COMPUTATIONAL ANALYSIS AND MODELING OF GENOME-SCALE AVIDITY DISTRIBUTION OF TRANSCRIPTION FACTOR BINDING SITES IN CHIP-PET EXPERIMENTS

VLADIMIR A. KUZNETSOV* [1]
kuznetsov@gis.a-star.edu.sg

YURIY L. ORLOV [1]
orlovy@gis.a-star.edu.sg

CHIA LIN WEI[1]
weicl@gis.a-star.edu.sg

YIJUN RUAN[1]
ruanyj@gis.a-star.edu.sg

**Correspondent author: V.A.K.*
[1] Genome Institute of Singapore, Biopolis street, 60, 138672 Singapore

Advances in high-throughput technologies, such as ChIP-chip and ChIP-PET (Chromatin Immuno-Precipitation Paired-End diTag), and the availability of human and mouse genome sequences now allow us to identify transcription factor binding sites (TFBS) and analyze mechanisms of gene regulation on the level of the entire genome. Here, we have developed a computational approach which uses ChIP-PET data and statistical modeling to assess experimental noise and identify reliable TFBS for c-Myc, STAT1 and p53 transcription factors in the human genome. We propose a mixture probabilistic model and develop computational programs for Monte Carlo simulation of ChIP-PET data to define the background noise of the sequence clustering and to identify the probability function of specific DNA-protein binding in the eukaryotic genome. Our approach demonstrates high reproducibility of the method and not only distinguishes bona fide TFBSs from non-specific TFBSs with a high specificity, but also provides algorithmic and computational basis for further optimization of experimental parameters of the ChIP-PET method.

Keywords: ChIP-PET, transcription factor binding sites, human genome, mixture probabilistic model, Kolmogorov-Waring process, Monte Carlo simulation

1. Introduction

Identification of gene regulatory elements for a given transcription factor is an important problem of computational genomics. The function of promoters, enhancers and other regulatory elements is mediated by DNA/protein interactions. The protein transcription factor binding sites (TFBS) serve as the basic units of gene functional activity. Computational prediction and high-throughput experimental validation of genome-scale sets of binding sites demands integrated approaches. Recently, great success has been achieved in the identification of TFBS for several essential regulators (p53 [9], c-Myc [2,11], STAT1 [3], p63 [10]) in human and Oct4, Sox2 and Nanog transcription factors (TFs) in mouse [6]. However, it has been difficult to identify all specific TFBS for several reasons. Currently available experimental information about the specificity of TF binding is essentially incomplete due to the difficulty of measuring the entire dynamical

range of avidities of large (and actually unknown) numbers of DNA binding sites for a given TF and high level background noises vs. signals.

A recent development of sequencing-cloning technology [8] entails the possibility of highly efficient and unbiased coverage of mammalian genome for large-scale identification of regulatory elements (Chromatin ImmunoPrecipitation Paired-End diTag, or in brief, ChIP-PET method). ChIP-PET provides a new powerful technique for localization of the most physically specific mammalian TF binding regions at a resolution of up to a few base pairs [6,9,11]. The software suite for comprehensive processing and managing of raw Paired-End diTag (PET) sequence data were recently described in [1].

Most unexpectedly, all studies using ChIP-PET data have shown that the TFs bind specifically to a surprisingly large number of genomic regions (extrapolated to 5,000-20,000 depending on the protein) [6,9,10,11]. Due to a large data volume, the major fraction of these TFBSs would not be validated by traditional experimental methods. Our knowledge about optimization of the relationship between the specific and noise events are still limited. Therefore, new mathematical and computational models are required in order to analysis of raw ChIP-PET data and correctly identify and predict specific TF binding regions and to optimize parameters of ChIP-PET method.

In this work, we present a probabilistic model of protein-DNA binding and computational simulations that model the ChIP-PET experiment concerned with specificity and sensitivity issues of TFBSs detection. We study the performance of a new analytical approach using ChIP-PET data for human p53, c-Myc, IFN-α induced STAT1 and IFN-γ induced STAT1 [9,11]. Finally, we discuss some problems that arise with the avidity function of TFBS when applied on the scale of the entire genome and with the functionality of revealed TF binding sites.

