1656–1665 *Nucleic Acids Research, 2011, Vol. 39, No. 5* *Published online 4 November 2010*
doi:10.1093/nar/gkq848

Sensitive and accurate identification of protein–DNA binding events in ChIP-chip assays using higher order derivative analysis

Christian L. Barrett[1,*], Byung-Kwan Cho[1] and Bernhard O. Palsson[1]

[1]Department of Bioengineering, University of California, San Diego, La Jolla, CA 92093, USA

Received March 30, 2010; Revised September 1, 2010; Accepted September 9, 2010

ABSTRACT

Immuno-precipitation of protein–DNA complexes followed by microarray hybridization is a powerful and cost-effective technology for discovering protein–DNA binding events at the genome scale. It is still an unresolved challenge to comprehensively, accurately and sensitively extract binding event information from the produced data. We have developed a novel strategy composed of an information-preserving signal-smoothing procedure, higher order derivative analysis and application of the principle of maximum entropy to address this challenge. Importantly, our method does not require any input parameters to be specified by the user. Using genome-scale binding data of two *Escherichia coli* global transcription regulators for which a relatively large number of experimentally supported sites are known, we show that ~90% of known sites were resolved to within four probes, or ~88 bp. Over half of the sites were resolved to within two probes, or ~38 bp. Furthermore, we demonstrate that our strategy delivers significant quantitative and qualitative performance gains over available methods. Such accurate and sensitive binding site resolution has important consequences for accurately reconstructing transcriptional regulatory networks, for motif discovery, for furthering our understanding of local and non-local factors in protein–DNA interactions and for extending the usefulness horizon of the ChIP-chip platform.

INTRODUCTION

Protein–DNA interactions are fundamental for cellular function. Comprehensive and accurate knowledge of protein-binding locations on a chromosome is a prerequisite for understanding transcriptional regulation, resolving the role of proteins in structuring the bacterial nucleoid and eukaryotic chromatin and revealing the dynamics of protein binding or translocations. The biological significance of *in vivo* protein–DNA interactions has been remarkably enhanced by the advent of the combination of chromatin immuno-precipitation with DNA microarrays (ChIP-chip) (1). In this technological framework, the DNA in proximity to binding events is obtained by protein–DNA complex fragmentation and immuno-precipitation. Hybridization of this DNA to a tiled DNA microarray produces an enrichment signal at particular locations of the chromosome. The data from a ChIP-chip experiment is information rich in that it is a report on quasi-digital protein–DNA binding events, but these binding event signals are shrouded in an analog signal due to the fact that the DNA flanking the actual binding event is also hybridized to the microarray. Furthermore, probe-level noise inherent in the microarray platform has a significant negative impact on the signal-to-noise ratio. The challenge, then, in ChIP-chip data analysis is to identify all protein–DNA binding events and to do so with high accuracy.

A number of methods, discussed elsewhere (2), have been developed to analyze ChIP-chip data sets. Many methods only aim to identify the broad regions of enrichment and not the precise location of binding events. ChIP-chip is a high-throughput technology, and to fully leverage its capabilities requires statistical significance calculations to be included with binding event information. Few methods provide this information. Furthermore, all available methods require user-specified parameters—such as window sizes and cutoff values—that are difficult for users to optimally set. As yet, there is no available method that identifies the locations of protein–DNA binding events with high accuracy, is sensitive to weak signals and to closely spaced binding events, can associate statistical significance values to the identified binding events and learns needed parameters from each individual ChIP-chip data set instead of requiring them as user input.

*To whom correspondence should be addressed. Tel: +1 858 822 0787; Fax: +1 858 822 0022; Email: cbarrett@ucsd.edu

Higher order derivative analysis has a long history in the analytical chemical sciences (3–6), having been applied to a large number of spectroscopic techniques (7) whose principal commonality is that their output is a curved spectrum comprising a single peak or, more typically, a number of overlapping peaks. Derivate analysis of zero-order spectra is a powerful technique for identifying weak peak signals from background noise and for resolving essentially hidden peaks in a spectrum that is composed of closely spaced peaks of different magnitudes. The power of derivative analysis resides in the fact that faint changes in the slope of a signal are revealed as separate, easily identifiable peaks in the signal's higher derivatives. Herein, we report on the development of a method for applying higher order derivative analysis (i.e. employing derivatives greater than two) for the first time to ChIP-chip data for the discovery of protein–DNA binding events. We evaluate the method by applying it to ChIP-chip data sets of two global regulators in *Escherichia coli*, for which a large number of experimentally supported binding sites (ESBSs) are known, and by comparison to widely used methods. In so doing, we demonstrate an approach, called DECODE (binding event discovery using derivatives), which accurately and sensitively identifies binding site locations without the need for user-specified parameter settings and which delivers a significant quantitative and qualitative performance gain over available methods.

MATERIALS AND METHODS

Defining ESBSs

We downloaded protein–DNA binding event data in the form of a table from RegulonDB and extracted only those entries involving the proteins Lrp and Fis. We then retained only those interactions whose support for existence included 'Binding of purified protein'.

Input data preparation

The data utilized in this work were from Nimblegen arrays with 50 bp probes that are overlapped by 25 bp. There are no issues that would preclude the use of other array platforms, provided that the array data is prepared (as described below) in a similar manner. The control channel corresponds to the probe signal intensities when only genomic DNA is hybridized to the array and the experiment channel is the probe-signal intensities when the immuno-precipitated DNA is hybridized to the array. For an experimental replicate, then, the two channel signals were normalized to have the same sum of signal intensities—a correction necessary to reflect that fact that the same amount of DNA was used for each channel's hybridization. All replicates' control channel signals were then quantile normalized together, as were the experiment channel signals. Each replicate's experiment channel signal was then quantile normalized with its corresponding control channel signal, and a final enrichment signal formed from the ratio of the experiment and control channel signals. The probe ratio values were not logarithmically transformed.

Cross-replicate equalization

We equalized the baseline signal across all replicates by, for each replicate in turn, histogramming the probe ratio values (in bin sizes of 0.01) and identifying the bin value corresponding to the histogram maximum—or the average value of the background noise distribution. We then computed a replicate-specific offset by subtracting the histogram maximum bin value from 1.0. A replicate enrichment signal was then baseline corrected by adding the offset to all probe ratio values. The effect of this procedure was to make the value of 1.0 correspond to the average background noise value in all replicates, making them all directly comparable.

