Published in final edited form as:
Mol Biosyst. 2009 December ; 5(12): 1429–1438. doi:10.1039/B906880e.

Scoring overlapping and adjacent signals from genome-wide ChIP and DamID assays†

Audrey Qiuyan Fu[a,b] and **Boris Adryan**[*,a,c,d]

[a]Cambridge Systems Biology Centre, University of Cambridge, Tennis Court Road, Cambridge, UK CB2 1QR

[b]Department of Physiology, Development and Neuroscience, University of Cambridge, Downing Street, Cambridge, UK CB2 3DY

[c]Department of Genetics, Department of Genetics, Downing Street, Cambridge, UK CB2 3EH

[d]Cambridge Computational Biology Institute, Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Wilberforce Road, Cambridge, UK CB3 0WA

Abstract

Much of the research utilising genome-wide ChIP and DamID assays aims to understand the combinatorial feature of transcription factor binding and the chromatin modification code.With these experimental methods becoming more affordable and widespread, the focus of research is shifting to making sense of the data. Amongst the many challenges arising from data analyses, we are concerned with identifying biologically meaningful co-occurrences of transcription factor binding or chromatin modifications, using genome-wide profiles generated from ChIP and DamID assays. Co-occurrences are reflected in overlapping and adjacent signals in multiple ChIP or DamID profiles. We review existing quantitative methods to score overlaps and to cluster binding events in ChIP and DamID profiles. For pairwise comparison, existing methods either are based on a single score at the genome level or take a genomic, region-specific view. To draw inference from many profiles simultaneously, methods exist to cluster regions by their regulatory importance or to infer *cis*-regulatory modules for a particular region. We provide a simple guide to some of the statistical tools used by these methods.

Introduction

The combinatorial binding of transcription factors (TFs) to gene regulatory regions is an integral component of gene regulation. The binding of a few different regulatory proteins to individual gene promoters can already be observed in prokaryotic organisms such as *E. coli*,[1,2] and simple eukaryotic organisms such as *S. cerevisiae*.[3-5] In higher organisms, the size and composition of regulatory regions seen for many genes involved in metazoan development provide a glimpse into the complexity of transcriptional regulation. Specifically, different enhancer and silencer sequences often integrate the binding of a variety of TFs to enable a precise context-specific target gene response. These complex regions are embedded into the chromatin structure, and the accessibility of TFs to their target sites is determined by an entire set of histone modifications.[6] Such a complex regulatory region can be found in the vertebrate *globin* locus that requires the integration of

[†]This article is part of a *Molecular BioSystems* themed issue on Computational and Systems Biology.

[*]ba255@cam.ac.uk.
qf205@cam.ac.uk

many *cis*- and *trans*-acting factors for the process of 'globin switching'.[7] The most intriguing and extensively studied examples of the control of spatio-temporal gene expression by the integration of various TF input signals include the regulation of individual stripes of the pair-rule gene *even-skipped* in *Drosophila* development,[8-10] and the control of *CyIIIa* and *Endo16* genes in the sea urchin.[11] Auto- and co-regulating TFs form functional modules, and ultimately, gene regulatory networks as those seen in early developmental processes in flies and Echinoids.[12,13]

In this review, we consider the problem of inferring potentially biologically meaningful co-occurrences in terms of transcription factor binding events or chromatin modifications (including histone modifications and DNA methylation). 'Co-occurrence' refers to multiple TFs binding to or several chromatin modifications occurring in the same genomic region or genomic regions close to each other (Box 1A). Insights from these investigations will provide an entry point to detecting functional modules formed by two or more TFs, to deciphering important combinations of histone modifications in gene regulation, as well as to identifying genomic regions with an elevated level of regulatory input ('co-localisation hotspots'[14]).

Chromatin immunopurification (ChIP) is currently the method of choice to detect transcription factor binding sites (TFBSs) and chromatin modifications on the genome-wide level. It is an *in vivo* method that captures genomic DNA bound and cross-linked to a DNA-binding protein.[15] It is versatile, allowing for the probing of site-specific TFs, the basal RNA polymerase machinery, other DNA-binding proteins, as well as chromatin modifications using specific antibodies. DamID, a variation of this technique, is useful when no suitable antibodies exist for ChIP, although it suffers a lower resolution than the ChIP method in general. Specifically, DamID employs a fusion of Dam methyltransferase to the DNA-binding protein of interest.[16] As there is no equivalent endogenous enzyme in most eukaryotic species, only DNA lying in the vicinity of the binding site is methylated and can be purified using a methyl-N^6-adenine-specific antibody. The ChIP method is being used in conjunction with microarrays (ChIP-on-chip) to probe large portions of the genome,[17-20] or with massively parallel and next-generation sequencing (ChIP-seq)[21-24] where no whole-genome array is available. These high-throughput techniques enable the genome-wide detection of ChIP-enriched sequences in an unprecedented manner. With current advances in genomic microarrays and high-throughput sequencing, they are likely to become even more widely used (for a comparative review, see Aleksic and Russell[25] in this issue).

Unfortunately, the resolution of these genome-wide techniques is nowhere near the actual DNA sequence recognised by TFs yet. Target sequences of TFs are thought to be less than 15 bp in length in most cases. Signal-enriched genomic regions in the ChIP and DamID profiles, however, often range between 500 bp and 5 kb, depending on chromatin preparation, detection technique and platform (more details below). Computational methods to improve the physical resolution of ChIP assays do exist[26,27] but have not seen broad application yet. Computer programs also exist to identify the causative TFBSs specifically in ChIP-enriched regions,[28-30] which resembles *de novo* motif finding and has not been solved sufficiently.[31,32]

Despite these shortcomings, ChIP and DamID methods have generated extremely valuable data, which one can use to study co-occurrences of TFBSs or chromatin modifications (together referred to as 'events' hereinafter). These co-occurrences are represented by overlapping and adjacent signals in the profiles of two or more factors (Box 1 and more details below). In this review, we review existing methods for scoring these overlapping and adjacent signals for pairs of factors and for many factors. Our review is by no means exhaustive, but aims to lay out some basic ideas, their connections to each other and their relative merit.

