

Chapter 2

Relative Avidity, Specificity, and Sensitivity of Transcription Factor–DNA Binding in Genome-Scale Experiments

Vladimir A. Kuznetsov

Abstract

One of the most crucial problems with genome-wide experimental analysis is how to extract meaningful biological phenomena from the resulting large data sets. Here, we present modeling and prediction techniques that are applied to genome-wide identification of in vivo protein–DNA binding sites from ChIP-based data sets. We develop a simple mixture probabilistic model of occurrence of non-specific and specific TF–DNA binding events for transcription factor binding to any site in the genome. We calculated the statistical significance of specific and non-specific random binding events using Kolmogorov–Waring and exponential functions, respectively. The binding events in the chromosome regions associated with non-specific, non-random binding loci were also identified and filtered out. The mixture model fits equally well to five different TFs (ERE, CREB, STAT1, Nanog, Oct4) data provided by ChIP-PET, SACO, and ChIP-Seq methods included in this study. We present a uniform methodology for estimating specificity, total number of binding sites, and sensitivity of data sets detected by these ChIP-based genome-wide experimental systems. We demonstrate strong heterogeneity of specific TF–DNA binding sites in terms of their avidity and by correlation between observed relative binding avidity of specific TF–DNA binding site and the level of mRNA transcription of the nearest gene target. Finally, we conclude that the sensitivity problem has not been resolved by current ChIP-based methods, including ChIP-Seq.

Key words: Transcription factor, avidity function, binding sites, mixture model, specificity, sensitivity, ChIP-PET, ChIP-Seq, SACO.

1. Introduction

1.1. Protein–DNA Interactions In Vivo and Its Relative Avidity

Identification of interactions of gene regulatory elements (e.g., evolutionarily conserved genome sequences, anti-sense transcripts) is an important problem in molecular biology and functional genomics. Among those elements, the protein transcription factor binding sites (TFBSs) are considered to be the basic units of

functional gene activity (1–4). A transcription factor is a sequence-specific DNA binding protein that binds to specific segments of DNA (binding sites (BSs), motifs) (5, 6). These binding proteins are a part of the cellular system that controls the transcription of genetic information from DNA to RNA, establishing the gene expression patterns that can determine cellular programming (7–9). About 10% of proteins in complex organisms such as humans are produced as a result of transcriptional activity in the genome of cells. Transcription factors are the largest regulatory set of proteins in the cell. One TF could physically bind and functionally control hundreds and thousands of specific BSs in cells of complex organisms. Moreover, some DNA motifs have the potential to serve as targets for different TFs, and different motifs are often clustered in a relatively small promoter (upstream gene) and genic regions. TFs are also capable of physically competing and/or synergistically interacting with each other and BSs (10–12).

For a given TF, the protein–DNA interactions *in vivo* link transcription factors (TF) with its direct DNA target to form a gene–protein “interactome” and corresponding TF–DNA binding network scaffold for transcription regulation. Links in this simplified network are often defined by a ranking system corresponding to the relative avidity values of the TF–DNA binding events (13, 14). However, the binding avidity potential of different genome loci for the same TF is not a constant function; it can vary within a genome by several orders of magnitude (14) and can depend on many determining factors including sequence composition of BS motif, location of BS in the genome domain, cell differentiation, physiological conditions, and environmental factors. In fact, due to biological complexity and detection limitation, information about relative binding avidity and total number of TF binding sites for TF in a given cell type is unknown. Usually, only high-specific TFBSs with relatively high binding avidity have been identified and biased by the prediction methods used.

1.2. High-Throughput Methods for Determining the DNA Binding Events

The TF–DNA interactions can be detected by chromatin immunoprecipitation (ChIP), and the power of this technique is in its ability to analyze protein–DNA interactions *in vivo* (15–22) (**Fig. 2.1**). In ChIP experiments, TFs are cross-linked with DNA while an immune reagent (antibody) specific to a DNA binding factor is used to enrich target DNA fragments to which the TF was bound in a living cell. The bound DNA fragments overlapped and enriched with TFBSs are then identified and mapped on reference genomes and can be quantified to produce additional computational results.

Historically, the first technical platform to conduct wide-scale TF–DNA experiments was ChIP-on-chip DNA microarrays that tiled significant regions of the genome ((23), *see* also references in (2)). In ChIP-on-chip experiments, the copy number of DNA

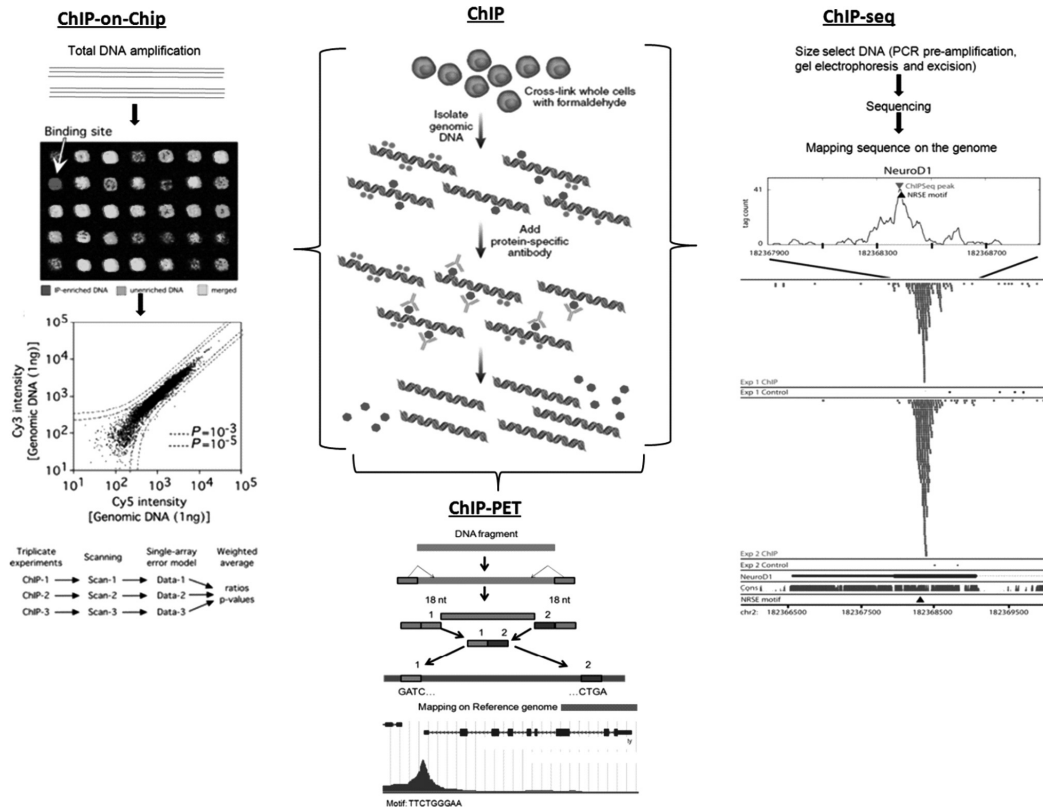


Fig. 2.1. Workflow of ChIP-based experiments for the study of TF–DNA binding sites in living cells: DNA and proteins are cross-linked and purified, then bound DNA is analyzed by ChIP-on-chip and massively parallel short-read sequencing, ChIP-PET and ChIP-Seq (1, 4, 15, 18, 19).

segments associated with TF of interest is compared to a reference sample, generally either genomic DNA or any DNA that might be immunoprecipitated with a negative control antibody (Fig. 2.1). Starting with a biological question, a ChIP-on-chip experiment can be divided into three major steps: First, to set up and design the experiment by selecting the appropriate array and probe type. Second, to conduct actual experiment in a wet lab. In the lab, the ChIP-purified DNA is amplified, fluorescently labeled, and hybridized to DNA sequences along with a fluorescently labeled control DNA sample usually corresponding to total genomic DNA. Array elements that correspond to the specific genomic-binding sites for the protein can be identified as those that display significantly stronger fluorescence signal in the ChIP DNA channel compared to the control. The last step involves gathering the data to be computationally analyzed with appropriate statistical test(s) for detection of significant ChIP-enriched DNA targets. Although over the years ChIP-chip approaches have significantly improved and have greatly expanded our understanding of gene-wide TF–DNA interactions, it seems difficult to make ChIP-chip

analyses affordable, reliable, and highly sensitive on the complex genome-wide scale (15–19). In particular, several technological drawbacks with this method include complications in array hybridization and probe design, low resolution of the BS location, and experimental standardization.

Another way to achieve genome-wide identification of protein–DNA interactions is to adapt high-throughput DNA tag sequencing for analysis of chromosome mapping of ChIP DNA. Serial analysis gene expression (SAGE) is a short sequence tag method which was originally developed to analyze transcriptome profiles (2). Several groups have modified the original SAGE protocol to isolate sequence tags from ChIP DNA and construct libraries of DNA tags for large-scale tag sequencing (21, 24). For example, SACO (21) combines chromatin immunoprecipitation (ChIP) with a modification of SAGE. This method has the potential to semi-quantitatively interrogate an entire metazoan genome by combining ChIP with a modification of long serial analysis of gene expression (Long-SAGE); a method normally used for transcriptome analysis (2, 21). By sequencing many thousands of concatemerized 21 bp genomic signature tags (GSTs) generated from anti-TF ChIPs, a genome map of TF binding sites can be identified and quantified. These and next-generation short-tag sequencing technologies (15–17) used to analyze protein–DNA fragment binding released after ChIP-chip have distinct advantages over standard microarray hybridization approaches. In particular, chromatin immunoprecipitation paired-end-ditag (ChIP-PET) (4, 9) and ChIP-Seq methods (15–17) entail the possibility of a highly efficient process with a potentially unbiased coverage of mammalian genome for large-scale identification of regulatory elements (promoters, enhancers, hyper-methylated regions, and other regulatory elements) mediated by DNA/protein interactions.

Current ChIP-based paired-end ditags ChIP-PET technologies are capable of producing up to 1 or 2 millions of sequence reads during each instrument run (4, 9). The advantage of using PET over single tags is that the PETs mark the start and end of each ChIP fragment. When PET fragments are mapped to the reference genome, the identity of each individual ChIP fragment can be inferred by the PET mapping location, and binding sites can be accurately defined by the common regions within clusters of overlapping PETs. Furthermore, duplicate PET fragments arising from fragment amplification events during cloning can be easily distinguished and removed by treating these multiple PETs that map to an identical location as a single fragment. It has been demonstrated that ChIP-Seq method provides the most powerful short-tag sequencing technique for accurate localization of the physically specific mammalian TF binding regions at a resolution of up to a few base pairs (16, 17).

Nevertheless, detecting TF–DNA interactions by ChIP-based sequence tag methods remains fraught with difficulties because it involves multiple and non-linear experimental steps, sampling procedures and unique data analysis methods. Our knowledge about optimization of the relationship between the specific and noise binding events, sampling errors defined by new technologies are still very limited. Difficulty in discerning a successful experiment from a failed one and in choosing appropriate data analysis methods is often a hurdle.

**1.3. Importance
of Statistical and
Computational
Bioinformatics
Analyses of Protein–
DNA Interaction Events
on the Genome-Wide
Scale SAGE-Coupled
ChIP Assays**

In many cases, it is desirable to know the *specificity* and *sensitivity* of genome-wide measurements of transcription factor–DNA binding events. If one has prior knowledge of a set of all transcription factor binding sites and sequences not bound by the factor in question, then conventional calculation of specificity and sensitivity of genome-wide TF-binding events is straightforward. However, in the absence of such knowledge, one needs to rely on statistical analysis, data-driven physical models, and computational predictions using currently available high-noisy and essentially incomplete DNA fragment samples. For example, one can rank the identified target sites based on the number of tags in a cluster of DNA fragments or in the cluster overlap (peak value of the DNA fragment cluster), and split the genomic regions into non-overlapped blocks and count the frequency of events considered above binding. After ranking the values of such “binding events”, this frequency information can be presented in a form of an *empirical* frequency distribution function which is an essential starting point for any further statistical analysis of data and planning of validation studies (13, 14). This function has been used to identify the adequate statistical models required in order to perform appropriate statistical analysis catered to different types of genome-wide sequence data sets or prediction of specific TF binding regions (4, 10, 14). This leads to an estimation of the sensitivity of genome-wide transcription binding events using incomplete samples (13).

The objective of this work is to develop a basic statistical model and computational estimation approach to quantitatively analyze different types of genome-wide tag-based TF–DNA binding experiments. We specifically propose a mixture probabilistic model of non-specific and specific TF–DNA avidity function. Our model estimates relative avidity function of TF–DNA binding. We also summarize the findings of a newly developed procedure which could be used to estimate specificity and sensitivity of genome-wide tag-coupled ChIP assays (SACO, ChIP-PET, and ChIP-Seq). We develop an uniform approach for quantitative analysis of such experiments which allows (i) to identify a reliable set of TFBS events in the experiment, (ii) to reconstruct a relative binding avidity function of TF–DNA interactions using essentially incomplete and high-noisy data, (iii) using such data to predict the

total number of real BSs for a given TF in mammalian genomes (e.g., rat, human), and (iv) to compare different ChIP-based tag sequencing approaches by uniform statistical parameters. In the end, we validate our results using TFBS motif search/prediction algorithms and microarray expression data.

