ChIP-on-chip significance analysis reveals large-scale binding and regulation by human transcription factor oncogenes

Adam A. Margolin[1,2,3,*], Teresa Palomero[4,5], Pavel Sumazin[2], Andrea Califano[1,2,4,†], Adolfo Ferrando[4,5,6,†], Gustavo Stolovitzky[2,6,†,**]

[1]Department of Biomedical Informatics, [2]Joint Centers for Systems Biology, [4]Institute for Cancer Genetics, [5]Department of Pathology, [6]Department of Pediatrics, Columbia University, New York, NY 10032

[3] Functional Genomics and Systems Biology Group, IBM T.J. Watson Research Center, Yorktown Heights, N.Y. 10598

[†] These authors contributed equally.

*1130 St. Nicholas Avenue Room 910, New York, NY 10032. E-mail: adam@dbmi.columbia.edu, Telephone: 917-414-7709

**IBM T.J. Watson Research Center, Yorktown Heights, N.Y. 10598. E-mail: gustavo@us.ibm.com, Telephone: 914-945-1292, Fax: 914-945-4217

ABSTRACT

ChIP-on-chip has emerged as a powerful tool to dissect the complex network of regulatory interactions between transcription factors and their targets. However, most ChIP-on-chip analysis methods use conservative approaches aimed to minimize false-positive transcription factor targets. We present a model with improved sensitivity in detecting binding events from ChIP-on-chip data. Biochemically validated analysis in human T-cells reveals that three transcription factor oncogenes, NOTCH1, MYC, and HES1, bind one order of magnitude more promoters than previously thought. Gene expression profiling upon NOTCH1 inhibition shows broad-scale functional regulation across the entire range of predicted target genes, establishing a closer link between occupancy and regulation. Finally, the resolution of a more complete map of transcriptional targets reveals that MYC binds nearly all promoters bound by NOTCH1. Overall, these results suggest an unappreciated complexity of transcriptional regulatory networks and highlight the fundamental importance of genome-scale analysis to represent transcriptional programs.

INTRODUCTION

The dysregulated activity of oncogenic transcription factors contributes to neoplastic transformation by promoting aberrant expression of target genes involved in regulating cell homeostasis. Therefore, characterization of the regulatory networks controlled by these transcription factors is a critical objective in understanding the molecular mechanisms of cell transformation. ChIP-on-chip (ChIP^2 for short)[1] has emerged as a promising technology in the dissection of transcriptional networks by providing high-resolution maps of genome-wide transcription factor-chromatin interactions.

A ChIP^2 experiment is typically performed using a two-channel microarray in which each arrayed probe matches a specific genomic region. The immunoprecipitate (IP) channel measures the relative concentration of genomic fragments bound by an immunoprecipitated protein (usually a transcription factor), while the whole cell extract (WCE) channel measures the relative concentration of fragments in a total chromatin preparation (input control) or in an immunoprecipitation with a non-specific control antibody[2]. Probes with an IP channel intensity significantly higher than the WCE channel intensity are expected to identify genomic regions bound by the protein, thus correct interpretation of these data depends critically on an accurate statistical model for the IP/WCE ratios in genomic regions that are not occupied by the protein. This so called null hypothesis model is used to compute the probability that a given ratio may be produced by background experimental noise. Ratios for which the null hypothesis can be rejected at a desired statistical significance level identify genomic regions likely to be bound by the transcription factor.

Recently, several elegant ChIP^2 analysis methods have been proposed to tackle problems such as integrating measurements from adjacent probes[3-6], or inferring binding site locations at sub-probe resolution[7]. However, the lower-level problem of developing an accurate null hypothesis model has received comparably little attention (see Supplementary Discussion on Previous Methods). Briefly, most current models either normalize or compute probe significance based on whole-dataset statistics[4, 8-19], which include both bound and unbound probes. This approach is based on the assumption that transcription factors have a limited number of direct target genes and therefore the number of bound probes is negligible compared to unbound ones. In this manuscript we demonstrate that transcription factors occupy a higher than anticipated fraction of the genome, rendering whole-dataset statistical approaches, which fail when transcription factor binding sites account for a large percent of all probes (i.e., > 5%)[2], unsuitable for ChIP^2 analysis. Thus the use of whole-dataset statistics can cause severely inaccurate modeling of the data (**Figure 1**). Without an accurate model to define meaningful statistical thresholds, these ChIP^2 data analysis methods use highly conservative approaches aimed to minimize the rate of false-positive predictions at the cost of missing the majority of actual binding sites. Thus while many ChIP^2 studies have produced novel target collections for specific transcription factors, only a small subset of their targets is likely to have been characterized. The problem is compounded by the lack of a clear correlation between strength of the ChIP^2 signal and the ability of the transcription factor to regulate the target, leading to potential misclassification of many targets that are both bound and functionally regulated[20].

Using an empirically determined null hypothesis model of the distribution of intensity ratios for non-IP enriched probes in ChIP^2 experiments, we developed an analytical method, called ChIP^2 Significance Analysis (CSA), that, when compared to other routinely used methods, increases the number of detected binding sites by up to an order of magnitude for all analyzed transcription factors. Biochemical validation of this analytical approach for three transcription factors (NOTCH1, MYC, and HES1) in human T-cell acute lymphoblastic leukemia (T-ALL) cells produced quantitative agreement between the percent of expected versus actual false-positive predictions. Furthermore, analysis of gene expression signatures associated with NOTCH1 inhibition indicates functional regulation by this transcription factor oncoprotein across the entire range of predicted targets. Finally, the ability to infer a more complete repertoire of targets for the studied transcription factors provides a much clearer picture of the interaction between regulatory programs controlled by MYC and NOTCH1. Overall, these results highlight the power of the proposed analysis framework for the identification of transcriptional networks and provide an improved and fundamentally different picture of the transcriptional programs controlled by NOTCH1, HES1 and MYC in T-ALL.

RESULTS

Probe statistics are accurately modeled by CSA

T-ALL is a malignant tumor characterized by the aberrant activation of oncogenic transcription factors[21, 22]. We have recently demonstrated that constitutive activation of NOTCH1 signaling due to mutations in the NOTCH1 gene activates a transcriptional network that controls leukemic cell growth[23-26]. These studies have also demonstrated a fundamental role for HES1 and MYC as transcriptional mediators of NOTCH1 signals, in particular characterizing an ensemble of feed-forward loops in which NOTCH1 directly regulates MYC, and both transcription factors regulate a common set of targets promoting leukemic cell growth, proliferation, and response to NOTCH1 inhibitor therapies[25, 27]. To characterize the structure of the oncogenic transcriptional network driven by activated NOTCH1 in T-cell transformation we sought to identify the direct transcriptional targets of NOTCH1, HES1 and MYC. We hypothesized that the development of an accurate null hypothesis statistical model would result in improved sensitivity in the identification of transcription factor targets and a more accurate description of the individual and combinatorial regulatory programs controlled by these transcription factors.

We first generated a data-derived empirical null hypothesis model of the distribution of IP/WCE intensity ratios for probes associated with unbound fragments (see Methods), and used it to assign a p-value to each probe in the analysis of ChIP^2 assays representing replicate experiments for NOTCH1, MYC, and HES1. ChIP^2 assays for these transcription factors were performed in HPB-ALL cells, a well characterized T-ALL cell line with high expression levels of activated NOTCH1, MYC and HES1. For NOTCH1, ChIP^2 assays were also performed in CUTLL1 cells, another NOTCH1-dependent T-ALL cell line. Analysis of the magnitude versus amplitude plots (**Figure 2a**) of the intensity dependent distribution of probe ratio values shows marked differences in the shape of the plots corresponding to the different experiments. These differences justify the use of data-specific, empirical null models, and suggest that analytical methods that use a predefined functional form for the null distribution may overlook important features of the data. In each case, CSA accurately models the left tail of the probe ratio probability density, where the contribution from bound probes is expected to be minimal (**Figure 2a,b**). Based on the inferred distributions, a large number of probe ratio values cannot be explained as background experimental noise and are therefore classified as bound genomic regions. In contrast, calculation of the null hypothesis model based on whole-dataset statistics shows that this approach has a limited capacity to differentiate bound-probe signals from background noise, especially for the MYC experiment (**Figure 1**). Finally, we note that the p-value distribution for all probes should be uniform between zero and one (unbound probes) with a single peak near zero (bound probes). Importantly, CSA accurately captures these statistical properties (Supplementary Figure 1).