ChIP and DamID profile data

Because of different designs, ChIP (ChIP-on-chip and ChIP-seq) or DamID assays generate different types of raw data that have different resolutions. The genomic resolution of microarrays used in ChIP-on-chip depends on the length and distribution of the probes (Box 1B). ChIP-seq, on the other hand, generates a large amount of overlapping short reads over the whole genome (see Aleksic and Russell[25]). Hence, the raw data from ChIP-seq are counted in many short genomic regions of varying lengths. As a variant of ChIP-on-chip assays, DamID assays also produce probe intensities. However, their resolution is determined by the chromatin preparation itself, but not the microarray resolution. In fact, early applications of DamID utilised cDNA arrays for detection, as the methylation tag tends to spread far into gene regions even in cases of intergenic TFBSs.

The raw and processed data represent ChIP and DamID profiles at three levels of resolution; in decreasing order: the probe level, the peak-region level and the gene level. Profiles at the probe level contain the raw intensities at all probes. Those at the peak-region level process the raw data to focus on ‘signal-enriched genomic regions’, associating each region with a score summarising the regional signal intensity or a binary value. This method is also applied for ChIP-seq data. Dichotomising the raw data and assigning intergenic events to nearby genes generates gene-level profiles. This approach implies that the closest gene is the regulatory target, which is probably fair to assume when the events are within the proximal promoter but can be arbitrary for binding events occurring just in between two genes.

ChIP and DamID profiles are thus imperfect representations of true binding events and chromatin modifications (Box 1C). These imperfections will affect the inference of co-occurrences and conclusions drawn from the inference.

Co-occurrences and their representations in ChIP or DamID profiles

To infer co-occurrences, we may compare ChIP or DamID profiles from the same microarray (*e.g.* intensities from the same probes), from different assays (*e.g.* binary peak-region profiles from different ChIP-on-chip assays) or across platforms (*e.g.* binary gene-level profiles between ChIP-on-chip and ChIP-seq assays).

Co-occurrences are represented by overlapping and adjacent signals in ChIP or DamID profiles. At high resolutions, such as at the peak-region and probe levels, completely overlapping, partially overlapping and adjacent signals are all informative of co-occurrences. At the gene level, one can usually observe only complete overlaps (Box 1C).

Methods of scoring overlapping and adjacent signals in ChIP or DamID profiles

The complexities with ChIP or DamID profile data, including uncertainty, measurement error and different resolutions, pose challenges to inferring and scoring co-occurrences between factors. How to score co-occurrences across different platforms? How to measure the amount of complete as well as partial overlaps? How close to each other should the events be to be considered adjacent? How to evaluate the statistical significance of the co-occurrence score? What assumptions underlie these statistical considerations? These are relevant questions in light of the current endeavours to map TFBSs and chromatin modifications on a large scale. In addition, the wealth of data previously generated may require the cross-platform integration, even at a comparatively low resolution.

The methods reviewed here are listed in Table 1. With only two profiles, existing methods to score co-occurrences generally take one of these two approaches: one summarises the genome-wide level of co-occurrence in a single score, which effectively ignores the variability of events across genomic regions; the other acknowledges this variability and assesses the level of co-occurrence in individual genomic regions. When the profiles are available for many factors, it is possible to identify ‘co-localisation hotspots’, or identify *cis*-regulatory modules for each of a small set of genes. In addition to reviewing these methods, we provide a simple guide to basic statistical concepts behind some of the approaches discussed here.

Methods for pairs of factors

Methods based on a genome-wide score

Simple counting—This strategy is suitable for scoring (complete) overlaps in two profiles. At each of the probe, peak-region and gene levels, one may simply count the number of genomic regions where either or both factors bind. These counts can be summarised in a contingency table (Box 2) and displayed in a Venn diagram showing the overlaps as the intersection of two circles.

The potential to learn about the modular binding of DNA-binding proteins on a genome-wide level was already recognised in some of the first ChIP-on-chip publications. A simple counting strategy was employed when Lieb *et al.*[18] probed the genomic binding of yeast Rap1 and its binding partners. They found a general agreement in the signals based on the counts and presented this in a Venn diagram. The same strategy was adopted in the pioneering work of Iyer *et al.*[17] on the binding profiles of yeast SBF and MBF complexes during cell cycle progression. In fact, most studies published to date have used a similar strategy. Recent examples include the comparison of TFBSs from different tissues at the gene and peak-region levels to account for the effect of different resolutions,[33] and the probing of chromatin complexes such as Polycomb and its DNA-binding partner Pleiohomeotic, the two of which are likely to co-occur.[34]

Simplicity is the main advantage of the counting strategy. It provides a quick and succinct summary of the level of co-occurrence for large profiles. This summary is also easily interpretable even to non-experts, but this simplicity leads to the following problems: (i) it does not work well for partial overlaps (*e.g.* partially overlapping peak regions; see Box 1). Treating a partial overlap as a complete one can lead to overestimation of co-occurrence, whereas treating it as no overlap can lead to underestimation. (ii) It ignores spatial information, *e.g.* information about recurrent clusters of binding sites may be lost. The counts treat all probed regions as equally spaced and their positions interchangeable. This treatment may lead to underestimation of co-occurrence, considering that TFBSs and certain chromatin modifications tend to cluster.

Correlation coefficient—Another summary statistic of the overall similarity between pairs of profiles is Pearson’s correlation coefficient r (Box 2), which can be computed for intensities as well as dichotomised data and at each of the three levels (probe, peak-region and gene). On a scale of 0 (independent) to 1 (fully correlated), scores higher than $r = 0.4$ are generally interpreted to indicate significant levels of similarity. For example, Orian *et al.*[35] compared the genomic binding of dMyc/dMax/dMad by calculating rs between microarray intensities for the cDNAs in their DamID assay. This quantity was also used to correlate ChIP and DamID binding data of *Drosophila* TFs on the probe level from a tiling array.[14,36]

As part of the ENCODE project, Zhang *et al.*[37] computed rs for pairs of dichotomised profiles (*i.e.* each base pair is scored '1' for binding and '0' for no binding) of more than 100 TFs.

