pr(r_i|l_j, l_k) = \Pr(r_i^{(1)}|l_j) \Pr(r_i^{(2)}|l_k)$ and $\Pr(r_i^{(\cdot)}|l_j)$ is uniform when $j = B$.



(a) (b) (c)

Fig. 1. These are examples of read distributions learned from CTCF ChIA-PET data. SPROUT is initially run with "generic" distributions and then the distributions are re-estimated using the strongest events and SPROUT is re-run with the empirically learned distributions to discover more accurate predictions. (a) The positions of the ends of self-ligation pairs are modeled using a two dimensional distribution. (b) The positions of the ends of inter-ligation pairs where both ends are assigned to the same anchor are also modeled using two dimensional distributions. Each of the four possible strand combinations has its own constraints in terms of where the ends are likely to be positioned relative to each other and to the anchor. This figure demonstrates the distribution associated with inter-ligation pairs where both ends map to the positive strand. (c) The positions of the ends of inter-ligation pairs are modeled separately using one dimensional distributions.

To avoid overfitting, we wish to find a minimal number of anchors that explain the data well while allowing the noise distribution to account for reads that are not accounted for by anchors. Additionally, we assume that among all possible pairs of anchors most pairs are not interacting. Thus, we wish to find a minimal number of interacting pairs of anchors that explain the observed data. To achieve both of these types of sparsity we introduce negative Dirichlet priors [3] on π and ψ as specified by Eq. 2 and Eq. 3.

$$\Pr(\pi|\alpha) \propto \prod_{j=1}^{M} \pi_j^{-\alpha} \tag{2}$$

$$\Pr(\psi|\beta) \propto \prod_{j=1}^{M} \prod_{k=1}^{M} \psi_{j,k}^{-\beta} \tag{3}$$

As will become apparent when the inference procedure is described, the α and β parameters have the effect of specifying the minimum number of pairs of reads that must be associated with an anchor or an interaction, respectively, in order to avoid being eliminated from the model.

We also introduce priors on l and ρ. For l we introduce a Bernoulli prior which reflects our prior belief that an anchor exists at a particular genomic coordinate and that at most one anchor exists at any genomic coordinate. Given L possible genomic coordinates,

$$\Pr(l|k) = \prod_{i=1}^{L} k_i^{\mathbf{1}(i\in l)}(1-k_i)^{\mathbf{1}(i\notin l)} \tag{4}$$

$$= \prod_{i=1}^{L}(1-k_i)\prod_{j=1}^{M}\frac{k_{l_j}}{1-k_{l_j}} \tag{5}$$

$$\propto \prod_{j=1}^{M}\frac{k_{l_j}}{1-k_{l_j}} \tag{6}$$

In this work we consider uniform k, but k could be made non-uniform to reflect any prior belief about where anchors should be located. For ρ we introduce a Beta prior

$$\Pr(\rho|a,b) \propto \rho^{a-1}(1-\rho)^{b-1} \tag{7}$$

In this work we let $a = 1$ and $b = 1$ which is a uniform prior on ρ.

Each pair of reads is either a result of a self-ligation event or an inter-ligation event and is associated with one or two anchors. We introduce latent variables $\mathbf{Z} = \{z_1, \ldots, z_N\}$ such that each $z_i = \langle z_i^{(1)}, z_i^{(2)}\rangle$ is a pair of anchor indices $1 \ldots M$ or special index B reflecting the noise distribution. Another special index is used to indicate that a pair of reads was generated by self-ligation i.e. $z_i = \langle j, -\rangle$.

The complete data likelihood is

$$\Pr(\mathbf{R},\mathbf{Z}|\pi,\psi,\rho,l) = \Pr(\mathbf{R}|\mathbf{Z},l)\Pr(\mathbf{Z}|\pi,\psi,\rho) \tag{8}$$

$$= \prod_{i=1}^{N}\left[\prod_{j=1}^{M}[\rho\pi_j \Pr(r_i|l_j)]^{\mathbf{1}(z_i=\langle j,-\rangle)}\prod_{k=1}^{M}[(1-\rho)\psi_{j,k}\Pr(r_i|l_j,l_k)]^{\mathbf{1}(z_i=\langle j,k\rangle)}\right] \tag{9}$$

