Journal of Physical Chemistry & Biophysics
ISSN: 2161-0398

Zhou et al. J Physic Chem Biophysic 2011, S:4
DOI: 10.4172/2161-0398.S4-002

Research Article Open Access

Bias in Genome Scale Functional Analysis of Transcription Factors using Binding Site Data

Jair Zhou[1,2], Haixia Rachel Li[1,3] and R. Krishna Murthy Karuturi[1]*

[1]*Computational & Mathematical Biology, Genome Institute of Singapore, Singapore*
[2]*Nanyang Technological University, Singapore*
[3]*Institute of Molecular & Cellular Biology, Singapore*

Abstract

Genome scale functional analysis of a transcription factor is performed by mapping its genomic binding sites to genes using a nearness criterion followed by enrichment analysis for the pre-defined functional categories such as GO categories. It can result in biased assessment of functionality of transcription factors. In the view of the enormous work underwent using this procedure and its simplicity and effectiveness, it is important to understand the bias in this procedure and the factors influencing it. We show that the bias originates from widely varying gene lengths, intergenic regions and gene density. We also demonstrate that the bias depends on multiple factors such as the nearness criterion and the choice of the associated parameters and the distribution of the binding sites of the transcription factor. Furthermore, we propose a resampling based procedure called reFABS for unbiased functional analysis of binding sites.

Keywords: GO Analysis; Bias analysis; Functional Analysis of Transcription Factors; Binding site mapping

Availability: The software reFABS will be available upon request.

Introduction

The identification of the DNA binding sites of a transcription factor and the genes potentially regulated is important in understanding its cellular functions. The high-throughput technologies such as ChIP-chip [1] and ChIP-seq [2] have made the genome wide identification of the binding sites of various transcription factors feasible. Many transcription factors have been found to bind, directly or through tethering mechanism, at thousands of DNA sites in higher eukaryotes. But, different transcription factors exhibit different distributions of their binding sites with reference to genes. For example, >70% of GABP and SRF binding sites are within 2 kbp of genes whereas only 53% of the NRSF (neuron-restrictive silencer factor) binding sites are within 2 Kbp of a gene [3]. Estrogen receptor is known to bind at many (>50%) sites far from genes [4,5]. Similarly, FOXA2 is also known to bind in extended gene regions [6], ~50% are intragenic and >50kbp away from genes.

Due to lack of hi-throughput experimental procedures and data for direct genome scale identification of the genes regulated by different transcription factor binding sites, the potentially regulated genes are identified by mapping the binding sites to genes using the k-nearest genes criterion. By this, a binding site is mapped to K nearest genes within a pre-defined distance i.e. within a window of pre-defined size. Its two extreme cases are commonly used: (1) nearest gene assignment (K=1) which maps a binding site to the nearest gene within a pre-defined distance or window i.e. a binding site is mapped at most to one gene; and, (2) nearest binding site assignment (K=∞) which maps a gene to a nearest binding site within a pre-defined window i.e. a binding site may be mapped to all genes within the window. The reference locus of a gene for such a mapping could be its transcription start site (TSS), intragenic boundaries (close to TSS or TES) or intragenic boundaries including the intragenic region (body). The hypothesis driving the k-nearest genes criterion is closer the binding site to a gene, higher the chance that it may be regulating the respective gene. After having mapped the binding sites of a transcription factor to genes, the analysis of the potentially regulated functions such as biological processes and its role in the responsiveness of genes is routinely performed. It is carried out by enrichment analysis on the mapped genes using one of the standard enrichment analysis procedures such as GSEA [7], GSA [8] and GO term finder [9].

The fundamental assumption in the above conventional procedure is that the underlying null-hypothesis of the overall procedure of mapping and enrichment is same as that of enrichment in the mapped genes. McLean et al. [10] pointed out the error in the assumption and presence of bias in the conventional functional analysis of transcription factors and proposed a procedure called GREAT to reduce or remove the bias. However, the sources and extent of the bias have not yet been analyzed and understood. In the view of the enormous number of studies used this procedure, it is important to understand the extent of the bias for different mapping criterion. Hence, in this paper, we analyze the sources of the bias and quantify it. Furthermore, we propose a resampling based procedure, called reFABS, for functional analysis of the transcription factors from their binding sites and apply the procedure on two transcription factor binding site datasets: Estrogen Receptor (ER), Serum Response Factor (SRF) and GA binding protein (GABP). Our reFABS procedure is fundamentally different from GREAT in the way the multiple binding sites mapped to a single gene are considered. We present our results for the cases K=1 and K=∞ for two different gene references (TSS and Body).