One needs to be careful when interpreting r. It is a measure of the overall similarity (co-occurrences and non co-occurrences) rather than co-occurrences only. Consider the following over-simplified example. Binding profile '10000' for both TF1 and TF2 leads to $r = 1$. Binding profiles '11110' and '11100' for TF3 and TF4, respectively, give $r = 0.6$. However, TF3 and TF4 clearly have more co-binding events than TF1 and TF2 do. In other words, high correlation coefficient values can be due to many co-occurrences as well as a large number of un-occupied regions. Other distance measures (*e.g.* the weighted Hamming similarity in the case of binary profiles) may therefore be more suitable.

Similar to simple counting, the use of correlation coefficients also discards spatial information in the data, treating probed regions as equally spaced and their positions interchangeable. To adjust for this effect, Zhang *et al.*[37] further applied a sliding-window approach to the binary binding profiles at the base pair level. They counted the number of 1s in each sliding window and then computed *rs* for these sliding windows for pairs of TFs. This adjustment, however, effectively modifies the original binding profile, putting most weight at the centre of each enriched region, although this is not necessarily the most suitable approach.

Hypothesis tests based on a single score—Counts of overlaps between profiles do not themselves indicate statistical significance. For example, a nearly ubiquitously binding factor may overlap with all events of a more rarely binding factor. Are their overlaps convincing evidence for a recurrent pairwise interaction? To answer this question, current methods generally set it up as a hypothesis testing problem, with the 'null' hypothesis being that co-occurrences between two factors are due to chance, and draw conclusions based on the p-value (Box 2). The first three tests reviewed below are based on contingency tables and make explicit distribution assumptions on the counts in the contingency table. Although having a theoretical appeal, these tests ignore adjacent signals and discard the spatial structure in the observed events along the two profiles. Permutation tests have been used or proposed to address these issues.

Hypergeometric test—A hypergeometric test (Box 2), as a simple case of Fisher's exact test, can be applied to the contingency table for two profiles to assess the significance of the counts. In one of the first genome-scale ChIP experiments, Simon *et al.*[20] reported the genomic binding to a target gene promoter of nine TFs involved in cell cycle control. The authors dichotomised the binding profiles and determined the statistical significance of the preference for a TF to bind to genes active during specific stages of the cell cycle assuming a hypergeometric distribution, thus establishing the connection of the cell cycle-regulating TFs with their target genes. On the other hand, the authors did not assess the statistical significance for TF co-occurrence.

The data of these initial experiments in yeast were complemented with binding profiles of a further set of TFs, yielding genome-wide binding information for 203 proteins, some of which were assayed in a variety of physiological conditions.[3] Several groups attempted 'regulatory code breaking' on these data, looking at either combinatorics at the promoter level[3] or the regulatory network structure.[5] By integrating ChIP profiles, phylogenetic conservation and published evidence, Harbison *et al.*[3] determined the binding specificity of 102 TFs and inferred putative TFBSs based on these position weight matrix (PWM) matches. Altogether, this yielded a map comprising of 3353 TF–DNA interactions at 1300 promoter regions. Amongst those hundreds of co-occuring TFs, the authors used the hypergeometric test to detect almost 100 pairs of TFs that may co-bind to some genomic regions more often than expected at random.

The hypergeometric test has long been used on gene-level counts as the examples above show. A more recent example is the mapping of binding sites of the stem cell chromatin remodelling complex esBAF by ChIP-seq.[38] The authors assigned esBAF binding as well as occurrences of Oct4, Sox2, Nanog and others to genes. Hypergeometric testing revealed a significant overlap between esBAF and these core pluripotency players.

Chi-square test—With improved microarray resolution, profiles on the peak level result in large counts in the contingency table. A chi-square test provides a good approximation to the hypergeometric test (Box 2). For the comparison of binding profiles for Polycomb PRC1 proteins on a 10 Mb tiling array, Nègre *et al.*[36] dichotomised probe signals and carried out a chi-square test to prove that PRC1 proteins significantly co-occur.

Log-linear model—Another approach is to fit a log-linear model to the contingency table. Datta and Zhao[39] used this idea to assess co-binding between TFs based on the ChIP-on-chip data from Harbison *et al.*[3] Instead of the contingency table obtained directly from the profiles, they used the *p*-values for the microarray intensity values as the data to infer the true binding states for two TFs. The inferred binary binding states across genes were summarised in a contingency table. They then fitted this inferred contingency table with a log-linear model (Box 2). Under standard theory, one can test whether the coefficient for the interaction term is significantly greater than 0. If so, then one may conclude significant co-binding and potential cooperativity between the two TFs. Their simulation study confirmed that this approach, working with those *p*-values rather than the dichotomized binding data, has more power. Meanwhile, the authors acknowledged that this regression approach does not easily generalise to comparing multiple TFs simultaneously, because in general, higher-order interactions (interactions between at least three TFs) are not easily detectable.

Permutation test—As mentioned before, permutation tests (Box 2) offer a flexible alternative to the above three tests. They are applicable to any co-occurrence score defined by the user and use a null distribution generated from permuting the observed profiles, which helps preserve the spatial structure inherent in the data.

How to permute the observed profile has been an issue though. Allowing a uniform random distribution over the genome may not well reflect its regulatory architecture, *e.g.* concentrated occurrences of TFBSs around core promoters. Haiminen *et al.*[40] is one of the few studies that explicitly tackles this problem. Through simulation, they demonstrated that the above naive scheme is prone to false positives. They therefore proposed two other permutation schemes. The first permutation scheme, also used in Hannenhalli and Levy,[41] and Klein and Vingron[42] in the analysis of computationally predicted binding sites, fixes the positions of binding sites of the two TFs and randomly assigns a site to a TF, according to the frequency of the binding events of each TF. The second permutation scheme retains the observed binding profile for one TF and permutes other profiles according to the first scheme. Both schemes produced much fewer false positives than the naive scheme did, based on simulated data. The second proposed scheme, however, applies only to three or more profiles, and may not be symmetric amongst profiles.

Methods to account for variability of events across genomic regions

The above methods based on a genome-wide score provide an overview of the level of co-occurrence. It is, however, reasonable to expect different amounts of co-occurrences across different genomic regions. Hence, methods that account for the spatial variability may offer more insights on the study of co-occurrences between pairs of factors, although not many methods are available for this purpose.