2. Data, Methods, Models, Algorithms and Software

2.1. Transcription factors

Transcription factor p53 regulates the expression of genes involved in a variety of cellular functions including cell cycle arrest, DNA damage repair, and apoptosis. ChIP-PET analysis of p53 binding in human colon cancer cells HCT116 was carried out as described in [9]. c-Myc is a proto-oncogene that regulates cell growth, cell proliferation, cell differentiation, and apoptosis [2]. In the ChIP-PET, we used human cell line that expresses high levels of exogenous c-Myc under the control of tetracycline [11]. STAT1 (signal transducer and activator of transcription) regulates proliferation by promoting growth arrest and apoptosis in response to interferon (IFN) signals [3]. ChIP-PET analysis of human cancer cells HelaS3 was carried out after treatment of these cells by IFN-α and INF-γ, as described in [3].

2.2. Basic Concept of ChIP-PET Method

Paired-End diTag (PET) method extracts a pair of 16-18 bp sequences from 5' end and 3' end of each cDNA clone, concatenates the PETs for efficient sequencing, and maps the resulting PET sequences to the genome. Such PET sequences characterize the ChIP enriched DNA fragments. Figure 1 shows a flow chart of ChIP-PET sequences

processing, mapping and clustering to the genome (using c-Myc library obtained from human P493 cells as the example) [11].

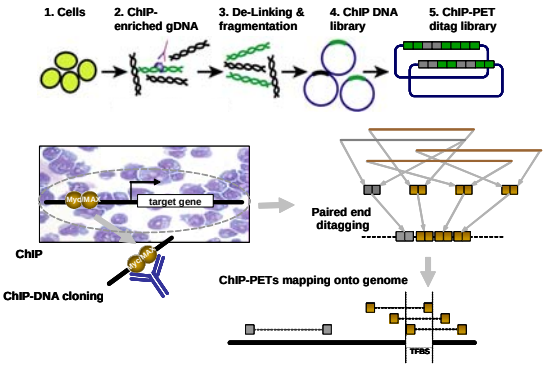

Figure 1. ChIP-PET. Top panel: Outline of the chromatin immunoprecipitation Paired-End diTag (ChIP-PET) method. Bottom panel: ChIP-PET analysis of c-Myc binding sites in human P493 cells. See details in [11]. Overlapping PET clusters define a more precise BS.

2.3. Definitions of DNA Fragment Cluster and Cluster Overlap

Let us define a *DNA fragment cluster* (or a cluster) as the overlapping PET DNA sequence fragments mapped to the genome (Figure 2A). More specifically, a PET sequence belongs to a cluster if it overlaps by at least 4 bp with any other sequence of the cluster in chromosome coordinates (Figure 2A). The number of PET sequences in a cluster is the *cluster size*. A total cluster span is defined as the genome region span covered by the cluster (Figure 2A). *The cluster overlap* is the most common PET DNA fragment in overlapping PETs in the given cluster. *The cluster member overlap count* (the peak) is the number of the overlapping PETs in a given cluster. The distribution of PET sequences within the cluster of sizes 3 and larger could be complex due to several cluster peaks (i.e., multimodal distribution of PET sequences). In this work, first, we count the highest peak (major mode) in the overlapping PET sequence cluster (Figure 2A). By examining the peaks observed for the rest of PET sequences in the cluster, we define the next highest peak and so on. To identify separate peaks in the given cluster, we use a strict definition of the cluster: every PET sequences in a cluster should overlap one another. To count the abundance of second peaks, we count the number of overlapped PET sequences excluding PET sequences from the first peak. If there are still sequences in the cluster, then we repeat the same procedure for third peaks and so on.

The number of cluster peaks occurrences is counted as the number of unique sequences containing at least one common nucleotide in a local PET sequence peak within a cluster. A cluster peak is more specific definition than a cluster overlap, because one cluster could contains more than one peak (local maximum of the sequence overlaps) and the peaks within a multimodal cluster could map the true protein-DNA interaction loci (Fig. 2A).