Potential binding region identification

Potential binding site regions were identified as those contiguous regions at least 400 bp in length wherein all probes had an enrichment value greater than 1.0. This cutoff value corresponds to a region size that is much smaller than the size of any positive signal that would be due to immuno-precipitated DNA in any ChIP-chip experiment, and so represents a very liberal criterion for identifying regions that might contain a binding event.

Enrichment signal smoothing

The enrichment signal in potential binding site regions was smoothed in a two-step manner. In step one, we removed spikes in the raw signal using an approach based on Poincaré maps (8,9). As suggested (9), we utilized Chauvenet's criterion and the median of the absolute deviations from the sample median to calculate the Universal Threshold. Because ChIP-chip signals vary over a greater range than the type of signal for which the Poincaré map procedure was originally intended, we modified the procedure to identify spikes in the enrichment signal, s, of a potential binding site region using a surrogate signal, s*. For each probe i in s, we computed the value for probe i in s* as

$$s_i^* = \frac{\mathrm{abs}(\mathrm{s}_i - \mathrm{m}_i)}{\mathrm{m}_i}$$

where

$$\mathrm{m}_i = \frac{(\mathrm{s}_{i-1} + \mathrm{s}_{i+1})}{2}$$

By normalizing each probe value in this way, we effectively removed the magnitude of the underlying signal while retaining the spike behavior—rendering all probes directly comparable. We then applied the Poincaré map procedure to s* and, additionally, computed a weight, w_i, for each probe i:

$$\mathrm{w}_i = \exp(-s_i^*)$$

A probe i was considered to be a spike if it was outside of the ellipse of the Poincaré map procedure. We used the weights, w_i, to replace the signal value for each probe i

considered to be a spike using the weighted average of it as well as its two neighboring probes:

$$s'_i = \frac{(w_{i-1} * s_{i-1} + w_i * s_i + w_{i+1} * s_{i+1})}{(w_{i-1} + w_i + w_{i+1})}$$

Finally, we computed the percent change in the sum of the values of the signals s and s′. By substituting s′ for s, the entire spike-removal procedure could be iterated, which we did until the percent change converged. In practice, convergence corresponded to a percent change of $\leq 0.1\%$.

The second step of the smoothing procedure was smoothed using the Savitzky–Golay filter (10) with a symmetric smoothing window whose of half-width was optimally computed using a modification the Durbin-Watson criterion (11). The output of the smoothing procedure was a smoothed enrichment signal *S*.

Potential binding site identification

The first three derivatives of *S* were calculated using the differential quotients derivative method—which simply computes the derivative at a point as the average of the slopes between it and its two adjacent neighboring points. Negative second derivative regions greater than five probes in length were then identified, and all positive-to-negative zero crossings of the third derivative within these regions were identified as local maxima positions. The local maxima so identified were the locations of apices of peaks that could be due to either *bona fide* binding events or to noise. We defined the set of all such apex locations as L.

Peak estimation

Peak estimation is the process of simultaneously estimating the shape and size of the peaks at all of the peak apex positions in L. We estimated these peaks using two objective functions concurrently. First, we required that the estimated peaks, when summed to form a reconstructed signal *R*, minimized the difference, *D*, between *R* and *S*. That is, we sought to estimate peaks such that, when summed, reconstructed the original enrichment signal as closely as possible. We computed *D* as

$$D = \sqrt{\sum_p (S_p - R_p)^2}$$

where *p* is the index over all probes in *S*. Second, we required that the estimated peaks maximize the entropy, *E*, over all of the probes in *S*. (The definition of *E* and how *D* and *E* were balanced are detailed below.) This second requirement is known as the principle of maximum entropy (12), and it states that the only justifiable (frequency) distribution that can be constructed from incomplete information is the one that has maximum uncertainty, subject to any constraints. As constraints, we required estimated peaks to be unimodal and symmetric.

It is necessary to explain how a binding region signal was reconstructed before describing how the entropy of *R* was calculated. The binding region probe values are real numbers. We converted the probe values to integers by multiplying them by a large integer (100) and then rounding to integers. In so doing, reconstruction of the binding region signal could be accomplished by incrementally adding or subtracting 'counts' to probes in an ongoing signal reconstruction. Since adding or subtracting counts to a probe was done in the context of estimating some peak, the counts had an explicitly attached peak label *l* from the set L. Thus the counts for a probe had a frequency distribution f_l over the different peak labels. The Shannon entropy for a probe p,

$$e_p = -\sum_{l \in L} fl * \log(fl)$$

could then be computed. The total entropy for the all probes in a region was then computed as

$$E = \sum_{\mathrm{p}} e_p$$

In regards to how the two mathematical objective functions governing the peak-estimation process were utilized simultaneously, counts were only added for a probe if *D* decreased or *D* remained unchanged and *E* increased.

The two constraints were enforced in regards to how counts could be added to estimate a peak. The first constraint was that probe values on either side of the peak maximum had to be decreasing with increasing distance from the peak maximum (to ensure a unimodal peak). The second constraint was that the count values for the symmetric probes about the probe position of the peak maximum had to be the same.

In a final step, the probe values were re-scaled by division with the same large integer used above in order to transform the probe intensity values back to real numbers.

Identifying peaks due to noise

We first identified the complementary regions to the potential binding site regions by identifying those regions that were at least 400 bp in length wherein all probe values were *less than* 1.0. (That is, we identified regions whose signal was below the average background noise level and so could confidently be assumed to not contain binding events.) We then inverted these regions about the probe value of 1.0 to create fake enrichment signals and proceeded with our algorithm to identify peak apex positions and perform peak estimation. We termed the identified peaks 'noise peaks', as they could assuredly be assumed to not be due to *bona fide* binding events.

Peak significance values

From each ChIP-chip replicate, we fit a γ distribution to the distribution of the noise peak heights. The parameters of the γ distribution were then used to calculate the *P*-value of peaks identified in the enrichment signal.

Distinguishing binding events from noise

Once all peaks have been identified in all potential binding event regions in a ChIP-chip replicate and their associated *P*-values calculated, we applied the local false discovery rate (FDR) (13) to distinguish binding event peaks from noise peaks. We used a local FDR value of 0.01.