Analysis and Methods

We demonstrate the bias in the conventional functional analysis

***Corresponding author:** R. Krishna Murthy Karuturi, Computational & Mathematical Biology, Genome Institute of Singapore, Singapore; Tel: +65-680808040; Fax: +65-68088303; E-mail: karuturikm@gis.a-star.edu.sg

Received December 12, 2011; **Accepted** January 20, 2012; **Published** January 24, 2012

Citation: Zhou J, Li HR, Murthy Karuturi RK (2012) Bias in Genome Scale Functional Analysis of Transcription Factors using Binding Site Data. J Physic Chem Biophysic S4:002. doi:10.4172/2161-0398.S4-002

procedure by applying it on randomly chosen binding sites from the human genome. We randomly sampled 10K sites from entire human genome (build Hg18) and mapped them to Refseq genes using nearest gene assignment criterion with window size of 100 Kbp and TSS as reference. For comparison and as a control, we randomly selected 4000 genes as 10K randomly selected binding sites map to ~ 4000 genes based on the mapping criterion. We performed gene ontology (GO) enrichment analysis using Fisher's exact test [11] (a popularly used test for enrichment analysis) for both sets of genes and selected GO terms at p-value < 0.05. We repeated this procedure 5 times and counted the number of GO categories that were repeatedly selected. Ideally, in the absence of bias, the number of GO categories repeatedly selected should be close to 0. Further, to demonstrate the variation of bias with the size of the window, we repeated the experiment for 430K randomly selected sites mapped using the same criterion but with a different window size of 1 Kbp to ~ 4000 genes.

In Figure 1, we show the cumulative distribution of the number of GO terms repeatedly selected for each selection of genes over all 5 runs. It shows that the direct random sampling of genes results in very few GO categories enriched indicating absence of bias. Whereas, the analysis of the genes obtained by mapping 10K random sites using window of 100 kbp resulted in 40 GO categories repeatedly selected in all 5 runs and ~ 200 were commonly selected in at least two runs which demonstrates large bias in the functional analysis. But it reduced substantially for the mapping using reduced window size of 1kbp. This simulation clearly shows that there is a bias for certain GO categories and discrepancy in the null-hypotheses used. It increases with the increased window size.

Non-uniform genomic feature lengths is the source of bias

The nearest gene assignment (K = 1) is analytically tractable for the bias quantification which we provide here. The nearest gene assignment (K = 1) criterion introduces so called assignment domain for each gene. It is a genomic range around the gene such that a binding site is assigned (mapped) to the gene if it falls within the assignment domain of that gene (Figure 2A). A variation in the assignment domain lengths from gene to gene results in the bias as the probability of a random site mapped to a gene is proportional to its assignment domain length. The assignment domain of a gene in a genome, as shown below, is function of the window size, gene reference and the lengths of flanking intergenic regions and its intragenic region in the genome of interest.

Precisely, if the window size is W and the reference is TSS then the assignment domain of gene X, denoted as D_X, is given by (depicted in Figure 2 (B))

$$D_X = \min(0.5|T_X - T_U|, W) + \min(0.5|T_D - T_X|, W)$$

Where, T_X is the locations of the TSS of gene X, T_U and T_D are the locations of the TSS of genes upstream and downstream of X. D_X varies from gene to gene as the lengths of intergenic and intragenic regions vary. The factor 0.5 signifies competition between genes flanking the upstream intergenic region if $|T_X\text{-}T_U| < 2W$, then the nearest gene is assigned. Similar interpretation holds for the downstream gene. If all $|T_X - T_U| \geq 2W$ and all $|T_D - T_X| \geq 2W$ then $D_X = 2W$, uniform assignment domain lengths, which implies even the conventional procedure can be expected to give unbiased results.

