Published online 13 September 2017 *Nucleic Acids Research, 2017, Vol. 45, No. 21* ***e173***
doi: 10.1093/nar/gkx799

Ritornello: high fidelity control-free chromatin immunoprecipitation peak calling

Kelly P. Stanton[1,2,†], Jiaqi Jin[3,†], Roy R. Lederman[4,5], Sherman M. Weissman[3] and Yuval Kluger[1,2,4,*]

[1]Department of Pathology, Yale University School of Medicine, 333 Cedar Street, New Haven, CT 06520, USA, [2]Interdepartmental Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT 06511, USA, [3]Department of Genetics, Yale University School of Medicine, 333 Cedar Street, New Haven, CT 06520, USA, [4]Program of Applied Mathematics, Yale University, 51 Prospect Street, New Haven, CT 06511, USA and [5]Department of Mathematics and PACM, Princeton University, Fine Hall, Washington Road, Princeton, NJ 08544-1000, USA

Received December 23, 2015; Revised August 07, 2017; Editorial Decision August 26, 2017; Accepted August 30, 2017

ABSTRACT

With the advent of next generation high-throughput DNA sequencing technologies, omics experiments have become the mainstay for studying diverse biological effects on a genome wide scale. Chromatin immunoprecipitation (ChIP-seq) is the omics technique that enables genome wide localization of transcription factor (TF) binding or epigenetic modification events. Since the inception of ChIP-seq in 2007, many methods have been developed to infer ChIP-target binding loci from the resultant reads after mapping them to a reference genome. However, interpreting these data has proven challenging, and as such these algorithms have several shortcomings, including susceptibility to false positives due to artifactual peaks, poor localization of binding sites and the requirement for a total DNA input control which increases the cost of performing these experiments. We present Ritornello, a new approach for finding TF-binding sites in ChIP-seq, with roots in digital signal processing that addresses all of these problems. We show that Ritornello generally performs equally or better than the peak callers tested and recommended by the ENCODE consortium, but in contrast, Ritornello does not require a matched total DNA input control to avoid false positives, effectively decreasing the sequencing cost to perform ChIP-seq. Ritornello is freely available at https://github.com/KlugerLab/Ritornello.

INTRODUCTION

Reliable and precise characterization of where proteins, such as transcription factors (TFs), interact with the genome, enables biologists to understand how gene expression is regulated at the molecular level. The human genome, for example, encodes about 1500 TFs (1) and many of them directly recognize and bind to specific DNA sequences to regulate gene expression. Therefore, identification of where each TF binds to the DNA is critical for reconstructing the complex regulatory network of gene expression. Chromatin immunoprecipitation (ChIP) followed by high-throughput sequencing (ChIP-seq) is a powerful tool for detecting protein–DNA interactions at the genome-wide scale and has become the method of choice. In a ChIP-seq experiment, first, proteins interacting with the DNA are chemically attached to the DNA using formaldehyde-mediated crosslinking. Then the DNA is fragmented into short pieces and antibodies specifically targeting the protein of interest are used to pull down DNA fragments bound by that protein. Finally, the immunoprecipitated DNA fragments are released from the protein of interest and subjected to high-throughput DNA sequencing. The resulting sequenced reads are mapped to a reference genome and computational peak calling algorithms are applied to process mapped reads and infer protein-binding positions.

TFs usually bind to short specific DNA sequences (motifs) and generate sharp point-source peaks (2). For most ChIP-seq experiments currently available, only one of the two 5′ ends of each double-stranded DNA fragment has been sequenced (single end sequencing), so the read coverage near the point-source peaks follow a characteristic bimodal shape. However, calling peaks accurately from a large quantity of mapped reads is nontrivial and over 40 algorithms have been developed (3–44) since the ChIP-seq technology was first introduced (45). Peak calling remains challenging due to the presence of artifactual-binding events (false positives) and background noise from reads outside of peaks, multi-binding events with overlapping read contributions and variability of experimental quality. Additionally,

*To whom correspondence should be addressed. Tel: +1 203 737 6262; Fax: +1 203 785 6486; Email: yuval.kluger@yale.edu
†These authors contributed equally to the paper as first authors.

for most peak calling algorithms, matched negative controls, which are usually DNA samples obtained without performing immunoprecipitation (total DNA input control) or immunoprecipitated by non-specific antibodies (IgG control), are often required to control the false positive rate (FPR).

Performing a negative control experiment for each sample, effectively doubles the sequencing cost of ChIP-seq, limiting the number of samples that can be run per experiment. Peak calling algorithms that do not use the control (including those that have the option to run with or without it) have been developed; however, they underperform due to the lack of a detailed characterization of ChIP-seq signal and noise.

Binding events can also occur in close proximity to one another and it is often difficult to resolve how many binding sites are present and precisely where binding occurs. BRACIL (46) and CSDeconv (47) use blind deconvolution algorithms to resolve individual peaks at multi-binding loci but are not scalable for peak calling and are thus used for post processing when peaks have been identified by other peak callers. GEM (23) incorporates *de novo* motif discovery into the peak identification process aiding in resolving individual peaks, but may not be suitable if the TF of interest does not bind to DNA directly or does not have any specific motif.

ChIP-seq experiments can also be of varying quality. Collective efforts by large consortia have provided guidelines on how to evaluate the quality and signal-to-noise ratio of ChIP-seq experiments. The opposing strand cross-correlation between the read coverage on the positive and that on the negative strands has been used to assess experimental quality by ENCODE (2). The cross-correlations of ChIP-seq as well as input control experiments exhibit two modes, one at or around their respective average fragment lengths and an additional one at or around their respective read lengths. High quality experiments tend to have a greater contribution from the fragment length mode, while low quality experiments and input controls tend to have a larger contribution from the read length mode. Specifically, to assess the quality of ChIP-seq experiments, ENCODE recommends two metrics, the normalized strand coefficient (NSC) and the relative strand correlation (RSC)(2). The NSC is the ratio between the the fragment length mode and the baseline for large offsets of cross-correlation. The RSC is the ratio between the fragment length mode and the read length mode of the cross-correlation. If the NSC or RSC scores are low, indicating poor experimental quality, ENCODE recommends repeating the experiment. Given the considerable cost of repeating a ChIP-seq experiment, it is useful to be able to 'rescue' samples with suboptimal quality for use as additional replicates, rather than discarding them.

Here, we present Ritornello, a novel algorithm for finding binding sites of TFs. Ritornello is based on both digital signal processing (DSP) and statistical techniques. In the current work, we contribute the following innovations and insights:

i) a peak caller, that does not require a matched control and still maintains a low FPR, outperforming even algorithms that use the control.
ii) an efficient method to perform full deconvolution of multi-binding events on a genome wide scale.
iii) samples of low quality can be 'rescued', instead of being discarded.
iv) a rigorous characterization of the binding signals and artifacts in the presence of noise in ChIP-seq data.
v) a non-parametric approach to calculate the fragment length distribution (FLD) for any single-end NGS experiment.

We benchmarked Ritornello against MACS2 (3) and GEM (23), two algorithms recommended by the ENCODE consortium (2). In the default modes each requires the matched control. We demonstrated that Ritornello, a matched control free method, outperformed MACS2 and GEM.

MATERIALS AND METHODS

We have developed Ritornello to find candidate peaks efficiently, with minimal use of memory and computation time, by using a DSP technique called a matched filter, classify candidate peaks as true-binding events or artifacts based on their shape and finally test candidate-binding positions for significance based on comparison to a model absent of binding at that position. The scheme of the Ritornello method is detailed in Figure 1.

Derive fragment length distribution via deconvolution

For fragments overlapping a binding position, the positive strand mapping reads will be upstream of the binding site whereas negative strand reads will be downstream of the site. As such, the distance between a read and the binding position is dependent on the fragment length, which most peak calling algorithms must estimate to obtain accurate predictions for binding locations. Specifically, most current peak callers only estimate the average fragment length rather than the whole distribution. Our first innovation for Ritornello is calculating, not just the mean fragment length, but the entire sample specific empirical FLD from single-end reads (step 2 of Figure 1). Ritornello utilizes this FLD, as a key component, for more accurate peak predictions, which we will describe in detail below.

The opposing strand 'cross-correlation' $\Psi_n = \Pr[R^n] * \Pr[-R^p]$ is the single strand 'autocorrelation' $\Phi_p = \Pr[R^p] * \Pr[-R^p]$ convolved against the FLD as has previously been shown (48):

$$\begin{aligned}\Psi_n &= \Pr[R^n] * \Pr[-R^p] \\ &= \Pr[R^p + F] * \Pr[-R^p] \\ &= \Pr[R^p] * \Pr[-R^p] * \Pr[F] \\ &= \Phi_p * \Pr[F],\end{aligned} \tag{1}$$

where, $\Pr[R^p]$ is the probability of choosing a read starting at a position R^p on the positive strand, $\Pr[R^n]$ is the probability of choosing a read start at a position R^n on the neg-

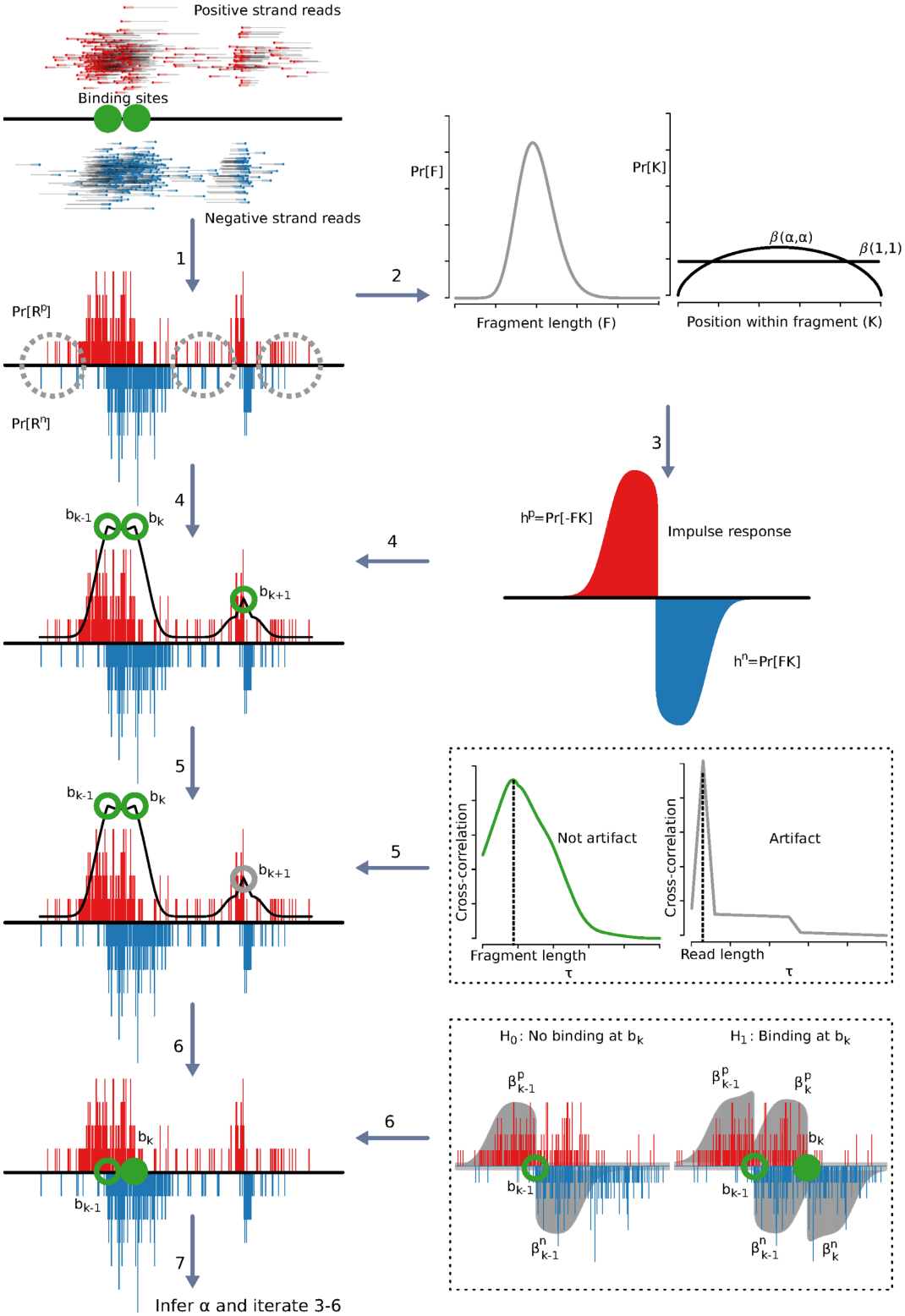

Figure 1. Overview of the Ritornello approach. Step 1: the reads are mapped to the genome and the distributions of read starts Pr[R^p] and Pr[R^n] are calculated. Step 2: the FLD Pr[F] is calculated using only those areas identified as background coverage (free of peaks and artifacts). Step 3: the expected distribution of reads around a binding event is calculated from a model including the FLD Pr[F] as well as the distribution of relative-binding positions within fragments Pr[K]. Initially, K is modeled with a uniform distribution or equivalently $K \sim \beta(\alpha, \alpha)$ with $\alpha = 1$. Step 4: the expected distribution of reads around a binding event is used to locate candidate-binding event peaks (e.g. b_{k-1}, b_k and b_{k+1}) by identifying the positions with the highest match with impulse response function. Step 5: candidate-binding events are classified as either read length artifacts (e.g. b_{k+1}) or are retained as putative-binding events (e.g. b_{k-1} and b_k) based on the shape of the local cross-correlation between opposing strands. Step 6: the binding intensity, $\beta_k^p + \beta_k^n$, of each peak, b_k are deconvolved from the mixture of local peaks b_{k-1} and background noise using a maximum likelihood approach. The likelihood ratio test

ative strand, Pr[F] is the FLD and * is the convolution operator. The FLD can then be obtained by deconvolution as follows:

$$\Pr[F] = \mathcal{F}^{-1}\left(\frac{\mathcal{F}(\Psi_n)}{\mathcal{F}(\Phi_p)}\right), \tag{2}$$

where, $\mathcal{F}$ is the Fourier transform operator and $\mathcal{F}^{-1}$ is its inverse.