By assuming that key parameters in different genomic regions vary according to a common probability distribution, hierarchical modelling is an effective approach to incorporate variability across genomic regions. Feng *et al.*[43] used this approach to compare binding profiles of RNA Polymerase II (PolII) before and after certain treatment for each gene using ChIP-seq. The authors modelled the count of PolII-targeted fragments for each gene with a Poisson distribution with a gene-specific mean parameter. These mean parameters were further assumed to come from two categories, different (category 1) and not different (category 2) after the treatment. The Poisson mean parameters in each category follow a distribution. They then estimated the posterior probability (*i.e.* the probability after accounting for the observations and prior knowledge) of the counts belonging to, say, category 1, for each gene. A high value for this posterior probability would indicate that the binding behaviour in that gene is more likely to be different before and after treatment. Using gene-specific Poisson distributions to retain the stochasticity in genes, this method allows for the identification of individual genes that are bound differently before and after the treatment.

The hierarchical model in Feng *et al.*[43] considers the gene-level profiles and assumes independence amongst the genes. This assumption, however, does not hold at higher resolutions. Xu *et al.*[44] applied the widely-used hidden Markov model (HMM) approach to compare two ChIP-seq profiles of histone modifications, accounting for spatial dependence across genomic regions, each of length 1 kb. The HMM consists of two layers: the hidden Markov chain, on which each variable represents which profile is significantly enriched in a genomic region; the observed layer, *i.e.* the ChIP-seq counts of the two profiles given rise to by the hidden variables. Inferring the hidden enrichment state in each region depends on the observed ChIP-seq counts in that region as well as the enrichment state in the neighbouring regions. Although it might be difficult to work with ChIP-seq data at even higher resolutions, the HMM can be a powerful tool to incorporate spatial variability in binding events or chromatin modifications.

Simple regional comparisons are also possible without assuming any underlying signal distribution. In their study to capture the binding preference of the histone acetyl-transferase MOF in *Drosophila*, Kind *et al.*[45] normalised ChIP signals within gene regions according to the gene lengths, binned the normalised signals into fragments of 10% the length of a ‘standard gene’, and thus established the differences in MOF binding to genes on the X chromosome and on autosomes.

Methods for many factors

With more ChIP profiles becoming available, pairwise comparison is not adequate to reveal the interplay between many factors. Binding information from many factors gives one more power than pairwise comparison does in terms of identification of ‘co-localisation hotspots’ and *cis*-regulatory modules.

Overall assessment of co-occurrence

One can count the number of observed events (from different factors) in each genomic region and summarise these counts in a histogram, also known as an empirical distribution. This empirical distribution reflects the genome-wide level of co-occurrence amongst many factors. A different type of permutation test from the one reviewed above can be used to assess whether the observation of this empirical distribution is due to randomness. The common strategy is to permute occurrences of different TFs or chromatin modifications amongst occupied regions, or the other way round. For example, in the first large-scale ChIP-on-chip experiment for the majority of an organism's TF repertoire, Lee *et al.*[4] obtained binding data for 106 yeast TFs. By randomly shuffling regulators amongst bound promoters, the authors found many promoters bound by four or more TFs, which is unlikely to be a result of chance, thus confirming the existence of multi-input motifs. Chen *et al.*[46] also used this strategy and established that four or more TFs of the core pluripotency network in stem cells occupy a highly significant proportion of enhancer sites. Cuddapah *et al.*[47] used a similar strategy to determine the significance of mapped CTCF binding sites falling into specific chromatin domains.

Identification of 'co-localisation hotspots'

Multiple testing based on Poisson distribution—One can test in each genomic region whether a statistically significant number of binding events from different factors co-occurred in this region. Chen *et al.*[46] termed regions with significant enrichment for multiple TFs 'multiple transcription factor binding loci' (MTLs). To identify MTLs, they used Poisson distributions combined with Fisher's method to assess the statistical significance of co-binding in individual regions. One can assume that the majority of genomic binding events of a factor occur independently at a relatively low frequency along the genome. This behaviour is best described with a Poisson distribution (Box 2). The authors calculated a *p*-value for each TF using a Poisson distribution to determine if the number of binding events in a genomic region under consideration appears more often than would be expected looking at the overall occurrences of this TF in the genome. The multiple *p*-values for all the TFs calculated this way provide the basic material for local clustering of TFs. The authors then derived a combined *p*-value using Fisher's method (Box 2) for each region. Whether each combined *p*-value is significant was further determined by applying the Benjamini–Hochberg method, which controls the genome-wide false positive rate.[48]

Clustering—Instead of determining the statistical significance of co-occurrence in individual regions, one can also cluster genomic regions according to their binding affinity to TFs or chromatin modifications. Here, we borrow the term 'co-localisation hotspots' from Moorman *et al.*[14] to refer to clusters with a high affinity for factors. Many available clustering algorithms[49,50] take one of the two approaches. One is the non-hierarchical clustering approach (*e.g.* the widely-used K-means clustering), which requires the user to specify the number of groups. This has been used, for example, for clustering of non-overlapping genomic windows of 100bp within 10 kb around gene transcription start sites, and has been used in Heintzman *et al.*[51] to correlate four classes of histone modification sets with gene activity. The other method is the hierarchical clustering approach, which does not make this requirement and generally produces a dendrogram by repeatedly pairing similar items (*e.g.* Zhang *et al.*[37] organised 18 frequently co-occuring TFs into a dendrogram based on the pairwise correlations amongst the profiles).

Moorman *et al.*[14] combined K-means clustering with a self-organising map (SOM; which, roughly speaking, can be considered as non-hierarchical clustering) in a two-step procedure to identify 'co-localisation hotspots' using the tiling array profiles for a set of factors including seven TFs. Their profile data consist of normalised intensity values of 1 kb genomic tiles of a 3 Mb region. They applied the SOM method in their first step. Genomic regions to which a similar set of TFs tends to bind are grouped together. Unlike other clustering techniques, this self-organising map approach locates the clusters of genomic regions on a hypothetical grid, which has a much lower dimension than the number of genomic regions. This grid then retains information of the distances between clusters. K-means clustering in the second step used this distance information to carry out 'super-clustering'—grouping similar nodes (genomic regions) on the grid together to further reduce dimensions. The authors found eight clusters, termed ‘chromatin types’, with different binding affinity for different sets of TFs. In particular, the authors defined genomic regions in the chromatin type to which essentially all the TFs investigated bound as ‘co-localisation hotspots’. Although this two-step clustering procedure may have over-simplified the problem as the authors pointed out, the identified chromatin types did provide a coarse summary of the binding affinity of the probed regions over the whole genome for many factors.