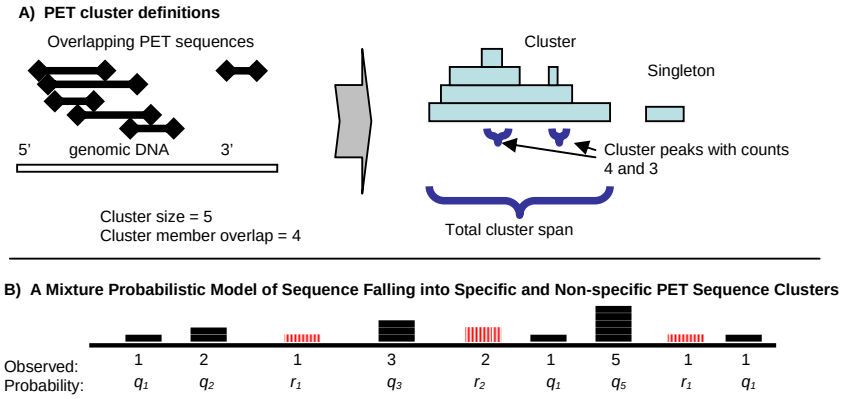

Figure 2. Two types of PET-clusters: definitions and statistical background of formation. A) Schematic example of sequence cluster, singleton, and cluster overlap. Cluster size is 5 (PET-5), and cluster member overlap is 4. Using strict criteria we define two clusters (peaks) by size 4 and 3. B) Schematic model of ChIP-PET sequences cluster overlaps on chromosome. q_1, q_2, q_3, q_5 are the binding probabilities for specific binding sites with avidity 1,2,3 and 5, respectively. r_1,r_2,r_3 are the probabilities of occurrence of sequence overlaps for non-specific PET sequences.

2.4. Characteristics of the ChIP-PET Libraries

In general, larger clusters (or peaks) often represent more specific binding sites [9,11] (see also Figure 2B). This correlation was observed for p53, ERE, c-Myc binding loci due to a concordance of the cluster (or picks) size with direct gene target expression data, direct qPCR-PET measurements and motif search analysis [9,11]. Nevertheless specific loci for the TFs could also be found in the smallest clusters (PET-2 peaks) and even in some singletons [11].

The number of PET sequences in the studied ChIP-PET libraries varied from 60 thousands (for p53 ChIP-PET library) to ~1 MB (for IFN-α activated STAT1 ChIP-PET library). The mean length of PET DNA sequences was in a range from ~400 bp to ~1700 bp (396 bp for c-Myc data; 623 bp for p53 data, 1385 bp for STAT1 data). Non-specific distinct PET DNA sequences represent a vast majority of PET sequences ranging from 75% to 95% of the total number of distinct sequences of ChIP-PET library.

2.5. Performance of ChIP-PET Data and Statistical Tasks

Significant amount of non-specific (background) genomic DNA is always present in the inmmunoprecipitated DNA material of ChIP-PET library. Some non-specific DNA might be easily filtered out after computer mapping of the DNA fragments on the genome [9]. Nevertheless background genomic DNA fragments that are uniquely mapped onto the genome still remain. With a larger sampling of DNA pool, the DNA fragments can be enriched by specific ChIP DNA sequences, and a larger number of true overlapping clusters might be observed. This sample size issue is related to the optimization of

performance of the method. We have preliminary analyzed an influence of variation of several parameters of the method (e.g, frequency distribution of the lengths of PET sequences, derived after sonication and fragmentation of DNA-protein complexes, avidity of specific immuno-precipitation binding, etc.) on quality of ChIP-PET libraries. We have recognized that, *sampling* and *erroneous sequence* are essential issues in the analysis and validation of ChIP-PET data.

Due to background noise and sampling errors, the following basic statistical tasks are becoming imperative: i) to estimate specificity of the ChIP-PET experiment, i.e. to predict the number of reliable TFBSs; ii) to assign quantity measure of reliability to every PET cluster overlap peak that forms a putative TF binding site; iii) to predict the total number of specific binding sites presented in the PET library.

Using several data sets on PET sequence mapping onto human genome presented in T2G database [1], we analyze these problems via probabilistic modeling and computational simulation of non-specific and specific binding sites loci for a given TF.

2.6. Distributions of PET Cluster Overlaps and Clusters

The number of PET sequences covering specific genome sites should roughly relate to site avidity of binding protein (Figure 2B). We assume that the distribution function of distinct cluster size (observed by number of PET sequences in a peak) could be modeled as a sum of distributions of specific and non-specific (background noise) clusters (peaks):

$$P_{obs}(X=m) = \alpha * P_{sp}(X=m) + (1-\alpha) * P_{ns}(X=m), \quad (1)$$

where P_{obs} is the probability distribution function of occurrence of a PET sequence cluster, X is the size of a given PET sequence cluster, m=1,2,3,... is the number of sequences in a cluster, P_{sp} is the probability distribution function of specific PET cluster occurrence, $0<\alpha<1$ is the fraction of specific clusters in the cluster population, P_{ns} is the probability distribution function of occurrences of the non-specific (background noise) cluster in the cluster population.