Similarly, if the window size is W from a gene including intragenic region (body as reference), then the assignment domain of gene X, D_X, is given by (Figure 2 (C))

$$D_X = L_X + \min(0.5UP_X, W) + \min(0.5DN_X, W)$$

Where L_X is the length of the intragenic region of gene X, UP_x and DN_x are the lengths of the flanking intergenic regions upstream and downstream of gene X respectively. It may result in even worse bias as shown by the new assignment domain calculation i.e. variation in L_X also contributes to the bias. If all intergenic regions are of length ≥ 2W and if $L_X = L$ (a constant which is not true for almost all eukaryotic genomes) then $D_X = L+2W$ which implies no bias for any gene and the conventional enrichment procedure can be carried out. As none of them would be true (especially the condition $L_X = L$) especially in case of the genomes of higher eukaryotes, we expect large variation in D_X from gene to gene as shown in Figure 3 for the human genome. As the lengths of the genomic features such as genes, intergenic regions (L_X, UP_X, DN_X, $|T_X - T_U|$ and $|T_D\text{-}T_X|$) vary highly from gene to gene in complex genomes, D_X is also expected to vary highly from gene to gene. The probability (p_X) of a gene being mapped by a random site is proportional to its D_X i.e.

$$p_x = \frac{D_X}{\sum_{\forall y} D_y} \frac{\sum_{\forall y} D_y}{GenomeLength} = \frac{D_X}{GenomeLength}$$

If D_X is not constant, but spread over a finite range, then p_X will not be

Figure 1: Demonstrating bias in the functional analysis of transcription factors using random selection of sites and mapping them to genes. The bias is higher for larger window and non-existent for direct gene based analysis.

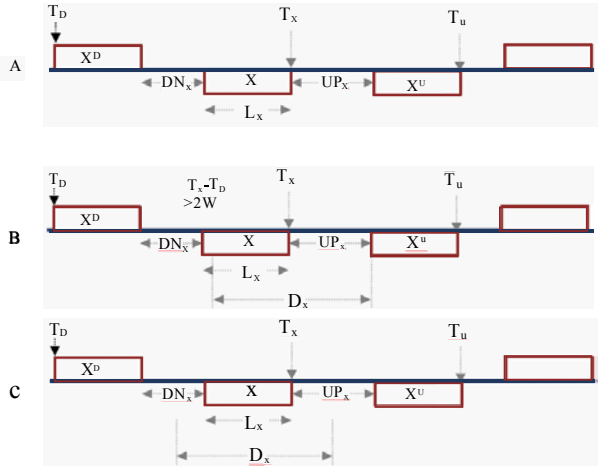

Figure 2: Illustration of assignment domains for nearest gene assignment criterion. (A) Notations used for the calculation of assignment domain; (B) Assignment domain (D_x) calculation for TSS reference for gene X, D_X = $0.5(T_U\text{-}T_X)$ + W as $(T_U\text{-}T_X) < 2W$ and $T_X\text{-}T_D > 2W$; (C) Calculation of D_X for Body reference, D_X is $L_X+DN_X/2+UP_X/2$ as $DN_X < 2W$ and $UP_X < 2W$.

same for different genes which results in genes with long D_X would be mapped more often even for randomly selected binding sites. It results in the false enrichment of gene categories containing genes with long D_X. This is in contrast to the equal probability $(p_X = p)$ assumption made in the conventional functional analysis of transcription factors through mapping their binding sites. Moreover, the bias would be stronger if the distribution of D_X is close to uniform distribution in a range of $[0, D_{max}]$ and D_{max} is large enough. The distributions of D_X in the human genome for different window sizes and references are shown in the Figure 3a. The expected bias may be quantified by the entropy of the distribution of D_X i.e. lower entropy implies lower bias and higher entropy results in higher bias. Figure 3b shows that smaller W with TSS as reference gives rise to lowest bias and gene body as reference is heavily biased irrespective of the choice of W due to large variation in L_X in the human genome. The bias can be reduced or eliminated if W is sufficiently small for the TSS reference, but it may result in a fewer binding sites being mapped for many transcription factors which tend to bind in extended regions resulting in lower power for the enrichment tests and failure to identify of some functions of the transcriptions factor under study.\

reFABS: Resampling procedure for unbiased functional analysis of transcription factor binding sites