2. Definitions, Models, and Methods

2.1. DNA Fragment Cluster, Cluster Overlap, and Cluster Peak

In ChIP-Seq, a ChIP-enriched SAGE-like tag is represented by either a single internal 21 bp tag sequence (SACO), by a 27 bp tag sequence (ChIP-Seq), or by a ~ 36 bp paired-end ditag (ChIP-PET in which the ditag is constructed from 18 bp 5' and 3' signature sequences extracted from each end of the ChIP DNA fragment). Thus, SACO and ChIP-Seq demarcate a single end of the length of the sonicated ChIP DNA fragment and ChIP-PET full length of the sonicated ChIP DNA fragment, respectively. The binding sites are then deduced based on the frequency with which tags in a given genome locus are extracted from ChIP DNA fragments relative to the background computational expectation or background control data.

A distinctive feature of *the binding event* defined by any large-scale ChIP-based technology is *the DNA fragment cluster*. Due to the highly dimensional and multivariate nature of TF–DNA binding data, cluster should be (i) accurately located (mapped) in a defined chromosomal region of the reference genome, (ii) additionally characterized as a putative binding site, and (iii) appropriately counted. Due to the fundamental properties of current high-throughput technology concepts, an identification of length of DNA sequences and aggregation procedures of the sequences into clusters are not addressed and hence technology specific. For example, the SACO method relies on the observation of 21 bp DNA fragments for mapping of specific region and forms clusters that identify and quantify the TF binding site by count of such GSTs in the “SACO cluster” as “any collection of GSTs that are within 2 kb of each other” (21). Additionally, most of SACO loci are confirmed by identification of chromosome location of that putative TFBSs near or within “transcriptional open regions”.

Different definitions of the cluster and corresponding TF–DNA binding event are used by ChIP-Seq (15, 16). For the ChIP-Seq TF–DNA binding data, called ChIP-Seq reads (SETs), we are considering the “extended and overlapped” DNA fragment clusters formed by distinct DNA fragments (i) which are overlapped due to computational extension of the original 27 nt sequence into a 174 nt (15, 16) extended SET (XSET) and (ii) which share

common loci (at least 4 bp). These overlapped clusters (observed as local peak heights on a genome coordinate) can provide a statistical evaluation of the number of TF binding events in the entire genome as represented by the ChIP-Seq sample.

According to ChIP-PET (4, 9), when PET DNA sequences share the same locus (four or more common nucleotides in our study) on the same chromosome region, they are recognized as a cluster and overlapping PET DNA sequences and corresponding chromosome loci are called the cluster overlap (Fig. 2.2, Section 2). A chromosome locus of the cluster overlap can contain TFBS region and the number of overlapped PET DNA sequences in that cluster overlap region can represent a semi-quantitative measure of relative avidity of BS (*see* below). If more than one statistically confident overlapped PET DNA region is included in the PET DNA sequence cluster, the cluster overlap region containing the largest number of sequences (largest peak) is counted as the binding event associated with the cluster.

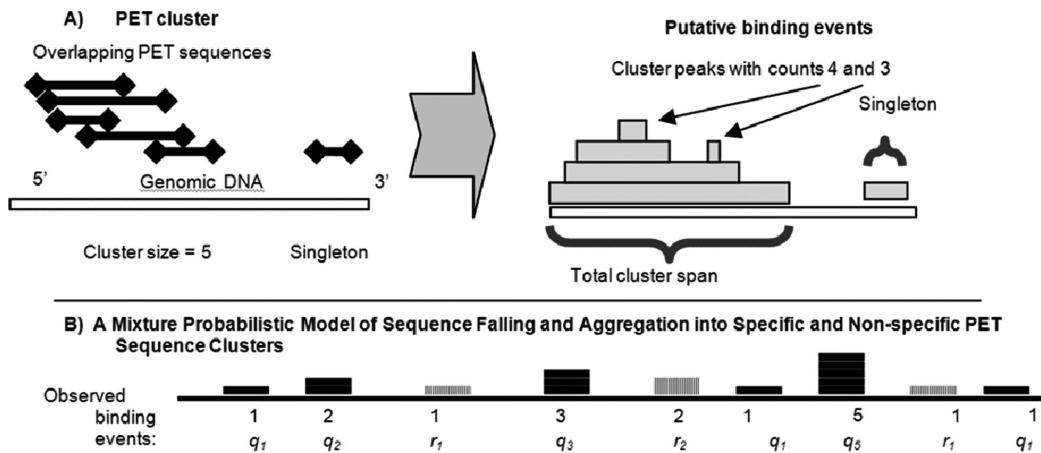


Fig. 2.2. Determination of binding events defined after ChIP-PET DNA fragment mapping onto genome. **(A)** Schematic example of sequence cluster, singleton, and cluster overlap (representing putative binding event). Cluster size is 5 (PET-5) and cluster member overlap (peak height) is 4. Using strict criteria we define two clusters (peaks) by sizes 4 and 3. **(B)** Illustration of a concept of a mixture frequency distribution of specific q_1, q_2, q_3, q_5 (black rectangles) and non-specific r_1, r_2 (gray rectangles) probabilities of binding events for a given TF. The down index of the symbols indicates the avidity potential value, presented also graphically by height of the rectangle. For simplification aim, we consider the avidity potential as the discrete function. Note that currently available ChIP-based sequencing method provides the samples which can be essentially incomplete due to limited deepness of available sequencing reads and high-noisy issue. DNA sequence overlaps can provide chromosome location of the most high-avidity binding sites; however, due to limited sample size and noise signals, the moderate- and low-avidity specific BSs could not be reliably detected (14) (*see* also Notes and Results sections).

2.2. A Mixture Probabilistic Model of the Distributions of TF-DNA Binding Events

The number of sequences covering specific genome loci should roughly relate to binding site avidity of TF-DNA complex (Fig. 2.2B). We assume that the probability distribution function of binding events (e.g., the number of PET sequences in a distinct

peak) could be modeled as a simple sum of the distributions of specific and non-specific (background noise) clusters (distinct peaks):

$$P(X = m) = \alpha P_{\text{sp}}(X = m) + (1 - \alpha) P_{\text{ns1}}(X = m), \quad [1]$$

where P is the mixture probability distribution function of occurrence of specific and random non-specific binding events, X is the number of binding events in the genome, $m = 1, 2, 3, \dots$ is the value of binding events, P_{sp} is the probability distribution function of specific binding events, $0 < \alpha < 1$ is the fraction of specific binding events, and P_{ns1} is the probability distribution function of occurrences of random non-specific (background noise) and/or “low-avidity” binding events. The parameter α can be estimated as a fraction of sequenced DNA fragments of the non-specific (false-positive) DNA fragment clusters or cluster overlaps observed in DNA fragment mapped uniquely onto genome.

Based on our analysis of ChIP-PET, ChIP-Seq, or SACO data sets, we could construct an empirical frequency distribution function of binding events (count of clusters/cluster overlaps/cluster peaks) in a given library. P_{ns} is the probability distribution function that describes low-specific and/or non-specific binding events, which on mass is mostly represented by singleton sequences and low-height peaks. We model P_{ns} using the exponential probability distribution function. We also model this function using Monte Carlo simulated frequency distribution of random clusters (and random cluster overlaps) constructed after random localization of ChIP-enriched DNA fragment set onto available regions of the human genome by sampling the mapped DNA fragments from uniform distribution (4, 14). We model P_{sp} using the Kolmogorov–Waring and the generalized Pareto probability functions (14). We estimated the parameters of the functions P_{ns} and P_{sp} and the weight parameter α using algorithm published in (25). The attributes of noise and the specific components of our mixture model will be presented in **Section 3**.

Notably some non-specific DNA fragment clusters can also be found among highly abundant and reproducible clusters and/or cluster overlaps with large enough heights (**Fig. 2.3**). These clusters contain a minor fraction of DNA fragments of ChIP-based library (by our estimates $\sim 3\text{--}7\%$ of all sequences mapped onto genome). Clusters with the problematic or “impossible” TFBS locations (mitochondria, Y chromosome in female cells, centromeric regions, no gene within 100 kb vicinity putative BS location, near genome gap localization, etc.) are seriously considered as an important source of systematic errors. In this case, more general mixture model could be considered:

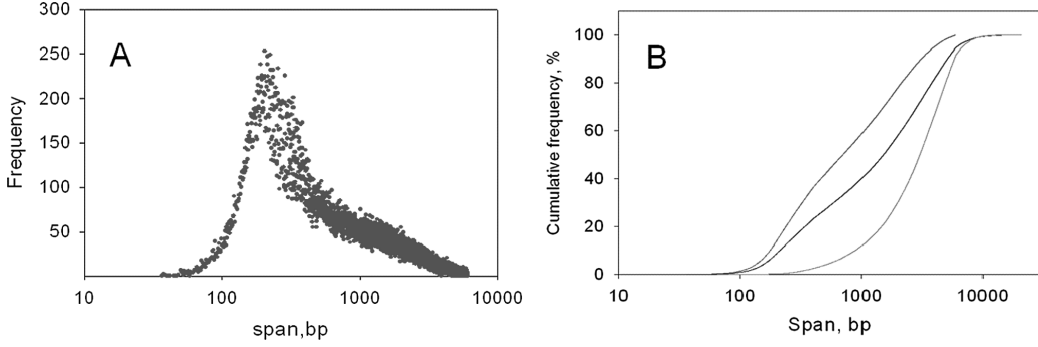


Fig. 2.3. False-positive “significant” clusters in ChIP-PET data sets can be occurred due to suboptimal ultrasound-generation of DNA fragments with very long and very short PET DNA fragment spans. **(A)** An example of the frequency distribution of span of PET DNA fragments found in INF- γ -induced STAT1 ChIP-PET binding event data set. This data contains a large fraction of long DNA fragments. **(B)** The cumulative functions of the spans of PET DNA fragments in the total set of DNA fragments which uniquely mapped onto reference genome (used in panel **A**), the spans of singleton PET DNA fragments sequences, and the spans of PET DNA fragments formed the clusters with size 2 and larger (PET-2+).

$$P(X = m) = \alpha P_{sp}(X = m) + \beta P_{ns1}(X = m) + (1 - \alpha - \beta) \times P_{ns2}(X = m),$$

where the probability functions $P_{ns1}(X = m)$ and $P_{ns2}(X = m)$ are the probability of non-specific almost random and non-specific systematic errors, respectively, and α, β are unknown weight parameters. In our analyses P_{ns2} can be defined as false-positive clusters by some rules and excluded at the prefiltering stage of data processing (*see Section 2*).

2.3. Empirical Model of Avidity Function of Specific Binding Events

We define a ChIP-enriched DNA fragment *library* as a list of ChIP-enriched DNA fragments which uniquely map the genome and contain distinct tags (ditag in the case of ChIP-PET method). The size of a library, M , is the total number of distinct DNA fragments observed in the library and uniquely mapped to the genome. Let $n(m, M)$ denote the number of ChIP-enriched DNA fragment sequences in cluster overlap, which have peak height m (no. of distinct sequence tags with common genomic locus) in the library of size M . Let J denote the maximum observed peak height of extended and overlapped cluster of the in the ChIP-enriched DNA fragment library. Let N denote the number of specific binding events:

$$N = \sum_{m=1}^J n(m, M). \quad [2]$$

Let c denote the given confidence threshold of binding event in the library ($c > 1$). This threshold could have different value in cases of model prediction and of experimental estimates. Let M_c be the cumulative mass (or number of distinct ChIP-enriched DNA fragments) observed at the given threshold c :

$$M_c = \sum_{m=c}^J n(m, M)m.$$

where $c \geq 1$. At the given specificity cutoff value (no. of observed binding sites),

$$N_c = \sum_{m=c}^J n(m, M).$$

Several classes of skewed probability functions (Poisson, exponential, standard power law, lognormal, logistic functions) are available to fit the empirically defined BS events. After performing goodness-of-fit analysis using the method presented in (25), we found the best fit was obtained by generalized discrete Pareto (GDP) function. We model the specific avidity distribution function using the truncated GDP function, which can be considered as a good asymptotic approximation of many random processes (13, 25):

$$f(X = m) = z_J^{-1} \frac{1}{(m + \beta)^{(k+1)}}, \quad [3]$$

where the random variable X is the discrete value of binding avidity in a given BS (e.g., the peak height of the ChIP-Seq DNA fragment cluster); $m = 1, 2, \dots, J$; $f(m)$ is the probability that a randomly chosen specific BS has an TF-DNA binding avidity value m . The parameter f involves two parameters, k and β , where $k > 0$ and $\beta > 0$; the normalization factor z is the generalized Riemann zeta-function value: $z_J = \sum_{m=1}^J 1/(m + \beta)^{(k+1)}$ ($m + \beta)^{(k+1)}$ (26). The parameter k characterizes the skewness of the probability function; the parameter β characterizes the deviation of the GDP function from a simple power law. Since in log-log plot this function exhibits a positive curvature and changes its slope when the sample size is changed (14, 25), function [3] is the so-called “scale-dependent network” statistical model (27).

