BIOINFORMATICS ORIGINAL PAPER

Vol. 31 no. 1 2015, pages 48–55
doi:10.1093/bioinformatics/btu568

Sequence analysis

Advance Access publication September 15, 2014

JAMM: a peak finder for joint analysis of NGS replicates

Mahmoud M. Ibrahim[1,2,*], Scott A. Lacadie[2] and Uwe Ohler[1,2,*]

[1]Department of Biology, Humboldt University, Invalidenstrasse 43, D-10115 Berlin, Germany and [2]The Berlin Institute for Medical Systems Biology, Max Delbrück Center for Molecular Medicine Berlin-Buch, Robert Rössle Str. 10, Berlin 13125, Germany

Associate Editor: Inanc Birol

ABSTRACT

Motivation: Although peak finding in next-generation sequencing (NGS) datasets has been addressed extensively, there is no consensus on how to analyze and process biological replicates. Furthermore, most peak finders do not focus on accurate determination of enrichment site widths and are not widely applicable to different types of datasets.

Results: We developed JAMM (Joint Analysis of NGS replicates via Mixture Model clustering): a peak finder that can integrate information from biological replicates, determine enrichment site widths accurately and resolve neighboring narrow peaks. JAMM is a universal peak finder that is applicable to different types of datasets. We show that JAMM is among the best performing peak finders in terms of site detection accuracy and in terms of accurate determination of enrichment sites widths. In addition, JAMM's replicate integration improves peak spatial resolution, sorting and peak finding accuracy.

Availability and implementation: JAMM is available for free and can run on Linux machines through the command line: http://code.google.com/p/jamm-peak-finder

Contact: mahmoud.ibrahim@mdc-berlin.de or uwe.ohler@mdc-berlin.de.

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on June 24, 2014; revised on August 8, 2014; accepted on August 18, 2014

1 INTRODUCTION

A common task in Genomics research is detecting enriched sites after alignment of next-generation sequencing (NGS) reads, which involves separating the genome into regions of high enrichment (i.e. peaks or clusters or binding sites) and regions of low enrichment (Pepke *et al*., 2009). Most peak and cluster finding programs are developed with a specific experimental protocol or dataset type in mind (Kumar *et al*., 2013). Therefore, it is usually difficult to apply the same analysis pipeline uniformly across all datasets in a given project.

Recently, there were attempts to develop *universal* peak finders by defining the problem as that of classical signal detection (Kumar *et al*., 2013). The main advantage of this approach is that it allows for uniform data analysis via theoretically proven optimal signal detection. However, it does not take into account that enrichment sites are often not expected to have the same shape or signal properties, even if in the same dataset. For example, DNase-I hypersensitive regions are expected to have different widths and signal-to-noise ratios (SNR; Natarajan *et al*., 2012). Therefore, there is a need for an approach that would not only focus on optimal detection of enrichment sites but would also be able to adapt to enrichment sites with different signal properties and to define their boundaries accurately (Xing *et al*., 2012).

Furthermore, while others focus on integration of multiple datasets to define co-occurrence or differential enrichment [see Li *et al*. (2014); Liu *et al*. (2013); Shen *et al*. (2013) and Zeng *et al*. (2013) for examples], there is no consensus on biological replicates integration to find accurate consensus peaks. One common approach is to determine enriched sites on each replicate separately, and then combine the results via union or intersection (Schweikert *et al*., 2012; Yang *et al*., 2014). Another common approach is to pool aligned reads from all replicates available and then detect enriched sites on the pooled alignments [see Tuteja *et al*. (2009) for an example]. Taking the intersect or union of separately detected sites mandates rescoring the peaks and leads to inaccurate enriched sites' widths. Pooling alignments before site detection obscures the differential spatial and intensity information in the replicates. As biological replicate experiments are not expected to be exactly reproducible, there is a need to develop a method for replicate integration that takes advantage of the differential information in the individual replicates to find accurate consensus peaks.

In this article, we introduce JAMM (Joint Analysis of NGS replicates via Mixture Model clustering): a universal peak finder that can integrate information from multiple replicates to find consensus peaks, determine accurate peak widths and resolve neighboring narrow peaks. We demonstrate JAMM using ChIP-Seq (Johnson *et al*., 2007), including transcription factor ChIP-Seq, punctate histone modification ChIP-Seq and broad histone modification ChIP-Seq as well as DNase-Seq (Crawford *et al*., 2006). We compare several programs that focus on different aspects of the peak finding problem (Table 1). MACS (Zhang *et al*., 2008) models read counts using a local Poisson distribution, PeakRanger (Feng *et al*., 2011) focuses on detecting neighboring narrow peaks at high resolution, PeakZilla (Bardet *et al*., 2013) is designed for uniform punctate transcription factor binding sites, BCP (Xing *et al*., 2012) develops explicit formulas to model read counts, CCAT (Xu *et al*., 2010) detects broad enrichment patterns with low SNR and DFilter (Kumar *et al*., 2013) is a universal peak finder based on optimal signal detection. We demonstrate that JAMM is widely applicable to different types of datasets, can define accurate peak boundaries and that JAMM's replicate integration improves peak finding resolution and accuracy.

*To whom correspondence should be addressed.