We are interested in inferring likely values for π, ψ, ρ, and l. To accomplish this we employ a variant of the EM algorithm [2] to maximize the complete data log posterior

$$\log\Pr(l,\pi,\psi,\rho|\mathbf{R},\mathbf{Z},k,\alpha,\beta,a,b) = \sum_{i=1}^{N}\left[\sum_{j=1}^{M}\left[\mathbf{1}(z_i=\langle j,-\rangle)\left(\log\rho + \log\pi_j + \log\Pr(r_i|l_j)\right)\right.\right.$$
$$\left.\left. + \sum_{k=1}^{M}\mathbf{1}(z_i=\langle j,k\rangle)\left(\log(1-\rho) + \log\psi_{j,k} + \log\Pr(r_i|l_j,l_k)\right)\right]\right] \tag{10}$$
$$-\alpha\sum_{j=1}^{M}\log\pi_j - \beta\sum_{j=1}^{M}\sum_{k=1}^{M}\log\psi_{j,k} + \sum_{j=1}^{M}\log\frac{k_{l_j}}{1-k_{l_j}} + (a-1)\log\rho + (b-1)(1-\rho) + C$$

E Step:

$$\gamma(z_i) = \frac{\prod_{j=1}^{M}\left[\rho\pi_j\Pr(r_i|l_j)\right]^{\mathbf{1}(z_i=\langle j,-\rangle)}\prod_{k=1}^{M}\left[(1-\rho)\psi_{j,k}\Pr(r_i|l_j,l_k)\right]^{\mathbf{1}(z_i=\langle j,k\rangle)}\Big]}{\sum_{j=1}^{M}\Big[\rho\pi_j\Pr(r_i|l_j)] + \sum_{k=1}^{M}\left[(1-\rho)\psi_{j,k}\Pr(r_i|l_j,l_k)\right]\Big]} \tag{11}$$

M Step:

$$\hat{l}_j = \underset{x}{\operatorname{argmax}} \left\{ \sum_{i=1}^{N} \left[\gamma(z_i = \langle j, - \rangle) \log \Pr(r_i|x) + \sum_{k=1}^{M} \left[\gamma(z_i = \langle j, k \rangle) \log \Pr(r_i|x, l_k) \right] \right] + \log \frac{k_x}{1 - k_x} \right\} \tag{12}$$

$$\hat{\pi}_j = \frac{\max(N_j - \alpha, 0)}{N_\pi} \tag{13}$$

$$N_\pi = \sum_{j=1}^{M} \max(N_j - \alpha, 0) \tag{14}$$

$$N_j = \sum_{i=1}^{N} \gamma(z_i = \langle j, - \rangle) \tag{15}$$

$$\hat{\psi}_{j,k} = \frac{\max(N_{j,k} - \beta, 0)}{N_\psi} \tag{16}$$

$$N_\psi = \sum_{j=1}^{M} \sum_{k=1}^{M} \max(N_{j,k} - \beta, 0) \tag{17}$$

$$N_{j,k} = \sum_{i=1}^{N} \gamma(z_i = \langle j, k \rangle) \tag{18}$$

$$\hat{\rho} = \frac{N_\pi + a}{N + a + b} \tag{19}$$

The E and M steps are repeated until the posterior approximately converges. The components of l that correspond to non-zero components of π are the estimated anchor locations. Non-zero components of ψ indicate pairs of anchors that are candidates for significance testing as interactions.

The algorithm is initialized with uniform π and l set at regular intervals throughout the genome. Components of π that do not assign probability to any pairs of reads are set to 0 and effectively eliminated from the model. Components with $N_j < \alpha$ are eliminated shortly thereafter. In the estimation of $\hat{l}_j$ during each M step the components of l other than the jth component are held fixed making this algorithm an instance of the expectation-conditional maximization algorithm [12]. Thus, the posterior is not necessarily maximized at each iteration but convergence to a local maximum is still guaranteed. The estimation of $\hat{l}_j$ is tractable, despite the lack of a closed form solution, because for the set of pairs of reads such that $\gamma(z_i = \langle j, \cdot \rangle) > 0$, $Pr(r_i|x) > 0$ for any pair of reads in the set for x in only a small neighborhood around the previous value of l_i. Only x in that neighborhood need be considered which reduces the search space for the optimal x considerably.