Based on ChIP-PET data, we could construct an empirical frequency distribution function of occurrence of PET clusters and corresponding PET cluster peaks. P_{sp} is related to the specific avidity of DNA-protein binding. We can estimate P_{sp} using the Generalized Pareto probability function [4,5], which can be derived from the Kolmogorov-Waring distribution function as an asymptotic solution [5]. We could estimate P_{ns} by a computer simulation of the non-specific sequence clustering model. We will discuss this model in the next section.

2.7. A Data-driven Model of Background Noise Sequence Overlapping

To simulate non-specific component P_{ns}, we propose a physical model dependant on the chromosome size, in particular, virtual PET sequences of the observed length randomly drop down into an interval equal to the chromosome size. The algorithm is as follows: 1) Virtual random sequences mapped on a given chromosome randomly drop down into sequence domain that equals to available sequence domain of the chromosome; 2) Virtual position in the chromosome was selected by a random number generator; 3) The

length of the virtual sequence is taken from the pool of observed PET sequences (we use empirical distribution of sequence lengths for our simulations (Figure 3)); 4) Virtual clusters are counted from overlapping virtual sequences (by at least 1 nt in the most common PET sequence overlap); 5) Y and M chromosomes and centromere regions of other chromosomes together PET sequences mapping these genome territories were excluded from modeling process.

The use of the observed length distribution is an important for modeling since longer sequences have a larger chance to form false clusters. An example of observed PET sequence length distribution can be found in Figure 3. It is possible to use a predefined fixed length for all virtual PET sequences (average length of observed PET sequences) [9]. We observed that such a model is oversimplified real data; in particular, it skips a number of large clusters (PET3+) which are important for unbias estimation of sensitivity of the method (Figure 3). Our data-driven noise model predicts a higher number of random clusters than the simplified model predicts.

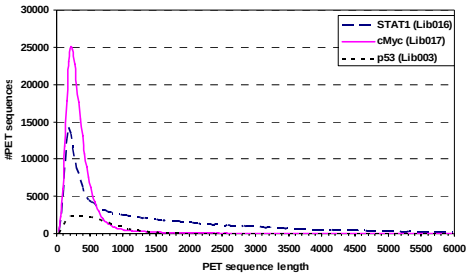

Figure 3. Observed PET sequence length distribution for p53, c-Myc and STAT1 libraries. The left size of the distributions could be fit well by the gamma distribution, however, the right long tail assumes more complex (mixture) distribution.

For the next step of estimation of model Eq (1), we performed a subtraction of the simulated distribution of random cluster size from the observed distribution of the cluster size. We estimated also a fraction (1-α) of non-specific sequences of model Eq (1) as a parameter to fit the observed frequency distribution of cluster size, assuming that frequencies of the non-specific clusters should be smaller than the one observed for each cluster size.

Notice. To overcome the problem of stochastic behavior, we fulfilled the simulation of random cluster formation many times (from 1 to 1000). Special attention was paid to quality of software for random number generation (RNG) [7].

2.8. A Model of Avidity Function of Specific Binding Site

We model the specific avidity distribution function using the truncated Generalized Discrete Pareto (GDP) function, which can be considered as a good limiting approximation of many random processes [4,5]:

$$f(m) := P_{sp}(X = m) = \zeta_J^{-1} \frac{1}{(m+b)^{k+1}} \tag{2}$$

where the *f(m)* is the probability that a randomly chosen specific BS has an avidity value *m*. The *f* involves two unknown parameters, *k*, and *b*, where *k*>0, and *b*>1; the normalization factor ζ_J is the generalized Riemann zeta-function [5], truncated in the interval [2, 200]. Eq(2) can be considered as asymptotic distribution function derived from Kolmogorov-Waring (KW) probability function [5]. This KW model could be used as possible exploratory model of aggregation TF on a DNA binding site. In particular, we could model the evolution of TF-DNA interaction as the random the random linear Kolmogorov process [5] of binding and detachment of TF on specific DNA binding sites taken to account at least two binding transition probabilities: due to specific "binding potential" (preferential attachment mechanism [5]) and "non-specific potential" (Poisson process mechanism [5]). Similar two processes but with different intensities are assumed for detachments transitions.