The bias resulting from non-constant D_X can be eliminated by using gene-category specific null distribution that accounts for variation in p_X or D_X which could be different for different gene categories as they are composed of different gene sets. We propose a resampling procedure, called reFABS, to estimate true null distribution for each gene category and the truly enriched categories. Our reFABS procedure (Table 1), given M binding sites and the mapping criterion **C** which is defined by the choice of K, W and reference in the context of this paper, has two major steps: (A) estimating number of random sites to be sampled; (B) estimating statistical significance of enrichment.

The step A is important as the distribution of given binding sites V may not be as uniformly distributed as the sampled binding sites which may result in another bias in the enrichment analysis. In practice, for a given **C**, M true binding sites of a transcription factor map to many more genes compared to M randomly drawn sites. But, as our **C** is a fixed window based criterion, it is sufficient to have the randomly drawn sites map to the similar number of genes as that of the M true binding sites.

Results

We investigated bias in the functional analysis of three transcription factors, Estrogen Receptor (ER), Serum Response Factor (SRF) and GA Binding Protein (GABP), in the context of Gene Ontology (GO) enrichment analysis. ER is a ligand-activated transcription factor composed of several domains important for hormone binding, DNA binding and activation of transcription. It binds to ERE (estrogen receptor element) elements and involves in many cellular processes including growth, differentiation and function of the reproductive system. It is a target of the drugs used to treat breast cancer [12]. SRF [12] encodes a ubiquitous nuclear protein that stimulates both cell differentiation and proliferation. It binds to the serum response element (SRE) and is required for cardiac differentiation and maturation. GABP [12] encodes the GA binding protein transcription factor, beta subunit. This protein forms a tetrameric complex with the alpha subunit and stimulates transcription of target genes. The encoded protein may be involved in activation of cytochrome oxidase expression and nuclear control of mitochondrial function. The summary of the datasets is given in Table 2. They were obtained from different technologies and

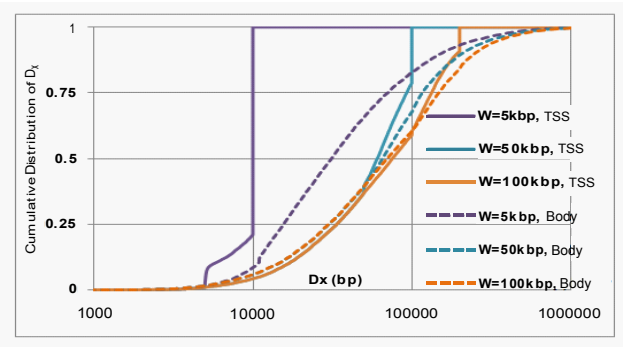

Figure 3a: Cumulative distributions of assignment domains in the human genome (build #18) for different W and reference for nearest gene assignment criterion (K=1). Variation in D_X for Body reference is large and does not change dramatically as in the case of TSS reference.

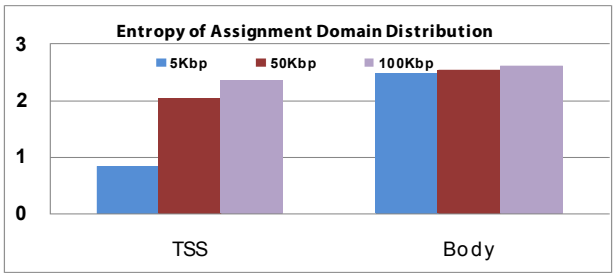

Figure 3b: Entropy of distribution of D_X in the human genome (build #18) for different choices of W (=5Kbp, 50Kbp and 100Kbp) and reference (TSS and Body) for nearest gene assignment criterion (K=1). Higher entropy may mean higher bias in the analysis. It shows that low entropy of TSS reference with W=5kbp implies lowest bias and the entropy (and hence the bias) increases with increasing window size. For body reference, the entropy is high and remains almost same irrespective of the choice of W which is indicative of the large bias for studies involving body reference even for smaller W.