Improved ChIP^2 sensitivity by CSA

CSA then incorporates the probe significance model with an analytical method that integrates the statistics for replicate experiments and probes with nearby genomic locations (to account for the ChIP^2 fragmentation lengths, see Methods). We used CSA to compute the false discovery rate (FDR) associated with the most significant 500 base pair region on each of the 16,697 promoters represented on the array. Analysis of NOTCH1, MYC and HES1 promoter occupancy in T-ALL showed a larger than anticipated number of candidate target genes for these transcription factors. Specifically, using CSA at a conservative FDR of 0.05, the number of promoters on the array bound by the transcription factors in this study are: MYC (8,016, 48.0%), NOTCH1 in CUTLL1 (3,154, 18.9%), HES1 (3,074, 18.4%), and NOTCH1 in HPB-ALL (2,471, 14.8%) (**Table 1**). To contrast these results with those obtained using standard analytical tools we analyzed our ChIP^2 data using the Single Array Error Model (SAEM)[1, 8], a widely used method for analysis of two-color ChIP^2 arrays, and the standard method packaged with the Agilent analysis software. This analysis reveals that CSA predicts approximately an order of magnitude more bound promoters than SAEM, which predicts the following numbers: MYC (127, 0.8%), NOTCH1 in CUTLL1 (647, 3.9%), HES1 (187, 1.1%), and NOTCH1 in HPB-ALL (410, 2.5%) (Table 1).

One would expect that for a fixed IP/WCE fold ratio (e.g., twofold), ChIP^2 experiments with a large number of probes above that fold ratio would yield a large number of predicted bound targets by any method. While this is the case for CSA, the opposite trend is observed for SAEM (**Table 1**), as the null hypothesis model variance for ubiquitously bound transcription factors, such as MYC, is grossly overestimated. Overall, these results demonstrate that CSA has improved sensitivity in the identification of transcription factor binding sites from ChIP^2 experiments and shows robust performance even for experiments performed on broadly bound transcription factors, for which other analyses fail using whole-dataset statistical approaches.

Accuracy of CSA binding predictions are supported by binding site enrichment analysis

As a first test of the broad transcription factor binding predictions generated by CSA, we performed motif analysis to evaluate the enrichment of MYC binding sites, using the TRANSFAC[28] position specific scoring matrix M00322, in the promoters of target genes identified by CSA and SAEM. The DNA-binding component of NOTCH1 transcriptional complexes, CSL, is not represented in TRANSFAC or JASPAR[29] and the only HES1-associated matrix was found to be of low quality and a poor predictor of HES1 binding, independent of the analysis method (see Methods). In this analysis, promoters were first ranked according to their CSA and SAEM p-values computed from the MYC ChIP^2 experiment. Enrichment analysis was performed by identifying MYC/M00322 matching sites in the 600bp fixed-length window centered on the most significant probe in the highest scoring promoter region identified by the algorithm. The match threshold was set so that a negative set, $S^{(-)}$, of 3,000 fragments showing the least amount of MYC binding, would produce a false-positive rate of 30%. Details on the procedure are given in Methods.

Analysis of the cumulative proportion MYC/M00322 matching fragments as a function of their ChIP^2 ranking by the corresponding method shows that fragments inferred by both CSA and SAEM are enriched in MYC/M00322 sites and that, in both cases, site enrichment is correlated with the ChIP^2 ranking (**Figure 3a**). However, for any *n*, the top *n* fragments inferred by CSA are considerably more enriched than the top *n* fragments inferred by SAEM. Note that even when all the probes are considered (largest *n* in **Figure 3a**), fragments detected by CSA are more enriched in MYC/M00322 matching sites than those detected by SAEM, indicating that the regions identified by CSA as the most significant are more likely to contain the motif than those chosen by SAEM.

To compare the predicted false-positive rate by ChIP^2 analysis with the significance of MYC binding site enrichment, we binned the fragments based on CSA and SAEM rankings (100/bin) and assessed whether the MYC/M00322 motif could be successfully used to distinguish the fragments in each bin from those in the negative set, $S^{(-)}$. The classification p-value based on binding site enrichment is in remarkable agreement with the expected false-positive rate by CSA analysis, suggesting significant enrichment of MYC sites in the promoters of ~7,000 genes, corresponding to the range of high confidence targets predicted by CSA (**Figure 3b**). Beyond this threshold, both the CSA-inferred false-positive rate and the p-value of sequence-based classification degrade very rapidly. By comparison, only the first ~1,800 SAEM-inferred fragments are well classified by the MYC/M00322 motif, suggesting that CSA significantly improves the ranking of MYC-bound fragments and that a meaningful statistical cutoff can be determined a-priori.

Experimental validation of CSA transcription factor binding predictions

To further test the accuracy of CSA-based transcription factor target predictions we performed independent Chromatin Immunoprecipitation (ChIP) experiments for each of the four ChIP^2 conditions and tested the IP enrichment of specific promoters by quantitative PCR (qPCR). We first analyzed eight predicted NOTCH1 targets in HPB-ALL cells, randomly sampled at an FDR ≤ 20%. Seven of these eight predicted fragments were validated as bound by NOTCH1 and only the least significant fragment failed validation (**Table 2**).

We further tested an additional twelve targets for HES1 and MYC in HPB-ALL and for NOTCH1 in CUTLL1, sampling predicted targets uniformly at an FDR of 20% (i.e., 20% expected false-positives) (**Table 2**). In this analysis, twenty-six of thirty-six

(72.2%) targets were positive by ChIP/qPCR and nine (25%) were negative. The remaining gene (the second least significant for MYC) could not be amplified by qPCR. Non-validated/false-positive targets are, in general, at the end of the ranked lists (**Table 2**). The only outlier is the first ranked fragment for HES1 (KIAA1407 gene promoter). To obtain experimental evidence on the robustness of our validation assay we randomly selected ten genomic regions not identified as bound by MYC and ten not identified as bound by HES1. Nine selected regions are within promoters and eleven are in intergenic regions. As expected none of these twenty regions showed evidence of binding by MYC or HES1 when tested by ChIP/qPCR.

For all experiments numerous validated genes had CSA ranks in the thousands. The lowest ranking validated genes before encountering a false-positive are as follows: 2,223 for NOTCH1 in CUTLL1, 2,958 for NOTCH1 in HPB-ALL, 4,901 for MYC, and while the top ranking gene for HES1 failed validation, the following seven, down to rank 3,247, were positive. Notably, many of the validated targets show relatively subtle ChIP^2 signals. For example, C6orf82, a validated HES1 target, had ChIP^2 binding ratios in replicate experiments of 1.37 and 1.68 for the most significant probe in its promoter, and there was no enrichment (ratios of .81 and 1.15) for its adjacent probe. However, upon ChIP/qPCR validation, this region showed binding ratios of 2.69 and 4.65. ChIP/qPCR results are available in the Supplementary Materials.

Overall, thirty-three of the forty-four (75%) genes selected from those with a FDR of 20% by CSA were validated by ChIP/qPCR. Notably, only six of the forty-four genes (13.6%) tested in these experiments were identified by SAEM. Overall, these biochemical validation results support our computationally-derived conclusions regarding the broad range of binding for all tested transcription factors and demonstrate the power of CSA for reducing the false negative rate in ChIP^2 experiments.

NOTCH1 regulates direct target genes predicted by CSA

To test whether CSA-predicted NOTCH1-bound genes are also functionally regulated by this transcription factor, we treated a panel of ten T-ALL cell lines with Compound E, a γ-secretase inhibitor that blocks an essential proteolytic cleavage step required for release of the intracellular domains of NOTCH1 from the membrane and their translocation to the nucleus[30]. Genome-wide expression profiles of cells treated for 72 hours with Compound E (100 nM) or vehicle only (DMSO) were measured by microarray profiling, and expression changes were compared with NOTCH1 promoter occupancy identified by CSA analysis of the ChIP^2 data. Overall, 11,606 genes are represented on both the ChIP^2 and the expression arrays. For each gene we computed: 1) the ChIP^2 FDR based on the highest scoring 500bp region in its promoter, 2) the number of experiments in which it is expressed (not called absent by MAS5), and 3) the log2 expression ratio of the control versus treatment, averaged over the ten cell lines and duplicate experiments.

