ChIP-on-chip significance analysis reveals ubiquitous transcription factor binding

Adam A. Margolin Teresa Palomero Adolfo A. Ferrando Andrea Califano
Gustavo Stolovitzky

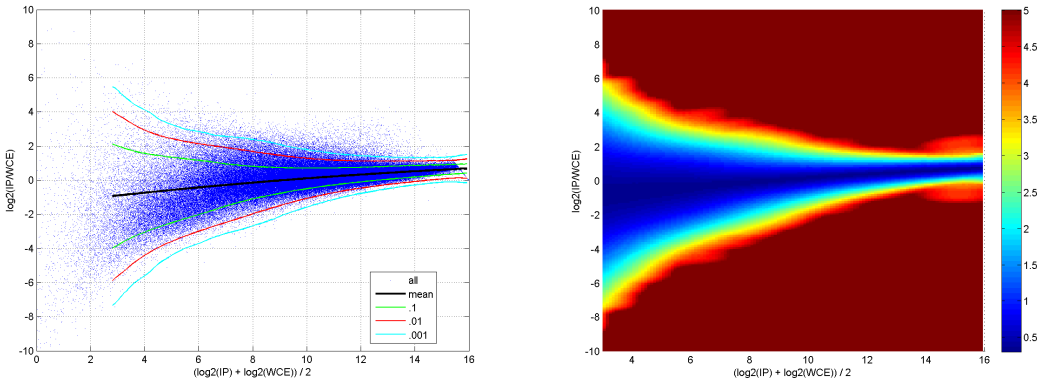

Figure 1: (Left) Magnitude versus amplitude (MA) plot of a ChIP-on-chip hybridization. The x-axis represents the average log2 intensity of the IP and WCE channels, and the y-axis represents the log2 ratio of IP/WCE. The black line represents the mean of the inferred null distribution, and the colored lines represent confidence intervals of .1, .01, and .001 probability. The model reveals an intensity dependent mean and variance of the null distribution, and a large number of probes are significantly enriched in the IP channel. (Right) The axes are the same as in the left panel, and colors represent the -log10 p-value of the null distribution.

ChIP-on-chip technology provides a genome-scale view of transcription factor (TF) / target interactions and a systems-level window into transcriptional regulatory networks. However, while many studies have used ChIP-on-chip data to effectively discover new TF targets, statistical methods have fallen short of developing an accurate model to disassociate signals caused by experimental noise from those caused by true biological variation, thus leveraging the technology to provide high confidence predictions of the full range of interactions.

This poster presents a novel method to accurately model the significance of binding events measured by ChIP-on-chip data. For each arrayed probe representing a genomic segment, a ChIP-on-chip microarray measures intensity levels for the IP channel, which is enriched in genomic fragments bound by an immunoprecipitated TF, and the WCE channel, which represents random genomic fragments. Statistical significance is inferred by computing the conditional probability, $p(M|A)$, where $M = \log 2\left(\frac{IP}{WCE}\right)$ and $A = \frac{\log 2(IP)+\log 2(WCE)}{2}$ (Fig. 1). A kernel density estimation procedure is used to calculate the joint probability, $p(M, A)$, and for each average intensity value, the mean of the null distribution (i.e. distribution for unbound probes) is inferred as $\hat{M}_A = \text{argmax}_M\, p(M|A)$. The distribution of $p(M|A)$, for $M < \hat{M}_A$, is then projected across $\hat{M}_A$ to yield the inferred null distribution, which is used to assign statistical significance scores. Probes for replicate experiments and probes with genomic locations within the fragmentation length ($\sim$ 500bp) are integrated to produce a single significance score for each genomic region.

The method is tested on six different ChIP-on-chip arrays representing replicate experiments for three different TFs (NOTCH1, MYC and HES1). For each experiment, this analysis reveals an order of magnitude more genomic binding events than detected by traditional analysis methods, predicting several thousand interactions for each TF and suggesting previously unappreciated complexity of transcriptional regulatory networks. Several independent experiments are used to provide evidence about the validity of these predictions. First, biochemical validation of more than 20 predicted targets by gene specific ChIP and qPCR confirm the accuracy of false discovery rate statistics computed by the method. Second, binding site enrichment analysis indicates that the strength of binding site signals are maintained over several thousand promoters. Finally, gene expression analysis reveals a coordinated downregulation of gene expression for the entire range of predicted NOTCH1 bound genes upon NOTCH1 inhibition experiments in cell lines, indicating that a large percentage of bound genes are also functionally regulated by the NOTCH1.