A Streamlined and Generalized Analysis of Chromatin ImmunoPrecipitation Paired-End diTag Data

Vinsensius B. Vega[1], Yijun Ruan[2], and Wing-Kin Sung[1]

[1] Computational and Mathematical Biology Group
[2] Clone and Sequencing Group
Genome Institute of Singapore
{vegav,ruanyj,sungk}@gis.a-star.edu.sg

Abstract. Comprehensive, accurate and detailed maps of transcription factor binding sites (TFBS) help to unravel the transcriptional regulatory relationship between genes and transcription factors. The recently developed sequencing-based genome-wide approach ChIP-PET (Chromatin ImmunoPrecipitation coupled with Paired-End diTag analysis) permits accurate and unbiased mapping of TF-DNA interactions. In this paper we outline a methodical framework to analyze ChIP-PET sequence data to identify most likely binding regions. Mathematical formulations were derived to streamline and strengthen the analysis. We established a more faithful noise distribution estimation that leads to the adaptive threshold scheme. The algorithms were evaluated using three real-world datasets. Using motif enrichment as indirect evidence and additional ChIP-qPCR validations, the overall performance was consistently satisfactory.

1 Introduction

Transcription factors (TF) play a pivotal role in controlling gene expression, and thus directing the cellular behavior and mechanisms. Part of the effort to fully decipher the intricate regulatory networks and pathway, a key intermediate goal would be to identify the relevant Transcription Factor Binding Sites (TFBS). Such is also one of the goals set out for the human genome[1].

Chromatin Immuoprecipitation (ChIP) is a powerful approach to study in vivo protein-DNA interactions. It consists of five major steps: (i) cross-link the DNA binding proteins to the DNA in vivo, (ii) shear the chromatin fibers using sonication or otherwise, (iii) immunoprecipitate the chromatin fragments using specific antibody against given protein targets, (iv) reverse the cross-linking of protein-bound DNA, and (v) analyze the ChIP enriched DNA fragments. These DNA fragments can then be characterized using high throughput approaches, such as hybridization-based ChIP-chip analysis [2],[3],[4],[5] or direct DNA sequencing. ChIP sequencing can be performed by sequencing individually cloned fragments [6],[7], concatenations of single tags derived from fragments (STAGE) [8],[9],[10],[11] or concatenations of pair-end-ditags to infer the linear structure

M. Bubak et al. (Eds.): ICCS 2008, Part III, LNCS 5103, pp. 130–139, 2008.

of ChIP DNA fragments (ChIP-PET)[12],[13]. The sequencing approaches have their advantages over the hybridization-based approaches by elucidating the exact nucleotide content of target DNA sequences.

In a ChIP-PET experiment, 5' (18bp) and 3' (18bp) signatures for each of the ChIP enriched DNA fragments were extracted and joined to form the paired end tag structure (PET) that were then concatenated for efficient sequencing analysis. The PET sequences were then mapped to the reference genome to infer the full content of each of the ChIP DNA fragments. The protein-DNA interaction regions enriched by ChIP procedure will have more DNA fragments representing the target regions than the non-specific regions. Therefore, with sufficient sequence sampling in the DNA pool of a ChIP experiment, multiple DNA fragments originated from the target regions will be observed.

In previous analyses of ChIP-PET data [12],[13], Monte Carlo simulations were employed to distinguish true signals from noise. This paper proposes mathematical formulations for performing the similar assessment as the Monte Carlo simulation in an efficient manner, and further generalizes the approach to resolve potential irregular noise arising from anomalous chromosomal copies and configuration.