Predicted NOTCH1-bound genes are both more likely to be expressed than genes not identified as bound by NOTCH1 and show a clear down-regulation upon NOTCH1 inhibition (**Figure 4**). The 2,000 most confident NOTCH1 targets (FDR<0.058) are expressed in 83.3% of experiments while the 6,000 least confident NOTCH1 targets are expressed in 38.7% of experiments (P=1.5x10^{-279}). The top 2,000 targets also show coordinated down-regulation upon NOTCH1 inhibition that is subtle in magnitude (mean=12.3%) but extremely significant (P=7.7x10^{-124}). The ChIP^2 analysis predicts a rapid increase in false-positives beyond the top 2,000 targets, and, correspondingly, their likelihoods to be expressed and regulated by NOTCH1 decrease. However, even for genes with ChIP^2 ranks between 4,000 and 5,000, there is significant enrichment for both the percent of expressed genes (59.4%, P=2.3x10^{-34}) and the expression change upon NOTCH1 inhibition (mean=3.9%, P=2.7x10^{-16}). These results demonstrate that, in contrast with previous analysis based on a limited number of targets[27], NOTCH1 directly contributes to the transcriptional activity of thousands of genes.

Interaction of NOTCH1 and MYC regulatory networks

NOTCH1 and MYC operate as highly interrelated regulators of cell growth, proliferation, and survival during T-cell development and transformation. In a recent study[24], we compared the regulatory networks controlled by NOTCH1 and MYC by using the ARACNE reverse engineering algorithm[31, 32] to predict fifty-eight and sixty-one targets of NOTCH1 and MYC, respectively, and observed a significant overlap of twelve genes between the two lists (P=2.4x10^{-52}). We went on to characterize a set of feed-forward loops in which NOTCH1 directly regulates MYC, and both transcription factors regulate a common set of targets promoting leukemic cell growth. Based on these findings we sought to further investigate the relationship between the genes bound by MYC and by NOTCH1, using the much larger list of inferred targets by CSA. Strikingly, the analysis predicts that MYC binds to 1,668 of the 1,804 (92.5%, P=3.6x10^{-12}) genes that are bound by NOTCH1, using a ChIP^2 FDR threshold of .01. In agreement with the fundamental role of NOTCH1 in controlling leukemia cell growth[24], the NOTCH1-bound genes are highly enriched in Gene Ontology (GO)[33] categories related to cellular growth and metabolism, such as cellular metabolism (*P*=8.2x10^{-42}), RNA metabolism (*P*=1.5x10^{-25}) and protein biosynthesis (*P*=2.4x10^{-10}). The complete output of the GO enrichment analysis, using the DAVID tool[34], is given in the Supplementary Materials.

DISCUSSION

We have shown that the choice of a realistic null hypothesis model can dramatically affect the result of ChIP^2 data analysis and its biological interpretation and proposed the CSA algorithm to assign meaningful statistical significance scores used to predict a more complete range of transcription factor-target interactions. The method of assessing probe statistical significance relies on minimal assumptions: that the null distribution is symmetric and that bound fragments do not significantly affect the left tail of the null hypothesis statistics. As a result, it should generalize well to ChIP^2 experiments performed using other platforms and cellular conditions. By using an independence model for replicate experiments and adjacent probes, we then incorporate the approach into a global analytical framework for the interpretation of ChIP^2 data. While the statistical independence assumption is valid for relatively sparse arrays, more dense arrays may introduce correlation for unbound nearby probes that are within the DNA fragmentation length. Therefore, the CSA method may be further improved by incorporating existing, more sophisticated models for the integration of nearby probes (e.g.[3-7]). However, for the arrays used in this study, we show that our results are in quantitatively good agreement with biochemical validation assays and that no correction seems to be required.

The analysis of three oncogenic transcription factors by two-color ChIP^2 arrays reveals that CSA, indeed, identifies an order of magnitude more bound gene promoters than standard analyses. Specifically, CSA predicts that each studied transcription factor binds to several thousand target genes, with MYC binding to roughly half of the assayed promoters, providing additional insight into the extreme pluripotency of this proto-oncogene[35]. These predictions might still be an underestimate because only the proximal promoter regions (-0.8KB to +0.2KB, relative to transcription start site) are represented on the arrays used in this study. CSA predictions were validated by three independent tests, including: ChIP/qPCR (experimental), binding site enrichment analysis for MYC (computational), and gene expression analysis after NOTCH1 inhibition (biological). ChIP/qPCR experiments are in excellent correspondence with CSA-inferred FDRs, especially considering that ChIP/qPCR itself has a 10%-20% false negative rate[9, 24, 36, 37]. Computational validation by sequence analysis further indicates that CSA-inferred FDRs are in striking agreement with MYC binding site enrichments. Finally, gene expression analysis after NOTCH1 inhibition both provides further support for the CSA predictions and creates a stronger than expected association between bound and regulated genes. We find that the NOTCH1 protein binds to a large number of promoters (>2,000) and that the set of corresponding genes is consistently, albeit weakly regulated upon NOTCH1 inhibition. These results are highly consistent with a previous study performed in yeast[20] that also observed correspondence of ChIP^2 results with both binding site enrichment and expression changes for a large number of genes.

GO enrichment analysis shows that NOTCH1 subtly regulates a large number of genes involved in the cellular growth machinery. These results add an additional layer of regulation to the effects of NOTCH1 signaling in promoting cell growth, with important implications for understanding the role of NOTCH1 signaling in development and transformation. Thus in addition to the established role of NOTCH1 in promoting growth through its interaction with MYC[27] and the PI3K-AKT[25] signaling pathway, NOTCH1 also has a direct effect in promoting cell growth. This irreversibly couples the developmental programs involved in stem cell homeostasis and lineage commitment activated upon NOTCH1 activation with the metabolic pathways needed for the expansion of stem cells and T-cell progenitors.

Finally, the availability of a more complete repertoire of bound promoters allows us, for the first time, to truly assess the extent of a transcription factor's regulatory program and the combinatorial overlap between independent programs. Our analysis shows that 92.5% of the promoters bound by NOTCH1 are also bound by MYC. Indeed, it appears that NOTCH1 co-regulates a specific subset of the MYC regulatory program. While this was previously hinted by statistical correlation of the regulatory programs inferred for the two transcription factors by expression analysis[27], the true extent of this overlap can only be grasped after resolving a more complete map of NOTCH1 and MYC targets. While contributing to our understanding of transcriptional regulation at the genome-scale, these findings suggest an even greater than expected complexity of these processes.

MATERIALS AND METHODS

CSA Algorithm

Single probe significance analysis: For each probe, the statistical significance of a two-channel measurement is inferred by computing the conditional probability of the magnitude (*M*) given the amplitude (*A*), $P_{null}(M \mid A)$, where $M = \log 2\left(\frac{IP}{WCE}\right)$ and $A = \frac{\log 2(IP) + \log 2(WCE)}{2}$, under the null hypothesis (i.e., no enrichment in the IP compared to the WCE channel). Here, *IP* and *WCE* represent respectively the probe intensity measurements for the IP and WCE channels. That is, we seek to compute the probability of observing a given ratio between the two channel intensities (IP versus WCE) given the total intensity measured in the two channels, assuming that no binding is present. This is because if the total intensity, A, is different, the same ratio, M, may have very different statistical significance. Typically, the higher the value of A, the higher the statistical significance for the same value of M. This is illustrated in the scatter plots of *M* versus *A* (**Figure 2a**), which correspond to a rotation and rescaling (to facilitate computation) of the scatter plots of *IP* versus *WCE*. The general idea is to first identify the intensity dependent mean of the null distribution (i.e., the distribution for unbound probes), and to use only ratio values below this mean to estimate the full conditional probability of *M* given *A*. This method makes two minimal assumptions. First, that bound probes have minimal effect on the distribution for ratio values below the identified mean, and, second, that the null distribution is symmetric around its mean (note that we do not assume a parametric shape of the null distribution).

The method begins by estimating the joint probability distribution, $P(M, A)$, using a Gaussian kernel density estimator[38, 39]. For computational efficiency, we truncate the kernels at four standard deviations in any direction and renormalize appropriately. Thus using kernels with diagonal covariance matrices and marginal variances h_A and h_M, the estimator is defined as:

$$P\left(M,A\right)=\frac{1}{2\pi h_A h_M NZ}\sum_{i=1}^{N}\exp\left\{-\frac{1}{2}\frac{\left(A-A_i\right)^2}{h_A^2}+\frac{\left(M-M_i\right)^2}{h_M^2}\right\},$$

where $Z=\left(1-2*Erf(4)\right)^2$ is the normalization constant that accounts for truncation of the Gaussians. The kernel width of the estimator is calculated using the AMISE criterion[40]. Conditioning on A yields the conditional distribution:

$$P(M\mid A)=\frac{P(M,A)}{P(A)}\text{, where } P\left(A\right)=\frac{1}{\sqrt{2\pi}h_A N}\sum_{i=1}^{N}\exp\left\{-\frac{\left(A-A_i\right)^2}{2h_A{}^2}\right\}$$

For a particular average intensity value, A_0, the conditional mean of the null distribution is inferred as $\hat{\mu}_{M|A_0}=\underset{M}{\operatorname{argmax}}\,P(M\mid A=A_0)$.