Identification of TFs for *cis*-regulatory modules for a gene—Instead of clustering genomic regions, Zeng *et al.*[52] used a regression-based approach to cluster TFs for a gene, thus identifying *cis*-regulatory modules (CRMs) based on ChIP measurements, as opposed to theoretical TFBS and CRM predictions.[53-56] The authors used time-course data of the yeast cell cycle and regressed the expression levels of a gene of interest on those of the genes corresponding to a set of TFs. The aim then was to select a subset of TFs whose gene expression levels are highly correlated with that of the gene of interest. During this selection process (variable selection in statistical terms), it was also desirable to keep all the TFs with high correlations with the gene, rather than just one or two as the representatives (which the authors showed was indeed the case with several variable selection techniques). Zeng *et al.*[52] applied factor regression to the TFs, regressing the TFs on a few factors (groups of TFs). Factors significantly correlated with the expression levels of the gene of interest were then considered to be co-binding at CRMs.

Conclusions

In this article we have reviewed a variety of methods for scoring co-occurrences across ChIP or DamID profiles. As a concluding remark, we would like to emphasise visual display of the profile data as an important aspect of exploratory data analysis.[57] As mentioned before, Moorman *et al.*[14] used a self-organising map to display the genome-wide binding landscape for several factors. Zhang *et al.*[37] displayed in a biplot three types of relationship for the ENCODE data: TFs and TFs, TFs and genomic regions, and between genomic regions. Clusters and other patterns are easily identified from those displays.

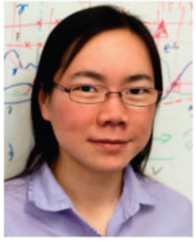
It is also important to apply different methods to the data before settling on the most suitable ones. This is because each method may offer a unique perspective (*e.g.* very different contingency tables may lead to the same small *p*-value), but depends on sometimes unsuitable assumptions (*e.g.* Pearson correlation coefficient measure similarity rather than co-occurrence).

Last but not least, we are interested in using ChIP measurements, rather than theoretically predicted binding sites, to score co-occurrences. Although most methods reviewed here should apply also to these computationally predicted profiles, we believe that observed profiles provide more direct information of the binding behaviours, and therefore may lead to more biologically sensible conclusions on the functional relationship between TFs and chromatin modifications.

Acknowledgments

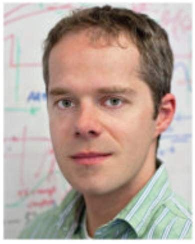
AF would like to acknowledge support from the Biology and Biotechnology Research Council through a project grant awarded to Prof. Sarah Bray (Department of Physiology, Development and Neuroscience, University of Cambridge). BA is a Royal Society University Research Fellow and wishes to acknowledge support from The Wellcome Trust through a project grant awarded for the study of transcription factor binding sites.

Biographies



Audrey Qiuyan Fu

Audrey Qiuyan Fu received her PhD in Statistics (Statistical Genetics) from the Department of Statistics at the University of Washington, Seattle, USA. She is currently a postdoctoral research associate in the Department of Physiology, Development and Neuroscience and the Cambridge Systems Biology Centre at the University of Cambridge. Her research interests include developing statistical methodologies for problems in biology and genetics to understand functional aspects of the genome.

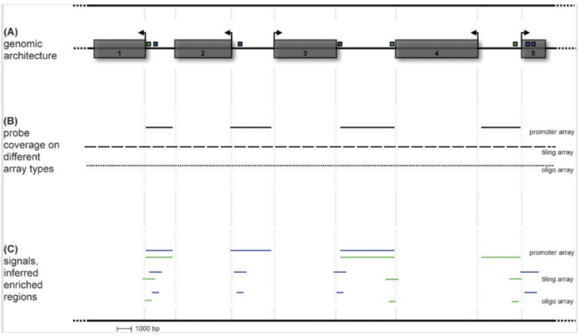


Boris Adryan

Boris Adryan is a developmental biologist by training, obtained a PhD for work at the Max-Planck-Institute for Biophysical Chemistry (Göttingen, Germany), and was an EMBO Long-Term Fellow at the MRC Laboratory of Molecular Biology in Cambridge. He is a Royal Society University Research Fellow at the Cambridge Systems Biology Centre and Member of the Cambridge Computational Biology Institute. His group works on transcriptional regulation in developmental processes, and genomics & computational methods to study them.

Box 1: Co-occurrences and their representations in ChIP/DamID profiles

Overlapping and close-by signals are representations of potentially meaningful co-occurrences between multiple factors. With our limited understanding of a eukaryotic regulatory logic, we do not yet know the precise spatio-temporal requirements for modulating a regulatory function. We can, however, begin to infer some of these rules based on genome-wide measurements from ChIP/DamID assays. We illustrate this with an example of two transcription factors and their binding sites:



(A) Co-occurrence of transcription factor (TF) binding refers to the binding sites of one TF being physically close to those of another TF. Displayed are the binding sites of two TFs (blue and green squares) on a piece of genome encoding five genes (grey boxes; arrows denoting their transcription start sites). While the factors exhibit some independent binding, the clustered occurrence of binding sites for both around the transcription start sites of genes 1 and 5 may well represent an important feature.