Comparison to other methods

We utilized the following algorithms for comparative evaluation: Mpeak (14), Nimblegen's windowed threshold-detection algorithm that is a component of their SignalMap software (15), MA2C (16), Chipotle (17) and TAMALPAIS (18). All algorithms were run with their default parameters. For TAMALPAIS, we used T02P05 predictions. For the algorithms that only predicted binding event regions, we used the center of the regions as the location of their binding event predictions.

RESULTS

We present as results the major aspects of our algorithm and an evaluation of its ability to identify ESBS for the global regulators Fis and Lrp in *E. coli*.

Algorithm

ChIP-chip experiments are usually performed using multiple replicates, and it is common to average these replicates to produce on enrichment signal that is then analyzed for binding event information. We find that different replicates can reflect non-trivial differences in molecular binding activity and that averaging can abolish strong enrichment signals or indicate binding event locations that are not supported by any individual replicate. Because our method is designed for high-resolution binding site identification, we did not average replicates but instead analyzed each on its own (18). Replicates, though, still need to be directly comparable. So, after normalizing replicates, first individually and then altogether, we computed and applied a baseline correction in the form of an offset for each replicate such that an enrichment signal of 1.0 corresponded to the average value of the background noise distribution (i.e. mean value of the non-enriched probes) (19).

The most prominent characteristic of ChIP-chip data is the non-smooth variation of signal between adjacent probes (Figure 1A). Derivative analysis is at the heart of our method, but it cannot be applied directly to unsmoothed data because it would indicate a derivative change between essentially all adjacent probes and would in effect be useless. The challenge, then, in applying derivative analysis is to reveal the underlying smoothly varying enrichment signal while simultaneously minimizing the loss of binding event information contained in the raw ChIP-chip signal. To address this problem, we developed a procedure involving Poincaré maps and the Savitzky–Golay smoothing filter that transforms a raw enrichment signal into one that varies smoothly over its entire domain while retaining subtle features (Figure 1B).

Derivative analysis identifies the apex positions of the constituent peaks underlying, and together composing, a ChIP-chip signal spanning a contiguous stretch of a chromosome. We utilized the second and third derivatives to precisely locate local maxima in each replicate's smoothed enrichment signal (Figure 1C and D). As a

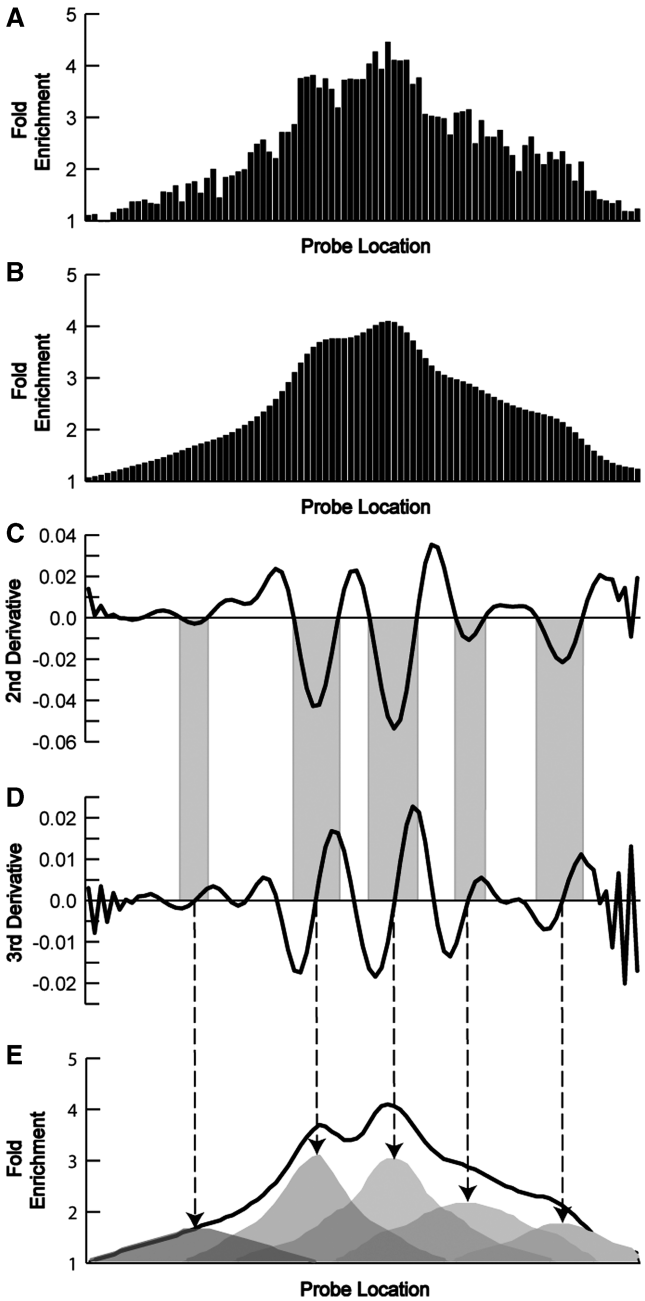

Figure 1. Identification and estimation of the protein-binding enrichment peaks underlying a ChIP-chip signal. From an unprocessed signal (**A**) that is de-noised and smoothed (**B**), the second (**C**) and third (**D**) derivative are calculated and used to identify the locations of underlying peaks. The maximum entropy principle is then applied to estimate the underlying peaks (**E**), which are due to *bona fide* protein–DNA binding events and to noise.