The conditional null distribution given $A=A_0$ is inferred by projecting $P(M\mid A=A_0)$ across $\hat{\mu}_{M|A_0}$ for $M<\hat{\mu}_{M|A_0}$. This procedure is used to calculate $P_{null}(M\mid A)$ for an evenly spaced grid of *A* and *M* values, excluding the 1% of probes with the lowest *A* values (which are assigned a p-value of one), and the complete conditional distribution is computed using two dimensional linear interpolation. For each probe, statistical significance is assessed using a one tailed test with reference to this distribution. Because the distribution is empirical, there is a limit to the inferable minimum p-value, which depends on the number of arrayed probes. For this array, we set the minimum p-value to 10^{-5}, which is roughly one divided by the number of probes on the array. We stress the importance of using an empirical distribution as we have observed that the empirical data generally displays significantly non-Gaussian tails.

That is, ChIP^2 data can often display fairly large deviations from the mean ratio by chance, whereas under a Gaussian probability model, which is characterized by very light tails, such large deviations would be assigned near zero probability. A major goal of this work is to accurately determine the probabilities of obtaining probe ratio values, and we believe that using a parametric distribution of probe statistics will produce significant errors in p-value estimation.

Combining replicates: Let p_i^j denote the p-value computed for the ith probe in jth replicate experiment. Assuming that replicates are independent in the null hypothesis, a test statistic for evaluating the probability of a joint observation of p-values across experiments is the product of the individual p-values, $\overline{s}_i = \prod_{j=1}^{M} p_i^j$, where $\overline{s}_i$ is the test statistic and *M* is the number of replicate experiments. If modeled correctly, p-values under the null hypothesis should be uniformly distributed (shown in Supplementary Figure 1). It is useful to log transform this equation such that we now evaluate $-\log\left(\overline{s}_i\right) = -\sum_{j=1}^{M} \log\left(p_i^j\right)$. Because the logarithm of a uniform distribution is exponentially distributed with mean one, this equation is a sum of exponentially distributed random variables, which is a Gamma distributed random variable itself, with mean one and M degrees of freedom. Thus significance can be evaluated as

$$\Gamma_{CDF}^{M}\left(-\sum_{j=1}^{M} \log\left(p_i^j\right)\right) \qquad (0.1)$$

Where Γ_{CDF}^{M} is the Gamma cumulative distribution function with mean one and M degrees of freedom.

Combining regions: Due to sonication, the DNA fragments hybridized to the microarray represent a distribution of varying lengths, while the distribution for immunoprecipitated fragments deriving from a particular binding site will be centered on that site. Therefore, the signal derived from a binding event may be detected by multiple probes in close genomic proximity to the binding site, and it is useful to combine the values from nearby probes to compute a combined statistic representing the probability of a binding event within the region spanned by those probes. Thus we adapt a commonly used strategy[11] of using a fixed sized sliding window and integrating the values of probes falling within this window. Based on published measurements of fragmentation lengths[7], we use a 500bp window and a step size of 150bp. Assuming that measurements from adjacent probes are independent in the null hypothesis, the same principle as in Eqn. (0.1) can be applied to integrate the values from nearby probes. That is, let *W* represent the set of probes falling within a given 500bp window. The integrated probability for this region is then calculated as:

$$p_{region} = \Gamma_{CDF}^{M*|W|}\left(-\sum\nolimits_{i\in W}\sum\nolimits_{j=1}^{M}\log\left(p_i^j\right)\right) \tag{0.2}$$

To compute the probability that any region within a gene's promoter is bound, we consider the most significant window, controlling for multiple tests using Bonferroni correction based on the number of probes on the promoter. This correction is not exact as the number of tests (i.e., the number of windows containing unique subsets of probes) is likely greater than the number of probes on a promoter, causing an underestimation of the significance, while the tests are not independent (i.e., the same probe may fall within multiple windows), causing an overestimation of the significance. However, since the number of probes for a promoter (and therefore the number of probes within each window) is relatively small, especially for the arrays used in this study, we expect this simplification to have little impact on the calculated statistics. For very dense arrays a more sophisticated multiple test correction procedure, such as those described in[41], may yield more accurate results.

FDR calculation: After computing a corrected p-value for each gene, representing the probability that the most significant region on the gene's promoter is bound, we control for multiple tests across genes and compute a false discovery rate using the Benjamini Hochberg procedure[42]. Let p_k represent the corrected p-value computed for gene *k*, let r_k represent the rank of gene *k* sorted by the ChIP^2 p-values, and let *G* represent the total number of genes on the array. Then the false discovery rate for gene *k* is computed as:

$$FDR_k = \frac{G * p_k}{r_k} \qquad (0.3)$$

ChIP^2: ChIP^2 analysis was performed in the HPB-ALL and CUTLL1 T-ALL cell lines. Briefly, 1x10^8 cells were subjected to chromatin immunoprecipitation using the following antibodies: anti NOTCH1 Val1744 polyclonal antibody (Cell Signaling; Danvers, MA) which specifically recognizes the intracellular activated form of NOTCH1, anti c-MYC (N-262) or anti HES1 (H-140) rabbit polyclonal antibodies (Santa Cruz Biotechnology; Santa Cruz, CA). ChIP^2 was performed following standard protocols provided by Agilent Technologies using Agilent Human Proximal Promoter Microarrays (44K features/array) and have been described previously (Vilimas et al., 2007). This platform contains ~4-5 probes/gene covering -0.8KB to +0.2KB (relative to transcript start site) of human transcripts from UCSC hg17/NCBI release 35 (May 2004). The arrays were scanned using an Agilent scanner and data was extracted using the Feature Extraction 8 software and analyzed using ChIP Analytics 1.1.

ChIP/qPCR: Validation of promoter occupancy was performed by quantitative PCR analysis of chromatin immunoprecipitates and their corresponding whole cell extracts (input control) in an ABI 7300 Real-Time PCR System (Applied Biosystems; Foster City, CA) using the SYBR® Green PCR Core Reagents (Applied Biosystems). Briefly, enrichment of candidate target promoters in chromatin immunoprecipitates was assessed by quantifying each promoter of interest in comparison with input chromatin control using an unrelated β-actin genomic sequence as reference. A 450bp region flanking the most significant probe in the region identified by CSA was selected to design the primers for the validation. Validation was considered positive if the mean enrichment, across three technical replicates, of the specific region in the immunoprecipitate versus the input control after correction by the β-actin levels was above two-fold. Regions with enrichments below three-fold were considered non-validated if they displayed high variability across the technical replicates. If biological replicates were performed we require at least one of the two replicates to meet this criteria[24], corresponding to the ChIP^2 analysis method, which assumes a null hypothesis model of no binding in any replicates. Data for all ChIP/qPCR validation experiments are given in the Supplementary Materials.

Microarray expression profiling: We treated duplicate cultures of 10 T-ALL cell lines (ALL-SIL, CCRF-CEM, CUTLL1, DND41, HPB-ALL, KOPTK1, MOLT3, P12 ICHIKAWA, PF 382, and RPMI8402) with Compound E (100 nM) ,a well characterized and highly active GSI, or vehicle only (DMSO) for 72 hours for microarray analysis with Affymetrix Human U133 Plus 2.0 Arrays. Samples for microarray analysis were prepared according to the manufacturer's instructions and as described[24].

Binding site enrichment analysis: We used vertebrate position weight matrices from TRANSFAC version 9.3 and JASPAR to model transcription factor sequence specificity. In total, we considered 675 position weight matrices. These include over thirty highly similar helix-loop-helix and Ebox-domain matrices that are predictive of MYC biding, one matrix that is predictive of HES1 binding, and no matrices that are predictive of NOTCH1 binding. We chose TRANFAC matrix M00322, which is derived from eighteen binding sites, for MYC binding site analysis. This choice is arbitrary and the reported results were replicated using other matrices as well. The only model available for HES1 (TRANSFAC matrix M01009) is derived from an insufficient number of validated sites (eight) and was a poor predictor of HES1 binding independent of the algorithm used to interpret HES1 ChIP[2] data. To identify high scoring sites for MYC using M00322, we scanned each potential site in a given sequence and assigned scores according to the Dirichlet-corrected log2 likelihood ratio of the probability that the site was generated from M00322 relative to the probability that it was generated from the base composition vector[39].