2 Results and Discussion

2.1 PET Clusters as the Readout for TFBS

Presence of PET clusters is clearly an initial indication of genomic loci enriched for TF-bound fragments. The more PETs that a cluster has, the more probable the TF binds to the region. We can set a minimum cut-off criterion, say

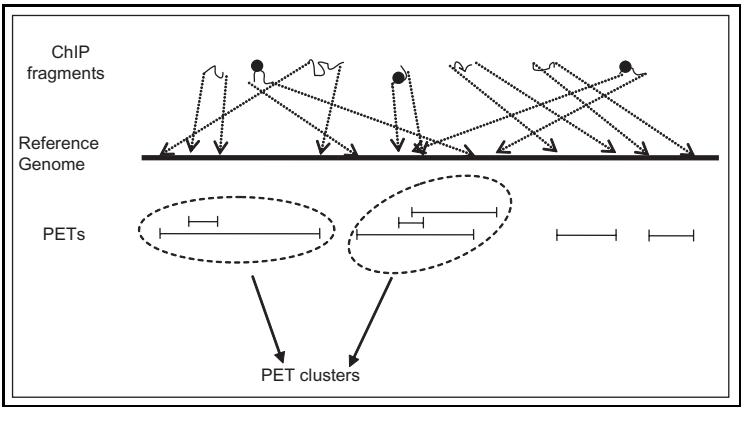


Fig. 1. ChIP fragments, PETs, and ChIP-PET clusters. ChIP fragments might be TF-bound (shaded circles) or simply noise. Mapped ChIP fragments are called PETs. Overlapping PETs are grouped into ChIP-PET clusters.

z, and classify clusters with at least z PETs (i.e. PETz+ clusters) to be the highly probable clusters with TF binding. A Monte Carlo approach has been successfully employed to determine this threshold [12].

More analytically, if we assume that the noisy PETs are randomly and uniformly distributed along the genome, then the distance, D, between any two consecutive random PETs is expected to follow the exponential distribution with rate $\lambda = \frac{F}{G}$, where F is the total number of PETs and G is the genome length. The probability of two PETs overlapping (i.e. the distance between them is less than or equal the (expected) PET length) by chance alone is $P_{\exp}(X \leq k; \lambda)$ where $P_{\exp}$ is the cumulative exponential distribution function whose rate is λ and k is the expected length of a PET. Two overlapping PETs can be found in a PET2 cluster and beyond. Thus, the probability $P_{\exp}(X \leq k; \lambda)$ is the probability of a PET2+ cluster to happen simply by chance. More generally, the probability of the occurrence of a PETn+ cluster by random is:

$$\Pr_{\text{PET}}(Y \geq n; \lambda) \approx (P_{\exp}(X \leq k; \lambda))^{(n-1)} = \left(1 - e^{-\lambda k}\right)^{(n-1)} \quad (1)$$

In place of the Monte Carlo simulations, one can readily compute the p-value of random PETn+ clusters using the above equation to determine the appropriate threshold for a given ChIP PET library. Additional empirical evidence for Eq. (1) will be provided later in the text.

2.2 Counting on Maximum Support for Defining Binding Regions

While number of PETs in a cluster is useful for assessing whether the cluster is likely to be a true signal, clusters with seemingly good number of PETs can still be generated by random noise. For a true binding region with many PETs, the exact position of the actual protein-DNA site will be more refined and appear as a sharp peak in the cluster. However, it is not uncommon to find big clusters whose overlapping regions are not well concentrated. This is an indication that they might have been formed simply by chance. Figure 2 shows a snapshot of two clusters from real libraries, contrasting a typical good cluster (left) with well defined core, to a configuration with scattered overlap region (right) which was most likely formed by random PETs.

We call a cluster as a moPETn (*maximum overlap* PET n) cluster if all of its sub-region is supported by at most n PETs. Similar to the previous definition, moPETn+ clusters represent the set of moPETm clusters where $m \geq n$. The left cluster in Fig.2 is PET5/moPET5, while the right cluster is PET5/moPET2. To some extent, PETn/moPETm (where $m < n$) clusters are doubtful. There are 96 of such clusters in the p53 library, while the Oct4 and ER libraries contain 4,929 and 910 such clusters.