In the current work, we clarify that Equation (1) assumes that R^p and F are statistically independent and thus Pr[$R^p + F$] can be simplified to Pr[R^p] * Pr[F]. If we assume that R^n (as opposed to R^p) and F are independent, then we can instead write the following relationship:

$$\begin{aligned}\Psi_p &= \Pr[R^p] * \Pr[-R^n] \\ &= \Pr[R^n + F] * \Pr[-R^n] \\ &= \Pr[R^n] * \Pr[-R^n] * \Pr[F] \\ &= \Phi_n * \Pr[F],\end{aligned} \tag{3}$$

and deconvolve as follows:

$$\Pr[F] = \mathcal{F}^{-1}\left(\frac{\mathcal{F}(\Psi_p)}{\mathcal{F}(\Phi_n)}\right). \tag{4}$$

Equations (2) and (4) describe how to obtain the FLD under two different and mutually exclusive assumptions (i.e. $R^p \perp F$ or $R^n \perp F$ where $\perp$ denotes statistical independence). To estimate the FLD empirically, Ritornello must locate genomic positions where either R^p or R^n is roughly independent of F and invoke the corresponding equations locally.

We note that the local coverage on either R^p or R^n is unlikely to be uniform when the FLD changes as a function of genomic position on that strand. Therefore, to estimate the fragment lenght distribution, we exclude regions that clearly do not satisfy the assumptions made in Equations (1) and (3) by retaining only regions whose coverage is roughly locally uniform for R^p in Equation (1) and R^n in Equation (3).

We identify these candidate regions by looking for read coverage that is locally uniform on either strand, using a χ^2 goodness of fit test (i.e. Pr[R^p] $\sim \mathcal{U}$ or Pr[R^n] $\sim \mathcal{U}$). For each window of size $2F_{\max}$ (twice the maximum fragment length) centered at position i on either strand, we calculate the χ^2 test statistic as follows:

$$\begin{aligned}z_i^p &= \sum_{j=i-F_{max}}^{i+F_{max}} \frac{(\Pr[R^p = j] - \mathcal{U}[j])^2}{\mathcal{U}[j]} \\ z_i^n &= \sum_{j=i-F_{max}}^{i+F_{max}} \frac{(\Pr[R^n = j] - \mathcal{U}[j])^2}{\mathcal{U}[j]}.\end{aligned} \tag{5}$$

We then sum the local empirical autocorrelations, $\Phi_{p,i}$ or $\Phi_{n,i}$, for those windows where either the positive or negative strand is independent of the fragment length, as determined by the χ^2 test for local uniformity, across the genome G. Additionally, we sum the local opposing strand cross-correlations, $\Psi_{p,i}$ or $\Psi_{n,i}$, associated with each autocorrelation according to Equations (1) and (3) as follows:

$$\begin{aligned}\Phi(\tau) &= \sum_{i=1}^{G} \Phi_{p,i}(\tau) I(V_i^p > \alpha) + \Phi_{n,i}(\tau) I(V_i^n > \alpha) \\ \Psi(\tau) &= \sum_{i=1}^{G} \Psi_{n,i}(\tau) I(V_i^p > \alpha) + \Psi_{p,i}(\tau) I(V_i^n > \alpha) \\ V_i^p &\equiv \Pr[z_i^p < \chi^2(2F_{max})] \\ V_i^n &\equiv \Pr[z_i^n < \chi^2(2F_{max})].\end{aligned} \tag{6}$$

where,

$$\begin{aligned}\Phi_{p,i}(\tau) &= \Pr[R^p = i] \Pr[R^p = i - \tau] \\ \Phi_{n,i}(\tau) &= \Pr[R^n = i] \Pr[R^n = i - \tau] \\ \Psi_{p,i}(\tau) &= \Pr[R^p = i] \Pr[R^n = i - \tau] \\ \Psi_{n,i}(\tau) &= \Pr[R^n = i] \Pr[R^p = i - \tau] \\ \tau &\in [-F_{max}, F_{max}].\end{aligned} \tag{7}$$

Using the global autocorrelation, Φ and cross-correlation, Ψ, functions from Equation (6), we estimate the FLD as follows:

$$\Pr[F] = \mathcal{F}^{-1}\left(\frac{\mathcal{F}(\Psi)}{\mathcal{F}(\Phi)}\right). \tag{8}$$

Why the fragment length distribution can be inferred from single end data?

We have just shown that the fragment length distribution can be derived via deconvolution from Pr[R^P] and Pr[R^n] when those distributions are related by Equations (1) or (3). This is intuitive in paired-end sequencing. However, in single-end sequencing, only one end is randomly selected from each fragment, and it is hard to imagine how the fragment length information can be inferred. For a given set of fragments, single end sequenced reads are a sub-sample of paired-end sequenced reads. Thus, Pr[R^P] and Pr[R^n], the only inputs to Equations (2) or (4), should be the same for the single-end reads as that in the paired-end, resulting in similar estimated FLDs, Pr[F].

To directly demonstrate that the computed Pr[F] of singled-end data is a reasonable approximation for the FLD, we have applied Equations (6) and (7) to a pseudo single-end data created by randomly sampling one end from each fragment of a paired-end experiment. In Figure 2, we show that the FLD calculated using Equations (6) and (7) from this pseudo single-end data (black) is similar to the true FLD of the paired-end sample (green).

←

is then applied to determine the significance of the peak. Step 7: the distribution of relative-binding positions within fragments, Pr[K], is updated using a maximum likelihood estimate of α, where $K \sim \beta(\alpha, \alpha)$. This is obtained using a combined likelihood model for the top 200 most significant peaks. Steps 3–6 are repeated using the new Pr[K].

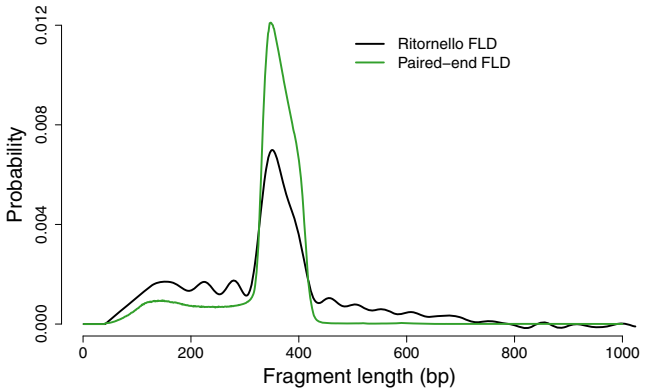

Figure 2. Ritornello captures the FLD from single-end sequencing data. Given a set of fragments, singled end sequenced reads are a subsample of paired-end sequenced reads. When enough single end fragments are sampled the distribution of read coverage on the positive and negative strands $\Pr[R^p]$ and $\Pr[R^n]$ are equivalent to their paired-end counterparts. A single read from each paired-end fragment of an EZH2 sample was randomly chosen to simulate singled end data. The FLD calculated by Ritornello from this data (black curve) closely approximates the true FLD (green curve).

Local fragment length distribution varies around binding events

The fragments generated from a binding event overlap the binding site, thus any given fragment must be at least as long as the distance from its start position to the binding site. This creates dependence between the fragment length and genomic position on both strands because reads that are further from the binding site are necessarily longer on aggregate. Consequently, neither the positive nor the negative strand read coverage is independent of the fragment length within a binding event, making it inappropriate to recover the FLD using Equation (2), as shown by simulation in Figure 3D. In contrast, in simulated event free regions, it is shown that the FLD can be correctly recovered using Equation (2) as seen in Figure 3A.

Local fragment length distribution varies around read length artifacts

In addition to binding events, we have observed local artifactual patterns that create dependence between fragment length and genomic position on both strands, preventing the reconstruction of FLD. These artifactual patterns fall into the following two categories:

i) a pile of reads whose start positions constitute a read length width column pattern on the positive strand, followed by a read length width column pattern on the negative strand, a read length downstream. We refer to this artifact as a column artifact. We simulate it in Figure 3B and show it in Figure 4A.
ii) a binding peak (or background read coverage) but with a read length width column pattern of missing reads on the positive strand followed by a read length width column pattern of missing reads on the negative strand, a read length downstream. We refer to this as a missing-column artifact. We simulate it in Figure 3C and show it in Figure 4B.

Based on paired end data, we have observed that the FLD is invariant with respect to genomic position with the exception of these artifacts and binding events which Ritornello excludes from the FLD calculation using Equations (6) and (7). Both of these artifacts cause local disturbances in the FLD. This is easier to see in paired-end sequenced reads. The paired-end column artifact shown in Figure 4C contains fragments with length distributed according to the sample's FLD. However, the fragments are organized such that longer fragments extend further from center of the artifact (denoted with an asterisk) than shorter fragments, implying F is dependent on R^p and R^n. The paired end missing-column artifact shown in Figure 4D is composed of fragments organized such that the range of possible fragment lengths is restricted based on genomic position. Specifically, on the positive strand, lengths of fragments (such as fragments A and B) must lie outside the range $[d^p, d^p + w]$ where d^p is the distance from the positive strand read to the center of the missing-column artifact (denoted with an asterisk). Likewise, on the negative strand, the lengths of fragments (such as fragments C and D) must lie outside the range $[d^n, d^n + w]$ where d^n is the distance from the negative strand read to the center of the missing-column artifact (denoted with an asterisk). Thus, column and missing-column artifacts are composed of fragments organized such that both strands are dependent on fragment length.

We note that if we were using Equations (2) or (4) without invoking Equations (5)–(8) at these artifactual regions, the resulting FLD would have two modes, one associated with the 'phantom peak' near the read length, and one associated with the predominant fragment length. ENCODE observed these two modes in the opposing strand cross-correlation, which is related to the FLD (Equations (1) and (3)). However, when these artifactual regions are filtered out as is done in equations (5)–(8), the 'phantom peak' is greatly attenuated. Thus, these artifactual areas give rise to a low quality RSC (2,49), the ENCODE measures of the 'phantom peak' using cross-correlation. The cause of these artifacts is likely incorrect mapping which is described in the next section.

Incorrect mapping leads to read length artifacts

The read length used in any sequencing experiment is determined subsequent to the collection of fragments. Therefore, read length artifacts are associated with sequencing or post-sequencing procedures. Each read in a ChIP-seq experiment is sequenced from the sample's genomic DNA, which can differ from the reference genome used in the alignment step. Comparative genome assembly algorithms used for sequence alignment work by comparing the nucleotide sequence for each read to the sequence of the reference genome and assigning the read to the location that gave the best alignment score, usually based on fuzzy string matching. If the nucleotide sequence of the sample's genome, from which the reads are sampled, disagrees with the reference genome at location x, the mapping algorithm can fail to assign the appropriate reads to that location. This could occur if the number of mismatched bases (or indels) per read ex-