(B) ChIP-on-chip arrays detect transcription factor binding sites (TFBSs) with different resolution and genomic coverage. ChIP assays detect TFBSs by extracting DNAs bound to the TFs and then identifying these DNAs using microarrays (ChIP-on-chip) or sequencing-based (ChIP-seq) methods. **Upper Panel:** Promoter arrays, the first-generation detection platform, capture TF binding with one probe per intergenic region of length on the order of kilobases. **Middle Panel:** Genomic tiling paths of PCR amplicons cover essentially the entire genome with a much higher resolution. **Lower Panel:** Oligonucleotide arrays cover the whole genome with probes of ~25–75 bp densely spaced (often less than 100 bp apart). ChIP-seq (not shown) detects binding events at slightly improved resolution over oligo arrays (see also Aleksic and Russell[25] in this issue).

(C) Dichotomised binding profiles are imperfect representations of the truth. These imperfections affect the scoring of overlapping and nearby signals as well as making inferences about co-occurrence. Microarrays following ChIP assays deliver one intensity value per probe as the readout. Sequencing methods deliver stacks of sequence reads over a genomic region. Shown here are inferred binary ‘enriched regions’ from dichotomising those raw intensities (this is also known as ‘peak finding’). Enriched regions can be represented also by a continuous value (*e.g.* the average intensity value over the region, or the *p*-value associated with the peak).

Raw and processed ChIP profiles are imperfect representations of the true binding profiles. ChIP assays may miss true binding events by design: the promoter array missed the two blue TFBSs within gene 5. This promoter array also does not allow separation of the distinct binding events between genes 3 and 4. Dichotomisation lowers the resolution: the enriched regions for the blue factor in gene 5 merge into each other and are indistinguishable, perhaps due to dichotomisation with an improper threshold.

The inference of co-occurrence from these overlapping and nearby signals is strongly affected by these imperfections. Whereas the complete overlap from the promoter array correctly detects co-binding upstream of gene 1, it would produce a false signal for the intergenic region between genes 3 and 4, and fail to detect the co-occurrence around gene 5. Partial overlaps from the tiling and oligo arrays provide more precise information on co-occurrence locations. Clustered but non-overlapping signals, *e.g.* blue and green enriched regions at gene 5, are also indicative of co-occurrence, but they are more sensitive to distance thresholds needed to capture such events.

Box 2: Distributions, tests and binding profiles

Here, we explain some basic concepts underlying the methods reviewed in the text. Numerical results given in some concepts are based on the dichotomised binding profiles of two transcription factors, TF1 and TF2, over 20 genomic regions given below:

TF1	0	0	0	1	1	1	1	1	0	1	1	1	0	1	1	1	1	1	1	0
TF2	0	0	0	0	1	1	1	0	0	1	0	0	1	1	1	1	1	1	1	1

1. Contingency table: counting the genomic regions occupied by TF1 and/or TF2 or neither gives the following table:

		TF2		
		0	**1**	**Sum**
TF1	0	4	2	6
	1	4	10	14
		8	12	20

2. Pearson correlation coefficient r: this quantity gives equal weight to co-binding (1,1) and co-non-binding (0,0). Hence, high values may not necessarily imply high levels of co-occurrence. For the above example, $r = 0.36$.

3. p-Value: under the null hypothesis of overlaps occurring at random, the p-value is the probability to observe, say, 10 and even more overlaps as in the example. That is,

$$p=\Pr(T \geq t|H_0),$$

where T is the test statistic (*e.g.* number of overlaps), t the observed value (10 here), and H_0 the null hypothesis. One rejects the null hypothesis for small p-values.

4. Hypergeometric test: it tests for co-occurrence based on the contingency table, which can be re-written using random variables:

		TF2		
		0	**1**	**Sum**
TF1	0	$n-k+t$	$m-t$	$m+n-k$
	1	$k-t$	t	k
		n	m	$m+n$

Assume that the row and column sums (m, n, k) are fixed. The probability of observing t is hypergeometric. The p-value for the example is

$$p=\Pr(T \geq 10|H_0, \quad m=12, \quad n=8, \quad k=14)=0.14.$$

Europe PMC Funders Author Manuscripts

5. Chi-square test: it tests for dependence (not co-occurrence) between TF1 and TF2, and applies to contingency tables with very large counts (when the calculation of hypergeometric probabilities becomes cumbersome). Under this test, the difference between observed and expected counts can be approximated by a chi-square distribution with one degree of freedom. The difference is defined as

$$D=\sum_{i=1}^{2}\sum_{j=1}^{2}\frac{\left(O_{ij}-E_{ij}\right)^2}{E_{ij}},$$

where O_{ij}s are the observed counts, and E_{ij}s are the expected counts under the null hypothesis, and are computed using the fixed row and columns sums. For example, E_{22} is calculated as

$$E_{22}=\frac{mk}{m+n}.$$

The example has small counts in most cells. Hence, the chi-square test does not apply.

6. Permutation test: it tests for co-occurrence through repeatedly permuting observed enriched regions (or binding events) in one or both profiles many times. A pre-defined co-occurrence score is calculated for each permutation. Many permutations produce a null distribution of the co-occurrence score. One can then use this null distribution to compute a p-value for the observed co-occurrence score.

7. Poisson distribution: it can be used to compute how likely it is for a single TF to have, say, three binding events in 1 kb with 300 events in 1 Mb. The formula is

$$\Pr\left(x=3;L=1\quad\text{kb},\rho=\frac{300}{1000\quad\text{kb}}\right)=e^{-L\rho\frac{(L\rho)^x}{x!}},$$

where ρ is the binding rate per bp.

8. Fisher's method for combining p-values: one can calculate a p-value for each TF in a genomic region to assess whether that TF has more binding sites than expected in this region. To assess whether both TFs bind to more sites than expected, p-values can be combined using Fisher's method

$$P=-2\sum_{i=1}^{n}\log\quad p_i,$$

where n is the number of TFs. The quantity P has a chi-square distribution with $2n$ degrees of freedom. As before, a small combined p-value associated with the quantity P suggests co-occurrence.

9. Log linear model: as a type of regression model, it fits the counts in a contingency table as follows:

$$\log\left(C_{ij}\right)=\beta_0+\beta_1\,(\text{TF1})+\beta_2\,(\text{TF2})+\beta_{12}\,(\text{TF1}\times\text{TF2})+\varepsilon_{ij},$$

in which C_{ij} is the count in the (i,j)th cell, βs the effects and ε_{ij} the error term. Hence, the logarithm of each count depends on the main effects (β_1 and β_2) of the two TFs as well as the interaction effect β_{12}. An estimate of β_{12} significantly greater than 0 indicates co-occurrence.