reFABS = function(V, C, R=5, Q=1000):
V: Set of binding sites
C: Mapping criterion
R: No. of sampling runs for step A
Q: No. of sampling runs for step B

A. Estimate number of sites to be sampled randomly
1. Map V using C and note down the number of genes (N) mapped
2. Treat all binding sites mapped to a gene as duplicates and let M be the number of such non-duplicate binding sites.
3. Sample M sites randomly
4. Map the M random sites using C and note down number of genes mapped
5. Repeat steps 3 to 4 for R times
6. N' is the average number of genes mapped in the steps 3-4 over R runs
7. M', the number of random sites to be sampled, = M×N/N'

B. Estimate Significance of Enrichment
1. Map V to genes using C. Let n_z be the number of genes mapped to the predefined gene category Z
2. Randomly sample M' sites from the reference genome
3. Apply the criterion **C**, map M' binding sites to genes
4. Let n_{iz}' be the number of mapped genes using M' in Z in i^{th} sampling run.
5. Repeat steps 2-4 for Q times
6. p-value of the category Z, P_z, is the fraction of Q sampling trials yielded $n_{iz}' \geq n_z$ i.e. $P_z = |\{i \mid i = 1 \ldots Q \text{ and } n_{iz}' \geq n_z\}|/Q$

Table 1: Pseudocode for reFABS procedure for unbiased enrichment analysis for binding sites of transcription factors.

TF	No. of binding sites	Reference	Mapping technology	Name of Dataset
ER	3665	Carroll et al. [5]	ChIP-chip	CarH
ER	6000	Carroll et al. [5]	ChIP-chip	CarL
ER	1234	Lin & Vega et al. [4]	ChIP-PET	LinH
ER	3000+	Lin & Vega et al. [4]	ChIP-PET	LinL
SRF	2429	Anton Valouev et al. [3]	ChIP-seq	SRF
GABP	6442	Anton Valouev et al. [3]	ChIP-seq	GABP

Table 2: Datasets of transcription factor binding sites for the GO enrichment analysis.

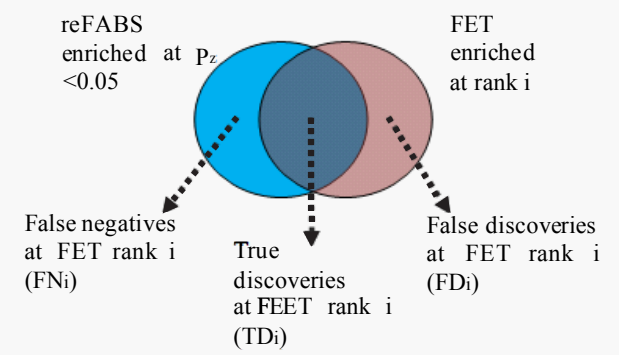

Figure 4: Definitions used for true discoveries (TD), false discoveries (FD) and false negatives (FN). The GO categories with p-value < 0.05 from reFABS were taken as complete true discoveries.

from different laboratories with number of binding sites widely ranging by ~5-fold, from 1234-6000. Furthermore, we used two sets of mapped binding sites for ER from two different research groups at two different levels of confidence using different binding site mapping technologies. LinH and CarH are high confidence mapped binding sites. Whereas, LinL and CarL are the low confidence mapped binding sites.

We mapped each set of binding sites to genes using nearest gene assignment criterion with reference to TSS as well as intragenic region (body) as described earlier. We analyzed bias for three different choices of W: 5 kbp, 50 kbp and 100 kbp. We used Fisher's exact test (FET) [11] to find the GO terms enriched for the set of genes mapped to the binding sites for each combination of W and reference in the conventional procedure. We also carried out our reFABS procedure and compared the results to quantify the bias in the functional analysis results obtained using the conventional procedure on real biological datasets.