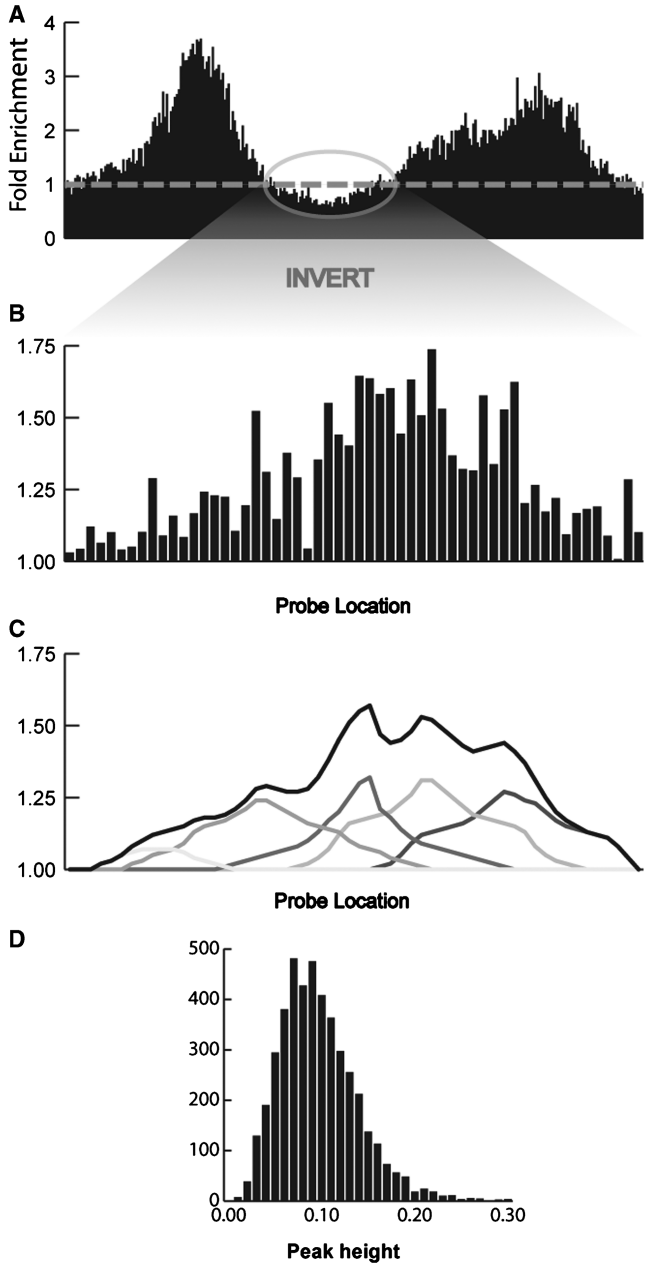

Figure 2. Learning the characteristics of noise peaks. (**A**) From a ChIP-chip signal that has been baseline corrected such that the mean background noise has a value of 1.0, regions wherein all probe values less than 1.0 are identified. (**B**) These regions are inverted to give fake 'enrichment signals', to which the peak identification procedure shown in Figure 1 is applied (**C**). (**D**) Distributions of noise peak characteristics are then computed, from which significance values can be calculated for the peaks identified in the real, non-inverted enrichment signal.

result of this strategy, local maxima positions corresponded to the apex positions of underlying peaks that were due to both *bona fide* protein-binding events and to noise. To discern between those maxima corresponding to noise and those corresponding to binding events, it was necessary to estimate the shape and size of the associated underlying peaks. We applied the principle of maximum entropy to resolve the shape and size of the underlying peaks, subject to the constraints that the peaks be unimodal and symmetric. The result of this step is the resolution of a smoothed enrichment signal into its underlying peaks (Figure 1E).

In order to quantify the probability that a peak is due to noise and not immuno-precipitated DNA, it was necessary to identify a large number of peaks that are with certainty due to noise from which noise peak statistics could be computed. Noise in this context is the background variation in signal that occurs among probes to which no immuno-precipitated DNA is complementarily bound. This noise is symmetrically distributed about the average non-enriched probe value (19), and because of how we baseline corrected each data replicate, this noise is symmetrically distributed about the enrichment signal value of 1.0. While probe values greater than 1.0 may be due in part to immuno-precipitated DNA, we can be certain that probe values less than 1.0 are due only to noise. We located regions wherein all probe values were less than 1.0 and reflected these about the value 1.0. This reflection effectively created false 'enrichment signals'. We then applied the peak-identification algorithm to these false enrichment signals and, with the identified peaks, computed null distribution statistics. This procedure is depicted in Figure 2. From these statistics we computed a *P*-value for every identified peak in a ChIP-chip replicate and used the local FDR to identify the peaks corresponding to *bona fide* binding events.

Performance evaluation of the algorithm

We have previously performed ChIP-chip analysis for the global regulators Fis (20) and Lrp (21) in *E. coli*. We used the large number of ESBS for these two DNA-binding proteins that are contained in the EcoCyc (22) and RegulonDB (23) databases to assess the sensitivity and accuracy of our method and its performance relative to other available methods.

We discriminated protein-binding events from noise using the local FDR values, which as can be seen in Figure 3A are strongly distributed toward the extreme values that the local FDR can assume (i.e. 0.0 and 1.0). These plots are for a representative single replicate, but are qualitatively very similar to the results for all replicates. The very clear split between peaks identified as being due to noise and to immuno-precipitated DNA means that the composition of the set of binding events is not very sensitive to the exact value of the local FDR cutoff value. This clear distinction implies that noise peaks have very different characteristics than *bona fide* binding event peaks. It also indicates that leveraging the symmetric nature of the background variation in non-enriched probe signals was an effective way to quantitatively discover these differentiating characteristics. The plots also indicate that peaks due to noise are much more numerous than predicted binding event peaks, underscoring the noisy nature of ChIP-chip data and the need for appropriate measures of significance.