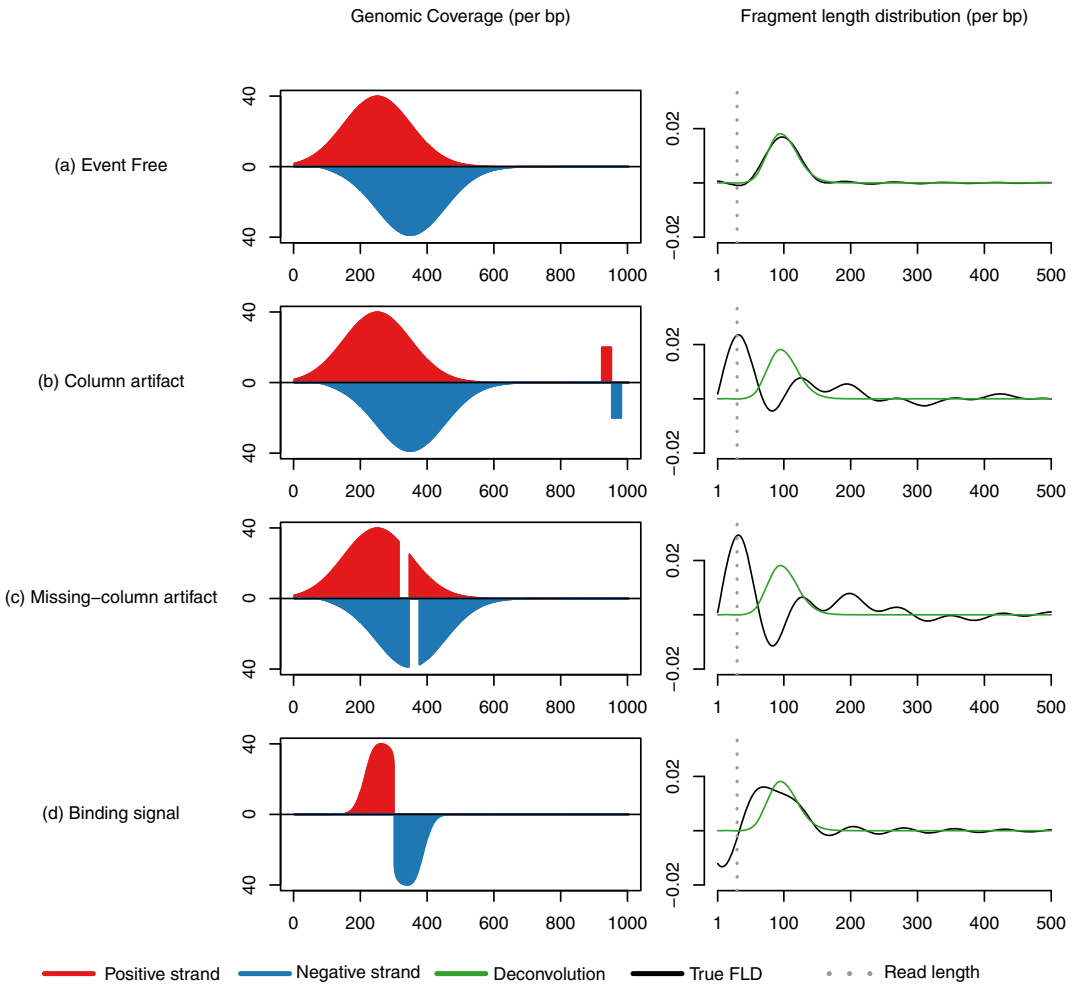

Figure 3. The presence of binding events and read length artifacts hinders reconstruction of the FLD by deconvolution. Coverage patterns (positive strand in red and negative strand in blue) were generated from reads sampled randomly from either end of simulated fragments. Equation (2) was applied to infer the FLD, FLD (black). The actual FLD was calculated from the simulated fragments (green). The read length is shown in gray. (**A**) The FLD, inferred from reads in simulated binding regions, deviates from the true FLD. (**B**) The FLD, inferred from genomic background coverage outside of binding events, agrees with the true FLD. (**C**) In the presence of read length column artifacts, the inferred FLD deviates from the true FLD and exhibits a 'phantom peak' at read length. (**D**) In the presence of read length missing-column artifacts, the inferred FLD deviates from the true FLD and exhibits a 'phantom peak' at read length.

ceeds a predetermined cutoff or simply if the reads belonging to that location map better to another region with higher sequence similarity. When this occurs there will be a discontinuity in coverage across w nucleotides (where w is the read length) because there are exactly w possible read start positions where the read would overlap the mismatched base (or indel). On the positive strand the discontinuity is upstream of x and on the negative strand the discontinuity is downstream of x. This results in an missing-column artifactual coverage pattern of incorrectly mapped reads to the upstream sequence highlighted in yellow as seen in Figure 5. Further, reads that fail to map to the correct location can instead map to another location with higher sequence similarity as determined by the mapping algorithm. This would result in a column artifact as seen in the downstream sequence highlighted in yellow in Figure 5. These artifacts tend to occur in interspersed repetitive regions such as the sequences shown in yellow in Figure 5.

For paired-end each relocated read in a column artifact, the associated read from the same fragment is also relocated as shown in Figure 4C. Likewise for each missing read in a missing-column artifact, the associated read from the same fragment is also missing as shown in Figure 4D. In principle, the missing column artifact problem may be mitigated by suggesting multiple possible mappings (which can be done using the bowtie −k option or efficiently using other algorithms (50)). However, while this may fill in some of the missing columns, it will create additional column artifacts, which would need to be resolved.

Derive the expected read coverage distribution around binding events

Once we estimate the FLD, we use it to derive a filter matched to the read coverage pattern characteristic to regions of true-binding events (step 3 of Figure 1). We denote putative ChIP target-binding sites by B_j, where j is an in-

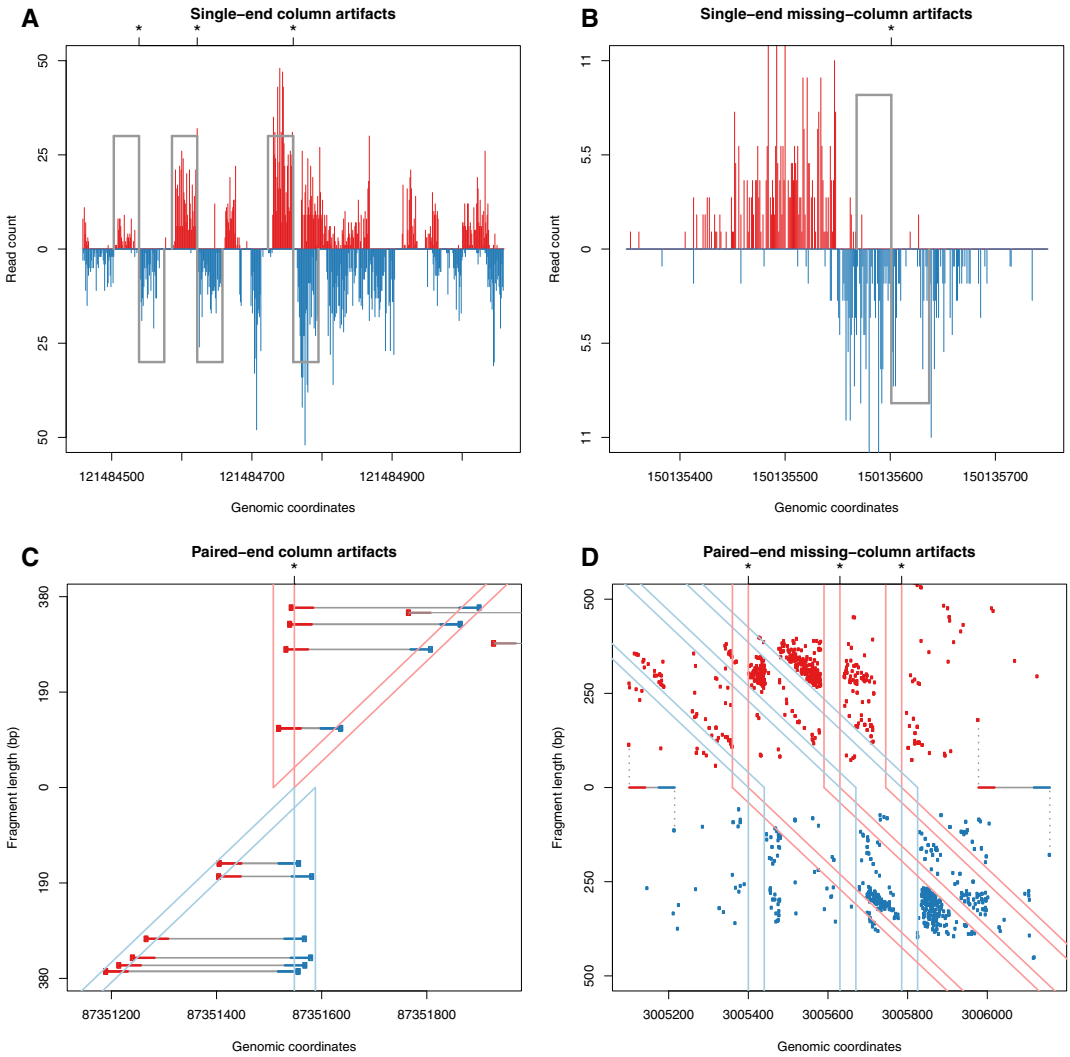

Figure 4. Column and missing-column artifacts and the nonrandom FLD in their neighborhoods. Examples of read length column (**A**) and missing-column (**B**) artifacts in a single-end human GM12878 cell anti-Serum Response Factor (SRF) ChIP sample on chromosome 1. Examples of a read length column (**C**) and missing-column artifact (**D**) in a paired-end MEF input control sample on chromosomes 1 and 18 respectively. Positive strand reads are shown in red while negative strand reads are shown in blue. The artifact center positions x where the sample genome differs from the reference are marked with asterisks. The paired-end scatterplots show each read's position (x-axis) and associated fragment length (y-axis) to demonstrate the dependence relationship between genomic position and fragment length. Every positive strand read (red point) is accompanied by a negative strand read (blue point) originating from the same fragment. For clarity, in the paired-end column artifact plot (C), we have plotted each fragment and separated them to two groups such that in one group all fragments have positive strand reads within the positive strand column (light red column) and in the other group all fragments have negative strand reads within the negative strand column (light blue column). Explicitly, reads from fragments contributing positive strand columns are shown between the pink lines, while those from fragments contributing to the negative strand column are shown between cyan lines. The paired-end missing-column artifact (D) has a column of missing reads on the positive strand followed by a column of missing reads on the negative strand. The positive strand column of missing reads is linked to a diagonal of missing reads on the negative strand, representing the associated missing downstream fragment ends and are together outlined in light red. Similarly, the negative strand column of missing reads is linked to a diagonal of missing reads on the positive strand, representing the associated missing upstream fragment ends, and are together outlined in light blue. We highlight two fragments to demonstrate the coupling between reads on the positive strand and reads on the negative strand.

dex representing the j-th putative-binding event along the genome. Each fragment originating from binding event j covers the binding position, B_j. The binding position B_j is then related to the read position as follows:

$$R_j^p = B_j - FK, \tag{9}$$

and

$$R_j^n = B_j + F(1-K), \tag{10}$$

where R_j^p is the start position of a read on the positive strand resulting from event j and R_j^n is the end position of a read on the negative strand resulting from event j. K is a random variable taking values between zero and one, describing the relative position of the binding site within a fragment. If K equals 0, then the binding position is at the most upstream end of the fragment, and if K equals 1, then the binding position is at the most downstream end of the fragment. If K is between 1 and 0, the binding position is at that location

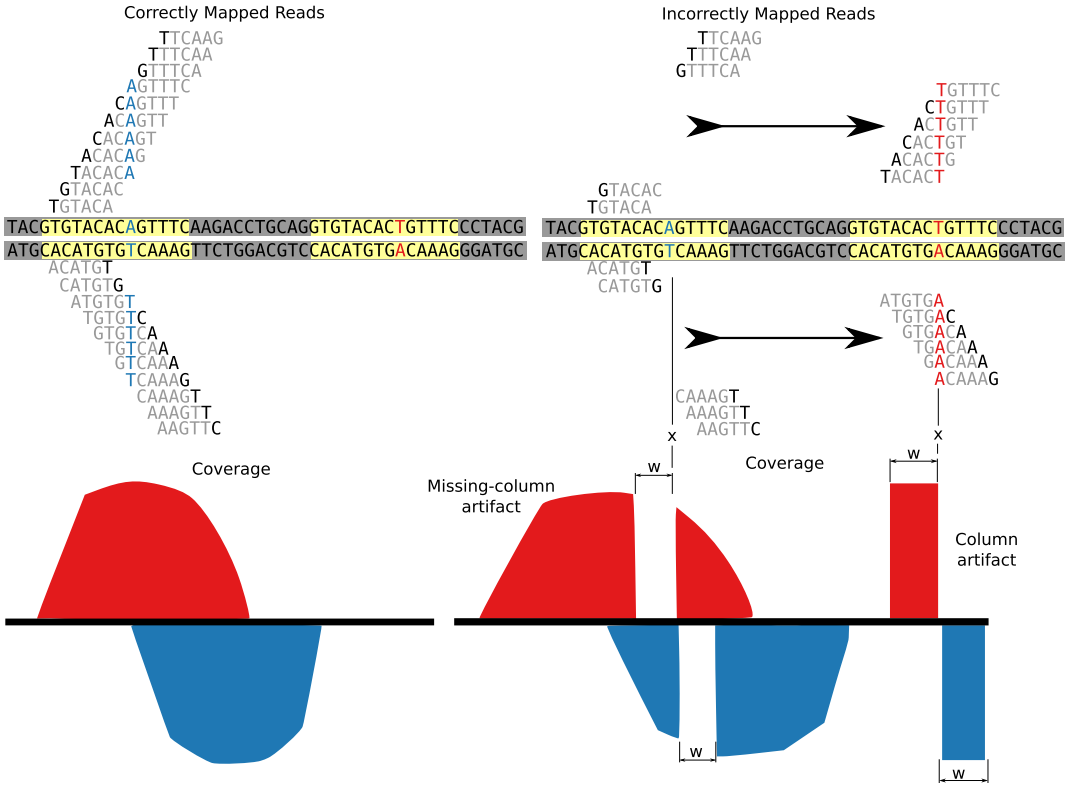

Figure 5. Read length artifacts likely stem from mapping problems. Reads that map to their correct locations do not give rise to artifacts (left). Differences between the reference and sequenced genomes can produce missing-column artifacts and additionally the coverage can be relocated to a region of higher sequence similarity forming a column artifact (right).

relative the fragment length. We model K as a β distributed random variable $\Pr[K] \sim B(\alpha, \beta)$. The β distribution is convenient for this purpose because it is flexible for modeling random variables which take values between 0 and 1. Additionally, we set $\alpha = \beta$, assuming that K is symmetrically distributed. We initialize K to a uniform distribution ($\alpha = 1$) as a natural choice in the absence of prior knowledge (step 3 of Figure 1) and as detailed subsequently we reevaluate it by optimizing α (step 7 of Figure 1). Next, applying algebra of random variables, it can easily be seen that:

$$\Pr[R_j^p] = \Pr[B_j] * \Pr[-FK], \tag{11}$$

and

$$\Pr[R_j^n] = \Pr[B_j] * \Pr[FK], \tag{12}$$

where * is the convolution operator, $\Pr[K] = \Pr[1 - K]$ due to the symmetry mentioned above and $\Pr[FK]$ is the product distribution as follows:

$$\Pr[FK = z] = \int_{-\infty}^{\infty} \Pr[F = x] \Pr[K = z/x] \frac{1}{|x|} dx$$

and (13)

$$\Pr[-FK = z] = \int_{-\infty}^{\infty} \Pr[-F = x] \Pr[K = z/x] \frac{1}{|x|} dx.$$

$\Pr[-FK]$ is the distribution of local read coverage on the positive strand with support upstream of a binding position (negative offset). $\Pr[FK]$ is the distribution of local read coverage on the negative strand with support downstream of a binding position (positive offset). We will use these local coverage patterns in both steps 4 and 6 of Figure 1 to locate and quantify candidate peaks respectively.