To quantitatively understand the bias in the conventional functional analysis of the above transcription factors, we used the GO categories enriched at the p-value < 0.05 using the reFABS procedure as true positives and evaluated their distribution in the enrichment ranking generated by conventional analysis using Fisher's Exact Test (FET). We generated plots of sensitivity vs. false discovery rate (FDR) curves. The definition of true discoveries (TD), false discoveries (FD) and false negatives (FN) are illustrated in the Figure 4.

FDR at rank i (FDR_i) is defined [13] as

$$FDR_i^* = \frac{FD_i}{i}$$

$$FDR_i = \min(FDR_i^*, FDR_{i+1}) \text{ where } 1 \le I \le n\text{-}1$$

Where n is number of GO terms analyzed

$$Sensitivity_i = \frac{TD_i}{TD_i + FN_i}$$

The sensitivity-FDR curves that can demonstrate bias in the functional analysis of transcription factors ER, SRF and GABP are shown in Figure 5. The solid curves are for TSS reference and broken curves are for body reference. Higher (lower) FDR is indicative of higher (lower) bias. In the absence of bias, the FDR should be very close to 0 for most of the sensitivity range. The sensitivity-FDR curves show increasing FDR with increasing window size for TSS reference, though it varies from dataset to dataset. The increase in FDR indicates that the bias in the conventional GO enrichment analysis is also increasing with increased window size as predicted from our analysis in the previous section (Figure 3) and correlates well with the entropy measure we proposed in the previous section. On the other hand, the mapping with gene body as reference appears to be biased even for a window of 5 kb due to largely varying gene lengths (L_X) as shown in Figure 3. This observation is consistent with our entropy calculation as a measure of bias. Mapping using gene body as reference appears to be worse than mapping with TSS as reference for W = 100 Kb which indicates that the variation in L_X is the major contributor to the bias for body reference. The bias is negligibly affected by the choice of W which indicates that the gene body length variation is the major factor if gene body is taken as reference. Further, the bias is more for low confidence sets of binding sites compared to the respective high confidence binding sites. It clearly indicates that the bias is a function of the quality of the binding site mapping.

Similarly, the analysis of nearest binding site assignment (K=∞) criterion also demonstrated larger bias than that of nearest gene assignment criterion with gene body reference (data not shown). This is mainly due to large variation of gene density across the human genome and high correlation structure resulted from the choice of K=∞.

Discussion

We demonstrated the statistical bias in the conventional functional analysis of transcription factors and the genomic factors influencing it. In the case of nearest gene assignment (K=1) criterion for mapping, the bias stems from the presence of so called assignment domain whose length can change considerably from gene to gene and it depends on the reference as well as the choice of the window size. We have shown that the bias increases with the increasing uniformity of the distribution of the assignment domain lengths. The nearest gene assignment criterion with TSS reference appears to be least biased and gene body reference is more biased. In case of nearest binding site assignment (K=∞), the correlation structure resulting from assigning a binding site to multiple genes also play a key role along with the choice of window size and reference. The choice of K=∞ gives rise to the most bias in the analysis. The analytical quantification of bias for different choice of K>1 may be discussed in our future work.

Our analysis has shown that the functional analyses that used TSS as reference with smaller W (≤ 5kbp) may not be biased except for false negatives due to mapping only a fraction of binding sites to genes. On the other hand, the bias introduced by larger window size might vary from study to study even for TSS reference. But, for body reference, the results of the most of the studies could be biased irrespective of the choice of the W.

However, the choice of W and the mapping criterion should be chosen with care considering the binding site distribution. For example, GABP has ~80% of binding sites at ± 2 Kbp of TSS of a gene