2.4. Explanatory Relative Avidity Model: The Kolmogorov– Waring Function

Equation [3] can be considered as asymptotic distribution function derived from the Kolmogorov–Waring (K–W) probability function, $P_{K-W}(X = m)$, where $m = 0, 1, 2, \dots$ (25). The K–W function has been derived in (25) and successfully used in modeling different types of events on genome, transcriptome, and proteome scales (14, 25, 27). We assume that P_{K-W} could also be used as a possible exploratory model of stochastic aggregation–dissociation TF on a specific DNA binding site. For statistical analysis of ChIP-based experiments, we consider the occupation of a BS by a TF as the series of Markov random chain events realizing TF-DNA interactions in terms of the random linear birth–death Kolmogorov process (25). Here, we specified this model for analysis of TF-DNA binding and dissociation processes.

In our model, we specified two aggregation transition probabilities: (i) “specific binding” due to preferential attachment mechanism of TF to specific DNA region on chromosome and (ii) “non-specific” binding driven by the Poisson process. We assume similar two types of processes for TF–DNA detachment transition events. However, the specific and non-specific dissociation processes could be realized with different intensities.

The exact steady-state solution of such binding–dissociation process can be described by Kolmogorov–Waring probability function, which is calculated via the following simple recursive formula (25):

$$\eta_m = p_{m+1}/p_m = \theta \frac{(a + m)}{b + m + 1}, \quad [4]$$

where $m = 0, 1, 2, \dots$ and the other three parameters a , b , and θ are unknown. Importantly, the K–W probability function allows us to estimate the value p_0 , which gives the fraction of lost (undetected) events in a given TF–DNA binding experiment.

$$p_0 = \frac{1}{{}_2F_1(a, 1; b + 1; \theta)} > 0,$$

where ${}_2F_1$ is the hypergeometric Gauss series (26). In this case

$$p_0 = \left(1 + \sum_{m=1}^{\infty} \left(\prod_{i=1}^m \frac{(a - 1 + i)}{(b + i)} \theta \right) \right)^{-1}.$$

Specifically, if $b > a > 0$ and $\theta \rightarrow 1 - 0$, then by (25)

$$\lim_{\theta \rightarrow 1-0} p_0 = \left(1 - \frac{a}{b} \right). \quad [5]$$

This formula can be used for extrapolation of the K–W model up to $m = 0$ (undetected event). By following a recursive formula we can estimate the TFBSs at each peak height intensity level as the following:

$$p_1 = p_0 \frac{a}{b + 1} \theta, \dots, p_{m+1} = p_m \frac{a + m}{b + m + 1} \theta. \quad [6]$$

Importantly, GDP function can be a fairly accurate approximation of K–W function throughout the entire dynamical range of random variable X ($m = 1, 2, \dots$) (25). We use this attribute of K–W probability function for (i) goodness-of-fit analysis of the model, (ii) estimation of specificity and sensitivity of studied of ChIP-based data set, and finally (iii) estimation of the total number of physically real specific BSs for a given TF in a given genome.

2.5. Prediction Method of the Total Number of Binding Events (BEs)

Here, we fit the truncated empirical distribution of binding event (BEs) (e.g., ChIP-Seq peak height value), starting the fitting of the specific part of the empirical distribution (equations [3] and [4]) and estimating the “empirical” threshold which provides minimum noisy component. Then after parameterization of specific and non-specific probability terms of the mixture probability function in the model [1] and estimating weight parameter α , we extrapolate the best-fit probability function of specific events into a noise-enriched event region of the empirical distribution to predict the entire specific frequency distribution of specific TFBSs in the given ChIP-based experiments. Using equation [5], we could estimate the total number of TFBSs in the entire genome.

2.6. SACO, ChIP-PET, ChIP-Seq Methods and Characterization of Data Sets

In this section, we will present brief descriptions of three TF–DNA methods, SACO, ChIP-PET, ChIP-Seq, and provide characterization of data sets which we used in the validation of these methods.

2.6.1. SACO (21): Detection of Transcription Factor DNA–CREB BSs in the Rat Genome

TF CREB binds to the cAMP-response element (CRE), a sequence identified in the promoters of many inducible genes (21). CREB has since been found to mediate calcium, neurotrophin, and cytokine signals as well as a variety of cellular stresses. The SACO library was prepared using ChIP DNA obtained from $\sim 10^8$ rat PC12 cells. Chromatin occupancy in DNA was isolated from PC12 cells that have been stimulated with forskolin to increase intracellular cAMP 15 min prior to DNA extraction. Forskolin-treated PC12 cells were subjected to a CREB–DNA binding assay using an anti-CREB antibody. For SACO experiments, sonicated CREB ChIP DNA was polished (protruding 3' and 5' ends were made flush) and ligated to adapters for limited amplification. The resulting DNA fragments in the assay averaging around 700 bp in length were represented in a SACO library by 21 nt SAGE tags. These chromatin DNA fragments were digested with NlaIII, which cleaves genomic DNA approximately every 120 bp, and a modified SAGE procedure was used to create concatemerized 21 bp *genomic signature tags* (GSTs). Approximately 5,000 plasmids were sequenced to obtain the sequence of $\sim 3 \times 10^6$ GSTs. The resulting distinct 21 bp SACO GSTs were matched to genomic sites. GSTs with exact matches or matches with one substitution error that were uniquely assignable to a genomic location were considered as positives. GSTs without a genomic match (SACO Tag0) or with multiple matches were not considered. GSTs within 2 kb of each other are taken to be associated with the same SACO locus and formed “SACO clusters”. GSTs are counted in that clusters and thus can be used for semi-quantitative profiling of BSs of a given TF. For additional details of SACO loci *see* (21).

After genomic mapping and noise sequence filtration the authors have selected $\sim 41,000$ GSTs that identified a single region in the rat genome. The authors considered at least duplicate SACO

tags (we called Tag-2) as high-specific clusters; clusters with size 2, 3, ..., 94 were found. We will call all that SACO clusters a Tag-2+ set. At this cutoff value, we used 6,269 SACO clusters represented by 24,082 GST sequences (<http://genome.bnl.gov/SACO/>).

2.6.2. ChIP-PET: Estrogen Receptor Element (ERE)-DNA Binding Sites

The estrogen receptor alpha (ER- α) is a member of the nuclear hormone family of intracellular receptors which is activated by the hormone 17 β -estradiol (estrogen) (12). The main function of ER- α is its role as a DNA binding transcription factor which regulates gene expression. ER- α interacts either directly with genomic targets encoded by ER elements (EREs) (5'GGTCAnnnTGACC3'), or indirectly by tethering to nuclear proteins, such as AP-1, Sp-1, or NF- κ B that are bound to DNA at their cognate regulatory sites (12).

ChIP-PET is a ChIP-based method which uses SAGE-like tags and 36 bp paired-end tags (PETs) (12). Briefly, hormone-deprived breast cancer cell line cells, MCF-7, were treated with 10 nM 17 β -estradiol for 45 min and then DNA-bound receptor complexes were isolated through chromatin immunoprecipitation (ChIP) using anti-ER- α antibodies. PET sequences were extracted from the raw reads and mapped to the human genome sequence assembly (hg18). A total of 95% of ChIP DNA fragments ranged from 100 bp to 2 kb. The distribution of the sequence span of these DNA fragments followed the log-normal function with a span average of 674 bp, median of 458 bp, and mode of 277 bp. Additional information can be found in **Table 2.3**.

To find relationships between relative binding avidity of ERE BSs and expression level of putative direct ERE TF gene targets, we used U133A&B expression profiling of transcripts of human ER-positive MCF7 cells defined before and after stimulation with 10 nM 17 β -estradiol (12). In these experiments, the RNA was extracted at 12, 24, and 48 h and hybridizations were performed in microarray triplicates according to manufacture protocol.

The 480,042 original ChIP-PET sequencing reads of INF- γ -stimulated STAT1-DNA binding (library shc016 in T2G DB; HeLa S3 cells (20)) were mapped to single location in the human genome assembly (hg17) and 327,838 distinct (non-redundant) PETs (68%) were identified. Of these unique fragments, the PET tags whose DNA fragment spans 5'- and 3'-ends <6 kb apart were selected. Then the PET DNA fragments mapped to the regions of unlikely locations (mitochondria, Y chromosome, centromeric localization of clusters, long gene-free chromosome regions, etc. (*see above*)) were excluded from that PET DNA fragments data set. We then selected 324,523 distinct sequences of the ChIP-PET library shc016 representing 260,953 DNA fragment cluster overlaps (including singleton fragment genome hits). From the untreated control data set (library shc019 in T2G DB; HeLa S3 cells), 507,828 original ChIP-PET sequencing reads were mapped to single location in the human genome assembly

(hg17) and 263,901 distinct PETs (52%) were identified. Of these unique fragments, the PET tags whose DNA fragment spans 5'- and 3'-ends <6 kb apart were selected. Finally, we selected 254,233 $((254,233/507,828) \times 100\% = 50\%)$ distinct sequences of the ChIP-PET library shc019 representing 212,982 DNA fragment cluster overlaps for our further analysis.

2.6.3. ChIP-PET: STAT1-DNA Binding in INF- γ -Stimulated and -Unstimulated HeLa S3 Cells

Interferons (INFs) are well-known cytokines that play pivotal roles in antiviral, antibacterial, cell proliferation and differentiation, and anti-tumor responses primarily by modulating gene expression via the JAK/STAT pathway (3, 16). STAT1 regulates proliferation by promoting growth arrest and apoptosis in response to INF signals. These effects have given the INFs major therapeutic value in the treatment of hepatitis, melanoma, leukemia, lymphoma, and multiple sclerosis, although their mechanism of action is still unclear (16, 24). The regulation of STAT1 by INF- γ provides a useful system to understand how transcription factors select specific binding sites and ChIP-PET has been used to study INF- γ induced DNA-STAT1 binding. In our analysis, we used original ChIP-enriched DNA fragments of INF- γ -induced DNA-STAT1 ChIP-PET library (sch016) and non-induced DNA-STAT1 (sch019) library stored in T2G DB (<http://t2g.bii.a-star.edu.sg>; (29)). Based on the ChIP-PET protocol (4), STAT1 ChIP-PET libraries were prepared from INF- γ -stimulated (sch016) and -unstimulated (sch019) HeLa S3 cells as described in (20). Briefly, for each biological replicate 12×10^8 cells were used in total. These cells were split into INF- γ -treated and -untreated halves for STAT1 ChIPs. The ChIP-enriched DNA fragments were cloned into the cloning vector pGIS3 to generate the ChIP DNA fragment library. Additional information about the libraries and mapping of ChIP-PET DNA fragments on human genome can be found in the UCSC browser track "GIS-PET" (<http://genome.ucsc.edu/>).

T2G DB contains 327,838 and 263,901 distinct ChIP-PET DNA fragments obtained from INF-stimulated and -unstimulated HeLa S3 cells, respectively. These sequences, which uniquely mapped to the reference human genome (hg18), were clustered by T2G DB algorithm into 247,846 and 212,982 clusters (including singletons), respectively. We have also identified and excluded probable false-positive clusters (and cluster overlaps) having a very large size and located distantly (>100 kb) from the most neighboring gene. Sequences, clusters located in centromere regions, Y chromosome and M chromosome as well as alpha satellite, and overlapping gap regions were also excluded from our analysis. Finally, we used 324,523 and 259,759 ChIP-PET fragments obtained from INF-stimulated and -unstimulated HeLa S3 cells, respectively. We aggregated computationally these

sequences in 246,644 and 211,619 clusters (including singletons) for INF-stimulated and -unstimulated data sets, respectively. To identify specific binding loci, we also recalculated the number of the PET peaks (*see* below); for INF-stimulated and -unstimulated HeLa S3 cells we defined 260,953 and 218,440 PET peaks, respectively.

To find relationships between relative TF binding avidity of BSs and expression level of putative direct STAT1 gene targets, we used U133 plus2 expression profiling of transcripts of HeLa cells at 0 and after 2 and 4 h after stimulation of HeLa S3 with INF- γ in microarray triplicates at standard experimental conditions (S. Hartman's microarray data set; (12))

To validate TF-DNA BS predictions, we provide computational identification of position weighted matrix (PWM) of canonical STAT1 motifs. PWM method of TF motifs is used to scan the locations of STAT1 canonical motifs in unstimulated and INF- γ -stimulated DNA sequence data sets derived by ChIP-PET method. The canonical STAT1 motifs are scanned by *nmscan* program from NestedMiCA with threshold-5 (Score Threshold-5.0) and both strand search (<http://www.sanger.ac.uk/Software/analysis/nmica/>).

2.6.4. ChIP-Seq: STAT1-DNA Binding in INF- γ -Stimulated and -Unstimulated HeLa S3 Cells

ChIP-Seq maps single-end 27 bp tags using the Illumina 1G system (1G) which provides a 1.5–2 orders of magnitude increase in the amount of sequence than other systems (e.g., SACO). The method is based on deep 1G sequencing of short-read *single-end tags* (SETs), which are simpler to prepare than PETs. Using ChIP-Seq, the authors in (16) compared STAT1-DNA binding in INF- γ -stimulated and -unstimulated human HeLa S3 cells, by generating approximately 47 million reads, totaling 1.26 GB of sequence data, from immunoprecipitated DNA fragments. For each sample, the authors transformed reads that mapped to unique genomic locations into a DNA fragment overlap profile. They identified significant peaks by thresholding profiles (with peaks ≥ 11 bp) at a height equivalent to an estimated false discovery rate (FDR) < 0.001 . The global properties of the resulting two profiles were distinct and consistent with high ChIP enrichment.