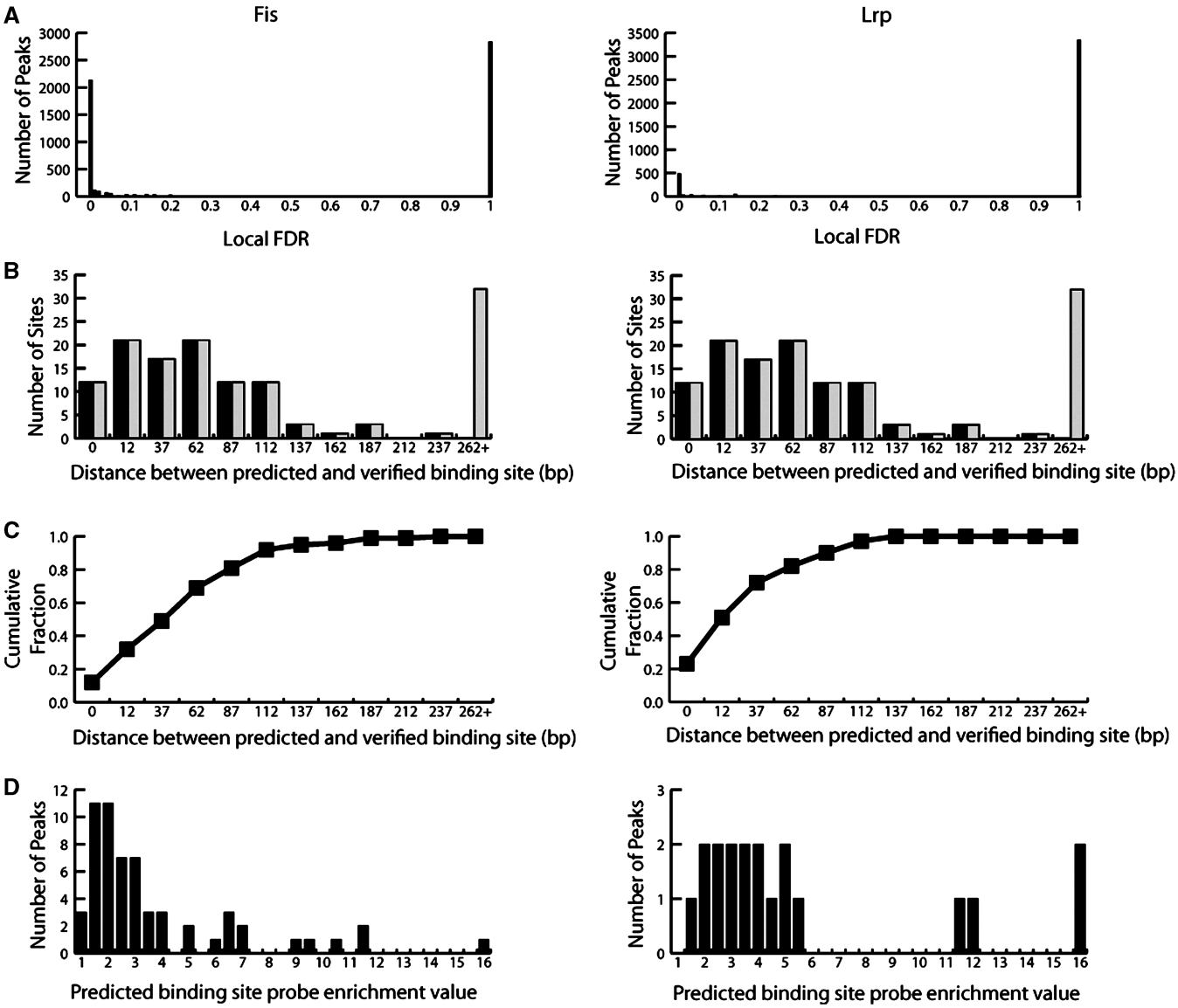

Figure 3. Evaluation of the algorithm using known binding sites. The algorithm was applied to ChIP-chip data sets of the global regulators Fis and Lrp in *E. coli*, for which a relatively large number of experimentally supported sites are known. (**A**) The local FDR shows a wide and clear gap between noise peaks and those likely due to protein–DNA binding events. (**B**) Histograms of the distance between predicted and all known binding locations (light bars) and only those known binding sites for which the ChIP-chip data showed some enrichment (dark bars). (**C**) The cumulative fraction of known sites identified as a function of the distance between known and predicted binding sites. (**D**) The enrichment value of the probe on which ESBSs are located, showing that 17% and 25% are weak signals with <2-fold enrichment over background noise.

For each ESBS data set (i.e. Fis or Lrp), we evaluated the accuracy of our method by tabulating the number of ESBS that were identified as a function of distance from the closest predicted binding site. Since we had multiple replicates, we defined the distance to the closest predicted binding site as the median distance between one of the ESBS and the closest predicted binding site in each of the replicates. The results for the Fis and Lrp data sets are shown by the light-colored bars in Figure 3B. The results show that, while a majority of the ESBS is predicted within a few probes, there were a large number of sites whose closest predicted binding event was relatively distant. We found that for these cases there was no evidence in our ChIP-chip data for the existence of these ESBS. A typical example is shown in Figure 4A, where the arrows indicate the location of four Fis ESBS in the *rnpB* promoter. (For the remainder of this study, we did not utilize these cases in any further analysis involving DECODE or other algorithms used for performance comparison.) We re-performed our accuracy assessment without cases like these, and display the results depicted by the black bars in Figure 3B. This reassessment significantly improves the accuracy for both Fis and Lrp data, especially the latter. Based on the results from these two data sets, we calculated that our method accurately identifies ~90% of the ESBS within ~88 bp. The majority of predicted sites were within ~38 bp of the ESBS. In terms of the number of probes on our tiled

array platform, these numbers correspond to four and two probes, respectively.

Sensitivity was a key performance goal in the development of our method, where sensitivity refers to the ability to identify weak and closely spaced binding events. Histograms of the median probe-enrichment values (across all replicates) on which the ESBS were located are shown in Figure 3D. The Fis histogram shows that 24% of the ESBS were on probes whose signal values were less than twice the average background noise signal. That a significant number of the ESBS with such weak signal were identified attests to our method's ability to work close to the background noise level. To demonstrate the ability to resolve closely spaced binding events, we show two examples. Note that, in each example, for clarity we show only the closest predicted binding events to the ESBS. The ESBS for Fis in the *osmE* promoter are shown in Figure 4B. While the larger peak is inaccurate by four probes from predicting the leftmost binding site, the smaller peak exactly identifies the probe of the two rightmost binding events. Nonetheless, these are difficult binding events to resolve since they occur on the shoulder of a larger enrichment signal. The example in Figure 4C is the Lrp enrichment peak at the invertible *fim* switch (24).