Matched filtering for rapid and accurate localization of candidate peaks

TFs bind to a small fraction of the genome, thus to improve efficiency we only test candidate peak positions (step 6 of Figure 1) that closely match our expected peak shape ($\Pr[-FK]$ and $\Pr[FK]$), as identified by a matched filter (51) (step 4 of Figure 1).

In signal processing, a filter is a function which selects for the desired output signal vector, s and suppresses the undesirable noise vector, v, of an observed input signal $x = s + v$. A matched filter (51) is a specialized filter whose time inverse is the impulse response function, h, where h is optimally parallel to the desired signal ($h \| s$) and orthogonal to the noise ($h \perp v$). The matched filter has the favorable property that when convolved with an observed signal, it will maximize the output signal to noise ratio. The time inversed matched filter, h, is defined as follows:

$$h = \gamma \Sigma^{-1} s, \tag{14}$$

where γ is a normalization constant and Σ^{-1} is the inverse covariance matrix of the noise.

In the context of ChIP-seq experiments, the observed input signal x is the read count at each position. For identi-

fying peaks, the desired output signal s is associated with an impulse response of $\Pr[-FK]$ for the positive strand and $\Pr[FK]$ for the negative strand. The read count noise is not globally stationary, but locally it is approximately a stationary process and is thus independent and identically distributed (i.i.d.) with the level of noise changing relatively slowly on a much larger length scale than the support ($2F_{\max}$) of the desired signal. When the noise is i.i.d., (14) is reduced to

$$h = \tilde{\gamma} s, \tag{15}$$

where the amplitude of the noise is absorbed in $\tilde{\gamma}$. From Equation (15) we see that the filter h is proportional to the impulse response of s. Specificaly, we use the following filters h^p and h^n for the positive and negative strands respectively:

$$h^p = \Pr[-FK] \quad h^n = \Pr[FK]. \tag{16}$$

We define the filtered signal Y at each position by the formula:

$$Y = h^p(-t) * \Pr[R^p] + h^n(-t) * \Pr[R^n]. \tag{17}$$

Usually γ is chosen to normalize the expected power of the noise after application of the filter to one. However, a given γ that normalizes the noise in low coverage areas to one, necessarily will give higher power in areas of higher coverage. Therefore, it is infeasible to specify a single γ when the noise is not globally stationary. As a result, the standard way of detecting the desired output signal by thresholding using a fixed signal to noise ratio is not applicable.

Instead, we identify local maxima using a Gaussian derivative filter, a technique commonly used for detecting local maxima (edges) in images as follows:

$$0 = Y * \frac{d}{dx}\mathcal{G}(0, \sigma^2), \tag{18}$$

where $\mathcal{G}$ is the Gaussian distribution with variance σ^2. Zero crossings (Equation (18)) from positive to negative of this smoothed first derivative are the local maxima. A minimum read count requirement is applied to avoid spurious low coverage local maxima. Those local maxima passing the threshold are selected as candidate-binding events, $b_j; j \in [1, N]$ (for N candidate events).

Remove false positive-binding events using cross-correlation

Genomic regions that have similarly high read coverage in both the ChIP sample and the negative control are false positive-binding events. Most current peak calling algorithms rely on negative controls (usually total DNA input) to control the FPR. To the best of our knowledge, if negative controls are not available, false positive events would not be filtered out by current peak calling algorithms. We have discovered that most of the significant false positive events are in fact the read length artifacts described earlier and exemplified in Figure 6. We have already shown that the cross-correlations and deconvolved FLDs of regions with read length artifacts exhibit the 'phantom peak' at read length (Figure 3).

Ritornello identifies regions containing false positive events by detecting the 'phantom peak' in their cross-correlations; true-binding events contain no such 'phantom peak' in thier cross-correlations. To this end, we employ a machine learning approach, building a classifier to distinguish between read length artifacts and true-binding events. To extract features for the classifiers, we calculated cross-correlation locally from $b_j - F_{\max}$ to $b_j + F_{\max}$ around each candidate peak. The features include: (i) the maximum value of the cross-correlation function in the range between zero and read length, which is denoted by c_1 and (ii) the maximum value of the cross-correlation function in the range between the read length and the maximum fragment length $F_{\max}$, which is denoted by c_2. In the neighborhoods of binding events c_2 is expected to be higher than c_1, whereas in neighborhoods of read-length artifacts we expect c_1 to be larger than c_2. We added additional features to account for consecutive read length artifacts of varying amplitudes and large amplifications such as due to polymerase chain reaction (PCR). For this purpose, we binarized the coverage in the positive and negative strands by setting positions with read count >0 to 1. We then performed a running mean smoothing on the binarized coverage, calculated cross-correlation and extracted the following features: (i) the maximum value of the binarized smoothed cross-correlation function in the range between zero and read length, which is denoted by d_1 and (ii) the maximum value of the binarized smoothed cross-correlation function in the range between the read length and the maximum fragment length $F_{\max}$, which is denoted by d_2.

We build a classifier using logistic regression, with features: $\{\frac{c_2}{c_1}, \frac{d_2}{d_1}\}$. The instances used to build this classifiers include manually classified peaks obtained as follows: we first applied MACS2 (negative control free mode) to four TF ChIP-seq datasets generated by ENCODE, subsequently selected the top 200 peaks for each of four samples and finally manually labeled regions with typical binding shape as true positives (see Figure 6A) and regions with characteristic read length artifact as false positives (see Figure 6B). We trained this model using a five fold cross-validation and achieved high performance with AUROC of 0.993. This set of features is scale-free and thus our trained classifier is generalizable to any ChIP-seq sample and does not need to be retrained. Ritornello incorporates this trained classifier (step 5 of Figure 1) to flag artifactual locations as false positives without the need for a paired total DNA input or IgG control.

Deconvolving single events from local coverage

The read coverage near an event is a mixture of reads generated by that event, noise and any neighboring events. In order to accurately quantify each event, it is essential to deconvolve its binding intensity (number of reads originating from each event) from this mixture. Fragments originating from different events in close proximity may overlap. Consequently, it is difficult to quantify the number of reads coming from each binding event. Further, fragments originating from non ChIP-ed DNA, off targeted sequencing due to antibody inefficiency, as well as other sources, contribute to background noise. We model the read coverage around

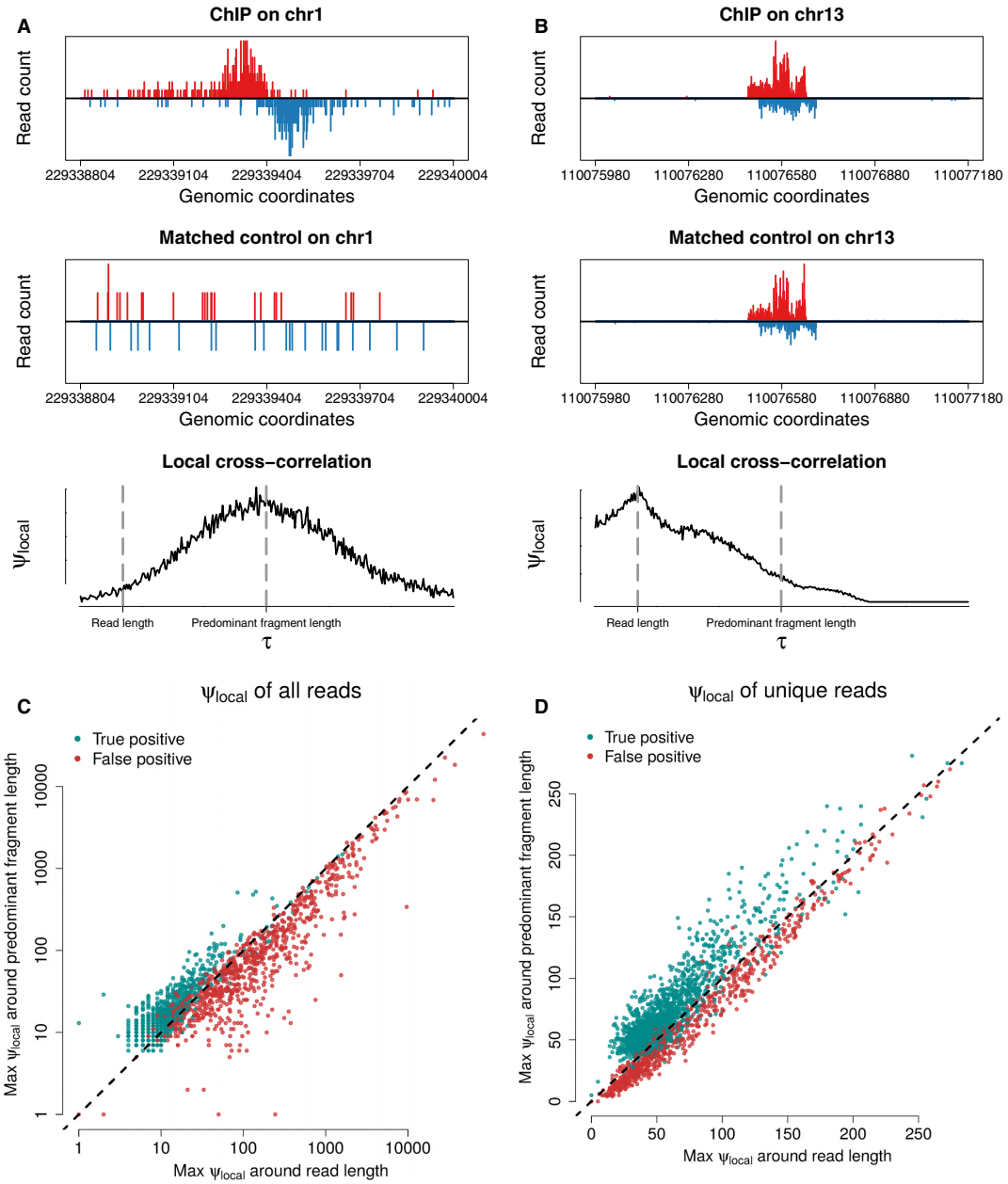

Figure 6. Local cross-correlation differentiates true-binding events from read length artifacts (false positives). (**A**) Local cross-correlation of a binding event in anti-ATF2 K562 ChIP sample peaks near the average fragment length. (**B**) Local cross-correlation of a column artifact in anti-ATF2 K562 ChIP sample peaks near the read length. (**C**) A scatterplot of the maximum local log cross-correlation up to 10 bp beyond the read length versus the maximum local log cross-correlation in the range of 10 bp beyond the read length to $0.75F_{\max}$ (**D**) A scatterplot of the maximum local binarized cross-correlation (using unique reads) up to 10 bp beyond the read length versus the maximum local binarized cross-correlation in the range of 10 bp beyond the read length to $0.75F_{\max}$.

each candidate peak using a generalized linear model to deconvolve its binding intensity.

The binding intensity, β_j, of each candidate peak, j, is only dependent on positions where that peak has support, $i \in [b_j - F_{\max}, b_j + F_{\max}]$. To efficiently deconvolve the signal at b_j we first discard peaks that do not overlap with b_j. We retain only the subset $q_k \in \{q_1 \ldots q_T\}$ of T peaks in close proximity to j (including j), such that the support of each peak in q overlaps with the support of j. The locally uniformly distributed noise associated with this neighborhood is indexed by q_0. Here we assume that read counts follow a Poisson distribution, a common assumption made by other algorithms, such as MACS and GEM (3,23). We can then model the number of reads on the positive and negative strands, C^p_{i,q_k} and C^n_{i,q_k}, at position i due to event q_k as follows:

$$C^p_{i,q_k} \sim Pois\left(\beta^p_{q_k} h^p(i - b_{q_k})\right), \tag{19}$$

and

$$C^n_{i,q_k} \sim Pois\left(\beta^n_{q_k} h^n(i - b_{q_k})\right), \tag{20}$$

where the parameters $\beta^p_{q_k}$ and $\beta^n_{q_k}$ denote the binding intensities (expected read counts) of event q_k. The impulse response functions $h^p(i - b_{q_k})$ and $h^n(i - b_{q_k})$ are the probabilities of observing a read at position i from event q_k. We note that different $\beta^p_{q_k}$ and $\beta^n_{q_k}$ values are used to account for local differences in read coverage between positive and negative strands.