The probability of a moPETn to be initiated by an arbitrary PET $< s, l >$ can be estimated by the probability of observing additional $(n-1)$ PET starting sites

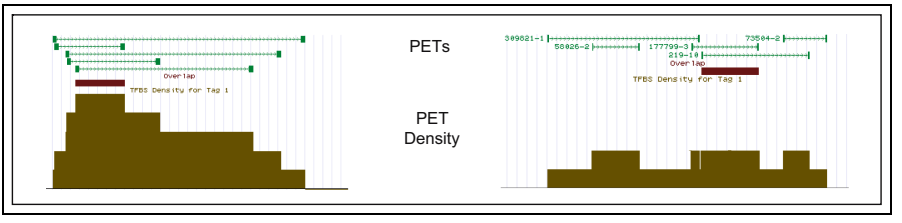


Fig. 2. A comparison of high fidelity cluster and noisy cluster. Good clusters are generally well-defined (left cluster), containing a strong overlapping region. Dispersed ChIP PET segments (right cluster) hint at the possibility of cluster formation purely at random and by chance alone.

at most l-bp away from s. This probability follows that of Poisson distribution for observing $(n-1)$ events whose rate is λ within the interval k (=expected PET length). More formally, the probability of an arbitrary PET to initiate a moPETn cluster:

$$\Pr_{\text{moPET}}(Y = n; \lambda) \approx P_{\text{Poisson}}(X = (n-1); \lambda k) = \frac{e^{-\lambda k}(\lambda k)^{(n-1)}}{(n-1)!} \qquad (2)$$

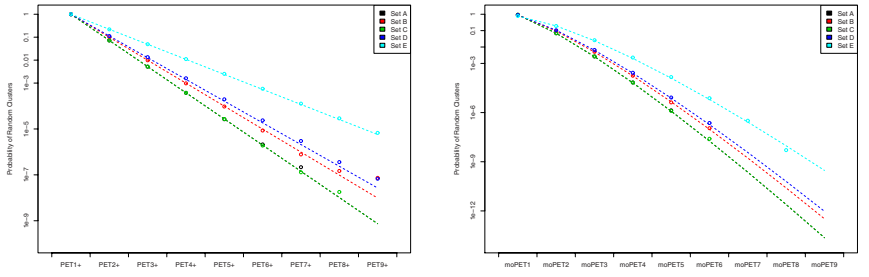
Using $\Pr_{\text{moPET}}(n)$ and given the acceptable p-value level, we can determine the appropriate cut-off of moPETn for identifying true TF-binding regions. Comparison with simulation results is presented below.

2.3 Comparing Empirical Results and Analytical Formulations

To evaluate the correctness of our statistical formulations in Eqs. (1) and (2), we compared the analytical estimations of PETn+ and moPETn clusters distributions to the empirical ones generated through a 100,000 runs of Monte Carlo simulations (see Methods) with different sets of parameter as listed in Table 1. The collected statistics were used to construct empirical distributions which were then compared with the proposed analytical framework. Figure 3(a) and 3(b) contrasts the empirical probability of PETn+ and moPETn occurrence (points) against the analytical estimations (dashed lines). The analytical curves tracks the empirical values very well, reconfirming the validity of the analytical distributions.

2.4 Adaptive Approach for Biased Genomes

The estimation of rate λ, i.e. the expected number of PETs per nucleotide, plays a critical role in Eqs. (1) and (2). This rate signifies the expected noise level of the dataset. A single global rate λ reflects the assumption that the



(a) (b)

Fig. 3. Comparing simulation and analytical results. Probability of a random (**A**) PETn+ or (**B**) moPETn cluster produced by chance, as estimated by Monte Carlo simulations (points) and computed analytically (dashed lines).

noisy PETs are randomly uniformly distributed across the genome. Although may be acceptable in general, libraries generated from aberrant genome require more refined estimation. Amplified genomic regions will have higher PET counts than the overall genome, making their purely random clusters bear stronger signal than those of normal regions. On the other hand, regions with significant deletions will contain less than expected PETs and their true binding loci will be much weaker. Using single global λ would result in higher false positive rates in amplified regions and higher false negative rates in deleted regions.

We devised a two-phase adaptive approach that takes account of local biases (see Fig.4) in predicting the most probable source (true binding vs. noise) of each PET cluster. Given a cluster c, the first phase considers the local window of some predefined size L centered on the cluster c, and, estimates the total number of noise PETs. The second phase computes the local λ and calculates a local moPET (or PET) cut-off T_{moPET}. Clusters c is considered to be binding region if its moPET (or PET) count is greater than T_{moPET}.

The noise estimation step (phase 1) counts the number of potentially noisy PETs within the window. Overestimation of noise would increase false negatives, while underestimation would add false positives. We adhere to two heuristics: (i) the current cluster should not be assumed as real and (ii) other clusters within the windows that seem to be real clusters should, as much as possible, not be counted as noise. The first rule stems from the fact that most of the clusters (especially PET1 clusters) are noise. Observations that binding sites are sometimes located proximal to each other motivated the second rule. For the window size L, we set it to be at least twice the expected distance between two PETs (i.e. λ^{-1}). Noise estimation starts by identifying the probable noisy clusters. Assuming that the current cluster c is noisy, clusters within the window L with higher moPET counts than the current cluster c are contextually considered non-noise (line 3 in Fig.4). Next, the rate of noisy PETs per cluster is estimated by taking the geometric mean [14],[15] of PET counts of (locally) noisy

```
GoodCluster(c, p, L)
1  Let D be the set of clusters that are located at most L/2 basepairs away
   (upstream or downstream) from c
2  G ← {}                                        ▷ Start of local noise estimation
3  for each cluster d ∈ D : if MO(d) ≤ MO(c) then G = G ∪ {PET(d)}
4  g ← GEOMEAN(G)
5  S ← Σ_{d∈D} min(PET(d), g)                    ▷ S is the estimated local noise

6  λ ← S/L                                       ▷ Start of threshold determination
7  T_moPET ← min({n | Pr_moPET(Y ≥ n; λ) ≤ p})
8  if MO(c) > T_moPET
       then return TRUE
       else return FALSE
```

Fig. 4. Pseudocode of the adaptive thresholding algorithm

clusters (line 4 in Fig.4). The final sum of noisy PETs, S, is calculated by adding the noisy PET counts of all the clusters within the current window (line 5 in Fig.4).

The second phase is quite straightforward through performing sufficient iterations of Monte Carlo simulations or the application of the Eqs. (1) or (2) using the local rate λ ($=S/L$) and considering the window length L.

2.5 Performance on Real Datasets

Three real datasets were used in our evaluation: p53 ChIP-PET [12], Oct4 ChIP-PET [13], and Estrogen Receptor (ER) ChIP-PET [16]. For each dataset, we applied our proposed algorithms to predict TF-bound regions. The predicted regions were then evaluated indirectly by enrichment of putative relevant binding motifs and (whenever available) directly through further ChIP qPCR validation assays (see Methods).

The p53 library was the first and the smallest dataset (65,714 PETs, average length: 625bp) and was constructed using the human HCT116 cancer cell lines. The ER ChIP PET library comprised 136,152 PETs (average length: 72bp), was assayed on human MCF-7 breast cancer cell lines. The largest library among the three, the Oct4 ChIP PET, was based on mouse E14 cell lines (366,639 PETs, average length: 627bp). The non-gapped genome lengths for human and mouse are estimated at 2.8Gbp (UCSC hg17) and 2.5Gbp (UCSC mm5).

Setting the cut-off of at $p = 10^{-3}$, the selected clusters for p53 is PET3+ or moPET3+, for ER is PET4+ or moPET3+, and for Oct4 is PET4+ or moPET4+. Table 2 gives the validations of each PET cluster group in each library and PET or moPET cluster group. Sharp motif enrichment can be seen

at the selected cut-offs in all libraries, compared to the PET2 or moPET2 group, which is expected to be noisy. Table 2 also shows high success rate of ChIP-qPCR validations. The p53 library had 100% of the ChIP-qPCR tested sites showing enrichment of p53 binding. The high ChIP-qPCR success rate ($> 95\%$) for the selected Oct4 moPET4+ clusters also increased our confidence of the validity of the cluster selection approach.

Prior to running the ChIP-qPCR validation for the ER library, we noticed unusual concentrations of PETs in regions, which correlated well with the regions previously reported to be amplified in the underlying MCF-7 cell lines [17]. This prompted us to employ the adaptive moPET thresholding algorithm to 'normalize' the amplified regions. We also applied the adaptive approach on the other two datasets, to see its effect on other libraries from relatively normal cell lines (i.e. the p53 and Oct4 libraries). The result is summarized in Table 3.

Adaptive thresholding might exclude clusters selected under the global thresholding and re-include clusters which would otherwise be excluded because they were below the global threshold. Global and adaptive moPET thresholding produced the same results for p53 (see Table 3). Interestingly, application of adaptive thresholding on the Oct4 library re-included some of the moPET3 clusters, with a higher motif enrichment. Only a tiny fraction of the moPET4 was rejected, with no significant impact on motif enrichment. The ChIP qPCR success rates for the adaptive-selected clusters were higher than before. A sizeable portion of the moPET3+ in ER ChIP PET library was rejected and the selected clusters had better motif enrichment, indicative of true binding. ChIP-qPCR assays on random samples of the selected clusters confirmed that further.

3 Methods

3.1 ChIP-PET Clustering

The primary ChIP-PET data is the locations and lengths of the ChIP-PET fragments. The tuple $< s, l >$ represents an l-bp long PET fragment mapped into location s. Two PET fragments $< s_1, l_1 >$ and $< s_2, l_2 >$, where $s_1 \leq s_2$, are said to be overlapping if $s_1 + l_1 \geq s_2$. A ChIP-PET cluster is defined as the largest set of cascading overlapping PET fragments.

3.2 Simulation Procedures

To generate a random PET library, we preformed a Monte Carlo simulation while taking into account the overall genome length (G), the total number of PETs (M), and the desired PETs' lengths (minimum and maximum lengths). In each Monte Carlo simulation, M points were randomly picked along the G-bp genome, mimicking the generation of a PET library containing completely random fragments. For each picked point, a random length was sampled from a uniform distribution within the given minimum and maximum bounds. Simulated PETs were clustered accordingly. Statistics of PETn+ and moPETn clusters were collected and averaged over a sufficient number of Monte Carlo iterations.

3.3 Evaluations of Selected Clusters from Real Datasets

The goodness of these clusters (i.e. whether these clusters were truly bound by the TF) was then doubly-assessed: (i) indirectly from the enrichment of putative relevant binding motifs among the selected clusters and (ii) directly through further ChIP qPCR validation assays. The putative motifs for p53 and Oct4 were identified based on the binding site models described in their respective papers [12],[13]. Putative ER binding motif were based on the consensus sequence, GGTCAnnnTGACC [18], and allowing for up to 2nt mismatches. Additional ChIP qPCR validations have also been carried out on some of the selected clusters [12],[13],[16].

4 Conclusions

We have described a more principled framework for analyzing ChIP-PET data to identify transcription factor binding regions. To dichotomize the PET clusters into potentially binding regions and likely non-binding regions, we utilized a random PET generation model and estimated the improbable concentration of PETs generated at random. The adaptive thresholding framework was introduced to handle aberrant genomes, e.g. due to amplifications and deletions in cancer cell lines, or other experimental conditions. These analyses might be further improved by taking into account other known inherent properties of the genome (e.g. prevalence of repeats). We also noticed a potential utility of the adaptive technique for identifying intrinsic features of the genome (e.g. for delineating amplified or deleted segments).

Acknowledgments. We would like to thank Jane Thomsen for providing the ChIP-qPCR validation data of the ER library. This work is supported by the Agency for Science, Technology and Research (A*STAR) of Singapore and an NIH/NHGRI ENCODE grant (1R01HG003521-01).

References

1. ENCODE Project Consortium: The ENCODE (ENCyclopedia Of DNA Elements) Project. Science 306, 636–640 (2004)
2. Iyer, V.R., Horak, C.E., Scafe, C.S., et al.: Genomic binding sites of the yeast cell-cycle transcription factors sbf and mbf. Nature 409, 533–538 (2001)
3. Ren, B., Robert, F., Wyrick, J.J., et al.: Genome wide location and function of dna binding proteins. Science 290, 2306–2309 (2000)
4. Horak, C.E., Mahajan, M.C., Luscombe, N.M., et al.: GATA-1 binding sites mapped in the β-globin locus by using mammalian chip-chip analysis. Proceedings of the National Academy of Sciences 99, 2924–2929 (2002)
5. Weinmann, A.S., Pearlly, S.Y., Oberley, M.J., et al.: Isolating human transcription factor targets by coupling chromatin immunoprecipitation and cpg island microarray analysis. Genes Dev. 16, 235–244 (2002)

6. Weinmann, A.S., Bartley, S.M., Zhang, T., et al.: Use of chromatin immunoprecipitation to clone novel E2F target promoters. Mol. Cell. Biol. 21, 6820–6832 (2001)
7. Hug, B.A., Ahmed, N., Robbins, J.A., Lazar, M.A.: A chromatin immunoprecipitation screen reveals protein kinase cbeta as a direct runx1 target gene. J. Biol. Chem. 279(2), 825–830 (2004)
8. Impey, S., McCorkle, S.R., Cha-Molstad, H., et al.: Defining the creb regulona genome-wide analysis of transcription factor regulatory regions. Cell 119, 1041–1054 (2004)
9. Kim, J., Bhinge, A.A., Morgan, X.C., Iyer, V.R.: Mapping dna-protein interactions in large genomes by sequence tag analysis of genomic enrichment. Nature Methods 2(1), 47–53 (2004)
10. Chen, J., Sadowski, I.: Identification of the mismatch repair genes PMS2 and MLH1 as p53 target genes by using serial analysis of binding elements. Proceedings of the National Academy of Sciences 102(13), 4813–4818 (2005)
11. Roh, T.Y., Cuddapah, S., Zhao, K.: Active chromatin domains are defined by acetylation islands revealed by genome-wide mapping. Genes Dev. 19(5), 542–552 (2005)
12. Wei, C.L., Wu, Q., Vega, V.B., et al.: A global map of p53 transcription-factor binding sites in the human genome. Cell 124(1), 207–219 (2006)
13. Loh, Y.H., Wu, Q., Chew, J.L., Vega, V.B., et al.: The oct4 and nanog transcription network regulates pluripotency in mouse embryonic stem cells. Nature Genetics 38(4), 431–440 (2006)
14. McAlister, D.: The law of the geometric mean. Proceedings of the Royal Society of London 29, 367–376 (1879)
15. Fleming, J., Wallace, J.: How not to lie with statistics: the correct way to summarize benchmark results. Communications of the ACM 29, 218–221 (1986)
16. Lin, C.Y., Vega, V.B., Thomsen, J.S., et al.: Whole-genome cartography of estrogen receptor alpha binding sites. PLoS Genet 3(6) (June 2007)
17. Shadeo, A., Lam, W.L.: Comprehensive copy number profiles of breast cancer cell model genomes. Breast Cancer Res 8(1), R9 (2006)
18. Klinge, C.: Estrogen receptor interaction with estrogen response elements. Nucleic Acids Research 29, 2905–2919 (2001)

Appendix: Tables

Table 1. Simulation setups. Five Monte Carlo simulation sets run to assess the analytical model of random PETn and moPETn clusters formations.

Simulation Set	A	B	C	D	E
Genome length	2Mbp	2Mbp	20Mbp	10Mbp	10Mbp
Number of PETs	300	300	3000	2000	5000
Min. PET length	500bp	700bp	500bp	200bp	300bp
Max. PET length	500bp	700bp	500bp	1000bp	700bp

Table 2. Motif enrichments and ChIP-qPCR validation rate of clusters selected by global thresholding

(A) p53 ChIP-PET clusters

	PET2	PET3	PET4	PET5	PET6	PET7	PET8+
Total clusters	1453	161	66	38	29	13	29
% with motifs	15.97%	59.63%	80.30%	65.79%	89.66%	84.62%	82.76%
ChIP-qPCR success rate	N/A	N/A	100.0%	100.0%	100.0%	100.0%	100.0%
	moPET2	moPET3	moPET4	moPET5	moPET6	moPET7+	
Total clusters	1489	140	69	30	26	35	
% with motifs	16.25%	67.14%	81.16%	70.00%	88.46%	88.57%	
ChIP-qPCR success rate	N/A	100.0%	100.0%	100.0%	100.0%	100.0%	

(B) Oct4 ChIP-PET clusters

	PET2	PET3	PET4	PET5	PET6	PET7	PET8+
Total clusters	29453	5556	1540	550	223	102	201
% with motifs	16.74%	24.62%	34.35%	42.36%	52.47%	49.02%	45.77%
ChIP-qPCR success rate	10.00%	9.68%	88.24%	90.48%	100.00%	100.00%	95.00%
	moPET2	moPET3	moPET4	moPET5	moPET6	moPET7+	
Total clusters	32739	3734	724	189	93	146	
% with motifs	17.57%	27.64%	41.57%	54.50%	70.97%	43.15%	
ChIP-qPCR success rate	10.00%	8.82%	95.00%	100.00%	100.00%	100.00%	

(C) ER ChIP-PET clusters

	PET2	PET3	PET4	PET5	PET6	PET7	PET8+
Total clusters	5704	930	341	181	124	78	216
% with motifs	40.06%	57.31%	65.69%	70.72%	76.61%	78.21%	83.33%
	moPET2	moPET3	moPET4	moPET5	moPET6	moPET7+	
Total clusters	6100	756	281	134	95	208	
% with motifs	41.02%	61.90%	64.77%	76.12%	78.95%	85.10%	

Table 3. Motif enrichments and ChIP-qPCR validation rate of clusters selected by adaptive thresholding

(A) p53 ChIP-PET clusters

	moPET2	moPET3	moPET4	moPET5	moPET6	moPET7+
Total clusters	0	140	69	30	26	35
% with motifs	N/A	67.14%	81.16%	70.00%	88.46%	88.57%
ChIP-qPCR success rate	N/A	100.0%	100.0%	100.0%	100.0%	100.0%

(B) Oct4 ChIP-PET clusters

	moPET2	moPET3	moPET4	moPET5	moPET6	moPET7+
Total clusters	0	524	717	189	93	146
% with motifs	N/A	36.83%	41.84%	54.50%	70.97%	43.15%
ChIP-qPCR success rate	N/A	16.7%	95.00%	100.0%	100.0%	100.0%

(C) ER ChIP-PET clusters

	moPET2	moPET3	moPET4	moPET5	moPET6	moPET7+
Total clusters	0	552	245	134	95	208
% with motifs	N/A	65.58%	68.57%	76.12%	78.95%	85.10%
ChIP-qPCR success rate	N/A	70.0%	83.3%	100.0%	100.0%	100.0%