The canonical STAT1 motifs are scanned by *nmscan* program from NestedMiCA with threshold-5 (Score Threshold-5.0) and both strand search.

2.6.5. ChIP-Seq Nanog and Oct4 Data: Sequence Filtering, Mapping Onto Reference Genome, and Sequence Clustering Are Essential Steps for Data Analysis

We used the similar data treatment procedure in our analysis of ChIP-Seq TFBS data sets for two ChIP-Seq libraries (GEO ID: GSE 11431) derived from mouse embryonic E14 cell lines (11).

3. Results

3.1. Computational Preprocessing of Raw Sequence Data, Sequence Mapping, Aggregation, and Preliminary Data Analysis

3.1.1. Long PET DNA Fragments Presented in the ChIP-Based Generated Library Can Form False Clusters and Produce Bias in the Count of Real Clusters and Its Sizes

The long tails of the frequency distribution of span of DNA fragments in ChIP-PET experiment can produce the errors in mapping, counting, and modeling of background (noise) events, since the longer DNA fragments have a larger chance to form random (false-positive) clusters and short fragments could be a source of multi-peak clusters. As an example, **Fig. 2.4A** shows the frequency distribution of a span (nt) of PET DNA fragments found in INF- γ -induced STAT1 ChIP-PET binding event data set. Goodness-of-fit analysis of the distribution showed that this distribution is a mixture of the relatively short DNA fragments (<700 nt) and very long DNA fragments (up to several thousands of nucleotides). **Figure 2.4B** demonstrates that the spans of DNA fragments clusters are often longer than the span of PET DNA fragments in PET fragment singletons. Obviously that the longer PET DNA fragments can form the false clusters and thus produce a bias in count of clusters and its peaks. It seems that prefiltering of very long ChIP DNA fragments before mapping of DNA fragments onto genomes will be a helpful procedure to reduce size of erroneous DNA clusters and its count.

3.1.2. Sequence Prefiltering and Compromise Between Specificity and Sensitivity in ChIP-Seq Data Analysis in Identification of TF STAT1 BSS in INF- γ -Stimulated HeLa S3 Cells

The STAT1–DNA binding ChIP-Seq assay produces millions of short (27 nt) unique sequence reads ($20\text{--}40 \times 10^6$) and thus it can contribute to high specificity and sensitivity of the TFBS mapping on the complex genomes (16). After ChIP for STAT1 in INF- γ -stimulated HeLa S3 cells, the 1G system produced a total of 24.1 million 27 bp sequence reads (SETs (16); **Table 2.1**). Of these, 15.1 million (63%) aligned uniquely to the non-repeat-masked human genome sequence. For unstimulated cells, the authors generated 22.7 million reads and uniquely mapped 12.9 million reads (57%).

A cluster peak “height” for the calculated XSET overlaps was the maximum number of overlapped XSETs for the cluster; height and FDR are inversely related. For both STAT1 libraries, the authors estimated FDR = 0.001 at height = 11. The resulting 41,582 stimulated peaks contained 17.9% of mapped reads, whereas the 11,004 unstimulated peaks contained only 4.2% of mapped reads ((16), **Table 2.1**). However, after the paper was published (16), the authors revisited their statistical criteria. In particular, new cutoff values of 9 and 10 were used for specific peaks in stimulated and unstimulated libraries, respectively, with the current data set. The authors also split largest clusters and redefined the number and heights of the peaks. In our analyses we used the revised sets as our raw data. With the authors’ new criteria, we

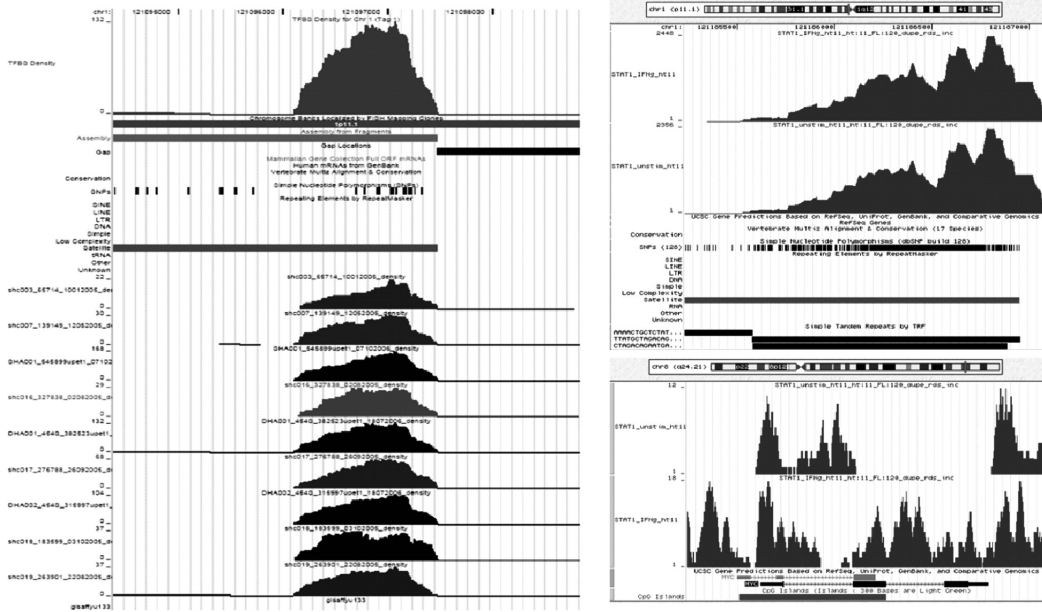


Fig. 2.4. Typical examples of non-specific binding regions. **(A)** False-positive ChIP-PET clusters can be often located within or nearby centromeric regions. Large clusters are occurred near chromosome gap regions in many ChIP-PET libraries for different TFs (T2G database). **(B)** False-positive ChIP-Seq clusters can be often located within or nearby centromeric regions. STAT1-DNA binding in INF- γ -stimulated and -unstimulated HeLa S3 cells ChIP-PET data sets. **(C)** False-positive group(s) of shot-height multiple clusters in ChIP-Seq-defined STAT1-DNA binding in the genomes of INF- γ -stimulated and -unstimulated HeLa S3 cells. Shot-height DNA fragment clusters might be observed in unusual multiple positions within genic or non-genic regions (including 3' ends of the gene and down-stream gene regions). Such unlikely real clusters often unstable in a given genome region across the different experiments and their number is very sensitive to the statistical method and criteria, which are used to estimate specificity of experiment (or FDR by (16), **Table 2.1**). Due to suboptimal experimental design, small number of samples and ignoring the genome sequence complexity information, ChIP-qPCR validation experiments could also provide an optimistic estimate of specificity (and lower cutoff peak value) of ChIP-Seq experiment (V.A. Kuznetsov, unpublished)).

counted 63,309 confidence-stimulated peaks and 16,470 unstimulated confidence peaks. We discarded all sequence reads that could not be uniquely mapped to the genome and did not follow our filtering criteria.

Figure 2.3 shows an example of a large false-positive multi-cluster DNA fragment mapping in centromere region of chromosome 1. These clusters exhibit the same location and the same shape of the clusters for unstimulated and INF- γ -stimulated samples. Such clusters were considered as false-positive clusters. The clusters (and related cluster peaks) with relatively long spans, large size, and that were distantly located (>100 kb) from most neighboring genes were also excluded from analysis (false-positive events due to systematic errors).

In total, we filtered out 1.3% ($853/63,309 \times 100\%$) and 4.4% ($727/16,470 \times 100\%$) problematic and false-positive stimulated and unstimulated peaks, respectively.

Table 2.1
Characteristics of updated ChIP-Seq libraries (16) after our revision

Parameter	Stimulated	Unstimulated
Reads		
Total sequenced (10^{-6})	24.1	22.7
Total, uniquely mapped onto genome (10^{-6})	15.1	12.9
In peaks (10^{-6})	2.71 (17.9%)	0.54 (4.2%)
Peak coverage (MB)	34.5 (1.12%)	9.7 (0.31%)
Peaks		
Median width (bp)	473	519
Peak height at revised threshold	10	9
No. of peaks	63,309	16,470
Average height	19.9	12.9
Median height	13	12
No. of peaks after filtering (used in our work)	62,456	15,743
No. of peaks in problematic clusters	853 (1.3%)	727 (4.4%)
Average height of clusters after filtering	20	12.6
Average height of clusters in problematic clusters	17.3	18
Peak coverage after filtering (MB)	34.1	9.3
Peak coverage of problematic clusters (MB)	0.45	0.4
Median width (bp)	474	522
Median width of problematic clusters (bp)	414	427
Number of sequences on peaks	1,246,120	198,566
Number of overlapping loci	14,874 (23.8%)	14,303 (90.8%)
Number of non-overlapping loci	47,582 (76.2%)	1,440 (9.2%)

The resulting 62,456 stimulated peaks contained 8.2% (1,246,120/15,100,000) of confidence-mapped reads, whereas the 16,470 unstimulated peaks contained 36.8% (198,566/540,000) of confidence-mapped reads (**Table 2.1**).

Finally, after all filters, we selected 62,456 and 15,743 ChIP-Seq fragments obtained from the libraries of INF-stimulated and -unstimulated HeLa S3 cells, respectively. **Table 2.1** summarizes our characteristics of revisited ChIP-PET sequence libraries derived from stimulated and unstimulated HeLa S3 cells.

3.1.3. ChIP-PET ERE Data: Sequence Filtering, Mapping Onto Reference Genome, and Sequence Clustering Are Essential Steps for Data Analysis

A significant amount of non-specific (background) genomic DNA is always present in the immunoprecipitated DNA material of any ChIP-based DNA sequence library. In fact, about 35–80% of original ChIP-based tag sequences are non-genomic/noise sequences. Fortunately, these DNA sequences can be easily filtered out after computer mapping of the DNA fragments on the genome. For example, in ERE–DNA binding study (12) (library shc07; <http://t2g.bii.a-star.edu.sg>), a total of 635,371 PETs were sequenced, of which 361,241 (~56.86%) were mapped unambiguously to unique loci in the reference genome (12). We filtered and reanalyzed ChIP-PET DNA fragment sequence data of ERE binding sequences using our algorithm of filtering, mapping, and clustering (**Section 2** (14)). Due to observed PET sequence duplication, the number of unique PET sequences mapped uniquely to the genome was reduced to ~137,500 distinct PET sequences. After additional filtering of sequences mapped to Y and M chromosomes, centromere, gap, and alpha satellite regions we finally selected 136,348 $((136,348/635,371) \times 100\% = 21.5\%)$ distinct sequences of the ChIP-PET library representing 124,756 DNA fragment cluster overlaps (including singleton fragment genome hits) (**Table 2.2**).

3.2. Real TF Binding Events in ChIP-PET and SACO Data Sets Are Following the Common Skewed Statistics

The empirical frequency distributions of the binding events for all ChIP-based data sets exhibit monotonically skewed shapes with a greater abundance of rare binding events and more gaps and irregularities among the higher-confidence binding events (**Figs. 2.3** and **2.4**). The form of the distributions of binding events are very similar to each other and the forms of the empirical distribution functions, which we have observed for DNA–TF binding events for ChIP-PET (4, 14, 25) and other data genome-related data sets (13, 27). In particular, in (4, 14, 25), we have demonstrated that the frequency distribution of the TF binding events defined in the ChIP-PET experiments specifically for the in the left-size of the empirical frequency distribution of the TF binding events for different TF, cell types, and environmental conditions are less sensitive to additive random noise and can be fitted well by the GDP function [3].

Figure 2.5A–D confirms these observations and demonstrates that shape of the frequency distribution of BEs is a common distribution for studied ChIP-based methods. **Figure 2.5A–C** shows the histograms specifying the frequency of distinct binding events in ChIP-PET experiments (panels A and B: INF- γ -stimulated and -unstimulated STAT1 and panel C: estradiol-induced ER–DNA binding events). Comparison of **Fig. 2.5D** with **Fig. 2.5A–C** demonstrates that the histogram specifying the frequency of binding events of TF CREB in SACO experiment (*see Section 2*) is very similar to the histograms for ERE TF and STAT1 TF ChIP-PET experiments.