To model the noise we will once again invoke our assumption of locally stationary noise, as in the discussion before Equation (16). Here we assume that the locally stationary noise is a uniformly Poisson distribution. We model the read counts due to noise at position i for the positive and negative strands as follows:

$$C^p_{i,q_0} \sim Pois\left(\beta^p_{q_0} U(i)\right), \tag{21}$$

and

$$C^n_{i,q_0} \sim Pois\left(\beta^n_{q_0} U(i)\right), \tag{22}$$

where U is a function that is locally uniform with support of $2F_{\max}$ around b_j and $\beta^p_{q_0}$ and $\beta^n_{q_0}$ are the expected number of reads due to noise on the positive and negative strands respectively.

The read count at position i is then given by the sum of read counts from all sources q_k as follows:

$$C^p_i = C^p_{i,q_0} + \sum_{k=1}^{T} C^p_{i,q_k} \sim Pois\left(\lambda^p_{i,q}\right), \tag{23}$$

where

$$\lambda^p_{i,q} = \beta^p_{q_0} U(i) + \sum_{k=1}^{T} \beta^p_{q_k} h^p(i - b_{q_k}), \tag{24}$$

and

$$C^n_i = C^n_{i,q_0} + \sum_{k=1}^{T} C^n_{i,q_k} \sim Pois\left(\lambda^n_{i,q}\right), \tag{25}$$

where

$$\lambda^n_{i,q} = \beta^n_{q_0} U(i) + \sum_{k=1}^{T} \beta^n_{q_k} h^n(i - b_{q_k}). \tag{26}$$

The relationships in Equations (23) and (25) use the following theorem: if $X_1 \ldots X_n$ are independent Poisson distributed random variables, $Pois(\lambda_1) \ldots Pois(\lambda_n)$, then their sum $X_1 + \ldots + X_n$ is Poisson distributed, $Pois(\lambda_1 + \ldots + \lambda_n)$.

In order to obtain the binding intensity for peak, b_j, we maximize the likelihood for the models (Equations (23) and (26)) of all nucleotides around b_j. The likelihood of local binding intensities β^p_q and β^n_q around the peak of interest, j, can be written as:

$$L(\beta^p_q, \beta^n_q | C^p_i, C^n_i) = \prod_{i=b_j - F_{max}}^{b_j + F_{max}} Pois\left(C^p_i; \lambda^p_{i,q}\right) Pois\left(C^n_i; \lambda^n_{i,q}\right). \tag{27}$$

We then find the maximum likelihood estimates for parameters β^p_q and β^n_q. The sum, $\beta^p_j + \beta^n_j$, is reported as the binding intensity for the peak at b_j.

We note that this is formally a Poissson generalized linear model with identity link function. Such a model has the advantage that it can resolve multiple peaks in close proximity, such as double-binding or triple-binding events. To our knowledge only BRACIL (46) and CSDeconv (47) are designed to deconvolve adjacent-binding events, in particular double-binding events and use different models. Those two algorithms are less efficient and therefore require as input a set of peaks from other peak callers.

Ritornello implements a dogleg optimization (the Newton–Raphson method coupled with initial gradient descent), which is much faster than traditional Expectation Maximization or Markov Chain Monte Carlo methods (52), enabling the rapid deconvolution of all loci detected in previous steps.

Testing candidate peaks for significance

In the previous section we quantified the intensity, β_j, of each candidate peak. Here we determine the significance of each of these candidate-binding events using a likelihood ratio test based on the likelihood we derived in Equations (27). The null model H^0_j is obtained by setting both $\beta^p_j = 0$ and $\beta^n_j = 0$. We note that we use the term null model at each position b_j to refer to the model involving a zero-binding intensity at b_j but with potentially nonzero $\beta^p_{q_0}$ and $\beta^n_{q_0}$ as well as β^p_q and β^n_q in neighboring candidate events. The alternative model H^1_j uses full parameterization including non vanishing β^p_j and β^n_j at the location of interest b_j.

Since the null model H^0_j is nested within H^1_j, we can employ the likelihood ratio test statistic (D) in the form:

$$\begin{aligned} D &= 2\ln\left(\frac{\max\left\{L(\beta^p_q, \beta^n_q | C^p_i, C^n_i)\right\}}{\max\left\{L(\beta^p_{q'}, \beta^n_{q'} | C^p_i, C^n_i)\right\}}\right) \\ q' &\equiv q \setminus \{j\} \\ H^0_j &: \beta_{q'} \in \Re, \beta_j = 0 \\ H^1_j &: \beta_q \in \Re. \end{aligned} \tag{28}$$

According to Wilke's theorem (53) the likelihood ratio test statistic for this nested model is distributed according to a χ^2 distribution with two degrees. We then calculate the p-value for each peak based on this χ^2 distribution.

Up to this point, we have obtained an initial list of putative-binding events. This was based on inferring an impulse response function given by the product distribution of the FLD and a uniformly distributed K (step 3 Figure 1). To further refine the impulse response function, we find the estimate of α that maximizes the combined likelihood of the 200 most significant putative events (step 7 of Figure 1). As shown in Figure 7, the $\Pr[-FK]$ and $\Pr[FK]$ derived from this procedure closely match the shape of highly abundant peaks. Finally, we repeat the peak identification, artifact testing, and likelihood ratio testing (steps 4–6 of Figure 1) using the updated h^p and h^n and report final list of significant peaks.

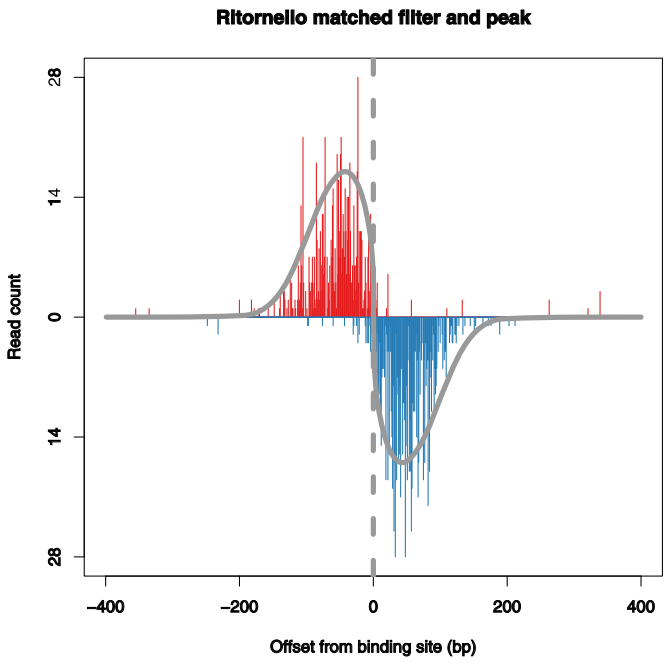

Figure 7. The parameterized filter closely matches the peak shape. Shown is an example peak in a human SRF sample. The parameterized filter for this sample is shown in gray. The peak location as determined by the filter is shown by a dashed line.

FDR correction

To control the false discovery rate, we adjust the *P*-values, p_i, associated with each hypothesis test by applying the Benjamini–Hochberg (54) approach to obtain *q*-values, q_i, which control the FDR as follows: $q_i = p_i \frac{m}{k_i}$, where k_i is the *P*-value rank by significance and m, is the total number of potential hypothesis tested. Finally, we ensure monotonicity using the Benjamini–Yekutieli correction (55).

Usually one performs all hypothesis tests and m is set to that number. For the likelihood ratio test that Ritornello performs, it is important to note that the null hypothesis is the absence of a peak at a position of interest in the presence of any amount of coverage due to uniform background signal or neighboring peaks. The appropriate m for Ritornello is then, not the number of candidate peaks tested, as these were previously enriched for peak shape coverage by the matched filter, but rather the total number of windows exceeding the minimum read count threshold before selecting those that match the filter.

RESULTS

To assess Ritornello's performance, we compared it against the MACS2 (3) and GEM (23) peak callers, which have both been recommended by the ENCODE consortium (56,57). We use 14 single-end TF ChIP-seq experiments from the ENCODE project (57), each with two biological replicates (see Table 1 and Supplementary Table S1). Matched DNA input or IgG controls were also available (Supplementary Table S1).

Although matched input controls are optional for some methods, they are used by ChIP-seq peak calling algorithms to avoid calling many spurious false positives. Both MACS2 and GEM have options to run with or without the matched input control. To show that Ritornello, which is a matched control free approach, avoids calling these false positives we benchmark it against MACS2 (with and without the matched control), and GEM (with and without the matched control) on each of the 28 samples. We refer to MACS2 with the control as MACS2-I and without the control as MACS2. We refer to GEM with the control as GEM-I and without the control as GEM. To control for the variable size of reported peaks, we used the interval from 100 bp upstream to 100 bp downstream of each peak summit identified by MACS2 or GEM, and of each peak location identified by Ritornello. If multiple peaks called from a single algorithm overlapped each other using this 200 bp window, only the peak with the more significant *q*-value is considered. Mitochondria were excluded from all analysis. We compare the performance of these algorithms in terms of:

i) the percentage of peaks unique to Ritornello with respect to each control utilizing sample.
ii) the similarity between characteristic coverages in binding events predicted uniquely by Ritornello, MACS2-I, or GEM-I to the coverages of strong binding events predicted by multiple algorithms.
iii) motif enrichment.

Additionally, we demonstrate that Ritornello predicts very few false positives in negative controls suggesting that Ritornello has a low FPR. We also observe that Ritornello produces results that are reproducible among technical replicates as determined by the irreproducible discovery rate algorithm (58) (Supplementary Table S2).

Most peaks called by Ritornello are common to MACS2-I or GEM-I

Control samples are used to eliminate false positive peaks during ChIP-seq peak calling. It is therefore important to measure how many peaks are unique to Ritornello with respect to the control using algorithms. The percentages of peaks unique to Ritornello with respect to MACS2-I and GEM-I are shown in Table 2. Ritornello tends to call few (<10%) unique peaks in all samples with the exceptions of E2F4 and ATF2. Although Ritornello tends to report fewer peaks than MACS2-I and GEM-I, these peaks are most often a subset of the peaks reported by those input using algorithms. Peaks that are called by MACS2-I and GEM-I but not by Ritornello typically have less reads and hence the shape of the read coverage at these loci is not well defined.

Comparing Ritornello and alternative methods based on unique peak coverage patterns

We next investigate the peak shape and motif content for samples where Ritonello called >10% unique peaks relative to MACS2-I or GEM-I. For this analysis, we compared the peaks that are uniquely reported by Ritornello to the corresponding algorithms in Figure 2. To be fair when comparing read coverage of unique peaks, we averaged the local distributions of read start positions (pileup) for the top

Table 1. TF ChIP-seq experiments (two replicates each)

TF	Cell type
NRF1	K562
SRF	GM12878
REST	H1
MAX	K562
ATF2	H1
E2F4	GM12878
GATA1	MEL
MYC	MEL
CTCF	Myotube
ELK1	K562
SRF	H1
MAX	H1
YY1	H1
REST	K562

Table 2. Percentage of peaks unique to Ritornello as compared to MACS2-I and GEM-I

	Vs–GEM–I	Vs–MACS2–I
ATF2 H1 rep1	5.2%	3.8%
ATF2 H1 rep2	29.2%	42.0%
CTCF Myotube rep1	<0.1%	0.6%
CTCF Myotube rep2	<0.1%	0.6%
E2F4 GM12878 rep1	0.8%	12.3%
E2F4 GM12878 rep2	1.8%	35.6%
ELK1 K562 rep1	0.8%	2.6%
ELK1 K562 rep2	0.8%	2.0%
GATA1 MEL rep1	0.4%	1.5%
GATA1 MEL rep2	0.4%	1.4%
MAX H1 rep1	1.9%	7.8%
MAX H1 rep2	3.1%	5.7%
MAX K562 rep1	0.5%	2.4%
MAX K562 rep2	1.9%	7.4%
MYC MEL rep1	3.7%	5.9%
MYC MEL rep2	5.0%	5.7%
NRF1 K562 rep1	0.3%	1.3%
NRF1 K562 rep2	0.5%	0.5%
REST H1 rep1	0.3%	0.9%
REST H1 rep2	0.6%	1.4%
REST K562 rep1	3.8%	1.1%
REST K562 rep2	0.5%	0.6%
SRF GM12878 rep1	<0.1%	0.7%
SRF GM12878 rep2	<0.1%	0.3%
SRF H1 rep1	1.2%	0.4%
SRF H1 rep2	<0.1%	0.6%
YY1 H1 rep1	0.8%	3.9%
YY1 H1 rep2	0.4%	3.4%

For each of the 28 experiments (replicates are denoted by rep1 and rep2), we highlight in dark blue where Ritornello had >10% unique peaks, light blue 5–10% unique peaks and gray <5% unique peaks. Ritonello calls few (<10%) unique peaks in most samples.