For each probed promoter, we considered the most significant region identified by CSA and SAEM. For CSA, we used the procedure described above, and, as a comparison, we used the probe p-values reported by SAEM, and used the CSA procedure for assigning FDRs to regions. We then ranked promoters by the FDR of the most significant region. The fragment boundaries were set to $\left[-300,300\right]$ from the center of the probe in the region with the lowest p-value. We constructed a negative set of fragments by selecting the lowest ranking 3,000 fragments according to each method. The FDR associated with the highest ranking negative fragment was greater than 0.8. We evaluated enrichment of M00322 sites in a given fragment based on the highest scoring site in this fragment or its anti-sense. When setting a fixed false-positive rate, we set an M00322 scoring threshold to permit no more than this percent of negative set fragments to have a site with a score greater than the threshold.

We used permutation testing to evaluate the enrichment of M00322 sites in the positive set against the negative set (lowest ranking 3,000 fragments). We compared the balanced error rate for M00322 to the balanced error rate for the top motif predictor (out of all TRANSFAC and JASPAR matrices) in each test, and iterated 10,000 times to evaluate enrichment p-values to within 0.0001 accuracy. The balanced error rate for a matrix was identified by enumerating all possible scoring thresholds for the matrix and selecting the threshold with the lowest average of the false positive and false negative errors. In each iteration we permuted the indicator vector that assigns each fragment to the positive or the negative set, computed the balanced error rate for each TRANSFAC and JASPAR matrix in the permuted positive and negative sets, and recorded the lowest balanced error rate across all matrices. Enrichment p-values were set by registering the frequency that the balanced error rate for M00322 was greater than balanced error rates recorded in permutation testing.

ACKNOWLEDGEMENTS

A.A.M. is supported by an IBM Ph.D fellowship. This work was supported by the National Cancer Institute (R01CA109755 to A.C. and RO1CA120196 to A.F.), the National Institute of Allergy and Infectious Diseases (R01AI066116), the Alex Lemonade Stand Foundation (T.P.), The Cancer Research Institute, the WOLF Foundation, the National Centers for Biomedical Computing NIH Roadmap Initiative (U54CA121852) and the Leukemia and Lymphoma Society (grants 1287-08 and 6237-08). A.F. is a Leukemia and Lymphoma Scholar.

REFERENCE LIST

1. Ren, B. et al. Genome-wide location and function of DNA binding proteins. *Science* **290**, 2306-2309 (2000).
2. Buck, M.J. & Lieb, J.D. ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. *Genomics* **83**, 349-360 (2004).
3. Glynn, E.F. et al. Genome-wide mapping of the cohesin complex in the yeast Saccharomyces cerevisiae. *PLoS Biol* **2**, E259 (2004).
4. Kim, T.H. et al. A high-resolution map of active promoters in the human genome. *Nature* **436**, 876-880 (2005).
5. Li, W., Meyer, C.A. & Liu, X.S. A hidden Markov model for analyzing ChIP-chip experiments on genome tiling arrays and its application to p53 binding sequences. *Bioinformatics (Oxford, England)* **21 Suppl 1**, i274-282 (2005).
6. Zheng, M., Barrera, L.O., Ren, B. & Wu, Y.N. ChIP-chip: data, model, and analysis. *Biometrics* **63**, 787-796 (2007).
7. Qi, Y. et al. High-resolution computational models of genome binding events. *Nat Biotechnol* **24**, 963-970 (2006).
8. Hughes, T.R. et al. Functional discovery via a compendium of expression profiles. *Cell* **102**, 109-126 (2000).
9. Lee, T.I. et al. Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science* **298**, 799-804 (2002).
10. Bieda, M., Xu, X., Singer, M.A., Green, R. & Farnham, P.J. Unbiased location analysis of E2F1-binding sites suggests a widespread role for E2F1 in the human genome. *Genome Res* **16**, 595-605 (2006).
11. Buck, M.J., Nobel, A.B. & Lieb, J.D. ChIPOTle: a user-friendly tool for the analysis of ChIP-chip data. *Genome Biol* **6**, R97 (2005).

12. Bernstein, B.E. et al. Genomic maps and comparative analysis of histone modifications in human and mouse. *Cell* **120**, 169-181 (2005).
13. Cawley, S. et al. Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell* **116**, 499-509 (2004).
14. Pokholok, D.K. et al. Genome-wide map of nucleosome acetylation and methylation in yeast. *Cell* **122**, 517-527 (2005).
15. Cleveland, W.S. Robust locally weighted regression and smoothing scatterplots. *J Amer Statist Assoc* **74**, 829-836 (1979).
16. Johnson, W.E. et al. Model-based analysis of tiling-arrays for ChIP-chip. *Proceedings of the National Academy of Sciences of the United States of America* **103**, 12457-12462 (2006).
17. Gibbons, F.D., Proft, M., Struhl, K. & Roth, F.P. Chipper: discovering transcription-factor targets from chromatin immunoprecipitation microarrays using variance stabilization. *Genome Biol* **6**, R96 (2005).
18. Huber, W., von Heydebreck, A., Sultmann, H., Poustka, A. & Vingron, M. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics (Oxford, England)* **18 Suppl 1**, S96-104 (2002).
19. Durbin, B.P., Hardin, J.S., Hawkins, D.M. & Rocke, D.M. A variance-stabilizing transformation for gene-expression microarray data. *Bioinformatics (Oxford, England)* **18 Suppl 1**, S105-110 (2002).
20. Tanay, A. Extensive low-affinity transcriptional interactions in the yeast genome. *Genome Res* **16**, 962-972 (2006).
21. Ferrando, A.A. & Look, A.T. Clinical implications of recurring chromosomal and associated molecular abnormalities in acute lymphoblastic leukemia. *Seminars in hematology* **37**, 381-395 (2000).
22. Ferrando, A.A. et al. Gene expression signatures define novel oncogenic pathways in T cell acute lymphoblastic leukemia. *Cancer cell* **1**, 75-87 (2002).
23. Palomero, T. et al. Activating mutations in NOTCH1 in acute myeloid leukemia and lineage switch leukemias. *Leukemia* **20**, 1963-1966 (2006).
24. Palomero, T. et al. Transcriptional regulatory networks downstream of TAL1/SCL in T-cell acute lymphoblastic leukemia. *Blood* **108**, 986-992 (2006).
25. Palomero, T. et al. Mutational loss of PTEN induces resistance to NOTCH1 inhibition in T-cell leukemia. *Nature medicine* **13**, 1203-1210 (2007).
26. Weng, A.P. et al. Activating mutations of NOTCH1 in human T cell acute lymphoblastic leukemia. *Science* **306**, 269-271 (2004).
27. Palomero, T. et al. NOTCH1 directly regulates c-MYC and activates a feed-forward-loop transcriptional network promoting leukemic cell growth. *Proceedings of the National Academy of Sciences of the United States of America* **103**, 18261-18266 (2006).
28. Matys, V. et al. TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic acids research* **31**, 374-378 (2003).
29. Sandelin, A., Alkema, W., Engstrom, P., Wasserman, W.W. & Lenhard, B. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic acids research* **32**, D91-94 (2004).
30. Miele, L. Notch signaling. *Clin Cancer Res* **12**, 1074-1079 (2006).

31. Basso, K. et al. Reverse engineering of regulatory networks in human B cells. *Nature genetics* **37**, 382-390 (2005).

32. Margolin, A.A. et al. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC bioinformatics* **7 Suppl 1**, S7 (2006).

33. Ashburner, M. et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics* **25**, 25-29 (2000).

34. Dennis, G., Jr. et al. DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol* **4**, P3 (2003).

35. Pelengaris, S. & Khan, M. The many faces of c-MYC. *Archives of biochemistry and biophysics* **416**, 129-136 (2003).

36. Polo, J.M. et al. Transcriptional signature with differential expression of BCL6 target genes accurately identifies BCL6-dependent diffuse large B cell lymphomas. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 3207-3212 (2007).

37. Odom, D.T. et al. Control of pancreas and liver gene expression by HNF transcription factors. *Science* **303**, 1378-1381 (2004).