To test the significance of a component $\psi_{j,k}$, the posterior is recomputed with that component removed. The greater the ratio of the posterior with the

component to the posterior without the component, the greater the significance of the corresponding interaction. Making the conservative assumption that all components with $N_{j,k} \leq 2$ are false positives, we set a threshold for the posterior ratio to be the value such that 5% of the components deemed significant have $N_{j,k} \leq 2$.

3 Results

We compared the performance of SPROUT to other methods by analyzing CTCF ChIA-PET data published by Handoko et al. [7]. Reads were processed using the LinkerRemover component of the ChIA-PET tool [10]. The pairs of reads that were positively identified as chimeric were discarded and the rest of the reads were aligned to the mouse genome as unpaired reads using BOWTIE [9]. Only pairs with both ends that map uniquely were considered for further analysis. In cases where more than one pair of reads aligns to the same location at both ends, only one pair is retained because such positional duplicates are likely to be PCR artifacts.

For comparison, we downloaded the significant intra-chromosomal interactions and CTCF binding events published by Handoko et al. SPROUT discovers both inter- and intra-chromosomal interactions, but for this analysis we limited our comparison to intra-chromosomal interactions only. SPROUT does not impose a lower bound on the distance between pairs of anchors that it will consider for identifying interactions. However, linearly proximal anchors are expected to be spatially proximal due to random polymeric movement of the chromosome in the space of the nucleus. Therefore, linearly proximal anchors are expected to be called interacting by SPROUT. To investigate the distance at which this effect diminishes, we looked at the frequency at which pairs of anchors detected by SPROUT interact as a function of distance between the anchors (Fig. 2). By 4000 bp, detected interactions become very infrequent suggesting that interactions of this distance or greater are unlikely to be due to the linear proximity effect. The shortest range interaction published by Handoko et al. is 5928 bp, so for comparison we only consider interactions discovered by SPROUT that span at least this distance. But, we note that Fig. 2 suggests that functional interactions may be discoverable by SPROUT at distances as low as 4000 bp.

By comparing the positions of the anchors discovered by SPROUT to matches to the CTCF motif, we discovered that SPROUT positions anchors with high accuracy, and is very sensitive compared to other methods of discovering CTCF binding events while maintaining a high degree of specificity. For comparison, we examined the CTCF binding event calls published by Handoko et al. as well as binding events identified by the GEM peak calling algorithm [6] which was run on an independent ChIP-Seq dataset [1]. It is worth noting that Handoko et al. based their binding event predictions on the ChIA-PET data but that their method for identifying interactions is independent of their binding event predictions. Overall there were more motif supported events in the set of events identified by SPROUT than the other two sets. The maximum height of each

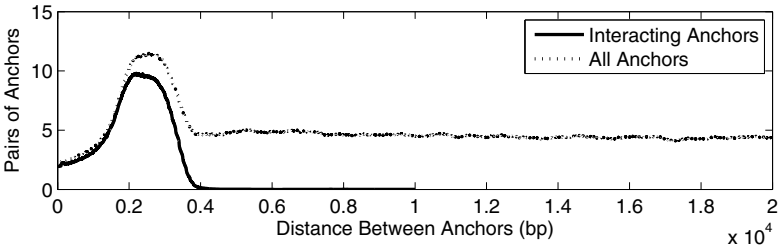


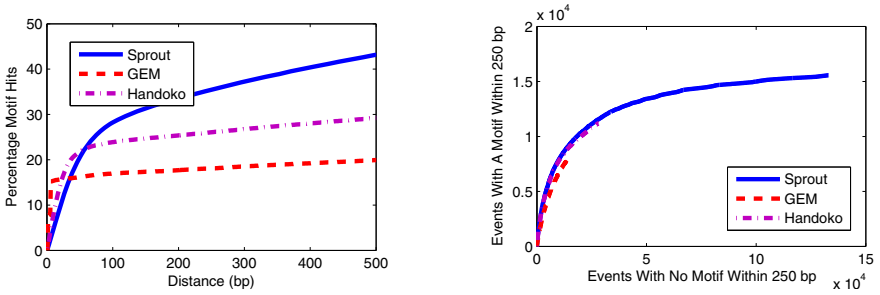
Fig. 2. Smoothed plots of the frequency at which interacting anchors identified by SPROUT exist at distances up to 20000 bp. Beyond 4000 bp anchors are very infrequently interacting relative to the number of possible interactions at a given distance and individual. This suggests that interactions that are detected by SPROUT that span more than 4000 bp are not explained by the linear proximity of the anchors.