References

1. Shen-Orr SS, Milo R, Mangan S, Alon U. Nat. Genet. 2002; 31:64–68. [PubMed: 11967538]
2. Thieffry D, Huerta AM, Perez-Rueda E, Collado-Vides J. BioEssays. 1998; 20:433–440. [PubMed: 9670816]
3. Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, Danford TW, Hannett NM, Tagne JB, Reynolds DB, Yoo J, Jennings EG, Zeitlinger J, Pokholok DK, Kellis M, Rolfe PA, Takusagawa KT, Lander ES, Gifford DK, Fraenkel E, Young RA. Nature. 2004; 431:99–104. [PubMed: 15343339]
4. Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM, Harbison CT, Thompson CM, Simon I, Zeitlinger J, Jennings EG, Murray HL, Gordon DB, Ren B, Wyrick JJ, Tagne JB, Volkert TL, Fraenkel E, Gifford DK, Young RA. Science. 2002; 298:799–804. [PubMed: 12399584]
5. Luscombe NM, Babu MM, Yu HY, Snyder M, Teichmann SA, Gerstein M. Nature. 2004; 431:308–312. [PubMed: 15372033]
6. Kouzarides T. Cell. 2007; 128:693–705. [PubMed: 17320507]
7. Guerrero G, Delgado-Olguin P, Escamilla-Del-Arenal M, Furlan-Magaril M, Rebollar E, De La Rosa-Velazquez IA, Soto-Reyes E, Rincon-Arano H, Valdes-Quezada C, Valadez-Graharn V, Recillas-Targa F. Comp. Biochem. Physiol., Part A: Mol. Integr. Physiol. 2007; 147:750–760.
8. Andrioli LPM, Vasisht V, Theodosopoulou E, Oberstein A, Small S. Development. 2002; 129:4931–4940. [PubMed: 12397102]
9. Harding K, Hoey T, Warrior R, Levine M. EMBO J. 1989; 8:1205–1212. [PubMed: 2743979]
10. Small S, Blair A, Levine M. Dev. Biol. 1996; 175:314–324. [PubMed: 8626035]
11. Kirchhamer CV, Yuh CH, Davidson EH. Proc. Natl. Acad. Sci. U. S. A. 1996; 93:9322–9328. [PubMed: 8790328]
12. Arnone MI, Davidson EH. Development. 1997; 124:1851–1864. [PubMed: 9169833]
13. Levine M, Davidson EH. Proc. Natl. Acad. Sci. U. S. A. 2005; 102:4936–4942. [PubMed: 15788537]
14. Moorman C, Sun LV, Wang JB, de Wit E, Talhout W, Ward LD, Greil F, Lu XJ, White KP, Bussemaker HJ, van Steensel B. Proc. Natl. Acad. Sci. U. S. A. 2006; 103:12027–12032. [PubMed: 16880385]
15. Solomon MJ, Larsen PL, Varshavsky a. Cell. 1988; 53:937–947. [PubMed: 2454748]
16. van Steensel B, Henikoff S. Nat. Biotechnol. 2000; 18:424–428. [PubMed: 10748524]
17. Iyer VR, Horak CE, Scafe CS, Botstein D, Snyder M, Brown PO. Nature. 2001; 409:533–538. [PubMed: 11206552]
18. Lieb JD, Liu XL, Botstein D, Brown PO. Nat. Genet. 2001; 28:327–334. [PubMed: 11455386]
19. Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG, Simon I, Zeitlinger J, Schreiber J, Hannett N, Kanin E, Volkert TL, Wilson CJ, Bell SP, Young RA. Science. 2000; 290:2306–2309. [PubMed: 11125145]
20. Simon I, Barnett J, Hannett N, Harbison CT, Rinaldi NJ, Volkert TL, Wyrick JJ, Zeitlinger J, Gifford DK, Jaakkola TS, Young RA. Cell. 2001; 106:697–708. [PubMed: 11572776]
21. Barski, a.; Cuddapah, S.; Cui, KR.; Roh, TY.; Schones, DE.; Wang, ZB.; Wei, G.; Chepelev, I.; Zhao, KJ. Cell. 2007; 129:823–837. [PubMed: 17512414]
22. Johnson DS, Mortazavi A, Myers RM, Wold B. Science. 2007; 316:1497–1502. [PubMed: 17540862]
23. Mikkelsen TS, Ku MC, Jaffe DB, Issac B, Lieberman E, Giannoukos G, Alvarez P, Brockman W, Kim TK, Koche RP, Lee W, Mendenhall E, O'Donovan A, Presser A, Russ C, Xie XH, Meissner

A, Wernig M, Jaenisch R, Nusbaum C, Lander ES, Bernstein BE. Nature. 2007; 448:553–552. [PubMed: 17603471]

24. Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao YJ, Zeng T, Euskirchen G, Bernier B, Varhol R, Delaney A, Thiessen N, Griffith OL, He A, Marra M, Snyder M, Jones S. Nat. Methods. 2007; 4:651–657. [PubMed: 17558387]
25. Aleksic J, Russell S. Mol. BioSyst. 2009 DOI: 10.1039/b906179g.
26. Qi Y, Rolfe A, MacIsaac KD, Gerber GK, Pokholok D, Zeitlinger J, Danford T, Dowell RD, Fraenkel E, Jaakkola TS, Young RA, Gifford DK. Nat. Biotechnol. 2006; 24:963–970. [PubMed: 16900145]
27. Reiss DJ, Facciotti MT, Baliga NS. Bioinformatics. 2008; 24:396–403. [PubMed: 18056063]
28. Ettwiller L, Paten B, Ramialison M, Birney E, Wittbrodt J. Nat. Methods. 2007; 4:563–565. [PubMed: 17589518]
29. Liu XS, Brutlag DL, Liu JS. Nat. Biotechnol. 2002; 20:835–839. [PubMed: 12101404]
30. MacIsaac KD, Gordon DB, Nekludova L, Odom DT, Schreiber J, Gifford DK, Young RA, Fraenkel E. Bioinformatics. 2006; 22:423–429. [PubMed: 16332710]
31. Das MK, Dai HK. BMC Bioinformatics. 2007; 8(Suppl 7):S21. [PubMed: 18047721]
32. Tompa M, Li N, Bailey TL, Church GM, De Moor B, Eskin E, Favorov AV, Frith MC, Fu YT, Kent WJ, Makeev VJ, Mironov AA, Noble WS, Pavesi G, Pesole G, Regnier M, Simonis N, Sinha S, Thijs G, van Helden J, Vandenbogaert M, Weng ZP, Workman C, Ye C, Zhu Z. Nat. Biotechnol. 2005; 23:137–144. [PubMed: 15637633]
33. Odom DT, Dowell RD, Jacobsen ES, Gordon W, Danford TW, MacIsaac KD, Rolfe PA, Conboy CM, Gifford DK, Fraenkel E. Nat. Genet. 2007; 39:730–732. [PubMed: 17529977]
34. Kwong C, Adryan B, Bell I, Meadows L, Russell S, Manak JR, White R. PLoS Genet. 2008; 4:e1000178. [PubMed: 18773083]
35. Orian A, Grewal SS, Knoepfler PS, Edgar BA, Parkhurst SM, Eisenman RN. Cold Spring Harbor Symp. Quant. Biol. 2005; 70:299–307. [PubMed: 16869766]
36. Nègre N, Hennetin J, Sun LV, Lavrov S, Bellis M, White KP, Cavalli G. PLoS Biol. 2006; 4:e170. [PubMed: 16613483]
37. Zhang ZDD, Paccanaro A, Fu YT, Weissman S, Weng ZP, Chang J, Snyder M, Gerstein MB. Genome Res. 2007; 17:787–797. [PubMed: 17567997]
38. Ho L, Jothi R, Ronan JL, Cui K, Zhao K. Proc. Natl. Acad. Sci. U. S. A. 2009; 106:5187–5191. [PubMed: 19279218]
39. Datta D, Zhao HY. Bioinformatics. 2008; 24:545–552. [PubMed: 17989095]
40. Haiminen N, Mannila H, Terzi E. BMC Bioinformatics. 2008; 9:336. [PubMed: 18691400]
41. Hannenhalli S, Levy S. Nucleic Acids Res. 2002; 30:4278–4284. [PubMed: 12364607]
42. Klein H, Vingron M. Genome Inform. 2007; 18:109–118. [PubMed: 18546479]
43. Feng WX, Liu YL, Wu JJ, Nephew KP, Huang THM, Li L. BMC Genomics. 2007; 9
44. Xu H, Wei CL, Lin F, Sung WK. Bioinformatics. 2008; 24:2344–2349. [PubMed: 18667444]
45. Kind J, Vaquerizas JM, Gebhardt P, Gentzel M, Luscombe NM, Bertone P, Akhtar A. Cell. 2008; 133:813–828. [PubMed: 18510926]
46. Chen X, Xu H, Yuan P, Fang F, Huss M, Vega VB, Wong E, Orlov YL, Zhang WW, Jiang JM, Loh YH, Yeo HC, Yeo ZX, Narang V, Govindarajan KR, Leong B, Shahab A, Ruan YJ, Bourque G, Sung WK, Clarke ND, Wei CL, Ng HH. Cell. 2008; 133:1106–1117. [PubMed: 18555785]
47. Cuddapah S, Jothi R, Schones DE, Roh TY, Cui KR, Zhao KJ. Genome Res. 2009; 19:24–32. [PubMed: 19056695]
48. Benjamini Y, Hochberg Y. J. R. Statist. Soc. B. 1995; 57:289–293.
49. Eisen MB, Spellman PT, Brown PO, Botstein D. Proc. Natl. Acad. Sci. U. S. A. 1998; 95:14863–14868. [PubMed: 9843981]
50. Sturn A, Quackenbush J, Trajanoski Z. Bioinformatics. 2002; 18:207–208. [PubMed: 11836235]
51. Heintzman ND, Stuart RK, Hon G, Fu YT, Ching CW, Hawkins RD, Barrera LO, Van Calcar S, Qu CX, Ching KA, Wang W, Weng ZP, Green RD, Crawford GE, Ren B. Nat. Genet. 2007; 39:311–318. [PubMed: 17277777]

52. Zeng LM, Wu J, Xie J. Stat. Appl. Genet. Mol. Biol. 2008; 7

53. Blanchette M, Bataille AR, Chen XY, Poitras C, Laganiere J, Lefebvre C, Deblois G, Giguere V, Ferretti V, Bergeron D, Coulombe B, Robert FO. Genome Res. 2006; 16:656–668. [PubMed: 16606704]

54. Elnitski L, Jin VX, Farnham PJ, Jones SJM. Genome Res. 2006; 16:1455–1464. [PubMed: 17053094]

55. Schroeder MD, Pearce M, Fak J, Fan HQ, Unnerstall U, Emberly E, Rajewsky N, Siggia ED, Gaul U. PLoS Biol. 2004; 2:e271. [PubMed: 15340490]

56. Vavouri T, Elgar G. Curr. Opin. Genet. Dev. 2005; 15:395–402. [PubMed: 15950456]

57. Tukey, JW. Exploratory Data Analysis. Addison-Wesley; Reading, MA: 1977.

Table 1

Methods for scoring overlapping and adjacent signals in two or more ChIP (or DamID) profiles. See text for details of these methods

Number of profiles under comparison	Accounting for spatial variability of events (Yes/No)	Method
Two	No	Simple counting[17,18,33,34]
		Pearson correlation coefficient[14,35-37]
		Hypothesis tests based on a single score
		Hypergeometric test[3,5,20,38]
		Chi-square test[36]
		Log-linear model[39]
		Permutation test[40-42]
	Yes	Poisson hierarchical model[43]
		Hidden Markov model[44]
		‘Standard gene’[45]
Many	Yes	Overall assessment of co-occurrence
		Permutation test[4,46,47]
		Identification of ‘co-localisation’ hotspots:
		Multiple testing based on Poisson distribution[46]
		Clustering[14,37,49-51]
		Identification of *cis*-regulatory modules
		Factor regression[52]