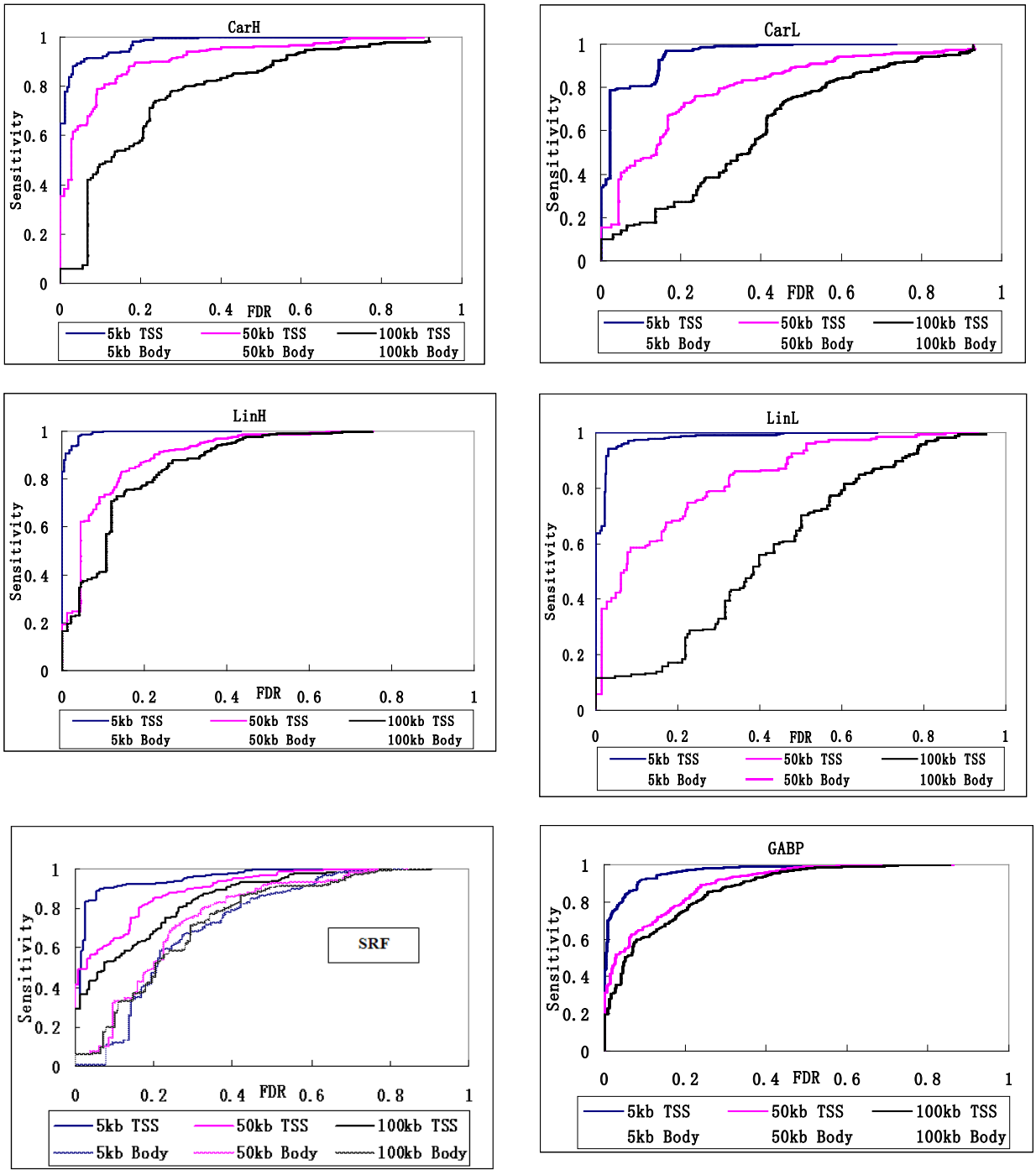

Figure 5: Sensitivity-FDR curves that demonstrate bias in the functional analysis of the transcription factors ER (CarH, CarL, LinH and LinL,), SRF and GABP.

which means it is better to choose nearest gene assignment criterion with TSS reference and W=2Kbp as the remaining genomic regions contain insignificant binding activity. In contrast, <10% of ER binding sites are within 5Kbp from TSS which means we need to consider extended regions from TSS to elicit the functions regulated.

Though our bias analysis is presented for FET based enrichment analysis, it is equally applicable for other methods of gene enrichment analysis. Our proposed resampling procedure, reFABS, is an alternative to GREAT in that we provide more choice for reference and K. Another major difference stems from how multiple binding sites mapped to a gene are treated: GREAT treats them as individual hits and our reFABS procedure treats them as one hit i.e. weight of a gene does not depend on the number of binding sites in D_x in our reFABS procedure. This might have an effect on the analysis of GO terms with genes that have

highly different assignment domains. Our future work will involve understanding the differences between the results obtained by GREAT and reFABS.