2.9. Analysis of Empirical Avidity Function of Specific Binding Sites

Due to our findings, the shape of the avidity probability function of TF-DNA binding on the genome scale should be described as skewed function of avidity (Figure 4). An example of observed avidity function for c-Myc binding sites defined by ChIP-qPCR method is presented in Figure 4A. We have also found that the distribution of avidity measured by qPCR as well as the tail of the empirical probability functions of cluster overlaps (Figure 4A, Figure 5) follow the GDP and correlate to each other (Figure 4B). One can see skewed distribution approached by the power law.

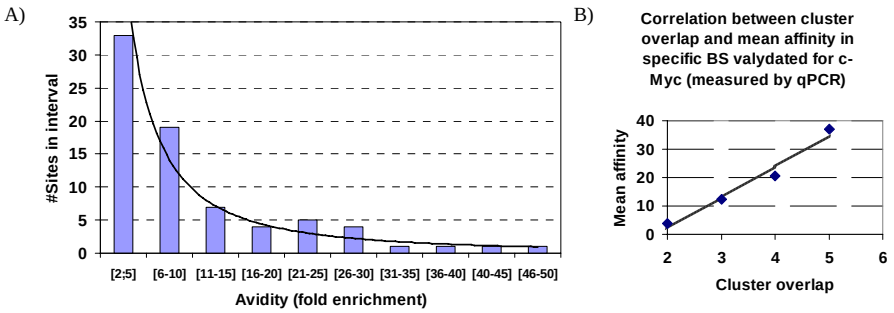

Figure 4. A) Avidity distribution function of PET sequences defined by ChIP-qPCR for c-Myc TFBS data. B) Correlation between ChIP-qPCR avidity and the number of sequences in specific cluster overlaps of c-Myc ChIP-PET library.

We argue that it is indeed a Pareto-like distribution. We have observed a similar avidity function for p53 [9] and Nanog transcription factor BS (not presented). Figure 4B shows a relation between the number of the sequences in cluster overlaps and the avidity value of c-Myc binding sites defined with ChIP-qPCR [11]. For 76 specific c-Myc binding sites in clusters, we found that the avidity of BS correlates with the number ChIP-PET sequences in cluster overlaps. The correlation coefficient between two data sets equals to 0.51 ($p<0.01$). Thus, we could use Eq. (2) as an empirical model of true avidity function of TF-DNA binding.

The avidity function of the ChIP-qPCR defining the specific TF binding sites can range from 2 to 200 fold enrichments [2]. Similar dynamical range was observed in our study of c-Myc mapping on the human genome [11].

Using this estimate, we calibrated enrichment fold from 2 to 200 and fitted the avidity function by the GDP function using the method presented in [5]. We constructed avidity function distribution for proposed number of binding sites that best fit the observed data based on minimal assumptions of statistical parameters of the distribution (data are not shown). For example, parameters *k* and *b* for *f(m)* function from Eq. (2) for c-Myc binding sites are equal to 3.4 and 1, correspondingly.

3. Results of Numerical Modeling

3.1. Goodness of Fit Analysis of Mixture Model and Estimations

To analyze the observed PET cluster peaks distribution, we parameterized the specific (Eq. 2) and nonspecific probability functions in Eq. (1) and estimated the parameter α. To accomplish that, we first used a goodness of fit analysis method presented below. Figure 5 shows an example of our tail-fitting and extrapolation method. We fit the theoretical (power law) function using only right (specific) part of the cluster size distribution (Figure 5A), and fit the proposed exponential function using only left (non-specific) part of the same distribution (Figure 5B).