38. Beirlant, J., Dudewicz, E., Gyorfi, L. & van der Meulen, E. Nonparametric entropy estimation: An overview. *Int. J. Math. Stat. Sci.* **6**, 17-39 (1997).

39. Hertz, G.Z. & Stormo, G.D. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics (Oxford, England)* **15**, 563-577 (1999).

40. Sheather, S.J. & Jones, M.C. A Reliable Data-Based Bandwidth Selection Method for Kernel Density-Estimation. *Journal of the Royal Statistical Society Series B-Methodological* **53**, 683-690 (1991).

41. Keles, S., van der Laan, M.J., Dudoit, S. & Cawley, S.E. Multiple testing methods for ChIP-Chip high density oligonucleotide array data. *J Comput Biol* **13**, 579-613 (2006).

42. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B-Methodological* **57**, 289-300 (1995).

TABLES

Table 1 – Number of predicted target genes for various methods

For each of the four $ChIP^2$ experiments, we compare the number of targets inferred by CSA (at 5% FDR) and by SAEM. The "SAEM Predicted" column refers to genes that are output as inferred targets using the standard Agilent software. This method uses a heuristic to define sets of p-value thresholds that a probe and its neighbors must pass to be considered a target. The following column uses the CSA procedure with a FDR cutoff of 5%. The last column lists the number of probes that are at least twofold enriched in the IP channel for both replicates, and provides a simple heuristic of the overall enrichment in the experiment. CSA infers dramatically more targets than SAEM for all experiments, and the number of targets predicted by CSA is correlated with the enrichment heuristic, while the predictions by SAEM are anti-correlated.

Transcription Factor	**Cell Line**	**SAEM Predicted**	**CSA (FDR 5%)**	**Probes with M>1 in both replicates**
MYC	*HPB-ALL*	127	8,016	8,534
HES1	*HPB-ALL*	187	3,074	1,470
NOTCH1	*CUTLL1*	647	3,154	907
NOTCH1	*HPB-ALL*	410	2,471	841

Table 2 – Validation of predicted targets at 20% FDR

We used ChIP/qPCR to test twelve genes each for NOTCH1 in CUTLL1 and for HES1 and MYC in HPB-ALL (one gene for MYC could not be amplified), and eight genes for NOTCH1 in HPB-ALL. The data columns are as follows: **Gene** – The gene name. **CSA FDR** – The FDR computed by CSA. That is, the percentage of genes with ranks lower than the current gene that are expected not to be bound by the transcription factor. **Validated** – Whether the gene was positively validated by ChIP/qPCR. **Rank** – The rank of the gene when all genes are sorted by their CSA-inferred statistical significance. **SAEM** – Whether the gene is identified as a target by SAEM.

NOTCH1 in HPB-ALL

Gene	CSA FDR	Validated	Rank	SAEM
FLJ13798	2.29E-12	yes	35	yes
RAB18	2.03E-05	yes	674	no
Porimin	0.0015	yes	1301	no
ZMAT2	0.0035	yes	1497	no
PSENEN	0.0068	yes	1675	yes
LMAN2	0.0379	yes	2316	no
XKR9	0.1054	yes	2958	no
THPO	0.125	no	3119	no

MYC

Gene	CSA FDR	Validated	Rank	SAEM
PRKACB	1.96E-13	yes	337	no
LOH12CR1	8.44E-11	yes	1131	no
HS3ST3B1	3.62E-09	yes	1885	no
POLR2I	6.11E-08	yes	2639	no
TXLNB	6.07E-07	yes	3393	no
PARP1	4.80E-06	yes	4147	no
KIAA1984	2.93E-05	yes	4901	no
ZNF233	1.91E-04	no	5655	no
KIF5B	0.0014	yes	6409	no
HIST1H2AK	0.0083	yes	7163	no
CPE	0.0417	N/A	7917	no

PEX16	0.126	no	8671	no

HES1

Gene	CSA FDR	Validated	Rank	SAEM
KIAA1407	2.96E-07	no	216	no
MGC3121	0.0001	yes	649	no
PRKDC	0.0015	yes	1082	no
GTF3C2	0.0057	yes	1515	no
FAM20B	0.0126	yes	1948	no
CHRM5	0.0239	yes	2381	no
BTBD9	0.0393	yes	2814	no
C6orf82	0.0577	yes	3247	no
WDSUB1	0.0815	no	3680	no
DACH2	0.1074	no	4113	no
NARG1L	0.1405	yes	4546	no
CHORDC1	0.1775	no	4979	no

NOTCH1 in CUTLL1

Gene	CSA FDR	Validated	Rank	SAEM
RNF139	5.41E-10	yes	171	yes
ELP3	2.76E-07	yes	513	yes
BAT2	5.22E-06	yes	855	yes
DDX5	7.45E-05	yes	1197	no
ZNF436	4.05E-04	yes	1539	no
ETFDH	1.44E-03	yes	1881	yes
DCP1A	4.41E-03	yes	2223	no
MRPL48	0.0109	no	2565	no
MYB	0.0303	no	2907	no
KCTD16	0.0599	yes	3249	no
ADAR	0.1022	yes	3591	no
BXDC5	0.1549	no	3933	no

FIGURE LEGENDS

Figure 1 – Whole-dataset statistics analysis

Blue bars represent a histogram of log2 IP/WCE probe ratio values from a MYC ChIP^2 experiment. The histogram displays distinct, overlapping distributions for bound and unbound probes. The dotted red curve shows the log2 ratio values after mean centering, a common normalization technique that, for this experiment, adjusts the mean of the null distribution to be negative in order to compensate for the large number of high ratio values for the bound probes. The green curve represents a Gaussian fitted to the overall distribution, demonstrating that analysis methods that fit a global error model to this data will significantly overestimate the variance of the null distribution and will incur a high false negative rate, as shown by the black arrow, which represents two standard deviations from the mean of the green curve and eliminates a large percentage of bound probes.