Our reFABS procedure can be optimized for running time, if higher precision is warranted in the calculation of p-value for the enrichment by increasing the number of resampling runs Q, by the following steps: (1) estimate p-value using 1000 resampling runs; (2) exclude all categories whose p-value > 0.1; (3) exclude all genes that do not contribute to the selected functional categories while keeping the original assignment domains of genes; and (4) run the remaining resampling runs for the higher precision in p-value estimation. The exclusion of about 90% of functional categories and the associated genes from the analysis will significantly improve computation time. Nonetheless, we are devising a computationally efficient method that does not require resampling of binding sites and mapping them. Towards this goal, our analytical understanding of the bias would greatly help.

Acknowledgments

We thank Huaien Luo, Ian Lee and Juntao Li for their valuable comments during this work. We also thank Prof Edison Liu and Prof Neil Clarke for their support. The research was supported by Genome Institute of Singapore, Biomedical Research Council and Agency for Science Technology and Research (A-STAR).

References

1. Buck MJ, Lieb JD (2004) ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. Genomics 83: 349-360.
2. Kharchenko PV, Tolsturukov MY, Park PJ (2008) Design and analysis of ChIP-seq experiments for DNA-binding proteins. Nat Biotechnol 26: 1351-1359.
3. Valouev A, Johnson DS, Sundquist A, Medina C, Anton E, et al. (2008) Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. Nat Methods 5: 829-834.
4. Lin CY, Vega VB, Thomsen JS, Zhang T, Kong SL, et al. (2007) Whole-Genome Cartography of Estrogen Receptor alpha Binding Sites. PLoS Genet 3: e87.
5. Carroll JS, Meyer CA, Song J, Li W, Geistlinger TR, et al. (2006) Genome-wide analysis of Estrogen Receptor binding sites. Nat Genet 38: 1289-1297.
6. Wederell ED, Bilenky M, Cullum R, Thiessen N, Dagpinar M, et al. (2008) Global analysis of *in vivo* Foxa2-binding sites in mouse adult liver using massively parallel sequencing, Nucleic Acids Res 36: 4549-4564.
7. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, et al. (2005) Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 102: 15545-15550
8. Efron B, Tibshirani R (2007) On testing the significance of sets of genes. Ann Appl Stat 1: 107-129.
9. Boyle EI, Weng S, Gollub J, Jin H, Botstein D, et al. (2004) GO: TermFinder--open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. Bioinformatics 20: 3710-3715.
10. McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, et al. (2010) GREAT improves functional interpretation of cis-regulatory regions. Nat Biotechnol 28: 495-501.
11. Agresti A (1992) A Survey of Exact Inference for Contingency Tables. Stat Sci 7: 131-153.
12. http://www.genecards.org/
13. Pawitan Y, Murthy KR, Michiels S, Ploner A (2005) Bias In the Estimation of False Discovery Rate and Sensitivity of Microarray Studies. Bioinformatics 21: 3865-3872.
14. Kubosaki A, Tomaru Y, Tagami M, Arner E, Miura H, et al. (2009) Genome-wide investigation of in vivo EGR-1 binding sites in monocytic differentiation. Genome Biol 10: R41.
15. Bodén M, Bailey TL (2008) Associating transcription factor-binding site motifs with target GO terms and target genes. Nucleic Acids Res 36: 4108-4117.
16. Koudritsky M, Domany E (2008) Positional distribution of human transcription factor binding sites. Nucleic Acids Res 36: 6795-6805.
17. Smeenk L, van Heeringen SJ, Koeppel M, van Driel MA, Bartels SJ, et al. (2008) Characterization of genome-wide p53-binding sites upon stress response, Nucleic Acids Res 36: 3639-3654.

This article was originally published in a special issue, **Computational and Mathematical Biology** handled by Editor(s). Dr. Kun Huang, The Ohio State University, Columbus; Dr. Ambarish Nag, National Renewable Energy Laboratory, Golden, Colorado, USA