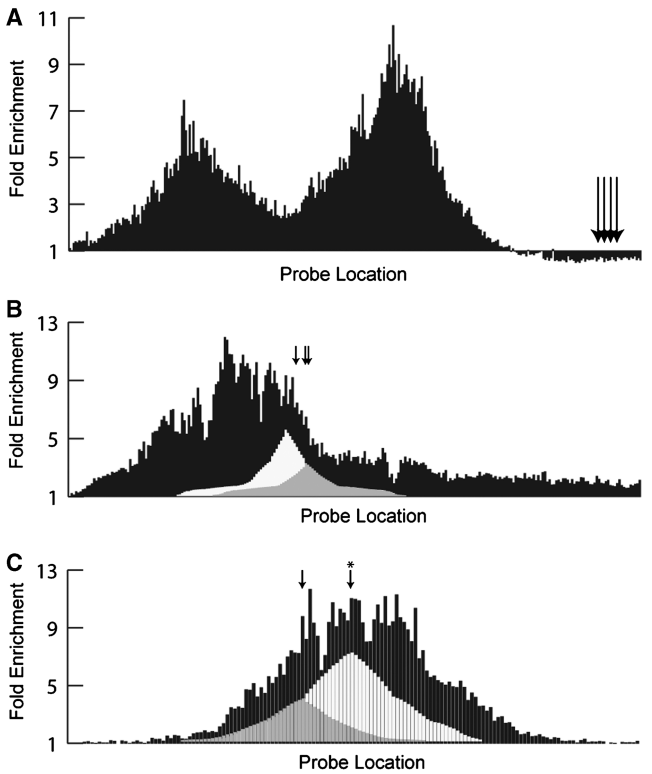

Figure 4. Predicted binding events in ChIP enrichment regions. Arrows mark the locations of ESBSs from literature. (**A**) The ChIP enrichment signal shows little to no support for the existence of some experimentally supported Fis binding sites reported in the literature (indicated by arrows, for the *rnpB* promoter). The method is able to identify closely spaced binding sites whose enrichment peaks overlap and/or occur on the shoulder regions of larger enrichment signals. Only the closest predicted peaks to the ESBSs (indicated by the arrows) are shown. (**B**) Fis binding at the *osmE* promoter and (**C**) Lrp binding in the *fim* switch. The left arrow identifies the Lrp binding site that is cataloged in RegulonDB and EcoCyc, and as can be seen the underlying predicted peak exactly locates this non-obvious binding site. The *fim* switch is a 314 bp, invertible DNA element. The invertible nature of this stretch of DNA means that it has two orientations in a population of cells, and so the Lrp binding site will actually be located in two chromosomal locations in a population. The inverted position of the Lrp binding site is marked by the starred arrow, which is also exactly located by a predicted peak.

To assess the performance of DECODE relative to other methods that are available, we performed the same analysis as above with widely used available methods. This comparative analysis, as shown in Figure 5, reveals that DECODE provides a marked improvement over all of the available methods in both accuracy and comprehensiveness. Furthermore, the performance of DECODE does not vary depending on the binding characteristics of the protein of interest—in distinction to all of the other methods. This performance difference is due to the fact that Fis binding signals are rarely isolated, unimodal and symmetric peaks [because of the Fis protein's propensity to oligimerize along the DNA into extended binding regions (25)] and that the other methods do not handle complicated enrichment signals as well as they do unimodal enrichment signals. The latter result highlights a strength of the derivative-based approach employed by DECODE.

DISCUSSION

We have developed a method to address the difficult challenge of extracting all of the information about protein–DNA interactions from a ChIP-chip data set. The difficulties in this challenge are due to the chemistry-based background hybridization noise inherent in the tiled array platform, to the incompletely fragmented DNA that flanks protein–DNA binding events that is also immunoprecipitated, and to the ambiguity of defining a binding event location due to the sometimes complicated biological interaction of proteins and DNA. These sources of difficulty together confuse the apex positions of enrichment peaks of isolated binding events, obscure weak binding event enrichment signals and mask closely spaced binding events.

Our method contains a number of novel components. Principal among these is the use of higher order derivative analysis for identifying local maxima—which correspond to underlying peaks—in a ChIP-chip signal. Second, we developed an information-preserving smoothing procedure that allowed us to apply derivative analysis to ChIP-chip signals. Third, we applied the principle of maximum entropy for discovering the underlying peaks, as opposed to deconvolution using a functional form as in most other approaches. Fourth, we leveraged the symmetric nature of the background noise to learn noise peak characteristics, allowing us to quantify the significance of the underlying peaks and to discriminate peaks due to

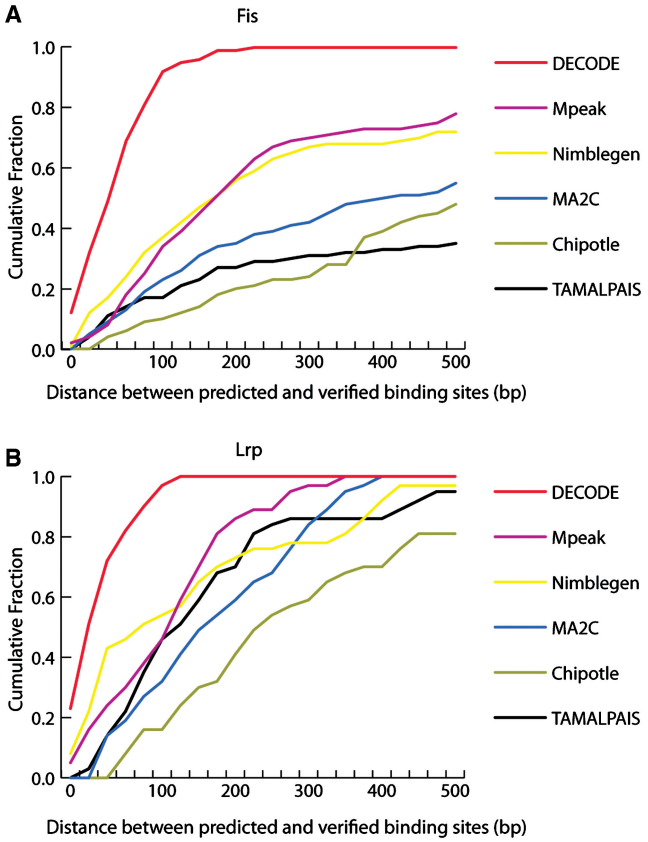

Figure 5. Comparison of DECODE to widely used available methods. The cumulative fraction of ESBSs identified as a function of the distance to predicted binding sites for ChIP-chip data sets of the *E. coli* global regulators (**A**) Fis and (**B**) Lrp.

binding events from noise peaks while controlling for false discovery rates.