curve in Fig. 3b indicates the total number of motif supported events discovered by each method. Furthermore, the weight assigned to events by SPROUT is a better classifier of motif supported events than the weights assigned by Handoko et al. to the ChIA-PET events or the weights assigned to ChIP-Seq events by GEM. The fact that the SPROUT curve in Fig. 3b is always greater than the other curves indicates that SPROUT achieves greater specificity.

The anchor regions identified by Handoko et al. tend to be relatively broad with an average width of 1997.7 bp (Fig. 4). By identifying binding events within the anchor regions, it may be possible to recover the true anchors for the interactions as a post-processing step. However, as an example of the difficulty in interpreting such broad interaction anchor regions, 63 of the 4077 interacting anchors identified by Handoko et al. contain more than one motif supported binding event. One of the strengths of SPROUT is that interactions called by SPROUT are directly associated with binding events, thereby reducing ambiguity in interpreting the results.

Upon comparing the significant interactions identified by SPROUT and Handoko et al., we noticed that certain significant interactions were missed by Handoko et al. Of the 420 significant interactions that span more than 5928 bp identified by SPROUT, 87 interactions lack a corresponding interaction identified by Handoko et al. with both anchors within 4 kb of the SPROUT identified anchors. Of these interactions, 64 have binding events identified by Handoko et al. within 250 bp of the SPROUT identified anchors. The fact that Handoko et al. failed to identify several interactions between binding events that they identify with their own method for detecting binding events indicates one of the benefits of SPROUT's approach of integrating interaction detection with anchor detection.

Handoko et al. identified 2241 significant interactions, however many of these interactions do not fit the model of an interaction between two distinct anchors as defined by SPROUT (Fig. 5). 200 of the Handoko et al. interactions do subsume SPROUT identified interactions and an additional 11 Handoko et al. interactions have SPROUT identified interactions with anchors within 4 kb of their anchors. 1181 of the Handoko et al. interactions that do not subsume SPROUT identified



(a) (b)

Fig. 3. Evaluation of the accuracy of CTCF binding events predicted by SPROUT and Handoko et al. from the ChIA-PET data as well as by GEM from an independent ChIP-Seq dataset. (a) The percentage of CTCF motif matches in the genome that have a binding event identified within distances up to 500 bp. (b) We used the presence of a CTCF motif match within 250 bp of an event as an approximate indicator of true positive anchor calls. As thresholds for significance are varied for each method, the number of true positive and false positive calls are plotted. This results in a receiver operating characteristic curve for each method.

.

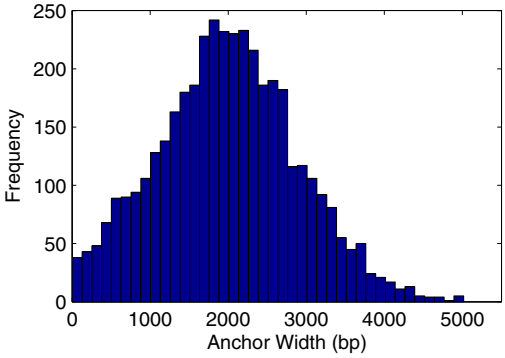


Fig. 4. A histogram of the widths of anchors identified by Handoko et al. illustrating the breadth of many of the anchor regions.

interactions do not contain a CTCF binding event (by their own definition) at one or both anchors. This clearly indicates that these are unlikely to reflect true interactions between CTCF-bound anchors. Of the remaining 860 Handoko et al. interactions, 52 involve 0 pairs of reads and 123 involve 1 pair of reads according to our alignment of the data. Handoko et al. used a rescue procedure in which reads that align to multiple locations are in some cases assigned to one location. We did not use this procedure when we aligned the reads which may explain why Handoko et al. assign significance to interactions that do not seem to be supported by enough read pairs without the rescue procedure. This leaves 685 Handoko et al. interactions that are supported by at least 2 pairs of reads that