200 most significant unique peaks of each algorithm (pileup plots in Figure 8 and additionally Supplementary Figures S1 and 2) and found that the characteristic patterns (black curves whose shapes match the impulse response functions) associated with the highest intensity peaks reported by all algorithms tend to be similar to the patterns obtained by aggregating the pileups of peaks uniquely reported by Ritornello but in contrast are less similar to the aggregated pileups of peaks uniquely reported by MAC2-I or GEM-I. These results provide evidence that peaks uniquely reported by Ritornello may be more likely to be indicative of binding events than those uniquely reported by MACS2-I or GEM-I.

Table 3. Percentage of the top 1000 peaks (or all peaks when <1000) within 100 bp of a motif obtained by Ritornello, MACS2-I, MACS, GEM-I and GEM

	GEM	GEM–I	MACS2	MACS2–I	Ritornello
ATF2 H1 rep1	29.4%	46.1%	41.3%	52.3%	47.9%
ATF2 H1 rep2	33.6%	25.0%	41.1%	23.8%	44.5%
CTCF Myotube rep1	89.9%	89.6%	88.7%	88.9%	89.6%
CTCF Myotube rep2	84.1%	87.6%	87.0%	87.4%	88.4%
E2F4 GM12878 rep1	25.5%	31.6%	28.1%	31.0%	30.9%
E2F4 GM12878 rep2	23.9%	31.3%	27.2%	29.8%	30.8%
ELK1 K562 rep1	38.5%	51.6%	42.6%	52.9%	49.1%
ELK1 K562 rep2	44.3%	50.6%	47.6%	52.1%	52.5%
GATA1 MEL rep1	13.6%	13.0%	13.4%	14.2%	13.4%
GATA1 MEL rep2	13.4%	13.5%	14.6%	13.4%	14.7%
MAX H1 rep1	3.6%	3.7%	5.0%	4.4%	5.6%
MAX H1 rep2	4.8%	5.3%	5.9%	6.1%	6.7%
MAX K562 rep1	2.5%	2.4%	3.7%	2.9%	4.2%
MAX K562 rep2	3.3%	3.2%	4.3%	3.6%	3.7%
MYC MEL rep1	15.3%	16.5%	17.0%	19.1%	20.1%
MYC MEL rep2	17.5%	20.0%	19.1%	23.0%	20.9%
NRF1 K562 rep1	85.8%	90.4%	89.1%	90.1%	89.7%
NRF1 K562 rep2	81.7%	90.1%	87.0%	90.5%	91.3%
REST H1 rep1	84.2%	84.6%	83.8%	84.4%	85.6%
REST H1 rep2	85.0%	85.2%	81.8%	82.9%	85.8%
REST K562 rep1	80.5%	81.1%	79.8%	81.4%	81.0%
REST K562 rep2	83.5%	83.6%	83.5%	83.7%	83.8%
SRF GM12878 rep1	17.5%	19.1%	17.7%	21.1%	19.0%
SRF GM12878 rep2	15.6%	17.4%	16.3%	17.9%	18.0%
SRF H1 rep1	43.7%	49.0%	46.6%	49.4%	49.6%
SRF H1 rep2	32.3%	34.3%	34.1%	36.4%	37.3%
YY1 H1 rep1	67.0%	70.1%	68.9%	71.1%	69.5%
YY1 H1 rep2	66.9%	70.9%	69.1%	73.4%	68.8%

For each of the 28 experiments (replicates are denoted by rep1 and rep2), we label in dark blue the algorithm that outputs the largest number of motif containing peaks, light blue the algorithm that outputs the second largest number of motif containing peaks and gray the algorithms that output the smallest number of motif containing peaks.

Comparing Ritornello and alternative methods based on motif occurrence rate

Availability of genuine validations of TF-binding events inferred by TF ChIP-seq peak callers is limited. Therefore, one of the measures used by practitioners for assessing the quality of peak callers is the fraction of predicted TF-binding events that overlap with the characteristic binding motif of the relevant TF. Employing the same 28 public ChIP-seq samples, we compare the motif enrichment for the top 1000 peaks reported by each algorithm or all peaks when less than 1000 are reported as shown in (Table 3). The annotated motifs specific to the TFs are downloaded from the JASPAR CORE 2014 motif library (59). Genomic locations that match the position weight matrix of each JASPAR motif were identified using PWMScan with default *P*-value cutoff at 0.00001 (60). Predicted peaks whose binding centers are within 100 bp from the corresponding JASPAR motif are classified motif containing. The peaks found by Ritornello had the highest motif occurrence rate compared with MACS2-I, MACS2, GEM-I and GEM in 15 out of the 28 samples. MACS2-I, MACS2, GEM-I and GEM had the highest motif occurrence rate in 8, 1, 3 and 1 samples respectively. This suggests that Ritornello, which is a control free peak caller, is able to identify the most significant true-binding events at comparable rates to input using peak callers.

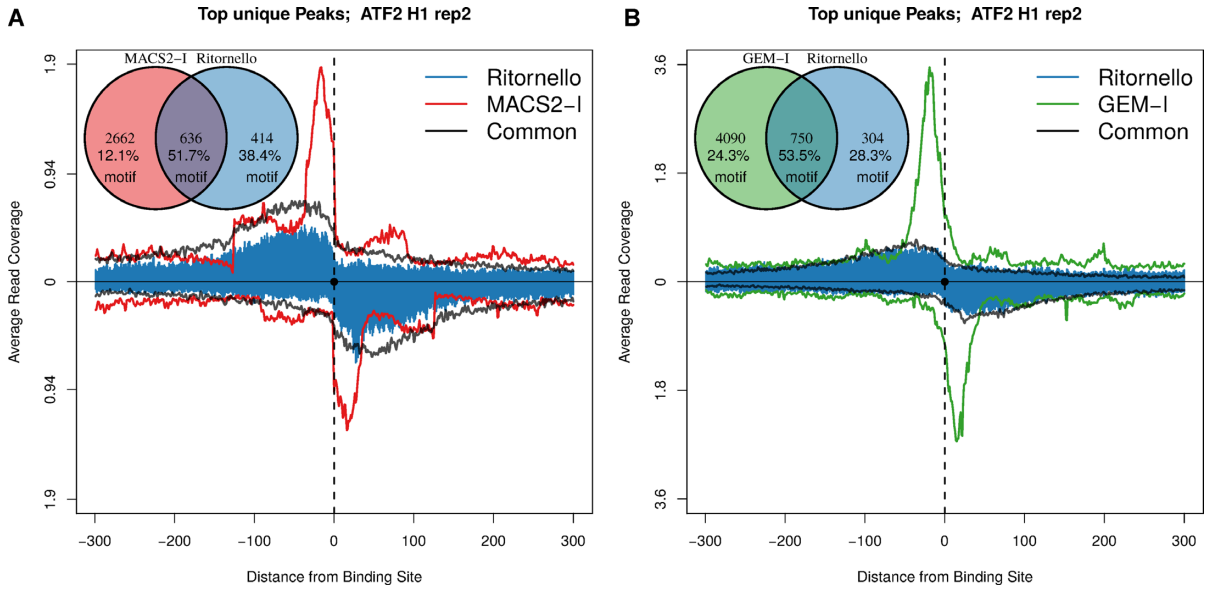

Figure 8. Pileup of read start positions for ATF2 rep2 peaks in H1 cells obtained by Ritornello only (blue) compared with that obtained from (**A**) MACS2 only(red) and (**B**) GEM only(green). The pileups of peaks common to Ritornello and MACS2 (A) or Ritornello and GEM (B) are shown in black. The pileups of read start positions for Ritornello best match the pileups of common peaks. Additionally Ritornello unique peaks have comparable levels of motif enrichment as compared to the other algorithms.

Additionally, we compare the number of motif containing peaks reported by each algorithm as a function of significance (Figure 9 and Supplementary Figures S3–9). We see that Ritornello exhibits improved motif enrichment, with respect to the other algorithms, for peaks reported in ATF2 rep2 and comparable enrichment to control utilizing algorithms in E2F4 rep1 and rep2. In general, Ritornello shows comparable motif enrichment to the input utilizing algorithms.

Ritornello identifies true-binding events in low quality samples

ENCODE recommends discarding low quality samples as determined by the NSC and RSC scores. Ritornello discards read length artifacts that give rise to low RSC and uses a matched filtering approach that maximizes the signal to noise ratio measured by the NSC, we therefore conjectured that Ritornello may be able to rescue these low quality samples. We compared the performance of Ritornello with that of MACS2-I, MACS2, GEM-I and GEM on samples with suboptimal quality based on the NSC and RSC scores. The ENCODE Consortium has suggested repeating experiments with NSC values <1.05 and RSC values <0.8. Using these criteria we identified that out of the 28 samples we investigated, four samples have suboptimal quality. These four experiments include: ATF2 H1 replicate 1 (NSC = 1.04, RSC = 0.62), ATF2 H1 replicate 2 (NSC = 1.04, RSC = 0.74), ELK1 K562 replicate 1 (NSC = 1.03, RSC = 0.64) and ELK1 K562 replicate 2 (NSC = 1.05, RSC = 0.73).

We observed that in these four samples, the pileups of peaks predicted by Ritornello have a characteristic bimodal shape of TF binding and have much stronger read coverage than their matched input controls (Figure 10 and Supplementary Figures S10–12). This suggests that there are numerous significant binding events that can be captured in low quality ChIP-seq samples. Additionally, the pileups of peaks reported by MACS2 or GEM have either non-uniform read coverage or a narrow bimodal shape (similar to column artifacts) as in Figure 10 and Supplementary Figures S10–12. It is worth noting, however, that when provided controls, the MACS2-I and GEM-I reported peaks also have a strong bimodal shape with less signal in the control sample. This illustrates that Ritornello reliably rescues peaks from low quality samples, while MACS2-I and GEM-I might be able to avoid calling false positive artifacts from low quality samples as well.

Ritornello obviates the need for a matched input control

To demonstrate that Ritornello calls few false positives that could otherwise be avoided with the aid of a match control, we show that Ritornello calls few peaks in control samples as seen in Table 4. We used the filter learned in the ChIP channel to look for regions in the corresponding control that match this pattern. This is performed by running Ritornello, supplying the filter found in the ChIP experiment rather than having it learned from the control data. The number of peaks Ritornello reported in each matched control is a proxy for the number of potential false positive peaks called by Ritornello in its corresponding ChIP-seq experiment. We found that Ritornello called very few peaks in the matched controls, which corresponds to a FPR in ChIP of <0.05 for 25 out of the 28 samples. The samples with the two highest FPR's use IgG as a control, so it is likely that Ritornello is picking up non-specific binding events, which are unlikely to be a source of noise in an actual ChIP-seq sample.

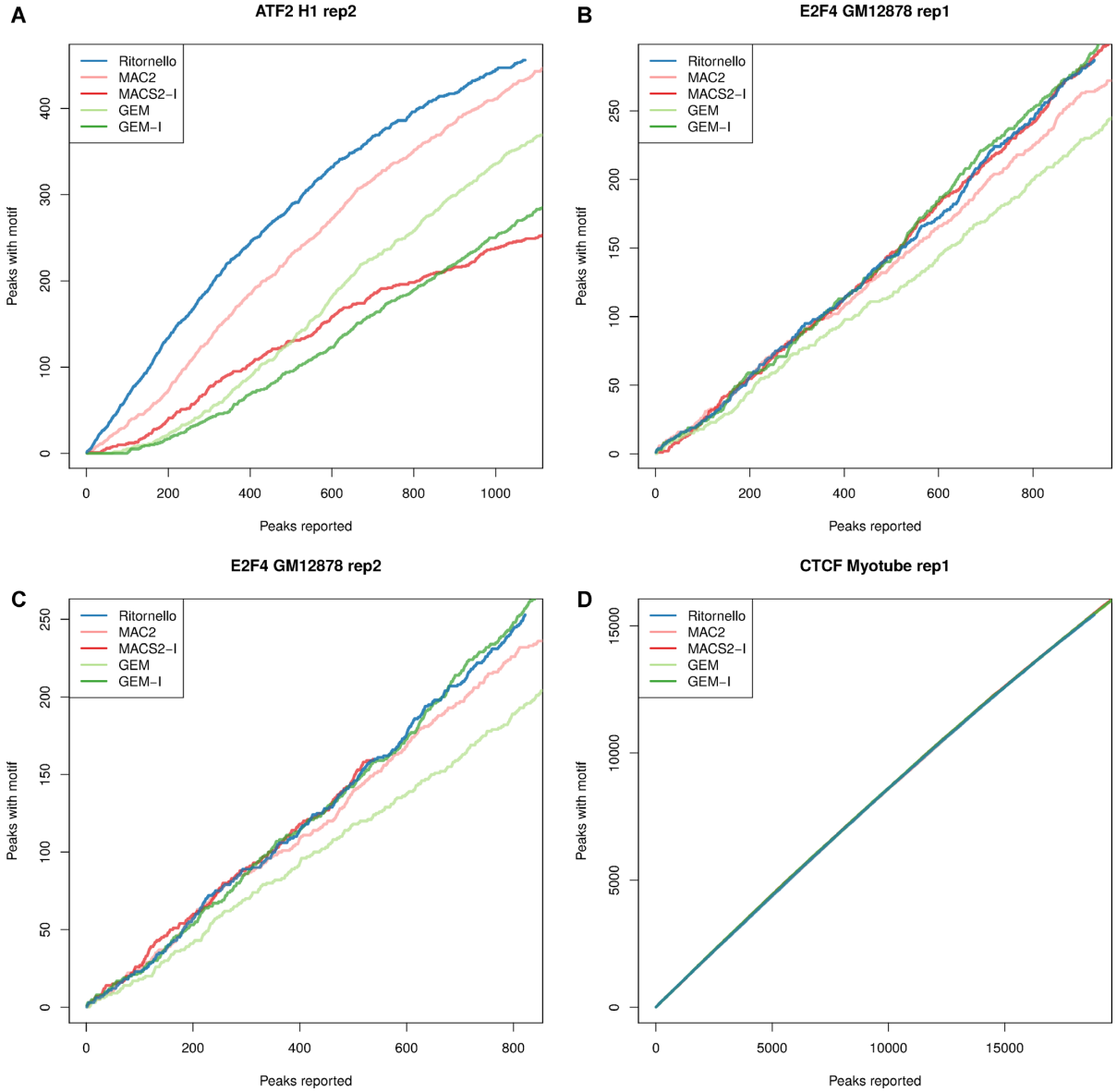

Figure 9. Motif containing peaks as a function of the rank of the *q*-value for (**A**) ATF2 rep2, (**B**) E2F4 rep1, (**C**) E2F4 rep2 and (**D**) CTCF rep1. The number of peaks displayed is slightly more than the number of peaks reported by Ritornello.