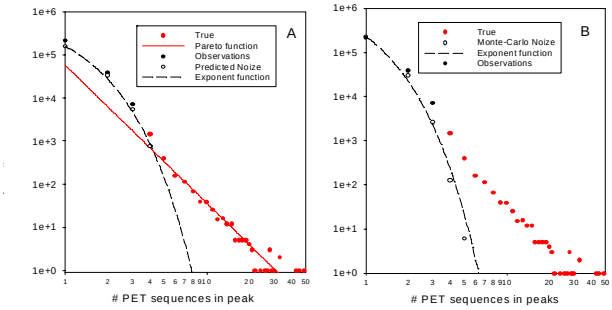

Figure 5. Decomposition of observed frequency distribution of the number of PET sequences in peaks in the STAT1 IFN-gamma activated library using (A) tail-fitting and extrapolation method and (B) noise peaks simulation and subtraction method. Red point: observed data; empty circles: restored background noise data; black circles data for the noise (nonspecific) peaks. Solid line on panel A: best-fit power law distribution with exponent parameter k=2.2 +/- 0.82 ($p<0.01$). Dashed lines: exponential function with slope parameter 1.75+/- 0.144 ($p=0.02$) and 2.32+/-0.010 ($p<0.01$) on panel A and panel B, respectively.

This figure shows the decomposition of the observed frequency distribution of the number of PET cluster peaks in the PET library for IFN-γ activated STAT1 (Lib016). To fit the parameters, we used Sigma-Plot software. In our model (Eq.1), best-fit power law distribution (Eq. 2) has exponent parameter *k*=2.2; the exponential function having a slope parameter 1.75 fits well to the non-specific (noise) component of the Eq.1. We assumed that all the cluster peaks of size greater than the cut-off value (3 or 4) are specific. By counting the total number of PET sequences associated with entire best-fit power distribution and the total

number of PET sequences associated with peaks of the observed distribution, we can also estimate the fraction of specific PET sequences in the library, α (Eq.1). For instance, for STAT1 data, α equals to 0.26 (Table 1).

For STAT1 IFN-γ activated ChIP-PET library, we used goodness of fit analysis [5](Figure 5) and computer simulations (Figure 6). Starting with a simulation model of

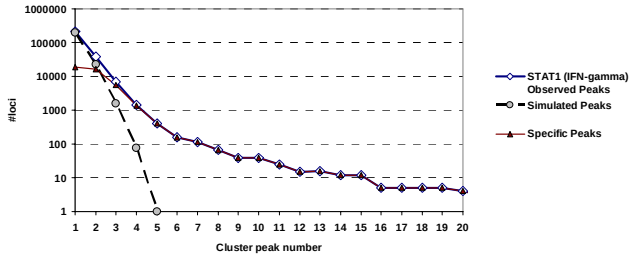

Figure 6. Observed and simulated noise distributions of cluster peak size for STAT1 library (Lib016). Specific cluster peak size distribution is estimated by subtracting the averaged simulated number of cluster peaks from observed number of cluster peaks. The noise distribution function was estimated by averaging of 20 Monte Carlo simulations. The fraction of noise PET sequences in our simulation was 75% of the STAT1 library size.

background noise distribution function, we can carry out a decomposition of the specific and non-specific components in Eq. (1). Figure 6 presents the noise peaks simulation and illustrates our subtraction method. To do that, we can calculate the distribution of non-specific (noise) cluster peaks and then after subtracting this distribution from observed distribution we could reconstruct the specific probability distribution function (Figure 6).

The both our fitting methods provide similar frequency distributions of the number of nonspecific (noise) PET sequence cluster peaks. Given the simulated and observed frequency distributions of the number of specific PET sequences within cluster peaks and subtracting one distribution from another, we can derive the true frequency distribution of the number of specific peaks, reporting here as the true TFBS loci (Figure 6). Thus, our results support the suggestion that the probability distribution of an avidity of specific binging sites follows a Pareto-like law while the non-specific components would be well approximated by exponential function.

We constructed also truncated cumulative function of the number of specific clusters (sum of number of clusters of size larger than given count). We used this cumulative function to estimate a fraction of specific TFBS loci for a given PET peak (Table 1). Due to estimated background cut-off value, this fraction equals to 1.0 for STAT1 PET6+ clusters (Table 1), that means that a fraction of non-specific cluster peaks larger than six is a negligibly small.

3.2. Specificity of PET Clusters

Based on our models, we estimate a cut-off critical value of cluster size which discriminates non-specific binding sites (noise) from specific binding sites in a given ChIP-PET library. The degree of reliability of every cluster in the library could be assigned

Table 1. Observed and simulated cluster peak size distribution for STAT1 IFN-γ activated library (Lib016).

PET count	1	2	3	4	5	6	7	8	9+
#Observed peaks	212447	39025	7103	1444	400	157	113	66	198
#Non-specific peaks	193528	22432	1565	75	1	0	0	0	0
Cumulative # observed peaks	260953	48506	9481	2378	934	534	377	264	198
Cumulative # specific peaks	43352	24433	7840	2302	933	534	377	264	198
Cumulative specific fraction	0.166	0.504	0.827	0.968	0.999	1	1	1	1

Table 2. Cut-offs of PET cluster size (95% level) and number of putative specific BS estimated by observed distributions.

TF	Total % of spec. PET	Cut-off value	% Spec. at cut-off	#BS at cut-off	% Spec. in PET2+	#BS in PET2+	#BS in PET1+
p53	5	3	98.6	284	53.2	936	1659
STAT1(IFN-γ)	25	4	96.8	2302	50.3	24433	43352
STAT1(control)	13	5	96.2	52	32.5	11735	18743
c-Myc	5	4	93.6	44	28.1	3683	9295

based on the probability that a cluster of the given size contains a binding site. Thus we can annotate every region in the clusters either as reliable (the probability that it contains BS is more than 99%), or probable (probability >95%), or potential (>70%).

We simulated background noise distribution by varying the number of specific PET sequences in the experiment from 0% to 30% and by selecting the best fit parameters. For p53 data we estimate the total percent of specific PET sequences in the ChIP-PET library to be at 5% (Table 2, second column). Comparing simulated and observed distributions of PET sequence cluster peak size, we can estimate the number of specific BS at any fixed cut-off value. For p53 data, for example, the cut-off value for cluster peak size equals 3 (third column) while specificity at this cut-off (number of specific BS in PET3+ cluster peaks) equals 284 (Table 2, fourth column). The cut off value of cluster peak size defines the specificity to be greater than 95%. The 6-th and 7-th columns in Table 2 demonstrate the specificity in PET2+ cluster peaks. The specificity is much lower for p53 BSs (53.2%), but the estimated total number of BSs in PET2+ cluster peaks is much higher (936 BSs for p53, Table 2). The total number of putative BS predicted by the PET singletons and clusters is estimated to be 1659. This is an estimate of the low limit of the total number of specific p53 BSs in the human genome. Our extrapolation of the GDP model predicts >5000 p53 BSs, most of which, however, should be very low-avidity BSs and therefore should be not functional in a cell.

The cut-off values of cluster peak size are relatively large for STAT1 and c-Myc libraries (Table 2). The total number of specific binding sites in the human genome is estimated to be as large as up to 43000 for IFN-γ induced STAT1 TF. By our computational simulations, the smallest number of specific BS estimated represented by the c-Myc library is ~ 9000. Base on extrapolation of the best-fit model (1), however, the total number of specific c-Myc BSs in the human genome should about 20000 or even larger [11].

4. Discussion and Conclusion

We have developed a computational method to estimate the number of specific binding loci in the genome for transcription factors studied in ChIP-PET experiments. Our probabilistic model of ChIP-PET clusters permits an accurate estimation of parameters of TFBS avidity function based on ChIP-PET experiment. The model explicitly uses information about the length of the ChIP-PET DNA fragments, the number of PET sequences in the library and the chromosome length. The summary is as follows:

1. We developed a statistical method for selecting specific **a** component in observed PET cluster sites distribution based on a novel mixture distribution model.
2. We developed a Monte Carlo simulation model and a program for non-specific binding events (background noise) in ChIP-PET cluster size distribution.
3. This computational model provides cut-off values for specific TFBS and supports estimates obtained by the goodness of fit method [9,11].
4. We have shown that the true and noise probability distributions of loci avidity are scale-dependent skewed functions: when the library size and/or the average sequence span become larger, the shape of each distribution changes. In particular, the tail of the distributions of occurrence of true BSs becomes longer.
5. The probability distribution of specific avidity of DNA-protein interactions for different TFs in mammalian cells can be described by a generalized Pareto function, while the random probability distribution is fitted by the exponential function.