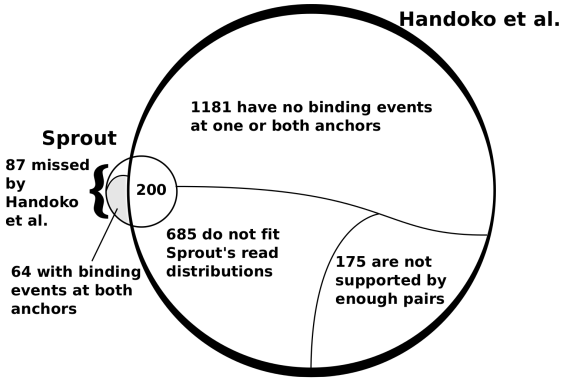


Fig. 5. Most of the interactions identified by Handoko et al. are not supported by pairs of reads with ends that fit SPROUT's read distribution.

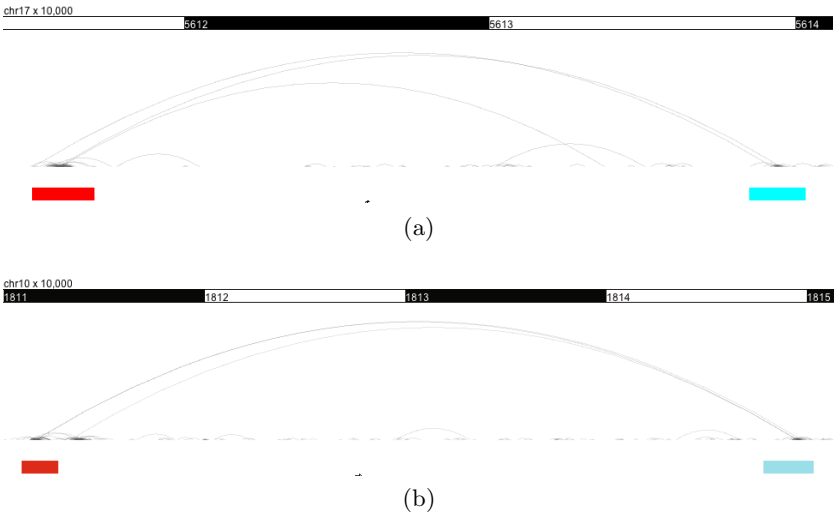
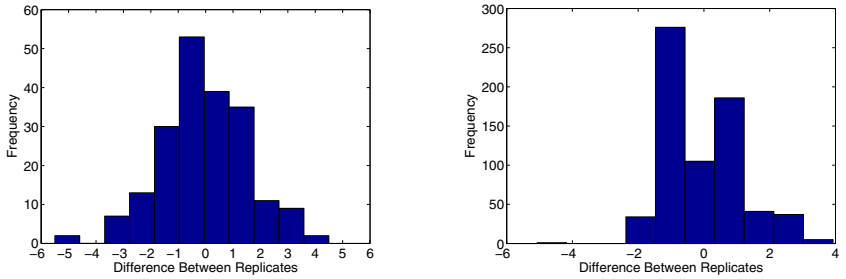


Fig. 6. Two interactions that are identified by Handoko et al. The boxes indicate the anchor regions that they identify. (a) This interaction is not called significant by SPROUT because the pairs of reads that connect the anchor regions do not fit SPROUT's model. (b) SPROUT does call a significant interaction between the anchors that fall within the Handoko et al. anchor regions because the pairs of reads that connect the regions were likely to have been generated by the anchors within the regions according to SPROUT's model. Note that there is a second potential anchor on the left side that falls outside of the Handoko et al. identified region. This anchor is identified by both SPROUT and Handoko et al. and is identified by SPROUT but not by Handoko et al. as an independent interaction with the anchor on the right.



(a) (b)

Fig. 7. Evaluation of biological replicate consistency in interactions discovered by both methods and in interactions identified by Handoko et al. that do not fit SPROUT's read distributions. (a) A histogram of the difference in the number of pairs of reads from each biological replicate that connect anchors identified by Handoko et al. that subsume interactions called by SPROUT. To account for the overall difference in signal strength, the values were subtracted by the mean per interaction difference. There are interactions that differ in support between the biological replicates. However, the normalized difference in pairs between the biological replicates is most frequently close to 0. (b) A histogram of the difference in the number of pairs of reads from each biological replicate that connect anchors identified by Handoko et al. that are supported by a plausible number of pairs of reads but do not fit SPROUT's read distributions. As in (a), the differences are subtracted by the mean difference. The biological replicates differ by one pair of reads much more frequently than they agree. This difference is significant given that 491 out of 685 interactions in this set are only supported by 2 pairs of reads total.

do not subsume SPROUT identified interactions. However, upon examination of many of these interactions (Fig. 6), the broadness of the interaction anchors allow pairs of reads to be considered together even though the positions of the reads do not fit SPROUT's model of how reads should be distributed around anchors.

Interactions supported by pairs of reads that fit SPROUT's read distributions are more consistent across biological replicates and therefore are more likely to represent true interactions. To demonstrate this we consider two sets of interactions. One set, which we call the good fit set, includes the 200 interactions identified by Handoko et al. that subsume SPROUT identified interactions. The other set, which we call the bad fit set, includes the 685 Handoko et al. interactions that contain binding events at both anchors and are connected by at least 2 pairs of reads but do not subsume interactions discovered by SPROUT. The first thing we noticed is that the interactions in the good fit set tend to be supported by more pairs of reads. The average number of pairs per interaction in the good fit set is 4.15 while for the bad fit set the average number of pairs is 2.73. We then identified which of the biological replicates each pair of reads came from. As can be seen in Fig 7, the biological replicates assign pairs of reads to the interactions in the good fit set more consistently than interactions in the bad fit set.

4 Conclusion

SPROUT uses all pairs of reads to estimate anchor positions and learns empirical interaction read distributions to more accurately assign pairs of reads to anchors. SPROUT interaction calls are more consistent across biological replicates than the method proposed by Handoko et al. Identifying high confidence interactions between accurately positioned anchors is a task that is increasing in importance as more genome structure data are produced. Utilizing data from various types of high throughput sequencing based approaches, several successful approaches to identifying regulatory elements have been developed. However, it is impossible to fully understand how these regulatory elements function without putting them in their spatial context in the nucleus. The interaction results produced by SPROUT from ChIA-PET data allow for a more accurate understanding of this spatial context.