This combination of novel components results in high accuracy and sensitivity for a number of reasons. Our strategy to resolve a signal into peaks—without identifying whether the peaks are due to noise or to enrichment from immuno-precipitated DNA—has two important consequences. First, it allows us to use liberal thresholds in delineating regions that 'might' contain binding events, and so with high certainty we do not miss any weak *bona fide* binding events. Second, it allows us to use significance testing to discriminate noise peaks from enrichment peaks. Furthermore, since our method is applied on a per-replicate basis, peak identification is based on the learned noise statistics of each individual experimental replicate. Consequently, parameters are optimally learned and set and are not required as input from the user. Since multiple replicates are unnecessary, our method is appropriate to use for exploratory ChIP-chip experiments.

We evaluated our method using ChIP-chip data sets of two DNA-binding proteins for which a relatively large number of ESBS are known (Figure 4). We demonstrated accuracy by showing that ~90% of the sites could be identified within four probes, and the majority could be identified within two probes. We demonstrated sensitivity by showing that 24% of the identified Fis ESBS were located on probes whose enrichment signal was <2-fold the background noise signal. We found that all of the ESBS that we did not closely predict did not have associated ChIP-chip signal enrichment to support the claim of their existence. These results demonstrate that our method was able to identify the local regions that had even very weak signals. Furthermore, they call into question a number of 'experimentally validated' binding sites that are cataloged in the literature—although the discrepancies could be due to different experimental conditions.

We also evaluated our method through a performance comparison involving widely used available methods (Figure 5). We found that DECODE is distinguished from the other methods both in its ability to accurately identify binding events and to comprehensively identify all of them. These accuracy and comprehensiveness characteristics were very similar for both of the qualitatively different ChIP-chip data sets utilized—stability not observed in the other methods. An important characteristic of a binding event discovery algorithm is that its performance does not vary with the protein under study, for such performance variance increases the uncertainty associated with all results that such an algorithm produces.

The ability to resolve binding event locations with high resolution and with associated statistical significance is important for many reasons. First, in genomic regions with a high density of genes or other sequence features, accurate localization helps disambiguate to which features the binding events are functionally related. Such accuracy is critical for accurate transcriptional regulatory network reconstruction, for instance. Second, it can dramatically improve the signal-to-noise ratio for motif discovery by identifying regions that are most likely to contain functional motifs. Third, it facilitates studies aimed at discovering principles of promoter architecture. Fourth, the statistical significance that DECODE associates with each predicted binding event (i.e. *P*-values) gives users the ability to integrate binding event predictions with other high-throughput data types (26). And fifth, the ability to localize binding events with improved accuracy and sensitivity will extend the usefulness horizon of the ChIP-chip platform, especially given that bacterial genomes can now be completely tiled with 1–5 bp resolution and that for bacterial genomes ChIP-chip still has cost and usability advantages over ChIP-seq.

A user of the DECODE software, which is freely available upon request, must bear in mind some caveats when interpreting the output. Our goal was to develop a method that could discover where binding events were occurring, with a minimum number of errors and localize the binding event more accurately than is possible with other methods. Realistically, though, one must recognize that defining a binding location entails ambiguity that will be a function of the size and overlap of the chip probes and of the nature of the interaction between the particular protein of interest and the DNA. So while some binding events can be expected to occur on a predicted probe, it would be

most appropriate to work with a narrow region around a predicted binding event location.

The application regime of DECODE encompasses both the resolution afforded by the chip tiling density and the range of genomes to which it can be applied. Given the high resolution and low cost of chip technology today, we did not design the algorithm for low-resolution arrays. Nonetheless, we demonstrated that our method is sensitive to weak enrichment signals, and so would be advantageous for discerning weak signals in low-resolution arrays. For low-resolution arrays whereon probes are widely spaced along the chromosome, the high-resolution advantages of our method are likely to be muted. Our method is not limited to bacterial genomes and would be appropriate for eukaryotic genomes since there are no genome-specific parameters in the software. The only issue that will arise in applying the method to eukaryotic genomes is the increased running time. Our method is 'embarrassingly parallel', though, so it could easily be run simultaneously on different portions of a eukaryotic Chip-chip data set.

CONCLUSIONS

In this work we have applied higher order derivative analysis to ChIP-chip data for the first time, and in so doing have extended the application regime of a powerful analytical technique. We limited our method to utilize only the third derivative, which may likely be the useful derivative limit given the signal-to-noise ratio of ChIP-chip data. Higher derivatives can be used for additional information gain, such as for resolving closely spaced binding events. Resolving more, and more closely spaced, binding events requires that the enrichment signal actually contain such discernable information, such as could be provided by chips with highly overlapping probes or by ChIP-seq (27,28). Chip-seq data is of a fundamentally different nature than ChIP-chip data, so application of our method to ChIP-seq would require changes to the raw data-processing aspect of our algorithm. Otherwise, the core elements of our method can be adapted to ChIP-seq data, and so could offer a consistent framework for maximizing the information gain from contemporary protein–DNA binding assay technologies.

FUNDING

National Institutes of Health Grant GM062791; the Office of Science-Biological and Environmental Research, U.S. Department of Energy, cooperative agreement DE-FC02-02ER63446; National Health Research Institutes-Taiwan Grant UCSD2008-3183R; National Institutes of Health Grant DE-AC05-76RL01830. Funding for open access charge: National Institutes of Health Grant GM062791.

Conflict of interest statement. None declared.