Significant correlation between ChIP-qPCR avidity and the number of sequences in specific cluster overlaps of ChIP-PET data suggests that the size of specific PET sequence clusters could indeed reflect of DNA-TF avidity. Validation of predicted “true” BS was shown for p53 and c-Myc binding sites in PET3+ clusters and verified by qPCR experiments [9,11]. We note that the binding specificity of TF sites defined by our model does not mean that the functionality of the BS in a given cell. The DNA-TF avidity is really very complex and skewed function (Eq (2), Figure 4) of many factors. Other factors such as the state of chromatin remodeling, epigenetic factors (acetylation and methylation of histones and DNA) may affect both ChIP-PET and ChIP-qPCR outcomes. Recently, Abcam ChIP Grade antibodies in combination with the new Solexa 1G sequencing technology have been used to identify the patterns of histone methylations on the human genome scale [13]. We assume that the computational modeling of binding avidity of specific TF BSs and mapping histone methylation regions could provide better understanding of TF binding and epigenetic control of DNA-TF avidity. It is interesting to note that recent ChIP-seq (ChIP-sequencing) methods have revealed 41582 putative STAT1-binding regions in IFN-γ stimulated HeLa S3 cells [12]. However, avidity of the vast majority of these BSs due to our model should be very low and perhaps these low-avidity BSs are unlikely functional. The addition of sequence information (consensus or weight matrix for given protein binding site) and genome information (as proximity to promoter regions, CpG islands) could significantly improve the quality of estimations for the binding sites [9,11,12]. ChIP-PET technology as well as ChIP-on-chip, ChIP-seq and STAGE (Sequence Tag Analysis of Genomic Enrichment) technologies [3,12] in combination with simulation analysis will allow us to identify thousands of novel binding sites in the human genome. A challenging statistical problem of estimation of specificity and sensitivity of these methods could be solved using the approach suggested in this work.

Acknowledgments
The authors would like to acknowledge Dr. Chiu Kuo Ping for useful discussions. This work is supported by the A*STAR of Singapore.

References

[1] Chiu, K.P., Wong, C.H., Chen, Q., et al, PET-Tool: a software suite for comprehensive processing and managing of Paired-End diTag (PET) sequence data, *BMC Bioinformatics*, 7:390, 2006.
[2] Fernandez, P.C., Frank, S.R., Wang, L., Schroeder, M., Liu, S., Greene, J., Cocito, A., and Amati, B., Genomic targets of the human c-Myc protein, *Genes Dev.*, 17(9):1115-29, 2006.
[3] Hartman, S.E., Bertone, P., Nath, A.K., et al. Global changes in STAT target selection and transcription regulation upon interferon treatments, *Genes Dev.*, 19(24):2953-68, 2005.
[4] Johnson, N.L., Kotz, S., and Balakrishnan, N., *Discrete Multivariate Distributions*, John Wiley&Sons, Inc., New York 299 p., 1997.
[5] Kuznetsov, V.A., Family of skewed distributions associated with the gene expression and proteome evolution, *Signal Processing*, 83:889-910, 2003.
[6] Loh, Y.H., Wu, Q., Chew, J.L., et al. The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells, *Nat Genet.*, 38(4):431-440, 2006.
[7] Matsumoto, M., and Nishimura, T., Mersenne Twister: A 623-dimensionally equidistributed uniform pseudo-random number generator, *ACM Transactions on Modeling and Computer Simulation*, 8:3-30, 1998.
[8] Ng, P., Wei, C.-L., Sung, W.K., et al., Gene identification signature (GIS) analysis for transcriptome characterization and genome annotation, *Nat Methods*, 2:105-111, 2005.
[9] 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): 207-19, 2006.
[10] Yang, A., Zhu, Z., Kapranov, P., McKeon, F., Church, G.M., Gingeras, T.R., and Struhl, K., Relationships between p63 binding, DNA sequence, transcription activity, and biological function in human cells, *Molecular Cell*, 24:593-602, 2006.
[11] Zeller, K.I., Zhao, X., Lee, C.W., et al. Global mapping of c-Myc binding sites and target gene networks in human B cells, *Proc Natl Acad Sci U S A*, 103(47):17834-17839, 2006.
[12] Robertson, G., Hirst, M., Bainbridge, M. et al., Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing, *Nat Methods*. 4(8):651-7, 2007.
[13] Barski A., Cuddapah S., Cui K. et al., High-resolution profiling of histone methylations in the human genome. Cell. 129(4):823-37, 2007.