References

1. Chen, X., Xu, H., Yuan, P., Fang, F., Huss, M., Vega, V.B., Wong, E., Orlov, Y.L., Zhang, W., Jiang, J., Loh, Y., Yeo, H.C., Yeo, Z.X., Narang, V., Govindarajan, K.R., Leong, B., Shahab, A., Ruan, Y., Bourque, G., Sung, W., Clarke, N.D., Wei, C., Ng, H.: Integration of External Signaling Pathways with the Core Transcriptional Network in Embryonic Stem Cells. Cell 133, 1106–1117 (2008)
2. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum Likelihood Estimation from Incomplete Data via the EM Algorithm. J. R. Stat. Soc. B 39, 1–38 (1977)
3. Figueiredo, M.A., Jain, A.K.: Unsupervised Learning of Finite Mixture Models. IEEE T. Pattern Anal. 4, 381–396 (2002)
4. Francastel, C., Schübeler, D., Martin, D.I.K., Groudine, M.: Nuclear Compartmentalization and Gene Activity. Nat. Rev. Mol. Cell Biol. 1, 137–143 (2000)
5. Fullwood, M.J., Liu, M.H., Pan, Y.F., Liu, J., Xu, H., Mohamed, Y.B., Orlov, Y.L., Velkov, S., Ho, A., Mei, P.H., Chew, E.G.Y., Huang, P.Y.H., Welboren, W., Han, Y., Ooi, H.S., Ariyaratne, P.N., Vega, V.B., Luo, Y., Tan, P.Y., Choy, P.Y., Wansa, K.D.S.A., Zhao, B., Lim, K.S., Leow, S.C., Yow, J.S., Joseph, R., Li, H., Desai, K.V., Thomsen, J.S., Lee, Y.K., Karuturi, R.K.M., Herve, T., Bourque, G., Stunnenberg, H.G., Ruan, X., Cacheux-Rataboul, V., Sung, W., Liu, E.T., Wei, C., Cheung, E., Ruan, Y.: An Oestrogen-Receptor-α-Bound Human Chromatin Interactome. Nature 462, 58–64 (2009)
6. Guo, Y., Mahony, S., Gifford, D.K.: High Resolution Genome Wide Binding Event Finding and Motif Discovery Reveals Transcription Factor Spatial Binding Constraints. P.L.O.S. Comput. Biol. 8, e1002638 (2012)
7. Handoko, L., Xu, H., Li, G., Ngan, C.Y., Chew, E., Schnapp, M., Lee, C.W.H., Ye, C., Ping, J.L.H., Mulawadi, F., Wong, E., Sheng, J., Zhang, Y., Poh, T., Chan, C.S., Kunarso, G., Shahab, A., Bourque, G., Cacheux-Rataboul, V., Sung, W., Ruan, Y., Wei, C.: CTCF-Mediated Functional Chromatin Interactome in Pluripotent Cells. Nat. Genet. 43, 630–638 (2011)
8. Hatzis, P., Talianidis, I.: Dynamics of Enhancer-Promoter Communication During Differentiation-Induced Gene Activation. Mol. Cell 10, 1467–1477 (2002)
9. Langmead, B., Trapnell, C., Pop, M., Salzber, S.L.: Ultrafast and Memory-Efficient Alignment of Short DNA Sequences to the Human Genome. Genome Biol. 10, R25 (2009)

10. Li, G., Fullwood, M.J., Xu, H., Mulawadi, F.H., Velkov, S., Vega, V., Ariyaratne, P.N., Mohamed, Y.B., Ooi, H., Tennakoon, C., Wei, C., Ruan, Y., Sung, W.: ChIA-PET Tool for Comprehensive Chromatin Interaction Analysis with Paired-End Tag Sequencing. Genome Biol. 11, R22 (2010)
11. Lieberman-Aiden, E., van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., Sandstrom, R., Bernstein, B., Bender, M.A., Groudine, M., Gnirke, A., Stamatoyannopoulos, J., Mirny, L.A., Lander, E.S., Dekker, J.: Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. Science 326, 289–293 (2009)
12. Meng, X., Rubin, D.B.: Maximum Likelihood Estimation via the ECM Algorithm: A General Framework. Biometrika 80, 267–278 (1993)
13. Meshorer, E., Misteli, T.: Chromatin in Pluripotent Embryonic Stem Cells and Differentiation. Nat. Rev. Mol. Cell Biol. 7, 540–546 (2006)
14. Misteli, T.: Beyond the Sequence: Cellular Organization of Genome Function. Cell 128, 787–800 (2007)
15. Misteli, T., Soutoglou, E.: The Emerging Role of Nuclear Architecture in DNA Repair and Genome Maintenance. Nat. Rev. Mol. Cell Biol. 10, 243–254 (2009)
16. Philips, J.E., Corces, V.G.: CTCF: Master Weaver of the Genome. Cell 137, 1194–1211 (2009)
17. Pombo, A., Branco, M.R.: Functional Organisation of the Genome During Interphase. Curr. Opin. Genet. Dev. 17, 451–455 (2007)
18. Rozowski, J., Euskirchen, G., Auerbach, R.K., Zhang, Z.D., Gibson, T., Bjornson, R., Carriero, N., Snyder, M., Gerstein, M.B.: PeakSeq enables systematic scoring of ChIP-seq Experiments Relative to Controls. Nat. Biotechnol. 27, 66–75 (2009)
19. van Steensel, B., Dekker, J.: Genomics Tools for Unraveling Chromosome Architecture. Nat. Biotechnol. 28, 1089–1095 (2010)
20. Zhao, R., Bodnar, M.S., Spector, D.L.: Nuclear Neighborhoods and Gene Expression. Curr. Opin. Genet. Dev. 19, 172–179 (2009)