Figure 2 – CSA determination of ChIP^2 target genes

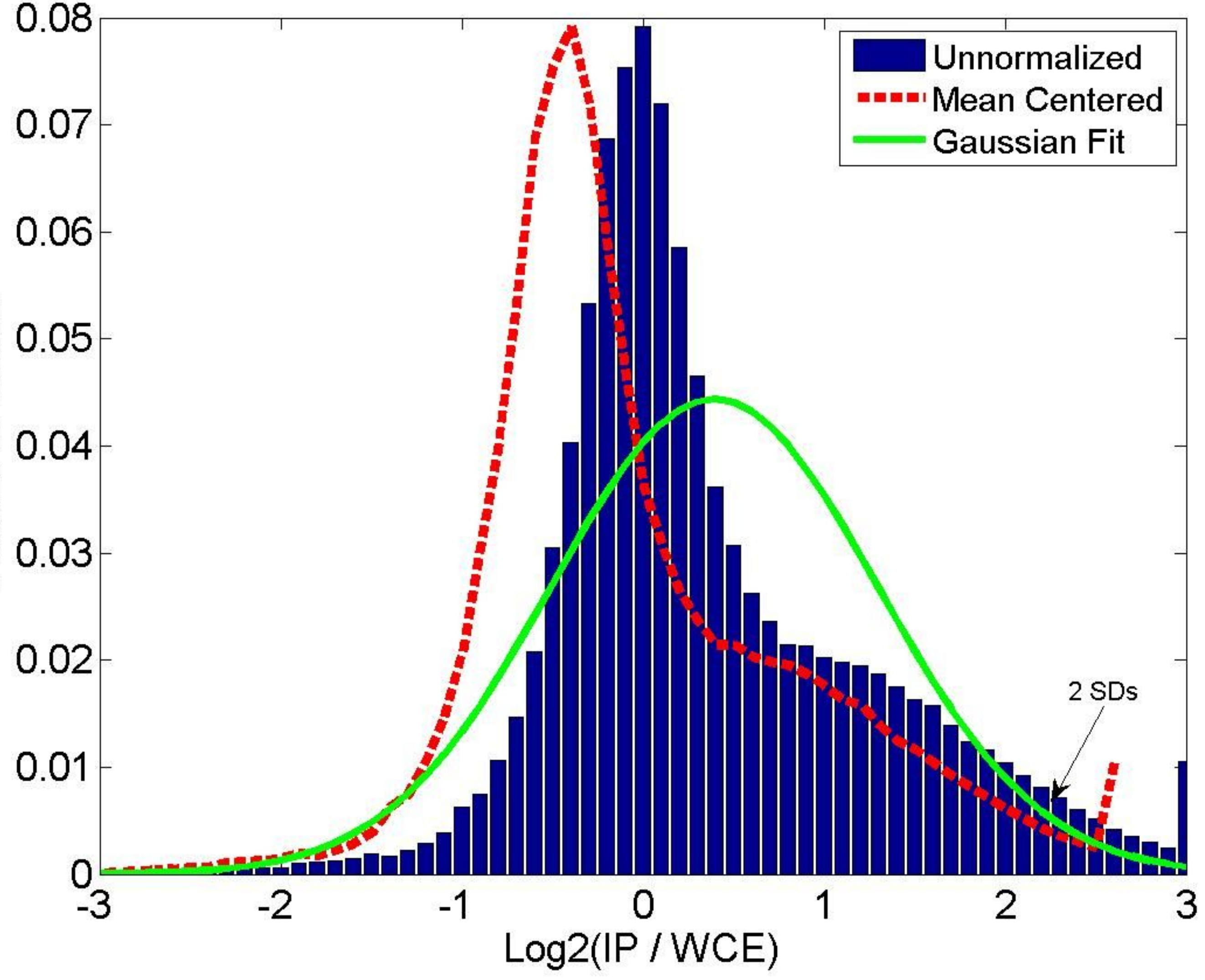
(a) Magnitude (*M*) versus amplitude (*A*) plots with confidence intervals inferred by CSA. The x-axis represents the amplitude, calculated as the average log2 intensity of the IP and WCE channels. The y-axis represents the magnitude, calculated as the log2 ratio of IP/WCE. The black line represents the intensity dependent mean of the inferred null distribution, and the colored lines represent confidence intervals of .1, .01, and .001 probability. Note that confidence intervals are computed based on a one tail test, thus the lower lines actually represent one minus the corresponding value. As shown, for all three transcription factors a large number of probes are significantly enriched in the IP channel (data points above the upper confidence interval lines), and MYC displays substantially more enrichment. **(b)** Graphic representation of the inferred distribution of $P(M \mid A=11)$. The blue curve represents the empirical conditional distribution of *M* computed by the kernel density estimation procedure at the particular value of *A*=11. The dotted black line represents the inferred mean of the null distribution and the dotted red line represents the inferred null distribution, computed by projecting the left side of the empirical conditional distribution across the inferred mean. For NOTCH1 and HES1 the bound probes manifest as a heavy tail on the right side of the empirical distribution compared to the null, whereas for MYC there is a clear bimodality, and separate distributions for bound and unbound probes can be observed by eye. **(c)** Heat map representation of the intensity dependent null distribution as inferred by CSA. The x-axis represents the average log2 intensity of the IP and WCE channels, and the y-axis represents the log2 ratio of IP/WCE. Colors represent the –log10 p-value of the null distribution (for values below the mean colors represent one minus the p-value for the one tailed significance test). As expected, the model reveals an intensity dependent mean and variance of the null distribution, with increased variance at low intensity levels, as well as sometimes for extremely high intensity levels due to saturation effects.

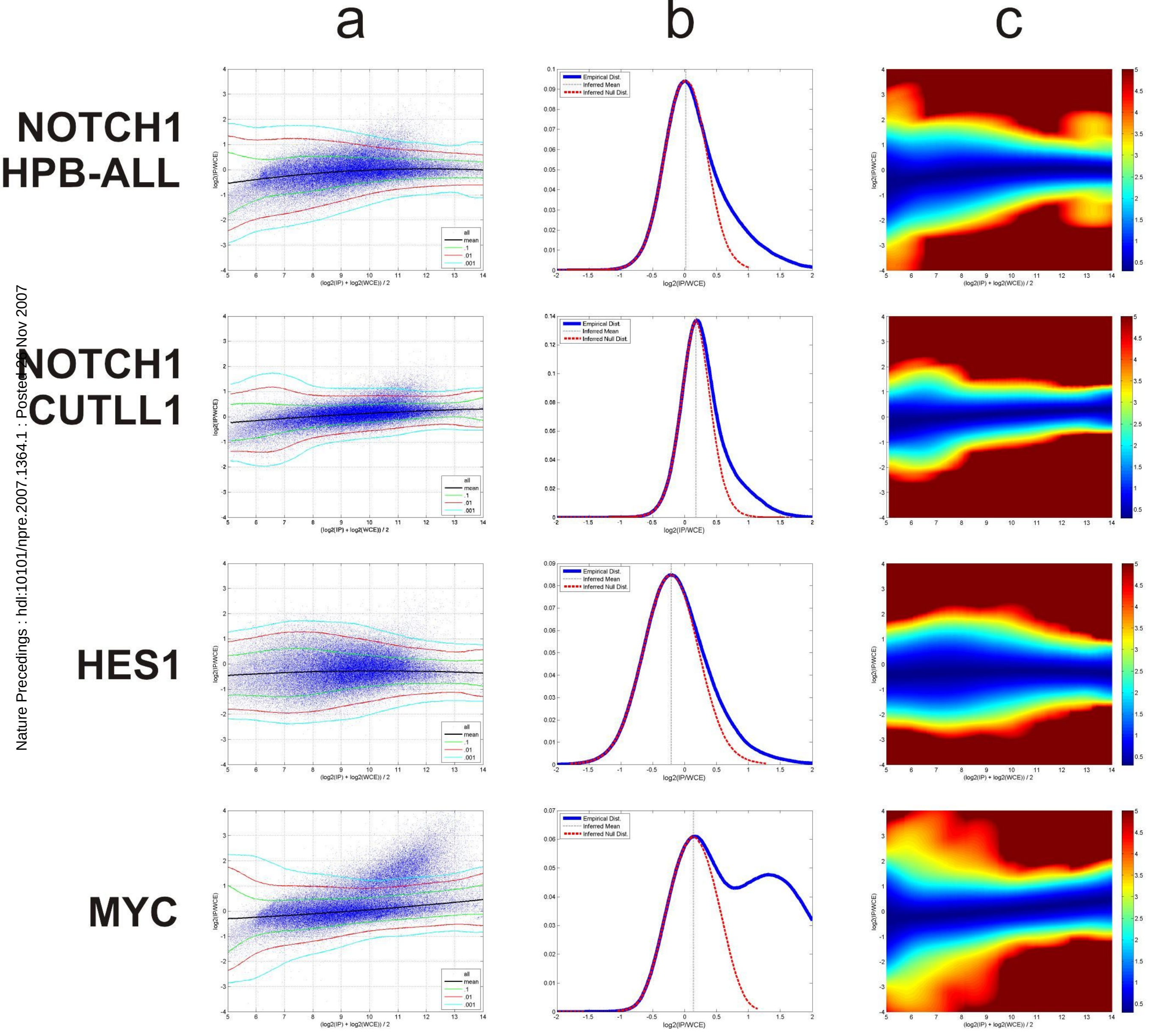
Figure 3 – Transcription factor binding motif enrichment analysis

(a) The percentage of identified sequences containing a binding site for MYC is plotted as a function of the total number of rank ordered sequences, using a threshold that yields a 30% false-positive rate (see Methods for a detailed description of this procedure). **(b)** For bins of 100 genes, sorted by the ChIP^2 rank, we computed MYC enrichment p-values relative to a background of unbound promoter fragments (solid blue curve). The x-axis represents the center of each bin. For each bin we also approximated the expected percent of bound genes, as computed by CSA, using the formula $FPR = \frac{FDR_r * r - FDR_l * l}{r - l}$, where FDR_r and FDR_l represent the FDRs for the genes at the right and left edge of the bin, respectively, and *r* and *l* represent their ranks. The dotted red curve displays this quantity, which is in excellent agreement with the sequence-based enrichment p-values.

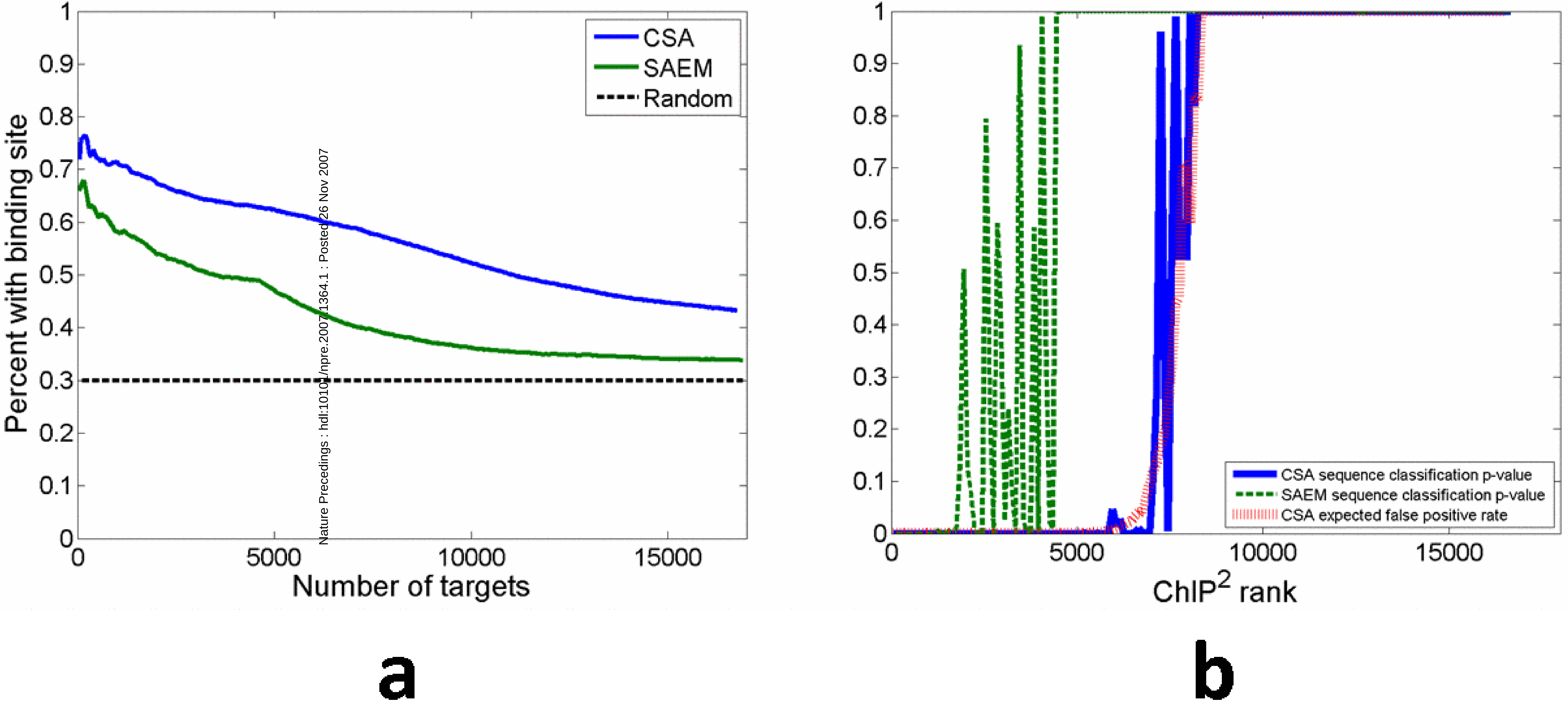
Figure 4 – Regulation of NOTCH1 target genes as a function of ChIP^2 rank

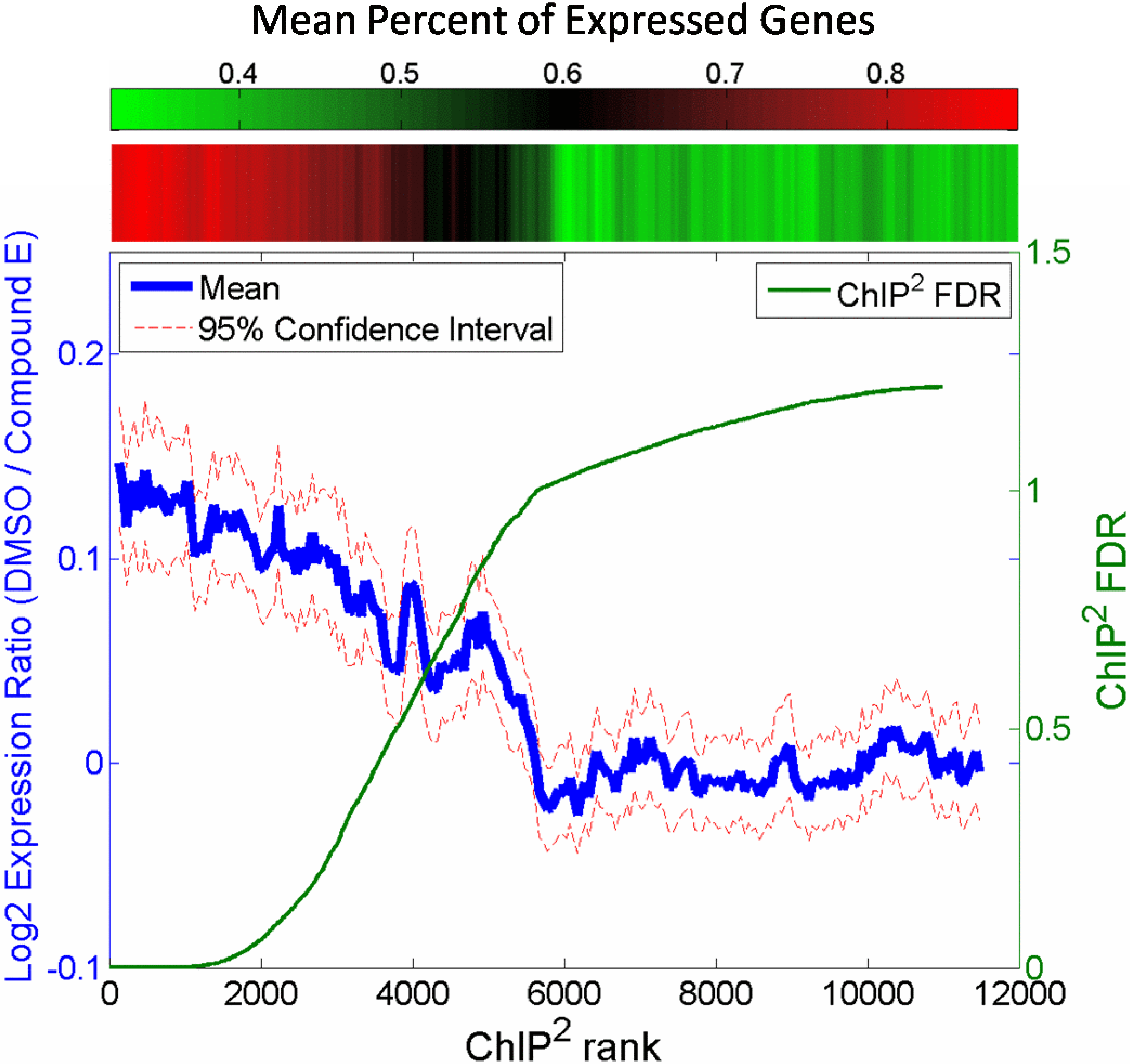
Genes are ranked according to their ChIP^2 FDR, plotted in green, as inferred by CSA (note that the Benjamini Hotchberg procedure can produce FDR values above one). The blue curve displays the mean log2 expression ratio of vehicle control compared to Compound E treatment, averaged over bins of 250 genes, with the 95% confidence interval plotted in red. The x-axis represents the center of each bin, and we use a sliding window with step size of fifty genes. Positive values indicate down-regulation upon NOTCH1 inhibition. The heat map above the plot displays the average percent of experiments in which the genes in the corresponding bin are expressed. Expression change upon NOTCH1 inhibition and the percent of expressed genes are highly correlated with the ChIP^2 ranking, and remain significantly enriched for over 5,000 predicted targets.


Fractional Count
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
-3
-2
-1
0
1
2
3
Log2(IP / WCE)
Unnormalized
Mean Centered
Gaussian Fit
2 SDs



a
b
c
NOTCH1
HPB-ALL
NOTCH1
CUTLL1
HES1
MYC
log2(IP/WCE)
(log2(IP) + log2(WCE)) / 2
all
mean
.1
.01
.001
Empirical Dist.
Inferred Mean
Inferred Null Dist.
log2(IP/WCE)


Percent with binding site
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
5000
10000
15000
Number of targets
CSA
SAEM
Random
a
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
5000
10000
15000
ChIP^2 rank
CSA sequence classification p-value
SAEM sequence classification p-value
CSA expected false positive rate
b


Mean Percent of Expressed Genes
0.4
0.5
0.6
0.7
0.8
Mean
95% Confidence Interval
ChIP^2 FDR
Log2 Expression Ratio (DMSO / Compound E)
0.2
0.1
0
-0.1
ChIP^2 FDR
1.5
1
0.5
0
0
2000
4000
6000
8000
10000
12000
ChIP^2 rank