

Supplementary Discussion on Previous Work

Unlike JBD, existing computational methods for analyzing ChIP-Chip data do not attempt to detect binding events at a spatial resolution that is higher than the underlying microarray probe resolution. First generation ChIP-Chip microarray designs used 500-2,000bp probes to represent ORFs or intergenic regions with gaps in genome coverage. Computational methods for analyzing these arrays used either simple immunoprecipitate (IP) enrichment ratio cut-offs or an error model to compute p-values for single array probes^{1,2}. Methods for analyzing more recent ChIP-Chip data from microarrays with better genome coverage have used similar techniques^{3,4}. In both cases, there was no attempt to achieve effective spatial resolutions below the spacing of probes on the array.

Some recent Affymetrix and Nimblegen microarrays contain densely tiled short probes, and several studies have detected binding events by finding a fixed sized window of enriched probes. IP enrichment is localized to windows by using the Wilcoxon rank sum test to compare experiments to controls^{5,6}. Keles *et al.* performed an analysis of ChIP-Chip data from Affymetrix arrays, and did consider how DNA fragment lengths might influence the outcome of statistical tests for binding⁷. Buck *et al.* use a similar method in their ChiPOTle software⁸. However, neither method improves the spatial resolution of the data beyond the 500 to 1000bp reported in the previous studies of Cawley *et al.*⁵ and Bernstein *et al.*⁶.

MPeak discovers binding events by fitting a predicted shape to the enrichment log-ratios to identify which microarray probe location corresponds to a binding event⁹. However, MPeak

does not perform joint learning of binding events and therefore does not handle nearby events effectively. Furthermore, MPeak is more vulnerable to noise than JBD because of its lack of explicit noise modeling.

Previous computational methods for motif discovery have integrated ChIP-Chip data at a coarse scale, associating a single transcription factor binding estimate with a relatively long sequence of DNA, such as a 500-2,000bp intergenic region. For example, MDScan is a motif discovery tool for analyzing sequences from ChIP-Chip data that makes use of the enrichment ratios associated with each sequence to improve performance¹⁰. Conlon *et al.* presented Motif Regressor, an algorithm used to discover motifs that correlate with mRNA expression changes, and suggested it could also be used to discover motifs that correlate with localization data¹¹. Hong *et al.* used Motif Regressor in this capacity and showed how the resulting motifs could be used to build sensitive classifier-based models of DNA-binding specificity¹². Smith *et al.* presented a method that identifies statistically over-represented motifs in ChIP-Chip data, and then uses multivariate regression to evaluate these motifs and identify pairs of interacting motifs¹³. All these methods integrate positional information at a segmental level, associating regions of sequence with a single localization datum in an attempt to improve motif discovery performance.

Supplementary Discussion on the Influence Function

We have analyzed data for many transcription factors and chromatin marks in several species to determine the influence function for each factor. As the influence functions in Supplementary

Figure 4 make clear, the influence functions for two factors are generally very similar if the protocols and samples used in the two experiments were the same. We have seen variation in the influence function arise from protocol differences (either formal differences or differences in who performed a procedure) and from biological sources such as the type of protein being profiled (eg, transcription factor vs chromatin mark).

Ideally, each researcher should determine the distribution of fragment sizes (and thus the influence function) for each biological replicate of each ChIP-Chip experiment. In practice, we expect that researchers will measure the DNA fragment size distribution until they are confident that their results are consistent from experiment to experiment.

1. Roberts, C. *et al.* Signaling and circuitry of multiple MAPK pathways revealed by a matrix of global gene expression profiles. *Science* **287**, 873–880 (2000).
2. Lee, T. *et al.* Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* **298**, 799–804 (2002).
3. Pokholok, D. K. *et al.* Genome-wide map of nucleosome acetylation and methylation in yeast. *Cell* (2005).
4. Lee, T. I. *et al.* The active and polycomb-repressed genome in human embryonic stem cells. *Science* (2005).
5. 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).
6. Bernstein, B. E. *et al.* Genomic maps and comparative analysis of histone modifications in human and mouse. *Cell* **128**, 169–181 (2005).
7. Keles, S., Dudoit, S., van der Laan, M. & Cawley, S. E. Multiple testing methods for ChIP-Chip high density oligonucleotide array data. *Berkeley Electronic Press* (2004). [Http://www.bepress.com/ucbbiostat/paper147](http://www.bepress.com/ucbbiostat/paper147).
8. Buck, M. J., Nobel, A. B. & Lieb, J. D. Chipotle: a user-friendly tool for the analysis of chip-chip data. *Genome Biology* (2005).
9. Kim, T. H. *et al.* A high-resolution map of active promoters in the human genome. *Nature* **436**, 876–880 (2005).

10. Liu, X. S., Brutlag, D. L. & Liu, J. S. An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. *ature Biotechnology* **20**, 835–839 (2002).
11. Conlon, E. M., Liu, X. S. & Liu, J. S. Integrating regulatory motif discovery and genome-wide expression analysis. *Proc Natl Acad Sci* **100**, 3339–3344 (2003).
12. Hong, P. *et al.* A boosting approach for motif modeling using ChIP-chip data. *Bioinformatics* **21**, 2636–2643 (2005).
13. Smith, A. D., Sumazin, P., Das, D. & Zhang, M. Q. Mining ChIP-chip data for transcription factor and cofactor binding sites. *Bioinformatics* **21**, i403–i412 (2005).