Table 1. Peak finders compared in this article

Peak Finder	ChIP-Seq (TF)	ChIP-Seq (HM-Punctate)	ChIP-Seq (HM-Broad)	DNase-Seq	Datasets integration	Suitable for IDR
MACS (Zhang *et al*., 2008)	Default	Default	–	–	–	No (score ties, peak widths)
CCAT (PeakRanger) (Xu *et al*., 2010)	–	–	Default	–	–	Caution (strict)
PeakRanger (Feng *et al*., 2011)	Default	Default	–	–	–	Caution (strict)
BCP (Xing *et al*., 2012)	BCP-TF	BCP-HM	BCP-HM	–	–	No (score ties, peak widths)
PeakZilla (Bardet *et al*., 2013)	Default	–	–	–	–	No (score ties)
JAMM	Default	-m narrow	-r region	-f 1	Biological replicates	Yes
DFilter (Kumar *et al*., 2013)	bs = 50 ks = 30 refine nonzero	bs = 100 ks = 100	bs = 100 ks = 30 nonzero	bs = 100 ks = 50 refine	Different experiments	Caution (peak widths)

Notes: TF, transcription factors; HM, histone modifications; Strict, can not *always* adapt to calling a large number of peaks; Peak widths, inaccurate peak widths; Score ties, peak scores have ties. Parameters mentioned here are those used in the article, unless otherwise stated. We varied the MACS -g parameter appropriately for different genomes.

2 JAMM PEAK FINDING METHODS

2.1 Overview

Figure 1 provides an overview of JAMM's analysis steps. Core peak finding steps involve selecting local windows that are enriched over background, followed by clustering the normalized extended-read counts in those windows into a peak cluster and noise cluster(s). Local clustering allows JAMM to adapt to peaks with different widths and signal properties and to accurately determine their boundaries. Furthermore, using clustering as an approach for peak finding extends naturally to multivariate clustering, which is useful for integrating datasets that are correlated but not expected to be exactly the same, such as biological replicates. We chose clustering via multivariate Gaussian mixture models (Banfield and Raftery, 1993), which allows for including information about the covariance of the replicates. Finally, JAMM scores peaks via the peak signal, represented by the geometric mean of the replicates peak signals, and how it compares to background.

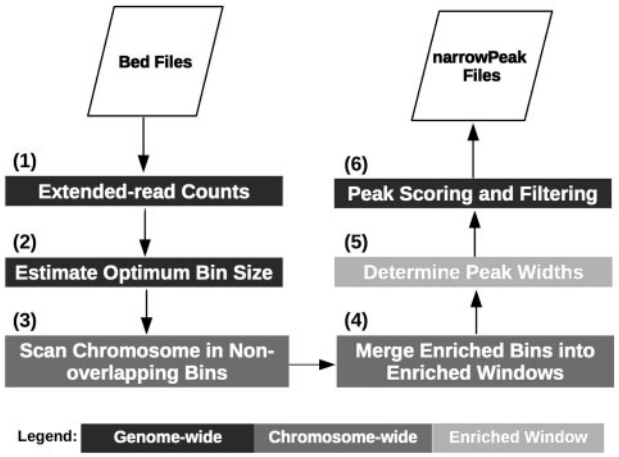

Fig. 1. An overview of JAMM's peak finding steps

2.2 Extended read counts

For ChIP-Seq datasets, JAMM uses cross-correlation analysis to estimate the average fragment length (Ramachandran *et al.*, 2013, see Supplementary Text). (Step 1, Fig. 1) Fragment length is calculated for each replicate, including biological control, separately.

Reads are extended/truncated to the average fragment length in the 5′-to-3′ direction. Extended-read counts at each base pair are divided by the mean extended-read count to produce normalized extended-read counts.

2.3 Enriched windows

JAMM selects enriched windows and then assigns peaks locally in those windows (Steps 2, 3 and 4, Fig. 1). To find enriched windows, JAMM divides the genome into small non-overlapping bins and makes a decision whether each bin is enriched over background. All book-ended enriched bins are merged into larger, non-overlapping, variable-width enriched windows. This approach ensures that enriched windows include entire binding sites and that JAMM can seamlessly adapt to broader enrichment domains. In addition, determining enrichment on the bin-level ensures maximized sensitivity, so that JAMM can easily adapt to reporting a large number of peaks.

Similar to Song and Smith (2011), JAMM selects the bin size Δ that minimizes the cost function $C_n(\Delta)$ (Shimazaki and Shinomoto, 2007):

$$C_n(\Delta) = \frac{2k - v}{(n\Delta)^2},$$

where n is the total number of reads, k is the average number of reads per bin for bins with width Δ and v is the variance (Shimazaki and Shinomoto, 2007). The user can also specify an arbitrary bin size. For multiple replicates, the optimum bin size is calculated separately for each replicate and the smallest bin size is selected.

A bin is enriched over background if

$$\mu_s > \mu_b$$

and

$$SNR_b > SNR_{chr},$$

where μ_s and μ_b are the average normalized extended-read counts in the sample bin and the corresponding background bin, respectively, and SNR_b and SNR_{chr} are the SNR in the bin and the corresponding chromosome, respectively. Any $SNR = \frac{\mu_s}{\sigma_b}$, where μ_s is the average sample normalized extended-read count and σ_b is the standard deviation of the control normalized extended-read count. For multiple replicates, all replicates have to pass this enrichment test for a bin to be considered enriched.

2.4 Peak finding

JAMM assumes that the signal (smoothed extended-read counts, see Supplementary Text) in enriched windows originated from a univariate Gaussian mixture model for single sample analysis or a multivariate Gaussian mixture model when integrating multiple replicates (Step 5, Fig. 1; Banfield and Raftery, 1993; Celeux and Govaert, 1995; Fraley *et al*., 2012):

$$\prod_{t=1}^{T} \sum_{k=1}^{K} w_k \times N_k(bp_t | \mu_k, \Sigma_k),$$

where T is the window size, K is the number of components (clusters), bp_t is the read signal value for base pair t, w_k is the weight of component k in the mixture and μ_k and Σ_k are the vector of means and the covariance matrix for component k, respectively (Fraley *et al*., 2012). The Gaussian mixture model is defined by the set of parameters $\Theta = w_1 ... w_k, \mu_1 ... \mu_k, \Sigma_1 ... \Sigma_k$.

To find accurate peak boundaries in enriched windows, JAMM fits a Gaussian mixture model to cluster the smoothed extended-read count in each enriched window separately, assuming either two mixture components (corresponding to peaks and noise, parameters: *-m normal*) or three mixture components (corresponding to peaks, peak tails and noise, parameters: *-m narrow*). In the univariate case, variance is assumed to be different between different components. In the mutivariate case, the covariance matrix is assumed to be different among different components and is parameterized according to its eigenvalue decomposition (Banfield and Raftery, 1993):

$$\Sigma_k = \Upsilon_k \times \Lambda_k \times \Upsilon_k^{-1}$$

and

$$\Lambda_k = \lambda_k \times A_k,$$

where Υ_k is the orthogonal matrix of the eigenvectors and Λ_k is a diagonal matrix with the eigenvalues at the diagonals, with λ_k being the first eigenvalue in Λ_k and A_k being a diagonal matrix with a vector at the diagonal that is proportional to the vector of eigenvalues. Therefore, Υ_k determines the orientation of the eigenvectors of k, while λ_k defines the volume k occupies in the n-dimensional space and A_k defines the shape of the contour lines (Banfield and Raftery, 1993).

Gaussian mixture clustering starts with chromosome-wide parameter initialization based on an imaginary large window W_I formed by concatenating the top-scoring windows in the chromosome. First, data points are assigned to clusters via k-means. Cluster assignments are then used to initialize an Expectation-Maximization (EM) algorithm to fit a Gaussian mixture model (Celeux and Govaert, 1995) starting with the maximization step. The Expectation step calculates the conditional probability that the read count signal at a given base pair bp_t originated from component k given Θ: $p_k(bp_t)|\Theta$. The Maximization step calculates the maximum likelihood estimates of the model parameters Θ given all $p_1 ... p_k$ for all $bp_1 ... bp_t$.

Assigning bp_t to a cluster k is derived directly from $p_{1...k}(bp_t)$ where bp_t is assigned to the mixture component k that maximizes $p_k(bp_t)$. Different structures of the covariance matrix (namely $\Upsilon_k \lambda_k A_k \Upsilon_k^{-1}$ and $\Upsilon_k \lambda_k A \Upsilon_k^{-1}$) are tested and the one maximizing the Bayesian Information Criteria is chosen (Fraley *et al*., 2012, see Supplementary Text). Formulas to update Θ (Maximization Step) and $p_k(bp_t)$ (Expectation Step) for both models are described by Celeux and Govaert (1995).

The model learned for W_I is used to initialize the EM Gaussian mixture clustering algorithm for every enriched window separately starting with the Expectation step. The mixture component with the highest mean is taken to be the enriched cluster and contiguous base pairs assigned to this cluster are taken to be the peaks. In the multivariate case, all replicates are required to agree on the mixture component mean ordering, otherwise the window is rejected (see Supplementary Text).

2.5 Peak scoring

The background signal in every peak is subtracted from the corresponding sample signal (Step 6, Fig. 1). When analyzing replicates, sample signal is taken to be the per-position geometric mean of the replicates signals. The resulting background-normalized signal values are averaged to produce the mean peak background-normalized signal (μ_{ns}). In addition, JAMM executes the Mann–Whitney U non-parametric test to compare the sample signal (not background normalized) with the corresponding background signal. A Benjamini–Hochberg correction (Benjamini and Hochberg, 1995) is applied to the full list of *P*-values after peak finding is complete. JAMM defines the peak score to be

$$S_p = \mu_{ns} \times -log10(p_{corrected}).$$

By default, each peak is scored and reported as a separate peak (parameters: *-r peak*). All peaks detected in one window can be merged to be scored and reported as one peak (parameters: *-r region*).

2.6 Implementation and output

JAMM is implemented as a bash script with peak finding and scoring implemented by R and Perl scripts. Other post- or preprocessing steps can be added to the pipeline easily if needed. JAMM outputs a sorted peak list in standard narrowPeak format with peak scores, *P*-values, corrected *P*-values and peak summits.

3 RESULTS

3.1 Accuracy and spatial resolution

First, we sought to establish that JAMM achieves a similar or better site detection accuracy compared with other peak finders. Accuracy refers to the extent to which peak finders can determine the correct locations of enriched sites. Because there is no gold standard for benchmarking peak finders (Szalkowski and Schmid, 2011), we analyzed five different ENCODE transcription factor ChIP-Seq datasets, including CTCF-HeLa, CTCF-K562, NRSF-K562, MAX-K562 and SRF-GM12878 (see Supplementary Tables S1 and S5—we refer to those datasets as 'accuracy-benchmark' datasets) using three different benchmarking methods: (i) motif finding precision (fraction of called peaks with motif matches) using FIMO (Grant *et al.*, 2011), which uses a uniform zero-order background model, (ii) maximum cumulative motif likelihood using SpeakerScan (Megraw *et al.*, 2009), which uses a first-order local background model and (iii) accuracy of recovery of manually curated positive peaks as reported by Rye *et al.* (2011) (see Section 5 and Supplementary Text). Regarding motif precision, we found that all peak finders perform comparably although DFilter and JAMM rank better when results are averaged across multiple datasets (Supplementary Table S1). Regarding motif likelihood, we also found that all peak finders, except BCP, perform rather comparably. However, JAMM ranks better than other peak finders when results are averaged across multiple datasets (Supplementary Table S1, Supplementary Figs S1–3). Finally, we found that PeakZilla is the best performing peak finder, followed by JAMM and PeakRanger, in terms of recovering manually curated positive peaks (Supplementary Table S1). When we average the results over all datasets and all benchmarks (a total of nine comparisons, Supplementary Methods), we found that JAMM and PeakRanger are the top ranking programs (Fig. 2a). JAMM ranked first for two benchmarks and third for one benchmark. PeakRanger ranked second for all benchmarks (Supplementary Table 1).

When comparing JAMM-I (JAMM with replicate integration) with JAMM running on pooled replicates (JAMM-P), we found that JAMM-I consistently outperforms JAMM-P (JAMM-P ranked better than JAMM-I in only one out of five datasets where there was a difference), indicating that JAMM's replicate integration improves peak finding accuracy over replicate read pooling. A main contributing factor is JAMM-I's better spatial resolution owing to replicate integration via multivariate mixture model clustering. Figure 2b provides a demonstration of JAMM-I's improvement over replicate pooling. Only JAMM-I can resolve two neighboring CTCF binding sites: the pooled replicate profile obscures the better spatial resolution of Replicate 1 owing to the poorer resolution of Replicate 2.

To further confirm peak finding accuracy, we analyzed datasets used by Bardet *et al.* (2013) via the peak finding precision benchmark (FIMO-based), including Twist, PHA-4, NFKB, CEBPA and Ste12 (we refer to those datasets as 'bardet-benchmark' datasets, see Supplementary Table S6). We found PeakZilla, PeakRanger and JAMM-I to be, on average, the top performing programs (see Supplementary Table S2).

Next, we asked whether peak finders can define accurate enrichment site widths. We found that BCP underestimates peak widths, while DFilter and MACS overestimate peak widths (Fig 2c and d). JAMM, PeakZilla and PeakRanger have accurate peak width determination to a large extent. PeakRanger slightly underestimated the peak widths of some sites with both CTCF and NRSF, while PeakZilla fixes peak widths at twice the estimated fragment length (Bardet *et al.*, 2013) (Supplementary Figs S4 and S5). Additionally, we observed a similar result with DNase-Seq: while JAMM can assign peak boundaries corresponding accurately to variable-width DNase-I-hypersensitive regions, DFilter can not (Supplementary Fig. S6).

Spatial resolution is especially relevant for histone modifications with punctate enrichment patterns. We analyzed peak coverage (see Section 5) of ENCODE HeLa-S3 H3K4me3. H3K4me3 is expected to be maximally enriched immediately

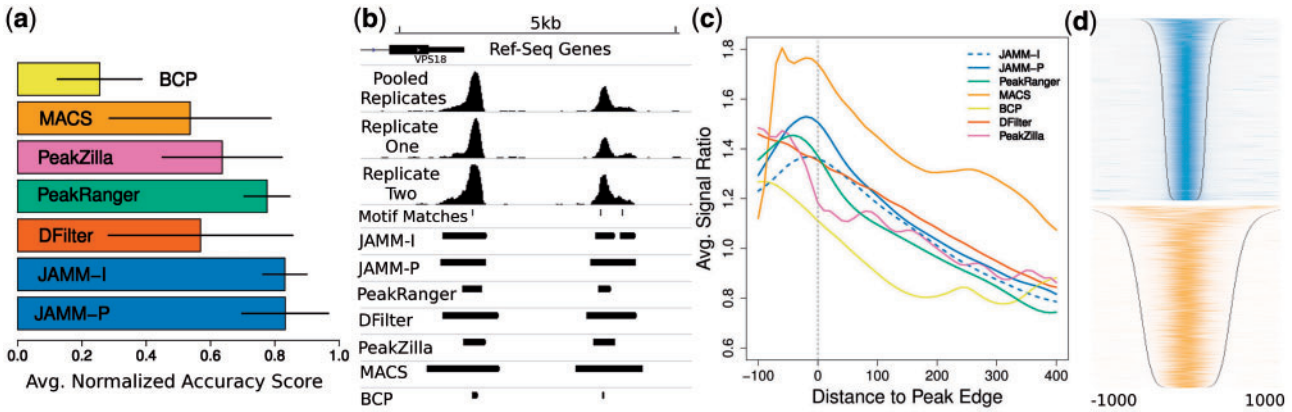

Fig. 2. Transcription Factor Peak Finding. (**a**) Average normalized accuracy score over three benchmarks (see Section 3, Supplementary Text, Supplementary Tables S1 and S5) (**b**) An example of JAMM-I's improved spatial resolution because of replicate integration (CTCF, K562) (**c**) Peak width determination: Average Signal Ratio indicates the ratio of extended-read counts in the 20 bp upstream to that in the 20 bp downstream of the indicated location, averaged over all peaks. Negative numbers on the x-axis indicate locations outside the peak and positive numbers indicate locations inside the peak (CTCF, HeLa-S3). Increased signal ratio outside the peaks indicates peak width underestimation. Increased signal ratio inside the peak indicates peak width overestimation. (**d**) The corresponding heatmaps to (c) for JAMM-P (top) and MACS (bottom). Heatmaps are centered on peak center, ranked by peak width and show extended-read count intensity and the corresponding peak edges (gray squares). See Supplementary Figures S4 and S5 for other peak finders and datasets

upstream and downstream of active transcription start sites (TSSs) (Guenther *et al.*, 2007). Although ChIP-Seq datasets typically have enough resolution to separate the signal upstream of TSSs from the signal downstream (Fig. 3a), many peak finders can not recover this resolution. Out of the peak finders we tested, only JAMM and PeakRanger can, on average, resolve neighboring H3K4me3 peaks, while other peak finders detect, on average, one large peak encompassing multiple enriched sites (Fig. 3b and c, see also Fig. 5b).

3.2 Broad enrichment patterns

Peak finders designed to process punctate enrichment sites are typically not able to capture broad enrichment domains, often because those domains feature relatively low SNR and stretch over thousands of base pairs. H3K27me3 and other histone modifications display broad enrichment patterns when assayed with ChIP-Seq. We tested CCAT (Xu *et al.*, 2010) (via PeakRanger's implementation of the CCAT algorithm) and BCP (Xing *et al.*, 2012) (both designed for broad enrichment domains) as well as DFilter and JAMM (both universal peak finders) in terms of their ability to capture broad enrichment patterns. We found that BCP assigns the most broad peaks. JAMM also assigns broad peaks but generally smaller than those called by BCP. DFilter and PeakRanger's CCAT are less suited for defining broader enrichment domains (Fig. 4).

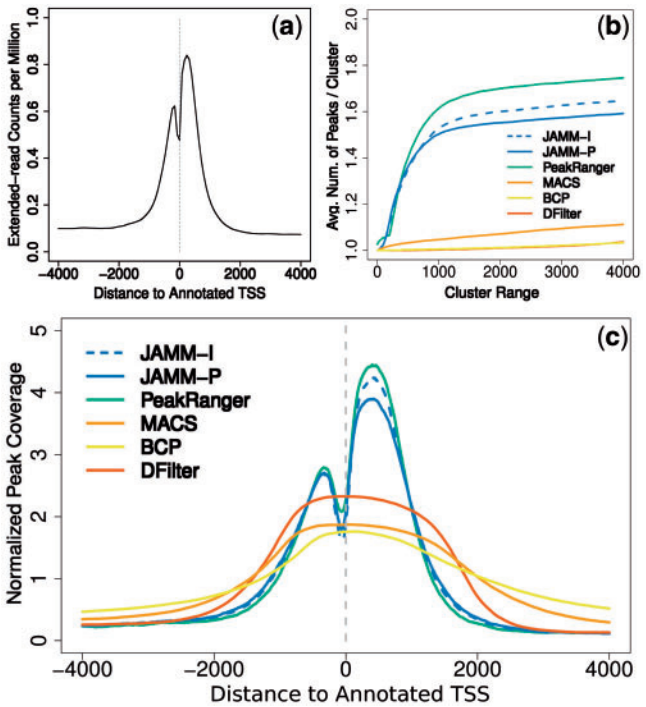

Fig. 3. Resolving Punctate Histone Modification Peaks (HeLa-S3 H3K4me3). Only JAMM and PeakRanger can recover the resolution of the dataset (**a**) in the peaks called (**b**) and (**c**). (b) Shows the average number of peaks per cluster at different cluster ranges. Cluster range is the maximum distance separating peaks in the same cluster (for example, if two peaks are 50 bp apart, they will be grouped together in one cluster if cluster range is 50 bp or more). See Supplementary Figures S8 and S9 for TSS peak coverage of other HeLa-S3 histone modification datasets

3.3 Peak scoring and sorting

JAMM can typically report a large number of peaks and relies on its peak scoring to robustly rank the reported peaks (see Section 5). This facilitates downstream analysis and gives users more flexibility in choosing a method to filter the peaks. Irreproducible Discovery Rate (IDR) (Li *et al.*, 2011) is an ENCODE-recommended method for filtering peak calls based on replicate reproducibility (Landt *et al.*, 2012). Briefly, the IDR pipeline involves calling peaks on the replicates separately, followed by applying the IDR statistical model to determine the number of reproducible peaks n given a certain IDR threshold. Peak reproducibility involves whether the peaks overlap and how their ranks compare in the replicates peak lists. Finally, peaks are called on the combined replicates and the top n peaks are taken to be the high-confidence reproducible peaks (see ChIP-Seq IDR Web page for more information: https://sites.google.com/site/anshulkundaje/projects/idr).

We applied the IDR analysis pipeline to HeLa-S3 CTCF ChIP-Seq ENCODE dataset. We found that sorting the peaks using JAMM's peak scores produces a clear phase shift between reproducible peaks and irreproducible peaks (Fig. 5a). To call peaks jointly on biological replicates (the final step in the IDR pipeline), aligned reads are usually pooled before peak finding, but pooling alignments obscures the differential signal intensities of the replicates, and, therefore, may lead to invalid peak sorting. Figure 5b shows an example of cases where JAMM-I's integrated sorting of peaks provides a more valid peak sort than sorting peaks called on pooled alignments. Pooling the replicates obscured the spatial information regarding peak 'b' (JAMM) as the two replicates do not agree on a specific peak location. JAMM relies on geometric averaging of replicates peak signal to score the peaks, which leads to more valid peak scores than those based on read pooling.

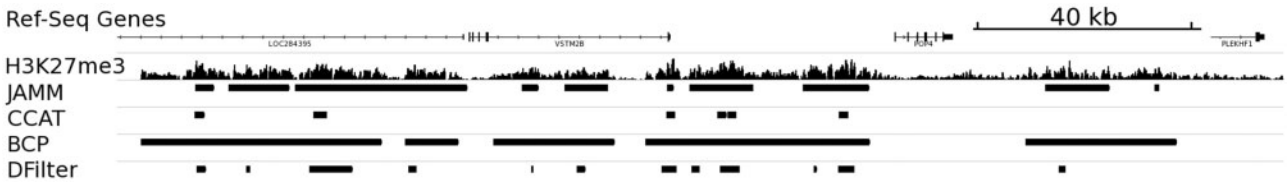

Fig. 4. ENCODE HeLa-S3 H3K27me3 (Pooled Replicates). Overview of peak calls, region shown is from chromosome 19. See also Supplementary Figure S7

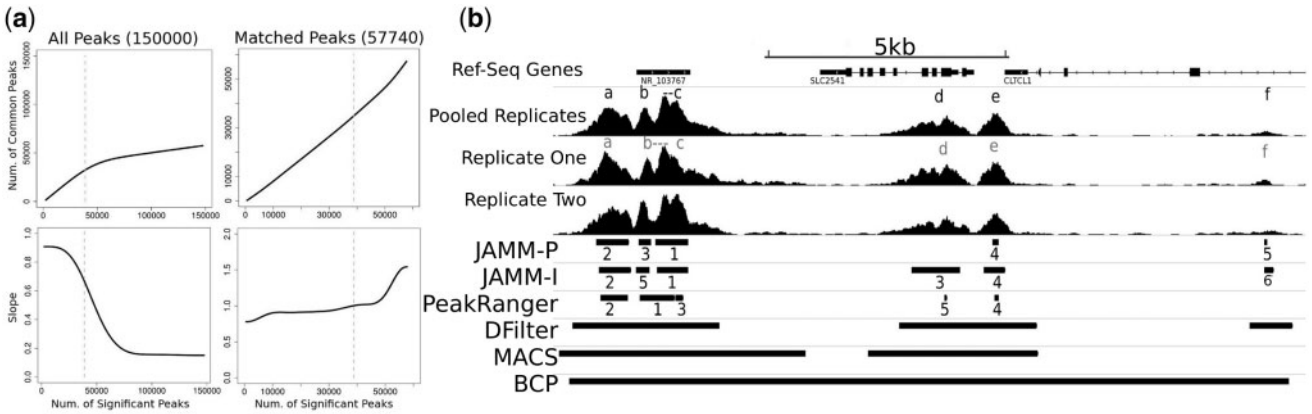

Fig. 5. JAMM's Peak Scoring. (**a**) Results for IDR analysis on biological replicates for ENCODE HeLa-S3 CTCF using JAMM. Dashed vertical line corresponds to the number of peaks selected with an IDR threshold of 0.02 (38 853 peaks). Recommendations for setting IDR thresholds are available on the IDR webpage. Input to the IDR pipeline included the top 150 000 peaks called by JAMM. The number of matched peaks increases as one descends through the sorted peak list up to the point where peaks become irreproducible between replicates. (**b**) ENCODE H3K4me3 HeLa-S3 ChIP-Seq. Black (JAMM) and gray (PeakRanger) letter labels indicate the peaks called. Numeric labels indicate peaks' relative rankings as defined by each peak finder

Taken together, JAMM provides a plausible approach to replicate integration that is widely applicable to different types of datasets and protocols. The analysis pipeline would start with peak calling on the replicates separately, followed by IDR analysis to select n (the number of reproducible peaks). Finally, peaks are called on the replicates jointly via JAMM's replicate integration and the top-scoring n peaks are taken as a highly confident set.

4 DISCUSSION

A desirable property in universal peak finders is detecting, and correctly determining the widths of, enrichment sites with different signal properties. JAMM fits a Gaussian mixture model for every local enriched window separately and only fixes the structure of the covariance matrix (see Section 2 and Supplementary Text). Therefore, JAMM can accurately determine widths of enrichment sites that have different signal properties, even if in the same dataset. Some peak finders start with learning an expected peak shape (Bardet *et al.*, 2013; Kumar *et al.*, 2013), making it more difficult to detect enrichment sites with varying widths or to assign their boundaries accurately. Other peak finders adapt specialized subroutines for refining peak widths after peak finding is complete (Rashid *et al.*, 2011). In some cases, this approach may be able to assign accurate peak boundaries. But when the original peak represents several closely spaced sites, this approach may result in choosing one site and missing the others (see Supplementary Fig. S5 in Rashid *et al.* (2011) for an example). We showed that JAMM's local clustering also avoids this caveat and can correctly resolve neighboring punctate sites, similar to programs specifically designed with this task in mind like PeakRanger (Feng *et al.*, 2011).

JAMM is a universal peak finder that can analyze different types of datasets with little change, if any, to the underlying method. This demonstrates that finding enriched sites in read-density–based NGS datasets is essentially the same task regardless of the sites' signal properties. Therefore, we propose that more attention could be directed toward developing universal peak finding solutions, refining preprocessing of read counts to correct for different biases (Hashimoto *et al.*, 2014) and toward developing solutions for biological replicates integration (Yang *et al.*, 2014).

Pooling reads from biological replicates before peak finding is part of the ENCODE consortium recommended guidelines (Landt *et al.*, 2012). However, when peaks are called on pooled replicates, the differential intensities and differential spatial coverage of the replicates are obscured. JAMM addresses replicate integration by looking at biological replicates as not being exactly reproducible and attempts to model their variability using information about their covariance in local enriched windows. Using various accuracy benchmarks, we demonstrated that this approach results in better peak finding accuracy over read pooling.

For peak scoring on replicates, JAMM uses the geometric average of the replicates peak signal. We demonstrated that this approach improves peak sorting. Additionally, we also show that peak finding on the geometric mean of separately normalized replicate signal profiles can improve peak finding accuracy over read pooling similarly to JAMM-I (see JAMM-G in Supplementary Text Section 1.2 and Supplementary Table S3). Geometric averaging of normalized signal profiles can potentially be implemented as a preprocessing step irrespective of the specific peak finding method. Therefore, although it may not be an optimal solution with increasing replicates variability, it is a plausible approach that other peak finders could easily implement for biological replicates analysis, without requiring a multivariate clustering framework.

Accuracy benchmarks are independent of read count densities, as opposed to peaks per cluster (Fig. 3) and peak width accuracy (Fig. 2c and d). However, motif content benchmarks do not represent a definite gold standard because of our incomplete understanding of protein–DNA interactions and potential biases in the benchmarking methods (Szalkowski and Schmid, 2011). We attempted to remedy this by using two different motif

scanning algorithms and by including a manually curated set of peak calls as an additional benchmark (Rye *et al.*, 2011). But manual curation may also be biased because the manually curated set represents only a small fraction of the peaks present in a dataset (345 peaks for MAX, 235 for NRSF and 198 for SRF), and because some peak finders (like PeakZilla) use peak detection methods similar to curation criteria (Bardet *et al.*, 2013; Rye *et al.*, 2011).

Many peak finders ignore being able to report a larger number of peaks and/or ignore providing appropriate peak scores (Table 1), both required criteria for assessing replicate reproducibility via IDR analysis (Li *et al.*, 2011). Appropriate peak scores would have few or no ties and represent the confidence in the peak accurately based on its read density and how it compares with background or biological control. JAMM can typically determine a large number of peaks, and it also provides robust peak scores with few score ties if any. This, in addition to its accurate peak width determination, makes JAMM potentially more applicable for different types of downstream analyses that rely on ranked peak lists.

Although a multivariate clustering framework can potentially be used for differential peak finding, JAMM can not find differential peaks across multiple conditions in its current implementation. Also, JAMM does not take into account mappability, GC content and Copy Number Variations (CNVs). CNVs are especially relevant for cancer cell lines (Pickrell *et al.*, 2011), while GC content bias is a known problem in high-throughput sequencing libraries, probably due to PCR amplification (Benjamini and Speed, 2012). We could not detect CNV bias in JAMM's peak calls in regions of loss when compared with a peak finder that corrects for CNVs (Ashoor *et al.*, 2013), but we noticed a possible increase in the proportion of peaks called by JAMM in regions of gain (see Supplementary Table S4). Explicit implementation of GC content bias and CNV correction could improve peak finding accuracy (Ashoor *et al.*, 2013; Rashid *et al.*, 2011), and we plan to incorporate appropriate correction subroutines in the near future. Finally, JAMM is typically slower than other peak finders with less complicated models, taking ~6–7 h on average to analyze a typical human ENCODE ChIP-Seq dataset when run using a single processor.

5 METHODS

5.1 Datasets, preprocessing and accuracy analysis

All "accuracy-benchmark" datasets are ENCODE datasets. (Bernstein *et al.*, 2012). 'bardet-benchmark' datasets were produced by Bardet *et al.* (2013) and He *et al.* (2011) for Twist, Schmidt *et al.* (2010) for CEBPA, Zhong *et al.* (2010) for PHA-4 and Kasowski *et al.* (2010) for NFKB and Zheng *et al.* (2010) for Ste12. CTCF-NHEK (used in JAMM-G analysis), H3K4me3, H3K27ac, H3K27me3 and DNase-Seq are ENCODE datasets (Bernstein *et al.*, 2012). Fastq files were aligned to the respective genomes using Bowtie2 (Langmead and Salzberg, 2012) (hg19: CTCF-HeLa, CTCF-K562, CTCF-NHEK, NFKB, H3K4me3, H3K27ac, H3K27me3 - mm9: CEBPA - dm3: Twist - ce6: PHA-4). Alternatively, we started with the alignments provided by Rye *et al.* (2011) (MAX, NRSF, SRF), Zheng *et al.* (2010) (Ste12) and ENCODE (DNase-Seq). PCR duplicates were removed using SAMTools (Li *et al.*, 2009). See Supplementary Text, section 1.1.

Transcription factor motifs were obtained from JASPAR (Mathelier *et al.*, 2014) (NRSF: MA0138.2, NFKB: MA0105.1, CEBPA: MA0102.2, PHA-4: MA0047.1, Ste12: MA0393.1), FlyReg (Bergman *et al.*, 2005) (Twist) and Schmidt *et al.* (2012) (CTCF). Motif precision analysis was done using FIMO (*P*-value, 'accuracy-benchmark': 0.0001/ *P*-value,'bardet-benchmark': 0.001) (Grant *et al.*, 2011), cumulative log likelihood analysis was done using SpeakerScan (background window: 150 bp) (Megraw *et al.*, 2009) and the curated peak set provided by Rye *et al.* (2011) was used for manual curation analysis. Manual curation results were defined as the number of peaks that intersect at least one manually curated positive peak after subtracting the number of peaks that intersect manually curated negative peaks exclusively. Results for 'accuracy-benchmark' transcription factor datasets were 0–1 scaled and averaged over all three benchmarks to produce Fig. 2a. See Supplementary Text, section 1.3.

5.2 Visualization

Extended Reads per Kilobase per Million reads mapped (RPKM)-normalized read counts were produced using deepTools (Ramírez *et al.*, 2014) at 10 bp resolution and visualized in IGV browser (Thorvaldsdóttir *et al.*, 2013).

Read coverage heatmaps were produced for peak regions (1000 bp in each direction centered at the peak center) using smoothed extended-read counts at 10 bp resolution.

For peak coverage plots of histone modification peak calls, we intersected each set of peak calls with annotated promoter regions from UCSC hg19 known genes (4000 bp in each direction centered at the annotated TSS), using BEDTools (Quinlan and Hall, 2010). Each base pair was assigned a score of 1 for each intersecting peak. Per-base pair scores were summed then divided by the mean per-base pair score to produce normalized peak coverage scores. Raw extended-read coverage was produced using ngs.plot (Shen *et al.*, 2014) on the same TSS regions.

See Supplementary Text, section 1.3 for more details.

Funding: MMI was supported by the Max-Delbrück-Center/New York University Exchange Program.

Conflict of interest: none declared.

REFERENCES

Ashoor,H. *et al.* (2013) HMCan: a method for detecting chromatin modifications in cancer samples using ChIP-seq data. *Bioinformatics*, **29**, 2979–2986.

Banfield,J.D. and Raftery,A.E. (1993) Model-based gaussian and non-gaussian clustering. *Bio-metrics*, **49**, 803–21.

Bardet,A.F. *et al.* (2013) Identification of transcription factor binding sites from ChIP-seq data at high resolution. *Bioinformatics*, **29**, 2705–2713.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B*, **57**, 283–300.

Benjamini,Y. and Speed,T.P. (2012) Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acid Res.*, **40**, e72.

Bergman,C.M. *et al.* (2005) Drosophila DNase I footprint database: a systematic genome annotation of transcription factor binding sites in the fruitfly, *Drosophila melanogaster*. *Bioinformatics*, **21**, 1747–1749.

Bernstein,B.E. *et al.* (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.

Celeux,G. and Govaert,G. (1995) Gaussian parsimonious clustering models. *Biometrics*, **28**, 781–793.

Crawford,G.E. *et al.* (2006) Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). *Genome Res.*, **1**, 123–131.

Feng,X. *et al.* (2011) PeakRanger: a cloud-enabled peak caller for ChIP-seq data. *BMC Bioinformatics*, **12**, 139.

Fraley,C. *et al.* (2012) *MCLUST Version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation,* Technical Report no. 597, Department of Statistics, University of Washington, June 2012.

Grant,C.E. *et al.* (2011) FIMO: scanning for occurrences of a given motif. *Bioinformatics*, **27**, 1017–1018.

Guenther,M.G. *et al.* (2007) A chromatin landmark and transcription initiation at most promoters in human cells. *Cell*, **130**, 77–88.
Hashimoto,T.B. *et al.* (2014) Universal count correction for high-throughput sequencing. *PLoS Comput. Biol.*, **10**, e1003494.
He,Q. *et al.* (2011) High conservation of transcription factor binding and evidence for combinatorial regulation across six Drosophila species. *Nat. Genet.*, **43**, 414–420.
Johnson,D.S. *et al.* (2007) Genome-wide mapping of in vivo protein-DNA interactions. *Science*, **316**, 1497–1502.
Kasowski,M. *et al.* (2010) Variation in transcription factor binding among humans. *Science*, **328**, 232–235.
Kumar,V. *et al.* (2013) Uniform, optimal signal processing of mapped deep-sequencing data. *Nat. Biotechnol.*, **31**, 615–622.
Landt,S.G. *et al.* (2012) ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.*, **22**, 1813–1831.
Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–359.
Li,H. *et al.* (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
Li,Q. *et al.* (2011) Measuring reproducibility of high-throughput experiments. *Ann. Appl. Stat.*, **5**, 1699–2264.
Li,Y. *et al.* (2014) T-KDE: a method for genome-wide identification of constitutive protein binding sites from multiple ChIP-seq data sets. *BMC Genomics*, **15**, 27.
Liu,B. *et al.* (2013) QChIPat: a quantitative method to identify distinct binding patterns for two biological ChIP-seq samples in different experimental conditions. *BMC Genomics*, **14** (**Suppl. 8**), S3.
Mathelier,A. *et al.* (2014) JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, **42**, D142–D147.
Megraw,M. *et al.* (2009) A transcription factor affinity-based code for mammalian transcription initiation. *Genome Res.*, **19**, 644–656.
Natarajan,A. *et al.* (2012) Predicting cell-type-specific gene expression from regions of open chromatin. *Genome Res.*, **22**, 1711–1722.
Pepke,S. *et al.* (2009) Computation for ChIP-seq and RNA-seq studies. *Nat. Methods*, **6** (**Suppl. 11**), S22–S32.
Pickrell,J.K. *et al.* (2011) False positive peaks in ChIP-seq and other sequencing-based functional assays caused by unannotated high copy number regions. *Bioinformatics*, **27**, 2144–2146.
Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.
Ramachandran,P. *et al.* (2013) MaSC: mappability-sensitive cross-correlation for estimating mean fragment length of single-end short-read sequencing data. *Bioinformatics*, **29**, 444–450.
Ramírez,F. *et al.* (2014) deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Res.*, **42**, W187–W191.
Rashid,N.U. *et al.* (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.
Rye,M.B. *et al.* (2011) A manually curated ChIP-seq benchmark demonstrates room for improvement in current peak-finder programs. *Nucleic Acids Res.*, **39**, e25.
Schmidt,D. *et al.* (2010) Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science*, **328**, 1036–1040.
Schmidt,D. *et al.* (2012) Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. *Cell*, **148**, 335–348.
Schweikert,C. *et al.* (2012) Combining multiple ChIP-seq peak detection systems using combinatorial fusion. *BMC genomics*, **13** (**Suppl. 8**), S12.
Shen,L. *et al.* (2013) diffReps: detecting differential chromatin modification sites from ChIP-seq data with biological replicates. *PloS One*, **8**, e65598.
Shen,L. *et al.* (2014) ngs.plot: Quick mining and visualization of next-generation sequencing data by integrating genomic databases. *BMC Genomics*, **15**, 284.
Shimazaki,H. and Shinomoto,S. (2007) A method for selecting the bin size of a time histogram. *Neural Comput.*, **19**, 1503–1527.
Song,Q. and Smith,A.D. (2011) Identifying dispersed epigenomic domains from ChIP-Seq data. *Bioinformatics*, **27**, 870–871.
Szalkowski,A.M. and Schmid,C.D. (2011) Rapid innovation in ChIP-seq peak-calling algorithms is outdistancing benchmarking efforts. *Brief. Bioinform.*, **12**, 626–633.
Thorvaldsdóttir,H. *et al.* (2013) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.*, **14**, 178–192.
Tuteja,G. *et al.* (2009) Extracting transcription factor targets from ChIP-Seq data. *Nucleic Acids Res.*, **37**, e113.
Xing,H. *et al.* (2012) Genome-wide localization of protein-DNA binding and histone modification by a Bayesian change-point method with ChIP-seq data. *PLoS Comput. Biol.*, **8**, e1002613.
Xu,H. *et al.* (2010) A signal-noise model for significance analysis of ChIP-seq with negative control. *Bioinformatics*, **26**, 1199–1204.
Yang,Y. *et al.* (2014) Leveraging biological replicates to improve analysis in ChIP-seq experiments. *Comput. Struct. Biotechnol. J.*, **9**, e201401002.
Zeng,X. *et al.* (2013) jMOSAiCS: joint analysis of multiple ChIP-seq datasets. *Genome Biol.*, **14**, R38.
Zhang,Y. *et al.* (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol.*, **9**, R137.
Zheng,W. *et al.* (2010) Genetic analysis of variation in transcription factor binding in yeast. *Nature*, **464**, 1187–1191.
Zhong,M. *et al.* (2010) Genome-wide identification of binding sites defines distinct functions for *Caenorhabditis elegans* PHA-4/FOXA in development and environmental response. *PLoS Genet.*, **6**, e1000848.