REFERENCES

1. Ren,B., Robert,F., Wyrick,J.J., Aparicio,O., Jennings,E.G., Simon,I., Zeitlinger,J., Schreiber,J., Hannett,N., Kanin,E. *et al.* (2000) Genome-wide location and function of DNA binding proteins. *Science*, **290**, 2306–2309.
2. Johnson,D.S., Li,W., Gordon,D.B., Bhattacharjee,A., Curry,B., Ghosh,J., Brizuela,L., Carroll,J.S., Brown,M., Flicek,P. *et al.* (2008) Systematic evaluation of variability in ChIP-chip experiments using predefined DNA targets. *Genome Res.*, **18(Suppl. 3)**, 393–403.
3. Hammond,V.J. and Price,W.C. (1953) Derivative spectroscopy: theoretical aspects. *J. Opt. Soc. Am.*, **43**, 924.
4. Tannenbauer,E., Merkel,P.B. and Hamill,W.H. (1953) *J. Phys. Chem.*, **21**, 311.
5. Morrison,J.D. (1953) Studies of ionization efficiency. Part III. The detection and interpretation of fine structure. *J. Chem. Phys.*, **21**, 1767–1772.
6. Giese,A.T. and French,C.S. (1955) The analysis of overlapping spectral absorption bands by derivative spectrophotometry. *Applied Spectroscopy*, **9**, 78–96.
7. Talsky,G. (1994) *Derivative Spectrophotometry: Low and High Order*. Weinheim, New York, VCH.
8. Goring,D.G. and Nikora,V.I. (2002) Despiking Acoustic Doppler Velocimeter Data. *J. Hydrolic Engineering*, **128(Suppl. 1)**, 117–126.
9. Wahl,T.L. (2003) Discussion of ''Despiking Acoustic Doppler Velocimeter Data'' by Derek G. Goring adn Vladimir I. Nikora. *H Hydraulic Engineering*, **129(Suppl. 6)**, 484–487.
10. Savitzky,A. and Golay,M.J.E. (1964) Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry 1964*, **36(Suppl. 8)**, 1627–1639.
11. Vivo-Truyols,G., Torres-Lapasio,J.R., van Nederkassel,A.M., Vander Heyden,Y. and Massart,D.L. (2005) Automatic program for peak detection and deconvolution of multi-overlapped chromatographic signals part I: peak detection. *J. chromatography*, **1096**, 133–145.
12. Jaynes,E.T. (1957) Information theory and statistical mechanics. *Physical Rev.*, **106(Suppl. 4)**, 620–630.
13. Efron,B., Tibshirani,R., Stock,J.D. and Tusher,V. (2001) Empirical Bayes analysis of a microarray experiment. *J. Am. Stat. Assoc.*, **96**, 1151–1160.
14. Kim,T.H., Barrera,L.O., Zheng,M., Qu,C., Singer,M.A., Richmond,T.A., Wu,Y., Green,R.D. and Ren,B. (2005) A high-resolution map of active promoters in the human genome. *Nature*, **436**, 876–880.
15. SignalMap. http://www.nimblegen.com (April 2010, date last accessed).
16. Song,J.S., Johnson,W.E., Zhu,X., Zhang,X., Li,W., Manrai,A.K., Liu,J.S., Chen,R. and Liu,X.S. (2007) Model-based analysis of two-color arrays (MA2C). *Genome Biol.*, **8(Suppl. 8)**, R178.
17. Buck,M.J., Nobel,A.B. and Lieb,J.D. (2005) ChIPOTle: a user-friendly tool for the analysis of ChIP-chip data. *Genome Biol.*, **6(Suppl. 11)**, R97.
18. Bieda,M., Xu,X., Singer,M.A., Green,R. and Farnham,P.J. (2006) Unbiased location analysis of E2F1-binding sites suggests a widespread role for E2F1 in the human genome. *Genome Res.*, **16(Suppl. 5)**, 595–605.
19. Margolin,A.A., Palomero,T., Sumazin,P., Califano,A., Ferrando,A.A. and Stolovitzky,G. (2009) ChIP-on-chip significance analysis reveals large-scale binding and regulation by human transcription factor oncogenes. *Proc. Natl Acad. Sci. USA*, **106(Suppl. 1)**, 244–249.
20. Cho,B.K., Knight,E.M., Barrett,C.L. and Palsson,B.Ø. (2008) Genome-wide analysis of Fis binding in *Escherichia coli* indicates a causative role for A-/AT-tracts. *Genome Res.*, **18(Suppl. 6)**, 900–910.
21. Cho,B.K., Barrett,C.L., Knight,E.M., Park,Y.S. and Palsson,B.O. (2008) Genome-scale reconstruction of the Lrp regulatory network in *Escherichia coli*. *Proc. Natl Acad. Sci. USA*, **105**, 19462–19467.
22. Keseler,I.M., Collado-Vides,J., Gama-Castro,S., Ingraham,J., Paley,S., Paulsen,I.T., Peralta-Gil,M. and Karp,P.D. (2005) EcoCyc: a comprehensive database resource for *Escherichia coli*. *Nucleic Acids Res.*, **33(Database issue)**, D334–D337.

23. Salgado,H., Gama-Castro,S., Peralta-Gil,M., Diaz-Peredo,E., Sanchez-Solano,F., Santos-Zavaleta,A., Martinez-Flores,I., Jimenez-Jacinto,V., Bonavides-Martinez,C., Segura-Salazar,J. *et al.* (2006) RegulonDB (version 5.0): Escherichia coli K-12 transcriptional regulatory network, operon organization, and growth conditions. *Nucleic Acids Res.*, **34(Database issue)**, D394–D397.
24. Kelly,A., Conway,C., Croinin,T.O., Smith,S.G. and Dorman,C.J. (2006) DNA supercoiling and the Lrp protein determine the directionality of fim switch DNA inversion in Escherichia coli K-12. *J. Bacteriol.*, **188**, 5356–5363.
25. Schneider,R., Lurz,R., Luder,G., Tolksdorf,C., Travers,A. and Muskhelishvili,G. (2001) An architectural role of the Escherichia coli chromatin protein FIS in organising DNA. *Nucleic Acids Res.*, **29**, 5107–5114.
26. Hwang,D., Rust,A.G., Ramsey,S., Smith,J.J., Leslie,D.M., Weston,A.D., de Atauri,P., Aitchison,J.D., Hood,L., Siegel,A.F. *et al.* (2005) A data integration methodology for systems biology. *Proc. Natl Acad. Sci. USA*, **102**, 17296–17301.
27. Johnson,D.S., Mortazavi,A., Myers,R.M. and Wold,B. (2007) Genome-wide mapping of *in vivo* protein–DNA interactions. *Science*, **316**, 1497–1502.
28. Robertson,G., Hirst,M., Bainbridge,M., Bilenky,M., Zhao,Y., Zeng,T., Euskirchen,G., Bernier,B., Varhol,R., Delaney,A. *et al.* (2007) Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat. Methods*, **4**, 651–657.