Table 2.2
Observed and predicted frequency distributions of binding events in CREB BS (SACO), ERE BS (ChIP-PET), and INF- γ STAT1 BS (ChIP-PET)

1. SAGO, cluster size															
CREB BSs	1	2	3	4*	5	6	7	8	9	10	c11+	c4	Spec. c4+	c3+	Spec. c3+
Observed loci	17,509	3,588	1,065	518	274	208	126	77	69	50	328	1,650	1	2715	1
Specific loci (GDP)	3,036	1,390	742	439	279	188	132	96	72	55	271	1,532	0.929	2274	0.838
Specific (K-W)	3,037	1,389	743	441	282	191	134	98	74	56	291	1,568	0.950	2311	0.851
Non-specific loci(NSL)	14,473	2,228	343	52.8	8.1	1.3	0.2	0	0	0	0	62.4	0.04	405	0.15
Fraction of NSL	0.827	0.621	0.322	0.102	0.030	0.006	0.002	0	0	0	0	non	non	non	non
2. ChIP-PET, cluster overlap															
ERE BSs	1	2	3	4*	5	6	7	8	9	10	c11+	c4	Spec. c4+	c3+	Spec. c3+
Observed loci	11,7024	6,245	771	279	133	93	70	40	22	24	55	716	1	1,487	1
Specific loci (GDP)	3,693	1,202	512	258	145	88	57	39	27	20	79	713	0.996	1,225	0.824

Specific (K-W)	3,693	1,200	513	257	144	87	56	37	26	19	68	693	0.969	1,206	0.811
Non-specific loci (NSL)	113,331	5,046	225	10	0.4	0.0	0.0	0.0	0.0	0.0	0.0	10	0.01	235	0.16
Fraction of NSL	0.9684	0.8081	0.291	0.036	0.003	0	0	0	0	0	0	non	non	non	non
3. ChIP-PET, overlap height															
INF-γ STAT1 BSs	1	2	3	4	5	6*	7	8	9	10	c11+	c6+	Spec. c6+	c5+	Spec. c5+
Observed loci	212,447	39,025	7,103	1,444	400	157	113	66	39	38	121	534	1	934	1
Specific loci (GDP)	2,763	1,368	734	420	253	159	104	70	48	34	110	526	0.985	779	0.834
Specific (K-W)	2,764	1,368	735	422	255	161	105	71	49	35	111	531	0.995	786	0.841
Non-specific loci (NSL)	209,697	37,516	6,712	1,201	215	38	7	0	0	0	0	45	0.085	260	0.279
Fraction of NSL	0.99	0.96	0.94	0.83	0.54	0.24	0.06	0.00	0	0	0	non	non	non	non

GDP: best-fitted (truncated) generalized discrete Pareto probability function; K-W: best-fitted Kolmogorov-Warning probability function; c: occupancy level of BS locus; *: cutoff value at selected specificity (Spec.).

3.3. Specificity and Critical Cutoff Value of Binding Events in ChIP-PET and SACO Experiments

Using our curve-fitting analysis method (Section 2) with the experimental data presented in Fig. 2.5A–D, we de-composited the noise and specific components of binding events in studied ChIP-PET and SACO data sets and estimated a cutoff value of binding events at the appropriate specificity. Table 2.2 contains detailed information about the number of binding events from 1 to 11 for CREB SACO library and 2 ChIP-PET libraries. The numerical results of goodness-of-fit analysis of GDP and K–W functions are also presented in Table 2.2 and both show an accurate approximation of the available data. The probability of specific events (specificity) in each experiment is also estimated at different reasonable cutoff values. A detailed description and

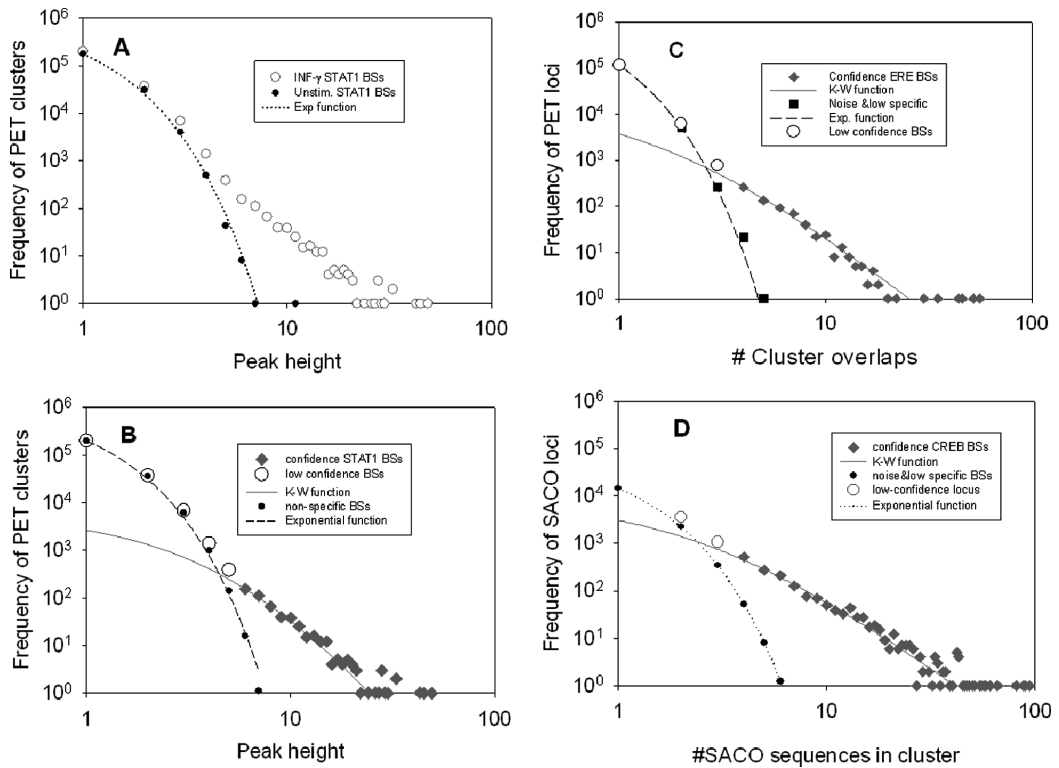


Fig. 2.5. Common statistical properties of binding events and goodness-of-fit analysis of the distribution of binding events across ChIP-based methods and different TF–DNA binding systems (ChIP-PET, SACO). (A) STAT1–DNA binding events in INF- γ -stimulated ($\circ$) and -unstimulated (control) ($\bullet$) cell samples. Binding event was defined by the highest peak in overlap of cluster ChIP-PET DNA sequences. Dotted line: best-fit exponent function at power coefficient, $s = 1.95 \pm 0.136$; $t = 14.4$; $p < 0.0001$. (B) A decomposition and estimation of the parameters of our mixture probability distribution model: ($\diamond$) best-fit GDP to specific (right) tail of the empirical distribution; ($\bullet$) model-predicted non-specific binding events; ($\circ$) total number of specific and non-specific events in the overlapped region of the mixture distribution. Exponential function with power coefficient $s = 1.86 \pm 0.072$ ($t = 25.96$; $p < 0.0001$) fits well to predicted frequency distribution and simultaneously fits to observed frequency distribution of non-stimulated HeLa S3 cells (panel A). The generalized discrete Pareto (GDP) functions [3] fits well to truncated statistics of specific binding events at cutoff value equal to 6 and $k = 4.446 \pm 0.0806$ ($t = 56.9$; $p < 0.0001$); $b = 6.258 \pm 0.1260$ ($t = 51.9$; $p < 0.0001$). See Table 2.3 for additional information and details.

numerical values of these and other important statistical characteristics of the libraries are presented in **Table 2.3**. Comparison of **Fig. 2.5A,B** show that the “noisy” component of the mixture distribution [1], modeled by exponential distribution provides a suitable fit to the empirical distribution of binding events of TF STAT1 in INF- γ -stimulated HeLa S3 cells.

Table 2.3

Comparison of performance of the ChIP-based Tag sequencing platforms. 1. CREB (SACO Clusters); 2. INF- γ -stimulated specific STAT1 BS Type I (ChIP-PET peaks); 3. INF- γ -stimulated specific STAT1 BS Type I (ChIP-Seq peaks); 4. Unstimulated specific STAT1 BS Type I (ChIP-Seq peaks)

Subset of STAT1 BSs	Avidity	Subset of BS	No. of loci	STAT1 motifs	Fraction of BSs
Stimulated	High	Peak height ≥ 61	2,322	1,898	0.82
Stimulated	Low	Peak height < 61	60,134	23,653	0.39
Unstimulated	High	Peak height ≥ 31	64	30	0.47
Overlapped with stimulated BSs	Low	Peak height < 31	14,239	4,452	0.31
Non-overlapped with stim. BSs	Low	Peak height < 31	1,440	282	0.20

*Data for high-occupancy (specific) BS Type I; θ , a , b : Estimated parameters of KW function; k , β : estimated parameters of GDP function; s : estimated power parameter in exponential function.

Figure 2.5A demonstrates that GDP fits well to the more biologically relevant right tail of the empirical distribution of binding events of TF STAT1 in INF- γ -stimulated HeLa S3 cells, as the exponential function fits well to the distribution of the binding events in the unstimulated STAT1 set (**Fig. 2.5B**). Our co-localization analysis of ChIP-PET BSs in the stimulated and unstimulated data sets shows that only a few BSs were common in the stimulated and unstimulated data sets. The canonical STAT1 motif (TTCCNGGAA) was also well defined in a significant fraction of the relevant tail of the empirical distribution of binding events of TF STAT1 in INF- γ -stimulated HeLa S3 cells but it was poorly defined in unstimulated STAT1 set (not presented).

These results suggest that of BEs in unstimulated data set cannot be reliably detected in the given ChIP-PET experiment.

Figure 2.5C,D show that for the SACO and ERE TF BEs analyses, the simple exponential distribution statistics also fits well the false-positive low-abundance cluster overlap peaks and GDP and K-W functions can approximate the tail of the empirical frequency distribution of specific binding events (after filtering out of non-specific clusters) efficiently. Parameters of the fitting are represented in **Table 2.3**. The estimates presented in **Tables 2.2** and **2.3** suggest that only a relatively small fraction of real BS in all these experiments can be identified within the experiments. We (4, 10, 12, 14) and others (8, 20, 30, 15, 16) have shown that in combination with motif search procedures (e.g., (6)) and expression data analysis, many ChIP-based specific TFBSs can be confirmed as the specific and functional BSs.

It is possible that there are many real low-abundant and moderate-abundant binding events representing BSs (presented in the filtered library), but these events cannot be considered as real physical binding sites due to the high fraction of noise in ChIP detection of those BSs. In the case of ChIP-PET data for several ChIP-PET binding events (c-myc, ERE, p53, oct4, Nanog), we have predicted these false-positive BSs theoretically and then proved the prediction experimentally in PET-ChIP assay (4, 13). For instance, we have found that in c-myc, p53, and in ERE ChIP-PET data, which library/sample sizes were comparable with the CREB SACO library, a critical (cutoff) level at conventional specificity ($\sim 90\text{--}95\%$) was associated with loci having 3, 3, and 4 overlapped clusters, respectively (14). For singletons PET data, a fraction of computationally predicted and ChIP-PCR-confirmed specific BSs was less than 10–20% of all observed single sequences mapped on the genome (9, 10).

Analysis of co-localization of the 6,303 loci identified by multiple GSTs reveals that only 40% of 6,303 GST clusters were within usual 2 kb of the transcriptional start of an annotated gene, and 72% were within 1 kb of a putative cAMP-response element (CRE). Most of these loci are represented by highly abundant clusters (not presented).

In general, the larger clusters (or cluster peaks) tend to map more specific binding sites (4, 14) (*see also Fig. 2.5; Tables 2.2 and 2.3*). This property has also been observed for p53, Stat1, c-Myc ChIP-PET TF binding loci in combination with microarray expression data analysis and by a combination of direct qPCR measurements with motif search analyses (4, 10, 12, 14, 20). Our combined data analyses, including validation of limited random subsets of PET clusters with ChIP-qPCR, show that specific loci for these TFs can also be determined in the corresponding PET data sets even in smallest clusters (PET-2 peaks) and in singletons (10). Moderate-size and large clusters could contain a relative small fraction of non-specific clusters, however, these clusters can be accurately validated in combination with other (mentioned above) independent methods (4, 9, 10).

3.4. Sensitivity and Estimates of the Total Number of TFBS in ChIP-PET and SACO Experiments

To estimate the number of specific BSs we begin with an estimation of the specific BS events in a high-noisy enriched region. For instance, for INF- γ -induced STAT1 binding data this region includes peak values 5, 4, 3, 2, 1 (Table 2.2; Fig. 2.5B). For this estimate, the best-fit GDP function can be back extrapolated to the high-noisy enriched region and then be fitted by K-W function. According to this method, the entire GDP function was accurately fitted by K-W function, equation [4] (discontinued line; Fig. 2.5B), with model parameters $\theta = 0.997$, $a = 4.341$, and $b = 8.757$. Finally, using equation [5] we can estimate the fraction of undetected specific BSs of STAT1 TF $p_0 = (1 - a/b) = 0.50$. Empirical distribution and K-W fitting results in ERE ChIP-PET experiments provide $p_0 = 0.77$. Empirical distribution and GDP fitting results in SREB SACO experiment provide $p_0 = 0.56$. Table 2.3 provides detailed information about the identified statistical distribution. For example, the total number of specific SACO-defined SREB BSs in the rat genome is 15,438 and the total number of specific PET-defined BSs for ERE and STAT1 in the human genome is 26,350 and 12,252, respectively. The dynamical range of these numbers consists of the previous publications (11–16, 20, 3).

3.5. ChIP-Seq DNA Fragments for TFs Nanog and Oct4 Binding DNA in Mouse Embryonic E14 Cells

Nanog and Oct6 are the key regulatory genes involved in self-renewing and development of the mouse and human embryonic stem cells (9, 11). Figure 2.6 shows the frequencies of peak height value distributed according to the number of occurrences of the 200-nt ChIP-Seq DNA fragments for Nanog and Oct4 TF binding DNA in mouse embryonic E14 cells mapped onto the reference mouse chromosomes (11) and collected in the T2G libraries (29). Both data sets show a similar form of the frequency distribution of binding events (peak heights). We used the probabilistic models [3], [4], and [5] to describe the distribution of specific binding events (e.g., the number of real peaks) observed in these ChIP-based experiments. The models [3], [4], and [5] allow us to fit the data well to the empirical distributions (Fig. 2.6). Additionally, Fig. 2.6 shows that an extrapolation of the best-fitted K-W function into the right side high-noisy region of the distribution (the left part of the distribution with height between 1 and cutoff value of ChIP-qPCR-defined specificity (11), $c = 11$ for Nanog and $c = 8$ for Oct4 TF binding events). Moreover, using equations [4] and [5], we can estimate a fraction of real specific BSs which has not been detected in the ChIP-Seq experiments (p_0). Estimated parameters of equation [4] for Nanog TF data: $\theta = 0.997$; $\alpha = 5.870$; $\beta = 7.465$, and thus by equation [5], $p_0 = 0.22$; for Oct4 data: $\theta = 0.998$; $\alpha = 5.681$; $\beta = 8.32$, thus by equation [5], $p_0 = 0.32$. Finally, we can estimate the total number of specific (physical) BSs in the mouse genome for a given TF. This estimate equals $\sim 66.7 \times 10^3$ BSs for Nanog TF and $\sim 28.2 \times 10^3$ TFBSs Oct4.

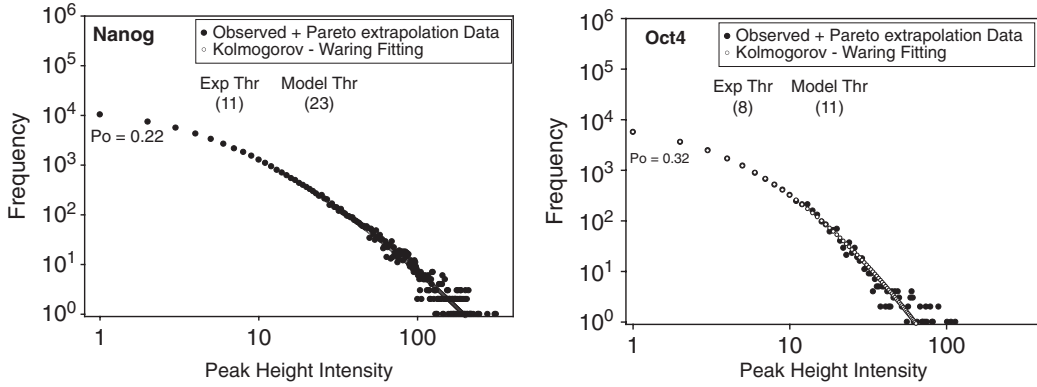


Fig. 2.6. Goodness-of-fit analysis of the frequency distribution of binding events (no. of ChIP-Seq peaks) fitted starting from cutoff peak value defined in ChIP-qPCR. **(A)** Nanog TF data set ($M_{c+} = 468,156$; $N_{c+} = 52,004$). **(B)** Oct4 TF peak data set ($M_{c+} = 84,810$; $N_{c+} = 19,160$). For these data (11), the K-W model fits to the observed tail (reliable part of the mixture frequency distribution), as well as the best-fit truncated GDP model data extrapolated to the smaller peak values including value 1 (belonging to the highly noise-enriched subset). K-W function fits to best-fit estimated GDP function values on the entire dynamical range of peak values. Due to high confidence of curve-fitting of K-W, the parameters of K-W function can be thus accurately estimated. Using the estimated parameters α and β allow us to calculate a fraction of TF BEs (p_0) which was not detected in a given ChIP-Seq library. Empty circle: fitted and extrapolated K-W function values. Close circle: the observed (truncated at $c = 11$) frequency distribution of BEs. Vertical dash line is represented by ChIP-qPCR defined threshold.

3.6. Sample-Size Dependence Analysis Can Define Rescaling in Binding Mechanisms of ChIP-Based Methods

ChIP-Seq (16) uses much larger number of sequence reads than other methods. Therefore the examination of TF binding sites might be more adequately represent real complexity and diversity of the TF binding patterns. We suggest that due to this advantage the empirical distribution of ChIP-Seq data set will allow us to find more diverse/complex patterns of STAT1-DNA binding than ChIP-PET and other ChIP-based methods.

We assumed that two types of binding events in the human genome could exist and associate with two distinct molecular mechanisms of Stat1-DNA binding. Therefore, we suggested that two distinct specific distributions of STAT1 binding should be observed in ChIP-Seq experiment. We hypothesized that our mixture distribution model could reflect the “low-avidity” and “high-avidity” binding events of two distinct classes of specific binding events. **Figure 2.7** shows that a mixture of two K-W functions fits well to the empirical frequency distribution of STAT1-DNA binding events. We found that a single K-W function could not fit to the data equally as well due to a decomposition of the mixture of “low-avidity” and “high-avidity” binding events (not presented). ChIP-Seq binding events are represented by XSETs. Detailed notations and estimated parameter values of the model are presented in **Tables 2.3** and **2.4**. **Table 2.4** shows that in a stimulated sample, the high-avidity STAT1 BS are strongly associated with location of canonical motif (82% BSs association), while small number of high-avidity binding sites (64 BSs) also exist

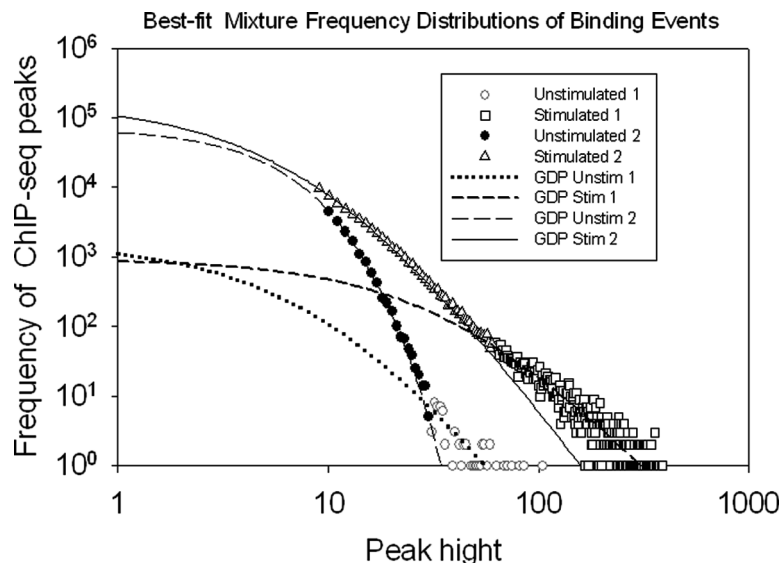


Fig. 2.7. The frequency distribution of INF- γ -stimulated and -unstimulated STAT1–DNA binding events in ChIP-Seq experiment (16). We consider the empirical distribution as a mixture of distributions of two distinct classes of binding events. Two K–W functions fits well to the empirical frequency distribution of STAT1–DNA binding events due to a decomposition of the mixture of “low-avidity” and “high-avidity” binding events. ChIP-Seq binding events are represented by XSETs. See detailed notations on the graphic and the estimated parameter values of the model in **Tables 2.3** and **2.4**.

in subpopulation of unstimulated sample and 47% of these BSs are supported by canonical motifs. A large number of stimulated and unstimulated BS exhibit a low-avidity pattern. The functional role of these binding sites is still unknown.

We also performed a comparison of ChIP-PET (20) and ChIP-Seq (16) detection of high-avidity binding sites of STAT1 TF in INF- γ -stimulated (**A,C**) and INF- γ -unstimulated HeLa S3 cells (**B,D**). For stimulated HeLa S3 cells, the ChIP-Seq and ChIP-PET data are consistent in chromosome location (common loci) and the relative amplitude of the signals. However, the ChIP-PET data (in which library size is much smaller than the library size of ChIP-Seq experiment) lost essential information about truly existing binding sites in the genome of unstimulated cells. (ChIP-PET binding events are presented by the peaks that correspond to the number of overlapping ditags at a given genomic coordinate. ChIP-Seq binding events are represented by XSETs.) However, the fraction and height values of specific BEs in ChIP-Seq data set are expressed much more strongly and significantly than in the ChIP-PET data set. We therefore demonstrate for the first time that ChIP-Seq data for unstimulated HeLa S3 cells can be used to detect 64 high-confidence specific binding sites in unstimulated HeLa cells (**Fig. 2.8**). We found 45 RefSeq genes predicted as direct targets for “basal” transcription regulation by

Table 2.4
High- and low-avidity BSs in stimulated and unstimulated subsets

Definition	SACO	ChIP-PET	ChIP-Seq(1)	ChIP-Seq(2)
Total number of loci, N	23812	260953	62456 at c_{9+}	15743 at c_{10+}
Total number of DNA fragments in N loci, M	41702	324523	1246120 at c_{9+}	198566 at c_{10+}
Mean of occupancy (e.g., height of peak). M/N	1.75	1.25	19.95 at c_{9+}	12.61 at c_{10+}
Fraction of DNA fragment singletons, p_1	0.74	0.81	NA	NA
Predicted specific loci in the library, N_{sp}	6737	6074	16872	4694
DNA fragments N_{sp} , M_{sp}	20962	15599	493992	25794
Mean of specific occupancy. M_{sp}/N_{sp}	3.11	2.57	29.28	5.50
Fraction of specific loci out of library, P_o	0.563	0.50	0.04	0.22
Predicted specific loci in the genome, N_{tot}	15438	12252	17661	5985
Confidence value of occupancy (e.g., peak height), ϵ	4	6	61	31
Observed loci at occupancy level $\geq c_+$, $N(c_+)$	1650	534	2322	64
DNA fragments in N_{c+} , $M(c_+)$	13822	4941	271521	2809
Predicted specific loci in N_{c+} , $N_{sp}(c_+)$ (by K-W)	1568	526	2134	58
DNA fragments in $N_{sp}(c_+)$, $M_{sp}(c_+)$	12918	4662	237790	2148

% Estimated specific loci in library, $N_{sp}/N*100\%$	28.3	1.87	NA	NA
% Specific DNA fragments in the library, $M_{sp}/M*100\%$	50.3	4.54	NA	NA
% Specific loci with occupancy value $\geq c$, $N_{sp}(c_+)/N_{sp}*100\%$	25.5	8.09	12.65	1.36
% specific sequences in N_{c_+} , $Msp(c_+)/M_{sp}*100\%$	61.6	29.88	48.14	9.49
Specificity = $N_{sp}(c_+)/N(c_+)*100\%$	95.0	91.52	91.90	90.70
Sensitivity = $N_{sp}(c_+)/N_{tot}*100\%$	10.16	4.36	12.08	0.97
θ (K-W)	0.999	0.997	0.996	0.969
a (K-W)	1.712	4.341	37.318	7.633
b (K-W)	3.924	8.757	39.064	9.737
k (GDP)	2.224	4.446	2.266	2.975
β (GDP)	2.649	6.258	43.347	10.290
s (exponential function)	1.871	1.721	Not fitted	Not fitted

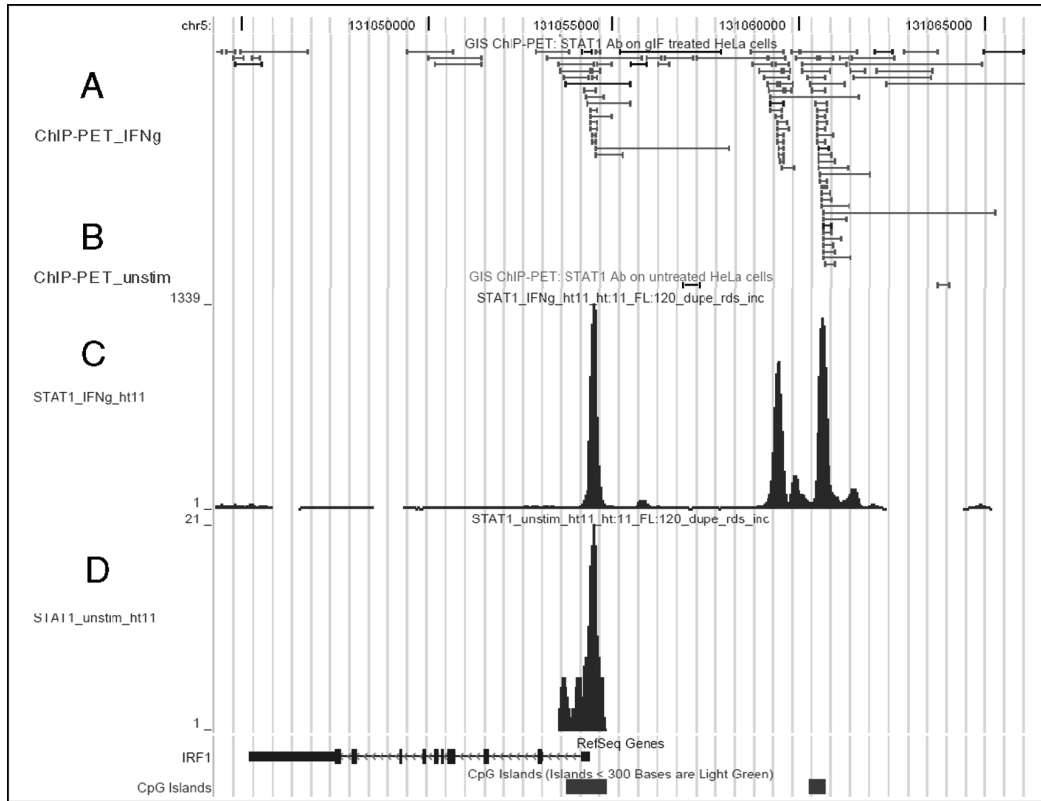


Fig. 2.8. Sample-size dependence of ChIP-based methods. Comparison of ChIP-PET (20) and ChIP-Seq (16) detection of high-avidity binding sites of STAT1 TF in INF- γ -stimulated (A,C) and INF- γ -unstimulated HeLa S3 cells (B,D). For stimulated cells, both data provide strong consistency in chromosome location (common loci) and relative amplitude of the signals, however, ChIP-PET data (which library size is much smaller than the library size of ChIP-Seq experiment) lost essential information about truly existed binding sites in the genome of unstimulated cells. ChIP-PET binding events are presented by the peaks corresponds to the number of overlapping PET DNA Sequences. ChIP-Seq binding events are represented by XSETs.

STAT1 in INF- γ unstimulated cells including interferon regulatory factor 9 (IRF9), ATP-dependent DNA helicase homolog (*Saccharomyces cerevisiae* (HFM1)), polyamine-modulated factor 1 (PMF1), WD repeat domain 74 (WDR74), polymerase (DNA directed), delta 1 catalytic subunit 125 kDa (POLD1) (Table S1), and many poorly described and in silico predicted genes, for instance, LOC284801.cApr07, kloypeybo.aApr07, RNU2P2.cApr07, LOC554226.bApr07.

We would like to indicate that noise BEs in ChIP-Seq and ChIP-PET data sets are increased when data set sample size becomes larger. This common property of ChIP-based sequencing methods is a difficult problem to resolve for reliable detection of real BSs events even in a very large-size samples.

Figure 2.9 provides useful information regarding comparative sensitivity of ChIP-Seq and ChIP-PET analyses. In our meta-analysis of co-localization of STAT1 binding sites both the

Table S1

ChIP-Seq-defined high-avidity STAT1 TFBSs (peak height >30 DNA fragments) in unstimulated HeLa S3 cells found in 50 kb upstream transcription start site (TSS) region and in downstream gene regions

Cluster	Max height in cluster (unstimulated cells)	Chr coordinate	RefSeq	Gene Symbol	Strand	Distance btw TSS of gene and start of cluster, nt
55	32	chr5:32803910-32804468	NM_000908	NPR3	+	-56488
37	35	chr5:37448874-37449213	NM_018034	WDR70	+	-33705
51	33	chr5:17302178-17302751	NM_006317	BASP1	+	-31428
5	71	chr1:154452793-154453347	NM_014635	SLC25A44	+	-22439
5	71	chr1:154452793-154453347	NM_001135672	SLC25A44	+	-22439
31	39	chr5:134287616-134288360	NM_032151	PCBD2	+	-18907
38	35	chr17:35390245-35390634	NM_178171	GSDMA	+	-17493
1	105	chr14:23700051-23700428	NM_017999	RNF31	+	-13552
52	33	chr11:47556913-47557379	NM_175732	PTPMT1	+	-13208
44	34	chr12:52131714-52132186	NM_00105354	PRR13	+	-10014
41	34	chr12:52131714-52132186	NM_018157	PRR13	+	10014
11	55	chr11:62365421-62365972	NM_003154	STX5	-	-9284
33	36	chr17:30502100-30502605	NM_001014445	NLE1	-	-8664
33	36	chr17:30502100-30502605	NM_018096	NLE1	-	-8664
3	71	chr1:154452793-154453347	NM_007221	PMF1	+	-3385
33	36	chr17:30502100-30502605	NM_001033576	UNC45B	+	-3151
33	36	chr17:30502100-30502605	NM_173157	UNC45B	+	-3151
29	40	chr5:40870875-40871529	NR_002583	SNORD72	-	-2280
49	33	chr5:10363449-10364226	NM_138809	CMBL	-	-2280
11	55	chr11:62365421-62365972	NM_018093	WDR74	-	-1216
22	45	chr17:37144858-37145318	NM_001079871	HAP1	-	67
22	45	chr17:37144858-37145318	NM_177977	HAP1	-	67
22	45	chr17:37144858-37145318	NM_001079870	HAP1	-	67
53	33	chr15:47235030-47235368	NM_001143887	COPS2	-	117
53	33	chr15:47235030-47235368	NM_004236	COPS2	-	117
57	32	chr5:36187640-36188151	NM_001007527	LMBRD2	-	133
52	33	chr11:47556913-47557379	NM_016506	KBTD4	-	185
1	105	chr14:23700051-23700428	NM_006034	IRF9	+	211
52	33	chr11:47556913-47557379	NM_018095	KBTD4	-	231
52	33	chr11:47556913-47557379	NR_024222	KBTD4	-	231
53	33	chr15:47235030-47235368	NM_001001556	GALK2	+	238
42	34	chr1:143807525-143807996	NM_004892	SEC22B	+	239
39	35	chr19:55571256-55571614	NM_007121	NR1H2	+	241
43	34	chr1:171950450-171950995	NM_014458	KLHL20	+	253
29	40	chr5:40870875-40871529	NM_000997	RPL37	-	270
52	33	chr11:47556913-47557379	NM_004551	NDUFS3	+	295
41	34	chr1:93317075-93317639	NM_007358	MTF2	+	305
57	32	chr5:36187640-36188151	NM_005933	SKP2	+	306
57	32	chr5:36187640-36188151	NM_032637	SKP2	+	306
43	33	chr1:161558037-161558717	NM_031423	NUF2	+	310
43	33	chr1:161558037-161558717	NM_145697	NUF2	+	310
33	35	chr17:35390245-35390634	NM_002809	PSMD3	+	341
34	35	chr2:74535384-74535835	NM_031288	INO80B	+	373
45	34	chr17:46585635-46586093	NM_000259	NME1	+	384
45	34	chr17:46585635-46586093	NM_198175	NME1	+	384
45	34	chr17:46585635-46586093	NM_001018136	NME1-NME2	+	384
44	34	chr12:52131714-52132186	NM_031939	PCBP2	+	439
44	34	chr12:52131714-52132186	NM_005016	PCBP2	+	439
44	34	chr12:52131714-52132186	NM_001098620	PCBP2	+	439
44	34	chr12:52131714-52132186	NM_001128914	PCBP2	+	439
44	34	chr12:52131714-52132186	NM_001128913	PCBP2	+	439
44	34	chr12:52131714-52132186	NM_001128912	PCBP2	+	439
44	34	chr12:52131714-52132186	NM_001128911	PCBP2	+	439
59	32	chr12:52355841-52356497	NM_005176	ATP5G2	-	536
24	44	chr5:324169324820	NM_013232	PDCD6	+	567
59	32	chr12:52355841-52356497	NM_001002031	ATP5G2	-	939
54	32	chr5:5474861-5475820	NM_015325	KIAA0947	+	946
34	35	chr2:74535384-74535835	NM_012477	WBP1	+	3751
29	40	chr5:40870875-40871529	NM_032537	CARD6	+	6292
39	35	chr19:55571256-55571614	NM_002691	POLD1	+	8149
34	35	chr2:74535384-74535835	NM_006302	GCS1	-	10712
17	51	chr17:38821136-38822164	NM_001651	ARL4D	+	10743
2	85	chr1:91625290-91625782	NM_001017975	HFM1	-	17785
35	35	chr5:16947382-16948205	NM_012334	MYO10	-	42004

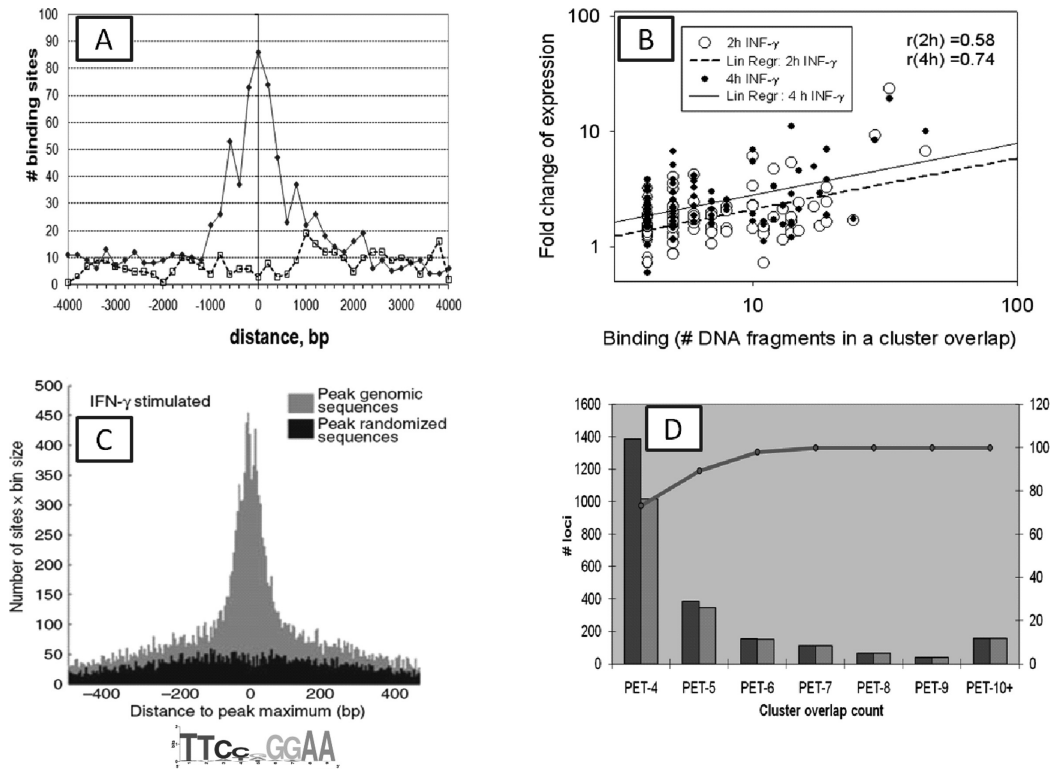


Fig. 2.9. Comparison analysis of the results of ChIP-Seq and ChIP-PET data sets. **(A)** ChIP-PET binding site distributions near a transcription start site (TSS) of gene. **(B)** Spearman correlation between relative avidity (peak height in ChIP-PET experiment) and expression signal value (time course expression micro-array Hartmann's data; (31)). **(C)** ChIP-Seq binding site distributions near transcription start site (TSS) of a gene a variation of the distribution in ChIP-Seq BEs is smaller in compare to ChIP-PET. **(C)** ChIP-Seq data (16). **(D)** Consistency between frequencies of ChIP-PET and ChIP-Seq binding events approaches to 100% level (solid line) at value PET-6 and above.

methods provide similar shape of binding site distributions near transcription start site (TSS) of a gene. However, spatial variation of the BSs in ChIP-Seq experiment is relatively smaller than in the ChIP-PET experiment (Fig. 2.9A; Fig. 2.9C). Consistency between frequencies of ChIP-PET and ChIP-Seq binding events approaches to 100% level (solid line Fig. 2.9D) at value PET-6 and above.

3.7. ChIP-Based Approach Allows Measuring a Relative Avidity Function

Interestingly, we determined a strong correlation between peak height in STAT1 ChIP-PET experiments and expression signal value (defined by micro-array Hartmann's data; (31)) in putative gene targets located in vicinity 2 kb from identified STAT1 BS. Figure 2.9C shows that this correlation (estimated by Spearman coefficient correlation) presents in time-course microarray experimental design (0 h, 2 h and 4 h). These observations suggest that the probability of binding events in ChIP PET clusters located in canonical promoter region could reflect a relative avidity of STAT1 binding and casuistically increase the probability of regulation of the transcription of STAT1 direct gene targets.

4. Discussion and Conclusion

In this work, a novel model of TF–DNA binding and algorithm for accurate identification of binding sites from short reads generated from ChIP-Seq, ChIP-PET, and SACO experiments is presented.

We demonstrated that the empirical distributions for all the studied ChIP-based data sets are well fitted by a mixture model with specific component of BEs described by the skewed Pareto-like distribution functions (generalized power law), whose shape depends in a predictable manner on the sample size (13). Such distributions can be generated as limiting distribution at K–W random process where the birth and death intensities are linear functions of (binding) events (25). The power law for analysis of ChIP-Seq data analysis was also recently used in (14, 32–34).

Our probabilistic model and computational algorithm allow us not only to mathematically describe a common law (Figs. 2.5–2.7) of TF–DNA binding but also to estimate the number and fraction of specific BSs for TF studied in ChIP-based experiment even if the data set is essentially incomplete and enriched with high-noisy events.

The sensitivity and the specificity of our model are demonstrated by applying the model to analysis of different ChIP-Seq data sets across different platforms. In this work, we studied binding statistics of five biologically essential and well-characterized human transcription factors: ERE (estrogene receptor- α), CREB (cAMF-response element), Nanog (Nanog homeobox), Oct4 (POU class 5 homeobox 1), and STAT1 (signal transducer and activator of transcription protein 1). By our estimates the number of BS in the genome are 66.7×10^3 , 28.2×10^3 , 15.5×10^3 , and 26.4×10^3 for Nanog (ChIP-Seq), Oct4 (ChIP-Seq), CREB (SACO), and ERE (ChIP-PET) TFs, respectively. For STAT1 TF the estimates of the number of BS in the human genome are 12.25×10^3 (ChIP-PET), 17.66×10^3 (ChIP-Seq, high-avidity BS), and 5.99×10^3 (ChIP-Seq; low-avidity BS). Our new result is the identification of 64 novel STAT1 TFBS and their potential gene targets in unstimulated cells.

These findings could provide an insight in the field of study of basal transcription machinery and predict new gene targets for direct STAT1 transcription control.

We determined that the specificity of all studied here experiments was high (91–96%). That results consist of published estimates of the specificity of ChIP-based methods (4, 11, 15, 16, 21). The BE sensitivity asks for a question: how many physical BSs, including low-avidity BSs present in the genome of a given cell in given environmental conditions. Sensitivity is difficult to assess experimentally because of the lack of well-reliable bench markers and methods for detection of low binding avidity. We used a

computational approach to estimate the sensitivity of ChIP-based experiments. By our estimations, the sensitivity of all ChIP-based methods is low: 6.3%, 4.8%, 10.2%, 4.6% for Nanog (ChIP-Seq), Oct4 (ChIP-Seq), CREB (SACO), ERE (ChIP-PET), respectively, and 4.36% (ChIP-PET), 12.1% (ChIP-Seq, high-avidity BS); 0.97% (ChIP-Seq; low-avidity BS) for STAT1 TF (**Table 2.4, Fig. 2.6**). That surprisingly low sensitivity levels of the current ChIP-based sequencing methods for identification of TF–DNA binding can be associated with (i) a large fraction of noisy sequences forming low- and moderate-avidity binding events and (ii) missing specific ChIP-derived sequences (not-detected sequences) due to the limited sample size experimental data set and/or suboptimal design of experiments. The efficiency of sequencing (the percentage specific DNA sequences at a given specificity cutoff peak height, **Table 2.4**), defined by qPCR-ChIP, was also low: 3.6, 1.0, 8.8, 50.3, 0.54% for Nanog, Oct4, ERE, SREB, STAT1 (ChIP-PET) TFs, respectively.

We could conclude that although ChIP-Seq is a powerful technique, it still produces essentially incomplete and noisy-rich sequences under representing the low- and moderate-avidity DNA–protein binding events of TF in complex genomes. While a higher number of reads may increase sensitivity and resolution, it may also increase the fraction of noise reads. In fact, subsampling in all ChIP-Seq and ChIP-PET data sets showed that the noise component increases when the data set sample size becomes larger. This common property of ChIP-based sequencing methods is a hard problem for a reliable detection of real BSs even in very large-size samples (**Figs. 2.6 and 2.7, Table 2.4**). In this case, other factors, like specificity of antibodies, optimal (shorter/homogeneous), length of ChIP DNA fragments, and better computational processing of raw data, have a direct impact on the sensitivity and the resolution of identified BSs are desired.

Similar to what was recently described in (10, 14, 32), we can suggest that the probability of binding events can be defined by the distance of BS from spatio gene transcription start site could reflect the distribution of relative avidity of binding (STAT1–DNA binding **Fig. 2.9A,B,C**). However, poor correlation between the relative binding avidity, expression pattern, canonical motif presenting, and enrichment of ChIP-derived TF–DNA binding events was also indicated for TF with non-canonical (distant, >10 kb) location of BS (for instance, ERE binding (12)).

Finally, we would like to indicate that (i) an adequate experimental design, standardization, and optimization; (ii) manual and automatic prefiltration of ChIP-derived sequences; (iii) non-redundant mapping of the sequences onto the complex genomes; (iv) filtering and clustering procedures of ChIP-derived DNA sequences in the different regions of the chromosomes; (v) adequate statistical methods to define a binding events (e.g., cluster peak height,

cluster overlap span); (vi) minimization of signal-to-noise ratio; (vii) adequate controls and statistical modeling of background (noise) signals; and (viii) deep data sampling as an essential in the context of improvement of sensitivity of ChIP-based experiment. Mapping of regulatory sequence (motifs, CpG, BSs for cooperative TFs, expression data, etc.) could provide an additional benefit in the identification of real binding events and its links with gene expression patterns.

Acknowledgments

I thank Chia Lin Wei, Chiu Kow Ping, and Ruan Yujin for providing access to ChIP-Seq data sets of T2G DB and for very useful discussions of their ChIP-PET method. I also thank Piroon Jenjaroenpoorn, Yuri Orlov, and Onkar Singh for partial but important computational support of analytical part of this work. I express my special acknowledgment to Yuri Nikolsky for his stimulated interest in this study. This work was supported by BII/A-Star, Singapore.

References

1. Ren, B., Robert, F., Wyrick, J.J., Aparicio, O., Jennings, E.G., Simon, I., Zeitlinger, J., Schreiber, J., Hannett, N., Kanin, E., Volkert, T.L., Wilson, C.J., Bell, S.P., and Young, R.A. (2000) Genome-wide location and function of DNA binding proteins. *Science* **290**(5500), 2306–9.
2. Kim, T.H. and Ren, B. (2006) Genome-wide analysis of protein-DNA interactions. *Annu. Rev. Genom. Human Genet.* **7**, 81–102.
3. Hartman, S.E., Bertone, P., Nath, A.K. et al. (2005) Global changes in STAT target selection and transcription regulation upon interferon treatments, *Genes Dev.* **19**(24), 2953–68.
4. Wei, C.L., Wu, Q., Vega, V.B. et al. (2006) A global map of p53 transcription-factor binding sites in the human genome. *Cell* **124**(1), 207–19.
5. Stormo, G. DNA binding sites: Representation and discovery. *Bioinformatics* **16**, 16–23.
6. Down, T.A. and Hubbard, T.J. (2005) NestedMICA: sensitive inference of over-represented motifs in nucleic acid sequence. *Nucleic Acids Res.* **33**, 1445–53.
7. Lovegrove, F.E., Peña-Castillo, L., Mohammad, N., Liles, W.C., Hughes, T.R., and Kain, K.C. (2006) Simultaneous host and parasite expression profiling identifies tissue-specific transcriptional programs associated with susceptibility or resistance to experimental cerebral malaria. *BMC Genomics* **7**, 295.
8. Fernandez, P.C., Frank, S.R., Wang, L., Schroeder, M., Liu, S., Greene, J., Cocito, A., and Amati, B. (2006) Genomic targets of the human c-Myc protein. *Genes Dev.* **17**(9), 1115–29.
9. Loh, Y.H., Wu, Q., Chew, J.L., et al. (2006) The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. *Nat. Genet.* **38**(4), 431–40.
10. Zeller, K.I., Zhao, X., Lee, C.W., et al. (2006) Global mapping of c-Myc binding sites and target gene networks in human B cells. *Proc. Natl. Acad. Sci. U S A* **103**(47), 17834–9.
11. Chen, X., Yuan, P., Fang, F., et al. (2008) Integration of external signalling pathways with the core transcriptional network in embryonic stem cells. *Cell* **133**, 1106–17.
12. Lin, C.Y., Vega, V.B., Thomsen, J.S., Zhang, T., Kong, S.L., Xie, M., Chiu, K.P., Lipovich, L., Barnett, D.H., Stossi, F., George, J., Kuznetsov, V.A., Lee, Y.K., Cham, T.H., Palanisamy, N., Katzenellenbogen, B.S., Miller, L.D., Ruan, Y., Bourque, G., Wei, C.L., and Liu, E.T. (2007) Whole-genome cartography of estrogen receptor α binding sites. *PLoS Genet.* **3**(6), e87.
13. Kuznetsov, V.A. (2002) Statistics of the numbers of transcripts and protein sequences

- encoded in the genome. In: *Computational and Statistical Methods to Genomics* (W. Zhang and I. Shmulevish, Eds.; 1st Ed.). Kluwer: Boston-Dordrecht, pp. 125–71.
14. Kuznetsov, V.A., Orlov, Y.L., Ruan, Y., and Wei, C.L. (2007) Computational analysis of genome-scale avidity distribution of TFBS in ChIP-PET experiments. *Genome Informatics* **19**, 83–94.
 15. Johnson, D.S., Mortazavi, A., Myers, R.M., and Wold, B. (2007) Genome-wide mapping of in vivo protein-DNA interactions. *Science* **316**(5830), 1497–502.
 16. Robertson, G., Hirst, M., Bainbridge, M. et al. (2007) Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat. Methods* **4**(8), 651–7.
 17. Barski, A., Cuddapah, S., Cui, K. et al. (2007) High-resolution profiling of histone methylations in the human genome. *Cell* **129**(4), 823–37.
 18. Massie, C.E. and Mills, I.G. (2008) ChIP-ping away at gene regulation. *EMBO Rep.* **9**(4), 337–43.
 19. Mardis, E.R. (2007) ChIP-Seq: Welcome to the new frontier. *Nat. Methods*, **4**, 613–4.
 20. Euskirchen, G.M., Rozowsky, J.S., Wei, C.L., Lee, W.H., Zhang, Z.D., Hartman, S., Emanuelsson, O., Stolc, V., Weissman, S., Gerstein, M.B., Ruan, Y., and Snyder, M. (2007) Mapping of transcription factor binding regions in mammalian cells by ChIP: comparison of array- and sequencing-based technologies. *Genome Res.* **17**(6), 898–909.
 21. Impey, S., McCorkle, S.R., Cha-Molstad, H., Dwyer, J.M., Yochum, G.S., Boss, J.M., McWeeney, S., Dunn, J.J., Mandel, G., and Goodman, R.H. (2004) Defining the CREB regulation: a genome-wide analysis of transcription factor regulatory regions. *Cell* **119**(7), 1041–54.
 22. Oszlak, F., Song, J.S., Liu, X.S., and Fisher, D.E. (2007) High-throughput mapping of the chromatin structure of human promoters. *Nat. Biotechnol.* **25**(2), 244–8.
 23. Lieb, J.D., Liu, X., Botstein, D., and Brown, P.O. (2001) Promoter-specific binding of Rap1 revealed by genome-wide maps of protein-DNA association. *Nat. Genet.* **28**, 327–34.
 24. Bhinge, A.A., Kim, J., Euskirchen, G.M., Snyder, M., and Iyer, V.R. (2007) Mapping the chromosomal targets of STAT1 by Sequence Tag Analysis of Genomic Enrichment (STAGE). *Genome Res.* **17**(6), 910–6.
 25. Kuznetsov, V.A. (2003) Family of skewed distributions associated with the gene expression and proteome evolution. *Signal Processing* **83**, 889–910.
 26. Johnson, N.L., Kotz, S., and Balakrishnan, N. (1997) *Discrete Multivariate Distributions*, John Wiley & Sons, Inc.: New York, p. 299.
 27. Kuznetsov, V.A. (2006) Emergence of size-dependent networks on genome scale. In: *Lecture Series on Computer and Computational Sciences* (Brill Acad. Publishers: The Netherlands), **7A**, pp. 754–7.
 28. Scafoglio, C., Ambrosino, C., Cicatiello, L., Altucci, L., Ardovino, M., Bontempo, P., Medici, N., Molinari, A.M., Nebbioso, A., Facchiano, A., Calogero, R.A., Elkon, R., Menini, N., Ponzone, R., Biglia, N., Simondi, P., De Bortoli, M., and Weisz, A. (2006) Comparative gene expression profiling reveals partially overlapping but distinct genomic actions of different antiestrogens in human breast cancer cell. *J. Cell. Biochem.* **98**(5), 1163–84.
 29. Chiu, K.P., Wong, C.H., Chen, Q. et al. (2006) PET-Tool: A software suite for comprehensive processing and managing of Paired-End diTag (PET) sequence data. *BMC Bioinformatics* **7**, 390.
 30. Wormald, S., Hilton, D.J., Smyth, G.K., and Speed, T.P. (2006) Proximal genomic localization of STAT1 binding and regulated transcriptional activity. *BMC Genomics* **7**, 254.
 31. Hartman, S.E., Bertone, P., Nath, A.K., Royce, T.E., Gerstein, M., Weissman, S., and Snyder, M. (2005) Global changes in STAT target selection and transcription regulation upon interferon treatments. *Genes Dev.* **19**(24), 2953–68.
 32. Jothi, R., Cuddapah, S., Barski, A., Cui, K., and Zhao, K. (2008) Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data. *Nucleic Acids Res.* **36**(16), 5221–31.
 33. Zhang, Z.D., Rozowsky, J., Snyder, M., Chang, J., and Gerstein, M. (2008) Modeling ChIP sequencing in silico with applications. *PLoS Comput. Biol.* **4**(8), e1000158.
 34. Kuznetsov, V.A., Singh, O., Huck, Ng, and Wei, C.L. (2008) Modeling and prediction of DNA-protein interaction events of transcription factors (TF) in ChIP-seq experiments. In: *The Sixth International Conference on Bioinformatics of Genome Regulation and Structure (BGRS'2008)*. Institute of Cytology and Genetics SB RAS: Novosibirsk, Russia, June 22–28, 2008, p. 131. ISBN 978-5-91291-005-0.



<http://www.springer.com/978-1-60761-174-5>

Protein Networks and Pathway Analysis

Nikolsky, Y.; Bryant, J. (Eds.)

2009, XIV, 410 p. 118 illus., Hardcover

ISBN: 978-1-60761-174-5

A product of Humana Press