DISCUSSION

In this work, we demonstrated that we could infer the entire FLD, rather than only the mean fragment length, using a deconvolution approach from single-end TF ChIP-seq experiments. We derived an experiment-specific probabilistic model to mathematically describe the well-known bimodal shape of TF binding. Using this bimodal shape, we applied the matched filter technique from signal-processing to identify potential TF-binding sites and used a Poisson GLM to deconvolve the binding intensities and test the significance of each putative-binding event. Our model efficiently deconvolves the effect of neighboring peaks as well as noise to resolve multiple adjacent-binding events. We compared Ritornello (a control-free approach) with two popular algorithms recommended by ENCODE, MACS2-I and GEM-I, which require matched controls to reduce false positives. We found that Ritornello outperforms these other methods when input is unavailable and performs similarly when it is in terms of reproducibility between biological replicates, motif enrichment and the coverage patterns of unique reproducible peaks.

We also identified artifactual-binding regions where the local cross-correlation peaks at read length instead of around fragment length. We elucidated that these artifactual regions contribute to the phantom peaks associated with poor experimental quality. Current peak calling algorithms, such as MACS2 and GEM, rely on matched control samples to remove a substantial fraction of these artifacts. We provide an extensive description of this specific category of artifacts and their origin, and offer an automated approach to filter out artifacts without requiring matched controls. Taken together, Ritornello offers an alternative

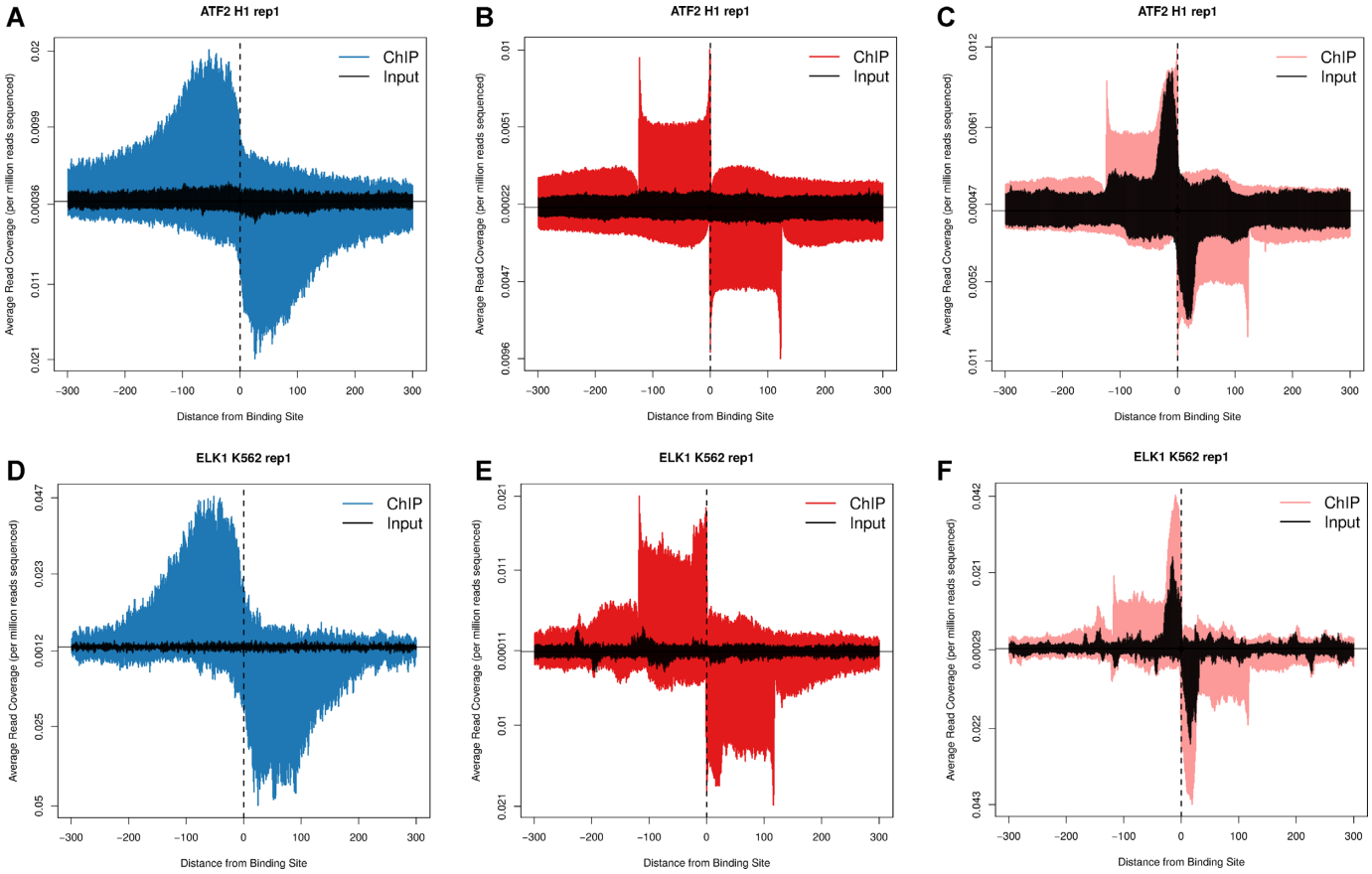

Figure 10. Pileup of read start positions for peaks identified by: Ritornello, (**A**) and (**D**) (blue); MACS2-I, (**B**) and (**E**) (dark red); and MACS2, (**C**) and (**F**) (light red) in samples of low quality and their matched controls. Matched controls are shown in black. The pileups of peaks detected by Ritornello show smooth bimodal shapes and similarly to MACS2-I, have stronger read coverage in the ChIP as compared to their matched controls. In contrast, the pileups of peaks detected by MACS2 have a narrower shape and irregular spikes similar to the negative controls.

that obviates the need for a match control, demonstrating that one can safely reduce the total experimental cost of TF ChIP-seq experiments, while providing superior analytic results.

ENCODE provides a blacklist of genomic regions which contains artifactually high read coverage in different ChIP-seq experiments (61). This manually curated blacklist largely overlaps with repetitive regions in the genome (61). The blacklist has several drawbacks: (i) it does not cover all artifactual regions, (ii) it is not generalizable to different cell types and (iii) it is only available for human and mouse. We note that a few peak calling tools, such as PePr (62), optionally remove artifacts in regions where the local read coverage in the ChIP is similar to that in the matched control. However, these methods still require a matched control. Ritornello is capable of removing artifacts independently without requiring either prior knowledge of a blacklist or matched negative controls.

Many variables influence the quality of ChIP-seq experiments and our ability to infer true-binding events from the data. These include factors such as antibody efficiency, DNA fragmentation, PCR amplification, sequencing depth, read mapping quality etc. Each of these factors may vary from one sample to another. Ritornello is designed to implicitly take into consideration these experiment specific parameters from raw data and is applicable to a wide variety of protocols. Additionally, it does not require any tuning of parameters. One limitation of Ritornello is that it is designed to detect point-source peaks such as TF-binding events. For broad-source peaks, such as epigenetic modifications, we recommend other peak callers.

We note that there are scenarios where Ritornello is expected to exceed the performance of control utilizing algorithms. These include: (i) when the control is of poor quality and (ii) positions where methods considering controls fail to eliminate read length artifacts. Additionally, in regions with very low coverage or high amounts of noise that obscure the binding shape of potential peaks, Ritornello will not call-binding events, and will thus report peaks more conservatively than MACS2-I and GEM-I.

We demonstrated that in TF ChIP-seq experiments that would otherwise be discarded due to low quality, Ritornello (as well as MACS2-I or GEM-I) might reliably recover true-binding events amidst the high levels of artifacts. Often repeat experiments with the same reagents are of poor quality according to ENCODE's metric and thus algorithms capable of handling such data are required. If, however, the repeated experiments were to improve results, the previous poor quality samples may still be of value to strengthen the findings of the higher quality samples.

Table 4. Ritornello tends to call many more peaks in ChIP-seq samples than in control samples

	ChIP	Control	FPR
ATF2 H1 rep1	992	45	4.5%
ATF2 H1 rep2	1072	2	0.2%
CTCF Myotube rep1	18847	24	<0.1%
CTCF Myotube rep2	24852	11	<0.1%
E2F4 GM12878 rep1	930	12	1.3%
E2F4 GM12878 rep2	822	10	1.2%
ELK1 K562 rep1	265	3	1.1%
ELK1 K562 rep2	394	5	1.3%
GATA1 MEL rep1	11220	73	0.7%
GATA1 MEL rep2	13128	66	0.5%
MAX H1 rep1	12054	31	0.3%
MAX H1 rep2	15509	1	<0.1%
MAX K562 rep1	2868	3	<0.1%
MAX K562 rep2	13115	3	<0.1%
MYC MEL rep1	3173	498	15.7%
MYC MEL rep2	3713	496	13.4%
NRF1 K562 rep1	1640	3	0.2%
NRF1 K562 rep2	774	3	0.4%
REST H1 rep1	5115	62	1.2%
REST H1 rep2	5617	58	1.0%
REST K562 rep1	6840	0	<0.1%
REST K562 rep2	3054	0	<0.1%
SRF GM12878 rep1	1987	1	<0.1%
SRF GM12878 rep2	2302	1	<0.1%
SRF H1 rep1	1409	50	3.5%
SRF H1 rep2	955	75	7.9%
YY1 H1 rep1	6385	57	0.9%
YY1 H1 rep2	3573	93	2.6%

ControlType
IgG
input

The FPR is <0.05 for 25 out of 28 samples. It is not surprising that the two samples where the FPR is highest use IgG controls. IgG can create many non-specific binding sites and is not representative of the background noise present in a ChIP-seq sample. Control samples were run using the filter learned from the corresponding ChIP sample.

SOFTWARE AVAILABILITY

Ritornello was programmed in C++ using the FFTW (63) library for fast computation of the Fourier transform and the Samtools (64) library for interfacing with the sequence alignment/map format which has become the standard in high throughput sequencing. Ritornello is freely available for download at https://github.com/KlugerLab/Ritornello together with a detailed tutorial. Further analysis and graphics were made using the R statistical language (65).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank Francesco Strino, Vladimir Rokhlin, Ronen Talmon and Ronald Coifman for insightful discussions.

FUNDING

National Institute of Health (NIH) [U54HG006996-03 to S.W., 1R01HG008383-01A1 to Y.K., R01 GM086852 to Y.K.]. Funding for open access charge: NIH [1R01 HG008383-01A1 to Y.K, R01 GM086852 to Y.K.].
Conflict of interest statement. None declared.

REFERENCES

1. Vaquerizas,J.M., Kummerfeld,S.K., Teichmann,S.A. and Luscombe,N.M. (2009) A census of human transcription factors: function, expression and evolution. *Nat. Rev. Genet.*, **10**, 252–263.
2. Landt,S.G., Marinov,G.K., Kundaje,A., Kheradpour,P., Pauli,F., Batzoglou,S., Bernstein,B.E., Bickel,P., Brown,J.B., Cayting,P. *et al.* (2012) ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.*, **22**, 1813–1831.
3. Zhang,Y., Liu,T., Meyer,C.A., Eeckhoute,J., Johnson,D.S., Bernstein,B.E., Nusbaum,C., Myers,R.M., Brown,M., Li,W. *et al.* (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol.*, **9**, R137.
4. Boyle,A.P., Guinney,J., Crawford,G.E. and Furey,T.S. (2008) F-Seq: a feature density estimator for high-throughput sequence tags. *Bioinformatics*, **24**, 2537–2538.
5. Chen,X., Xu,H., Yuan,P., Fang,F., Huss,M., Vega,V.B., Wong,E., Orlov,Y.L., Zhang,W., Jiang,J. *et al.* (2008) Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell*, **133**, 1106–1117.
6. Valouev,A., Johnson,D.S., Sundquist,A., Medina,C., Anton,E., Batzoglou,S., Myers,R.M. and Sidow,A. (2008) Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nat. Methods*, **5**, 829–834.
7. Fejes,A.P., Robertson,G., Bilenky,M., Varhol,R., Bainbridge,M. and Jones,S.J. (2008) FindPeaks 3.1: a tool for identifying areas of enrichment from massively parallel short-read sequencing technology. *Bioinformatics*, **24**, 1729–1730.
8. Zang,C., Schones,D.E., Zeng,C., Cui,K., Zhao,K. and Peng,W. (2009) A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. *Bioinformatics*, **25**, 1952–1958.
9. Rozowsky,J., Euskirchen,G., Auerbach,R.K., Zhang,Z.D., Gibson,T., Bjornson,R., Carriero,N., Snyder,M. and Gerstein,M.B. (2009) PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nat. Biotechnol.*, **27**, 66–75.
10. Guttman,M., Garber,M., Levin,J.Z., Donaghey,J., Robinson,J., Adiconis,X., Fan,L., Koziol,M.J., Gnirke,A., Nusbaum,C. *et al.* (2010) Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat. Biotechnol.*, **28**, 503–510.
11. Wang,C., Xu,J., Zhang,D., Wilson,Z.A. and Zhang,D. (2010) An effective approach for identification of in vivo protein-DNA binding sites from paired-end ChIP-Seq data. *BMC Bioinformatics*, **11**, 81.
12. Qin,Z.S., Yu,J., Shen,J., Maher,C.A., Hu,M., Kalyana-Sundaram,S., Yu,J. and Chinnaiyan,A.M. (2010) HPeak: an HMM-based algorithm for defining read-enriched regions in ChIP-Seq data. *BMC Bioinformatics*, **11**, 369.
13. Boeva,V., Surdez,D., Guillon,N., Tirode,F., Fejes,A.P., Delattre,O. and Barillot,E. (2010) De novo motif identification improves the accuracy of predicting transcription factor binding sites in ChIP-Seq data analysis. *Nucleic Acids Res.*, **38**, e126.
14. Rashid,N.U., Giresi,P.G., Ibrahim,J.G., Sun,W. and Lieb,J.D. (2011) ZINBA integrates local covariates with DNA-seq data to identify broad and narrow regions of enrichment, even within amplified genomic regions. *Genome Biol.*, **12**, R67.
15. Newkirk,D., Biesinger,J., Chon,A., Yokomori,K. and Xie,X. (2011) AREM: aligning short reads from ChIP-sequencing by expectation maximization. *J. Comput. Biol.*, **18**, 1495–1505.
16. Cairns,J., Spyrou,C., Stark,R., Smith,M.L., Lynch,A.G. and Tavare,S. (2011) BayesPeak–an R package for analysing ChIP-seq data. *Bioinformatics*, **27**, 713–714.
17. Halbritter,F., Vaidya,H.J. and Tomlinson,S.R. (2012) GeneProf: analysis of high-throughput sequencing experiments. *Nat. Methods*, **9**, 7–8.
18. Muino,J.M., Kaufmann,K., van Ham,R.C., Angenent,G.C. and Krajewski,P. (2011) ChIP-seq analysis in R (CSAR): an R package for the statistical detection of protein-bound genomic regions. *Plant Methods*, **7**, 11.
19. Song,Q. and Smith,A.D. (2011) Identifying dispersed epigenomic domains from ChIP-Seq data. *Bioinformatics*, **27**, 870–871.
20. Hower,V., Evans,S.N. and Pachter,L. (2011) Shape-based peak identification for ChIP-Seq. *BMC Bioinformatics*, **12**, 15.

21. Feng,X., Grossman,R. and Stein,L. (2011) PeakRanger: a cloud-enabled peak caller for ChIP-seq data. *BMC Bioinformatics*, **12**, 139.
22. Zhang,X., Robertson,G., Krzywinski,M., Ning,K., Droit,A., Jones,S. and Gottardo,R. (2011) PICS: probabilistic inference for ChIP-seq. *Biometrics*, **67**, 151–163.
23. Guo,Y., Mahony,S. and Gifford,D.K. (2012) High resolution genome wide binding event finding and motif discovery reveals transcription factor spatial binding constraints. *PLoS Comput. Biol.*, **8**, e1002638.
24. Micsinai,M., Parisi,F., Strino,F., Asp,P., Dynlacht,B.D. and Kluger,Y. (2012) Picking ChIP-seq peak detectors for analyzing chromatin modification experiments. *Nucleic Acids Res.*, **40**, e70.
25. Narlikar,L. and Jothi,R. (2012) ChIP-Seq data analysis: identification of protein-DNA binding sites with SISSRs peak-finder. *Methods Mol. Biol.*, **802**, 305–322.
26. Kumar,V., Muratani,M., Rayan,N.A., Kraus,P., Lufkin,T., Ng,H.H. and Prabhakar,S. (2013) Uniform, optimal signal processing of mapped deep-sequencing data. *Nat. Biotechnol.*, **31**, 615–622.
27. Wang,R., Hsu,H.K., Blattler,A., Wang,Y., Lan,X., Wang,Y., Hsu,P.Y., Leu,Y.W., Huang,T.H., Farnham,P.J. *et al.* (2013) LOcating non-unique matched tags (LONUT) to improve the detection of the enriched regions for ChIP-seq data. *PLoS One*, **8**, e67788.
28. Kruczyk,M., Umer,H.M., Enroth,S. and Komorowski,J. (2013) Peak Finder Metaserver: a novel application for finding peaks in ChIP-seq data. *BMC Bioinformatics*, **14**, 280.
29. Wang,J., Lunyak,V.V. and Jordan,I.K. (2013) BroadPeak: a novel algorithm for identifying broad peaks in diffuse ChIP-seq datasets. *Bioinformatics*, **29**, 492–493.
30. Nakato,R., Itoh,T. and Shirahige,K. (2013) DROMPA: easy-to-handle peak calling and visualization software for the computational analysis and validation of ChIP-seq data. *Genes Cells*, **18**, 589–601.
31. Sun,G., Chung,D., Liang,K. and Keles,S. (2013) Statistical analysis of ChIP-seq data with MOSAiCS. *Methods Mol. Biol.*, **1038**, 193–212.
32. Kim,N.K., Jayatillake,R.V. and Spouge,J.L. (2013) NEXT-peak: a normal-exponential two-peak model for peak-calling in ChIP-seq data. *BMC Genomics*, **14**, 349.
33. Taskesen,E., Hoogeboezem,R., Delwel,R. and Reinders,M.J. (2013) Hypergeometric analysis of tiling-array and sequence data: detection and interpretation of peaks. *Adv. Appl. Bioinform. Chem.*, **6**, 55–62.
34. Chung,D., Park,D., Myers,K., Grass,J., Kiley,P., Landick,R. and Keles,S. (2013) dPeak: high resolution identification of transcription factor binding sites from PET and SET ChIP-Seq data. *PLoS Comput. Biol.*, **9**, e1003246.
35. Ashoor,H., Herault,A., Kamoun,A., Radvanyi,F., Bajic,V.B., Barillot,E. and Boeva,V. (2013) HMCan: a method for detecting chromatin modifications in cancer samples using ChIP-seq data. *Bioinformatics*, **29**, 2979–2986.
36. Zeng,X., Sanalkumar,R., Bresnick,E.H., Li,H., Chang,Q. and Keles,S. (2013) jMOSAiCS: joint analysis of multiple ChIP-seq datasets. *Genome Biol.*, **14**, R38.
37. Kallio,A. and Elo,L.L. (2013) Optimizing detection of transcription factor-binding sites in ChIP-seq experiments. *Methods Mol. Biol.*, **1038**, 181–191.
38. Bardet,A.F., Steinmann,J., Bafna,S., Knoblich,J.A., Zeitlinger,J. and Stark,A. (2013) Identification of transcription factor binding sites from ChIP-seq data at high resolution. *Bioinformatics*, **29**, 2705–2713.
39. Li,Y., Umbach,D.M. and Li,L. (2014) T-KDE: a method for genome-wide identification of constitutive protein binding sites from multiple ChIP-seq data sets. *BMC Genomics*, **15**, 27.
40. Wu,H. and Ji,H. (2014) PolyaPeak: detecting transcription factor binding sites from ChIP-seq using peak shape information. *PLoS One*, **9**, e89694.
41. Lund,E., Oldenburg,A.R. and Collas,P. (2014) Enriched domain detector: a program for detection of wide genomic enrichment domains robust against local variations. *Nucleic Acids Res.*, **42**, e92.
42. Hansen,P., Hecht,J., Ibrahim,D.M., Krannich,A., Truss,M. and Robinson,P.N. (2015) Saturation analysis of ChIP-seq data for reproducible identification of binding peaks. *Genome Res.*, **25**, 1391–1400.
43. Ibrahim,M.M., Lacadie,S.A. and Ohler,U. (2015) JAMM: a peak finder for joint analysis of NGS replicates. *Bioinformatics*, **31**, 48–55.
44. Jalili,V., Matteucci,M., Masseroli,M. and Morelli,M.J. (2015) Using combined evidence from replicates to evaluate ChIP-seq peaks. *Bioinformatics*, **31**, 2761–2769.
45. Barski,A., Cuddapah,S., Cui,K., Roh,T.Y., Schones,D.E., Wang,Z., Wei,G., Chepelev,I. and Zhao,K. (2007) High-resolution profiling of histone methylations in the human genome. *Cell*, **129**, 823–837.
46. Gomes,A.L., Abeel,T., Peterson,M., Azizi,E., Lyubetskaya,A., Carvalho,L. and Galagan,J. (2014) Decoding ChIP-seq with a double-binding signal refines binding peaks to single-nucleotides and predicts cooperative interaction. *Genome Res.*, **24**, 1686–1697.
47. Lun,D.S., Sherrid,A., Weiner,B., Sherman,D.R. and Galagan,J.E. (2009) A blind deconvolution approach to high-resolution mapping of transcription factor binding sites from ChIP-seq data. *Genome Biol.*, **10**, R142.
48. Stanton,K.P., Parisi,F., Strino,F., Rabin,N., Asp,P. and Kluger,Y. (2013) Arpeggio: harmonic compression of ChIP-seq data reveals protein-chromatin interaction signatures. *Nucleic Acids Res.*, **41**, e161.
49. Ramachandran,P., Palidwor,G.A., Porter,C.J. and Perkins,T.J. (2013) MaSC: mappability-sensitive cross-correlation for estimating mean fragment length of single-end short-read sequencing data. *Bioinformatics*, **29**, 444–450.
50. Lederman,R. (2013) A random-permutations-based approach to fast read alignment. *BMC Bioinformatics*, **14**(Suppl. 5), S8.
51. North,D.O. (1963) An analysis of the factors which determine signal/noise discrimination in pulsed-carrier systems. *Proc. IEEE*, **51**, 1016–1027.
52. Ruud,P. (1989) A Comparison of the EM and Newton–Raphson Algorithms. *Economics Working Papers 89-105*. University of California at Berkeley, Berkeley.
53. Wilks,S.S. (1938) The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Ann. Math. Statist.*, **9**, 60–62.
54. Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B Stat. Methodol.*, **57**, 289–300.
55. Benjamini,Y. and Yekutieli,D. (2001) The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.*, **29**, 1165–1188.
56. Consortium,E.P. (2004) The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science*, **306**, 636–640.
57. Consortium,E.P. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
58. Li,Q., Brown,J.B., Huang,H. and Bickel,P.J. (2011) Measuring reproducibility of high-throughput experiments. *Ann. App. Stat.*, **5**, 1752–1779.
59. Mathelier,A., Zhao,X., Zhang,A.W., Parcy,F., Worsley-Hunt,R., Arenillas,D.J., Buchman,S., Chen,C.-y.Y., Chou,A., Ienasescu,H. *et al.* (2014) JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, **42**, D142–D147.
60. Iseli,C., Ambrosini,G., Bucher,P. and Jongeneel,C.V. (2007) Indexing strategies for rapid searches of short words in genome sequences. *PLoS One*, **2**, e579.
61. Carroll,T.S., Liang,Z., Salama,R., Stark,R. and de Santiago,I. (2014) Impact of artifact removal on ChIP quality metrics in ChIP-seq and ChIP-exo data. *Front. Genet.*, **5**, 75.
62. Zhang,Y., Lin,Y.H., Johnson,T.D., Rozek,L.S. and Sartor,M.A. (2014) PePr: a peak-calling prioritization pipeline to identify consistent or differential peaks from replicated ChIP-Seq data. *Bioinformatics*, **30**, 2568–2575.
63. Frigo,M. (1999) A fast Fourier transform compiler. In: Ryder,BG, Zorn,BG and Berman,AM (eds). *Acm Sigplan Notices*. ACM, NY, Vol. **34**, pp. 169–180.
64. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G. and Durbin,R. (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
65. R Core Team (2013) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna.